University of Catania

Department of Physics and Astronomy
PhD in Complex Systems for Physical, Socio-economic and Life Sciences

Thesis For The Degree Of Doctor Of Philosophy

Models and Tools for the analysis of genomic data in clinical practice

Supervisor
Prof. Alfredo Pulvirenti

Correlatore
Dott. Stefano Forte

Candidate
Grete Francesca Privitera

Academic Year MMXIX-MMXXI (XXXIV cycle)

"I have no special talents. I am only passionately curious." [Albert Einstein]

Abstract

Precision Medicine is growing every day as a key methodology to decide which therapy is the best choice, in terms of efficacy and adverse effects, for each patients. The Personalized Medicine much rely on molecular data produced by Next Generation Sequencing (NGS). The extraction of knowledge from such kind of data makes use of Bioinformatics. Unfortunately, bioinformatics analysis results quite uncomfortable and difficult for non expert users. For this reason, pipelines and tools have been created to easily analyze cancer data. However, most of them are still not suitable for clinicians and life scientists in general. To going beyond such limitation in this thesis we will introduce the following tools: (i) Oncoreport, a system for the analysis of NGS data in clinical context; (ii) TMBCalc a pipeline for the construction of a light weight targeted gene panel for immunoterapy; (iii) a tool for variant prioritization based on machine learning techniques. The results presented here provide an advances in the actual usage of NGS data in clinical setting.

Keywords: NGS, Tumor, Report, Precision Medicine, Immunotherapy, Genome, Microbiome, Tools.

Contents

List of Figures

List of Tables

Nomenclature

AF Allele Fraction

CTLA-4 Cytotoxic T-Lymphocyte Antigen number 4

DEGs Differential Expression Genes

DP Depth

GLM Generalized Linear Model

HBV Hepatitis B virus

HCV Hepatitis C virus

HIV Human Immunodeficiency Virus

HLA Human Leukocyte Antigen

HNPCC Hereditary Non-Polyposic Colorectal Cancer

HPV16 Human Papillomavirus 16

HR Hazard Ratio

HRV16 Human Rhinovirus 16

HSV1 Human alphaherpesvirus 1

LCA Last Common Ancestor

MDSCs Myeloid-Derived Suppressor Cells

MHC Major Histocompatibility Complex

MMR MisMatch Repair

MSI Microsatellite Instability

NGS Next Generation Sequencing

OMIM Online Mendelian Inheritance in Man

ONT Oxford Nanopore sequencing

PacBio Pacific Biosciences

PD-L1 Programmed Death Ligand-1

PPV Positive Predicted Value

RPART Recursive Partitioning

SMRT Single-Molecule Real-Time sequencing

TAA Tumor-Associated Antigens

TAMs Tumor-Associated Macrophages

TCGA The cancer genome atlas

TIL Tumor-Infiltrating lymphocyte

TMB Tumor Mutational Burden

TME Tumor Microenviroment

TSA Tumor-Specific Antigens

VUS Variant of Uncertain Significance

WES/WXS Whole Exome Sequencing

WGS Whole Genome Sequencing

WHO World Health Organization

Aim of the project

The aim of my PhD research project was to improve technologically and methodologically the study of cancer to enhance the cancer therapies setting. A particular focus was put on developing software to help to reach this goal. In this thesis I expose several results. First of all, a tool called OncoReport used for the study of NGS results of patients for the development of a personalized therapy. The idea under this software development was to help laboratories and clinicians in the fast comprehension of big data deriving from sequencing. Secondly, a tool called TMBCalc implemented for the calculation of the TMB through a docker container. This pipeline was developed jointly with a small genes signature for the research of the TMB in all type of cancers to try to harmonize this methodology. Next, an algorithm called VarPrAl built to prioritize variants thanks to survival curves and pathogenity score. This in order to help also in the therapies construction focusing on those mutations that are believed to worsen the patient situation. At the end, I show a literature review of several bioinformatic tools for the research of viruses in RNAseq of patients. The oncovirus study is fundamental since they can provoke cancer due to the mutations that they can cause inside the human genome. A summary of that is showed in Fig 1. The results displayed here have been reached thanks to the collaboration with the IOM Ricerca and the International Agency for Research on Cancer (IARC) of Lyon.

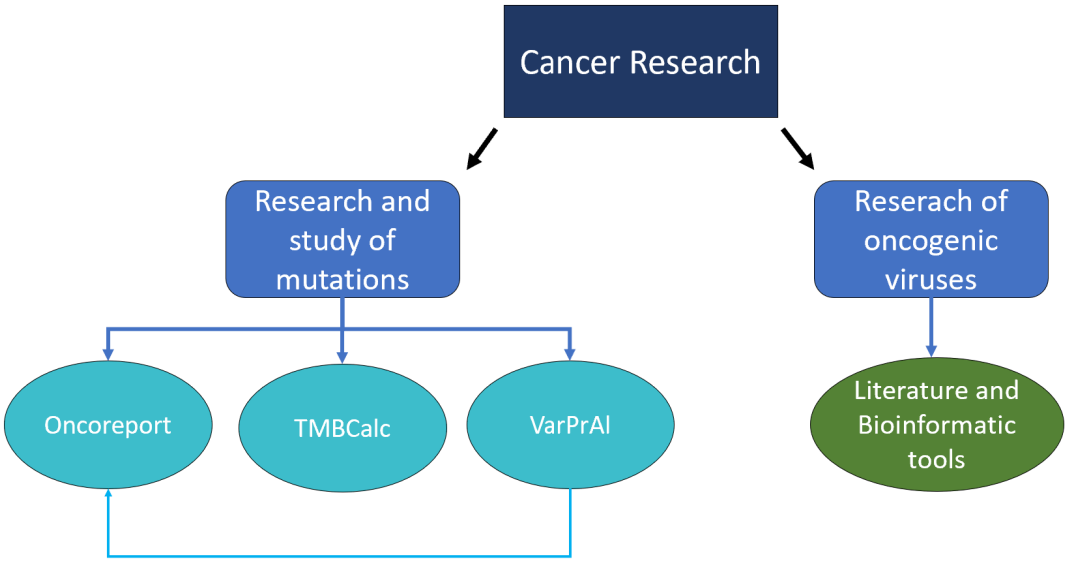

Figure 1: Aim of the project in the cancer research

Chapter 1

Introduction

1.1 Cancer, Oncogenes and Oncosuppressors

Cancer is one of the leading diseases in the world. The cancer cases rise as the population become older and adopt behaviors that increase cancer risk. According to the World Health Organization (WHO) in 2020 near 10 million people died of cancer in the world. The most common cases of cancer can be found in breast, lung, colon and rectum, prostate, skin and stomach. While the most common causes of death are lung, colon and rectum, liver, stomach and breast Fig. 1.1.

Cancer arises because of the transformation of normal cells in tumor cells as a cause of mutations in human body especially due to risk factors such as tobacco, air pollution, unhealthy diet, lack of physical activity, radiation, chemical carcinogens, etc. A mutation is an alteration of the DNA which can bring to different types of changes. The ones of our interest are those that occur on oncosuppresors or oncogenes. The genes involved in cancer are commonly divided in oncogenes and oncosuppressors. The former genes when activated are involved in tumor transformation while the latter ones lead to cellular proliferation when inactivated. As a matter of fact, oncogenes are

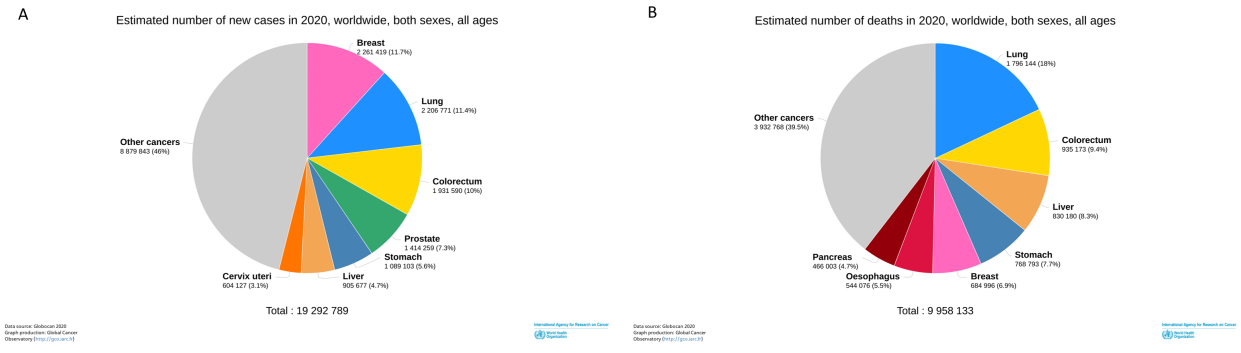

Figure 1.1: A. Number of new cancer cases in 2020 worldwide. B. Estimated number of cancer deaths in 2020 worldwide.

genes involved in the cellular cycle whereas oncosuppressors are normally designated to the cellular apoptosis. The mutations in these genes can be both inherited or, as said before, acquired by DNA replication errors or/and for exposure to carcinogens. In particular, an oncogene is a gene that, when over-expressed or mutated, has the potential to cause cancer. Normal oncogene are indicated as proto-oncogenes since they are involved in cell growth, proliferation or inhibition of apoptosis. Their mutations cause the cells survival even if cells are altered or malfunctioning. The first oncogene discovered was src in chicken rous sarcome virus. After this one 40 highly oncogenic retrovirus have been discovered in animals. These retrovirus contain oncogenes responsible for cells transformation, but not for virus replication. These oncogenes derive from host cells proto-oncogene, that are cells regulatory genes which control cells proliferation, that have been incorporated in viral genome. However, not all oncogenes are derived from retrovirus. Retrovirus indeed might cause cancer especially in animals, the human oncogenes, most of the time, derive from mutation or over-expression. Examples of other oncogenes are RAS; WNT, MYC and ERK. For instance, MYC codes for transcription factors that are produced at higher rates when it is mutated [11]. Instead, tumor suppressor genes or oncosuppressors are genes that are inactivated in tumor development. These genes normally inhibit cellular proliferation. With a loss or reduction in their function the negative control is absent and the cells proliferate out of control. The first tumor suppressor gene, Rb, was identified in retinoblastoma. The second and most important tumor suppressor identified is p53, it can be found inactivated in a wide of human cancers such ad leukemia, brain tumors, colon cancer and so on. Most of the time oncosuppressors and oncogenes work together in tumor proliferation [12]. The identification of mutations that bring to the tumor raise is essential to establish anti-tumoral therapies. In order to identify these mutations we need to apply first of all a biopsy and after that a Next Generation Sequencing (NGS) analysis.

1.1.1 Biopsy and Precision medicine

Biopsies A Biopsy is a medical exam consisting in the collection of a portion of tissue to analyze it. It is done to confirm a suspect of a disease like cancer. The biopsies in cancer allow the histological recognition of the disease and, thanks to the sequencing techniques, also the study of the genetic profile of the tumor. It is now under study the possibility to use multiple biopsy for tumor since tumors show an extensive heterogeneity. The classic solid biopsy is difficult to obtain because of the pain that costs to the patient and the clinical risk, more or less dangerous depending on the site of the tumor. In the last years is more and more common to use liquid biopsy, but the study of the genomic composition of the tumor with this it is still under study. The liquid biopsy consists in the collection of a blood sample to search for circulating tumor DNA (ctDNA). This biopsy is

possible because tumor DNA enters into the circulation through the apoptosis of the cells. ctDNA is higher in sick patients because in normal condition phagocytes clear the apoptotic debris [13, 14] but in tumor the clearance is not efficient [15]. Furthermore, it is also believed that there is a release of DNA fragments into the circulation fixed and circulating tumor cells (CTCs) [16]. ctDNA is obtained from serum or plasma deriving from blood. Plasma is collected putting blood in a tube treated with an anticoagulant and then removing cells by centrifugation. Serum is collected after the blood form clot and after centrifugation. The ctDNA is then extracted through a commercial kit [17].

Tumoral Biomarkers A biomarker is "A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention" [18]. Find new biomarkers is a critical step in scientific research, especially in cancer. New biomarkers could lead to the discovery of new treatments or new prevention methods. However, it takes time to approve their clinical use and most of the time researching them is difficult and expensive. A biomarker can be diagnostic, so it confirms the presence of the disease of interest. In particular, it can be an early biomarker useful to prevent the disease. Another kind of biomarker is the pharmacodynamic/response biomarker. It is a biomarker of which the level changes in response to exposure to a medical product. Examples of these biomarkers are the genes for which we can associate a drug. A predictive biomarker is a biomarker that identifies an association with the effect of an intervention, it identifies the patient that will respond or not to a therapy. A prognostic biomarker is a biomarker used to identify the likelihood of a clinical event, disease recurrence or disease progression. This is used after there has been already a diagnosticated disease and it is associated with differential disease outcomes. The susceptibility/risk biomarker instead indicates the potential of developing a disease in a patient without clinically apparent medical conditions. A biomarker to be good should not only be correlated with a clinical outcome but it should also explain the change in the clinical outcome [19].

Precision medicine Nowadays, thanks to the identification of specific genomics alterations due to the biopsy and the follow high-throughput analysis we are able to identify targeted therapies. We are focusing our attention on precision and personalized medicine. Precision medicine tries to identify which therapy is the most suitable for a specific patient considering lifestyle, cancer staging and biological characteristics. Personalized medicine is another face of the same coin because, while precision medicine relies only on data and information, it refers to the genetic of the patient focusing on attitudes, knowledge and social context. These personalized therapies are necessary because each cancer type indeed underlies a huge heterogeneity, the clinicians analyse patient biological

features and alteration to make treatment decisions. The National Research Council's Toward Precision Medicine adopted the definition of precision medicine in 2008 from the President's Council of Advisors on Science and Technology. The Definition is "The tailoring of medical treatment to the individual characteristics of each patient...to classify individuals into sub-populations that differ in their susceptibility to a particular disease or their response to a specific treatment. Preventative or therapeutic interventions can then be concentrated on those who will benefit, soaring expense and side effects for those who will not". Precision medicine is used to guide health care decision to the most effective treatment for a specif patient in that specific time, improving in this way health-care quality. This permit to avoid useless diagnostic test and possibly threatening therapies for the patient. The use of precision medicine is possible thanks to the discover of biomarkers, like the markers for efficacy or for adverse effects of a therapy. The discover of the biomarkers themselves and the possibility to approach to precision medicine is due to the advance in these last years of NGS. This technology have made researchers able to analyze thoroughly cancer genome profile using Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES or WXS) or target sequencing or even RNA-Seq. Precision medicine is also used to make a prognosis, prevent and diagnose the disease, not only for treating it [20]. Its development has been helped by the use of the multiomics technology through the production of Big Data. The aim of precision medicine is to develop a specific treatment for each individual and for his condition to maximize the power of that therapy and minimize adverse reactions [21]. The hope is that in few years the sequencing of tumor will be a routine exam for patients, in this way we will be able to increase patients' quality of life and life expectancy.

1.2 Human genome Sequencing

The human genome sequencing has been finished in draft form in 2001 [22, 23] and definitively in 2003 [24] after 13 years of working thanks to the Sanger DNA sequencing technology. This technique was very expensive and permitted only the sequencing of small parts of the genome per time, this brought to a cost between 0.5 and 1 billion dollars. In those years in fact the personal genome sequencing was unthinkable. After the release of the completed human genome in 2004 the National Human Genome Research Institute (NGHRI) created a 70 million dollar DNA sequencing technology initiative setting the goal toward a cost per genome of one thousand dollars [25]. The decrease was fast, in 2006 the cost of a human genome sequencing was between 20-25 million dollars. This price had a significant decrease with the advent of NGS sequencing. Indeed, in 2020 it was under 1 thousand dollars Fig.1.2. Though, all these costs do not consider the one of the clinical interpretation of the data. [26]

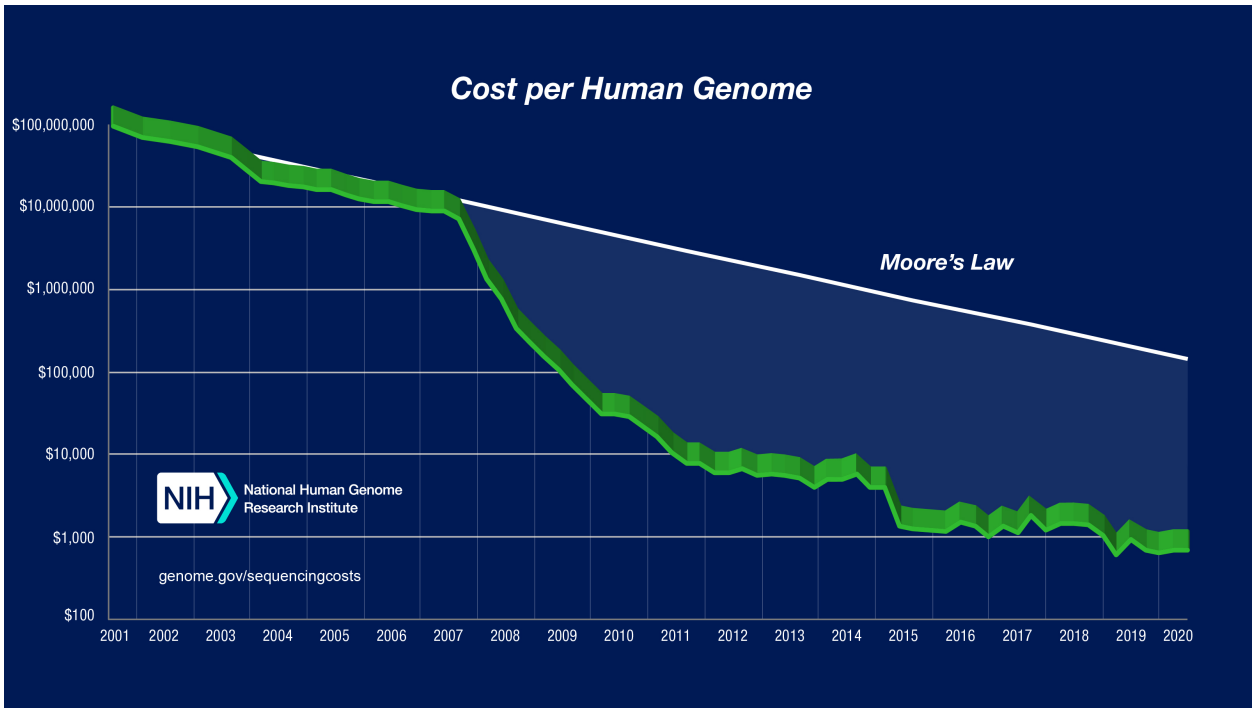

Figure 1.2: Human Genome sequencing cost from 2001 to 2020

The DNA sequencing is the process that is need to determine each bases of the DNA. It is now fundamental because the comparison between the genome reference with a new sequenced genome can help in the diagnosis of diseases. The sequencing is used also for forensic analysis, virology, metagenomics and so on. Starting from 1970 to today we can find three generations of sequencing methods, each of them created to fasten the analysis, reducing costs and amplify quality.

1.2.1 First generation methods of sequencing

The first sequencing methods were the Sanger dideoxy synthesis and the Maxam-Gilbert chemical cleavage invented in 1977 Fig.1.3. In particular, the Sanger method consists in specific chain terminating dideoxy nucleotides that lack the 3' -OH group, thus they cannot form the DNA chain. These nucleotides possess a fluorescent label, different for each nucleotide, that is detected to understand the specific sequence of the DNA. The first genomes sequenced with Sanger have been ΦX174 [27] and bacteriophage λ [28]. Even if now the method is automatic and uses an electrophoresis capillary instead of a slab gel, it is still employed in laboratories for some analysis. The main limitations of this method are that it does not allow the sequencing of complex genomes, its speed and the cost [27]. The Maxam-Gilbert method instead cleave the nucleotides leading to a series of marked fragments that are separated with electrophoresis by their size. Here the DNA is not cloned [29].

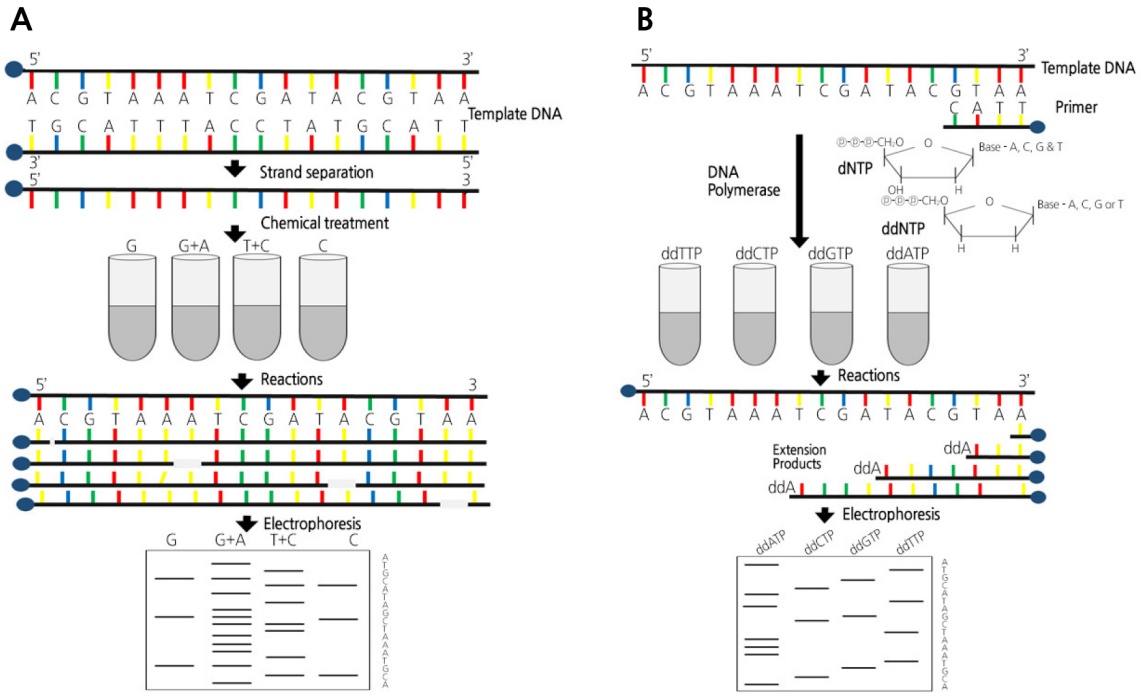

Figure 1.3: Process of sequencing using A) Maxam Gilbert method B) Sanger method Fig.1 from Kang et al. (2009) [1]

1.2.2 Second generation methods of sequencing

From 2005 a new generation of sequencers has emerged, the second generation methods have been invented because of the need of higher throughput sequencing of large genomes at a lower cost. This second generation is able to generate many million of short reads in parallel, it is faster than the first generation, it cost less and it does not need electrophoresis for the output detection. These methods can be grouped in two categories: sequencing by hybridization (SBH) and sequencing by synthesis (SBS). The most commonly used platforms of this generation are Roche/454 [30], Illumina/Solexa [31] and ABI/SOLiD [32] launched between 2005 and 2007 and also Ion Torrent launched [33] in 2010.

The first methods use pyrosequencing technique Fig.1.4A. It is based on the detection of pyrophosphate released after each nucleotide incorporation in the new strand of DNA that is formed. Each DNA fragment is attached to a bead in which there are primers. The reads generated by this methods are long between 100 to 700 bp depending on the instrument, the main errors seen in this methods are insertion and deletion in homopolymers regions [34]. The Ion Torrent Technology Fig.1.4B is conceptually similar to the Roche/454 pyrosequencing. It is based on the detection of the hydrogen ion released during the sequencing process. When a nucleotide is incorporated in the DNA chain, while it is duplicated on the bead, a hydrogen ion is released changing the pH of the solution that is detected by a sensor and converted into a voltage signal. Avoiding optical scanning to discern nucleotide the process is faster. It takes between 2 and 8 hour to sequence producing

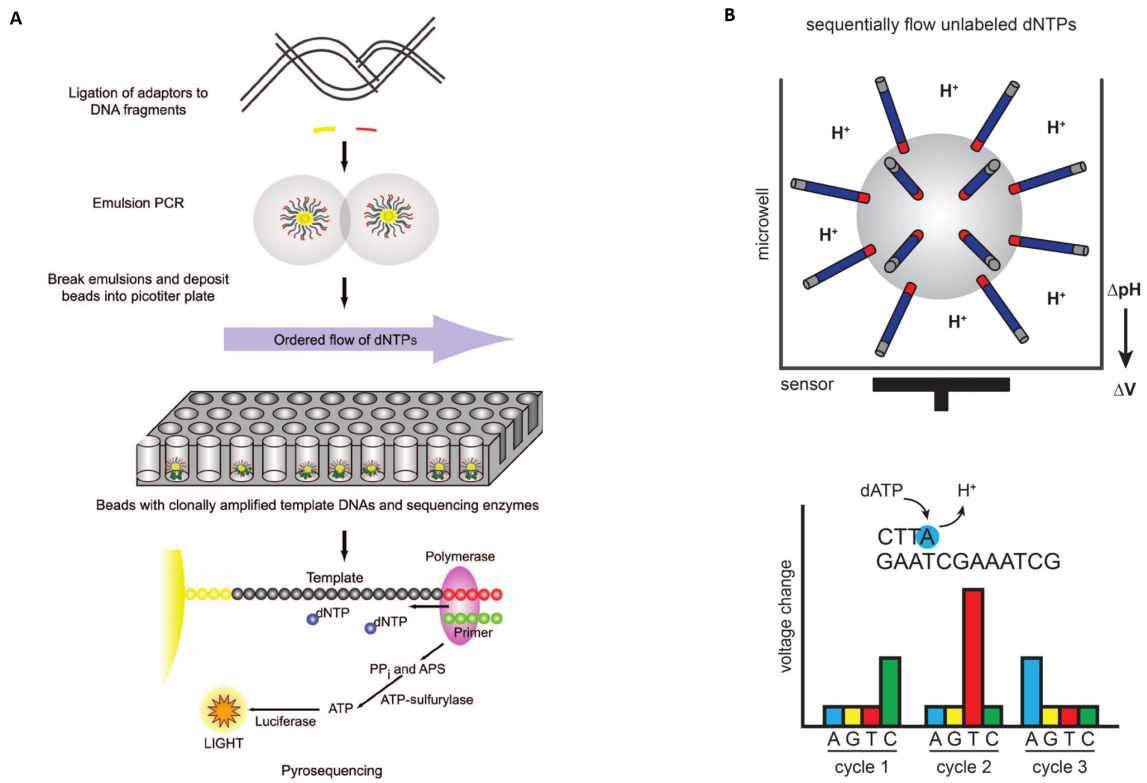

Figure 1.4: Process of sequencing using A) Roche/454 method Fig.1 from Voelkerding et al. (2009) [2] B) Ion Torrent Fig.2B from Reauter et al. (2015) [3]

reads of length between 200 and 600 bp. Also here it is difficult to interpret the homopolymer sequences with an error rate of about 1%. The most used system of sequencing in this moment is the Illumina technology Fig.1.5A. It consists in one step in which samples are randomly fragmented into sequences in which adapters are ligated. These adapters ligate to other adapters on a solid plate. Each fragment is amplified by a bridge PCR which creates cluster of the same sequence. Modified nucleotides are employed to identify each base, in particular fluorescently-labeled 3'-O-azidomethyldNTPs. The light of each nucleotide incorporated is detected by a coupled-charge device (CCD) camera. Now the length of short reads sequenced by Illumina is around 125bp. The error rate of this technology is about 1% and most of the time consist in the substitution of nucleotides. The ABI/SOLiD (Supported Oligonucleotide Ligation and Detection) Fig.1.5B sequencing process consists in attaching adapters to the DNA fragments, put it on beads and cloning it by PCR. These beads are placed on a glass where fluorescent labels are ligated to the DNA sequentially. After the ligation the nucleotide is removed and the ligation is done on the nucleotide -1. The cycle keeps going until the sequence is finished. Four different fluorescent colors are also used here. Each based is sequenced twice. ABI/SOLiD reads length is about 75 bp. In this platform the main error type is substitution, most of the time due to the noise during the ligation. All these methods described above are SBS.

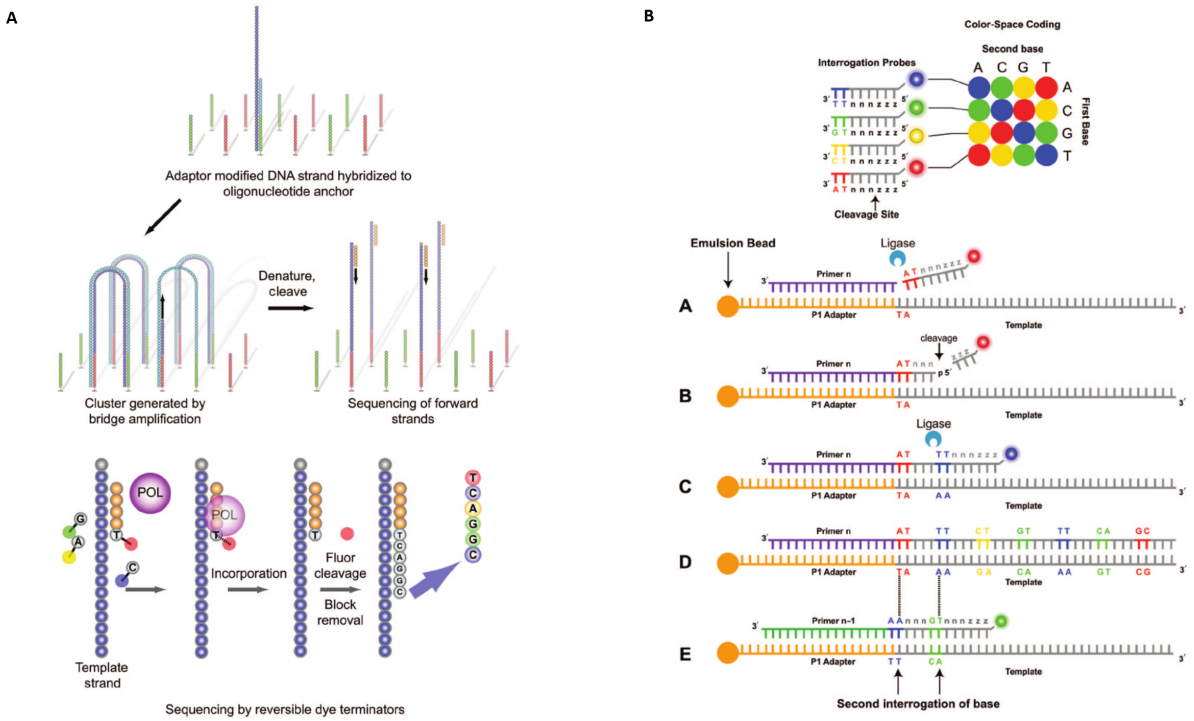

Figure 1.5: Process of sequencing using A) Illumina B) ABI/SOLiD Fig.2-3 from Voelkerding et al. (2009) [2]

1.2.3 Third generation methods of sequencing

The second generation methods need the PCR amplification step which is a long and expensive procedure. Moreover, the presence of repetitive areas causes problems in the amplification and the short reads analyzed make it difficult to assembly the genome. For this reason, third generation methods were created with the aim to sequence long DNA and RNA molecules without the PCR process [35]. These new methods have a lower cost and an easier sample preparation and they can produce longer reads simplifying the assembly of the genome [36]. They can be split in two approaches, the Single-Molecule Real-Time sequencing (SMRT) approach [37] and the nanopore approach. These approaches are employed respectively by Pacific Biosciences (PacBio) and Oxford Nanopore sequencing (ONT) [38]. PacBio [39] Fig.1.6A is the most used third generation sequencer. The DNA synthesis occurs in a zeptoliter-sized chambers, called zero-mode waveguides (ZMW), with a polymerase immobilized in the bottom. This is done to reduce background noise. It uses the same fluorescent labelling as the ones of the second generation methods detecting the signal in real time when they are incorporated, avoiding in this way the amplification step. The signal released by the nucleotide incorporation is recorded in real time by a sensor. Unfortunately, for this method the error rate is about 13% most of the time given by insertion and deletions errors. The length of the reads reach 60 kbp with an average of 10 kpb. A sequencer with the ONT technology Fig.1.6B is the MinION sequencing device [40]. This technique consists in link the first strand of the DNA with

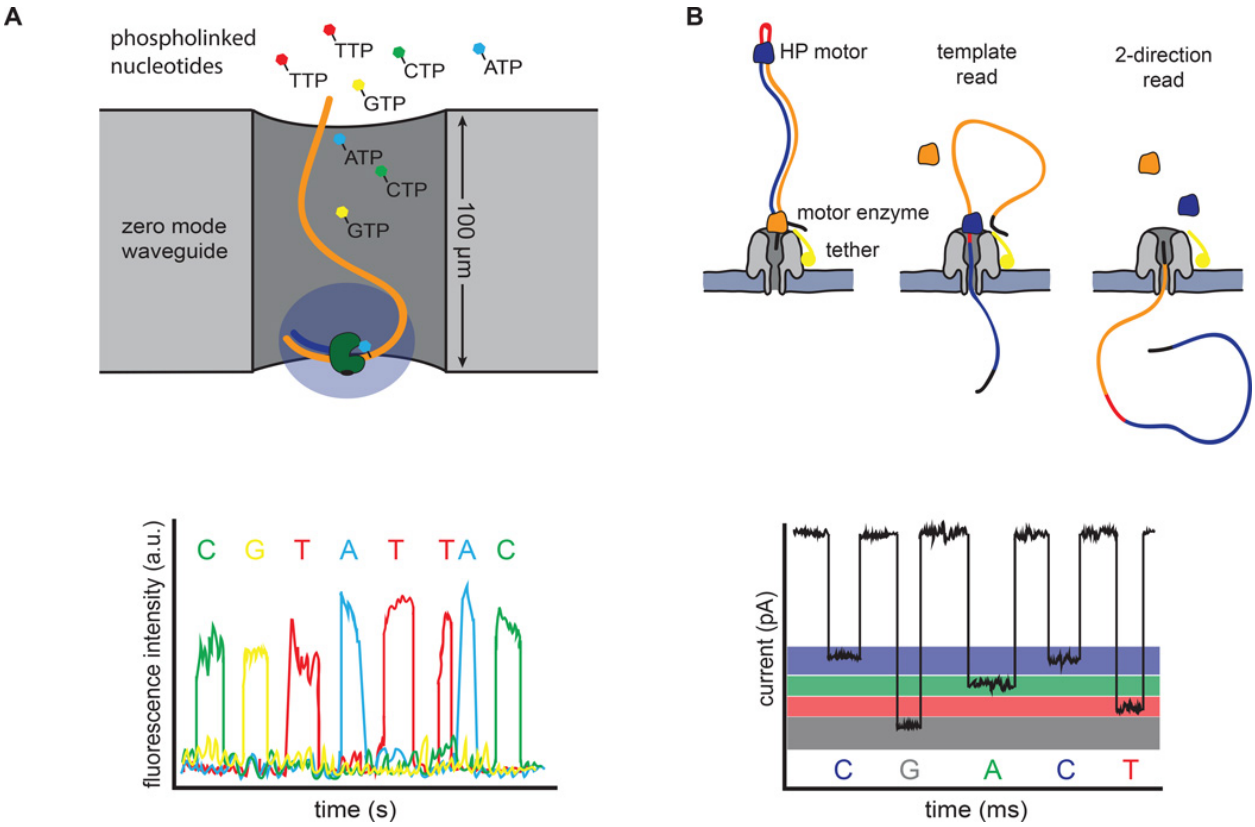

Figure 1.6: Third generation sequencing workflow. A) Pacific Biosciences. B) Nanopore Sequencing Fig.3 from from Reauter et al. (2015) [3]

its complementary by an hairpin. The DNA fragment is passed through a protein nanopore. When the DNA passes through the pore it generates a variation of a ionic current caused by the different nucleotides. This variation is recorded and then interpreted. Both strands are read. The reads produced by MINiON exceed the 150 kbp, but the error rate is of 5.1% for substitution, 4.9% for insertion and 7.8% for deletions [41]. The best characteristic of this sequencer is that it is portable and USB-powered [42].

1.2.4 Next Generation Sequencing: Importance and applications

The release of Second generation methods, better known as Next Generation Sequencing (NGS) methods, in the mid-2000s has revolutionized the way we do science. It brought to a huge cut in the cost of sequencing and opened to new perspective for genome analysis and exploration. The NGS platforms produce a huge quantity of data with an error rates between 0.1-1% and the read length is shorter than those used in Sanger. Since they were commercialized as technology capable of producing very high throughput sequence they are also called "High Throughput Sequencing Technologies". These techniques can sequence in parallel billions of reads in a single run. As told before, their main disadvantage is that they require high computer resource to do genome assembly [3, 36]. In this moment the main machines used are the ones of Illumina, example of them are

"MiSeq", "MiniSeq", "iSeq" and "NetxSeq". The uninterrupted decreasing of the cost is bringing more and more to the insertion of the NGS analysis in clinical practice, making possible study and understand deeply and fast all the living organisms. The NGS advances gave the possibility to perform also different kind of analysis, such as the metagenomic ones. In this way we now have a better vision on the insight of the diversity of the microbial species from the human body to the soil. For this reason, it was also possible to characterize the human microbiome through The Human Microbiome Project [43] between 2007 and 2013. Thanks to these analysis we have been able to understand the difference that exist between the microbiome of different individuals and the possibility to restore a healthy microbioma. As well as we understood which bacteria can cause specific disease. Bringing also to the discovery of new bacteria that are not culturable in laboratories. Regarding human the NGS analysis performed can be split in WGS, WES and target sequencing. From WGS we obtain a complete information since we have the sequencing of all the genome of the individual. Even if we have not the complete information, but only the one of the exome with WES we can have a major depth of the study thanks to the dimension of the sequencing. In particular, for cancer NGS analyses are performed both on solid tumor biopsy and on liquid biopsy. Another way of study tumors with NGS is the Single Cell Sequencing (SCS). It can be done both on DNA and RNA. This system consist in isolating a single cell, avoiding the mixed signal that it is obtained from normal sequencing obtaining cell specific information. In this way it is possible to investigate the tumor heterogeneity. A NGS experiment on a patient produces large amount of data, requiring time consuming tasks to extract knowledge. Users have to be specifically trained to be able to install and use tools able to interpret the results. The road to bring genomics analysis integrated into clinic for good is open, even if the cost and the computer resource needed have still to be reduced.

NGS tools

Seeking for a user-friendly way to digest NGS data in the last few years many pipelines have been developed for their analysis. In particular:

- In 2019 Joo T. et al. introduced *SEQprocess* [44], a tool implemented in R for the analysis of patients NGS samples in FASTQ format. It can be applied on both DNAseq and RNAseq and has six pre-customized pipelines. The user can analyze WGS, WES or liquid biopsy DNA-seq samples. It allows allele frequencies estimation for each variant by determining read depths of the mutated and wild-type sequences withing an RNA sample. However, the tool has been releases only for Linux Operating System (OS).
- *SEQVitA* [45] is an open source platform for the prediction of SNVs and small INDELs in

WGS, WES and in custom next generation sequencing data. The user can load one or multiple BAM or mpileup files to start the analysis. It can also analyze vcf outputs produced directly by the user. According to the number and type of samples, the detection of the mutations is divided in Germline, Population and Somatic, if the user employs paired tumor-control samples. Variants annotation is done both for coding and non-coding variants and they are provided with a functional impact score allowing prioritization.

- *DNAscan* [46] can annotate and visualize variants in a NGS sample. The user is able to analyze a sub-region of the genome. The annotation step is performed by *Annovar* [47] involving the *Clinvar* [48, 49], *Exac* [50], *dbSNP* [51] and *dbNSFP* [52] databases. At the end of the analysis a quality control and a results reports are generated with a tab-delimited list of the variant found in the analysis process. The pipeline is available on Amazon web Services (AWS), as a Singularity image and as a Docker image.

- *VARIFI* [53] is a web-based pipeline for the automatic variant identification, filtering and annotation. The user can load a sample up to 400MB, making impossible the analysis of larger samples. It combines different aligners and variant callers to improve accuracy. It is easy-to-use and does not need any users computational power because it is conducted on the *VARIFI* server. The user can load two input files (a BAM or FASTQ file and a BED file). Reads mapping is done with *bowtie2* [54], *BWA* [55] and *NextGenMap* [56], then reads are realigned using Genome Analysis Toolkit. Variants are called using UnifiedGenotyper and *bcftools*. Finally, annotation is done with *Annovar*, removing potential false positive. A final report is produced containing variants sorted by a confidence score and a plot with amplicon coverage information.

- *iWhale* [57] is a pipeline that allows the identification of somatic SNVs and indels in tumors. It is based on Docker. It analyzes fastq samples using *BWA* with the possibility to use both GRCh37 or GRCh38 reference genome. To improve the quality of the alignment several *GATK* command are used. After that, various variant callers are combined to obtain more reliable results. For variants annotation it employs databases such as *Clinvar*, *dbSNP*, *gnomAD* [58] and so on, to identify relevant variants to help the interpretation of the patient's tumor landscape and to create specific therapies based on driver mutations.

Unfortunately, the results produced by such systems are often strenuous to be obtained and their interpretation in a clinical context could result challenging and therefore useless to support decisions. To try to fill this gap, several commercial systems have been proposed. One example is TGex [59], an online knowledge-driven clinical genetics analysis platform that combines variant annotation and

filtering capabilities with a user interface that permits the interactive interpretation and filtering of the variants by scientist without any bioinformatics skills. To perform the analysis the user needs to upload the patient VCF and as output he can have a report file in PDF or Word with a detailed variant annotation file in Excel.

1.3 Immunity and Immunotherapy in cancer

1.3.1 Immunity

Each human owns two types of immunity, innate immunity and acquired immunity. The innate immune system is the natural defense mechanism encoded by the genes of the host. It includes physical barriers such as cell junctions, the secretes mucus layer (which we can find for instance in the respiratory epithelium), the epithelial cilia, small molecules such as cytokines, chemokines, etcetera. Some parts of the innate immunity are constantly active, but others are activated when there is an interaction with chemical structures invading from external. In the innate immune response the white blood cells that participates are neutrophils and macrophages. They secrete destructive substances such as enzymes that digest proteins and chemicals. After that, they engulf and digest through the process of phagocytosis. One of the first reaction after innate immunity response is the inflammation, it is stimulated by the chemical factors released and it is needed to establish a barrier against the spread of the infection. The inflammation starts thanks to macrophages, dendritic cells, histiocytes and mast cells. The chemical factors produced during inflammation are histamine, brakydinin, serotonin which sensitize pain receptor, cause vasodilation and attract macrophages. Macrophages produce cytokines that mediate inflammatory response. The infection that survive to this attack brings to the activation of the lymphocytes. These cells are involved in adaptive immunity to make specific response and to remember the infection, so after a reinfection the attack is faster and more effective. The adaptive or acquired immune system is based on antigen-specific receptors expressed on T and B lymphocytes. Antigens are usually peptides processed from proteins. The antigenic receptors are formed by genes that had somatic rearrangement of germ-line genes elements. This assembly of the antigens receptor permits the formation of million of antigen receptors each with a unique specificity. In the adaptive immunity we have the recognition of the "non-self" antigens thanks to the Antigen Presenting Cells (APC). All cells except the erythrocytes can present antigens, but the specialized APC are Dendritic cells, B-cells and macrophages. We have two types of responses of adaptive immunity one for exogenous antigens and one for endogenous antigens. The former are displayed from dendritic cells on its surface coupling them with the Major Histocompatibility Complex (MHC) that it is also called Human Leukocyte Antigen (HLA). Thus MHC-antigen complex is after recognized by T-cells. The MHC exists in two forms I and II,

normally the exogenous antigens are displayed on MHC II and activate the CD4+ T helper cells. T helper cells cannot kill infected cells or pathogens but they direct other cells to accomplish these tasks. The latter ones are displayed on MHC I and activate instead CD8+ cytotoxic T-cells that kill infected cells directly. When cytotoxic T-cells are activated they undergo a process of clonal selection in which they split rapidly searching into the body for cells with the couple MHC I + peptide. B cells have the role of create antibodies. Antibodies (called also immunoglobulin) are Y-shaped proteins that identify and neutralize foreign objects. In mammals we can find IgA, IgD, IgE, IgG and IgM each can handle different kinds of antigens. The bond between antibody and antigen makes the antigens a target for phagocytes. The B cells that produce antibodies are called plasma cells and part of them differentiate in memory B cells, which act when there is a re-infection in the host. The innate response is our first line of defense, while adaptive immunity comes after T and B cells have undergone to clonal expansion. Synergy between the two immunity is essential for a fully immune response [60, 61]. The main goal of the immunity system is to discriminate the appropriate target. The immunotherapy activation of the immunity system might be risky since a too strong reaction can be deleterious for the host, it might be even deadly [62].

1.3.2 Immunotherapy and Tumor Mutational Burden

William Bradley Coley is the man considered as the father of Immunotherapy. In 1891 he tried for the first time to use the immune system to treat bone cancer. This idea of taking advantage of immune system to fight cancer came to his mind by observing cases of patients that went into spontaneous remission after streptococcal infection. So he decided to inject mixtures of live and inactivated *Streptococcus pyogenes* and *Serratia marcescens* into patients' tumors. However, the so called "Coley's toxins" were not so appreciated by oncologists that preferred to continue to using surgery and radiotherapy avoiding the risk of infecting already sick patients with pathogenic bacteria. After 1945 advances in immunity and cancer research brought to the re-discovery of the immune system as possible therapy. In that period in fact interferon was discovered and Ruth and John Grahams invented the first cancer vaccine. Between 1860 and 1980 T cells, dendritic cells and natural killer cells were discovered and studied enhancing the knowledge about the immune system [63]. In the meantime, at the University of Minnesota there was the first bone marrow transplant as a treatment for hematological cancers [64]. About 50 years ago professor Lloyd J. Old, considered the pioneer of cancer immuno-onclogy, noted that between cancer cells and normal cells there are differences which can be recognized by the body's immune system [65]. In 1957 Thomas and Burnet proposed a theory about immunosurveillance, they suggested that lymphocytes act as sentinels in order to identify and eliminate the cells that had been transformed by mutations

[66]. This theory re-emerged in 1974 when Stutman showed that mice with impaired immune system developed cancer faster than wild type ones [67, 68]. Even the identification of natural killer cells providing additional support to the immunotherapy possibility [69]. Only at the end of 1900 Schreiber, Dunn, Old and their teams proved that T cells were able to provide anti-tumor surveillance and immune response. Subsequent discoveries were about immunoediting, cancer cell escape and the higher risk of cancer development in immunosuppressed patients [70, 71, 72, 73, 74]. Since a lot of tumors block their own recognition by the immunity system the fundamental role of the immunotherapy it is to make the human immunity system recognize the tumor as an alien. Immunotherapy in cancer is used to activate or boost the patient immune system to let it attacks tumor cells. This is necessary because tumor cells evade recognition and elimination by T cells that bring to their uncontrolled growth and to clinical progression. The only way for our immune system to recognize the tumor cells is its production of two class of antigens called Tumor-Associated Antigens (TAA) and neoantigens or Tumor-Specific Antigens (TSA) that are over-expressed in tumor cells. These antigens are produced only in tumor cells as a cause of amino acid change due to mutations. The mechanism of immune elimination of tumor is explained by the fact that the adaptive immunity system recognized these neoantigens released by necrotic and/or apoptotic tumor cells. Tumor specific cytotoxic T cells go back to the tumor and destroy their target. Several drugs have already been approved by FDA for more than nine cancer types. The ones used in this moment are immune checkpoint inhibitors that hinder molecules which suppress immune response. The first immune checkpoint molecule was discovered in 1987 by Brunet and his team and was named Cytotoxic T-Lymphocyte Antigen number 4 (CTLA-4) [75]. However, only in 1995 Jim Allison et al. discovered the crucial role of this molecule as immune checkpoint molecule [76]. Nowadays more than 2000 cancer immunotherapy agents exist. [77]. In 2011 FDA [78] approved ipilimumab as the first checkpoint inhibitor for treating advanced melanoma. Over than 20% that have been enrolled in the first ipilimumab trial before 2011 are still alive and they have no evidence of disease. Nivolumab, instead, was in 2014 the first PD-1 inhibitor to gain approval for the treatment of Melanoma in Japan [79]. Other actually approved inhibitors of the PD1-receptor or of its ligands, PD-L1 and PD-L2, are pembrolizumab, atezolizumab, durvalumab and avelumab. In these last years cancer immunotherapy is revealing as the therapy with the most durable and outstanding benefits across tumors, with actual treatments between 16% and 30% of patients survive melanoma and lung cancer. However, often immunotherapy results useless or toxic for patients leading to unpleasant side effects, such as skin rash, colitis, hepatotoxicity, pneumonitis, endocrinopathies and autoimmune disease [80]. Nowadays, only 20% of patients benefits from cancer immunotherapy [81, 82] for two reasons, the lack of a proper Biomarker and the lack of the patients of anti-tumor immunity. The Tumor Microenviroment (TME) is fundamental for immunotherapy success, if the

TME has region of hypoxia and elevate lactate levels that do immunosuppression. The TME is composed by cancer cells, stromal features, blood vessels and infiltrating immune cells. It is highly variable between individuals and different tumors [83]. An example of the pathways correlated to the TME are the downregulation of the MHC class I on the surface of tumor cells[84]. In the TME we can found upregulated cells such as Myeloid-Derived Suppressor Cells (MDSCs), tumor-associated macrophages (TAMs) and mast cells that prevent the immune system from eliminating tumor cells [85]. The tumor associated macropahes are capable of suppress immune response [86, 87]. As routine biomarkers we can already find Programmed Death Ligand-1 (PD-L1), MisMatch Repair (MMR), Microsatellite Instability (MSI) and Tumor-infiltrating Lymphocyte (TIL). The most used biomarker for immunotherapy decision is the PD-L1, the higher its expression the higher the probability of response, but its performance are not satisfying. The rate of response is between 19 and 30% [88]. PD1 and PDL1 are indeed involved in immune evasion [89] because PDL1, when expressed in high levels, binds on PD1 and can cause the inactivation of T cell and apoptosis of them, the blockage of their pathway can led the tumor system to attack the tumor. Moreover, some PDL1-negative patients respond also positively to the immunotherapy leading to the idea that there are other contributors to the therapy [90]. Another important biomarker, especially in colorectal cancer, is MSI. Over 80% of cases of Hereditary Non-Polyposic Colorectal Cancer (HNPCC) show this instability of microsatellites [91]. MSI is a unique molecular alteration and hypermutable phenotype, which is the result of a defective DNA MMR system. It can be defined as the presence of repetitive DNA sequences of alternating sizes that are not present in the corresponding Germline DNA. Determining MSI status in tumors can show prognostic, therapeutic implications and can even be used as a diagnostic tool for tumor classification. A recently used biomarker is the Tumor Mutational Burden (TMB), it counts the number of neoantigens existing in the tumor to calculate the possibility to activate immunotherapy in the patient. The goal of TMB searching is its use in clinical practice. This searching technique makes possible to obtain a real complete photograph of the tumor, from the molecular point of view, for each patient: this allows to realize precision medicine for immunotherapy. The immune response is able to fight tumors mainly in the initial phase, but then the tumors learn to react and escape from it through the production of a series of molecules. Immunotherapy removes the brakes of the immune response through the inhibitors of the immune check points. Tumors which have a greater load of neoantigens and mutations, are precisely those that benefit the most from this strategy, as they are more able to stimulate the immune response.

TMB TMB is one of the most promising emerging biomarker. The TMB calculation is commonly used for melanoma, that is one of the tumor with the higher mutation rate, but now the common

goal is to make it feasible for lung and colon cancer that are the two forms of cancer with the worst prognosis. TMB is a very complex biomarker that requires sophisticated NGS techniques for its determination. It is calculated counting the mutations found in a tumor sample, excluding the ones that are already known as cancer mutation, and dividing the sum by the total length of the sequenced sample DNA in Megabase (Mb). Calculating the TMB patients are split in two groups, the high TMB group where patients would benefit from immunotherapy and the low TMB group where patients would not. Higher is the TMB of the patient higher is its possibility to respond to immunotherapy because a higher TMB corresponds to a higher number of neoantigens. There is not a standardized value to decide which TMB is high or low because the threshold depends from the study, the drugs [92], but especially on the tumor treated [93]. Unfortunately, its use is not standardized yet, bringing to the use of different methods of calculation and thresholds. Nowadays, the gold standard to measure the TMB is using the WES analysis with tumor and normal sample, but it is under study the possibility to employ specific panels [94, 95] to speed up the analysis with the same precision and sensitivity and to make them feasible for clinical practice. As a matter of fact, WES analysis have high cost and require extensive data management, moreover for the TMB analyses two sample are needed, the tumor and the normal one to discard germline mutations. Unfortunately, the availability of this matched sample in clinical practice varies across organization. Germline variants in a tumor-only sequencing can be filtered out using available databases, but this procedure needs a high level of standardization for each type of tumor and for each population [96]. In addition, there is a lack of standardization of current TMB methods both by research laboratories and bioinformatic analyses. The actual commercial panels endorsed for TMB research are The FoundationOne CDx assay approved by FDA and the MSK-IMPACT (Memorial Sloan Kettering Cancer Center) which has been authorized by the 510k pathway [97, 98, 99]. Furthermore, panel such as "Thermo Fisher Scientific Oncomine Tumor Mutation Load Assay" [100], "TruSight Tumor 170" [101] and "Foundation Medicine FoundationOne" [95] are used in clinical, but are not approved yet. Some authors recommend to use targeted gene panel assays that have larger genome coverage (ideally with ~ 1 megabase as lower limit) because they yield more reliable TMB estimation than smaller panels [102, 103, 104]. Notwithstanding the existance of all this panels, according to Wu et al. [105] the current available panels can assess TMB accurately only in several particular cancer types. They explain that the correlation itself is unreliable to evaluate the performance of panels and that accuracy is a superior index of the situation.

Differential expression of genes in TMB Differential Expression Genes (DEGs) analysis have permitted to clarify the role of the genes in cancer patients, and between high and low TMB patients. Comparing tumor and normal colon sample Gao et al. [106] found that differential

expressed genes (DEGs) were mainly involved in protein transport and apoptotic and neurotrophin signaling pathway. Wang et al. [107] screened TCGA-BRCA data-set splitting patients in TMB high and TMB low and analysed them with KEGG and GO databases. They found that DEGs were mostly enriched in epidermis development, extracellular matrix, and receptor-ligand activity among Biological process, Cellular Components, and Molecular Functions, respectively and showing that 343 genes were expressed differential in the 2 groups of TMB. Zhang et al. [108] found that in bladder urothelial carcinoma differential genes were involved in catalytic activity, acting on DNA, single-stranded DNA-dependent ATPase activity. Moreover, TMB-enrichment of related signature correlates with multiple cancer-related crosstalk, including cell cycle, DNA replication, cellular senescence, and p53 signaling pathway.

1.4 Variant Prioritization

Thanks to NGS we are able to collect every day huge quantities of genomics data which are employed in clinical practice. The study of the genomics variants inferred from NGS is fundamental for predictive and precision medicine. Therefore, to arrange a specific not toxic therapy for a patient is decisive to recognise pathogenic variants responsive to specific drugs. The interpretation of these variants results complex due to the need to integrate a lot of bioinformatics tools that require a computer science expertise. Moreover, to perform such analysis is needed a huge computational power that it is not always disposable. There is a demand for both easy to use bioinformatic pipeline and for variants prioritization databases. The recognition of a causative variant is difficult because WES and WGS produce thousands of sequence variants for which the detection rate of casual variant is lower than 20-30% [109, 110]. Nowadays, multiple unrelated individuals with similar phenotype with the same gene mutation are required to define a variant causative [111]. For all these reasons, a lot of variants are classified as "VUS" i.e. "Variant of Uncertain Significance", which means that a variant that damages a gene is not necessarily damaging to an individual's health, or as "Unknown". This class is the most common in personal genome sequences and can include both novel variants on coding sequence of disease-causing genes but it can refer also to variants in genes unlinked to disease. Usually, to predict the deleteriousness of a variant, so to do a variant prioritization, a pathogenicity score is employed. As a matter of fact, a variant can be considered pathogenic only when its DNA alteration has a role in the disease process. A right assignment of variants is fundamental, wrong assignment indeed could lead to severe consequences for patients, resulting in incorrect prognosis and therapy. Some recent analysis showed that the 27% of mutations found in 104 individuals were either common polymorphisms or lacked direct evidence for pathogenicity [112]. The "Variant prioritization" or "Variant Filtration" is the practice of annotate and interpret the variants to

identify variants conceivably associated with pathological condition. Most of the time this process requires manual curation by expert clinician. Often, different institutions and operators apply different criteria and filters that limit the overall reproducibility of the results of these analyses. The identification of pathogenic variants could lead to correlate variants and phenotype identifying causative variants for a specific disease. A lot of gene and variants prioritization methods have been implemented, all this tools filter, evaluate and prioritize thousand of variants using public databases. Unfortunately, there are a great number of intergenic or intronic regulatory variations or unidentified structural variants that are not prioritized yet. Example of the Gene and/or Variant Prioritization methods are:

- *GeneDistiller* [113] can be used as a prioritization tool or with other prioritization tools to display rich information on human candidate's genes obtained with those. It offers different approaches such as Projection, Selection, Sorting and Prioritisation. In the first approach it is the user that chooses the genes that are of interest to him. In the second approach the user applies filters to the genes decreasing them to a smaller group. In the third approach the genes are sorted according to certain parameters. Finally, the forth approach, that is the prioritisation one, offers a function which ranks genes according to the researcher's specifications. This methods can be combined.

- *MutationDistiller* [114] prioritize monogenic disease variant, with the help of GeneDistiller. It filters the polymorphism using databases such as ExAC and 1000Genome and use Clinvar to identify known disease-causing mutations. After the analysis MutationDistiller presents a prioritized list of the most likely candidate variants with information about them and their genes that can be downloaded as a summary table. The table shows the variant in class so the user can focus on certain types of alteration more than in others.

- *VINYL* [115] derives a pathogenicity score aggregating different public databases. The idea that stands behind the tool construction is that affected individuals have and excess of deleterious variants compared to a matched population of unaffected ones. The tool is highly flexible permitting the incorporation of different types of annotation and resources by the user. It seems to have highly levels of sensitivity and specificity.

- *KGGSeq* [116] does an analysis procedure for the discovery of human Mendelian disease genes combining filtration and prioritization functions. It filters and prioritizes the variants at three levels, genetic, variant-gene and knowledge according to the resource used. Such as MutationDistiller it filters out common variants using public databases like 1000Genome and the allele frequency threshold.

- *vPot* (variant prioritization ordering tool) [117] is a command line python-based program that creates a pathogenicity score making the user able to prioritize variants. Taking the ANNOVAR-annotated file the tool will annotate the variant based on the annotation elements found using several databases as CADD, LRT, etc. At the end the tool calculates a score that is normalized.

- Variant Ranker [118] is a web based tool to interpret genomic data that gives back a list of prioritized variants generated computing a score thanks to several databases.

In clinical practice and in this tools is common to use SIFT [119], CADD [120], PolyPhen-2 [121] and other pathogenicity score predictor to help variants interpretation. They derive an impact score using amino acid or nucleotide conservation and most of the time classify the variants as "Damaging or tolerate".

1.5 Human Microbiome

Human microbiome is a complex machine composed by bacteria, virus and archaea that interact with each other and with the host permitting the maintenance of the function of the host organism. The microbiome of each individual is diversified by multiple factors such as diet and environment. Studying the bacteria of microbiome nowadays is pretty easy thanks to the standardization of the tool QIIME [122] that permits to analyze the 16s rRNA gene of the bacteria, region that is highly conserved. The database of 16s rRNA sequences is constantly updated and curated. Instead, since viruses do not possess phylogenetically conserved region it is not possible to build a phylogenetic tree and so their study is a difficult task to achieve. Therefore, little is known about the function of virome in human microbiome. [123, 124, 125] Viruses are parasites that infect cells thanks to their surface proteins that bind with cellular receptors [126]. They can have single or double stranded DNA or RNA. In particular, RNA viruses form a highly diverse group, they can be single-standed (ss) plus or minus oriented, or ssRNA with a dsFNA as an intermediate product, or double-stranded (ds). Their genomes are small going from 3400nt to 31000nt [127]. A study of 2013 of Anthony et al. [128] estimated the existence of at least 320000 species of viruses that infect mammals. The 2016 database release from the International Committee for the Taxonomy of Viruses classified just 8 orders, 122 families, 735 genera, and 4404 species. In 2020, after only 4 years, ICTV have affirmed that 6 realms, 10 kingdoms, 17 phyla, 2 subphyla, 39 classes, 59 orders, 8 suborders, 189 families, 136 subfamilies, 2224 genera, 70 subgenera, 9110 species exist [129]. Thus, it is normal to think that we cannot even image the extent of viral diversity that exist in the environment. Even if viruses have been the first biological system to be sequenced (bacteriophage MS2 in 1976

[130] and ΦX174 in 1977) the recognition of viruses nowadays is still difficult. There are a lot of causes that can explain this problem, first of all the cultivation of virus in lab it is a complicated and long task. Secondly, not all existing viruses are currently correctly annotated in databases. Moreover, the high number of viruses that already exists does not allow to create a comprehensive database. Finally, the tools for virus research are not consistent with each other, not user-friendly and demand a huge quantity of computing resources [131]. The development of tools for the research of viruses it is necessary especially for clinical diagnostic and public health. At this time, two methods exist for the discovery of virus: sequence-dependent and sequence-independent methods. Between the former ones are included PCR, using consensus primers, and hybridization methods such as microarrays. These methods, though, require the knowledge of specific nucleic acids like consensus sequences of previously known viruses. They permit to discover novel virus, but with a common root with others. For instance, this method was used for Human Immunodeficiency Virus (HIV) [132] and for simian retrovirus [133]. The latter ones do not require the knowledge of viruses in the samples. The methods utilized are suppression subtractive hybridization (SSH) and representational difference analysis (RDA). However, the best choice for viruses discovery is viral metagenomics. Metagenomics is the study of microbial community genomes taking them directly from the environment. This approach is culture-independent and sequence-independent. It is not only less biased, but it is also helpful for the research of both known and novel virus. These new viruses could be recognized as potentially infectious agents and associated with human diseases. A metagenomic analysis is composed by three steps: (i) sample preparation, (ii) high-throughput sequencing and (iii) bioinformatic analysis. At the beginning metagenomic was applied through the Sanger method, nowadays NGS methods such as Illumina/Solexa and Roche 454 are employed. Normally a tool for virus research does an alignment between the host genome and the sample, so it deletes all the host sequences leaving only the microorganisms sequences. After that, it can both use unassembled reads or reconstructed contigs, arising from the assembly process, to classify viruses. Most of the time this classification is done using BLAST or through a homology research with already existing sequences in database such as ncbi one. This latter technique, though, does not permit the identification of new viruses and requires precise post-process to keep only meaningful classification without risking the loss of real important data. Regarding BLAST, even if it permits to discover new viruses it takes a lot of time and CPU to classify virus, so now is common to use abundance estimation programs and k-mer programs. Abundance estimation programs create a database that is smaller than the collection of the entire genomes so the classification is faster, but they classify only a small part of the sequences of a metagenomics sample. These tools are only meant to characterize the distribution of the organisms present in a given sample. K-mer tools, instead, are based on kmer methods which find exact matches between small substrings (k-mers),

from the reads sequenced, and a viral reference sequence from a database. This method permits rapid sequence classification, but struggles identifying divergent sequences of viral origin because is less sensitive and specific when it has to identify species [134]. Some tools use high-sensitivity protein alignment, but it takes too much time and RAM use to be done. Certainly, there is a desperate need for a standard protocol and algorithm for virus metagenomic downstream analysis to study and understand the large number of sequences that have not yet similarity with anything already known [135]. The major pros and cons of metagenomic are indeed that it produces a huge quantity of data that have to be analysed. As a matter of fact, it is esteemed that in each metagenomic virus study there are between 40% to 90% of not detectable sequences that are probably virus sequences [136] limiting the possibility to understand the virome structure and function in health and disease. Each of the tools that exists in this moment has its advantages and disadvantages and it is necessary to evaluate which is the best tool to use depending on the analysis situation [137]. Furthermore, most of the time software are not updated and/or they are not well developed and so cannot be employed for long time. Software should be not only functional, but they should be also sustainable developed, documented and tested and be distributed through robust and user friendly channels such as non-online application. Additionally, it is difficult to find pre-built indexes for virus that are up-to-date and built them require big memory usage.

Chapter 2

OncoReport

OncoReport is a flexible and easy-to-use tool able to generate reports from DNA NGS data for supporting decisions in clinical settings. It focuses on the gene annotations with drugs to understand which can be the best patients' customized therapy. *OncoReport* was developed using both bash and R. Its Graphical User Interface (GUI) Fig.2.2 has been developed in Javascript using the Electron framework (https://electronjs.org/). Electron is an open-source framework developed and maintained by GitHub, allowing the development of desktop GUI applications using web technologies. As it can be seen in subsection 1.2.4 called "NGS tools" there is a lack of a free platform pipeline with an user interface which is able to produce a readable report with information about drugs sensible/resistant mutations. So, we developed this new software usable by clinicians and laboratories. Its aim is to be feasible for non-informatic expert user in order to be implemented in every day clinical practice.

2.1 The Pipeline

The pipeline is thought to be as smart as possible, so the user can upload different kinds of samples to start the analysis. It can upload directly the FASTQ resulting from the NGS analysis, the SAM, the BAM, the UBAM, the VCF or even the varianttable file built from illumina experiment. The pipeline will automatically recognize the extension of the file and will start the analysis. The time of the analysis depends on different factors. Firstly, if the user decides to analyse a WGS or a WES sample the analysis will be longer than an analysis of a targeted panel. Secondly, an analysis starting from a VCF is faster than an analysis starting from a FASTQ. Finally, a lot depends also on the computer resource available for the user. For sure, for WGS analysis are needed at least 64Gb of RAM, 8 core and 1Tb HDD. The oncoreport pipeline Fig.2.1 comprises four main part:

1. *Pre-processing*: The pre-processing step consists in removing the sequencing adapters with

TrimGalore[138, 139], aligning the sample with *bowtie2*[54] and sorting and cleaning the BAM file, with the erasing of duplicates in case of tumour-normal analysis, via *Picard* [140].

2. *Variant Call-Filtering*: The variant calling is done with *Mutect2 GATK*[141], the parameters are set according to the different types of samples analysis. *Mutect2* has been chosen for its performance in the detection of the somatic variants that are the main focus in this research. However, in a future version we are planning to add other variant caller such as *VarScan* [142] to create a consensus variant file that will be even more reliable. The Variant filtering is done with GATK tools. In particular, the filtering in the liquid biopsy and only tumour pipeline comprise the Depth (DP) and the Allele Fraction (AF) filters, the former to specify the depth required for the analysis and the latter to divide the germline and the somatic mutations. For instance, in liquid biopsy analysis we can set AF to 0.3, while in a solid tumour analysis it can be set to 0.4. This because normally a variant with an $AF < 0.3$ is more probably a new somatic variant showing less than a germline variant. This variant splitting in germinal and somatic is instead done automatically if the user pass to the pipeline two samples for the same patient, one normal sample and one tumor sample. This makes the analysis more accurate.

3. *Variant Annotation*: The Variant annotation is done through a custom R script using the following databases, *Civic*, *CGI*, *Refgene*, *PharmGKB* and *Cosmic*.

4. *Report Generation*: The report is created using *R*, css and HTML, it is developed to be easily understood by clinicians and even by oncology patients. It consists on a description of the most important mutations variants found in the patient's tumor. These mutated variants are reported with their associated drugs, the details of the specific mutation and the clinical trials that are available for the mutation and its associated drug.

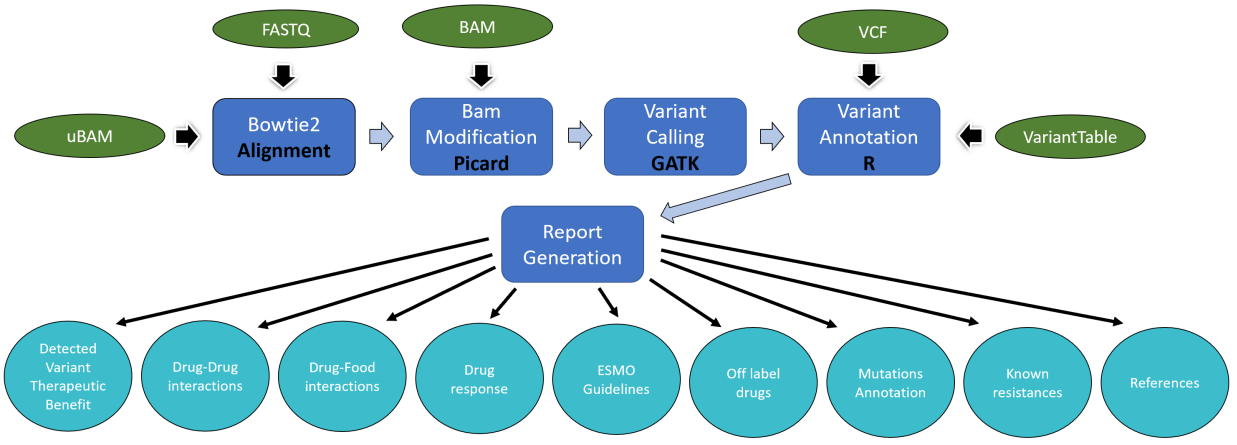

Figure 2.1: Oncoreport pipeline

Drugs Score Calculation Throughout the report creation, each drug coupled with a variant acquire a confidence score. This score calculation depends on three considerations: (i) the year of publication of the association between it and the variant, (ii) the pathogenicity of its mutation, (iii) how many times it is repeated in our report. Depending on such score each drug will have a different color, such as green which will suggests that the drug is the best choice for the patient. In particular the drug will receive 3 points if the paper of the drug-variant association has been published in the last years (between 2021 and 2019), 2 points if has been published between 2018 and 2016, 1 point if has been published between 2012 and 2010 and 0 point if it has been published before 2010. To calculate the score of pathogenicity we make use of *dbsnfp41c*. For each of the 8 pathogenicity predictors inside this database we assign to the variant a score between 1 and 0. Concerning the presence in our report, the drug will obtain 1 point for each variant mutation in which it is mentioned. At the end all these temporary scores are summed and give back the final score.

2.2 The Report

The final output is a report composed by different sections which are reachable from a Table Of Contents (TOC) Fig. 2.3A: At the beginning we can find a horizontal panel which holds patients' information supplied by the user while uploading the NGS files. Such information include patient's name and a code to identify uniquely the specific patient. The user could supply also the information about the drugs taken in that moment by the patient to understand the interaction between the drug already taken and the drugs suggested by the report. The section are split in this way:

- **Detected Variant Therapeutic Benefit** Fig. 2.3B,C,D consists of a description of the most important variants mutation found for the patient's tumor. These variants are reported with their associated drugs, the details of the specific mutation, the clinical trials that enroll patients harbouring them (found in clinicaltrials.gov), the confidence score and the year in which the paper of the drug-variant association was published. Higher is the confidence score, more reliable is the suggestion. This section has a division in two tables. The first one with all the drugs stated as "Clinical Evidence" or in general approved, for instance by FDA and/or NCCN, and already used. The second one with drugs used only in clinical trials, in case study or tested only in vitro. The user can also find the approval information for each drug, understanding which one is feasible in his country. The institutions that we have already integrated are FDA[78], EMA[143] and AIFA[144].

- **Drug-Drug interactions** Fig. 2.4A lists the drug-drug interaction among the drugs recom-

mended by our tool and the other drugs found in drugbank[145]. This permits the patient to avoid unpleasant effects caused by the drug mixting. Moreover, the user can upload the drugs already taken by the patient and have the drug interactions information also about them.

- **Drug-Food interactions** Fig. 2.4B lists the possible drug-food interactions derived from our system suggestion using drugbank.

- **ESMO Guidelines** Fig. 2.4C reveals the European Society for Medical Oncology (ESMO) [146] guidelines related to the patient's cancer type. The ESMO is the main entity in Europe for the dissemination of best practices for the prevention, diagnosis, treatment and follow-up of cancer diseases. This area of the report shows on the left a scrollable list of all clinical practice guidelines limited to patient's disease. The information related to each element of the list are displayed at the center of the page in the form of text or dynamic algorithm.

- **Mutations' Annotation** Fig.2.5B lists all the mutations that have been found in the patient's sample describing their role as pathogenic, benign and so on. It is built using *Refgene* and *Clinvar* [48]. It focuses on the function and the clinical significance of the mutation.

- **Drug response** Fig. 2.5A lists the mutations found in the *PharmGKB* [147] database. It gives information about the efficacy or the toxicity of a drug associated with a mutated variant, but it does not include specification of the diseases.

- **Off label drugs** Fig. 2.5C includes the variants that, according to the current knowledge, are associated with drugs in tumors distinct from the patient's one.

- **Known resistance** Fig. 2.5D contains annotations from the Cosmic [148] database that are useful to discover the existence of drug-resistant mutations.

- **Reference** Fig. 2.4D the user will find the literature references for each feature in the report. In this way he can go directly in the pubmed archive to check all the information about the couple mutation-drug.

OncoReport can be used in connection to NGS data coming from Liquid biopsy and tumor tissue both alone or with normal pair sample. It can be used for WGS, WES and/or target panel analysis. We prepared a user manual that explains step-by-step how to use the tool which is provided in the github page. OncoReport can be downloaded both with its Graphic Interface or, for informatic expert users, also from Docker Hub. The system is completely offline to assure the security of the patients data.

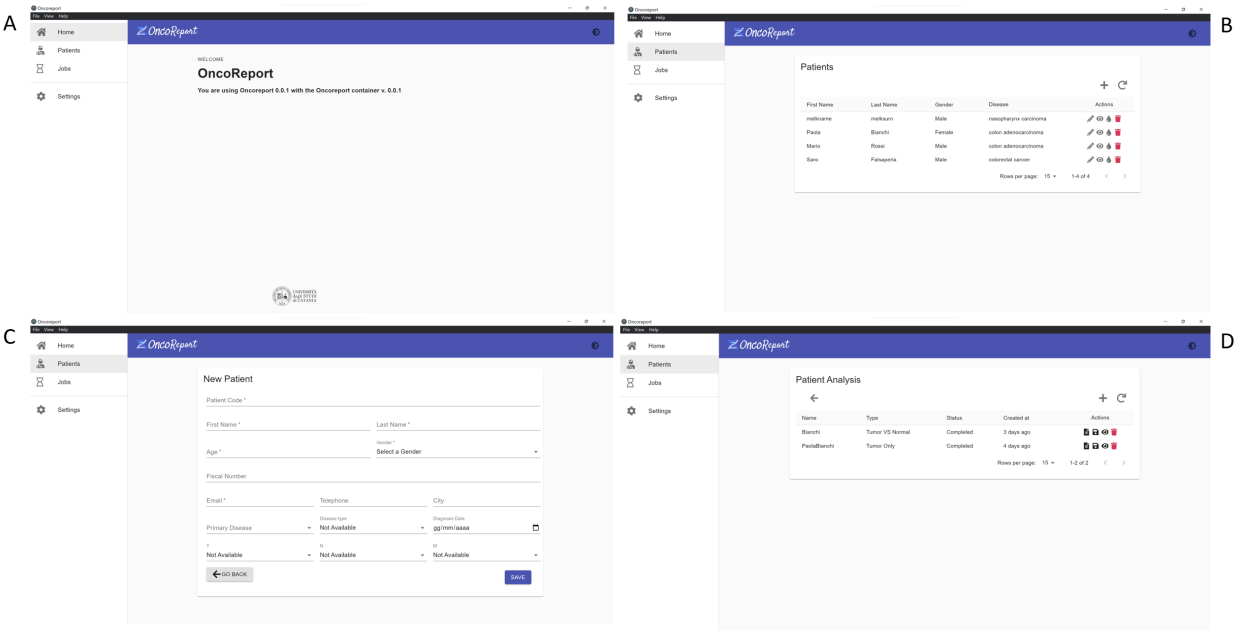

Figure 2.2: OncoReport Interface. A. Home page B. Patients list C. New patient creation C. List of patient analysis

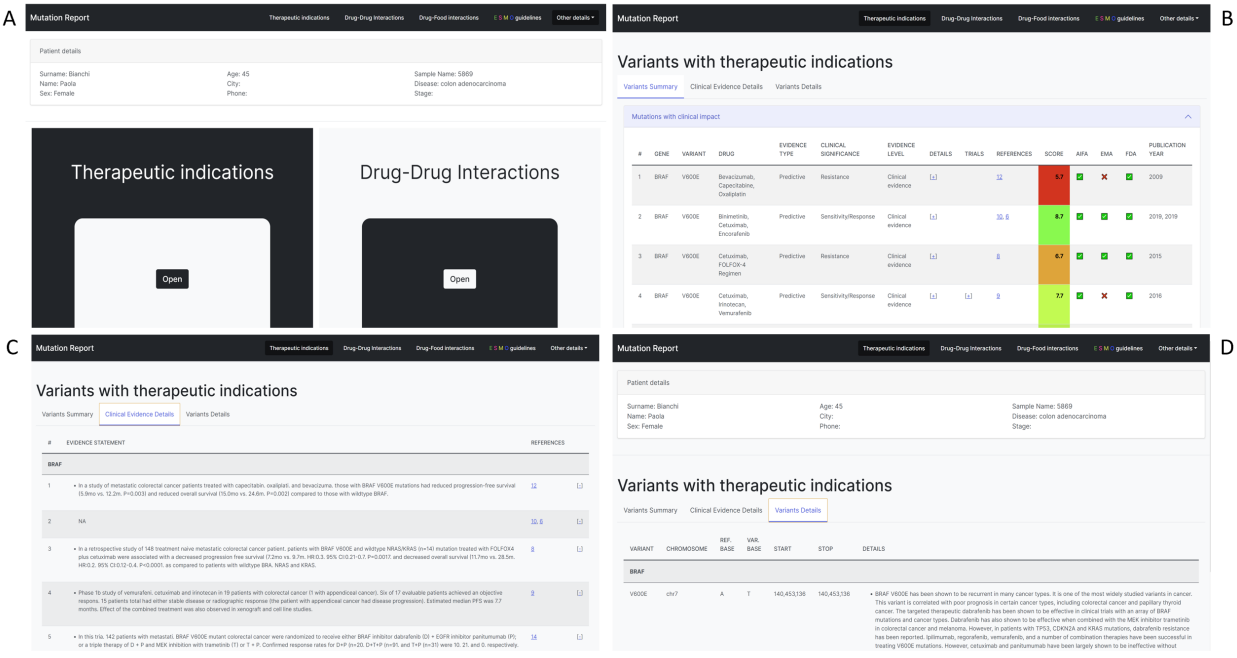

Figure 2.3: Report. A) Patient information B) Detected Variant Therapeutic Benefit section with drug-mutation information C) Detected Variant Therapeutic Benefit section with Evidence details D) Detected Variant Therapeutic Benefit section with Variant Details

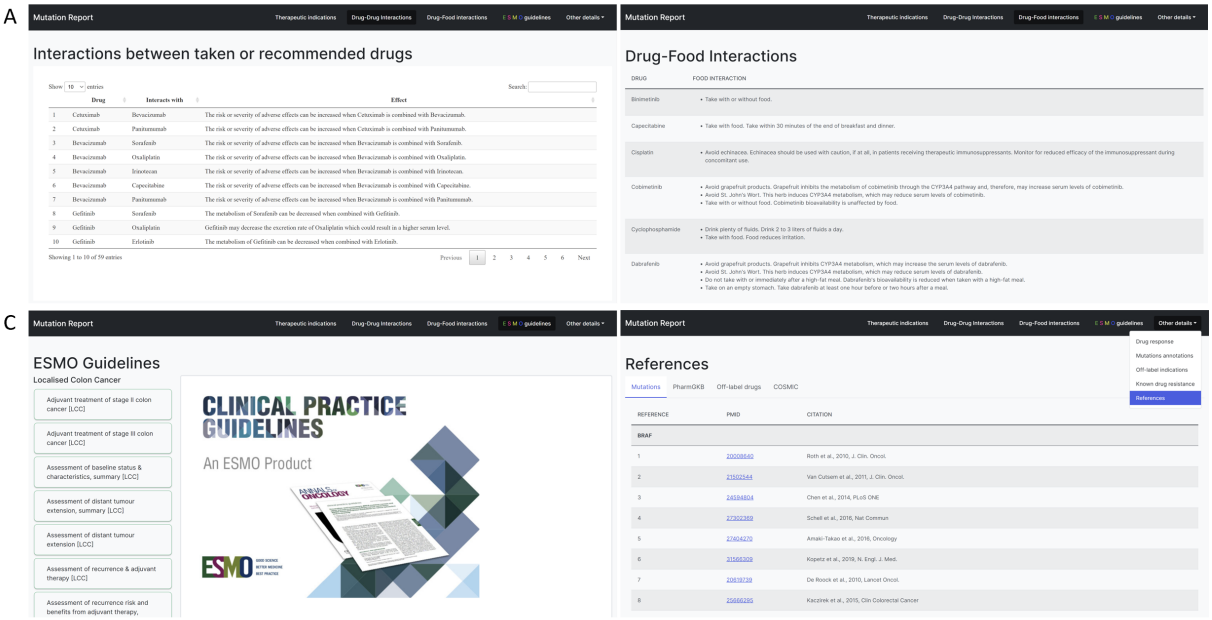

Figure 2.4: Report A) Drug-Drug interaction B) Drug-Food Interaction C) ESMO Guidelines for patient disease D) Reference

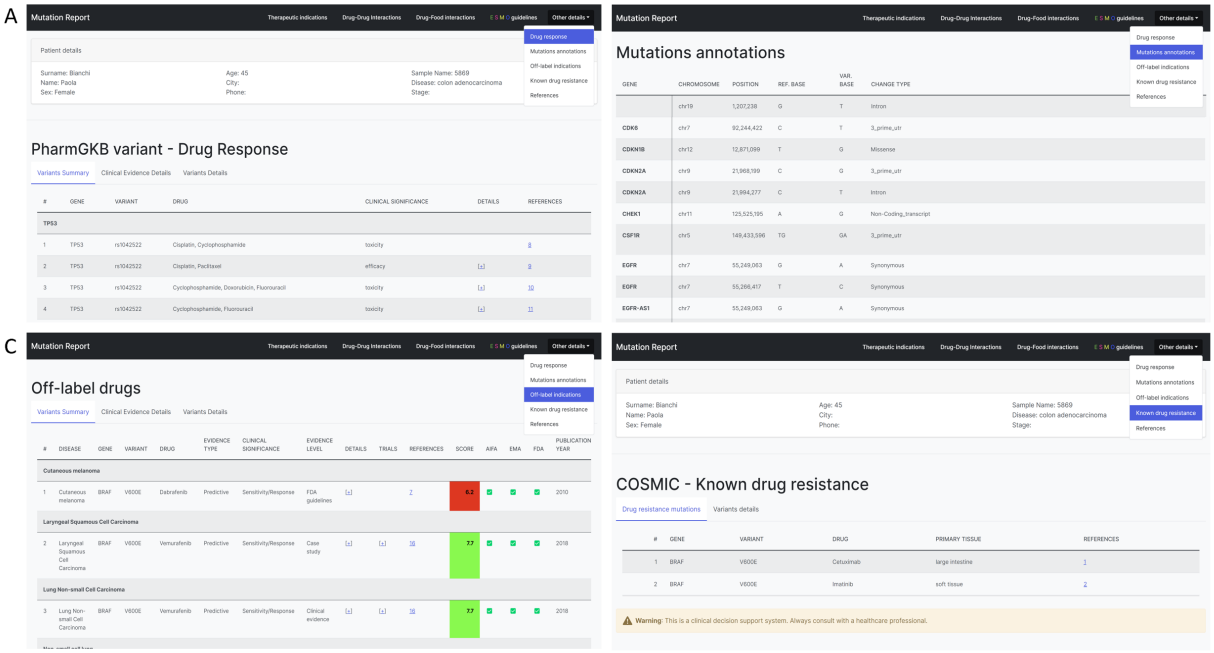

Figure 2.5: Report. A) Drug Response section with PharmGKB database's information about the mutations found in the patient B) Annotation of all the mutations found in the patient C) Off label Drugs D) Known Mutations Resistance found in Cosmic database

2.3 Databases

Oncoreport employs several free databases to understand the role of the mutations found in the patient. Most of them include databases used for coupling variants with drugs to find the perfect therapy for the patient's mutational profile. Each of these databases was modified in a specific way to be used inside the pipeline. A summary of these can be seen in table 2.1.

Figure 2.6: Databases used by Oncoreport

CIViC Clinical Interpretation of Variants in Cancer [149] *(CIViC)* is an open source database that describes the therapeutic, prognostic, diagnostic and predisposing relevance of inherited and somatic variants of all types. It was released in 2015. Currently it contains 2969 variants and 460 genes for 8441 items. Each of its interpretations are matched with fundamental information for the study of the variable such as the disease were we can find the variant, its clinical action, its clinical significance, its evidence type and its evidence level. The association between the gene and the variant has different levels depending on the force of their relation. In particular we found 5 levels:

(A) Validated associations: those which have proven/consensus associations in human medicine;

(B) Clinical evidences: associations supported by Clinical trials or other primary patients data;

(C) Case study: Variants found in case reports from clinical journals;

(D) Preclinical evidences: associations supported by in vivo or in vitro models;

(E) Inferential associations: Indirect evidences.

Each *CIViC* mutation is manually curated. Every month there is a new *CIViC* release, but it is also

possible to download a nightly version of the database. The download can be done both directly or via API.

Cancer Genome Interpreter Cancer Genome Interpreter *(CGI)* [150] is a free platform that annotates all variants of the tumor that constitute state-of-the-art biomarkers of drug response organized using different clinical evidence. It comprises 5601 validated oncogenic alterations, 1631 biomarkers of drug response, 765 cancer genes. Its catalog, which it is downloadable directly or via API, was obtained using bioinformatic analysis and manually curated literature. As in *CIViC* each gene is annotated with its mode of action in tumorigenesis, the diseases and the drugs that act on it.

PharmGKB Pharmacogenomics Knowledgebase *(PharmGKB)* [147] is a resource that collects and curates information about human genetic variant and their drug responses. It provides clinically relevant information such as dosing guidelines, annotated drug labels, potentially actionable gene-drug associations and genotype-phenotype relationships. Also here the gene-drug-disease relationships are extracted from literature using manual curation and natural-language-processing techniques. We took the feature with the relationship between variant and drug followed by the attribute of sensitivity or resistance and the respective pubmed reference. As in *CIViC* each clinical association has a level of evidence that goes from 1, a clinical evidence, to 4 case report or in vitro study associations. *PharmGKB* was enriched using ensembl to add the position of the variant and the alternative base and using the efetch API to add literature information.

Cosmic The Catalogue Of Somatic Mutations In Cancer *(Cosmic)* [148] is a resource for exploring the effects of somatic mutations in human cancer. It exists since 2004 when it was only a survey of four genes. In the actual version (v95) it comprises 9.215.470 gene expression variants curated over 28.551 papers. It covers non-coding mutations, gene fusions, copy-number variants and drug-resistance mutations. In particular, we use the drug-resistance mutations database to search for the drugs that can be deleterious for the patient's health. *Cosmic* includes this database since 2016. From Cosmic we took two files, one that contains all the known mutation related to cancer and one which consists on all the resistance mutations. These two are merged to identify the variants that have a correlation of resistance with a specific drug.

ClinVar *ClinVar* [48] provides a freely available archive of reports of relationships among medically important variants and phenotypes. It gives the interpretations of the relationship of the variation to human health and the evidence supporting each interpretation. It was created to provide a centralized, public open-access database for data needed to interpret variants. It was created

in 2012. We use it to give an interpretation of all the mutations found in each patients along with *RefSeq*.

RefSeq NCBI Reference Sequence Database *(Refseq)* [151] is a project that maintains and curates publicly available databases at the National Center for Biotechnology Information (NCBI).

DrugBank *DrugBank* [145, 152] is a freely available web resource which contains information about drugs. It has been created in 2006. It evolved year by year, in its first version it included drugs data with their drugs targets. In the last versions it was updated adding pharmacogenomic data, molecular data, pharmacometabolomics data, pharmacotranscriptomics data, pharmacoproteomics data and so on. At the moment it possess information about 14594 drugs. The aim of *DrugBank* is to help to achieve major advancement in medicine industry.

Database	Full Name	Description	Link	Reference	Version
CIViC	Clinical Interpretation of Variant in Cancer	«CIViC is an open access, open source, community-driven web resource for Clinical Interpretation of Variants in Cancer.»	https://civicdb.org/home	https://doi.org/10.1038/ng.3774	December 2021 (latest)
CGI	Cancer Genome Interpreter	«CGI identifies potentially oncogenic alterations and it flags genomic biomarkers of drug response with different levels of clinical relevance.»	https://www.cancergenomeinterpreter.org/home	https://doi.org/10.1101/140475	2020
COSMIC	Catalogue Of Somatic Mutations	«COSMIC, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.»	https://cancer.sanger.ac.uk/cosmic	https://doi.org/10.1093/nar/gky1015	V95 (latest)
PharmGKB	Pharmacogenomics Knowledgebase	«PharmGKB is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers.»	https://www.pharmgkb.org/	https://doi.org/10.1038/clpt.2012.96	2020
ClinVar		«ClinVar aggregates information about genomic variation and its relationship to human health. »	https://www.ncbi.nlm.nih.gov/clinvar/	https://doi.org/10.1093/nar/gkv1222	2021-10-25 (latest)
Refseq	NCBI Reference Sequence Database	«Refseq is a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.»	https://www.ncbi.nlm.nih.gov/refseq/	https://doi.org/10.1093/nar/gkr1079	2021-12-08 (latest)
ESMO	European Society for Medical Oncology	«ESMO is the leading European professional medical oncology organization. Its Clinical Practice Guidelines (CPG) are intended to provide the user with a set of recommendations for the best standards of cancer care, based on the findings of evidence-based medicine.»	https://www.esmo.org/		2021 (latest)
DrugBank		«DrugBank is a unique bioinformatics/cheminformatics resource that combines detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information.»	https://go.drugbank.com/	https://doi.org/10.1093/nar/gkj067	5.1.8 (latest)

Table 2.1: OncoReport Databases information

Chapter 3

Tumor Mutational Burden

Aim of the TMB research We sought to better comprehend the TMB calculation in particular in colon cancer, trying to simplify its calculation designing a specific pipeline, testing a commercial panel and designing a new possible signature. Our research starts from survival analysis and arrives to enrichment analysis with the purpose of enhancing our knowledge about TMB and patient classified as TMB high (H-TMB) and TMB low (L-TMB).

Dataset To reach this goal we have used The Cancer Genome Atlas (TCGA) samples of several cancers. We downloaded raw samples of Colon adenocarcinoma (COAD, $n = 298$) which has been used as main tumor in our study. BAM files of tumor and normal tissues biopsies were analyzed to extract the somatic mutations. Other solid cancer types downloaded were: Ovarian serous cystadenocarcinoma (OV, $n = 441$); Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, $n = 305$); Thyroid carcinoma (THCA, $n = 496$); Bladder Urothelial Carcinoma (BLCA, $n = 412$); Uterine Corpus Endometrial Carcinoma and Uterine Carcinosarcoma (UCEC and UCS, $n = 628$); Esophageal carcinoma (ESCA, $n = 181$); Kidney renal papillary cell carcinoma (KIRP, $n = 288$); Kidney renal clear cell carcinoma (KIRC, $n = 339$); Liver hepatocellular carcinoma (LIHC, $n = 415$); Stomach adenocarcinoma (STAD, $n = 450$); Pancreatic adenocarcinoma (PAAD, $n = 183$); Prostate adenocarcinoma (PRAD, $n = 497$); Adrenocortical carcinoma (ACC, $n = 240$); Skin Cutaneous Melanoma (SKCM, $n = 466$); Lung Squamous Cell Carcinoma (LUSC, $n = 494$); Lung Adenocarcinoma (LUAD, $n = 512$). For these we directly downloaded the VCF samples supplied by TCGA. We also used, as independent dataset to test our signature, a dataset of 101 WES samples of breast cancer obtained from dbGaP [153]. Such samples were analyzed with our pipeline starting from the hg19 BAM.

3.1 TMBCalc

The analysis of patients samples for the calculation of TMB takes a lot of time since it is necessary to use two samples per patient, one normal and one tumoral. The calculation of TMB is done through the somatic mutations which are the ones that cause the creation of neoantigens. The somatic mutations are the mutation that we can find in the tumor sample excluding the germline mutations already existing in the normal sample. Since a standardized pipeline for the study of TMB does not exist we decided to create a pipeline using already existing tools Fig. 3.1. Our pipeline can be started through FASTQ or BAM files. It comprises four modules listed below.

1. *Alignment:* The samples have been aligned with *bowtie2* [54] using the Genome assembly hg38. It is also possible to use hg19 reference genome.

2. *Bam Processing:* Each bam was modified adding the right readgroups, sorting and cleaning it erasing the possible duplicates derived from the NGS analysis using picard.

3. *Variant calling:* The Variant calling has been done using *Gatk Mutect2* with its subsequent Filtration, keeping only the mutation labeled as "PASS"; and *VarScan* with the specific command "somatic" and the consecutive "somaticFilter" with min-var-freq set to 10. A feature of *Varscan* ("processSomatic") allows to split the vcf file in several vcf with two principal groups, snp and indel. This allows users to apply some filter only to the former one. To get a more precise variant calling, the two vcf resulting from the callers at the end are intersected.

4. *Annotation:* The annotation has been done using *Annovar* [110, 47], an annotation tool written in perl that holds several databases helping to annotate variants in gene based way, region-based way or to filter them. We used the databases 1000genome (2015_08) [154], *snp146* [51], *cosmic88* [155], *NHLBI Exome Sequencing Project (ESP6500)* [156]. All the variants found in this databases were filtered out. We have as output two simple test file one with all the variants that contribute to the TMB calculation and one with the TMB calculation itself.

In our study we had to perform a previous step of conversion from bam to fastq by making use of *bedtools'* command bamtofastq after a sorting did with *samtools* for both TCGA and dbGaP samples.

TMB calculation and Thresholds Although the TMB is commonly defined as «the number of the counted non-synonymous mutations that alter the amino acid sequence of a protein» we decided

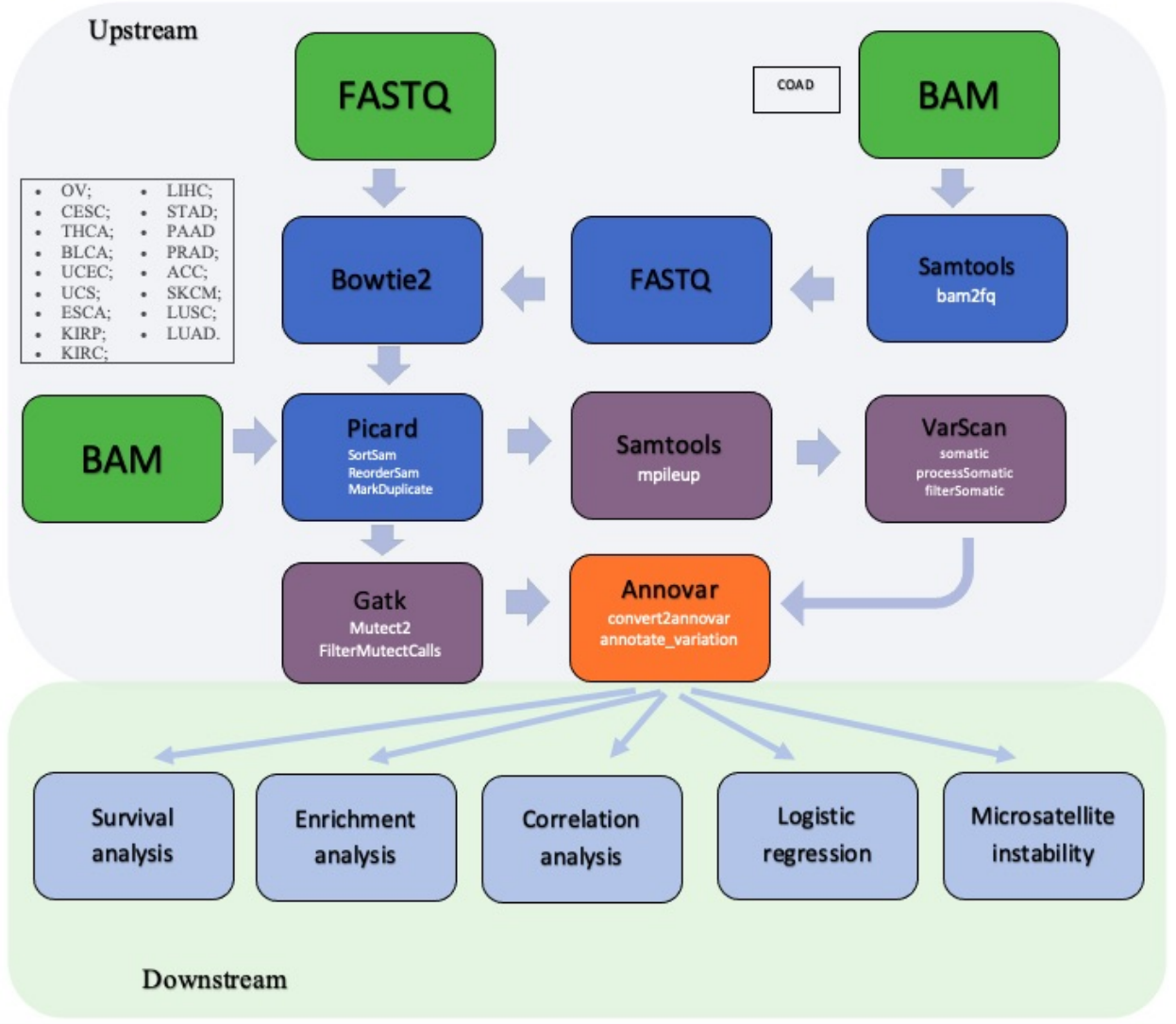

Figure 3.1: Upstream analysis with TMBCalc Pipeline and Downstream analysis for the study of a new gene panel for the TMB research

to include also the synonymous mutations since they result useful to improve sensitivity, because they are explanatory of a mutational process [157, 97]. The TMB has been calculated esteeming the size of genome to 38 MB [157] and applying the formula: $TMB = \frac{\#mutations}{GenomeSize}$. All our analysis on the panels consist of correlation and logistic regression analysis. Regarding Colon cancer each analysis was done for five different threshold found in literature. These threshold were 5, 10, 20, 25.29 and 36.66. The first one was used as the default threshold. For the other tumors we used the threshold of 20 as suggested by Calmers et al. [157] except for Breast Cancer samples of *dbGaP* were we used a threshold of 10 as suggested by Sammons et al. [158] and Meara et al [159].

Panels We perform our analysis on several panels. (i) The Illumina commercial panel named "Ampliseq for Illumina Comprehensive Cancer Panel" that is constituted by 409 cancer-associated genes exons to understand if it was possible to run simultaneously an analysis of mutations and

TMB with this panel used in clinical routine. This panel has a length of 1.7Mb so it exceeds the minimal length suggested in previous work for a TMB panel. (ii) We built several custom panels to understand which are the characteristic that a TMB panel should have. In particular, we built 1000 panels with for each 50, 100, 200 and 300 genes using the gene found in the WES analysis. (iii) Additionally, we built a panel with the 500 most frequently mutated genes.

3.2 Survival Analysis and Microsatellite Instability

First of all we analyzed the overall survival of patients stratified with different TMB thresholds. We performed survival analysis using the R packages survival and survminer generating survival curves. The analysis of MSI were then conducted using the information about the MSI found in TCGA portal. In this occasion the survival analysis on colon cancer patients were done employing the TMB thresholds 5 and the presence or absence of MSI and also with the presence or absence of metastasis. The resultant curves show that there is not a significative difference in the survival both between patients with H-TMB and L-TMB and between patients with TMB lower than the threshold and no MSI and patients with a TMB higher than the threshold and MSI Fig. 3.2 A,B. A significative difference is detected in the Fig. 3.2 C where the patients with H-TMB, MSI and Metastasis have a lower survival. Splitting the patients in High and Low TMB for the Panel of the 500 mostly mutated genes Fig. 3.3A the difference of survival is significative, with a decrease of the survival for the H-TMB patients with an Hazard Ration (HR) of 1.68. In Fig. 3.3 we can see that with the MSI information the results remain stable with the WES one. For the Ampliseq for Illumina Comprehensive Cancer Panel the curves with and without MSI information had both a not significative p.value Fig. 3.4.

3.3 Panel analysis

Correlation analysis Using the Pearson Correlation without splitting the patient in H-TMB and L-TMB our results show that even a small TMB panel with 50 randomly selected mutations give a strong correlation between WES-TMB and panel TMB. Though in Fig.3.5 we can see that such a correlation of course increases while the number of genes increase in the panel.

Stratifying samples in H-TMB and L-TMB this correlation remain strong only for H-TMB patients. Even if this trend is similar for the 100, 200 and 300 genes panels groups we can see that L-TMB correlation raises when the number of genes increase. Regarding the 500 most frequently mutated genes the correlation coefficient for low TMB patients its higher than 0.7 Fig.3.6. We can conclude that a panel built with the most frequently mutated genes in WES could be a good

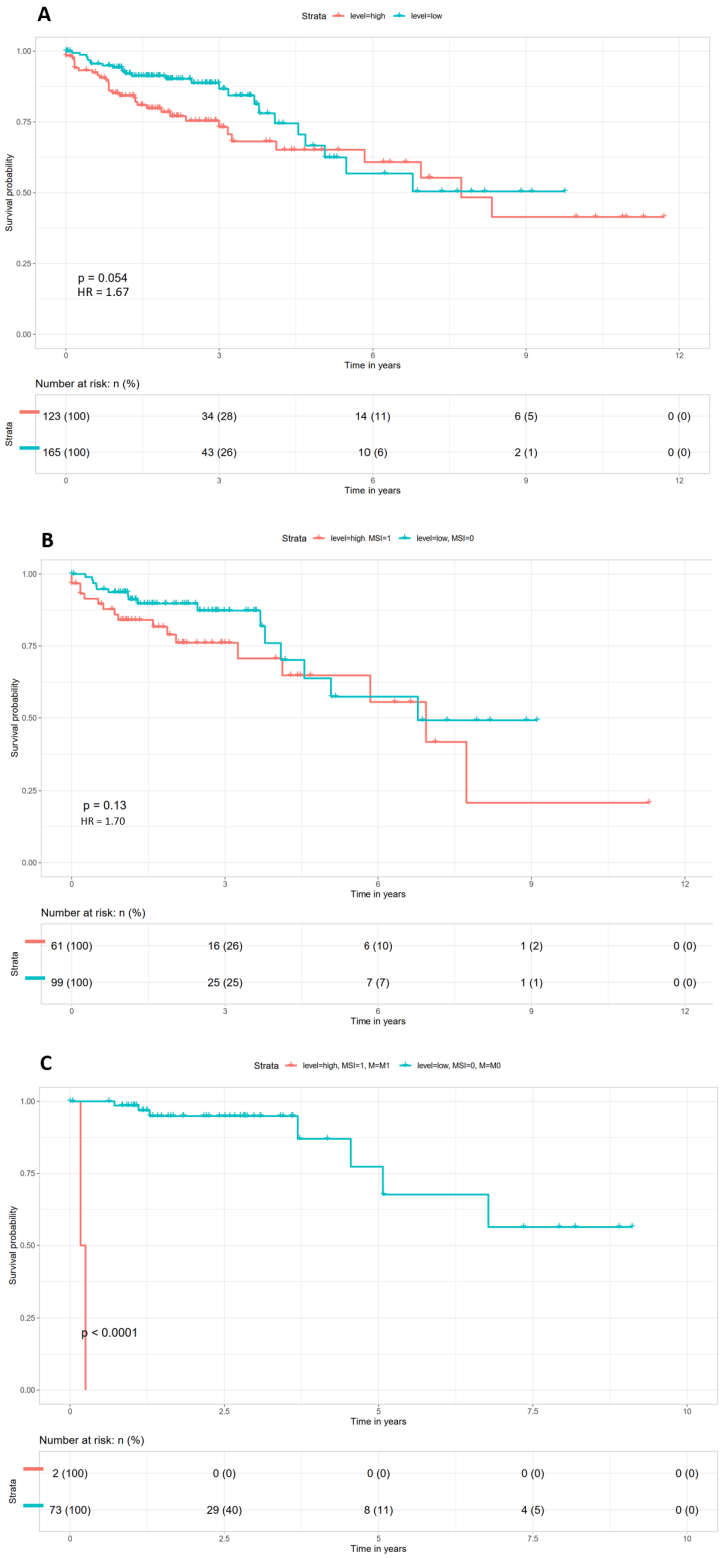

Figure 3.2: Survival curves. A) Survival curves at threshold 5 for all patients with survival information. B) Survival curves with TMB threshold 5 and MSI information. C) Survival curves with threshold 5, MSI and Metastasis information.

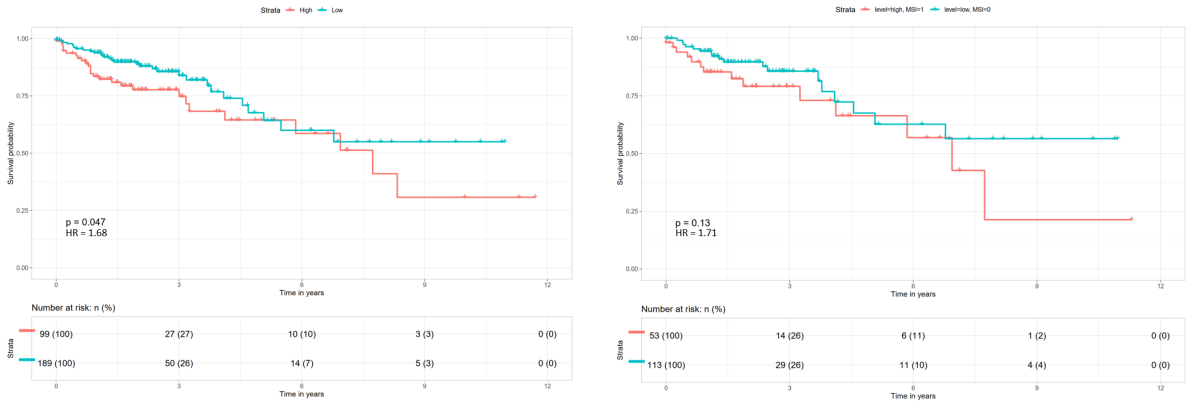

Figure 3.3: Survival curves of the 500 most frequently mutated genes in colon cancer A) With threshold 5. B) With Threshold 5 and MSI information.

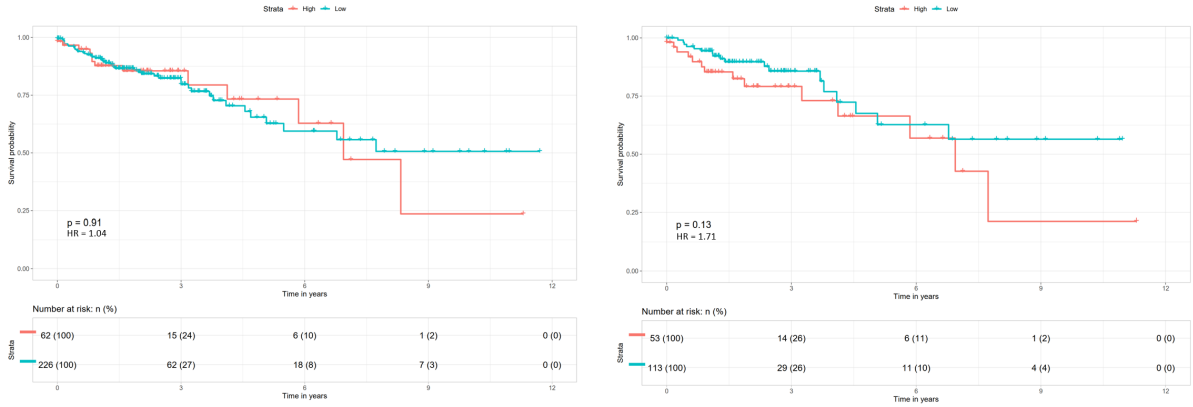

Figure 3.4: Survival curves of the Ampliseq for Illumina Comprehensive Cancer Panel in colon cancer A) With threshold 5. B) With Threshold 5 and MSI information.

Number of Genes (Panel mean length)		**50** *(0.273)*	**100** *(0.548)*	**200** *(1.095)*	**300** *(1.642)*
Pearson	**Mean (SD)**	0.938 (0.017)	0.968 (0.008)	0.983 (0.004)	0.989 (0.003)
Spearman	**Mean (SD)**	0.646 (0.029)	0.686 (0.027)	0.736 (0.026)	0.770 (0.023)
Pearson High TMB	**Mean (SD)**	0.929 (0.021)	0.963 (0.010)	0.981 (0.005)	0.987 (0.003)
Pearson Low TMB	**Mean (SD)**	0.169 (0.067)	0.240 (0.063)	0.328 (0.060)	0.391 (0.056)

Figure 3.5: Correlation between TMB calculated with WES data and TMB calculated with custom panels with 50,100,200 and 300 genes in colon cancer

solution for TMB analysis, but unfortunately, this specific panel has a length of 6.44 Mb resulting too big for the use in clinical. Concerning the "Ampliseq for Illumina Comprehensive Cancer Panel" we can see in Fig. 3.7 that the total correlation is high, but when we split the patients in H-TMB and L-TMB with threshold 5 this correlation decrease dramatically.

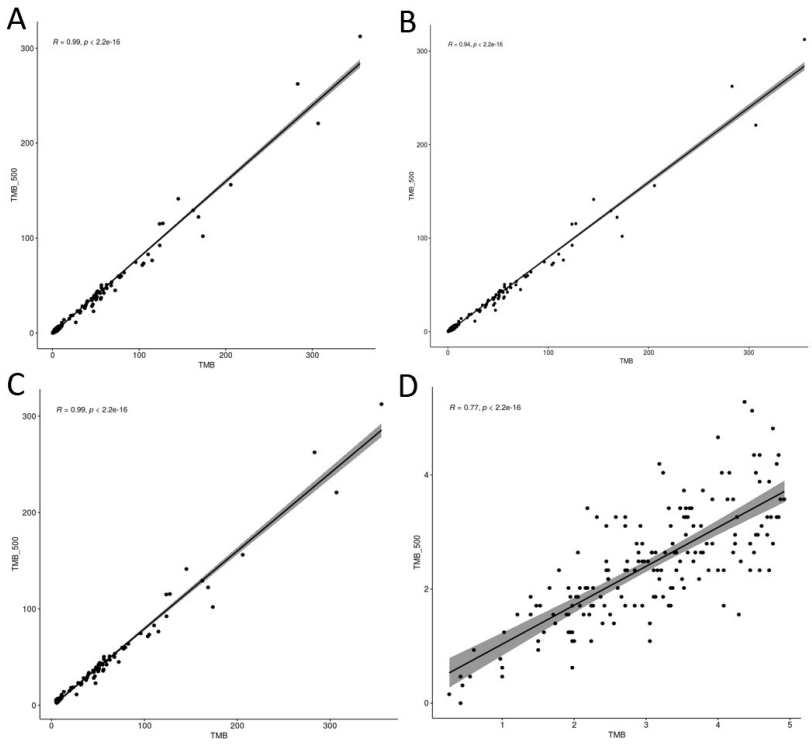

Figure 3.6: Correlation between TMB calculated with WES data and TMB calculated with the custom panel with the 500 most frequently mutated genes in colon cancer with threshold 5. A) Pearson correlation of all patients. B) Spearman correlation of all patients. C) Pearson correlation of H-TMB patients. D) Pearson correlation of L-TMB patients

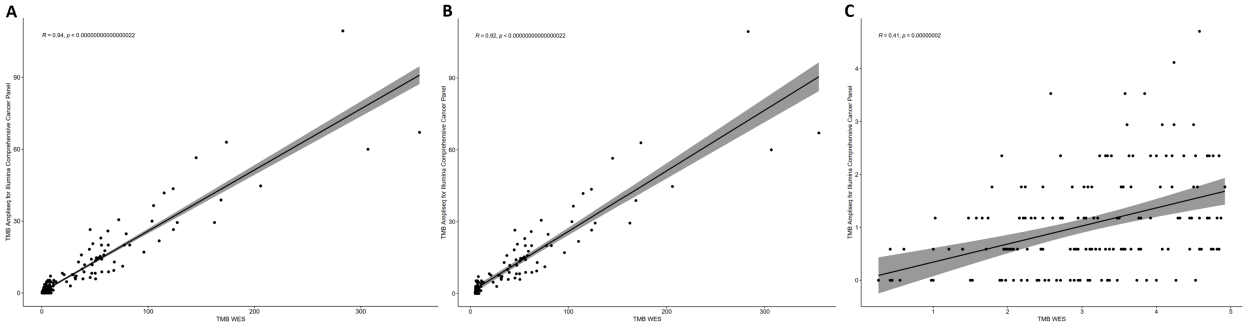

Figure 3.7: Correlation between TMB calculated with WES data and TMB calculated with Ampliseq for Illumina Comprehensive Cancer Panel in colon cancer with threshold 5. A) Pearson correlation of all patients. B) Pearson correlation of H-TMB patients. C) Pearson correlation of L-TMB patients

3.4 Ten most frequently mutated genes in colon cancer

Next we analysed the 19269 mutated genes found in the samples using the refSeq gene Annovar database [160]. Among these we selected the ten top frequently mutated genes in colon cancer: *TTN, SYNE1, MUC19, RYR2, NEB, LRP1B, MUC16, DYNC2H1, RYR3, COL11A1.* Such genes were then analyzed to establish their actual TMB predicting power. For each gene we counted the actual number of mutations in each patient, the number of patients with such mutations, the mean number of mutations for each patient. Then we split the patients according to a TMB threshold, and

we reported the same statistics for each patient group. We can clearly see that the mean number of mutations for each gene in the patient with TMB higher than the threshold is significantly higher than the corresponding one in the low TMB group. By making use of the number of genes mutations in each class (high vs low TMB) we computed Odds Ratio. The results show the strong discriminative power of such signature of genes. In tables 3.1, 3.2, 3.3, 3.4, 3.5 we can see the results for each threshold. This high number of mutations per patients in each of the ten single genes is not explicable with the length of such genes. Except for TTN, that is the longest gene between the ones that contribute to the TMB in WES, the other genes are not the longest. We found SYNE1 as the fifth longest gene, MUC16 as the seventh and the other cannot even be found between the first twenty longest genes. In the past, some authors have already studied the role of several of these genes in TMB samples. For instance Kang et al. found that TTN (72%), MUC16 (67%), and LRP1B (38%) are between the 10 more frequent gene in melanoma. [161]. The same study of the ten most frequently mutated genes was conducted for each of the other sixteen tumors with threshold 20 and it can be seen in Tables 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, 3.14, 3.15, 3.16, 3.17, 3.18, 3.19, 3.20, 3.21.

3.5 Classification Tree and logistic regression

To further understand the actual predictive power of TMB computed using the Ampliseq for Illumina Comprehensive Cancer Panel, we computed the relation between its value and real one (computed using WES). The real TMB was then partitioned in two values: High TMB and Low TMB according to different thresholds ranging from 5 to 34.66. Such a binary variable was used as dependent variable within two different prediction models: decision trees and logistic regression. The reliability of the classification has been computed using a 10-fold cross validation with the *caret* R package. Measures such as Sensitivity (the proportion of positive that are correctly identified, in our case the H-TMB patients), Specificity (the proportion of negatives that are correctly identified, in our case the L-TMB patients) and Accuracy were computed. In table 3.22 we report the results of the classification of the two models using different thresholds. Both models show that using a threshold of 20 the TMB calculated with the panel matches perfectly the one calculated with WES. Therefore, such a panel could be suitable for TMB computation in clinical setting. A second analysis with logistic regression was conducted using the ten most frequently mutated genes (see Table 3.23). This showed that patients can be classified in high and low TMB using only these data. In each of the studied threshold we have a specificity higher than 90% and a sensitivity above the 60%, the overall accuracy was greater than 70%.

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations	Mean of mutations per patient (SD)	Patients with mutations and H-TMB level (Out of 126)	Mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 172)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	369	117 (39%)	3.15 (5.1)	82 (65%)	328	4 (5.9)	35 (20%)	41	1.17 (0.38)	4.6e-15	7.2	4.4e-05
SYNE1	301	109 (36%)	2.76 (3.5)	74 (59%)	263	3.55 (4)	35 (20%)	38	1.08 (0.28)	1.1e-11	5.5	1.6e-06
MUC19	214	91 (30%)	2.35 (2.7)	63 (50%)	180	2.86 (3.1)	28 (16%)	34	1.21 (0.56)	5.3e-10	5.1	1.3e-4
RYR2	209	92 (31%)	2.27 (2.9)	70 (55%)	185	2.64 (3.3)	22 (13%)	24	1.09 (0.93)	3.3e-15	8.4	2.1e-4
NEB	192	68 (23%)	2.82 (3.1)	55 (44%)	177	3.21 (3.4)	13 (7%)	15	1.15 (0.37)	2.4e-13	9.4	4.4e-05
LRP1B	169	88 (29%)	1.92 (1.9)	63 (50%)	142	2.25 (2.2)	25 (14%)	27	1.08 (0.28)	4.8e-11	5.8	8.2e-05
MUC16	168	75 (25%)	2.24 (2.9)	59 (47%)	152	2.58 (3.2)	16 (9%)	16	1 (0)	2.5e-13	8.5	4e-4
DYNC2H1	152	83 (28%)	1.83 (1.6)	67 (53%)	134	2 (1.7)	16 (9%)	18	1.12 (0.35)	< 2.2e-16	11	2.7e-4
RYR3	143	64 (21%)	2.23 (2.8)	52 (41%)	131	2.52 (3.05)	12 (7%)	12	1 (0)	8.6e-13	9.3	7.4e-4
COL11A1	141	74 (25%)	1.90 (2.2)	51 (40%)	118	2.31 (2.6)	23 (13%)	23	1 (0)	1.5e-07	4.4	7.9e-4

Table 3.1: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 5

Ten most frequently mutated genes	Patients with mutations and high TMB level (Out of 66)	Mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 232)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	58 (89%)	301	5.19 (6.7)	59 (25%)	68	1.15 (0.4)	< 2.2e-16	21	2.4e-05
SYNE1	57 (86%)	240	4.21 (4.4)	59 (25%)	61	1.03 (0.5)	< 2.2e-16	21.6	2.9e-06
MUC19	48 (73%)	164	3.42 (3.4)	43 (18%)	50	1.16 (0.5)	4.6e-16	11.6	3.3e-05
RYR2	56 (85%)	169	3.02 (3.6)	36 (18%)	40	1.11 (0.3)	< 2.2e-16	30	2.1-e4
NEB	46 (70%)	166	3.61 (3.6)	22 (9%)	26	1.18 (0.4)	< 2.2e-16	21.6	3.7e-05
LRP1B	49 (74%)	125	2.55 (2.4)	39 (17%)	44	1.13 (0.3)	< 2.2e-16	14.1	1.4e-4
MUC16	45 (68%)	135	3 (3.6)	30 (13%)	33	1.1 (0.3)	< 2.2e-16	14.2	9.6e-4
DYNC2H1	54 (82%)	121	2.24 (1.9)	29 (12%)	31	1.06 (0.3)	< 2.2e-16	30.9	3.2e-05
RYR3	41 (62%)	117	2.85 (3.3)	23 10%	26	1.13 (0.4)	< 2.2e-16	14.7	2.3e-3
COL11A1	40 (61%)	104	2.60 (2.9)	34 (15%)	37	1.09 (0.3)	7.1e-13	8.9	2.1e-3

Table 3.2: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 10

Ten most frequently mutated genes	Patients with mutations and high TMB level (Out of 60)	Mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 238)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	55 (92%)	298	5.42 (6.8)	62 (26%)	71	1.14 (0.4)	< 2.2e-16	31	2.1e-05
SYNE1	55 (92%)	236	4.29 (4.4)	54 (23%)	65	1.20 (0.5)	< 2.2e-16	37	4.1e-06
MUC19	45 (75%)	160	3.55 (3.4)	46 (19%)	54	1.17 (0.5)	9.6e-16	12	3.5e-05
RYR2	53 (88%)	166	3.13 (3.6)	39 (16%)	43	1.10 (0.3)	< 2.2e-16	38	1.8e-4
NEB	46 (77%)	166	3.61 (3.6)	22 (9%)	26	1.18 (0.4)	< 2.2e-16	31	3.7e-05
LRP1B	45 (75%)	121	2.69 (2.4)	43 (18%)	48	1.12 (0.3)	< 2.2e-16	13	9.3e-05
MUC16	42 (70%)	132	3.14 (3.7)	33 (14%)	36	1.09 (0.3)	< 2.2e-16	14	8.2e-4
DYNC2H1	52 (87%)	119	2.89 (1.9)	31 (13%)	33	1.06 (0.2)	< 2.2e-16	42	2.6e-05
RYR3	40 (67%)	116	2.9 (3.4)	24 (10%)	27	1.12 (0.4)	< 2.2e-16	17	2.1e-3
COL11A1	35 (58%)	99	2.83 (3)	39 (16%)	42	1.08 (0.3)	2.8e-10	7	1.7e-3

Table 3.3: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20

Ten most frequently mutated genes	Patients with mutations and high TMB level (Out of 58)	Mutations in high TMB patients - Mean of mutations per high patient		Number and percentage of patients with mutations and low TMB level (Out of 240)	Number of mutations in low TMB patients - Mean of mutations per low patient		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	53 (91%)	293	5.53 (6.9)	64 (27%)	76	1.19 (0.4)	< 2.2e-16	28.8	2.9e-05
SYNE1	54 (93%)	235	4.35 (4.5)	55 (23%)	66	1.20 (0.5)	< 2.2e-16	44.7	3.7e-06
MUC19	44 (76%)	158	3.59 (3.5)	47 (19%)	56	1.19 (0.5)	1.1e-15	12.8	4.3e-05
RYR2	51 88%	164	3.21 (3.7)	41 (17%)	45	1.10 (0.3)	< 2.2e-16	34.7	1.6e-4
NEB	46 79%	166	3.61 (3.6)	22 (9%)	26	1.18 (0.4)	< 2.2e-16	37.1	3.7e-05
LRP1B	45 75%	121	2.69 (2.4)	43 (18%)	48	1.17 (0.3)	< 2.2e-16	15.6	9.3e-05
MUC16	41 71%	131	3.19 (3.7)	34 (14%)	37	1.09 (0.3)	< 2.2e-16	14.4	7.8e-4
DYNC2H1	51 (88%)	118	2.31 (1.9)	32 (13%)	34	1.06 (0.2)	< 2.2e-16	46.3	2.4e-05
RYR3	39 (67%)	115	2.95 (3.4)	25 (10%)	28	1.12 (0.4)	< 2.2e-16	17.4	2.0e-3
COL11A1	33 (57%)	96	2.90 (3.1)	41 (17%)	45	1.10 (0.3)	3.4e-09	6.3	2.1e-3

Table 3.4: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 25.29

Ten most frequently mutated genes	Patients with mutations and high TMB level (Out of 51)	Mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 247)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	49 (96%)	283	5.77 (7.1)	68 (27%)	86	1.26 (0.7)	< 2.2e-16	63.6	5.3e-05
SYNE1	49 (96%)	228	4.65 (4.6)	60 (24%)	73	1.22 (0.5)	< 2.2e-16	75.2	3.7e-06
MUC19	40 (78%)	153	3.82 (3.6)	51 (21%)	61	1.20 (0.5)	6.2e-15	13.8	3.8e-05
RYR2	46 (90%)	156	3.39 (3.8)	46 (19%)	53	1.15 (0.4)	< 2.2e-16	39.5	2.e-4
NEB	44 (86%)	164	3.73 (3.6)	24 (10%)	28	1.17 (0.4)	< 2.2e-16	56.8	2.8e-05
LRP1B	41 (80%)	115	2.80 (2.5)	47 (19%)	54	1.15 (0.7)	< 2.2e-16	17.2	1.5e-4
MUC16	39 (76%)	129	3.31 (3.8)	36 (14%)	39	1.08 (0.3)	< 2.2e-16	18.7	7.1e-4
DYNC2H1	46 (90%)	111	2.41 (2)	37 (15%)	41	1.11 (0.3)	< 2.2e-16	51.1	5.7e-05
RYR3	37 (72%)	112	3.03 (3.5)	27 (11%)	31	1.15 (0.4)	< 2.2e-16	21.1	2.5°-3
COL11A1	31 (61%)	94	3.03 (3.2)	43 (17%)	47	1.09 (0.3)	1.3e-09	7.3	1.9e-3

Table 3.5: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 34.66

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations Out of 408	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 46)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 362)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	212	130 (32%)	1,63 (1.4)	32 (69%)	80	2,5 (2.5)	98 (27%)	132	1,35 0,41	3.2e-08	6.1	1.5e-02
SNHG14	127	98 (24%)	1,29 (0.7)	28 (61%)	48	1,71 (1.2)	70 (19%)	79	1,13 0,11	1.2e-08	6.4	1.5e-02
SYNE1	126	82 (20%)	1,54 (2)	24 (52%)	55	2,29 (3.5)	58 (16%)	71	1,22 0,25	2e-07	5.7	1.5e-01
CSMD3	102	73 (18%)	1,4 (1.5)	17 (37%)	41	2,41 (3)	56 (15%)	61	1,09 0,08	9e-04	3.2	8.6e-02
HMCN1	101	81 (20%)	1,25 (0.6)	20 (43%)	30	1,5 (0.9)	61 (17%)	71	1,16 0,21	1e-04	3.8	1.2e-01
MYHAS	97	71 (17%)	1,37 (1.1)	14 (30%)	29	2,07 (2.2)	57 (16%)	68	0,37 1,19	2.1e-02	2.3	1.6e-01
MUC16	96	72 (18%)	1,33 (0.8)	18 (39%)	35	1,94 (1.2)	54 (15%)	61	1,13 0,15	2.7e-04	3.6	1.4e-02
RYR2	95	75 (37%)	1,27 (0.8)	23 (100%)	34	1,48 (1.3)	52 (29%)	61	1,17 0,18	1.6e-07	5.9	2.8e-01
LRP1B	90	68 (17%)	1,32 (1)	17 (37%)	31	1,82 (1.9)	51 (14%)	59	1,16 0,13	4.5e-04	3.5	1.6e-01
SYNE2	89	65 (16%)	1,37 (0.7)	23 (50%)	40	1,74 (0.8)	42 (12%)	49	0,29 1,67	6e-09	7.5	4.6e-03

Table 3.6: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Bladder Cancer

Ten most frequently mutated genes	#Mutation in all patients	Patients with the mutations (Out of 302)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 43)	Mutations in high TMB patients - Mean of mutations per high patient		Patients with mutations and low TMB level (Out of 259)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	208	90 (30%)	2,31 (5.6)	33 (77%)	145	4,39 (9)	57 (22%)	63	1,1 (0,3)	6e.12	11.6	4.4e-02
DMD	116	82 (27%)	1,41 (1.2)	29 (67%)	58	2 (1.8)	53 (20%)	58	1,09 (0.3)	2.1e-09	8	1.1e-02
MUC16	114	67 (22%)	1,70 (1.5)	27 (63%)	63	2,33 (2.1)	40 (15%)	51	1,27 (0.7)	3.6e-10	9.1	1.7e-02
RYR2	103	60 (20%)	1,72 (3)	26 (60%)	63	2,42 (4.5)	34 (13%)	40	1,18 (0.4)	1.3e-10	10	1.7e-01
NEB	101	62 (20%)	1,63 (1.8)	26 (60%)	60	2,31 (2.7)	36 (14%)	41	1,14 (0.4)	3.4e-10	9.3	3.7e-02
MUC19	98	60 (20%)	1,63 (1.8)	28 (65%)	60	2,14 (2.5)	32 (12%)	38	1,19 (0.5)	1e-12	13	6e-02
SNHG14	90	57 (19%)	1,58 (1.9)	20 (46%)	51	2,55 (3)	37 (14%)	39	1,05 (0.7)	5.6e-06	5.2	4.1e-02
CSMD3	88	66 (22%)	1,33 (0.9)	23 (53%)	42	1,83 (1.4)	43 (17%)	46	1,07 (0.3)	6.8e-07	5.7	1.7e-02
OBSCN	88	47 (15%)	1,87 (2)	21 (49%)	57	2,71 (2.7)	26 (10%)	31	1,19 (0.5)	1.4e-08	8.4	1.8e-02
SYNE1	88	51 (17%)	1,72 (1.9)	24 (56%)	59	2,46 (2.6)	27 (10%)	29	1,07 (0.3)	1.3e-10	10.7	1.7e-02

Table 3.7: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Cervix Cancer

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations	Mean of mutations per patient	Patients with mutations and high TMB level (Out of 38)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 181)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	273	111 (51%)	2.46 (1.9)	35 (92%)	132	3.77 (2.3)	76 (42%)	141	1.85 (1.3)	4.8e-09	15.9	4.6e-05
MUC16	166	96 (44%)	1.73 (1)	27 (71%)	59	2.18 (1.2)	69 (38%)	107	1.55 (0.8)	2.6e-04	3.9	2e-02
SYNE1	95	64 (29%)	1.48 (1)	22 (58%)	41	1.86 (1.3)	42 (23%)	54	1.28 (0.7)	5.3e-05	4.5	6.1e-02
CSMD3	84	64 (29%)	1,31 (0.6)	18 (47%)	26	1,44 (0.6)	46 (25%)	58	1,26 (0.6)	0.01	2.6	0.3
MYHAS	82	64 (29%)	1,28 (0.6)	23 (60%)	38	1,65 (0.9)	41 (23%)	44	1,07 (0.3)	9.3e-06	5.2	5.3e-03
LRP2	79	63 (29%)	1,25 (0.5)	18 (47%)	27	1,5 (0.7)	45 (25%)	52	1,15 (0.4)	9.5e-03	2.7	6.3e-02
UBR4	79	56 (25%)	1,41 (0.9)	20 (53%)	31	1,55 (1.2)	36 (20%)	48	1,33 (0.6)	7.4e-05	4.4	0.4
RYR2	76	63 (29%)	1.21 (0.5)	18 (47%)	21	1.67 (0.4)	45 (25%)	55	1.22 (0.5)	9.5e-04	2.7	0.6
PKHD1	74	61 (28%)	1,21 (0.5)	23 (60%)	31	1,35 (0.6)	38 (21%)	43	1,13 (0.3)	4e-06	5.7	0.1
DST	73	54 (25%)	1.35 (0.6)	15 (39%)	22	1.47 (0.8)	39 (21%)	51	1.31 (0.6)	2.4e-2	2.4	05

Table 3.8: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Ovarian serous cystadenocarcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 180)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 20)	Mutations in high TMB patients - Mean of mutations per high patient		Patients with mutations and low TMB level (Out of 160)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	185	100 (55%)	1.85 (1.1)	19 (95%)	52	2.74 (1.4)	81 (51%)	133	1.64 (0.9)	7.4e-05	18.3	4e-03
MUC19	131	74 (41%)	1.77 (1)	17 (85%)	38	2.23 (1)	57 (36%)	93	1.63 (1)	3.9e-05	10.1	4.3e-02
LRP1B	128	79 (44%)	1.62 (1.1)	14 (70%)	32	2.28 (1.7)	65 (41%)	96	1.47 (0.9)	0.02	3.9	0.1
CSMD3	103	69 (38%)	1,49 (1)	12 (60%)	28	2,33 (1.9)	57 (36%)	75	1,31 (0.5)	0.05	2.7	9e-02
SNHG14	102	64 (35%)	1,59 (1)	15 (75%)	35	2,33 (1.4)	49 (31%)	67	1,37 (0.7)	2.5e-04	6.7	2e-02
SYNE1	96	66 (37%)	1.45 (0.8)	15 (75%)	33	2.2 (1.1)	51 (32%)	63	1.23 (0.5)	3.2e-04	6.3	6e-03
MUC16	85	63 (35%)	1.35 (0.7)	13 (65%)	23	1.77 (1)	50 (31%)	62	1.24 (0.6)	5.1e-03	4	9.2°e02
RYR2	80	59 (33%)	1.35 (0.6)	15 (75%)	25	1.67 (0.9)	44 (27%)	55	1.25 (0.5)	5.1e-05	7.8	0.1
HMCN1	79	62 (34%)	1,27 (0.6)	12 (60%)	18	1,5 (0.9)	50 (31%)	61	1,22 (0.5)	2.2e-02	3.3	0.3
COL11A1	75	58 (32%)	1.29 (0.6)	9 (45%)	16	1.78 (0.8)	49 (31%)	59	1.20 (0.4)	0.2	1.8	7.5e-02

Table 3.9: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Esophageal Carcinoma

Ten most frequently mutated genes	Number of mutation in all patients	Number and percentage of patients with the mutations (Out of 512)	Mean of mutations per patient (SD)	Number and percentage of patients with mutations and high TMB level (Out of 29)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Number and percentage of patients with mutations and low TMB level (Out of 483)	Number of mutations in low TMB patients - Mean of mutations per low patient		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	531	199 (39%)	2.67 (2.7)	29 (100%)	190	6.55 (4.7)	170 (35%)	341	2 (1.3)	3.3e-13	Inf	1.8e-05
LRP1B	481	231 (45%)	2.08 (1.7)	27 (93%)	110	4.07 (3)	204 (42%)	371	1.82 (1.3)	4.1e-08	18.4	5.5e-04
CSMD3	455	225 (44%)	2.02 (1.6)	27 (93%)	112	4.15 (2.7)	198 (41%)	343	1.73 (1)	1.3e-08	19.3	8.8e-05
SNHG14	441	224 (44%)	1.97 (1.4)	25 (86%)	83	3.32 (2.2)	199 (41%	358	1.8 (1.2)	2.3e-06	8.9	2.7e-03
RYR2	439	209 (41%)	2.1 (1.6)	29 (100%)	114	3.93 (3)	180 (37%)	325	1.8 (1)	1.5e-12	Inf	7.5e-04
MUC16	365	171 (33%)	2.13 (1.4)	27 (93%)	105	3.89 (2)	144 (30%)	260	1.8 (1)	6.6e-12	31.6	2.1e-05
USH2A	342	179 (35%)	1.91 (1.4)	25 (86%)	98	3.92 (2.4)	154 (32%)	244	1.58 (0.9)	6.7e-09	13.3	5.4e-05
SPTA1	305	174 (34%)	1.75 (1.1)	26 (90%)	75	2.88 (1.6)	148 (31%)	230	1.55 (0.8)	2.2e-10	19.5	3.7e-04
MYHAS	248	167 (33%)	1.48 (0.9)	19 (65%)	33	1.74 (0.9)	148 (31%)	215	1.45 (0.9)	3e-04	4.3	2.2e-01
COL11A1	229	144 (28%)	1.59 (0.9)	25 (86%)	62	2.48 (1.3)	119 (25%)	167	1.4 (0.6)	2.7e-11	19	5e-04

Table 3.10: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Lung Adenocarcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 494)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 19)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 475)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	1055	341 (69%)	3.09 (2.3)	19 (100%)	135	7.1 (3.2)	322 (68%)	920	2.86 (2)	1.4e-03	Inf	1.6e-05
SNHG14	585	296 (60%)	1.98 (1.5)	15 (79%)	63	4.2 (3.4)	281 (59%)	522	1.86 (1.2)	9.8e-02	2.6	2e-02
CSMD3	473	259 (52%)	1.82 (1.3)	16 (84%)	59	3.69 (3.3)	243 (51%)	414	1.7 (1)	4.5e-03	5.1	2.9e-02
LRP1B	468	254 (51%)	1.84 (1.3)	17 (89%)	59	3.47 (2.7)	237 (50%)	409	1.72 (1)	6.4e-04	8.5	1.8e-02
SYNE1	410	225 (45%)	1.82 (1.1)	18 (95%)	53	2.94 (1.7)	207 (43%)	357	1.72 (1)	5.4e-06	23.2	8.6e-03
MUC16	384	209 (42%)	1.84 (1.2)	17 (89%)	54	3.18 (2.3)	192 (40%)	330	1.72 (1)	2e-05	12.5	2.2e-02
RYR2	367	224 (45%)	1.64 (1)	15 (79%)	44	2.93 (1.6)	209 (44%)	323	1.54 (0.9)	3.7e-03	4.7	5.3e-03
USH2A	327	214 (43%)	1.53 (0.8)	16 (84%)	29	1.81 (0.8)	198 (42%)	298	1.5 (0.8)	2.6e-04	7.4	1.7e-01
COL22A1	285	189 (38%)	1.51 (1)	11 (58%)	33	3 (1.6)	178 (37%)	252	1.41 (0.8)	9.2e-02	2.3	8.6e-03
SPTA1	282	191 (39%)	1.48 (0.8)	13 (68%)	25	1.92 (1)	178 (37%)	257	1.44 (0.7)	8.3e-03	3.6	1.3e-01

Table 3.11: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Lung Squamous Cancer

Ten most frequently mutated genes	#mutations in all patients	Patients with the mutations (Out of 338)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 3)	Mutations in high TMB patients - Mean of mutations per high patient		Patients with mutations and low TMB level (Out of 335)	Number of mutations in low TMB patients – Mean of mutations per low patient		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	74	48 (14%)	1.54 (2.3)	3 (100%)	20	6.67 (9)	45 (12%)	54	1.2 (0.6)	2.7e-03	Inf	4e-01
PBRM1	64	61 (18%)	1.05 (0.2)	1 (33%)	2	2 NA	60 (18%)	62	1.03 (0.2)	4.5e-01	2.3	NA
LRP2	50	43 (13%)	1.16 (0.5)	2 (67%)	4	2 (1.4)	41 (12%)	46	1.12 (0.4)	4.4e-02	14.1	5.4e-01
NEB	39	27 (8%)	1.44 (1.9)	3 (100%)	14	4.67 (5.5)	24 (7%)	25	1.04 (0.2)	4.6e-04	Inf	3.7e-01
RNR2	38	37 (11%)	1.03 (0.2)	0 (0%)	0	0 (0)	37 (11%)	38	1.03 (0.2)	1	0	NA
COL11A1	37	30 (9%)	1.23 (0.9)	1 (33%)	6	6 (NA)	29 (9%)	31	1.07 (0.3)	2.4e-01	5.2	NA
HMCN1	37	26 (8%)	1.42 (1.2)	2 (67%)	8	4 (4.2)	24 (7%)	29	1.21 (0.5)	1.6e-02	25.2	5.2e-01
VHL	37	37 (11%)	1 (0)	0 (0%)	0	0 (0)	37 (11%)	37	1 (0)	1	0	NA
SYNE1	36	30 (9%)	1.2 (0.8)	1 (33%)	5	5 (NA)	29 (9%)	31	1.07 (0.3)	2.4e-01	5.2	NA
SYNE2	33	28 (8%)	1.18 (0.6)	2 (67%)	5	2.5 (2.1)	26 (8%)	28	1.08 (0.3)	1.9e-02	23.2	5.2e-01

Table 3.12: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Kidney Renal Clear Cell Carcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 287)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 4)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 283)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	87	66 (23%)	1.32 (0.7)	4 (100%)	10	2.5 (1.3)	62 (22%)	77	1.24 (0.6)	2.6e-03	Inf	1.4e-01
LRP2	62	51 (18%)	1.21 (0.5)	3 (70%)	5	1.67 (1.1)	48 (17%)	57	1.19 (0.4)	1.9e-02	14.5	5.5e-01
SYNE1	48	36 (12%)	1.33 (0.5)	4 (100%)	5	1.25 (0.5)	32 (11%)	43	1.34 (0.5)	2.1e-04	Inf	7.4e-01
CUBN	44	37 (13%)	1.19 (0.4)	1 (25%)	2	2 NA	36 (13%)	42	1.17 (0.4)	4.3e-01	2.3	NA
SYNE2	44	39 (13%)	1.13 (0.5)	3 (75%)	6	2 (1.7)	36 (13%)	38	1.05 (0.2)	8.5e-03	20.2	4.4e-01
UBR4	41	35 (12%)	1.17 (0.4)	2 (50%)	3	1.5 (0.70)	33 (12%)	38	1.15 (0.4)	7.4e-02	7.5	6.1e-01
NEB	36	31 (11%)	1.16 (0.4)	2 (50%)	2	1 (0)	29 (10%)	34	1.17 (0.4)	6e-02	8.6	2.2e-02
PKHD1	36	33 (11%)	1.09 (0.3)	2 (50%)	2	1 (0)	31 (11%)	34	1.1 (0.3)	6.6e-02	8	8.3e-02
OBSCN	35	33 (11%)	1.06 (0.2)	3 (75%)	5	1.67 (0.6)	30 (11%)	30	1 (0)	5.1e-03	24.7	1.8e-01
DST	32	29 (10%)	1.1 (0.3)	1 (25%)	2	2 (NA)	28 (10%)	30	1.07 (0.3)	3.7e-01	2.8	NA

Table 3.13: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Kidney Renal Papillary Cell Carcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 398)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 12)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 386)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
ALB	187	133 (33%)	1,41 (0.8)	5 (42%)	12	2,4 (1.1)	128 (33%)	175	1,37 (0.8)	5.4e-01	1.4	1.1e-01
TTN	150	109 (27%)	1.38 (0.7)	9 (75%)	18	2 (1.6)	100 (26%)	132	1.32 (0.6)	6.8e-04	8.5	2.5e-01
CSMD3	121	86 (22%)	1,41 (0.8)	7 (58%)	18	2,57 (1.9)	79 (20%)	103	1,3 (0.6)	5.4e-03	5.4	1.3e-01
RYR2	117	90 (23%)	1.3 (0.7)	8 (67%)	20	2.5 (1.2)	82 (21%)	97	1.18 (0.4)	1.1e-03	7.4	1.7e-02
MUC19	112	87 (22%)	1.28 (0.7)	9 (75%)	19	2.11 (1.6)	78 (20%)	93	1.19 (0.4)	9.8e-05	11.7	1.3e-01
OBSCN	187	133 (33%)	1.4 (0.8)	5 (42%)	12	2.4 (1.1)	128 (33%)	175	1.4 (0.8)	5.4e-01	1.4	8.2e-02
HMCN1	101	70 (17%)	1,44 (1)	7 (58%)	23	3,28 (2.3)	63 (16%)	78	1,24 (0.5)	1.5e-03	7.1	5.6e-02
COL11A1	96	76 (19%)	1.26 (0.7)	7 (58%)	17	2.43 (1.6)	69 (18%)	79	1.14 (0.4)	2.5e-03	6.4	8e-02
MUC16	96	79 (20%)	1.21 (0.5)	8 (67%)	14	1.75 (0.9)	71 (18%)	82	1.15 (0.4)	4.4e-04	8.8	1e-01
LRP1B	95	76 (19%)	1.25 (0.7)	7 (58%)	15	2.14 (1.8)	69 (18%)	80	1.16 (0.4)	2.5e-03	6.4	1.9e-01

Table 3.14: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Liver Hepatocellular Carcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 176)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level Out of 4	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level Out of 172	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	91	22 (12%)	4,14 (11.7)	4 (100%)	70	17,5 (25.8)	18 (10%)	21	1,17 (0.4)	1.9e-04	Inf	2.9e-01
SYNE1	48	15 (8%)	3,2 (7.7)	4 (100%)	37	9,25 (14.6)	11 (6%)	11	1 (0)	3.5e-05	Inf	3.4e-01
MUC16	45	15 (8%)	3 (7.5)	2 (50%)	32	16 (19.8)	13 (7%)	13	1 (0)	3.7e-02	11.8	4.8e-01
HMCN1	37	11 (6%)	3,36 (7.2)	3 (75%)	28	9,33 (13.6)	8 (5%)	9	1,12 (0.3)	7.1e-04	56.5	4e-01
LRP1B	37	18 (10%)	2,05 (3.8)	4 (100%)	22	5,5 (7.7)	14 (8%)	15	1,07 (0.03)	7.9e-05	Inf	3.3e-01
SNHG14	36	15 (8%)	2,4 (4.9)	4 (100%)	24	6 (9.3)	11 (6%)	12	1,1 (0.3)	3.5e-05	Inf	3.7e-01
MACF1	35	9 (5%)	3,89 (6.1)	4 (100%)	30	7,5 (8.3)	5 (3%)	5	1 (0)	3.3e-06	Inf	2.2e-01
RYR3	35	14 (8%)	2,5 (4)	3 (75%)	23	7,67 (7.2)	11 (6%)	12	1,09 (0.3)	1.5e-03	41.3	2.6e-01
USH2A	34	13 (7%)	2,61 (5)	4 (100%)	24	6 75,33	9 (5%)	10	1,11 (0.33)	1.8e-05	Inf	3.4e-01
NEB	33	11 (6%)	3 (6.3)	4 (100%)	26	6,5 107	7 (4%)	7	1 (0)	8.5e-06	Inf	3.6e-01

Table 3.15: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Pancreatic adenocarcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 465)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 256)	Mutations in high TMB patients - Mean of mutations per high patient (SD)		Number and percentage of patients with mutations and low TMB level (Out of 209)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	1877	350 (75%)	5,36 (7.2)	239 (93%)	1685	7,05 (8.2)	111 (53%)	192	1,73 (1)	< 2.2e-16	12.3	< 2.2e-16
MUC16	1058	286 (62%)	3,7 (4.8)	210 (82%)	951	4,53 (5.3)	76 (36%)	107	1,41 (0.7)	< 2.2e-16	8	8.4e-15
SNHG14	856	286 (61%)	2,99 (3.4)	208 (81%)	737	3,54 (3.8)	78 (37%)	119	1,52 (0.8)	< 2.2e-16	7.2	6.2e-12
MYHAS	793	266 (57%)	2,98 (3.4)	208 (81%)	713	3,43 (3.7)	58 (28%)	80	1,38 (0.6)	< 2.2e-16	11.2	6.2e-13
DNAH5	773	269 (58%)	2,87 (3.2)	203 (79%)	686	3,38 (3.5)	66 (31%)	87	1,32 (0.6)	< 2.2e-16	8.2	7.5e-14
MGAM	745	280 (60%)	2,66 (3.1)	203 (79%)	637	3,14 (3.5)	77 (37%)	108	1,4 (0.7)	< 2.2e-16	6.5	1.3e-10
LRP1B	600	234 (50%)	2,56 (3.2)	181 (71%)	535	2,95 (3.5)	53 (25%)	65	1,23 (0.6)	< 2.2e-16	7	1.3e-09
CSMD2	546	236 (51%)	2,31 (2.5)	184 (72%)	482	2,62 (2.7)	52 (25%)	64	1,23 (0.6)	< 2.2e-16	7.7	8.9e-10
DNAH9	449	206 (44%)	2,18 (2.7)	163 (64%)	392	2,4 (3)	43 (20%)	57	1,32 (0.7)	< 2.2e-16	6.7	4e-05
RYR1	448	213 (46%)	2,1 (2.4)	160 (62%)	385	2,41 (2.7)	53 (25%)	63	1,19 (0.4)	8.4e-16	4.9	2.1e-07

Table 3.16: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Skin Cutaneous Melanoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 496)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 2)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 494)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
SYNE1	55	38 (8%)	1.45 (2.6)	1 (50%)	17	17 (NA)	37 (7%)	38	1.03 (0.2)	0.15	12.2	NA
TTN	54	33 (7%)	1.64 (3.1)	2 (100%)	21	10.5 (12)	31 (6%)	33	1.06 (0.2)	4.3e-03	Inf	4.7e-01
LRP1B	41	31 (6%)	1.32 (1.4)	2 (100%)	10	5 (5.6)	29 (6%)	31	1.07 (0.3)	3.8e-03	Inf	5e-01
SNHG14	37	25 (5%)	1.48 (2.2)	2 (100%)	13	6.5 (7.8)	23 (5%)	24	1.04 (0.2)	2.4e-03	Inf	5e-01
MUC16	35	22 (4%)	1.59 (1.9)	2 (100%)	11	5.5 (6.4)	20 (4%)	24	1.2 (0.4)	1.9e-03	Inf	5e-01
CSMD3	32	30 (6%)	1.07 (0.4)	1 (50%)	3	3 (NA)	29 (6%)	29	1 (0)	0.12	15.8	NA
HMCN1	30	27 (5%)	1.11 (0.4)	2 (100%)	4	2 (1.4)	25 (5%)	26	1.04 (0.2)	3.7e-03	Inf	5e-01
MYHAS	30	21 (4%)	1.43 (1.1)	2 (100%)	7	3.5 (3.5)	19 (4%)	23	1.21 (0.4)	3.1e-03	Inf	5.3e-01
KMT2D	27	22 (4%)	1.23 (0.7)	1 (50%)	4	4 (NA)	21 (4%)	23	1.09 (0.3)	8.7e-02	22	NA
WASL	26	25 (5%)	1.04 (0.2)	0 (0%)	0	0 (0)	25 (5%)	26	1.04 (0.2)	1	0	NA

Table 3.17: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Prostate adenocarcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 494)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 13)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 481)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	105	33 (7%)	3.18 (7.3)	11 (85%)	77	7 (12)	22 (4%)	28	1.27 (0.5)	1.4e-12	111	1.4e-01
SYNE1	57	27 (5%)	2.11 (2.7)	8 (61%)	34	4.25 (4.4)	19 (4%)	23	1.21 (0.4)	2.9e-08	37.9	9.2e-02
HMCN1	56	23 (5%)	2.43 (3.7)	9 (69%)	37	4.11 (5.5)	14 (3%)	19	1.36 (0.8)	1.2e-10	72.1	1.7e-01
MACF1	47	22 (4%)	2.14 (3.2)	7 (54%)	28	4 (5.4)	15 (3%)	19	1.27 (0.6)	1.8e-07	35.2	2.3e-01
KIAA1109	45	19 (4%)	2.37 (2.1)	11 (85%)	36	3.27 (2.4)	8 (2%)	9	1.12 (0.3)	6e-16	297.8	1.3e-02
TG	45	39 (8%)	1.15 (0.5)	4 (31%)	6	1.5 (1)	35 (7%)	39	1.11 (0.4)	1.4e-02	5.6	1.5e-01
MUC16	44	17 (3%)	2.59 (4.1)	10 (77%)	37	3.7 (5.1)	7 (1%)	7	1 (0)	2.4e-14	207.2	1.3e-01
SNHG14	40	25 (5%)	1.6 (1.5)	7 (54%)	19	2.71 (2.5)	18 (4%)	21	1.67 (0.5)	5e-07	29.3	1.5e-01
UTP20	38	16 (3%)	2.37 (2.6)	10 (77%)	30	3 (3.2)	6 (1%)	8	1.33 (0.5)	1e-14	239.9	1.3e-01
MYHAS	36	17 (3%)	2.12 (2.5)	11 (85%)	30	2.72 (3)	6 (1%)	6	1 (0)	< 2.2e-16	392.3	8.2e-02

Table 3.18: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Thyroid carcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 415)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 84)	Number of mutations in high TMB patients - Mean of mutations per high patient		Number and percentage of patients with mutations and low TMB level (Out of 331)	Mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	387	176 (42%)	2,2 (2.2)	71 (84%)	216	3,04 (3.1)	105 (32%)	171	1,63 (1.1)	< 2.2e-16	11.7	4e-04
SYNE1	308	141 (34%)	2,18 (1.8)	65 (77%)	199	3,06 (2.3)	76 (23%)	109	1,43 (0.8)	< 2.2e-16	11.4	6.2e-07
LRP1B	266	154 (37%)	1,73 (1.4)	58 (69%)	124	2,14 (2.1)	96 (29%)	142	1,47 (0.7)	3e-11	5.4	2.7e-02
SNHG14	260	139 (33%)	1,87 (1.3)	63 (75%)	158	2,51 (1.5)	76 (23%)	102	1,34 (0.6)	< 2.2e-16	10	2.5e-07
CSMD3	248	136 (33%)	1,82 (1.7)	58 (69%)	150	2,59 (2.3)	78 (23%)	98	1,26 (0.6)	1.7e-14	7.2	6.7e-05
MYHAS	189	99 (24%)	1,91 (1.7)	57 (68%)	141	2,47 (2)	42 (13%)	48	1,14 (0.3)	< 2.2e-16	14.4	1e-05
HMCN1	186	117 (28%)	1,59 (1.5)	54 (64%)	114	2,11 (2.1)	63 (19%)	72	1,14 (0.4)	4.7e-15	7.6	1.5e-03
MUC16	167	144 (27%)	1,46 (1.1)	48 (57%)	87	1,81 (1.5)	66 (20%)	80	1,21 (0.5)	1.2e-10	5.3	1e-02
DNAH5	165	97 (23%)	1,7 (1.4)	52 (62%)	112	2,15 (1.8)	45 (13%)	53	1,18 (0.4)	< 2.2e-16	10.2	4e-04
SPTA1	158	110 (26%)	1,44 (1)	43 (51%)	68	1,58 (1.3)	67 (20%)	90	1,34 (0.7)	5.9e-08	4.1	2.8e-01

Table 3.19: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Stomach adenocarcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 227)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 3)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 224)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
NF1	26	24 (10%)	1.1 (0.3)	1 (33%)	2	2 (NA)	23 (10%)	24	1.04 (0.2)	2.9e-01	4.3	NA
MUC16	21	17 (7%)	1.23 (0.6)	3 (100%)	6	2 (1)	14 (6%)	15	1.07 (0.3)	3.5e-04	Inf	2.5e-01
TTN	21	17 (7%)	1.23 (0.7)	2 (67%)	6	3 (1.4)	15 (7%)	15	1 (0)	1.5e-02	26.8	3e-01
MUC19	16	12 (5%)	1.33 (0.6)	3 (100%)	4	1.33 (0.6)	9 (4%)	12	1.33 (0.7)	1.1e-04	Inf	1
RYR1	16	10 (4%)	1.6 (0.7)	3 (100%)	7	2.33 (0.6)	7 (3%)	9	1.28 (0.5)	6.2e-05	Inf	6.3e-02
LRP1	15	9 (4%)	1.7 (1)	1 (33%)	4	4 (NA)	8 (3%)	11	1.4 (0.5)	1.1e-01	13	NA
CSMD1	14	11 (5%)	1.27 (0.6)	2 (67%)	4	2 (1.4)	9 (4%)	10	1.11 (0.3)	6.2e-03	44.7	5.3e-01
ABCA13	13	12 (5%)	1.08 (0.3)	2 (67%)	3	1.5 (0.7)	10 (4%)	10	1 (0)	7.5e-03	40.3	5e-01
HMCN1	13	12 (5%)	1.08 (0.3)	2 (67%)	2	1 (0)	10 (4%)	11	1.1 (0.3)	7.5e-03	40.3	3.4e-01
LRP1B	13	10 (4%)	1.3 (0.7)	2 (67%)	4	2 (1.4)	8 (3%)	9	1.12 (0.3)	5.1e-03	50.2	5.4e-01

Table 3.20: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Adrenocortical carcinoma

Ten most frequently mutated genes	#mutation in all patients	Patients with the mutations (Out of 540)	Mean of mutations per patient (SD)	Patients with mutations and high TMB level (Out of 227)	Number of mutations in high TMB patients - Mean of mutations per high patient (SD)		Patients with mutations and low TMB level (Out of 313)	Number of mutations in low TMB patients - Mean of mutations per low patient (SD)		p.Value Fisher's Exact Test for Count Data	Odds ratio Fisher's Exact Test for Count Data	p.Value Welch Two Sample t-test
TTN	2745	265 (49%)	10,36 (17.2)	197 (87%)	2667	13,54 (18.9)	68 (22%)	78	1,15 (0.4)	< 2.2e-16	23.5	< 2.2e-16
NEB	1503	217 (40%)	6,93 (9.5)	170 (75%)	1445	8,5 (10.2)	47 (15%)	58	1,23 (0.5)	< 2.2e-16	16.7	< 2.2e-16
RYR2	1383	233 (43%)	5,93 (8.6)	175 (77%)	1315	7,51 (9.4)	58 (18%)	68	1,17 (0.4)	< 2.2e-16	14.7	8.5e-16
DMD	1374	216 (40%)	6,36 (9)	169 (74%)	1321	7,82 (9.6)	47 (15%)	53	1,13 (0.3)	< 2.2e-16	16.4	4.9e-16
SYNE1	1333	216 (40%)	6,17 (9)	173 (76%)	1286	7,43 (9.8)	43 (14%)	47	1,09 (0.3)	< 2.2e-16	20	7.2e-15
MUC16	1177	193 (36%)	6,1 (8)	153 (67%)	1126	7,36 (8.6)	40 (13%)	51	1,27 (0.7)	< 2.2e-16	14	5.2e-15
CSMD3	1141	215 (40%)	5,31 (7.7)	165 (73%)	1084	6,57 (8.4)	50 (16%)	57	1,14 (0.5)	< 2.2e-16	13.9	5.9e-14
LRP1B	1088	189 (35%)	5,76 (7.7)	150 (66%)	1045	6,97 (8.2)	39 (12%)	43	1,1 (0.3)	< 2.2e-16	13.6	4.7e-15
HMCN1	1080	193 (53%)	5,59 (7.2)	160 (70%)	1042	6,51 (7.7)	33 (10%)	38	1,15 (0.4)	< 2.2e-16	20.1	1.9e-15
DST	1056	181 (33%)	5,83 (8.1)	153 (67%)	1024	6,69 (8.6)	28 (9%)	32	1,14 (0.4)	< 2.2e-16	20.9	3.6e-13

Table 3.21: Ten most frequently genes percentage, mean, standard deviation, fisher and Welch's t-Test with threshold 20 for Uterus Corpus Endometrial Carcinoma and Uterine Carcinosarcoma

Threshold	Accuracy	Sensitivity	Specificity
RPART			
5	0.76	0.44	1
10	0.98	1	0.98
20	1	1	1
25.29	0.98	1	0.98
34.66	0.97	0.9	0.98
GLM			
5	0.76	0.52	0.94
10	0.98	1	0.98
20	1	1	1
25.29	0.97	0.91	0.98
34.66	0.97	0.9	0.98

Table 3.22: Accuracy, Sensitivity and Specificity of all colon cancer's thresholds calculated with GLM and RPART

Threshold	Accuracy	Sensitivity	Specificity
GLM			
5	0.78	0.6	0.91
10	0.95	1	0.93
20	0.97	0.92	0.98
25.29	0.95	1	0.94
34.66	0.95	0.9	0.96

Table 3.23: Ten most frequently genes GLM results

3.6 The 44 genes panel

For the ten genes signature of each of the seventeen tumors that we have studied, we built a 44 genes panel which has a size of 1.05 Mb. We tested this panel on each of the seventeen tumors using threshold 20. This pan-cancer signature shows promising results for each cancer here under study. Indeed, as shown in table 3.24, the Pearson correlation in all tumors is major than 80%. We performed the Pearson correlation also splitting the samples in H-TMB and L-TMB. The results of this show that the correlation is always higher than 70% in more than 80% of the tumors. Moreover, we calculated the accuracy, the sensitivity and the specificity of this panel in each tumor using glm (Generalized Linear Model). In all tumors the accuracy of the panel is more than 80%. Therefore, this signature is promising to be used as a new TMB specific panel easy to build and analyze since its length. Moreover, we tested this new possible panel on the dataset of 101 samples of breast cancer from *dbGaP*. As we can see in Fig 3.8A the Pearson correlation is 0.77 and if we split the samples in H-TMB (Fig. 3.8B) and L-TMB (Fig. 3.8C) groups with a threshold of 10 [158, 159] we have respectively a correlation of 1 for H-TMB and 0.79 for L-TMB. We can conclude that this panel is the best solution to calculate the TMB with a good accuracy for almost each type of cancer. With this panel is possible to speed the analysis of TMB and cut the cost of it reaching almost the same results as WES.

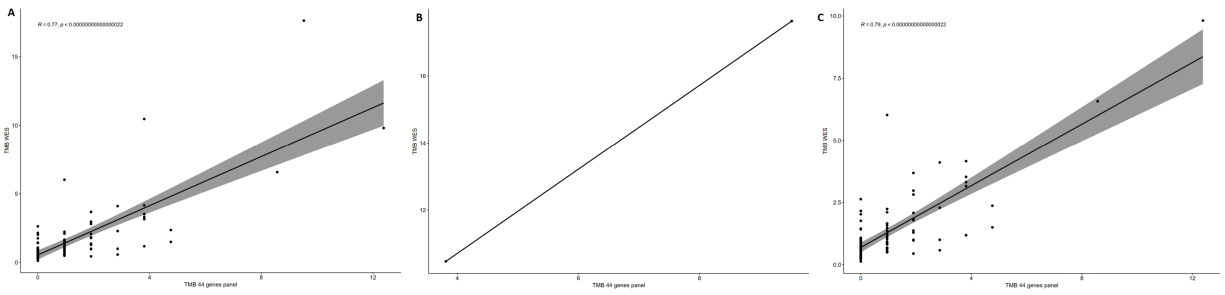

Figure 3.8: Correlation of dbGaP samples TMB calculated with WES and the 44 genes panel. A) All patients. B) High TMB patients. C) Low TMB patients

	Correlation	Correlation H-TMB	Correlation L-TMB	Accuracy	Specificity	Sensitivity
COAD	0.97	0.96	0.74	0.97	1	0.83
ESCA	0.84	0.86	0.79	0.88	0.9	0.75
STAD	0.97	0.97	0.79	0.96	0.97	0.94
SKCM	0.98	0.98	0.86	0.88	0.88	0.88
BLCA	0.93	0.95	0.74	0.97	1	0.78
UCEC-UCS	0.98	0.98	0.77	0.97	1	0.93
OV	0.92	0.88	0.80	0.88	0.92	0.71
CESC	0.99	0.99	0.70	0.96	0.98	0.87
PRAD	0.99	1	0.55	1	1	NA
PAAD	1	1	0.65	1	1	NA
LUAD	0.95	0.90	0.89	0.89	0.91	0.83
LUSC	0.91	0.89	0.78	0.93	0.99	0.72
KIRC	0.98	1	0.60	1	1	NA
KIRP	0.83	0.98	0.74	0.96	0.96	NA
THCA	1	1	0.87	0.99	1	0.5
ACC	0.94	0.93	0.81	1	1	NA
LIHC	0.90	0.83	0.79	0.99	1	0.5

Table 3.24: Correlation, Accuracy, Sensitivity and Specificity of each tumor between TMB analyzed with the panel built with the 44 most mutated genes and WES TMB

3.7 Transcriptome and Enrichment analysis

DEGs analysis on RNA-seq (HTSeq-Counts) data of TCGA has been performed using the R packages TCGAbiolinks and Limma. Two different analyses have been conducted: (i) a comparison between the H-TMB group vs the L-TMB group; (ii) a comparison of the patients with or without one of the ten most frequently genes at the time. Genes with $logFC > 0.6$ and with an adjusted $p.value < 0.05$ were selected. Enrichment analysis has been performed using MITHrIL [162]. The results of the perturbation analysis, showed in Tab 3.25, yielded several perturbed pathways in each of the threshold studied for colon cancer. In particular, among the pathways that are implicated in immunitary and inflamatory response we have found that "Cytokine-cytokine receptor interaction" and "Viral protein interaction with cytokine and cytokine receptor" were upregulated in all threshold, "Chemokine signaling pathway", "Antigen processing and presentation", "Intestinal immune network for IgA production", "Th1 and Th2 cell differentiation", "Th17 cell differentiation", "Natural killer cell mediated cytotoxicity", "Fc epsilon RI signaling pathway", "T cell receptor signaling pathway", "JAK-STAT signaling pathway", "Leukocyte transendothelial migration" have been found in 4 out of 5 threshold and "Toll-like receptor signaling pathway" in 3 out of 5 threshold. Only in threshold 34.66 analysis we also found the pathways "B cell receptor signaling pathway" and "NF-Kappa B signaling pathway".

Pathways	Perturbation	Th 5	Th 10	Th 20	Th 25.29	Th 34.66
Cytokine-cytokine receptor interaction	+	✓	✓	✓	✓	✓
Viral protein interaction with cytokine and cytokine receptor	+	✓	✓	✓	✓	✓
Chemokine signaling pathway	+	×	✓	✓	✓	✓
Antigen processing and presentation	+	×	✓	✓	✓	✓
Intestinal immune network for IgA production	+	×	✓	✓	✓	✓
Th1 and Th2 cell differentiation	+	×	✓	✓	✓	✓
Th17 cell differentiation	+	×	✓	✓	✓	✓
Natural killer cell mediated cytotoxicity	+	×	✓	✓	✓	✓
Fc epsilon RI signaling pathway	+	×	✓	✓	✓	✓
T cell receptor signaling pathway	+	×	✓	✓	✓	✓
JAK-STAT signaling pathway	+	×	✓	✓	✓	✓
Leukocyte transendothelial migration	+	×	✓	✓	✓	✓
Toll-like receptor signaling pathway	+	×	×	✓	✓	✓
B cell receptor signaling pathway	+	×	×	×	×	✓
NF-Kappa B signaling pathway	+	×	×	×	×	✓

Table 3.25: COAD Perturbated Pathways with $p.value \leq 0.05$, analyzed with MITHrIL

Chapter 4

Variant Prioritization

Understanding which variants sustain the tumor development it is a hard path. Databases such as clinvar and intervar are built to help researcher to understand the role of each variant in cancer. Unfortunately, it is not always possible to understand each variant's role and some of them are classified as VUS or as "Unknown". To help to understand the function of them in cancer, we developed an algorithm which prioritized variants found in TCGA for each tumor.

4.1 VarPrAl: Variant Prioritization Algorithm

Dataset used We collected all the VCF samples available in TCGA using as primary site: Adrenal gland (ACC, $n = 240$), Bladder (BLCA, $n = 412$), Brain (LGG, $n = 911$), Breast (BRCA, $n = 1034$), Cervix (CESC, $n = 305$), Colorectal (COAD, $n = 389$), Esophagus (ESCA, $n = 181$), Head and neck (HNSC, $n = 511$), Kidney (KIRP, KIRC, $n = 695$), Liver (LICH, $n = 419$), Lung (LUAD, LUSC, $n = 1067$), Ovary (OV, $n = 411$), Pancreas (PAAD, $n = 183$), Pleura (MESO, DLBC, $n = 115$), Prostate (PRAD, $n = 498$), Skin (SKCM, $n = 470$), Soft Tissue (SARC, PCPG, $n = 126$), Stomach (STAD, $n = 450$), Thyroid (THCA, $n = 496$) and Uterus (UCEC, UCS, $n = 628$). In particular we downloaded all the four types of VCF in TCGA generated with different variant caller, namely *MuSe* [163], *MuTect2* by GATK [141], *VarScan2* [142] and *SomaticSniper* [164]. We intersected the VCFs of the same patient to obtain a more reliable variants calling prediction.

Algorithm Our algorithm is implemented in R and bash, the VCFs are annotated with *Annovar* [47] using the *RefSeq* [160], *intervar* [165], *dbscnv11* and *dbNSFP* v4. [166] databases. The databases *Clinvar* [48], Open Regulatory Annotation database *(ORegAnno)]* [167], the Ensembl regulatory build annotation [168], genome-wide association studies *(GWAS)* [169] and Genotype-Tissue Expression *(GTEx)* [170] are employed in R to obtain a substantial pathogenicity score.

4.1.1 Databases for variant prioritization

- **Clinvar** is a public archive that reports the relationships among human variants and their phenotype. It was employed to classify the mutations in pathogenic, benign and unknown.

- **Intervar** together with Clinvar was employed for the first classification of the variants. It is a tool developed to help to interpret the clinical significance of variants using 18 criteria. We have used the last Annovar Intervar version.

- **ORegAnno** is a resource for curated regulatory annotation. It contains information about distinct regulatory elements such as regulatory regions, transcription factor binding sites, RNA binding sites, regulatory variants, haplotypes and others. It includes annotations schemes that describe both the elements and outcomes of regulatory events. The current release of ORegAnno include for human 261 660 516 bp in the GRCh38/hg38 genome assembly version.

- **GWAS** is a Catalog which delivers a high-quality curated collection of published genome-wide association studies enabling the user to identify causal variants, understand disease mechanism and establish targets for new therapies thanks to the strong association between common genetic variation at loci and human traits. In June 2021 it contained 5106 publications and 258738 associations.

- **GTEx** is a project that has established tissue specific databases to study the relationship between genetic variation and gene expression in human tissues thanks to Expression quantitative trait loci (eQTL). The V8 release include 17382 samples.

- **dbNSFP** The actual version of dbnsfp (v4.1) is based on hg38 and includes 81,782,923 nsSNVs and 2,230,170 ssSNVs and it includes several score predictor. Precisely, we employed the predictions of SIFT, SIFT4G, LRT [171], MutationTaster [172], MutationAssessor [173], FATHMM [174], PROVEAN [175], MetaSVM and MetaLR [176], M.CAP [177], PrimateAI [178], DEOGEN2 [179], BayesDel_addAF, BayesDel_noAF [180], LIST.S2 [181], fathmm-MLK [182] and fathmm.XF [183].

- **dbscnv11** [184] is a database of pre-computed scores for all potential scSNVs across the human genome, advantageous to identify splice-altering scSNVs. This database is fundamental to compute a pathogenicity score because mutations on splicing-site can be very deleterious for protein function.

4.1.2 Mutation classification and Validation

We started our analysis intersecting each patient's mutation set with the *Clinvar* database in R. Next, the variants that could not be found in *Clinvar* were annotated with *Intervar* using *Annovar*. The variants were split in three big categories: (i) Pathogenic, which include "Pathogenic" and "risk_factor" variants from *Clinvar* and "Pathogenic" variants from *intervar*; (ii) Benign, which include "Benign" and "protective" from *Clinvar* and "Benign" from *Intervar* and (iii) Unknown, which include "interpretations_of_pathogenicity","Uncertain_significance", "not_provided", "drug_response", "other", "association" and "Affects" from *Clinvar*, "Uncertain significance" and "Unknown" from *intervar* and also the variants that could not be classify by the two databases. As it can be seen in Fig. 4.1, we noticed by building survival curves that unknown mutations seemed to be more pathogenic or benign in comparison to the ones classified by *Clinvar* and *Intervar*, so we decided to build prioritization databases using these curves. These curves were assembled using the genes of one commercial Illumina panel called "AmpliSeq for Illumina Comprehensive Cancer Panel" and, for colon cancer, also customised prepared matrices built with the most significative gene in colon cancer. These matrices were validated to assure the worth of the method. To validate the matrices we employed a leave-one-out method splitting each of them in a train and a test matrix. The ratio of the training set to the validation set was 2. We tested in this way two different matrices and a group of similar matrices. In detail, a matrix created intersecting the TGA's VCFs the AmpliSeq for Illumina Comprehensive Cancer Panel and, specifically for colon cancer, a series of matrices created using the most meaningful genes founded in OMIM (Online Mendelian Inheritance in Man) [185] and the most frequently mutated genes in this specific cancer type. The significative genes found in OMIM are "PLA2G2A", "NRAS", "BUB1", "CTNNB1", "PIK3CA", "FGFR3", "TLR2", "APC", "MCC", "PTPN12", "DLC1", "PDGFRL", "RAD54B", "PTPRJ", "CCND1", "MLH3", "AKT1", "BUB1B", "TP53", "FLCN", "AXIN2", "DCC", "BAX", "SRC", "AURKA", "EP300", "MSH2", "MLH1", "PMS1", "PMS2", "MSH6", "TGFBR2", "MUTYH", "CHECK2", "KRAS", "BRAF", "MYH11", "PARK2" and "RNF43". The most frequent genes have been added to these matrices using the probability of mutation. The two matrices obtained from the commercial panel after its splitting in train and test were tested with a survival analysis between these two groups and ROC curves were built to inspect the sensitivity and the positive predictive value (PPV). The matrices specific for colon instead were validated splitting each one in two initial groups, a train and validation set, moreover, the validation set was also prioritized and the result were used to check if there was over-fitting. Later, the train group was validated with ROC curves, as in the first two matrices. The results of each analysis were compared and they proved that the matrices could be used as prioritization ones. The analysis were all conducted

only on somatic variants thanks to the presence of the matched normal sample for each tumor that permitted to exclude from the analysis the germline variants.

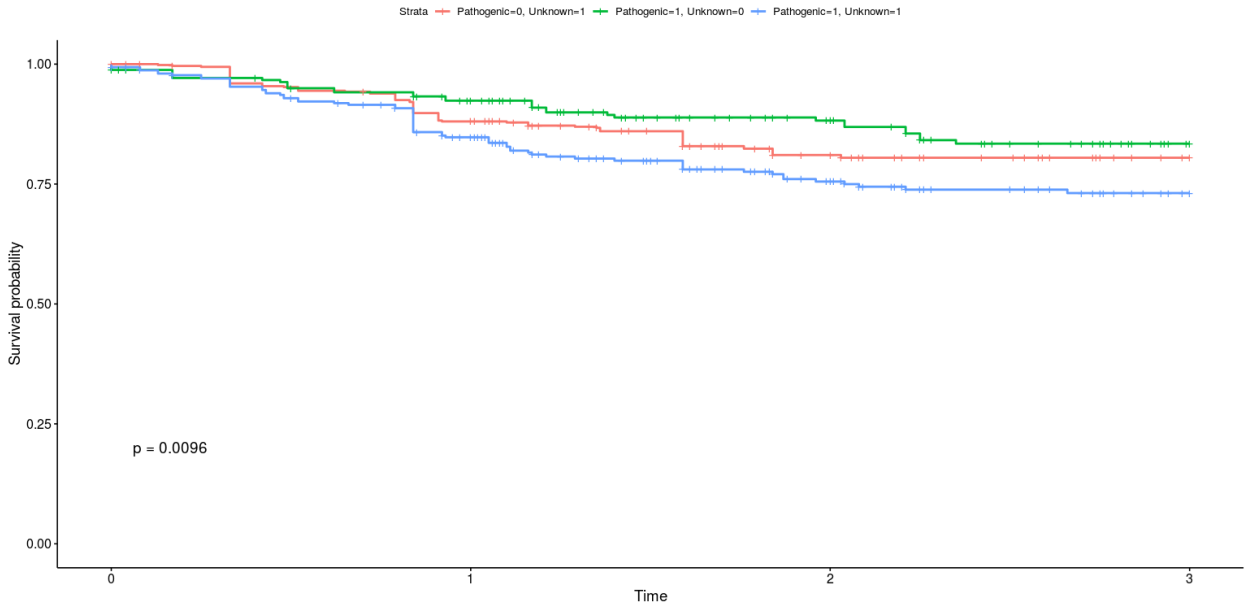

Figure 4.1: Survival curves of colon cancer patients with pathogenic and unknown mutations. It can bee seen that the patients with at least one pathogenic and one unknown mutation (blue line) have the worst survival compared with the patients with only unknown mutations (red line) and patients with only pathogenic mutations (green line) that appeared to be the ones with the best outcome.

4.1.3 Unknown Classification

Each unknown variant was compared with both pathogenic and benign variants through a survival analysis. In this way we have been able to assign to each an Hazard Ration (HR) and a p.value. The comparison was done with the survival and survminer R packages comparing the survival curves of groups of patients with and without these "unknown" mutations. Since in this way some mutations could result classified both as Pathogenic and as Benign we also assigned a pathogenicity score to each unknown variant. This score was calculated using different databases to decrease the possible False positive variant classification.

Computation of the Pathogenicity score The Pathogenicity score was computed using the databases *OregAnno*, *Refseq Function*, *GTEx*, *GWAS*, *dbscnv11* and *dbsnpf35*. We attribute 1 point to each variant that in dbsnpf database is assigned as deleterious more than 8 times, -0.5 point if the variant is assigned as deleterious less than 7 times, 0 point if a variant is assigned as tolerate 8 times. 1 point to the variant that in the database dbscnv11 has the score "ADA" and "RF" higher than 0.6. 1 point if a variant is assigned as "Non-synonymous" or "frameshift", 0.5 point if a variant is assigned as "synonymous SNV", 0 point if a variant is assigned as "Unknown". 1 point if the

variant exists in the database OregAnno, 0 if it does not exist. 1 point if the variant exists in the database GWAS, 0 if it does not exist. 1 point if the variant exist in the database GTEx, 0 if it does not exist.

4.1.4 Results

The aim of VarPrAI is to classify the "Unknown" Variants found in the NGS analysis to predict their deleteriousness in cancer. Thanks to Kaplan-Meyer survival curves and our pathogenic score we have been able to classify the unknown mutations as likely pathogenic or likely benign. In particular, we have been able to prioritize 101 variants as benign and 40 variants as pathogenic in colon cancer (Fig. 4.1) using the commercial panel and the pathogenity score. Our results indicate that for colon cancer the prioritization has a sensitivity of 1 and a PPV of ~ 0.6 with a $p.value \leq 0.05$ and a sensitivity between 0.99-0.67 and a PPV of 0.65 with a p.value between 0.06 and 0.99.

Colon Cancer Likely Pathogenic mutations	
Gene	*Variant Code or Start Position and Alternative Base*
AKT1	rs121434592
	rs774836044
ATM	rs140263969
	rs1555122944
	rs376521407
	108327726 - A
ATR	rs886058056
BAP1	rs770654975
BRCA2	rs786202373
	rs80358955
BTK	rs782338603
CDKN2A	rs971657556
CREBBP	rs140133512
MED12	rs199469669
MLH1	rs1274810165
	rs63750430
	rs776643257
MTOR	rs863225264
NF1	rs771706364
NOTCH1	136500667 - T
NOTCH2	rs747138507
NRAS	rs121913254
PALB2	rs876659475
PIK3CA	rs1553821144
PMS2	rs766811365

Table 4.1: Some of the mutations prioritizated as "Likely Pathogenic" in colon cancer

Chapter 5

Virus in Disease and in Cancer

Viruses research in vitro is not always feasible. Microorganisms cultivation is not always possible, so it now common their study using metagenomics. This research of viruses is important especially in cancer since a lot of them can cause tumors ad a cause of their integration inside the human genome. New discoveries of oncovirus could lead to better understand several tumors or to the development of new possible therapies. Among the oncovirus we can find the Human Papilloma virus that cause cervical tumors and head and neck tumors. To study these viruses in the organism it is common to use tools that analyze the DNAseq and the RNAseq of the pratients. Here we have analyzed 8 of them, comparing the performance of each of these tools on both a simulated dataset and a real dataset using the same computational resources. Between these tools two have been already inspected by Nooji et al. [186] review (ViruSeq and VirusFinder) but they were not tested with the same dataset. We inspect their sensitivity to understand which is the tool that has currently the best performance and to understand which is the most suitable for different situation.

5.1 Tools

- *VirusFinder*: VirusFinder [187] goal is to detect virus in a host sample, both integrated or un-integrated, and to detect virus integration. It works on RNAseq, WGS or target sequencing data and it accepts both raw sequencing reads (FASTQ) or alignment file (BAM). It is able also to detect novel virus thanks to the use of the viral database of BLAST. Its pipeline is constituted by three steps: (i) preprocessing, (ii) virus detection and (iii) virus integration site detection. It maps the raw sequencing reads with the host genome, keeping aside only the reads that are not mapped with the human genome itself. This reads are then used for viruses research. In the second step, in fact, the tool align the unmapped reads to a virus database, this databases can be downloaded from VirusFinder page. It is the one included in the RINS

package and it has 32,102 viruses [188]. This database can be replaced by the user with a custom one. After that, the tool de novo assembles the reads aligned to the virus DB into contigs and maps these to host genome and virus database. The virus that is ranked as the most abundant is used for the integration step. If the users know the virus and only want to discover the integration sites, they only need to skip this step and pass directly to the detection integration step. In the integration step VirusFinder combines human reference genome and virus sequence and than, employing BWA, aligns the reads recruited in the preprocessing to the new reference. From the result VirusFinder calls intercromosomal structural variants (SVs) using SVDetect [189] and CREST [190] reporting the breakpoint of the SVs that involve both Virus and Human. VirusFinder gives as output the candidates viruses identified, the contigs mapped to these viruses, the virus insertion sites detected and, if they exist, the possible novel contigs. The pipeline can be used splitting the three steps or launched with only one command that calls all the three scripts. Unfortunately, it has not be updated since 2014 and new updates in Trinity make impossible to use the second step where trinity is involved. We managed to solve the problem changing a part of the script and using the trinity version 2.8.5. In this way the tools is only feasible for advanced bioinformatic users, but not for non-expert user. The tool is implemented in Perl and it uses Bowtie2 [54], BWA, BLAST [191], BLAT [192], Samtools [193], Trinity [194], SVDetect and CREST. It is able to work with both Single-end (SE) samples and paired-end (PE) samples. The last update of the tools was in 2014.

- *VirusSeq*: VirusSeq [195] subtracts human reads and identifies virus and their potential integration sites. It uses MOSAIK [196] as alignment software. As input the user can insert both WGS or RNAseq samples. The non-human reads that are generated are aligned against a database that include all known viral sequence from Genome Information Broker for Viruses (http://gib-v.genes.nig.ac.jp/ - momentarily unavailable) and mapped reads are quantified. A cutoff set by the user is used to cut the classification that are probably wrong. The cutoff script though it is not perfectly functioning, so the tool give back only the most abundant virus found. Fortunately, it is possible to find a list of all the virus found in the sample in the log. This tool it is not able to discover novel viruses and it works on both SE and PE even if it is claimed to work only on PE. There are no information about the last update of the tool, probably 2013.

- *DAMIAN*: DAMIAN (Detection & Analysis of viral and Microbial Infectious Agents by NGS) [197] is called as a user-friendly open source software that enables clinical personnel to identify potentially pathogenic agents in clinical specimens. It has the ability to analyze also cohorts

and it is able to identify novel pathogens. It works both with DNA and RNA samples and the users can choose the host of their interest. It assembles reads into longer contigs prior to classification and annotation, because the longer sequences increase the sensitivity and the specificity of sequence similarity searches and the quality of the taxonomic assignments. It requires reads in FASTQ-format that can be both gzip compressed or not. It runs both with PE and SE. It removes low quality bases and adapters using trimmomatic. For each contig the tools determine length, circularity, GC-content and ORFs, which sequences are translated into amino acid sequences. So DAMIAN not only searches for nucleotide correspondence but also for known protein domains that can be specific for bacteria, virus or fungi in some occasion. As Database it employs the complete NCBI nt and nr database to perform the classifications. It can be run with different settings, searching matches only with nt or with nr, with both nt and nr or start searching with nt and if the match cannot be found searching with nr (independently, redundantly and iteratively). DAMIAN report is built with a summary page of all pathogens found, a page with the software information and a page for each of the pathogen. The report includes information as contig length, abundance, taxID and so on. The entries are sorted with a color code with six different categories. One category is colored in grey and it includes the pathogens that DAMIAN considers as contaminant or artifacts. This "contaminators" are specific viral sequence that are frequently detected as contaminators in DNA or RNA experiments. As output DAMIAN gives also three files for each pathogen two are fasta which contain respectively the contig sequence and the amino acids sequence and one bed file which contains the list of the orfs. DAMIAN has also another optional analysis that allows the identification of sequences from pathogens shared among groups of samples. The samples should be split in unclassified, positives and negatives. The pipeline performs pairwise BLAST alignment among the assembled contigs and create clusters that sort by a score. As output we have a spreadsheet with results and FASTA file with the contig sequence for each cluster, even for those that are not classified, so it can be easy to find novel pathogens. Bowtie2 is used as alignment software, IDBA-ud [198] is used for the assemble of sequence reads with a modification of the code to support reads up to a length of 250bp. The sequence complexity is assessed using dustmasker [199] from the NCBI Blast+ suite and contigs abundance is calculated based on the alignment of sequence reads to contigs using Bowtie2. Using HMMER and PFAM database a screening of the amino acid sequence is done. BLAST is used to identify similar sequences in NCBI's nt and nr database performing only MEGABLAST search if the user not specify anything. Though, it is also possible to do BLASTN and BLASTP analysis on all contigs or only in that contigs that not match with anything in the megablast search. NCBI's taxnames and taxnodes are used to determine the

lowest common ancestor (LCA). The software is written in Ruby and can be used on Linux. A PostgreSQL database is need to store analysis results and associated metadata. The last update of the tool was done in 2020.

- *VirTect*: VirTect [200] detects viruses in RNAseq data. It includes filters to discriminate real viral sequences from noise and artifacts minimizing false positive rates. The filters used in VirTect are three: (I) A threshold for the number of mapped reads (500), (II) A threshold for the coverage of mapped reads (5X) and (II) A threshold of the length of continuous mapped regions for any pathogen genome in their virus database (100). The inputs that should be provided are reads in FASTQ. This FASTQ sequences are mapped against human genome using TopHat2 [201]. The unaligned sequences are then aligned using BWA-MEM against the VirTect virus database that is composed by 757 viruses. It is possible to change this database with a custom one building the index for bowtie2 and bwa. At the end, the filtration are performed removing noise/artifact and poly(A) sequence that are well known to have high coverage with thousands of reads mapped to virus genomes. The main limitations of this methods are that we have not the possibility to detect new viruses and that we cannot use SE samples. To be able to use VirTect we had to add in the row 162 of the script a back slash because the tool stopped itself in the samtools depth transition. The last update of the tool was done two years ago.

- *VirDetect*: VirDetect [202] is also a tool based on subtraction that detects viruses from RNA-seq data. It aligns the reads to the human genome using STAR (Spliced Transcripts Alignment to a Reference) [203]. The reads not aligned to human genome are mapped to a database of viral genomes. The VirDetect databases has 1893 manually-curated vertebrate virus reference genome from GenBank from 16 December 2015. It is possible to use a custom database that has to be modified with a script supplied by VirDetect authors. The authors performed specific modification to the database used because RNAseq data have problems regarding low complexity reads that can lead to false positive. They have optimized the database to increase specificity masking the viral genomes for areas of human homology and areas of low complexity. The tool cannot be used to find novel viruses, but it takes both SE (with slide modification) and PE reads. The last update of tool was done 17 months ago.

- *MetaMap*: Metamap [204] is a k-mer program. It is not released as a command line tool, but is an ensemble of two tools, STAR and CLARK [205], that the user should run separately. Its goal is to classify not only virus but also bacteria and archaea coming from RNAseq data. Also in this tool the reads are aligned against the human genome and only the unmapped reads

are classified with CLARK-S [206]. The authors claim to have chosen these tools for their scalability and accuracy. The reads are classified using a set of uniquely discriminative short sequences at species level. CLARK-S output are a OTUs (Operational taxonomic units) count matrix and two csv. In the first csv the user can find all the reads matched with a pathogen, in the second each pathogen name and reads number. STAR is set to give as output not only the alignment of the genome but also the gene expression quantification. CLARK needs the generation of a large index file consisting of discriminative k-mers. This means that "it assigns a read r to a reference genome G if r and G share more discriminative k-mers than other genomes in the database". In CLARK the authors allowed mismatched between shared k-mers in a limited number of position to increase the sensitivity of the classification. Metamap can be used both with SE and PE samples. Unfortunately, it cannot be used to find novel, it identifies some reads as "Unknown" in its report, but there is no the possibility to use them to do a blast research for new viruses. The tool was update 3 years ago.

- *ViGen*: The aims of the pipeline called ViGen [207] is to detect and quantify read counts at the individual viral-gene level and to detect variants from human RNAseq. The input file necessary for this pipeline are reads in fastq format. It can be used not only for viruses, but also to detect other microbes if the information can be found in NCBI. The pipeline comprises four modules, in the first module, called "filtered human sample input", the RNAseq is aligned to human genome using RSEM tool which takes advantage of bowtie [208] or bowtie2 or STAR. Even though the authors recommend the use of bowtie, we have used bowtie2 for our analysis. The unaligned sequence are aligned with Bowtie2 and BWA against viral reference file. The unaligned sequences that result from step 1 are then re-aligned to viral reference file using Bowtie2 in Module 2 called "unfiltered human sample input". The alignment of the virus is done two times to be more comprehensive in viral detection. The reads aligned are used to obtain reads counts for each viral genome using samtools. The third module called "Viral Gene Expression Analysis" was not important for our analysis, but it calculates quantitative read counts at the individual viral-gene level. The same thing for module 4 called "Viral RNA Variant Calling Module" that is used to detect mutations in the transcripts from the viruses obtained in step1 and step2. The database used to detect the viruses can be easily replaced with another customed one creating for it the bowtie2 indexes. The database used by the authors include 745 human viruses. This tool does not permit the discovery of novel viruses since it is dependent from reference genome. It can manage both SE and PE samples. The tool was updated 3 years ago.

- *Kraken*: Kraken [209] is a k-mer tool that use memory-intensive algorithm that associate

	VirusFinder	VirSeq	VirTect	ViGen	VirDetect	DAMIAN	Metamap	Kraken2
Trimming	×	×	Cutadapt	×	×	Trimmomatic	×	×
Alignment	Bowtie2	MOSAIK	TopHat2	RSEM (Bowtie1, Bowtie2, STAR)	STAR	Bowtie2	STAR	×
Virus Alignment	Trinity	MOSAIK	Bwa mem	Bowtie2	STAR	×	CLARK	×
Virus Detection	Blast	Perl script	Samtools	Samtools/R	STAR	Megablast	CLARK	Kraken2
Language	Perl	Perl	Python	Bash/R	Bash/Java	Ruby	Bash	Perl
Website	https://bioinfo.uth.edu/VirusFinder/	https://odin.mdacc.tmc.edu/~xsu1/VirusSeq.html	https://github.com/WGLab/VirTect	https://github.com/ICBI/viGEN	https://github.com/dmarron/virdetect	https://sourceforge.net/p/damian-pd/wiki/Home/	https://github.com/theislab/MetaMap	https://github.com/DerrickWood/kraken2/wiki
Input	FASTQ/BAM	FASTQ	FASTQ	FASTQ	FASTQ	FASTQ	FASTQ	FASTQ/FASTA
Accepted data	RNAseq DNAseq	RNAseq DNAseq	RNAseq	RNAseq DNAseq	RNAseq	RNAseq DNAseq	RNAseq	RNAseq DNAseq
Sample type	SE/PE	PE	PE	SE/PE	SE/PE	SE/PE	SE/PE	SE/PE
Novel viruses discovery	YES	NO	NO	NO	NO	YES	NO	YES

Table 5.1: Tools features comparison

the k-mers with the lowest common ancestor (LCA) taxa. Unfortunately the tool has a high memory requirement. Kraken2 [210] was developed with a reduction in memory usage and to perform a quicker classification. It introduces a probabilistic and compact hash table that maps minimizers [211] to LCAs using one third of the memory of a standard hash table. The speed of Kraken2 is also achieved because it stores only minimizers with length $\ell(\ell \leq k)$, this minimizer will be the substring compared against the reference. Compared to Kraken it indexes with about 6 times of giga less. The authors claim the database of kraken2 to be about 85% smaller than the database of Kraken 1. Kraken2 does some modification on the ncbi taxonomy, it finds a minimal set of nodes that consist of all the nodes to which a reference sequence is assigned. The vertices between nodes are not modified. Then the tool assign to the nodes sequentially increasing internal taxonomy ID numbers, in this way ancestor nodes will have smaller ID numbers than their descendants. Naturally, a map of its internal taxonomy numbers with the external taxonomy numbers is saved to make the results interpretable. These internal IDs are used to facilitate the research of the LCA, since two nodes that have near number are near in the three. The Kraken2 database is built using the NCBI taxonomy [212], but it is possible for the user to create a custom database. Kraken2 takes both SE and PE samples and it is able to find novel viruses, because it gave back also the unclasiffied reads. It does not need the step of alignment and this makes it faster that the other tools that we have analysed. The tool was update 5 months ago.

The features comparison of these tools can be found in table 5.1.

5.2 Methods

To test each tool we used two different types of datasets, one simulated and one real. The simulated dataset was built with more than one virus. The real dataset samples have only one confirmed virus.

Simulated dataset creation The simulated dataset was built using the tool fluxsimulator [213]. We took thirty viruses' genome from NCBI and the human genome (version GRCh38) and we created a simulated paired-end RNAseq sample. We downloaded the genomes of these viruses from NCBI nucleotide, downloading both the FASTA and the GFF3 format. Subsequently, we used the tool AGAT [214] to transform the GFF3 files in GTF files. After running fluxsimulator to simulate RNAseq for each virus we joint each of the thirty viruses together with the simulated human RNAseq and we performed the virus research with all the tools cited above. The virus chosen for the simulated dataset are: Human Rhinovirus 3 (NC_038312), Human Rhinovirus 1 (NC_038311), Tomato mosaic virus (NC_0026921), Molluscum contagiosum virus (NC_001731), Apple mosaic virus (NC_003480), Encephalomyocarditis virus (X743121), Human papillomavirus 52 (MT815274), Hepatitis C virus (NC_004102), Human papillomavirus type 31 (U37410), Human papillomavirus type 54 (NC_001676), JC polyomavirus (NC_001699), Marine RNA virus SF-2 (NC_043518), Marine RNA virus JP-B (NC_009758), Hepatitis A virus (M14707), Human immunodeficiency virus 1 (NC_001802), Anguillid herpes virus (MW580855), Apis mellifera virus 14 isolate BFH508NG (MH973769), Human enterovirus (AB807826), Escherichia phage T7 isolate T7 (MZ318363), Human herpesvirus 6B (NC_000898), Human measles virus (NC_001498), Cyprinid herpesvirus 3 (NC_009127), Rotavirus C segment 8 (AJ549087), Japanese encephalitis virus (NC_001437), Human papillomavirus 116 (NC_013035), Influenza A virus (NC_007366), Rotavirus RCU (AF181864), Human parvovirus B19 (NC_000883), Hepatitis A virus (M14707) and Human papillomavirus 16 (NC_001526). We tried to mix human, animal and vegetable viruses to give an inclusive view of the power of these tools.

Real dataset To assure the functionality of the methods in real condition, we used 6 real datasets from which we took from 3 to 5 samples. These datasets assure the real presence of one virus verified with other methods or inserted intentionally in the cellular samples. We took for good only the annotation of the single virus and we discarded the other annotation as we do not have trustworthy information about them. Each dataset was picked from GEO and downloaded through ENA Browser Fig.5.2.

Virus	Geo project code	Samples
Human Papillomavirus 16 (HPV16)	GSE74949	SRR2932830, SRR2932844, SRR2932845, SRR2932846, SRR2932847
Hepatitis B Virus (HBV)	GSE65486	SRR1946683, SRR1946685, SRR1946686, SRR1946687
Human Rhinovirus 16 (HRV16)	GSE61141	SRR1565938, SRR1565939, SRR1565940, SRR1565941, SRR1565942
Human Alphaherpesvirus 1 (HSV1)	GSE59717	SRR1523654, SRR1523655, SRR1523656, SRR1523657, SRR1523668
Sars-CoV-2	GSE148729	SRR11550033, SRR11550033
Hepatitis C virus (HCV)	GSE84346	SRR3898707, SRR3898708, SRR3898709
Zaire Ebolavirus		SRR1553464

Table 5.2: Real dataset code: HPV16 [4], Hepatitis B Virus [5], Human Rhinovirus 16 [6], Human alphaherpesvirus 1 [7], Sars-CoV-2 [8], Hepatitis C virus [9], Ebola [10]

Database As explained in the section Tools each tool has its own database with some modification. Since viGen and VirTect were the two tools for which was easier to use a custom database and with default database with less of 800 viruses we decided to test them with two other databases. For the simulated dataset we used virusite [215] database and a database composed by the virtect database and the ncbi viral database. For real datasets we used only the latter one since the results of simulated dataset were better. We simply downloaded the ncbi viral database and we intersected the VirTect database with it. We chose not to use only the ncbi viral database since virtect database possess also PaVE [216, 217] database papillomavirus so the union of them give back better performance.

5.3 Results

Performance With 64 Gb of RAM and 64 cpu the Simulated dataset composed by two paired fastq each of 110.7 Mb took to be analyzed: 44 minutes with VirTect, 20 minutes with VirusSeq, 3 minutes with viGEN, 21 seconds with VirDetect, 42 minutes with DAMIAN, 12 minutes with Metamap, 7 minutes with VirusFinder and 6 minutes with Kraken2. For one of the HPV16 sample with two fastq of about 11 Gb for each it took 4h and 16 minutes with VirTect, 2h and 42 minutes with VirSeq, 6h and 43 minutes with Virus Finder, 3h and 7 minutes with ViGen, 1 day and 16h with DAMIAN, 25 minutes with VirDetect, 15 minutes with Metampap and 8 minutes with Kraken2. In particular DAMIAN was launched using 32 threads, VirusFinder with 8 threads, VirusDetect, viGen, Kraken2 and VirTect with 16 and VirSeq with 14. So evaluating the two different situation the fastest tool is kraken2.

Classification Only DAMIAN and VirTect have a step of trimming. In the first tool there is trimmomatic [218] included in the pipeline, in the second cutadapt [139] it has another script which needs adapter sequence to be executed. For this reason each of the real dataset samples was trimmed using Trim-Galore [138] before using the tools. We used for each tool the default options for filtering. For Kraken2 where there was no information about the filter threshold we used 10, for viGen where there was only information for the filter of copy number, we cut read counts at 50. VirTect give back two output called "continuous region" and "final continuous region" where the second one is filtered from the first one. We took the results from the non-filtered one as it appear to be more precise, indeed the "final continuous regions" file looses a lot of real existing viruses. The results obtained employing the simulated dataset and the results obtained with the real one are consistent. As it can be seen in table 5.3 we have calculated for each tool the true positive, the false positive and the false negative values having as result sensitivity and PPV. The best results for the PPV have been obtained by DAMIAN, Virtect, VirDetect and viGen with 1. However, viGen, VirTect and VirDetect have only respectively 5, 12 and 13 true positive of 30, so the best result is obtained by DAMIAN. The best 3 sensitivity can be found in Metamap, viGen with the merge database and with VirTect with the merge database with a value of 0.93. Overall the best result is achieved by DAMIAN and by viGen using the database merge. The user needs to consider that viGen is about 13 times faster than DAMIAN and requires less computer resources, but that DAMIAN is able to discover new viruses.

Concerning the real dataset in table 5.4 we can notice that for the HRV16 samples only VirusSeq and DAMIAN are able to classify the virus correctly. The other tools are not able to classify it maybe because they miss the Human Rhinovirus 16 in their databases. Instead, they have found rhinovirus that have sequence similarity with HRV16 as Human Rhinovirus 89 and Human Rhinovirus 1 which are part of the same species (Rhinovirus A), they have also found Human Rhinovirus 14 and Human Rhinovirus NAT001 that are part respectively of Rhinovirus B and Rhinovirus C species. Metamap classifies these samples as Human Rhinovirus A. This last tool, indeed, does not give always specific serotype classification, but it stop at higher classification level. In addition to HRV, it classified the Human papillomavirus 16 as alphapapillomavirus 9 namely the species name and not the serotype. This can be considered a problem since alphapapillomavirus 9 species includes 7 serotypes and Rhinovirus A even 82 different serotypes. Metamap is not specific also regarding the SARS-CoV-2 samples. For the sample 11550056 it classifies the virus as bat coronavirus instead of SARS coronavirus 2. As it can be seen in the sample SRR1946685 for viGen with its database the virus is not detected with the filter cut, but without the filter we are able to find the virus.

	VirusFinder	VirusSeq	DAMIAN	VirTect	VirTect (merge db)	VirTect (ViruSite db)	VirDetect	ViGen	ViGen (merge)	ViGen (ViruSite db)	MetaMap*	Kraken2
Human immunodeficiency virus	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Human papilloma virus 16	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Rous Sarcoma	✓	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓
Marine RNA Virus JP-B	✓	×	✓	×	✓	✓	×	×	✓	✓	✓	✓
Anguillid Herpesvirus	×	✓	×	×	✓	✓	✓	×	✓	✓	✓	✓
Marine RNA Virus SF-2	✓	×	✓	×	✓	✓	×	×	✓	✓	✓	✓
Human papillomavirus 31	×	✓	✓	✓	✓	×	×	×	✓	×	✓	×
Human papillomavirus 52	×	✓	✓	✓	✓	×	×	×	✓	×	✓	×
Human papillomavirus 54	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
JC polyomavirus	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hepatitis A virus	✓	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓
Hepatitis C virus	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Apple mosaic virus	✓	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓
Tomato mosaic virus	✓	✓	×	×	✓	✓	×	×	✓	✓	✓	✓
Molluscum contagiosum 1	×	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓
Rhinovirus 1	✓	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓
Rhinovirus 3	✓	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓
Encephalomyocarditis virus	✓	✓	✓	×	✓	✓	×	×	✓	✓	✓	✓
Rotavirus RCU	×	×	✓	×	×	×	×	×	×	×	×	×
	VirusFinder	**VirusSeq**	**DAMIAN**	**VirTect**	**VirTect (merge db)**	**VirTect (ViruSite db)**	**VirDetect**	**ViGen**	**ViGen (merge)**	**ViGen (ViruSite db)**	**MetaMap***	**Kraken2**
Human Parvovirus B19	✓	✓	✓	×	✓	✓	×	✓	✓	✓	✓	✓
Escherichia phage T7 isolate T7	✓	✓	×	×	×	✓	×	×	✓	✓	✓	✓
Enterovirus	×	×	×	×	✓	✓	×	✓	✓	✓	✓	✓
Herpes 6B	×	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓
Apis mellifera virus 14	×	×	✓	×	×	×	×	×	×	×	×	×
Influenza A virus	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Measles virus	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Cyprinid Herpesvirus 3	×	✓	×	×	×	×	✓	×	✓	✓	✓	×
Rotavirus C	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Japanese encephalitis virus	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
Human Papillomavirus 116	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓
TRUE POSITIVE	21	23	24	12	26	25	5	13	28	26	28	25
FALSE POSITIVE	1	100+	0	0	18	22	0	0	3	4	19	3
FALSE NEGATIVE	9	7	6	18	4	5	25	17	2	4	2	5
SENSITIVITY	0.7	0.77	0.8	0.4	0.86	0.83	0.16	0.43	0.93	0.86	0.93	0.83
PPV	0.95	0.19	1	1	0.59	0.53	1	1	0.90	0.86	0.59	0.89

Table 5.3: Simulated dataset results

Virus	Samples	VirusFinder	VirSeq	DAMIAN	VirTect	VirTect (merge db)	VirDetect	viGen	viGen (merge db)	Metamap*	Kraken2
HPV16	SRR2932830	✓	✓	✓	✓	✓	✓	✓	✓	✓*	✓
	SRR2932844	✓	✓	✓	✓	✓	✓	✓	✓	✓*	✓
	SRR2932845	✓	✓	✓	✓	✓	✓	✓	✓	✓*	✓
	SRR2932846	✓	✓	✓	✓	✓	✓	✓	✓	✓*	✓
	SRR2932847	✓	✓	✓	✓	✓	✓	✓	✓	✓*	✓
HRV16	SRR1565938	×	✓	✓	×	×	NA	×	×	✓*	×
	SRR1565939	×	✓	✓	×	×	NA	×	×	✓*	×
	SRR1565940	×	✓	✓	×	×	NA	×	×	✓*	×
	SRR1565941	×	✓	✓	×	×	NA	×	×	✓*	×
	SRR1565942	×	✓	✓	×	×	NA	×	×	✓*	×
HBV	SRR1946683	✓	✓	×	×	✓	✓	✓	✓	✓	×
	SRR1946684	✓	✓	✓	✓	✓	NA	×	✓	✓	×
	SRR1946685	×	✓	×	✓	✓	×	×	×	✓	×
	SRR1946686	✓	✓	✓	✓	✓	×	✓	✓	✓	×
	SRR1946687	×	✓	×	×	×	NA	×	×	✓	×

Virus	Samples	VirusFinder	VirSeq	DAMIAN	VirTect	VirTect (merge db)	VirusDetect	ViGen	ViGen (merge)	Metamap*	Kraken2
HSV1	SRR1523654	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	SRR1523655	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	SRR1523656	✓	✓	NA	✓	✓	✓	✓	✓	✓	✓
	SRR1523657	✓	✓	✓	✓	✓	NA	✓	✓	✓	✓
	SRR1523668	✓	✓	✓	✓	✓	NA	✓	✓	✓	✓
EBOLA	SRR1553464	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SARS-CoV-2	SRR11550056	✓	✓	NA	✓	✓	✓	✓	✓	✓*	✓
	SRR11550033	NA	✓	✓	✓	✓	✓	✓	✓	✓*	✓
HCV	SRR3898707	×	✓	NA	NA	NA	×	✓	✓	×	×
	SRR3898708	×	✓	NA	NA	NA	×	✓	✓	✓	×
	SRR3898709	×	✓	NA	NA	NA	×	✓	✓	✓	✓

Table 5.4: Real datasets results

Chapter 6

Discussions and Conclusions

As a matter of fact, the pipelines developed in the scientific and research context need expert users due to the peculiarity of the installation, the databases download, and the usage of specific tools. Furthermore, in most of the cases such pipelines are command line tools which result uncomfortable for life scientists and clinicians. The aim of my PhD project was to realize a technological and methodological advancement for answering questions on cancer therapies that nowadays have an important social and economic impact. The goal has been reached by developing a sensitive and efficient approach to identify oncological signature and suggest oncological therapies.

- The *OncoReport* software allows to acquire, store and analyse clinical and NGS cancer patients' data. OncoReport makes possible the clinical interpretation of NGS data through the generation of rich and comprehensive reports. Indeed, it clearly represents an effective decision support tool for oncologists. It has been developed to (i) be User-friendly, (ii) speed up the clinician work, (iii) help to prescribe a personalized therapy, (iv) be used in a clinical context without the need of technical knowledge [219].

- The pipeline *TMBcalc* has been developed for the study of TMB in patients to help the decision on the possibility to use immunotherapy. Immunotherapy is one of the most promising therapies of the last years, with a full regression for the patients where it works. The TMB study indeed permits to predict the quantity of neoantigens produced by a patients, higher its that number, higher is the possibility to respond to the therapy. The TMB standardization protocols are not mature yet, so our goal was to develop a pipeline that permitted a faster study of the biomarkers assisted by a pan-cancer panel as smaller as possible to speed up the analysis maintaining the same accuracy of WES analysis. Even if our pipeline has not an easy-to-use user interface yet, it is the first free pipeline to calculate the TMB through a docker container [220].

- The *VarPrAl* algorithm was developed to try to comprehend the role of single variants in specific diseases. This knowledge can be useful in cancer studies when the role of a variant is unknown or undefined. For instance, for patients for which does not exist a specific drug-mutation interaction the clinician can decide to study the role of each tumor variant to understand to what to focus on. Moreover, the variants listed by the system can be used as prognostic or diagnostic markers. The key point of a prioritization system is to separate genuine disease causing or disease-associated genetic variants from other variants that could be rare or not pathogenic for the disease under investigation. Unfortunately, the deleteriousness of a variant it is not sufficient to implicate a variant as playing a casual role in disease, but this data could aid clinicians in the study of possible new pathogenic variants. Seeing that, many pathogenicity-prediction algorithms exists, but no one is universally accepted as the best. So, using score predictors and survival information we developed an algorithm that prioritizes several variants from different type of tumors, particularly focusing on colon cancer. This type of prioritization, being tumor-specific, could bring to a more reliable identification of pathogenic variants.

- The study of the tools of RNAseq analysis for viruses research allowed us to understand the best characteristic of each tools to suggest which one is suitable for specific clinical analysis. The study of viruses metagenomics indeed it is still an hard path. Even if different tools exist there is no one that is able to classify perfectly all the viruses. The choice of which tool utilize it is imposed by the type of analysis that the user needs to do. If the user needs to discover new viruses and he has enough available computer resources it is recommended the use of DAMIAN. VirusFinder or Kraken2 can be a good alternative with less computer resource, even though this two tools need a more expert user for the study of novel viruses. If the user, instead, need only to research the already known virus existing in the sample viGen with the database formed by VirTect database and ncbi database is the best choice.

Bibliography

[1] Kang Yuna, Kang Chon-Sik, and Kim. Changsoo. History of nucleotide sequencing technologies: Advances in exploring nucleotide sequences from mendel to the 21st century. *Horticultural Science and Technology*, 37(5):549–558, 2019. doi: https://doi.org/10.7235/HORT.20190055.

[2] Karl Voelkerding, Shale Dames, and Jacob Durtschi. Next-generation sequencing: From basic research to diagnostics. *Clinical chemistry*, 55:641–58, 03 2009. doi: 10.1373/clinchem.2008.112789.

[3] Snyder MP. Reuter JA, Spacek DV. High-throughput sequencing technologies. *Mol Cell.*, 58 (4):586–97, 2015. doi: https://doi.org/10.1016/j.molcel.2015.05.004.

[4] Zhang Y, Koneva LA, Virani S, Arthur AE, Virani A, Hall PB, Warden CD, Carey TE, Chepeha DB, Prince ME, McHugh JB, Wolf GT, Rozek LS, Sartor MA. Subtypes of HPV-Positive Head and Neck Cancers Are Associated with HPV Characteristics, Copy Number Alterations, PIK3CA Mutation, and Pathway Signatures. *Clin Cancer Res.*, 22(18):4735–45, 2016. doi: https://doi.org/10.1158/1078-0432.CCR-16-0323.

[5] Dong H, Zhang L, Qian Z et al. Identification of HBV-MLL4 Integration and Its Molecular Basis in Chinese Hepatocellular Carcinoma. *PLoS One*, 10(4), 2015. doi: https://doi.org/10.1371/journal.pone.0123175.

[6] Bai J, Smock SL, Jackson GR Jr, MacIsaac KD, Huang Y, Mankus C, Oldach J, Roberts B, Ma YL, Klappenbach JA, Crackower MA, Alves SE, Hayden PJ. Phenotypic responses of differentiated asthmatic human airway epithelial cultures to rhinovirus. *PLoS One*, 10(2), 2015. doi: https://doi.org/10.1371/journal.pone.0118286.

[7] Rutkowski AJ, Erhard F, L'Hernault A, Bonfert T, Schilhabel M, Crump C, Rosenstiel P, Efstathiou S, Zimmer R, Friedel CC, Dölken L. Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun*, 6, 2015. doi: https://doi.org/10.1038/ncomms8126.

[8] Wyler E, Mösbauer K, Franke V, Diag A, Gottula LT, Arsiè R, Klironomos F, Koppstein D, Hönzke K, Ayoub S, Buccitelli C, Hoffmann K, Richter A, Legnini I, Ivanov A, Mari T, Del Giudice S, Papies J, Praktiknjo S, Meyer TF, Müller MA, Niemeyer D, Hocke A, Selbach M, Akalin A, Rajewsky N, Drosten C, Landthaler M. Transcriptomic profiling of SARS-CoV-2 infected human cell lines identifies HSP90 as target for COVID-19 therapy. *iScience*, 24(3), 2021. doi: https://doi.org/10.1016/j.isci.2021.102151.

[9] Boldanova T, Suslov A, Heim MH, Necsulea A. Transcriptional response to hepatitis C virus infection and interferon-alpha treatment in the human liver. *EMBO Mol Med.*, 9(6):816–834, 2017. doi: https://doi.org/10.15252/emmm.201607006.

[10] Stephen K. Gire, Augustine Goba, Kristian G. Andersen, Rachel S. G. Sealfon, Daniel J. Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, Shirlee Wohl, Lina M. Moses, Nathan L. Yozwiak, Sarah Winnicki, Christian B. Matranga, Christine M. Malboeuf, James Qu, Adrianne D. Gladden, Stephen F. Schaffner, Xiao Yang, Pan-Pan Jiang, Mahan Nekoui, Andres Colubri, Moinya Ruth Coomber, Mbalu Fonnie, Alex Moigboi, Michael Gbakie, Fatima K. Kamara, Veronica Tucker, Edwin Konuwa, Sidiki Saffa, Josephine Sellu, Abdul Azziz Jalloh, Alice Kovoma, James Koninga, Ibrahim Mustapha, Kandeh Kargbo, Momoh Foday, Mohamed Yillah, Franklyn Kanneh, Willie Robert, James L. B. Massally, Sinéad B. Chapman, James Bochicchio, Cheryl Murphy, Chad Nusbaum, Sarah Young, Bruce W. Birren, Donald S. Grant, John S. Scheiffelin, Eric S. Lander, Christian Happi, Sahr M. Gevao, Andreas Gnirke, Andrew Rambaut, Robert F. Garry, S. Humarr Khan, and Pardis C. Sabeti. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014. doi: 10.1126/science.1259657.

[11] Croce CM. Oncogenes and cancer. *N Engl J Med*, 385(5):502–11, 2008. doi: https://doi.org/10.1056/NEJMra072367.

[12] Cooper GM. *The Cell: A Molecular Approach.* Sinauer Associates, 2000. ISBN ISBN-10: 0-87893-106-6.

[13] Pantel K. Schwarzenbach H, Hoon DS. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer.*, 11(6):426–37, 2011. doi: https://doi.org/10.1038/nrc3066.

[14] Fairhurst AM. Pisetsky DS. The origin of extracellular dna during the clearance of dead and dying cells. *Autoimmunity.*, 40(4):281–4, 2007. doi: https://doi.org/10.1080/08916930701358826.

[15] Alberto Mantovani, Silvano Sozzani, Massimo Locati, Paola Allavena, and Antonio Sica. Macrophage polarization: tumor-associated macrophages as a paradigm for polarized m2 mononuclear phagocytes. *Trends in Immunology*, 23(11):549–555, 2002. ISSN 1471-4906. doi: https://doi.org/10.1016/S1471-4906(02)02302-5. URL `https://www.sciencedirect.com/science/article/pii/S1471490602023025`.

[16] M Stroun, J Lyautey, C Lederrey, A Olson-Sand, and P Anker. About the possible origin and mechanism of circulating dna: Apoptosis and active dna release. *Clinica Chimica Acta*, 313 (1):139–142, 2001. ISSN 0009-8981. doi: https://doi.org/10.1016/S0009-8981(01)00665-9. URL `https://www.sciencedirect.com/science/article/pii/S0009898101006659`. Proceedings of the Chinese Congress of Clinical Chemistry and Laboratory Medicine 2000.

[17] Loupakis F Crowley E, Di Nicolantonio F and Bardelli A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol.*, 10(8):472–84, 2013. doi: https://doi.org/10.1038/nrclinonc.2013.110.

[18] Elena Mishina. *FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Monitoring Biomarker.* Food and Drug Administration (US); Bethesda (MD), 11 2020. URL `https://www.ncbi.nlm.nih.gov/books/NBK326791/`.

[19] Robert M Califf. Biomarker definitions and their applications. *Experimental Biology and Medicine*, 243(3):213–221, 2018. doi: https://doi.org/10.1177/1535370217750088.

[20] Phillips KA. Ginsburg GS. Precision medicine: From science to value. *Health Aff (Millwood)*, 37(5):694–701, 2018. doi: https://doi.org/10.1377/hlthaff.2017.1624.

[21] Tung On Yau. Precision treatment in colorectal cancer: Now and the future. *JGH Open*, 3 (5):361–369, 2019. doi: https://doi.org/10.1002/jgh3.12153.

[22] J. Craig Venter and Mark D. Adams et al. The sequence of the human genome. *Science*, 291 (5507):1304–1351, 2001. doi: https://doi.org/10.1126/science.1058040.

[23] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409,411,412(6822,6838,6846):860–921,720,565, 2001. doi: https://doi.org/10.1038/35057062.

[24] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004. doi: https://doi.org/10.1038/nature03001.

[25] Schloss JA. How to get genomes at one ten-thousandth the cost. *Nat Biotechnol.*, 26(10): 1113–5, 2008. doi: https://doi.org/10.1038/nbt1008-1113.

[26] Fermont JM et al. Schwarze K, Buchanan J. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the united kingdom. *Genet Med.*, 22(1):85–94, 2020. doi: https://doi.org/10.1038/s41436-019-0618-7.

[27] F. Sanger, S. Nicklen, and A. R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977. doi: https://doi.org/10.1073/pnas.74.12.5463.

[28] F. Sanger, A.R. Coulson, G.F. Hong, D.F. Hill, and G.B. Petersen. Nucleotide sequence of bacteriophage λ dna. *Journal of Molecular Biology*, 162(4):729–773, 1982. doi: https://doi.org/10.1016/0022-2836(82)90546-0.

[29] Gilbert W. Maxam AM. A new method for sequencing dna. *Proc Natl Acad Sci U S A*, 74 (2):560–4, 1977. doi: https://doi.org/10.1073/pnas.74.2.560.

[30] Altman WE et al. Margulies M, Egholm M. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, 2005. doi: https://doi.org/10.1038/nature03959.

[31] Swerdlow HP et al. Bentley DR, Balasubramanian S. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, 2008. doi: https://doi.org/10.1038/nature07517.

[32] Tonthat T et al. Valouev A, Ichikawa J. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.*, 18(7):1051–63, 2008. doi: https://doi.org/10.1101/gr.076463.108.

[33] Rearick TM et al. Rothberg JM, Hinz W. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–52, 2011. doi: https://doi.org/10.1038/nature10242.

[34] Morrison HG et al. Huse SM, Huber JA. Accuracy and quality of massively parallel dna pyrosequencing. *Genome Biol.*, 8(7):R143, 2007. doi: https://doi.org/10.1186/gb-2007-8-7-r143.

[35] Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, 122(1):e59, 2018. doi: https://doi.org/10.1002/cpmb.59.

[36] McCombie WR. Goodwin S, McPherson JD. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.*, 17(6):333–51, 2016. doi: https://doi.org/10.1038/nrg.2016.49.

[37] John Eid, Adrian Fehr, and Jeremy Gray et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009. doi: https://doi.org/10.1126/science.1162986.

[38] Yue Wang, Qiuping Yang, and Zhimin Wang. The evolution of nanopore sequencing. *Frontiers in Genetics*, 5:449, 2015. doi: https://doi.org/10.3389/fgene.2014.00449.

[39] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015. ISSN 1672-0229. doi: https://doi.org/10.1016/j.gpb.2015.08.002. URL `https://www.sciencedirect.com/science/article/pii/S1672022915001345`. SI: Metagenomics of Marine Environments.

[40] Alexander S. Mikheyev and Mandy M. Y. Tin. A first look at the oxford nanopore minion sequencer. *Molecular Ecology Resources*, 14(6):1097–1102, 2014. doi: https://doi.org/10.1111/1755-0998.12324. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12324`.

[41] Miga KH et al. Jain M, Fiddes IT. Improved data analysis for the minion nanopore sequencer. *Nat Methods.*, 12(4):641–58, 2015. doi: 10.1038/nmeth.3290.

[42] Mehdi Kchouk, Jean-François Gibrat, and Mourad Elloumi. Generations of sequencing technologies: From first to next generation. *Biology and medicine*, 9:1–8, 2017.

[43] Hamady M et al. Turnbaugh PJ, Ley RE. The human microbiome project. *Nature.*, 18(449):7164, 2007. doi: https://doi.org/10.1038/nature06244.

[44] Lee J et al. Joo T, Choi J. SEQprocess: a modularized and customizable pipeline framework for NGS processing in R package. *BMC Bioinformatics*, 20(90):170–174, 2019. doi: https://doi.org/10.1186/s12859-019-2676-x.

[45] Dharanipragada P, Seelam SR, Parekh N. SeqVItA: Sequence Variant Identification and Annotation Platform for Next Generation Sequencing Data. *Front Genet*, 14(9):537, 2018. doi: https://doi.org/10.3389/fgene.2018.00537.

[46] Iacoangeli, A., Al Khleifat, A., Sproviero, W. et al. DNAscan: personal computer compatible NGS analysis, annotation and visualisation. *BMC Bioinformatics*, 20(213), 2019. doi: https://doi.org/10.1186/s12859-019-2791-8.

[47] Hakonarson H. Wang K, Li M. Annovar: Functional annotation of genetic variants from next-generation sequencing data. *"Nucleic Acids Research"*, "38"(16):e164, 2010. doi: https://doi.org/10.1093/nar/gkq603.

[48] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42:D980–D985, 2014. doi: https://doi.org/10.1093/nar/gkt1113.

[49] Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44:D862–D868, 2016. doi: https://doi.org/10.1093/nar/gkv1222.

[50] Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60000 exomes. *Nucleic Acids Research*, 45:D840–D845, 2017. doi: https://doi.org/10.1093/nar/gkw971.

[51] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311, 2001. doi: https://doi.org/10.1093/nar/29.1.308.

[52] Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat*, 32(8):894–899, 2011. doi: https://doi.org/10.1002/humu.21517.

[53] Krunic, M., Venhuizen, P., Müllauer, L., Kaserer, B., von Haeseler, A. VARIFI-Web-Based Automatic Variant Identification, Filtering and Annotation of Amplicon Sequencing Data. *"Journal of personalized medicine"*, "":9(1):10, 2019. doi: https://doi.org/10.3390/jpm9010010.

[54] Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. *"Nat Methods"*, "9": 357–359, 2012. doi: https://doi.org/10.1038/nmeth.1923.

[55] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. . *"Bioinformatics"*, "25"(14):1754–60, 2009. doi: https://doi.org/10.1093/bioinformatics/btp324.

[56] Sedlazeck, Fritz J. and Rescheneder, Philipp and von Haeseler, Arndt. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21):2790–2791, 2013. doi: https://doi.org/10.1093/bioinformatics/btt468.

[57] Binatti Andrea, Bresolin Silvia, Bortoluzzi Stefania, and Coppe Alessandro. iWhale: a computational pipeline based on Docker and SCons for detection and annotation of somatic variants in cancer WES data. *Briefings in Bioinformatics*, 22(3), 05 2020. doi: https://doi.org/10.1093/bib/bbaa065.

[58] Karczewski K.J., Francioli L.C., and G. et al. Tiao. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443, 2020. doi: https://doi.org/10.1038/s41586-020-2308-7.

[59] Mazor Y. et al. Dahary D., Golan Y. Genome analysis and knowledge-driven variant interpretation with tgex. *BMC Med Genomics*, 12(200), 2019. doi: "https://doi.org/10.1186/s12920-019-0647-8".

[60] Chaplin DD. Overview of the immune response. *J Allergy Clin Immunol.*, 125:3–23, 2010. doi: https://doi.org/10.1016/j.jaci.2009.12.980.

[61] Walport M et al. Janeway CA Jr, Travers P. *Immunobiology: The Immune System in Health and Disease. 5th edition.* Garland Science, 2001. ISBN ISBN-10: 0-8153-3642-X.

[62] Nicholson LB. The immune system. *Essays Biochem.*, 60(3):275–301, 2016. doi: https://doi.org/10.1042/EBC20160017.

[63] Pross H Kiessling R, Klein E and Wigzell H. "natural" killer cells in the mouse. ii. cytotoxic cells with specificity for mouse moloney leukemia cells. characteristics of the killer cell. *Eur J Immunol.*, 2:117–21, 1975. doi: https://doi.org/10.1002/eji.1830050209.

[64] Armitage JO. Pavletic ZS. Bone marrow transplantation for cancer - an update. *Oncologist.*, 1(3):159–168, 1996.

[65] Old LJ. Cancer immunology. *Sci Am.*, 236(5):62–70, 72–3, 76, 79, 1977. doi: https://doi.org/10.1038/scientificamerican0577-62.

[66] Ikeda H et al. Dunn GP, Bruce AT. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol.*, 3(11):991–8, 2002. doi: https://doi.org/10.1038/ni1102-991.

[67] Stutman O. Tumor development after 3-methylcholanthrene in immunologically deficient athymic-nude mice. *Science*, 183(4124):534–6, 1974. doi: https://doi.org/10.1126/science.183.4124.534.

[68] Stutman O. Chemical carcinogenesis in nude mice: comparison between nude mice from homozygous matings and heterozygous matings and effect of age and carcinogen dose. *J Natl Cancer Inst.*, 62(2):353–8, 1979.

[69] Thakar MS Abel AM, Yang C and Malarkannan S. Natural killer cells: Development, maturation, and clinical utilization. *Front Immuol*, 9:1869, 2018. doi: https://doi.org/10.3389/fimmu.2018.01869.

[70] Burnet FM. The concept of immunological surveillance. *Prog Exp Tumor Res.*, 13:1–27, 1970. doi: https://doi.org/10.1159/000386035.

[71] Schreiber RD. Dunn GP, Old LJ. The three es of cancer immunoediting. *Annu Rev Immunol.*, 22:329–60, 2004. doi: https://doi.org/10.1146/annurev.immunol.22.012703.104803.

[72] Robert D Schreiber, Lloyd J Old, and Mark J Smyth. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science (New York, N.Y.)*, 331(6024): 1565—1570, March 2011. ISSN 0036-8075. doi: https://doi.org/10.1126/science.1203486.

[73] Sanabria MH et al. Decker WK, da Silva RF. Cancer immunotherapy: Historical perspective of a clinical revolution and emerging preclinical animal models. *Front Immunol.*, 8:829, 2017. doi: https://doi.org/10.3389/fimmu.2017.00829.

[74] What is biotechnology?, immunotherapy: Timeline of key events. `https://www.whatisbiotechnology.org/index.php/timeline/science/immunotherapy`.

[75] Luciani MF et al. Brunet JF, Denizot F. A new member of the immunoglobulin superfamily–ctla-4. *Nature*, 328(6127):267–70, 1987. doi: https://doi.org/10.1038/328267a0.

[76] Allison JP. Leach DR, Krummel MF. Enhancement of antitumor immunity by ctla-4 blockade. *Science*, 271(5256):1734–6, 1996. doi: https://doi.org/10.1126/science.271.5256.1734.

[77] J. Tang, A. Shalabi, and V.M. Hubbard-Lucey. Comprehensive analysis of the clinical immuno-oncology landscape. *Annals of Oncology*, 29(1):84–91, 2018. ISSN 0923-7534. doi: https://doi.org/10.1093/annonc/mdx755. URL `https://www.sciencedirect.com/science/article/pii/S0923753419350203`. Focus on liquid bicopsy.

[78] Office of the commissioner. u.s. food and drug administration. `https://www.fda.gov/home`.

[79] Beckman A et al. (College of American Pathologists Personalized Health Care Committee.) Walk EE, Yohe SL. The cancer immunotherapy biomarker testing landscape. *Arch Pathol Lab Med.*, 144(6):706–724, 2020. doi: https://doi.org/10.5858/arpa.2018-0584-CP.

[80] Golshani G, Zhang Y. Advances in immunotherapy for colorectal cancer: a review. *Therap Adv Gastroenterol.*, 13, 2020. doi: https://doi.org/10.1177/1756284820917527.

[81] Ghanem I et al. Carretero-González A, Lora D. Analysis of response rate with anti pd1/pd-l1 monoclonal antibodies in advanced solid tumors: a meta-analysis of randomized clinical trials. *Oncotarget.*, 9(9):8706–8715, 2018. doi: https://doi.org/10.18632/oncotarget.24283.

[82] Chamoto K Iwai Y, Hamanishi J and Honjo T. Cancer immunotherapies targeting the pd-1 signaling pathway. *J Biomed Sci.*, 24(1):26, 2017. doi: https://doi.org/10.1186/s12929-017-0329-9.

[83] Gough MJ. Tormoen GW, Crittenden MR. Role of the immunosuppressive microenvironment in immunotherapy. *Adv Radiat Oncol.*, 4(3):520–526, 2018. doi: https://doi.org/10.1016/j.adro.2018.08.018.

[84] Kerstin Steinbrink, Edith Graulich, Sebastian Kubsch, Jürgen Knop, and Alexander H. Enk. Cd4+ and cd8+ anergic t cells induced by interleukin-10–treated human dendritic cells display antigen-specific suppressor activity. *Blood*, 99(7):2468–2476, 2002. ISSN 0006-4971. doi: https://doi.org/10.1182/blood.V99.7.2468. URL `https://www.sciencedirect.com/science/article/pii/S0006497120380344`.

[85] Mikhail V Blagosklonny. Immunosuppressants in cancer prevention and therapy. *OncoImmunology*, 2(12):e26961, 2013. doi: https://doi.org/10.4161/onci.26961.

[86] Jeremy J.W. Chen, Yi-Chen Lin, Pei-Li Yao, Ang Yuan, Hsang-Yu Chen, Chia-Tung Shun, Meng-Feng Tsai, Chun-Houh Chen, and Pan-Chyr Yang. Tumor-associated macrophages: The double-edged sword in cancer progression. *Journal of Clinical Oncology*, 23(5):953–964, 2005. doi: https://doi.org/10.1200/JCO.2005.12.172.

[87] Dzieciątkowski T. Dobosz P. The intriguing history of cancer immunotherapy. *Front Immunol.*, 10:2965, 2019. doi: https://doi.org/10.3389/fimmu.2019.02965.

[88] Anders RA Topalian SL, Taube JM and Pardoll DM. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat Rev Cancer.*, 16(5):275–87, 2016. doi: https://doi.org/10.1038/nrc.2016.36.

[89] Chen L. Zou W, Wolchok JD. Pd-l1 (b7-h1) and pd-1 pathway blockade for cancer therapy: Mechanisms, response biomarkers, and combinations. *Sci Transl Med.*, 8(328):328rv4, 2016. doi: https://doi.org/10.1126/scitranslmed.aad7118.

[90] Wang E. Zou J. Cancer biomarker discovery for precision medicine: New progress. *Curr Med Chem.*, 26(42):7655–7671, 2019. doi: https://doi.org/10.2174/0929867325666180718164712.

[91] Valero C, Lee M, Hoen D, et al. The association between tumor mutational burden and prognosis is dependent on treatment context. *Nature Genetics.*, 53(1):11–15, 2021. doi: https://doi.org/10.1038/s41588-020-00752-4.

[92] A. Stenzinger, J.D. Allen, J. Maas, M.D. Stewart, D.M. Merino, M.M. Wempe, and M. Dietel. Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions. *Genes Chromosomes Cancer.*, 58(8):578–588, 2019. doi: https://doi.org/10.1002/gcc.22733.

[93] Michael S et al. Lawrence. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013. doi: https://doi.org/10.1038/nature12213.

[94] D.B. Johnson, G.M. Frampton, M.J. Rioth, and et al. Targeted Next Generation Sequencing Identifies Markers of Response to PD-1 Blockade. *Cancer Immunol Res.*, 4(11):959–967, 2016. doi: https://doi.org/10.1158/2326-6066.CIR-16-0143.

[95] Luís Felipe Campesato, Romualdo Barroso-Sousa, Leandro Jimenez, Bruna R Correa, Jorge Sabbaga, Paulo M Hoff, Luiz F L Reis, Pedro Alexandre F Galante, and Anamaria A Camargo. Comprehensive cancer-gene panels can be used to estimate mutational load and predict clinical benefit to pd-1 blockade in clinical practice. *Oncotarget*, 6(33):34221–34227, 2015. ISSN 1949-2553. doi: https://doi.org/10.18632/oncotarget.5950. URL `https://www.oncotarget.com/article/5950/`.

[96] B. Meléndez, C. Van Campenhout, S. Rorive, M. Remmelink, I. Salmon, and N. D'Haene. Methods of measurement for tumor mutational burden in tumor tissue. *Transl Lung Cancer Res.*, 7(6):661–667, 2018. doi: https://doi.org/10.21037/tlcr.2018.08.02.

[97] Klempner SJ, Fabrizio D, Bane S, Reinhart M, Peoples T, Ali SM, Sokol ES, Frampton G, Schrock AB, Anhorn R, Reddy P. Tumor Mutational Burden as a Predictive Biomarker for Response to Immune Checkpoint Inhibitors: A Review of Current Evidence. *Oncologist.*, 25 (1):e147–e159, 2020. doi: https://doi.org/10.1634/theoncologist.2019-0244.

[98] U.S. Food and Drug Administration. FDA unveils a streamlined path for the authorization of tumor profiling tests alongside its latest product action. 2017. URL `https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm585347.htm`.

[99] U.S. Food and Drug Administration. FDA announces approval, CMS proposes coverage of first breakthrough-designated test to detect extensive number of cancer biomarkers. 2017. URL `https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm587273.htm`.

[100] Thermofisher. thermofisher oncomine™ tumor mutation load assay user guide. `https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0017042_TumorMutLoad_UG.pdf`.

[101] Trusight, tumor 170. `https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/trusight-tumor-170-data-sheet-1170-2016-017.pdf`.

[102] I. Buchhalter, E. Rempel, V. Endris, et al. Size matters: dissecting key parameters for panel-based tumor mutational burden (TMB) analysis. *Int J Cancer.*, 144:848–858, 2018. doi: https://doi.org/10.1002/ijc.31878.

[103] M. Allgäuer, J. Budczies, P. Christopoulos, et al. Implementing tumor mutational burden (TMB) analysis in routine diagnostics-a primer for molecular pathologists and clinicians. *Transl Lung Cancer Res.*, 7(6):703–715, 2018. doi: https://doi.org/10.21037/tlcr.2018.08.14.

[104] V. Endris, I. Buchhalter, M. Allgauer, et al. Measurement of tumor mutational burden (TMB) in routine molecular diagnostics: in-silico and reallife analysis of three larger gene panels. *Int J Cancer.*, 144(9):2303–2312, 2018. doi: https://doi.org/10.1002/ijc.32002.

[105] H.X. Wu, Z.X. Wang, Q. Zhao, F. Wang, and R.H. Xu. Designing gene panels for tumor mutational burden estimation: the need to shift from 'correlation' to 'accuracy'. *J Immunother Cancer.*, 7(1):206, 2019. doi: https://doi.org/10.1186/s40425-019-0681-2.

[106] Ping Gao, Miao He, Chunling Zhang, and Changhui Geng. Integrated analysis of gene expression signatures associated with colon cancer from three datasets. *Gene*, 654:95 – 102, 2018. ISSN 0378-1119. doi: https://doi.org/10.1016/j.gene.2018.02.007. URL `http://www.sciencedirect.com/science/article/pii/S0378111918301379`.

[107] F. Wang, C. Tang, X. Gao, and J. Xu. Identification of a six-gene signature associated with tumor mutation burden for predicting prognosis in patients with invasive breast carcinoma. *Ann Transl Med.*, 8(7), 2020. doi: https://doi.org/10.21037/atm.2020.04.02.

[108] Qi F et al. Zhang C, Shen L. Multi-omics analysis of tumor mutation burden combined with immune infiltrates in bladder urothelial carcinoma. *J Cell Physiol.*, 235(4):3849–3863, 2020. doi: https://doi.org/10.1002/jcp.29279.

[109] Hane Lee, Joshua L. Deignan, Naghmeh Dorrani, Samuel P. Strom, Sibel Kantarci, Fabiola Quintero-Rivera, Kingshuk Das, Traci Toy, Bret Harry, Michael Yourshaw, Michelle Fox, Brent L. Fogel, Julian A. Martinez-Agosto, Derek A. Wong, Vivian Y. Chang, Perry B. Shieh, Christina G. S. Palmer, Katrina M. Dipple, Wayne W. Grody, Eric Vilain, and Stanley F.

Nelson. Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA*, 312(18):1880–1887, 11 2014. doi: https://doi.org/10.1001/jama.2014.14604.

[110] Wang K. Yang H. Genomic variant annotation and prioritization with annovar and wannovar. *"Nat Protoc"*, "10":1556–1566, 2015. doi: doi:10.1038/nprot.2015.105.

[111] Sobreira N, Schiettecatte F, Boehm C, Valle D, Hamosh A. New tools for Mendelian disease gene identification: PhenoDB variant analysis module; and GeneMatcher, a web-based tool for linking investigators with an interest in the same gene. *Hum Mutat.*, 36(4):425–31, 2015. doi: https://doi.org/10.1002/humu.22769.

[112] MacArthur DG, Manolio TA, Dimmock DP et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–76, 2014. doi: https://doi.org/10.1038/nature13127.

[113] Schuelke M. Seelow D, Schwarz JM. Genedistiller–distilling candidate genes from linkage intervals. *PLoS One.*, 3:e3874, 2008. doi: https://doi.org/10.1371/journal.pone.0003874.

[114] Daniela Hombach, Markus Schuelke, Ellen Knierim, Nadja Ehmke, Jana Marie Schwarz, Björn Fischer-Zirnsak, and Dominik Seelow. MutationDistiller: user-driven identification of pathogenic DNA variants. *Nucleic Acids Research*, 47(W1):W114–W120, 05 2019. doi: 10.1093/nar/gkz330. URL `https://doi.org/10.1093/nar/gkz330`.

[115] Matteo Chiara, Pietro Mandreoli, Marco Antonio Tangaro, Anna Maria D'Erchia, Sandro Sorrentino, Cinzia Forleo, David S. Horner, Federico Zambelli, and Graziano Pesole. Vinyl: Variant prioritization by survival analysis. *bioRxiv*, 2020. doi: 10.1101/2020.01.23.917229. URL `https://www.biorxiv.org/content/early/2020/01/24/2020.01.23.917229`.

[116] Kwan JS et al. Li MX, Gui HS. A comprehensive framework for prioritizing variants in exome sequencing studies of mendelian diseases. *Nucleic Acids Res.*, 40:e53, 2012. doi: https://doi.org/10.1093/nar/gkr1257.

[117] Eddie Ip, Gavin Chapman, David Winlaw, Sally L. Dunwoodie, and Eleni Giannoulatou. Vpot: A customizable variant prioritization ordering tool for annotated variants. *Genomics, Proteomics & Bioinformatics*, 17(5):540–545, 2019. ISSN 1672-0229. doi: https://doi.org/10.1016/j.gpb.2019.11.001. URL `https://www.sciencedirect.com/science/article/pii/S1672022919301494`. Bioinformatics Commons.

[118] Georgitsi M. et al. Alexander J., Mantzaris D. Variant ranker: a web-tool to rank genomic

data according to functional significance. *BMC Bioinformatics*, 18, 2017. doi: https://doi.org/10.1186/s12859-017-1752-3.

[119] Pauline C. Ng and Steven Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 07 2003. ISSN 0305-1048. doi: 10.1093/nar/gkg509. URL `https://doi.org/10.1093/nar/gkg509`.

[120] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 10 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1016. URL `https://doi.org/10.1093/nar/gky1016`.

[121] Peshkin L et al. Adzhubei IA, Schmidt S. A method and server for predicting damaging missense mutations. *Nat Methods.*, 7(4):248–249, 2010. doi: https://doi.org/10.1038/nmeth0410-248.

[122] Dillon MR et al. Bolyen E, Rideout JR. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.*, 37(8):852–857, 2019. doi: https://doi.org/10.1038/s41587-019-0209-9.

[123] Pfeiffer JK. Robinson CM. Viruses and the Microbiota. *Annu Rev Virol.*, 1:55–69, 2014. doi: https://doi.org/10.1146/annurev-virology-031413-085550.

[124] Raoult D. There is no such thing as a tree of life (and of course viruses are out!). *Nat Rev Microbiol.*, 7(8), 2009. doi: https://doi.org/10.1038/nrmicro2108-c6.

[125] Brüssow H. The not so universal tree of life or the place of viruses in the living world. *Philos Trans R Soc Lond B Biol Sci.*, 364(1527):2263–74, 2009. doi: https://doi.org/10.1098/rstb.2009.0036.

[126] Zipursky S.L. et al. Lodish H., Berk A. *Viruses: Structure, function, and uses.* In Molecular Cell Biology, 4th ed. New York: W. H. Freeman, 2000. ISBN 0-7167-3136-3.

[127] Drosten C et al. Marz M, Beerenwinkel N. Challenges in RNA virus bioinformatics. *Bioinformatics.*, 30(13):1793–9, 2014. doi: https://doi.org/10.1093/bioinformatics/btu105.

[128] Murray KA et al. Anthony SJ, Epstein JH. A strategy to estimate unknown viral diversity in mammals. *mBio.*, 4(5), 2013. doi: https://doi.org/10.1128/mBio.00598-13.

[129] International committee on taxonomy of viruses. URL `https://talk.ictvonline.org/taxonomy/p/taxonomy_releases`.

[130] Fiers, W., Contreras, R., Duerinck, F. et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, "260":500–507, 1976. doi: https://doi.org/10.1038/260500a0.

[131] Rebecca Rose and Constantinides et al. Challenges in the analysis of viral metagenomes. *Virus Evolution*, 2(2), 08 2016. doi: https://doi.org/10.1093/ve/vew022.

[132] Carr JK et al. Mokili JL, Rogers M. Identification of a novel clade of human immunodeficiency virus type 1 in Democratic Republic of Congo. *AIDS Res Hum Retroviruses.*, 18(11):817–23, 2002. doi: https://doi.org/10.1089/08892220260139567.

[133] Bikandou B et al. Takemura T, Ekwalanga M. A novel simian immunodeficiency virus from black mangabey (Lophocebus aterrimus) in the Democratic Republic of Congo. *J Gen Virol.*, 86:1967–1971, 2005. doi: https://doi.org/10.1099/vir.0.80697-0.

[134] MacDonald ML, Polson SW, Lee KH. k-mer-Based Metagenomics Tools Provide a Fast and Sensitive Approach for the Detection of Viral Contaminants in Biopharmaceutical and Vaccine Manufacturing Applications Using Next-Generation Sequencing. *mSphere.*, "6"(2), 2021. doi: https://doi.org/10.1128/mSphere.01336-20.

[135] Dutilh BE. Mokili JL, Rohwer F. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol.*, 2(1):63–77, 2012. doi: https://doi.org/10.1016/j.coviro.2011.12.004.

[136] Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. *Virus Res*, "239": 136–142, 2017. doi: https://doi.org/10.1016/j.virusres.2017.02.002.

[137] Hollister EB. Santiago-Rodriguez TM. Human Virome and Disease: High-Throughput Sequencing for Virus Discovery, Identification of Phage-Bacteria Dysbiosis and Development of Therapeutic Approaches with Emphasis on the Human Gut. *Viruses*, 11(7), 2019. doi: https://doi.org/10.3390/v11070656.

[138] Felix Krueger. Trim galore: a wrapper tool around cutadapt and fastqc to consistently apply quality and adapter trimming to fastq files, 2012. URL `https://github.com/FelixKrueger/TrimGalore`.

[139] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 2011. doi: https://doi.org/10.14806/ej.17.1.200.

[140] Picard toolkit, 2018. URL `http://broadinstitute.github.io/picard/`.

[141] Banks E et al. McKenna A, Hanna M. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res.*, 20(9):1297–1303, 2010. doi: https://doi.org/10.1101/gr.107524.110.

[142] Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576, 2012. doi: https://doi.org/10.1101/gr.129684.111.

[143] European medicines agency. `https://www.ema.europa.eu/en`.

[144] Italian medicines agency. `https://www.aifa.gov.it/en/web/guest/home`.

[145] Guo AC et al. Wishart DS, Knox C. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 1(34):D668–72, 2006. doi: https://doi.org/10.1093/nar/gkj067.

[146] European society for medical oncology. `https://www.esmo.org/guidelines`.

[147] M. Whirl-Carrillo, E.M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C.F. Thorn, R.B. Altman and T.E. Klein. Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology*, 92(4):414–417, 2012. doi: http://dx.doi.org/10.1038/clpt.2012.96.

[148] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2019.

[149] Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, et al. Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics*, 49(2):170–174, 2017.

[150] Tamborero, D., Rubio-Perez, C., Deu-Pons, J. et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(25), 2018. doi: https://doi.org/10.1101/140475.

[151] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[152] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 11 2017. doi: https://doi.org/10.1093/nar/gkx1037.

[153] dbgap/database of genotypes and phenotypes/ national center for biotechnology information, national library of medicine (ncbi/nlm). `https://www.ncbi.nlm.nih.gov/gap`. Accessed: 2021-13-10.

[154] A. 1000 Genomes Project Consortium, Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, and G.R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi: https://doi.org/10.1038/nature15393.

[155] John G Tate and Bamford et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2018. doi: 10.1093/nar/gky1015. URL `https://doi.org/10.1093/nar/gky1015`.

[156] Nhlbi go exome sequencing project (esp). `http://evs.gs.washington.edu/EVS/`.

[157] Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, Schrock A, Campbell B, Shlien A, Chmielecki J, Huang F, He Y, Sun J, Tabori U, Kennedy M, Lieber DS, Roels S, White J, Otto GA, Ross JS, Garraway L, Miller VA, Stephens PJ, Frampton GM. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.*, 9(34), 2017. doi: https://doi.org/10.1186/s13073-017-0424-2.

[158] Sarah Sammons, Andrew Elliott, Jeremy Meyer Force, Nicholas C. DeVito, Paul Kelly Marcom, Sandra M. Swain, Antoinette R. Tan, Evanthia T. Roussos Torres, Jia Zeng, Mustafa Khasraw, Justin M. Balko, Wolfgang Michael Korn, and Carey K. Anders. Genomic evaluation of tumor mutational burden-high (tmb-h) versus tmb-low (tmb-l) metastatic breast cancer to reveal unique mutational features. *Journal of Clinical Oncology*, 39(15_suppl):1091–1091, 2021. doi: https://doi.org/10.1200/JCO.2021.39.15_suppl.1091.

[159] Tess A O'Meara and Sara M Tolaney. Tumor mutational burden as a predictor of immunotherapy response in breast cancer. *Oncotarget*, 12(5):394–400, 2021. ISSN 1949-2553. doi: https://doi.org/10.18632/oncotarget.27877. URL `https://www.oncotarget.com/article/27877/`.

[160] Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *"Nucleic Acids Res"*, "40":D130–D135, 2012. doi: https://doi.org/10.1093/nar/gkr1079.

[161] Kai Kang, Fucun Xie, Jinzhu Mao, Yi Bai, and Xiang Wang. Significance of tumor mutation burden in immune infiltration and prognosis in cutaneous melanoma. *Frontiers in Oncology*, 10:1801, 2020. doi: 10.3389/fonc.2020.573141. URL `https://www.frontiersin.org/article/10.3389/fonc.2020.573141`.

[162] S. Alaimo, G.P. Marceca, A. Ferro, and A. Pulvirenti. Detecting disease specific pathway substructures through an integrated systems biology approach. *Noncoding RNA.*, 3(2), 2017. doi: https://doi.org/10.3390/ncrna3020020.

[163] Hughes D.S. et al. Fan Y., Xi L. Muse: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol*, 17(178), 2016. doi: https://doi.org/10.1186/s13059-016-1029-6.

[164] Larson, David E et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–7, 2012. doi: https://doi.org/10.1093/bioinformatics/btr665.

[165] Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet.*, 100(2):267–280, 2017. doi: https://doi:10.1016/j.ajhg.2017.01.004.

[166] Mou C. et al. Liu X., Li C. dbnsfp v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site snvs. *Genome Med*, 12(103):235–241, 2020. doi: https://doi.org/10.1186/s13073-020-00803-9.

[167] Wang G et al. Lesurf R, Cotto KC. Oreganno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic acids research*, 44(D1):D126–D132, 2016. doi: https://doi.org/10.1093/nar/gkv1203.

[168] Daniel R Zerbino, Steven P Wilder, Nathan Johnson, Thomas Juettemann, and Paul R Flicek. The ensembl regulatory build. *Genome biology*, 16:56, March 2015. ISSN 1474-7596. doi: 10.1186/s13059-015-0621-5. URL `https://europepmc.org/articles/PMC4407537`.

[169] Cerezo M et al. Buniello A, MacArthur JAL. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, 47(D1):D1005–D1012, 2019. doi: https://doi.org/10.1093/nar/gky1120.

[170] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6): 580–585, 2013. doi: https://doi.org/10.1038/ng.2653.

[171] Fay JC. Chun S. Identification of deleterious mutations within three human genomes. *Genome Res*, 19(9):1553–61, 2009. doi: https://doi.org/10.1101/gr.092619.109.

[172] Schuelke M Schwarz JM, Cooper DN and Seelow D. Mutationtaster2: mutation prediction for the deep-sequencing age. *Nat Methods.*, 11(4):361–362, 2014. doi: https://doi.org/10.1038/nmeth.2890.

[173] Sander C. Reva B, Antipin Y. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, 39(17):361–362, 2011. doi: https://doi.org/10.1093/nar/gkr407.

[174] Cooper DN et al. Shihab HA, Gough J. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum Mutat.*, 34:57–65, 2013. doi: https://doi.org/10.1002/humu.22225.

[175] Murphy S et al. Choi Y, Sims GE. Predicting the functional effect of amino acid substitutions and indels. *PLoS One.*, 7(10), 2012. doi: https://doi.org/10.1371/journal.pone.0046688.

[176] Jian X. et al. Dong C., Wei P. Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Human molecular genetics*, 24(8):2125–2137, 2015. doi: https://doi.org/10.1093/hmg/ddu733.

[177] Berger MJ et al. Jagadeesh KA, Wenger AM. M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet.*, 48(12):1581–1586, 2016. doi: https://doi.org/10.1038/ng.3703.

[178] Padigepati SR et al. Sundaram L, Gao H. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet.*, 50(8):1161–1170, 2018. doi: https://doi.org/10.1038/s41588-018-0167-z.

[179] Ferté J et al. Raimondi D, Tanyalcin I. Deogen2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.*, 45(W1): W201–206, 2017. doi: https://doi.org/10.1093/nar/gkx390.

[180] Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. *Hum Mutat.*, 38(3): 243–251, 2017. doi: https://doi.org/10.1002/humu.23158.

[181] Nawar Malhis, Matthew Jacobson, Steven J M Jones, and Jörg Gsponer. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. *Nucleic Acids Research*, 48 (W1):W154–W161, 04 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa288. URL `https://doi.org/10.1093/nar/gkaa288`.

[182] Hashem A. Shihab, Mark F. Rogers, Julian Gough, Matthew Mort, David N. Cooper, Ian N. M. Day, Tom R. Gaunt, and Colin Campbell. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10): 1536–1543, 02 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv009. URL `https://doi.org/10.1093/bioinformatics/btv009`.

[183] Mark F Rogers, Hashem A Shihab, Matthew Mort, David N Cooper, Tom R Gaunt, and Colin Campbell. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, 34(3):511–513, 09 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx536. URL `https://doi.org/10.1093/bioinformatics/btx536`.

[184] Liu X. Jian X, Boerwinkle E. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.*, 42(22):13534–13544, 2014. doi: https://doi.org/10.1093/nar/gku1206.

[185] MD) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore. Online mendelian inheritance in man, omim®. `https://omim.org/`. Accessed: 2021-06-09.

[186] Vennema H et al. Nooij S, Schmitz D. Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Front Microbiol.*, 23(9), 08 2018. doi: https://doi.org/10.3389/fmicb.2018.00749.

[187] Zhao Z. Wang Q, Jia P. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One.*, 8(5), 2013. doi: https://doi.org/10.1371/journal.pone.0064465.

[188] Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics*, "28"(8):1174–5, 2012. doi: https://doi.org/10.1093/bioinformatics/bts100.

[189] Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, Barillot E. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, "26"(15):1895–6, 2012. doi: https://doi.org/10.1093/bioinformatics/btq293.

[190] Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, Holmfeldt L, Payne-Turner D, Fan X, Wei L, Zhao D, Obenauer JC, Naeve C, Mardis ER, Wilson RK, Downing JR, Zhang J. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods.*, 8(8):652–4, 2011. doi: https://doi.org/10.1038/nmeth.1628.

[191] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.ncer genomes with base-pair resolution. Basic local alignment search tool. *J Mol Biol.*, 215(3):652–4, 1990. doi: https://doi.org/10.1016/S0022-2836(05)80360-2.

[192] Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res*, 12:656–664, 2002.

[193] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 02 2021. ISSN 2047-217X. doi: https://doi.org/10.1093/gigascience/giab008.

[194] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol*, 29(7):644–652, 2011. doi: https://doi.org/10.1038/nbt.1883.

[195] Thompson EJ et al. Chen Y, Yao H. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics*, 29(2): 266–7, 2013. doi: https://doi.org/10.1093/bioinformatics/bts665.

[196] Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *"PLoS ONE"*, "9"(3), 2014. doi: https://doi.org/10.1371/journal.pone.0090581.

[197] Indenbirken D et al. Alawi M, Burkhardt L. DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. *Sci Rep.*, 9(1):16841, 2019. doi: https://doi.org/10.1038/s41598-019-52881-4.

[198] Yu Peng, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, 04 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts174. URL `https://doi.org/10.1093/bioinformatics/bts174`.

[199] Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol.*, 13(5):1028–40, 2006. doi: https://doi.org/10.1089/cmb.2006.13.1028.

[200] Deng M Xia Y, Liu Y and Xi R. Detecting virus integration sites based on multiple related sequencing data by VirTect. *BMC Med Genomics*, 19(9), 2019. doi: https://10.1186/s12920-018-0461-8.

[201] Trapnell C et al. Kim D, Pertea G. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4), 2013. doi: https://doi.org/10.1186/gb-2013-14-4-r36.

[202] Hollern D. et al. Selitsky S.R., Marron D. Virus expression detection reveals RNA-sequencing contamination in TCGA. *BMC Genomics*, 21(79), 2020. doi: https://doi.org/10.1186/s12864-020-6483-6.

[203] Schlesinger F et al. Dobin A, Davis CA. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.*, 29(1):15–21, 2013. doi: https://10.1093/bioinformatics/bts635.

[204] Westermann AJ et al. Simon LM, Karg S. MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. *Gigascience*, 7(6), 2018. doi: https://doi.org/10.1093/gigascience/giy070.

[205] Close TJ Ounit R, Wanamaker S and Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics.*, 16(1):236, 2015. doi: https://10.1186/s12864-015-1419-2.

[206] Lonardi S. Ounit R. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics.*, 32(24), 2016. doi: https://10.1093/bioinformatics/btw542.

[207] Madhavan S Bhuvaneshwar K, Song L and Gusev Y. viGEN: An Open Source Pipeline for the Detection and Quantification of Viral RNA in Human Tumors. *Front Microbiol*, 5(9), 2018. doi: https://10.3389/fmicb.2018.01172.

[208] Pop M Langmead B, Trapnell C and Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3), 2009. doi: https://10.1186/gb-2009-10-3-r25.

[209] Salzberg SL. Wood DE. Kraken: ultrafast metagenomic sequence classification using exact alignments. . *Genome Biol*, 15(3), 2014. doi: https://10.1186/gb-2014-15-3-r46.

[210] Langmead B. Wood DE, Lu J. Improved metagenomic analysis with Kraken 2. *Genome Biol*, 20(1), 2019. doi: https://10.1186/s13059-019-1891-0.

[211] Hunt BR et al. Roberts M, Hayes W. Reducing storage requirements for biological sequence comparison. *Bioinformatics.*, 20(18):3363–9, 2004. doi: 10.1093/bioinformatics/bth408.

[212] Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res.*, 40:D136–43, 2012. doi: https://doi.org/10.1093/nar/gkr1178.

[213] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–10083, 09 2012. ISSN 0305-1048. doi: 10.1093/nar/gks666. URL `https://doi.org/10.1093/nar/gks666`.

[214] Dainat J. Agat: Another gff analysis toolkit to handle annotations in any gtf/gff format. `https://www.doi.org/10.5281/zenodo.3552717`. Version v0.4.0.

[215] Matej Stano, Gabor Beke, and Lubos Klucar. viruSITE—integrated database for viral genomics. *Database*, 2016, 12 2016. doi: https://doi.org/10.1093/database/baw162.

[216] The papillomavirus episteme (pave). URL `pave.niaid.nih.gov`.

[217] Xirasagar S et al. Van Doorslaer K, Li Z. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res.*, 45(D1):D499–D506, 2016. doi: https://doi.org/10.1093/nar/gkw879.

[218] Usadel B. Bolger AM, Lohse M. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics.*, 30(15):2114–20, 2014. doi: https://doi.org/10.1093/bioinformatics/btu170.

[219] Privitera, GF. and Alaimo, S. and Micale, G. and Mare, M., and Martorana, E. and Villa, R., Ferro, A. and Forte, S. and Pulvirenti, A. . OncoReport: a system for integrative NGS analysis in precision medicine. *Submitted 2021*, 2021.

[220] Privitera, GF. and Caruso, A. and Alaimo, S. and Ferro, F and Forte, S. and Pulvirenti, A. . A small and reliable pan-cancer TMB gene signature. *Submitted 2021*, 2021.

Appendix A

Appendix

All these works have been conducted in partnership with the IOM ricerca of Viagrande and with the International Agency for Research on Cancer (IARC) in Lyon with the help respectively of Dott. Stefano Forte and Dott. Massimo Tommasino.