U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º
CICLO

FCUP
ANO

U.PORTO

New RNA signatures of therapy evasion in cancer

Ana Filipa Pacheco Fonseca Lopes de Mendonça

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

New RNA signatures of therapy evasion in cancer

Ana Filipa Pacheco Fonseca Lopes de Mendonça

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional

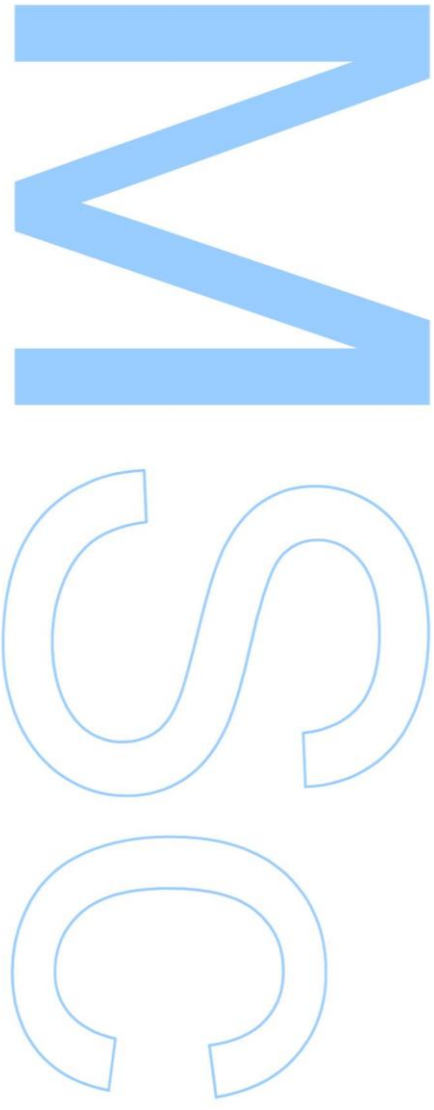
2022

New RNA signatures of therapy evasion in cancer

Ana Filipa Pacheco Fonseca Lopes de Mendonça
Mestrado de Bioinformática e Biologia Computacional
Departamento de Ciências dos Computadores
2022

Supervisor
Doutora Alexandra Moreira, Investigadora Principal, ICBAS, i3s, UP

Co-supervisors
Doutora Isabel Pereira-Castro, Investigadora Júnior, i3s, UP
Professor Pedro Ferreira, Professor Assistente, FCUP, i3S, INESC TEC



U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Sworn Statement

I, Ana Filipa Pacheco Fonseca Lopes de Mendonça, enrolled in the Master degree in Bioinformatics and Computational Biology at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offence.

Filipa Lopes de Mendonça

Author Statement

The work included in this thesis was executed by the candidate with close revisions by the supervising team and was supported by Fundação para a Ciência e a Tecnologia (FCT) under the project EXPL/SAU-PUB/1073/2021, by the “Cancer Research on Therapy Resistance: From Basic Mechanisms to Novel Targets”—NORTE-01–0145-FEDER-000051 project, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 952334 , byPrograma Operacional Regional do Norte and co-funded by
European Regional Development Fund under the project “The Porto Comprehensive Cancer Center” with the reference NORTE-01–0145-FEDER-072678— Consórcio P.CCC—Porto Comprehensive Cancer Center

Leveraged by the development of this work, one article is in preparation:

Pro-inflammatory polarization induces 3’UTR shortening and intronic polyadenylation in primary human macrophages

Joana Wilton[1-3], Filipa Lopes de Mendonça[2,3] Michael Tellier[4], Takayuki Nojima[4], Isabel Pereira-Castro[2,3], Angela M Costa[5,6], Jaime Freitas[2], Shona Murphy[4], Maria Jose Oliveira[5,6,7], Nicholas J Proudfoot[4], Alexandra Moreira[2,3,8]

[1]Graduate Program in Areas of Basic and Applied Biology (GABBA) PhD Program, ICBAS-Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Portugal; [2]Gene Regulation - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal, [3]IBMC-Instituto de Biologia Molecular e Celular, Portugal; [4]Sir William Dunn School of Pathology, University of Oxford, United Kingdom; [5]Tumor and Microenvironment Interactions Group – Instituto de Investigacao e Inovacao em Saude, Universidade do Porto, Portugal; [6]INEB-Instituto Nacional de Engenharia Biomédica, Portugal; [7]Faculdade de Medicina, Universidade do Porto, Portugal,; [8]ICBAS-Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Portugal

.

Filipa Lopes de Mendonça

Abstract

The blueprint of a cell is its transcriptional network. In eukaryotic systems and for protein coding genes, the 3'end mRNA-processing mechanism involves the pre-mRNA cleavage at the 3'end and the addiction of a poly(A)denosine (poly(A)) tail. The use of different Polyadenylation signals (PAS) results in several mRNA transcripts with different 3'ends. This co-transcriptional mechanism, known as alternative Polyadenylation (APA), provides a wide variability of mRNA isoforms from a single gene. When APA occurs in the three prime untranslated regions (3'UTR-APA), several transcripts with alternative 3'UTRs are produced that may control the expression of the mRNA.

Additionally, APA can occur within introns, by recognition of Intronic Polyadenylation (IPA) signals. The generated transcripts lose a portion of the coding regions, resulting in different proteins. Several studies have shown a correlation between APA events and cancer. Short 3'UTRs correlate with proliferative and cancer cells, activating proto-oncogenes, and truncated mRNAs caused by IPA may affect tumour-suppressive functions. Hence, the relationship between APA and cancer is physiologically relevant despite being still deciphered.

Our primary goal was to perform a bioinformatic analysis to find specific APA and IPA mRNA signatures in cancers that are triggered by tobacco smoking, to comprehend how cancer cells affect APA and IPA in immune cells and to reveal new diagnostic/therapeutic targets.

First, we investigated how colorectal cancer (CRC) cells affect APA and IPA in macrophages. We performed a sequence alignment data analysis obtained from 3'RNA-Seq data of primary human macrophages. We analysed data from unpolarised macrophages (M0) and pro-inflammatory macrophages (M1), isolated from 3 healthy donors and in macrophages co-cultured with two CRC cell lines (HCT15 and RKO). To analyse 3'UTR-APA and IPA events, we employed the APAlyzer algorithm, which uses the Poly(A)_DB database, a PAS collection, to examine APA events in coding regions and untranslated regions and introns. Our results show more 3'UTR-APA shortening events and more IPA occurring upon macrophage polarisation (M1) that is maintained when M1 cells are co-cultured with CRC cell lines. Next, to comprehend how the immune system affects 3'UTR-APA in CRC, 3'RNA-Seq of eight CRC samples were performed in mice, four from immunocompetent (IC) mice and the other four from immunodeficient (ID) mice. We developed a workflow with state-of-the-art tools to

analyse 3' RNA-Seq raw data and compare differential gene expression, 3'UTR-APA and IPA events in CRC from immunocompetent and immunodeficient mice. Immunodeficient mice also present more 3'UTR-APA shortening and more IPA events when compared to immunocompetent mice.

Next, we aimed to find common 3'UTR-APA events and APA signatures in lung, bladder, and head and neck cancers, which are triggered by tobacco smoking. No head and neck data were publicly available. Hence, we collected RNA-Seq data processed by the DaPars2 algorithm from GSE174330 (Bladder data), PRJNA356345 (Lung data), APAatlas (Processed RNA-Seq database from Genotype-Tissue Expression - Bladder and Lung data from normal tissues) and TC3A (Processed RNA-Seq database from the Cancer Genome Atlas - Bladder and Lung data from tumour tissues). No common genes were found.

To compare normal and tumour tissues, we paired the samples from TC3A and APAatlas and intersected all datasets. No common gene was found. As TCGA data require great computing power and time to analyse, we are currently working on APAtizer, an automated workflow built with Snakemake workflow manager that analyses RNA-Seq data and analyses APA events using Dapars2 or APAlyzer algorithms. This new tool was tested with three types of inputs: FASTQ single-end data, FASTQ paired-end data and BAM files.

Finally, to our knowledge, our results describe for the first time immune-related APA events in the CRC environment. However, more research on APA events in different types of cancer is required to better define common RNA signatures. The APAtizer will be a fundamental tool to analyse TCGA data and give continuity to exploring APA patterns in cancers triggered by tobacco smoking. The candidate genes will then be evaluated to validate their biological activity by our team.

Keywords: Alternative Polyadenylation, RNA Signatures, Cancer, RNA-Seq, Data Analysis, APAtizer

Resumo

O mecanismo celular provém da sua componente transcriptómica. Nos sistemas eucarióticos, o mecanismo de processamento do mRNA da extremidade 3' envolve a clivagem da extremidade 3' do pré-mRNA e a adição de uma cauda de poliadenosina (Poly(A)). O uso de diferentes sinais de poliadenilação (PAS) resultam em transcriptos de mRNA com diferentes extremidades. Esse mecanismo co-transcricional, conhecido como poliadenilação alternativa (APA), proporciona grande variabilidade de isoformas de mRNA a partir de um único gene. Quando o mecanismo de APA ocorre na região não traduzida do 3’ (3'UTR-APA), são produzidos vários transcritos que variam no comprimento do 3'UTRs.

Adicionalmente, o APA pode reconhecer sinais de poliadenilação ao nível do intrão (IPA). Os transcritos gerados perdem parte das regiões codificantes, resultando em proteínas diferentes. Estudos mostram uma correlação entre eventos APA e cancro. 3'UTRs curtos correlacionam-se com células proliferativas, cancerígenas e ativação de proto-oncogenes, enquanto mRNAs truncados por mecanismos de IPA podem afetar as funções supressoras de tumor. Assim, a relação entre APA e cancro ainda está por decifrar. O nosso principal objetivo foi realizar uma análise bioinformática para encontrar assinaturas de mRNA especificamente no 3’UTR-APA e IPA em células do sistema imune de forma a revelar novos alvos de diagnóstico/terapêuticos.

Primeiramente, investigamos como o cancro colorretal (CRC) afeta APA e IPA em macrófagos. Foi realizada uma análise a partir de dados 3'RNA-Seq.

Analisamos macrófagos não polarizados (M0) e macrófagos pró-inflamatórios (M1) isolados três dadores humanos saudáveis e estudamos M1 em co-cultura com duas linhas celulares de CRC (HCT15 e RKO) que diferem em marcadores importantes que afetam as vias tumorgénicas. Para analisar eventos 3'UTR-APA e IPA, aplicamos o APAlyzer, que usa a base de dados Poly(A)_DB para examinar eventos de APA em regiões codificantes, regiões não traduzidas e em intrões. Os resultados mostram mais eventos de encurtamento do 3'UTR-APA e mais IPA durante a polarização de macrófagos (M1) que é mantida quando células M1 são co cultivadas com linhas celulares CRC.

Seguidamente, procuramos compreender como é que o sistema imunológico afeta 3'UTR-APA e IPA na presença de CRC. Assim, oito amostras de ratinhos com CRC foram sequenciadas, quatro amostras são de ratinhos imunocompetentes (IC) e quatro de ratinhos imunodeficientes (ID). Desenvolvemos um protocolo com

ferramentas de última geração para analisar dados de 3'RNA-Seq para comparar eventos do 3'UTR-APA e IPA. Os nossos resultados mostram que os genes de ratinhos ID apresentam um aumento de número de genes que sofrem um encurtamento no 3'UTR-APA e mais IPA quando comparados a ratinhos IC.

O passo seguinte, consistiu em procurar eventos no 3'UTR comuns em cancro desencadeados pelo tabagismo. Recolhemos dados de RNA-Seq processados pelo DaPars usando dados dos projetos GSE174330 (dados de bexiga) e PRJNA356345 (dados de pulmão), e das bases de dados do APAatlas (base de dados de RNA-Seq processados de Genotype-Tissue Expression - dados de bexiga e pulmão de tecidos normais) e TC3A (base de dados de RNA-Seq processados - Dados de tecido tumoral de Bexiga e do Pulmão). Para comparar tecidos normais e tumorais, emparelhamos as amostras de TC3A e APAatlas e intersetamos todas os *datasets*. Não foram encontrados genes em comum.

Os dados do TCGA exigem uma grande quantidade de poder de computacional e de tempo para analisar. Atualmente, estamos a desenvolver o APAtizer, um protocolo automatizado usando o Snakemake que analisa dados de RNS-seq e analisa eventos de APA usando Dapars2 ou APAlyzer. Esta nova ferramenta está testada para três tipos de input: *FASTQ single-end*, *FASTQ paired-end* e ficheiros no formato BAM.

Finalmente, este é, ao nosso conhecimento, o primeiro estudo computacional sobre APA relacionado com o sistema imunitário em ambiente de CRC. No entanto, é necessário mais trabalho de investigação sobre APA e cancro. O APAtizer terá um papel fundamental como ferramenta de análise de dados do TCGA para dar continuidade à exploração de padrões de APA e cancro do tabaco. Com este trabalho, propomos encontrar novos candidatos para funcionar como biomarcadores através de métodos computacionais. Os genes candidatos serão posteriormente estudados e a sua atividade biológica validada pelo grupo.

Palavras-chave: Poliadenilação Alternativa, Assinaturas de RNA, Cancro, RNA-Seq, Análise de Dados, APAtizer

Agradecimentos

Em primeiro lugar, agradeço à Alexandra por me ter lançado o desafio que foi esta dissertação, ter-me dado a oportunidade de crescer num grupo tão bom que é o GR e fazer ciência de uma maneira tão especial e cativante. Agradeço também à Isabel que esteve sempre disponível para tentar perceber o meu "bioinformatiquês" e por todos os conselhos valiosos que levaram este trabalho a bom porto. Mesmo que muitas vezes tenha sido à distância, o GR será sempre sinónimo de boas discussões sobre ciência com otimismo e de como fazer boa ciência. Sou muito grata por todo o crescimento proporcionado, não só como cientista, mas também como pessoa. Um obrigada especial ao Alexandre Ferreira por ter embarcado comigo neste projeto com quem aprendi muito, espero ter ensinado muito também e a quem desejo muito sucesso no seu percurso.

Quero agradecer ao professor Pedro Ferreira por ter estimulado a minha independência enquanto bioinformática e por ter incentivado a minha autonomia, serão sem dúvida skills muito importantes para o meu futuro enquanto profissional.

Quero agradecer aos que fizeram este percurso comigo. Alice, Manuel, Filipe, Rafaela e Helena obrigada por terem tornado esta jornada tão especial, vocês foram a melhor surpresa que este percurso me trouxe. Quero agradecer aos meus amigos que estiveram ao meu lado durante este percurso académico, em especial à Ana Fernandes por ter ouvido as minhas crises existenciais e pela paciência para me descomplicar. Aos meus *Gostosinhos*, por serem os amigos mais antigos e queridos, por vibrarem comigo as vitórias e levantarem-me nas derrotas. Este trabalho também é resultado do vosso carinho, do qual não me podia sentir mais sortuda em merecê-lo. Aos meus avós, que mesmo não sabendo muito bem o que é bioinformática, sentem um enorme orgulho. Em especial, obrigada à minha avó Irene por todos os seus sermões sobre a importância de estudar para ter uma vida bonita, por se preocupar acima de tudo com a minha felicidade. Um agradecimento especial aos meus pais e irmãos, pois sem eles nada disto seria possível. Sou grata pela vossa presença, pelo carinho, apoio e por todos os ensinamentos, que sem dúvida, se refletem na pessoa que me tornei. Obrigada, também por me ouvirem e partilharem do meu entusiasmo durante todo este ano. Ao Ricardo, por ser o meu apoio nos melhores e momentos menos bons, do meu percurso académico e pessoal. Obrigada pelo teu carinho diário, pelos momentos em que foste a minha consciência, a minha motivação e por toda a força que me deste.

Esta dissertação, também é vossa.

“Hiding within those mounds of data is the knowledge that could change the life of a patient or change the world.”
— Atul Butte

Contents

List of Tables

List of Figures

List of Abbreviatures

3'UTR Three prime untranslated region

APA Alternative Polyadenylation

AS Alternative Splicing

aUTR Alternative 3'UTR

BAM Binary Alignment Map

Bp Base pairs

cDNA Complementary DNA

CDS Coding sequence

CFIIm Cleavage factor II

CFIm Cleavage factor I

CLL Chronic lymphocytic Leukaemia

CPSF Cleavage and polyadenylation specific factor

CSTF Cleavage stimulation factor

CR-APA Coding region APA

CRC Colorectal Cancer

cUTR Constitutive 3'UTR

dbGap Database of Genotypes and Phenotype

ddNTPs Unmodified dideoxy-nucleoside triphosphate

DES Downstream sequence elements

DGE Differential Gene Expression

DNA Deoxyribonucleic acid

GDC Genomic Data Commons

GO Gene ontology GRE

GTEx Genotype-Tissue Expression

HTS High-throughput sequencing

IC Immunocompetent

ID Immunodeficient

IGV Integrative Genomics Viewer

INSDC International Nucleotide Sequence Database Collaboration

IPA Intronic Polyadenylation

IPF Internal priming filter

M0 Unpolarized macrophages

M1 Polarized macrophages

MMPs Maximal Mappable Prefixes

mRNA Messenger RNA

NCBI National Center for Biotechnology Information

NCI National Cancer Institute

NGS Next Generation Sequencing

PAS Polyadenylation signal

PE Paired End

PDUI Percentage of Distal Usage Index

PMI Post-mortem-interval

Poly(A) Polyadenylation

Pre-mRNA Precursor mRNA

RAM Random Access memory

RNA Ribonucleic acid

RNA-Seq RNA Sequencing

SAM Sequence Alignment Map

SE Single End

SI Small intestine

snRNAs Small nuclear RNAs~

snRNPs Small nuclear ribonucleic particles

UCSC University of California at San Cruz

USE Upstream sequence elements

TCGA The Cancer Genome Atlas

UTR Untranslated region

VCF Variant Call Format

WHO World Health Organization

Thesis Outline

The developed work that culminates in this dissertation is compiled in five main chapters. Each of them is briefly described below.

Chapter 1 – Introduction: Provides a general overview of the topic investigated throughout the thesis. Includes a detailed review of the approaches proposed over the years enables, the broad identification and the methodologies underlying the identification of APA and IPA events.

Chapter 2 – Aims and Contributions: Delivers the primary motivation of the developed work, goals, contributions.

Chapter 3 – Methodologies: Starts with a description of the databases and algorithms we used to conduct the analysis. Following, a detailed description of the pipeline used to achieve each aim for this work is provided. Finally, we show a comprehensive description of the developed tool – the APAtizer – used to perform some of the previous analyses.

Chapter 4 – Results and Discussion: Addresses the results obtained through the methodology described in Chapter 4 and consequent discussion about the meaning of those results. This Chapter includes the APAtizer performance evaluation.

Chapter 5 – Conclusion: Summarizes and concludes the produced work, culminating with future perspectives and ongoing tasks to complete.

Chapter I - Introduction

Introduction

1.1 - Gene expression and regulation

The blueprint of a cell is its transcriptional regulatory network. The transcriptional regulatory network controls gene expression, one of the most fundamental cellular processes for all living organisms [1], [2]. Consequently, gene expression is the mechanism by which the information encoded in a gene produces mRNA molecules that code for proteins or originate non-coding RNA molecules with other functions. This mechanism requires careful regulation that can be done at transcriptional, co-transcriptional, or post-transcriptional levels.

When the final product is a protein, the synthesis machinery encompasses two main steps: *Transcription* and *Translation* (Figure 1). During the transcription, the formation of mRNA molecules occurs in three major steps. The transcription of protein-coding genes begins with the initiation step marked by the RNA polymerase II binding to the promoter gene [1] [3]. Next is the elongation step with the continuous addiction of nucleotides to the nascent pre-mRNA. Finally, the termination step with the ending of the transcription and the release of the pre-mRNA and RNA polymerase from the DNA template [1][3]

Hence, the product of the transcription of a protein-coding gene originates a precursor mRNA (pre-mRNA). The synthesized pre-mRNA comprises exons (coding sequences) and introns (non-coding sequences) that needs to be processed into a fully mature mRNA [1], [4]. This maturation process is called pre-mRNA processing (Figure 1), a co-transcriptional event that consists of three main processes: capping, splicing and polyadenylation [1][5]. First, a cap structure is added at the 5'end of the transcript preventing the degradation of the pre-mRNA. Next, to remove the introns from the pre-mRNA, the spliceosome is formed, involving the assembly of small nuclear RNAs (snRNAs) and proteins located in nuclear ribonucleic particles (snRNPs), essential elements that catalyses the excision of introns and the ligation of exons to form a mature mRNA [6]–[8]. Finally, the pre-mRNA is cleaved, and the polymerization of an adenosine tail occurs (polyadenylation tail), constituting the mechanisms of pre-mRNA 3'end formation [6]. The processed mRNA is then exported to the cytoplasm through nuclear pores to be translated by ribosomes into proteins [1].

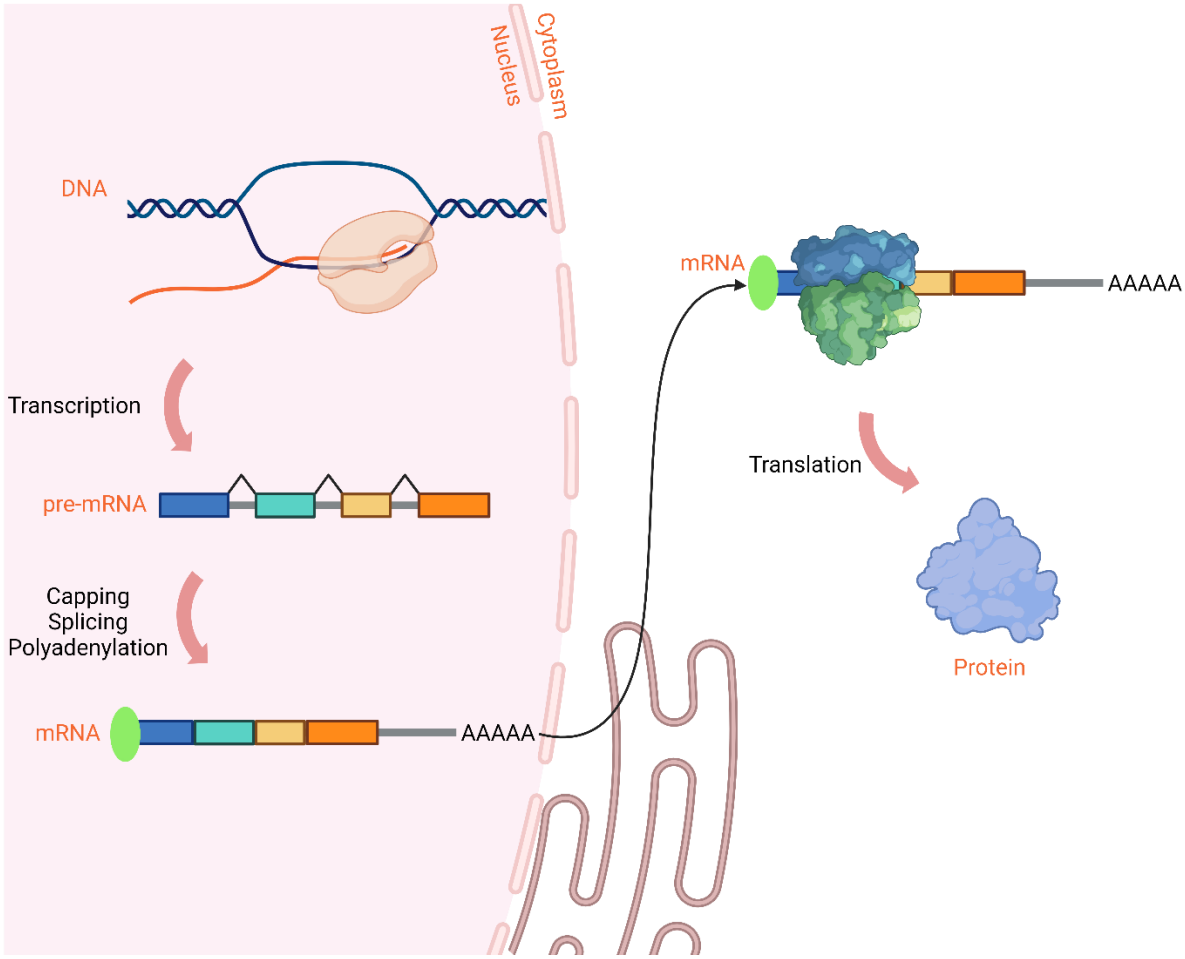


Figure 1 – Pre-mRNA processing in eukaryotic cells. In the nucleus, DNA transcription is mediated by RNA Pol II responsible for the generation of the pre-mRNA transcripts. These pre-mRNA molecules are processed by capping, splicing and polyadenylation giving rise to mature mRNA transcripts. The matured mRNAs are exported to the cytoplasm to be translated into proteins. Adapted from Alberts, Wilson et al. 2008

One single transcriptional unit can produce several mRNA molecules. Transcriptomic diversity found in human cells cannot be explained by the number of human genes, but it is mostly due to the production of different transcripts from one single gene. Alternative transcription initiation, alternative splicing (AS), alternative polyadenylation (APA), and alternative translation initiation can all result in multiple transcripts [1], [9]. These transcripts may possess different stabilities, subcellular localizations, expression levels, produce different proteins, thereby increasing the genome's coding capacity [10].

1.2 - Polyadenylation

The maturation of the mRNA requires a two-step mechanism encompassing pre-mRNA's endonucleolytic cleavage and the addiction of a poly(A) tail at the cleavage site. The sequence that signals 3′ end formation is the canonical sequence that composes the PAS. The PAS comprised of the hexanucleotide AAUAAA is present in almost 60% of the human pre-mRNA and is highly conserved in mammalian transcripts [11]. However, there are at least ten other weaker signals, including AUUAAA, AGUAAA, UAUAAA, and CAUAAA hexamers, function as noncanonical variants of this hexamer [11]–[14].

Four multi-subunit protein factors comprise the mammalian Polyadenylation and cleavage machinery that recognize the PAS: CPSF, a complex containing different subunits responsible for either interacting with the PAS or contributing to the specificity and strength of binding; CstF, which interacts with the DSE and works cooperatively with CPSF to improve binding and affinity to the PAS; cleavage factors Im and IIm (CFIm and CFIIm) that are required to direct the cleavage of pre-mRNA - the first interacts with the pre-mRNA while the second appears to be required for the cleavage reaction - and poly(A) polymerase (PAP) is recruited to the cleavage site and together with CPSF, catalyse poly(A) addition [11][16][17].

The CPSF-73 cleaves the pre-mRNA at the cleavage site, which is usually 15-30 nts downstream of the PAS and up to 20 nts upstream of DSE of the cleavage location site (preferentially a CA dinucleotide), and it is flanked by *cis*-auxiliary elements such as U-rich upstream sequence elements (USEs) and U/GU rich downstream sequence elements (DSEs) [9] [12]. Both USE and DSE have been shown to play a role in enabling PAS usage, and these elements' relative positions and sequence variations can influence the PAS's power efficiency. They form the USE-PAS-DSE pattern for signalling 3′ end processing, which is widely conserved in eukaryotes [15][11], [13], [16].

The poly(A) tail added during the polymerization step remains in the mature mRNA and it is length-specific, depending on the species (eg 150-250 adenines in most mammals, 250-300 in humans and 55-90 adenines in yeasts) [17], [18].

1.3 - Alternative Polyadenylation

Due to advanced next generation sequence technologies, it is acknowledged that more than 70% of human genes produce a great variability of mRNA isoforms due to the selection of different PAS, a process known as alternative polyadenylation (APA) [11], [19]. The choice of different PAS by APA originates mRNA isoforms with

alternative 3' UTRs and this may lead to differences in mRNA stability, localization, or protein efficiency [20].

1.3.1 - Types of APA

There are three major types of APA (Figure 2):

Coding region APA (CR-APA) occurs when a different PAS is chosen in the exon region. This mechanism involves alternative splicing mechanisms combined with APA mechanisms generating different proteins [21].

The 3' Untranslated Region APA (3'UTR-APA) is the most common type of APA and is responsible for most mRNA isoforms. It occurs when mRNA isoforms have different 3'UTR lengths without affecting the resulting protein. It is a cell and gene-type specific mechanism and depends on the cellular physiology. Longer 3'UTRs are more prone to regulation as they contain more *cis*-regulatory elements than shorter 3'UTRs [11]. Long 3'UTRs are correlated to cell differentiation [22]–[24], while short 3'UTRs with cell proliferation and cancer [25]–[27]. Although the protein sequence remains unchanged, 3-UTR-APA affects the resulting protein's production, localization, and function [28].

Intronic APA (IPA) occurs when the chosen PAS is located within introns. Similar to CR-APA, IPA isoforms result in different mRNA isoforms and coding region's truncation lead to the production of proteins with different functions or non-functional proteins. Truncated mRNAs were discovered primarily on tumour-suppressor genes, and the isoforms produced by IPA lacked the usual tumour-suppressive activities reported in full-length isoforms [29]. Several isoforms developed *de novo* oncogenic potential, implying that global IPA alterations inside a cell may be an unknown cancer driver.

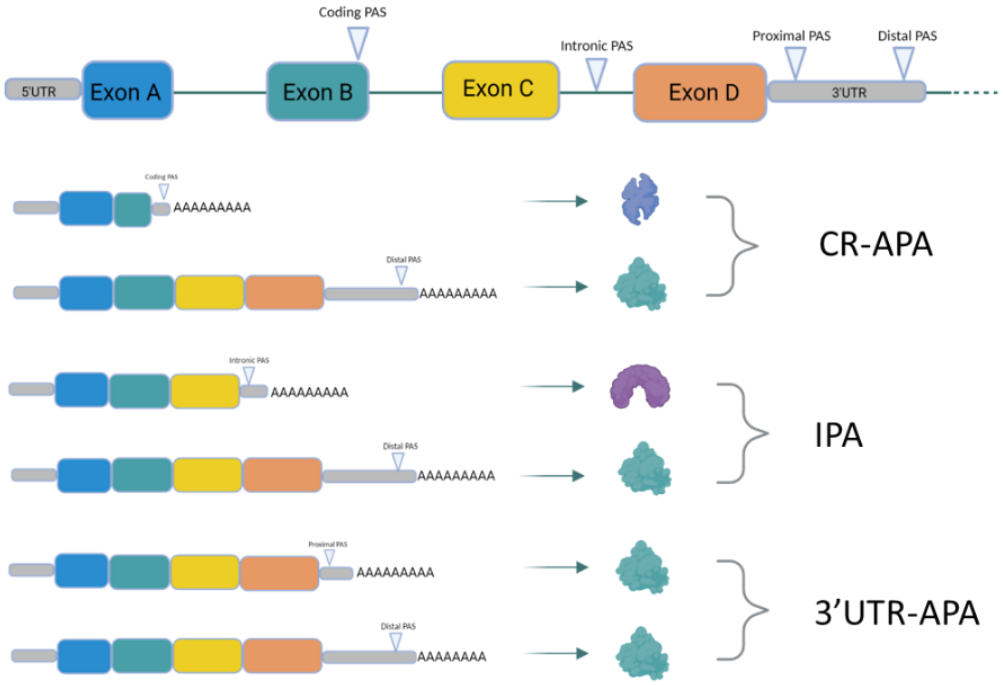


Figure 2 – The types of APA mechanisms. Coding Region APA (CR-APA) and Intronic Polyadenylation (IPA) differs in the c-terminal and consequently produces different proteins. The three untranslated region APA (3'UTR-APA) differs in the size of the 3'UTR without changing the protein structure The 3'UTRs are represented in grey; the coloured rectangles represent RNA exons and the arrows the Polyadenylation signals. - Adapted from Pereira-Castro,I. Moreira,A, Wires RNA 2021

1.3.2 - 3'UTR-APA and cancer

According to World Health Organization (WHO), cancer is the second cause of death worldwide [30]. Therefore, scientists have been studying and improving their methodologies to fight this global health problem by focusing on novel therapies.

The dysregulation of mRNA 3'end processing affects human health and is related to various human diseases, including cancer. Genome-wide studies revealed 3'UTR-APA patterns for certain physiological states and diseases, specifically cancer [26]. Interestingly, most of cancer cells tend to exhibit shorter 3'UTRs [31]. It is suggested that this systemic change helps proto-oncogenes to escape regulation by suppressive elements in their 3'UTRs that normally maintain balanced expression and prevent malignancy [31][32]. Additionally, the shortening of the 3'UTR is related to a mechanism in response to cellular stress to preserve mRNAs and is consistently found in cancers [26], [32].

On the other hand, APA can repress expression by promoting longer isoforms of particular genes. In glioblastoma, one of the most aggressive types of brain cancer,

APA tend to express longer 3'UTR isoforms that mediates lower mRNA levels of the MGMT gene that encodes for a protein that repairs DNA [32], [33]. Despite all the knowledge already acquired in the APA field, there is still a long way to go to fully understand the function of APA in cancer.

1.3.3 – IPA and cancer

The occurrence of IPA in cancer is also widespread. Recent studies with leukaemia patients suggest that truncated proteins originated by an increase of IPA events are prevalent in blood-derived immune cells [29], [34]. In multiple myeloma, the particular pattern of IPA observed suggest that IPA events are upregulated in blood cancers such as leukaemia [35]. Comparing normal and malignant B cells with chronic lymphocytic leukaemia (CLL), it was observed a widespread upregulation of truncated mRNAs and proteins in primary CLL cells caused by IPA. Notably, the truncated mRNAs and proteins lacked the tumour-suppressive functions [29], [36]. Thus, this study unlocked new perspectives for diagnosing cancer, considering mRNA events as significant contributors to cancer proliferation through the inactivation of tumour suppressor genes due to the upregulation of IPA events [29][35].

APA is a complex mechanism regulated by several factors, which altogether control cell physiology [37] and cancer is dependent on many factors. Therefore, it is still unclear if APA is a cancer cause or a consequence. However, with the collaboration of new technologies and bioinformatic tools, science will be able to unveil new APA and IPA profiles or RNA signatures that may be used in the future to investigate their function in cancer, and to be used as biomarkers and therapeutic targets.

1.4 - Next generation sequencing methods and technologies

Since the discovery of the DNA structure in the 1950s, scientists have aimed to determine various genomics. In 1975, Sanger and Coulson developed the first DNA developed sequencing method that allowed them to be the first to sequence the first whole genome of a virus named φiX174, a bacteriophage with single-stranded DNA that infects *Escherichia coli*. The Sanger sequencing methodology represented in Figure 3 consists of 5 main steps and allowed to sequence the human genome in ten years (Human Genome Project). It begins with the primer annealing and chain extension, where the DNA sequence of interest is used as a template for a chain-termination PCR. This PCR has the particularity of the addition of modified

deoxyribonucleotides (ddNTPs) that terminates the reaction. Each of the four ddNTPs has a singular fluorescent label. Next, each chain-terminated oligonucleotide undergoes an electrophoresis to separate the chains. The last step is the gel analysis and the determination of DNA sequence. It involves reading the gel to determine the sequence of the input DNA.

Reading the Sanger sequencing results will depend on which of the two complementary DNA strands is of interest and what primer is available. If the two strands of DNA are A and B and strand A is of interest, but the primer is better for strand B, the output fragments will be identical to strand A. On the other hand, if strand A is of interest and the primer is better for strand A, then the output will be identical to strand B. Accordingly, the output must be converted back to strand A.

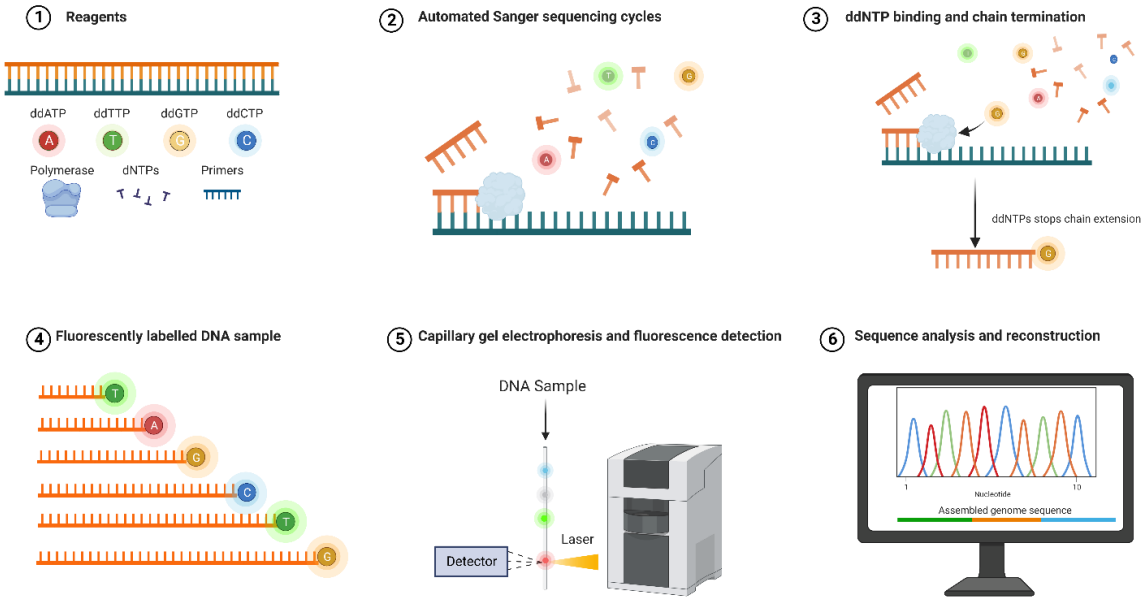


Figure 3 – Sanger sequencing method. 1-Reagens needed to preform Sanger sequencing. 2- primmer annealing and chain extension. The polymerase catalyses the DNA chain by incorporation of the available nucleotides 3- Addiction of fluorescent ddNTPS to terminate each DNA fragment. 4-The DNA fragments are labelled with fluorophores. 5-Electrophoresis to separate the DNA fragments by molecular mass. 6-Sequence analysis. – Adapted from "Sanger Sequencing", by BioRender.com (2022).

The Sanger sequencing method had a crucial role in revolutionising genomics and molecular biology studies. Nevertheless, this technology has limitations such relation cost per sequencing versus quality after 700-900 base pairs (bp) and time per sequencing [38]. In 2000, this technology gave space to Next Generation Sequencing (NGS) methods that are continually improving since its invention.

NGS methods comprise several modern sequencing technologies, including Illumina, Roche, Ion Torrent, PGM, SOLiD, PacBio and Nanopore [33]-[37]. These technologies allow DNA and RNA sequencing with fewer cost and time. High-throughput sequencing technologies have lately become a preferred tool for measuring gene expression with and gene structure. Nowadays, the NGS methods have a significant role in transcriptomic studies [38]-[40]. RNA sequencing (RNA-Seq) is a well-known deep-sequencing method for transcriptome profiling. This powerful approach provides an exact quantification and analysis of transcript levels and mRNA isoforms.

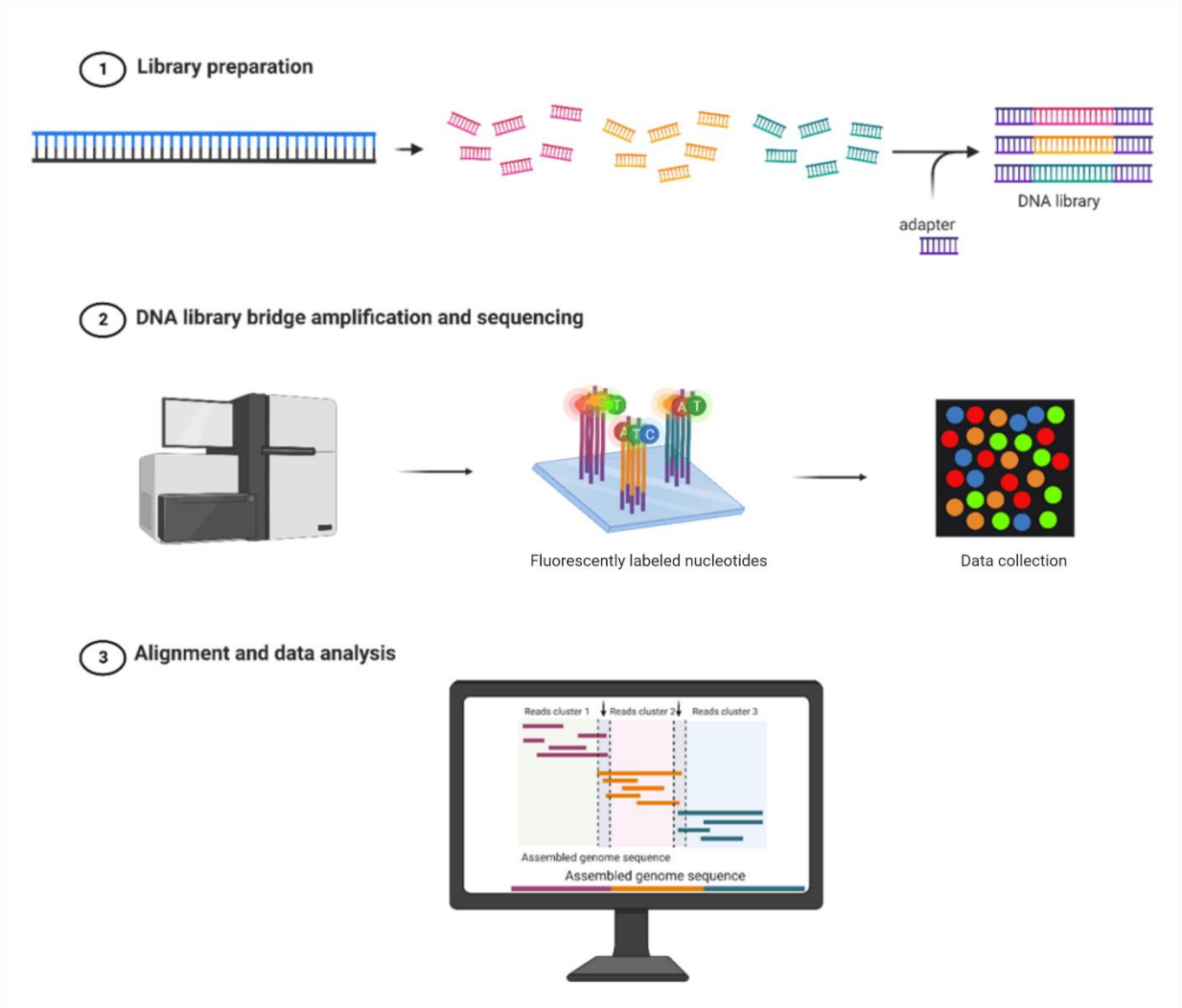


Figure 4 - Next Generation Sequencing method. 1- Library preparation by random fragmentation of DNA and binging the adapters to the DNA fragments. 2-Amplification of the library using PCR and clonal amplification methods and library sequencing. 3-Data Processing and Analysis from sequence reads. - Adapted from "Next Generation Sequencing (illumina)", by BioRender.com (2022).

1.4.1 – RNA-Seq

RNA-Seq provides single-base resolution for annotation and 'digital' gene expression levels at the genome-scale, frequently at a much lower cost than Sanger sequencing (Figure 5).

Transcriptome assembly i.e the reconstruction of the transcriptome is an essential step. After millions of short reads generated by sequencing platforms, they are mapped on a reference genome. Comprehensive transcriptome annotation and assembly are required to analyse transcript expression or perform gene quantification (Figure 6) [39], [40].

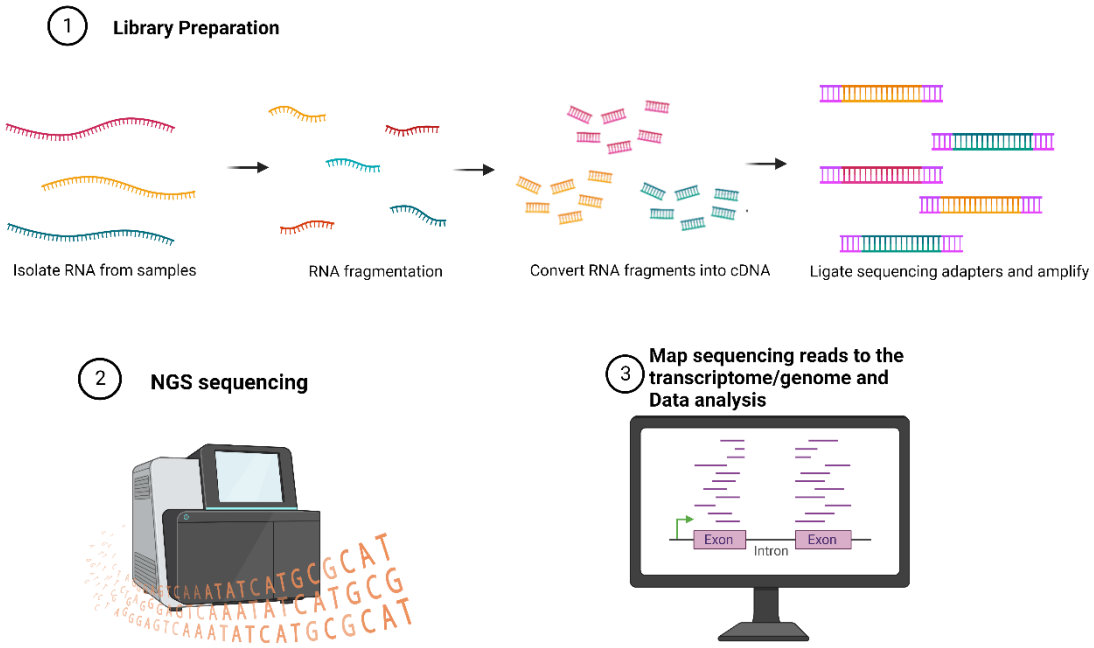


Figure 5 – RNA-Seq workflow. The cDNA library is generated from isolated and fragmented RNA. The cDNA library is sequenced and the reads are mapped against a reference genome or transcriptome. The resultant output data is analysed recurring to bioinformatic methods. - Adapted from "RNA Sequencing", by BioRender.com (2022).

A *de novo* assembly technique does not require a reference genome and can be applied when a high-quality reference genome is unavailable. Typically, a De Bruijn graph is applied to reconstruct transcripts by combining overlapping reads [41]. Most *de novo* methodologies require significant processing resources, with random access memory (RAM) being the most common restriction [41], [42].

The reference-based assembly technique requires the usage of a reference genome and consumes far fewer computing resources than the *de novo* strategy. When employing the reference-based technique, the quality of the genome and the capacity of the short-read aligner to align reads across introns will directly influence the correctness of the assembled transcripts [42].

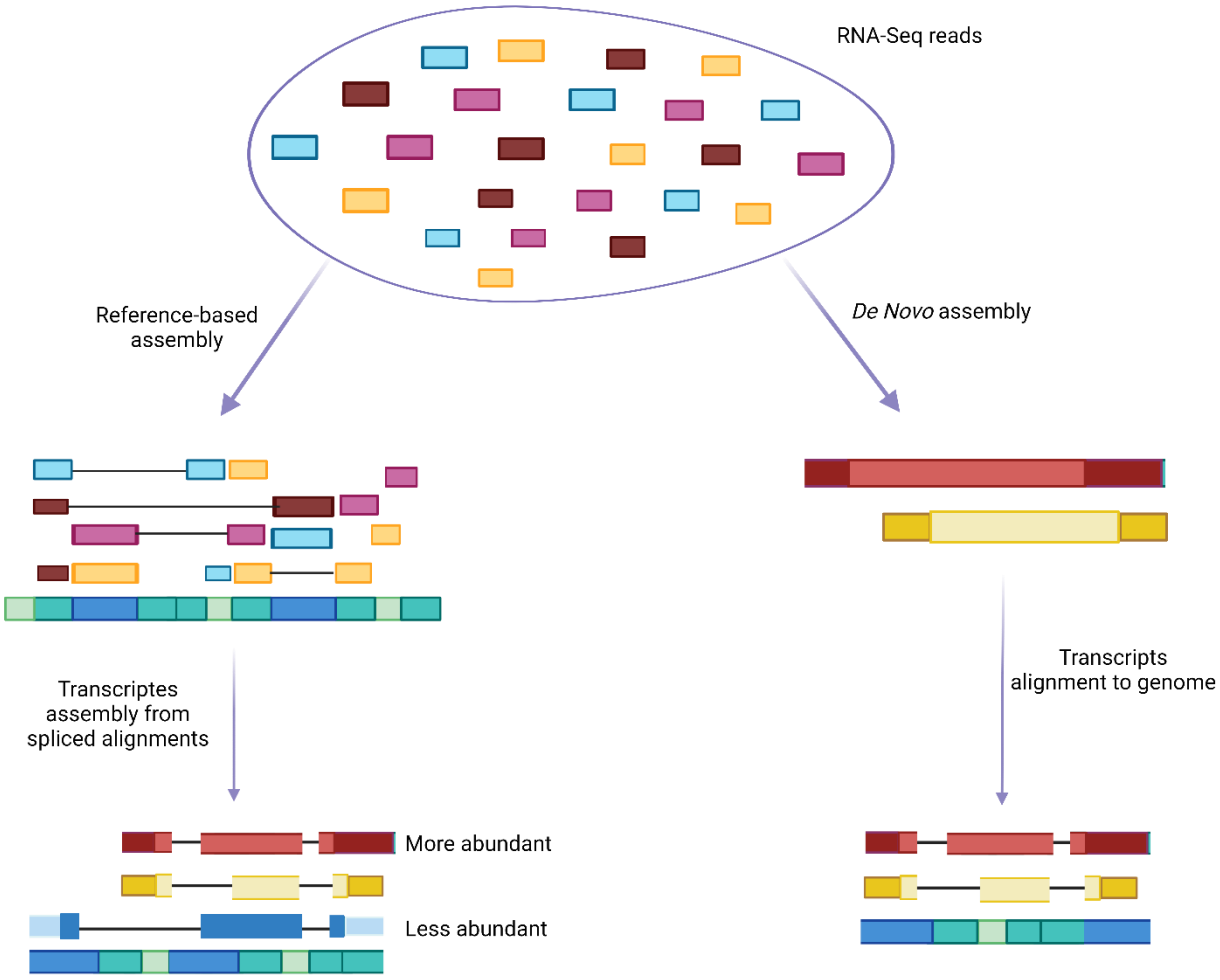


Figure 6- Reference-based assembly vs de novo assembly. The colored rectangles represents RNA-Seq reads. The ab initio approach (left) starts with the alignment of RNA-Seq reads to the genome considering alternative splicing events and then reconstruct transcripts from the alignment. The de novo approach (right) assembles the transcripts directly from the RNA-seq reads. Adapted from Nature Biotechnology. Haas BJ and Zody MC. Advancing RNA-seq analysis. 28:421-423

Nowadays platforms such NCBI, UCSC Genome Browser and ENSEMBL provides the annotation genome assemblies that have been deposited into the International Nucleotide Sequence Database Collaboration (INSDC) [43]–[45].

1.4.2 – QuantSeq 3'mRNA sequencing for RNA quantification

RNA-Sequencing and dedicated bioinformatic algorithms become the finest method for transcriptome profiling of differential gene expression and to analyse APA and IPA events [40].

Despite significant price reductions per base, the processes of sample preparation, sequencing, and data analysis remain significant cost factors in high-throughput screenings [46]. QuantSeq reduces spending in these areas. QuantSeq is an RNA sample preparation method for quantify gene expression accurately and cost-effectively (Figure 7). QuantSeq highly generates strand-specific next-generation sequencing (NGS) libraries close to the 3′ ends of polyadenylated RNAs. Only one fragment is generated for each transcript, directly linking the number of reads mapping to a gene to its expression. QuantSeq shortens data analysis time and allows for more multiplexing per run [46].

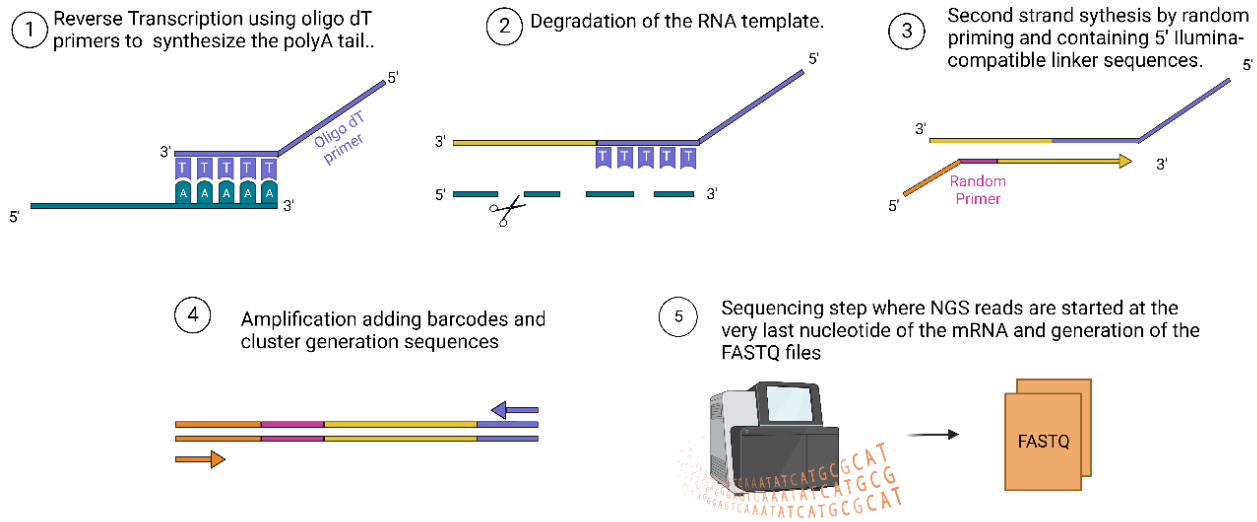


Figure 7- QuantSeq 3'mRNA data workflow. It is divided in three main steps: it starts with the sample preparation where libraries are prepared without the need for poly(A) enrichment. In the sequencing step reads started at the very last nucleotide of the mRNA. Sequencing generates only one fragment per transcript. Finally, the data analysis where the FASTQ are processed by bioinformatic tools, such as APAlyzer, which allowto quantify differential gene expression and to analyse APA and IPA events

1.5 - Bioinformatic identification of APA events

Bioinformaticians have attempted to identify PAS usage information from sequencing data by inferring reading coverage from conventional RNA-Seq or by quantifying read coverage data. Some methods rely on known annotations from curated databases, while others identify peaks from scratch [47].

1.5.1 - Databases for PAS and APA isoforms Storage

Due to the need to store experimentally annotated PAS and 3'UTR-APA mRNA isoforms, public databases were created (Table 1). The primary data came from EMBL annotations records giving place to the UTRdb database and transcript genome alignment in complementary DNA (cDNA) and ESTs that resulted in databases such PACdb, Poly(A)_DB3, and Poly(A) site track [47]–[50]. Thus, primary data inferred from RNA-Seq databases such as TCGA and GTEx projects resulted in TC3A, APAatlas, APAdb, APASdb and Poly(A)Site [51]–[56].

Table 1 – Bioinformatic Databases to store and analyse APA events. – Adapted from Nitika Kandhari et al, Int. J. Mol. Sci. 2021.

Database	Primary Data Collection	Organism	URL
UTRdb	5' and 3'UTR regions in EMBL/GenBank records	Human, Rodent, Plant, Fungi	http://utrdb.ba.itb.cnr.it/
PACdb	cDNA/ESTs	human, mouse, rat, dog, chicken, zebrafish, fugu, fruit fly, mosquito, nematode, Arabidopsis thaliana, rice and baker's yeast	http://harlequin.jax.org/ pacdb/
Poly(A)_DB	Aligned cDNA/ESTs	human, mouse, rat, chicken and zebrafish	http://poly(A)-db.org/v3/
GENCODE Poly (A) site teack	cDNA/ESTs	human	https://genome.ucsc. edu/cgi-bin/hgTrackUi? db=hg19&g= wgEncodeGencodeV19
APASdb	SAPAS	human (22 normal and cancer tissues), mouse, zebrafish	http: //mosas.sysu.edu.cn/utr
APAdb	Mace-Seq	Human, mouse, chicken	http://tools.genxpro.net/ apadb/
TC3A	RNA-Seq in TCGA	32 human cancer types	
APAatlas	RNA.seq in GTEx project	>50 human normal tissue	https: //hanlab.uth.edu/apa/
Poly(A)site	3'-seq, 3'READS, DRS, QUANTSEQ_REV, PAPERCLIP, Poly(A)-seq, PAT-seq	Human, mouse, and worm	https://poly(A)site.unibas.ch/

However, available RNA-Seq data focused on the 3'UTR remains limited. Many cell, tissue, and disease types remain missing. Raw RNA-Seq data from TCGA and GTEx projects requires a controlled access application, in order to protect patients' clinical data, resulting in a long and bureaucratic process with The Database of Genotypes and Phenotype (dbGap) [54], [55], [57].

To address this limitation in cancer research, the APAatlas and TC3A databases provide a resource database of APA inferred from RNA-Seq data. APAatlas employs the DaPars bioinformatic approach on the Genotype-Tissue Expression (GTEx) project, TC3A resorts a similar method, using RNA-Seq data from The Cancer Genome Atlas (TCGA), and the inferred APA genes are available in TC3A [52]. These databases' annotations help to visualize and interpret APA in genome browsers like the

UCSC Genome Browser or the Integrated Genome Browser. Unfortunately, it is not possible to use UCSC Genome Browser to compare normal and tumoral data due to the lack of GTEx data in the browser [55].

1.5.2 – Bioinformatic tools for APA detection and quantifications

The growing interest in 3'UTR dynamics and the development of associated technologies required the development of specific bioinformatics tools. Multiple methods for inferring APA from conventional RNA-Seq were developed (Table 2) [58]. Some APA detection methods rely on prior knowledge, while others involve discovering PAS from scratch [47]. Moreover, many APA detection and quantification tools depend on database annotations to guide bioinformatic analysis, as discussed in the section below.

Using APA information from databases improves the accuracy of *in silico* APA detection. Several tools were developed to detect APA events. Apalyzer, QAPA and Roar are three of the tools that find APA events using previously annotated 3'UTR isoforms, to quantify 3'UTR-APA [51]. DaPars, APAtrap and TAPAS allow a *de novo* poly(A) site identification.

Dynamic analysis of alternative polyadenylation from RNA-Seq (DaPars) identifies APA in RNA-Seq experiments from scratch [59]. The method scans the read coverage and identifies a distal PAS at the endpoint of the sample with the longest 3'UTR. The model that provides the best least-squares fit of the read coverage along the gene up to the identified distal PAS is then sought. This model comprises the location of a proximal PAS and the expression levels of the short and long isoforms in each of the two conditions. This best model provides the proximal PAS location and the information needed to calculate the Percentage of Distal poly(A) site Usage Index (PDUI) for each condition [59].

APAlyzer is a Bioconductor package that calculates the RNA-Seq read density (RD) after splitting the transcript 3'end regions based on Poly(A)_DB3 annotations to identify APAs in 3'UTR and intronic regions. This recent algorithm is the first able to detect IPA events [60].

Table 2 – Bioinformatic approaches to analyse APA mechanism. – Adaped from Bin Tian et al. ,2021

Feature	APAlyzer	APAtrap	DaPars	QAPA	Roar	TAPAS
Reference Database	Poly(A)_DB	-	-	Poly(A)_Site/Genecode	Poly(A)_DB & APASdb	-
***De novo* poly(A) site identification**	-	✓	✓	-	-	✓
3'UTR-APA	✓	✓	✓	✓	✓	✓
IPA	✓	-	-	-	-	-
Vizualization	✓	-	-	-	-	-

Recent scientific and technological advances have paved the way for the discovery and investigation of APA signatures and their relationship to cancer. An immune- or cancer-specific APA signature has the potential to be a clinical biomarker and open doors to new therapies. This possibility can solve many cancer treatment problems, such as circumventing cancer treatment resistance and presenting alternatives to conventional treatments such as chemotherapy. All the reported potential clinically relevant APA genes detected by computational power have to be validated in the laboratory and their capacity to be a therapeutic target or biomarker in a clinical setting.

Chapter II – Aims and Contributions

Thesis Aims and Contributions

Bioinformatics is increasingly needed in modern science. Through computational means, it is possible to process and analyse biological data efficiently, an indispensable requirement for the success of any research project. Notably, it can also be used for processing and analysing RNA-Seq data and exploring mRNA signatures in cancer.

Although it is known that alterations in APA mRNA isoforms can be related to cancer and to the activation of proto-oncogenes, the relation between APA and cancer is still to be unveiled. This work aimed to discover new mRNA signatures in cancer, comprehend how cancer cells impact the transcriptome of the immune cells and how the immunological background affect cancer cells, and to disclose new potential diagnostic/therapeutic targets.

The existence of a bioinformatician in a research group to analyse RNA-Seq data is not always possible as processing data can be time-consuming and, most of the time, requires prior computational knowledge. Therefore, motivated by promoting computational research independence to colleagues without computational knowledge, I also aimed to build a new user-friendly tool to analyse 3'UTR-APA signatures from raw RNA-Seq data.

In this regard, the specific aims of this work were to:

1- Investigate how the polarization of primary human macrophages to pro-inflammatory macrophages affects the transcriptome. Moreover, understand in what manner 3'UTR-APA and IPA expression of human macrophages are affected when co-cultured with colorectal cancer (CRC) cells.
2- Discover in what way the immune system affects 3'UTR-APA and IPA in CRC in mice.
3- Identify common 3'UTR-APA events in the lung, bladder, and head & neck cancers, which are triggered by tobacco smoking.
4- Create and optimise the pipeline to analyse 3'UTR-APA and IPA, transforming the automated pipeline into a user-friendly tool – the APAtizer.

Chapter III - Methodology

Methodology

3.1 – Databases

TC3A

TC3A is a database with data processed by DaPars2 software that comprises 10537 RNA-seq tumour samples from 32 TCGA cancer types from the UCSC Cancer Genomics Hub [52]. The data is deposed in synapse, an open-source platform for collaborative data analysis. (www.synapse.org, synapse ID: syn24982198) [52], [59].

APAatlas

APAatlas is a database with data processed by DaPars software that comprises 9611 RNA-seq samples from normal tissues and encompasses 53 normal tissues from the GTEx portal. The RNA-seq samples were aligned with hg19 genome using HISAT2. The BAM files were sorted and converted to BEDgaph using BEDtools. The BEDgaph files were submitted to DaPars algorithm. The data is deposed in a txt format in the APAatlas user-friendly database interface (https://hanlab.uth.edu/apa/) [53].

The Cancer Genome Atlas (TCGA)

The TCGA is a project that analyse human cancer samples using different high-throughput technologies to provide publicly available datasets to improve the cancer research [54]. The project was engaged by the National Institute of Health (NIH), NIH's National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) funded by the United States (US) government. The TCGA project is responsible for sample collection and processing, high-throughput sequencing (HTS) and bioinformatic analysis [54].

The RNA-seq data was sequenced based on the Illumina system and promotes the total transcriptome profiling. The repository of sequence data is NCBI dbGap responsibility [57].

The TCGA data is deposited on NCI Genomic Data Commons (GDC) (https://gdc-portal.nci.nih.gov). The RNA-seq raw data was harmonized through the alignment of the raw sequencing data to the human genome and is available in a binary alignment Map (BAM) file format.

Genotype-Tissue Expression (GTEx)

Similarly, to the TCGA project, the GTEx is a project launched by NIH and NCBI dbGap is responsible for the sequencing data repository [55].

The project built a database available on the GTEx data portal (https://gtexportal.org/home/) with the aim of helping the scientific community to study gene expression and genetic modifications in human tissues. Hence, the project is responsible for sample collection and processing, storage high-throughput sequencing (HTS) using Illumina system and the release of data on dbGap.

GTEx samples were obtained through low-post-mortem-interval (PMI) autopsy or organ and tissue transplant settings. Similarly, with TCGA database, the RNA-seq raw data is available in a BAM file format.

Database of Genotypes and Phenotypes (dbGap)

The dbGap is (http://www.ncbi.nlm.nih.gov/gap) is a NIH sponsored repository produced by genotype and phenotype studies [57]. The dbGaP website provides free access to publicly accessible metadata about submitted research, summary level data, and study-related papers. Individual-level data are available to scientists worldwide via the Controlled Access application [57].

3.2 – Programs & Softwares

Internal Priming Filter (IPF)

The IPF software, developed by Lexogen, uses thirteen hexamers sequences and comprises three modules:

1- The Transcript Termination Site (TTS) primary filter verifies if the start of a read lies within a sliding window (default +-10 nt) around species-dependent, annotated transcription termination sites. Alignments which satisfy this criterion are considered valid (hence not resulting from internal priming). The remaining reads will be the input for the next step.

2 - Next, the motif filter was used to check the reads for the presence of the thirteen motifs in a region up to 40 nt upstream of the read start position. Reads with a motif are considered as potentially belonging to a poly(A) site and will pass to the last filter.

3 - Internal priming occurs if a poly(A) sequence appears within a transcript. Sequences within transcripts are contained in the reference of the respective organism. Priming at the 3' end of transcripts occurs due to the poly(A) tail, which is not included in

the reference. Hence, the downstream A content (DAC) filter investigates the downstream region of the read start. If the number of A(s) in a reference sliding window (default 18 nts) downstream of the read start is larger than a given threshold (default nine nts), this indicates internal priming, and the read is excluded. Otherwise, the read is accepted.

Therefore, all reads close to an annotated TTS that have a motif upstream and a low A content downstream are kept. All other reads are removed.

In both datasets, quality control was performed using FASTQC software (v.0.11.9), and adapters were trimmed with Cutadapt (v.3.7), discarding reads under ten bases.

The Spliced Transcripts Alignment to a Reference (STAR) aligner

The STAR aligner was developed in C++ and promoted the alignment of high-throughput RNA-seq data to the reference genome, using FASTQ files as the input.

The program maps the reads through two main steps: Seed searching and Clustering, Stitching, and Scoring [61].

Seed searching consists of searching for the longest matching sequences for each read in different parts of the genome. Those sequences will be the Maximal Mappable Prefixes (MMPs). The different locations where the match happens are the seeds. The program repeats the cycle with the seed searching with the next longest sequence.

After the seed searching the seeds clustered based on proximity to a set of seeds. Finally, seeds are stitched together based on the best score[1] for the read alignment.

The output of the alignment it's a BAM file which contains the sequence alignment data.

Dapars2

DaPars2(v.2.1) is the next generation of DaPars. It is a Python-based open-source program that identifies and quantifies dynamic APA events from scratch, regardless of prior APA annotation. This program allows to compare normal and tumour pairs samples through RNA-seq data [59].

Based on RNA-seq data, the program identifies a distal poly(A) site, employs a regression model to infer the position of the proximal APA site after correcting for

[1] The best alignment score is based on mismatches, insertions, deletions and gaps.

potential RNA-seq non-uniformity along the gene body, detects statistically significant dynamic APAs, and has the potential to detect more than two dynamic APA events.

Given two or more RNA-seq samples, the distal poly(A) site refers to the endpoint of the longest 3′UTR among all samples. Therefore, the distal poly(A) sites will be used to identify the proximal poly(A) within this longest 3′UTR region. Hence, the annotated gene 3′end by up to 10 kb before reaching a neighbouring gene to identify potential distal poly(A) sites that may exist outside gene annotation. To obtain a combined coverage along the extended gene model, RNA-seq data from all input samples were merged [62].

To quantify the relative poly(A) site usage, the program uses the Percentage of Distal Usage Index (PDUI) for each sample (3).

$$PDUI = \frac{w_L^i}{w_L^i + w_s^i} \quad , \in [0,1] \qquad (3)$$

Here w_L^i and w_s^i are the estimated distal and proximal poly(A) sites expressed on transcripts of a sample i. Hence, the PDUI is the abundance of the long isoform divided by the total abundance (Figure 8). Each expressed transcript in the sample i has a PDUI value comprised in an interval between 0 and 1, where 0 indicates no usage of the distal poly(A) site and 1 suggests only usage of the distal usage; therefore, no usage of the proximal poly(A) site [59], [62].

$$\Delta PDUI = (PDUI_{Tumor} - PDUI_{Normal}) \quad , \in [-1,1] \qquad (4)$$

Lastly, to compare the 3'UTR-APA between normal and tumour samples pairs, i.e. normal and tumour samples from the same donor can be calculated using ΔPDUI (4). ΔPDUI is comprised in an interval between -1 and 1, where it is possible to observe a 3'UTR lengthening in case of a positive index and a 3'UTR shortening in case of a negative index. A significant difference between normal and tumour sample is considered if transcripts with a ΔPDUI > 0.2 were considered lengthened and ΔPDUI < -0.2 were considered shortened [62].

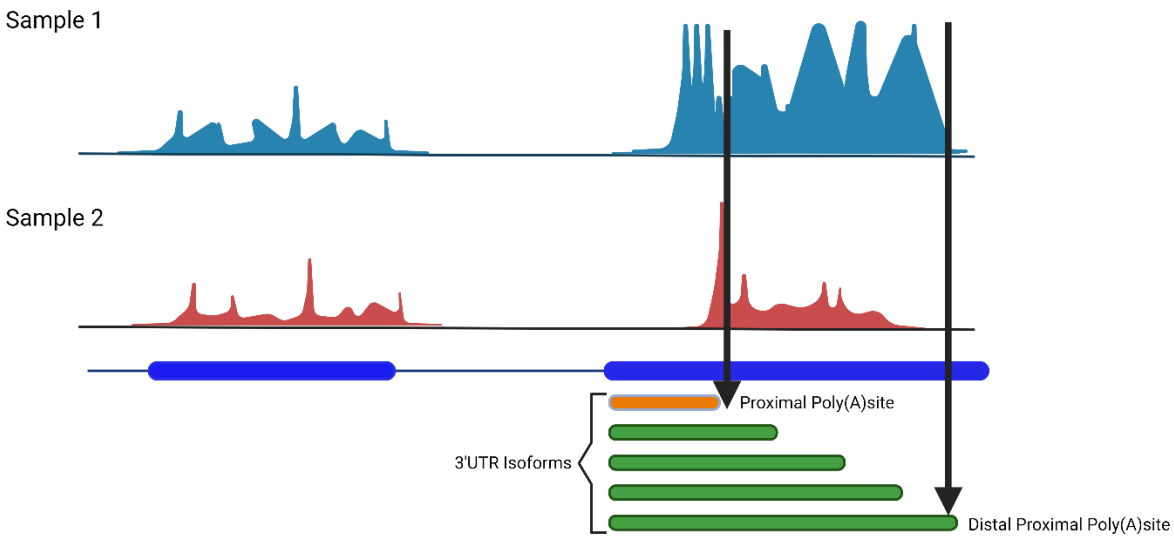


Figure 8- Scheme of DaPars2 algorithm. The DaPars2 predicts the proximal Poly(A) site (short isoform) and the Distal Proximal Poly(A) site (long isoform). To quantify the relative APA usage the PDUI is calculated by dividing the abundance of the long isoform by the total abundance. - Adapted from Genome Research. Wei Li et all. 2021

APAlyzer

APAlyzer is a Bioconductor package and a bioinformatic toolkit for analysing APA events, developed in R statistical programming language [60]. The premise of this program is the analysis of the 3-UTR-APA and IPA isoforms based on conserved poly(A) sites (PAS) annotated in the Poly(A)_DB database (Figure 9) [60].

The first step is to sample data and search for PAS references. This package reads coverage data from various genomic regions using BAM files from RNA-seq data as input. It loads the pre-built 3'UTRs and introns PAS references for the chosen genome. The pre-built mm10 version was used for mice samples and the hg38 version for the human primary macrophages' samples.

The next step is calculating the 3'UTR APA relative expression density (RE) using the PASEXP_3UTR function (1). The region between the stop codon and the first PAS, known as the constitutive 3′UTR (cUTR), and the region between the first and last PASs, known as the alternative 3′UTR (aUTR), are measured using the RNA-seq read density (RD) determined by dividing the number of mapped reads by the length of the region.

$$\mathrm{PASEXP}_{3\mathrm{UTR}} = log_2\left(\frac{RD_{aUTR}}{RD_{cUTR}}\right) \qquad (1)$$

The output data frame of 3'UTR and IPA functions covers reads count and RPKM for aUTR and cUTR and RE for each gene.

To calculate the IPA RE, the APAlyzer uses the PASEXP_IPA function (2). RD_{IU} represents the read density of the intronic upstream region of the IPA site, RD_{ID} represents the read density of the intronic downstream region of the IPA site, and RD_{TE} represents the read density of the constitutive region in the 3' terminal exon.[60]

$$\mathrm{PASEXP_{IPA}} = log_2\left(\frac{RD_{IU} - RD_{\mathrm{ID}}}{RD_{TE}}\right) \qquad (2)$$

The output data frame of 3'UTR and IPA functions covers reads count and RPKM for aUTR and cUTR and RE for each gene.

To calculate the IPA RE, the APAlyzer uses the PASEXP_IPA function (2). RD_{IU} represents the read density of the intronic upstream region of the IPA site, RD_{ID} represents the read density of the intronic downstream region of the IPA site, and RD_{TE} represents the read density of the constitutive region in the 3'terminal exons.

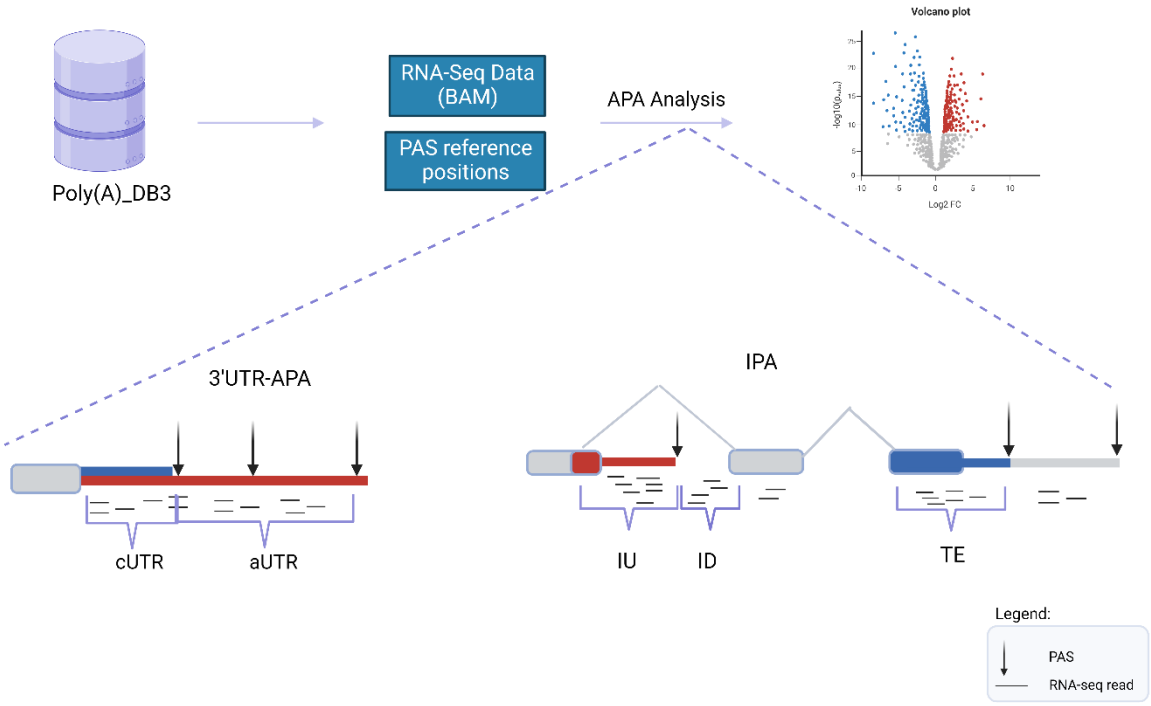


Figure 9 – Scheme of APAlyzer algorithm. The arrows represent PAS and the lines RNA-seq reads. There are two types of APA analysis: The 3'UTR-APA analysis are performed with the calculation of the relative expression between aUTR and cUTR. In IPA analysis the analysis is performed by computing the relative expression between IU and ID dividing by the relative expression of TE – Adapted from Bioinformatics. Tian B. et all. 2020

SAMtools

SAMtools is a library and a well stablished software package to manipulate alignments in Sequence Alignment Map (SAM) and Binary Alignment Map (BAM) formats [63]. This multifaceted toolkit provides diverse features such sort, merge, and

index alignments, remove PCR duplicates, convert BAM/SAM files from other alignment formats and inspect information from BAM/SAM files [63].

GATK, Picard and MarkDuplicates

GATK is a genomic analysis toolkit developed by the Broad Institute (https://gatk.broadinstitute.org/hc/en-us). It analyses genomic and clinical data, including RNA-seq data. GATK also englobes many tools to process HTS data, including the Picard open-source toolkit [60]. Therefore, Picard is a set of command line tools for manipulating HTS in SAM/BAM and variant call format (VCF) files. The MarkDuplicates tool is one of Picard's tools. This tool compares sequences in the reads' five prime positions. After collecting duplicate reads, the tool distinguishes between primary and duplicate reads using an algorithm that ranks reads based on the sums of their base-quality scores. MarkDuplicates also generates a metrics file that contains the number of duplicates for both single-end and paired-end reads [60]. SAMtools can also remove duplicates with rmdup tool. However, unlike MarkDuplicates, rmdup cannot remove intrachromosomal duplicates.[64]

BEDtools

The BEDtools toolkit provides a vas set of utilities for common operations on genomic features. This software suite for the comparison, manipulation, and annotation of genomic features in BAM, VCF, Browser Extensible Data (BED) and General Feature Format (GFF) format [65]. It provides a fast and flexible feature such as conversion from BAM to BED or FASTQ formats, computes the coverage over an entire genome, merge and intersect between two sets of genomic features [65].

HTSeq

The HTSeq is a Python Based Library that promotes the analysis of high-throughput sequencing (HTS) data. In the analysis of RNA-seq data for differential expression analysis, the HTSeq-count tool was used to count the overlap of reads with genes.

This function counts how many aligned reads overlap each gene's exons using the BAM files obtained on the data pre-processing step as input. Reads that align to multiple positions or overlap with more than one gene are discarded [66].

The output is a text file with a column with the name of genes and the respective count. These counts can then be used to perform gene-level differential expression analyses with tools like DESeq2 [67].

DESeq2

The DESeq2 is a Bioconductor package developed in R statistical programming language and estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution [67]–[69]. As a result, this package analyses comparative RNA-seq data using shrinkage estimators for dispersion, shrinkage fold change, standard error estimates, improved gene ranking and visualisation, hypothesis tests, and the regularised logarithm transformation for quality assessment and clustering of dispersed count data [67].

Snakemake

Snakemake is a python workflow manager that organize and automatize pipelines. It is a complete and flexible tool that can be configured to accept files stored on disk or in a remote storage such AWS, Azure, or Google [70].

The main idea is that workflows can be decomposed into small steps represented as rules. Each rule has three major steps: Input, Output and Shell. Each rule has defined the output obtained from a certain input file. In the shell step, use the command for the y enclosing the respective keywords in curly braces. Rules can be generic using wildcards. Cards allow the iteration per samples replacing the wildcards names with any non-empty script [70]. The tool requires technical knowledge in python and domain knowledge about the workflow that is going to be implemented [71]. However, the Snakemake script should allow maximum readability.

3.3 - APA signatures from the Immune system and relation to CRC.

3.3.1 - Dataset Preparation

For 3’mRNA-Seq (3’Seq), the QuantSeq 3’ mRNA-Seq Library Prep Kit REV for Illumina (Lexogen) was used to prepare the mRNA libraries using the RNA extracted from the macrophage and mice samples. The libraries were prepared at i3S and sequenced by the Lexogen company [72].

The primary human macrophage dataset is comprised of nine samples from three different healthy donors. For each donor, samples are constituted by two types of

macrophages, M0 and M1 and M1 co-cultured with two distinct cancer cell lines: HCT15 and RKO (Figure 10).

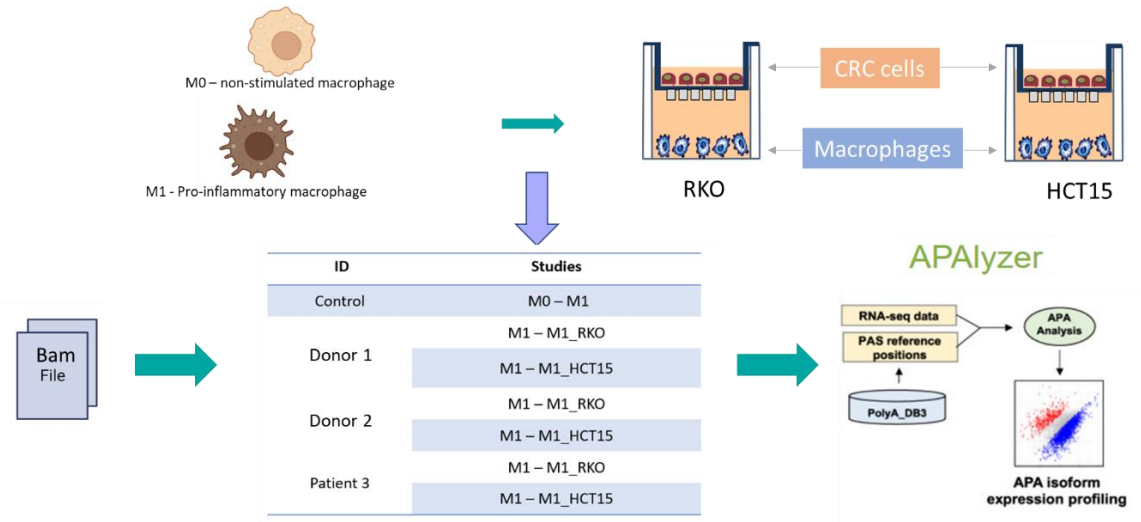


Figure 10- APA signatures analysis of Immune Cells workflow. The M1 samples were co- cultures with RKO and HCT15, two colorectal cancer (CRC) cell lines. The libraries were prepared by the group, using QuantSeq 3'mRNA-Seq Illumina kit and sequenced by lexogen. We employed the APAlyzer for examining 3'UTR APA and intronic APA changes, using RNA sequence aligned data.

The mice dataset was composed of eight CRC samples prepared at i3S from four immunocompetent (IC) mice and four immunodeficient (ID) mice, to study the mice immune system influence CRC. The IC mice was prepared using the VCMsh2LoxP/LoxP model where Msh2 was knock-out to promote CRC and ID mice were prepared using the VCMsh2LoxP/LoxP;Rag2-/-;Il2rg-/- model, where genes Msh2 *Rag*[2] and *Il2rg* were knock-out to promote CRC and immunodeficiency. CRC samples were retrieved from tumours located in the colon and small intestine (SI), two for each mice group. All mice experiments followed standard animal care guidelines and national legislation for animal experimentation (Figure 11).

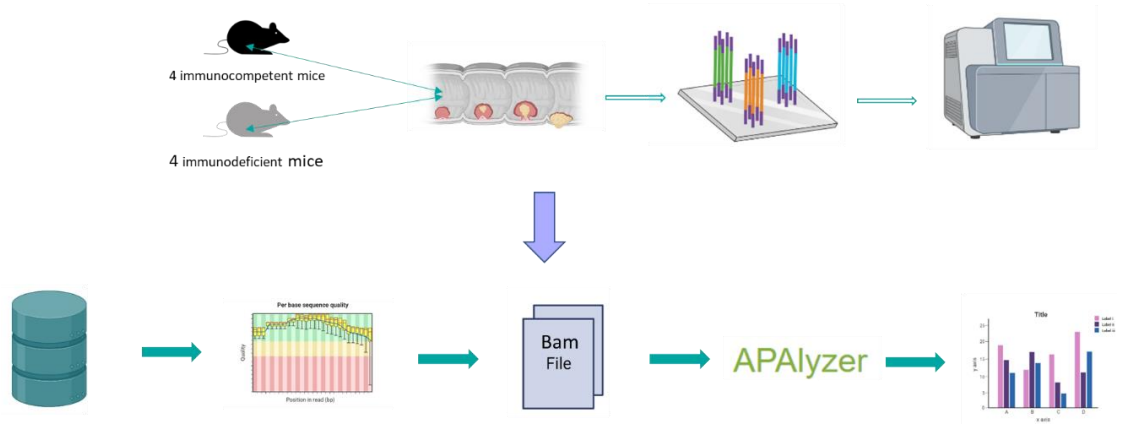


Figure 11 – Workflow to study the effect of the immune system on CRC APA events. We analysed four samples from immunocompetent mice and four from immunodeficient mice with colorectal cancer. The samples were sequenced by lexogen. We develop a workflow that analyses 3' RNA-seq raw data, the FASTQ files, and compares immunocompetent and immunodeficient mice to find new RNA signatures

3.3.1 - Data Pre-processing

All 3'seq raw data were filtered to detect accurate PASs with the IPF program develop and applied by Lexogen, to filter reads resulting from internal priming [72]. The IPF was applied on FASTQ samples for the primary human macrophage dataset and BAM files for the mice CRC cells dataset. Hence, the pre-processing data methodology was slightly different for the two studies. Next, the mice samples were mapped and aligned to the *Mus musculus* (mm10) genome reference, while the human primary macrophage samples to the *Homo sapiens* (hg38) genome reference using STAR (v.2.7.9). Finally, the quality of the aligned reads was controlled with Flagstat using SAMtools (v.1.12).

3.3.2 - Alternative Polyadenylation Analysis

The mRNA isoform differences were studied using APAlyzer software (v.3.0). To filter out genes that have a small number of reads the APAlyzer author, Bin Tian, suggested a cutoff = 5. The cutoff was applied on the 3'UTR-APA and IPA analysis.

3.3.3 - Differential gene expression analysis

A differential gene expression analysis was performed to comprehend the immune system effects on CRC expression in mice. The htseq-count from HTSeq (v.0.11.2) was used to count the number of reads mapped to each gene. The gene expression differences were calculated using DESeq2 (v.3.15). The analysis only considered genes with >10 mean TPM, absolute Log2Fold Changes of 2 or -2 and adjusted p-values (p-adj) to 0.05 ,to increase the statistical robustness of our analysis.

3.3.4 - Gene Ontology and Gene Pathways Analysis

The genes enrichment analysis was performed using the Gene Ontology (GO) (http://geneontology.org/), using the PANTHER classification system (http://pantherdb.org/) [73][74].

3.3.5 - Integrative Genomics Viewer (IGV)

The lists of genes obtained from the APAlyzer analysis and differential gene expression were submitted were explored using the IGV (v.7.0). The IGV is an interactive tool for the visual exploration of genomic data. The sequencing read distribution through the genome uses the BAM files as input and BAM index files (BAI). The BAI files were obtained through the index function from SAMtools [63].

3.4 - APA signatures in cancers triggered by tobacco smoking

3.4.1 - Sample Collection and Characterization

Blader (BLCA) and Lung (LUAD) processed data from GTEx (normal samples) are available in APAatlas (Hong et al., 2020), and TCGA data (tumour samples) are available at TC3A (Feng et al., 2018). The datasets are available in text files.

The FASTQ files of the Lung dataset comprises 98 normal/tumour sample pairs and are available in the GEO repository under accession number GSE174330 [75]. The bladder dataset comprises 15 normal/tumour pairs in FASTQ file format and can be found in Bioproject under the accession number PRJNA356345 [76]. The *FASTQ* files were analysed using a RNA-seq workflow with state-of-the-art tools. The quality control was performed using *FASTQc.* Next, Cutadapt was used to trim poly(A) tails, adapters, and quality trimming [77]. The trimmed sequences were chosen according to the studies where samples were sequenced primarily, and STAR was used to map the reads to the human genome (hg38) [61]. The output files were created as BAM files, sorted by index.

The BAM files were converted to *BEDgaph* using *BEDtools* [65]. The *BEDgaph* files were submitted to DaPars2 program [59]. The output was in text format.

3.4.2 - Alternative Polyadenylation Analysis

To analyse *DaPars2* data from all databases, an R script was developed to obtain the genes that have undergo 3'UTR shortening or lengthening. Data analysis the pre-processed data retrieved from APAatlas and TC3A was used as a normal/tumour combination. The Lung (GSE174330) dataset and the bladder (PRJNA356345) dataset contained samples from normal and tumour tissue from the same patient. The mean PDUI values from each gene were calculated, and the ΔPDUI (PDUI tumour – PDUI normal) was computed. Genes with a ΔPDUI > 0.2 were considered lengthened and ΔPDUI <- 0.2 were considered shortened.

3.4.3 - Integrative Genomics Viewer (IGV)

The top ten of genes obtained from the *APAlyzer* analysis and differential gene expression and explored using the *IGV* (v.7.0). The *IGV* is an interactive tool for the visual exploration of genomic data. The sequencing read distribution through the genome uses the *BAM* files as input and *BAM* index files (*BAI*). The *BAI* files were obtained through the index function from *SAMtools* [63].

3.5 - APAtizer – An Automated APA Analysis Workflow

3.5.1 - Overview of APA's structure and workflow

To analyse APA shortening and lengthening events and compare results from different studies and different software's, it is necessary to create a workflow to process all data uniformly. However, running each step at a time and waiting for the process to finish for all samples is time-consuming. Moreover, the data available on databases can be in different formats, such as FASTQ files with raw data or deposed with a previous alignment in a BAM file format. Therefore, my solution was to create a tool that provides an automated workflow that englobes different types of inputs considering different workflow specificities to uniformise different samples as input and obtain the analysis of APA.

3.5.2 – APAtizer Pipeline

The *APAtizer* is an automated tool is built with *Snakemake* workflow manager [70]. The *APAtizer* analyses human samples using hg38 as reference genome and uses as input two file formats: *BAM* and *FASTQ* (SE and PE files). The input is composed of three arguments: 1) the type of file used as input (*BAM, FASTQ* SE or

PE), 2) the software we want to analyse APA events (*DaPars2* or *APAlyzer*) and 3) the number of cores to run the process (Figure 12).

The tool comprises existent tools to analyse RNA-seq data using Python functions. The workflow consists in four major steps: Quality Control, Trimming, Genome Alignment and/or sorting aligned samples and APA analysis. *BAM* files from TCGA are unsorted. Hence, the first step is to sort the *BAM* files using *SAMtools.* Both type of inputs undergoes for a quality control step before and after the trimming step. *BAM* files are trimmed by Picard software to remove the duplicates and *FASTQ* files are trimmed by Cutadapt. *FASTQ* files are used for a genome alignment. Next, if the samples are going to be analysed by *Dapars2* there is an extra step were *BAM* files are converted in *BEDgaphs* and the indexes of *BEDgaphs* are created using *SAMtools*. Then the indexes built in the previous step and the *BEDgaphs* will be used as input to *Dapars2* software. On the other hand, if *APAlyzer* is chosen then the APAtizer will run the R script using the *BAM* files as input.

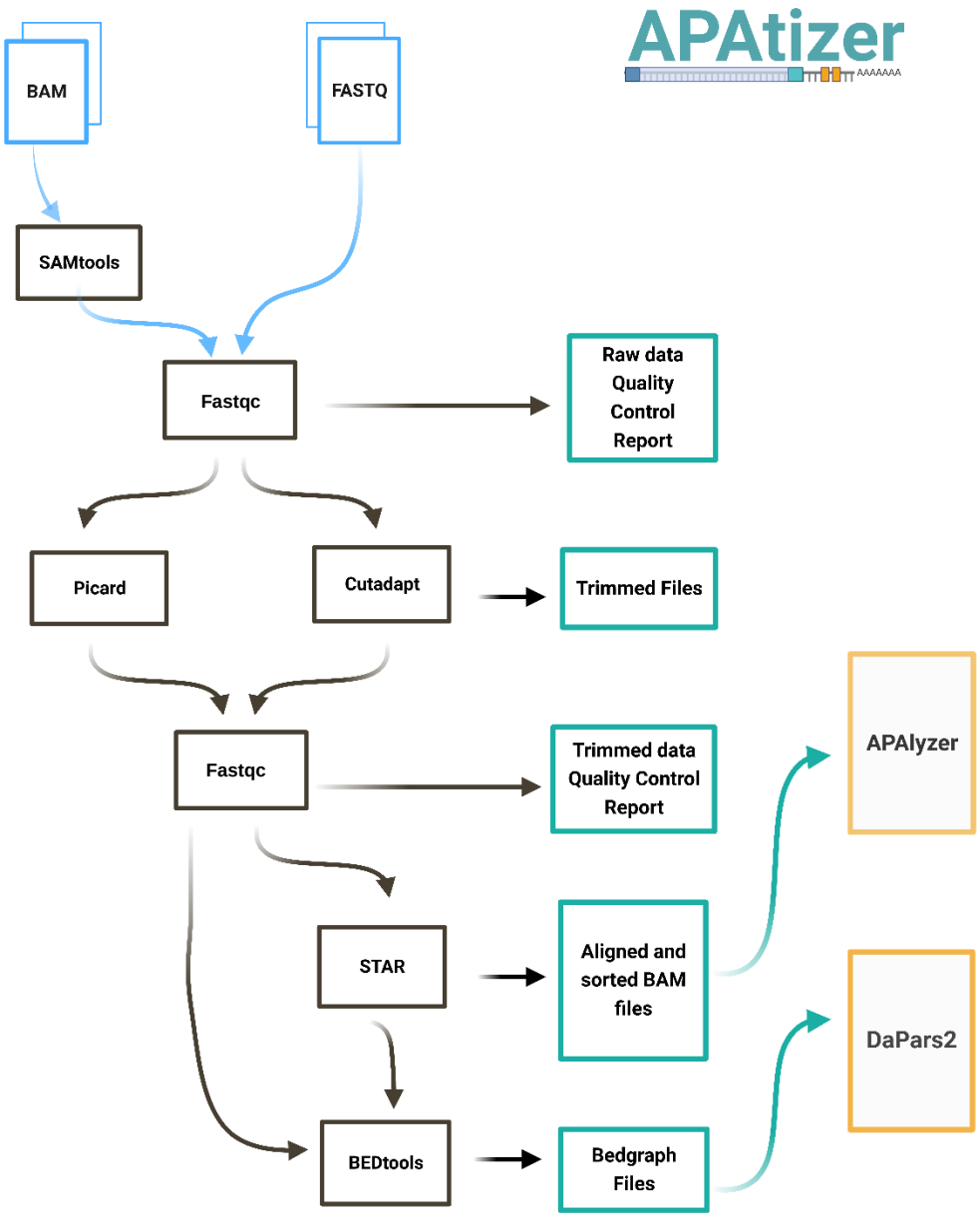


Figure 12 - APAtizer: the automated pipeline diagram. BAM and FASTQ files are used as input to analyse APA events recurring to existent tools.

3.5.3 - Dependences

For each rule, it is possible to define a software environment that will be automatically deployed via the *conda* package manager environment described by a *YAML* file. The *YAML* file is used by *conda* to install constituent software defined in the environment takes precedence over the same software in the operating system, is isolated and independent from the same software in other *Conda* environments.

3.5.4 - Illustrative examples of analysis with APA

To test the *APAtyzer* tool we used real case samples and datasets used to study cancers triggered by tobacco smoking. We analysed 30 samples to test each type of input. Bladder dataset to test *FASTQ* PE samples, lung to test *FASTQ* SE samples and Head and Neck from TCGA data deposed on GDC data portal to analyse *BAM* data. All datasets were submitted to *Dapars2* and *APAlyzer* program.

3.6 - Code Availability

All scripts and the *APAtizer* tool developed to perform the bioinformatics analyses during this project are protected by the intellectual property of i3S and are not publicly available. The opponents can request access to the folder with the developed work, for consultation purposes, here: Bioinformatics_GR.

Chapter IV – Results and Discussion

Results and Discussion

4.1 – Pro-inflammatory polarization of primary human macrophages and co-culture with CRC induce APA events

Macrophages are myeloid cells that play a significant role in the immune response. Unpolarized macrophages (M0) can be differentiated and polarized *ex-vivo* into pro-inflammatory (M1) or anti-inflammatory (M2) macrophages.

We used QuantSeq 3' mRNA-Seq to investigate how human macrophages 3'UTR-APA and IPA expression are modified upon M1 polarization and when co-cultured with CRC cells. This method quantifies gene expression based on RNA levels at the 3' ends of genes and APA profiles for each gene.[78]

4.1.1 - M1 macrophage polarization promotes APA events that are maintained in CRC co-culture

We performed two steps to investigate mRNA 3'UTR-APA and IPA isoforms upon macrophage M1 polarization and M1 macrophage co-culture with CRC cell lines. We prepared the samples to be sequenced and, consequently, analysed the data by computational methods.

The libraries were prepared using RNA extracted from M0, M1 and M1 macrophages co-cultured with CRC HCT15 and RKO cell lines from three healthy human donors. The two cell lines used in the co-cultures with macrophages differ in important markers that affect tumorigenic pathways. We resort to these two cell lines to exclude bias inherent to the characteristics of each cell line regarding the effect of the CRC co-culture. The files obtained from the sequencing step were processed using bioinformatic procedures.

We developed an R script to examine the 3′UTR-APA and IPA changes using APAlyzer. We compared APA isoforms in distinct mRNA isoforms between two populations: M1 vs M0 and M1 with CRC co-culture vs M1. The analysis of M1 with CRC co-culture was performed for both HCT15 and RKO cell lines. The differences between samples were calculated using t-test and p-value < 0.05. Considering a significant relative expression difference (RED) = 5% and *p*-values < 0.05 we are able to determine shortening (UP) and lengthening (DN) events for 3'UTR-APA and Upregulated (UP) or Downregulated (DN) for IPA.

As it can be observed in figure 13, our results show an upregulation of mRNA isoforms with short 3'UTR upon macrophage polarization. Hence, during the polarization from M0 to M1, there is a preference for genes to select proximal PAS by 3'UTR-APA leading to shortening of the 3'UTR. Moreover, we can also observe an increase in intronic Polyadenylation (IPA) during the polarization. Comparing these results to the macrophage co-cultured with cancer cells, we observe that APA events that occur during the polarization of macrophages are maintained in the co-culture for both cell lines (Figure 13).

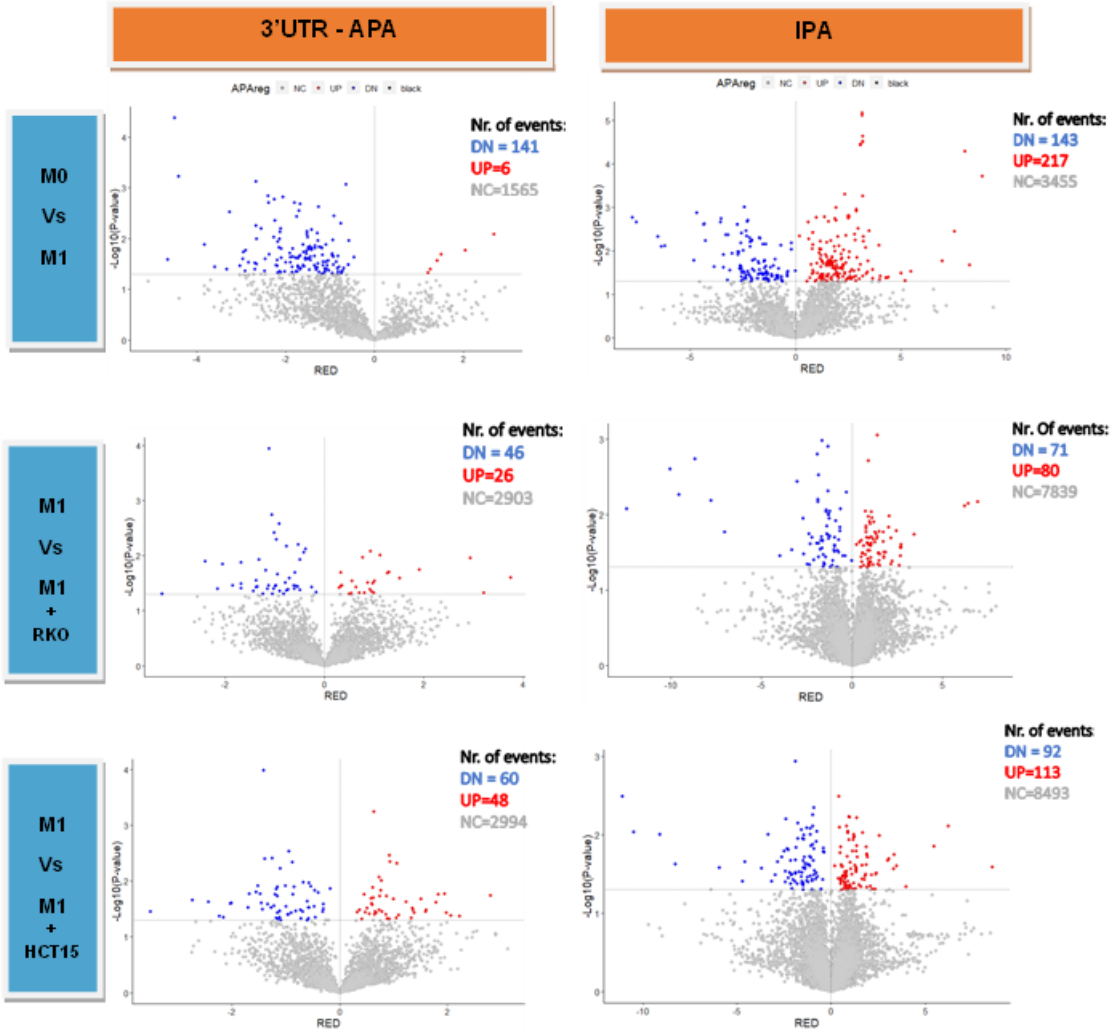


Figure 13 – Volcano plots with 3'UTR-APA and IPA from APAlyzer analysis of M0 versus-M1 and M1 macrophages versus M1 co-cultured with CRC cancer cell lines. In the 3'UTR-APA events red dots indicate 3'UTR-APA lengthening and blue dots 3'UTR-shortening events. For IPA events red dots indicates genes that preformed less IPA events and in red genes that preformed more IPA events. The grey dots are the non-significant gene events (less than 5% of difference between two sample groups).

Finally, we investigated several genes on IGV dynamic tool to the visual exploration of our results (Figure 14). In Figure 14a) we present PBX*3* gene,

associated with inflammation during sepsis and targeted by the tumour metastasis suppressor *let-7c* in the presence of CRC [79]. This gene is an example of 3'UTR-APA shortening when M1 macrophages are co-cultured with CRC cell lines. Figure 14b) is an illustrative example of a gene with upregulated IPA when M1 is co-cultured with CRC cell-lines. MAPK8 is a gene widely expressed in immune cells and tumours[80].

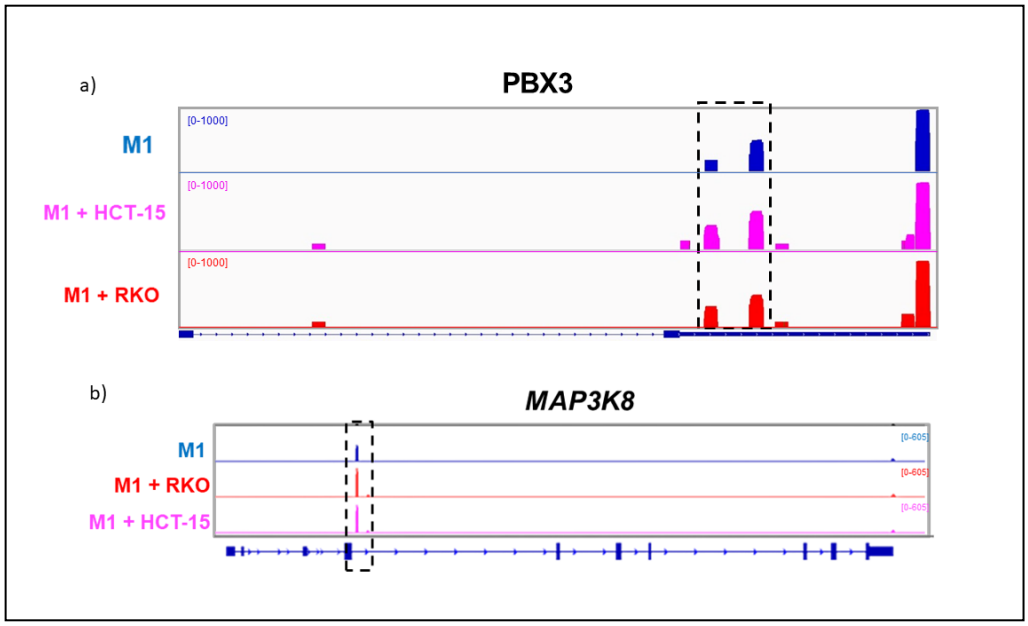


Figure 14- Dynamic exploration of 3'UTR-APA events comparing M1 macrophages (blue) M1 co-cultured with RKO cell line (Red) and with HCT15 (Pink). a) Exploration of PBX3 gene to illustrate a 3-UTR-APA shortening event on immunodeficient mice comparing to immunocompetent mice. b) Exploration of MAP3k8 gene to illustrate an upregulated IPA event on M1 when co-cultured with CRC cell lines.

4.1.2 – Discussion

In this study, resorting to computational analysis, we unveiled new insights about the transcriptome of pro-inflammatory polarized macrophages and understand how the expression of macrophages is altered by the presence of CRC.

Our RNA-seq data analysis revealed 28742 APA events in primary human macrophages after polarization and CRC co-culture. After the polarization we observed 141 3'UTR-APA shortening events, six 3'UTR-APA lengthening events, 143 downregulated IPA events and 217 upregulated IPA events. When M1 macrophages are co-cultured with RKO cell line we observed 46 3'UTR shortening events, 26 3'UTR-APA lengthening events, 71 downregulated IPA events and 80 upregulated IPA events.

Finally, when M1 macrophages are co-cultured with HCT15 cell line we observed 60 3'UTR shortening events, 48 3'UTR-APA lengthening events, 92 downregulated IPA events and 13 upregulated IPA events.

These results show a correlation between APA and macrophage inflammatory functions and increase the spectrum of known alterations in immune cells after pro-inflammatory polarization and cancer cell exposure.

Although there are many studies on APA about the relationship between immune system cells and cancer [27][35][75][76], there are no studies about the effects of cancer on macrophages. One of the reasons there are no studies on macrophages is the challenge of handling macrophages, due to their plasticity. Therefore, to our knowledge, this is the first time a computational study of the relationship between APA modulation, in primary human macrophages, in the presence of cancer has been performed. The findings clarify the crucial role of macrophage pro-inflammatory polarization in gene expression and APA and identify lists of genes that may be used as new biomarkers for CRC diagnosis in the future.

4.2 – The effect of the immune system on CRC APA events.

One of the main problems in cancer is the immune system evasion. However, the role of the immune system in CRC gene expression and APA is still incognito. We analysed the RNA-seq data and performed differential gene expression, 3'UTR-APA and IPA analysis on CRC tumors developed in immunodeficient mice and compared with immunocompetent mice.

4.2.1- Gene expression changes in immunodeficient mice with CRC

Gene expression analysis:

We performed a differential gene expression analysis. We used htseq to count RNA-seq reads and DESeq2 to analyse the BAM files data.

Our results, presented on Figure 16, show an enhanced volcano plot that allows the readability of gene expression analysis. We pre-filtered genes with low counts and kept genes that had at least 10 reads total[67] while only considering genes with a FDR p-adjusted value <0.05 and Log2Fold changes ≥ 2 or ≤ -2.

Comparing ID mice with IC mice, there are 19,416 analysed genes where 308 genes are significantly downregulated (blue dots), 305 genes are significantly upregulated (red dots) and the remaining genes are non-significant (grey dots) (Figure 15). The values for each event type are represented on the Venn-diagram in Figure 16.

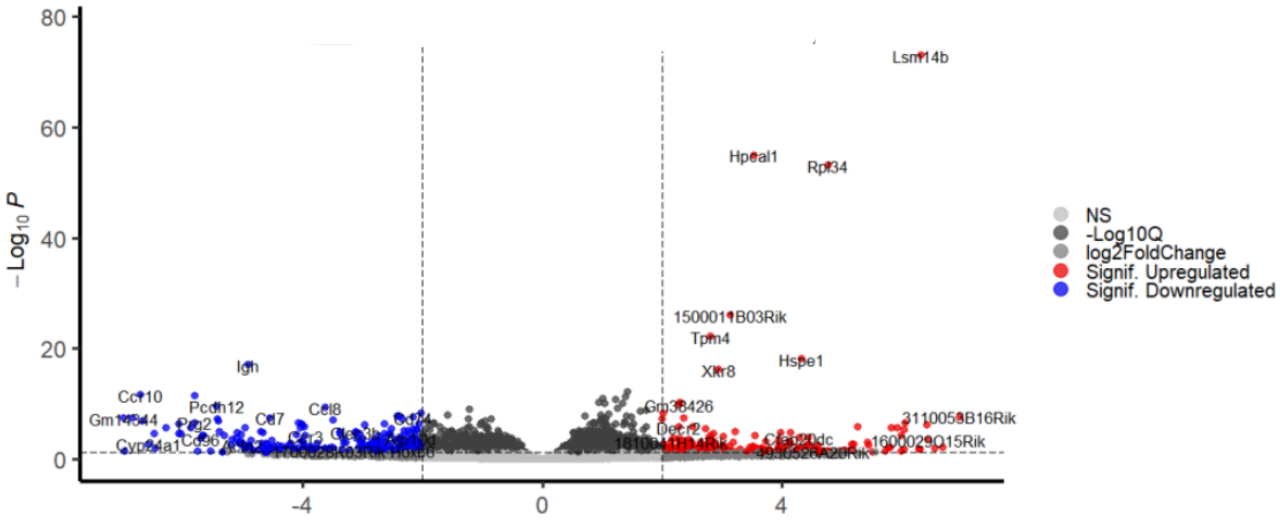


Figure 15– Plot with the comparison between Immunocompetent and immunodeficient mice. Vertical lines represent the fold change ≥ 2 or ≤ -2. The horizontal line represents the p-adjusted value = 0.05. Grey dots are non-significant values. Red dots are the upregulated significant dots and blue dots the downregulated significant values.

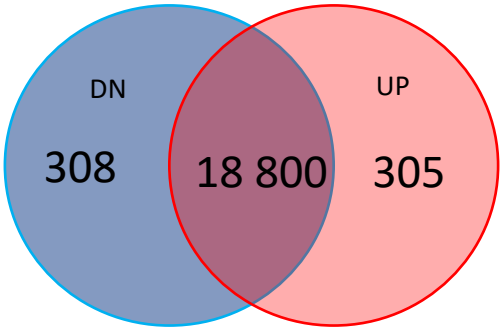


Figure 16-– Number of genes that are upregulated (UP) and downregulated (DN) when comparing the Immunocompetent mice vs Immunodeficient mice.

Next, to understand which genes were differentially expressed between IC mice and ID mice, we built two heatmaps combined with clustering methods to compare CRC differential gene expression in ID mice and IC mice for the top 50 upregulated (Figure 17a)) and downregulated genes (Figure 17b)). downregulated genes (a) and upregulated genes (b).

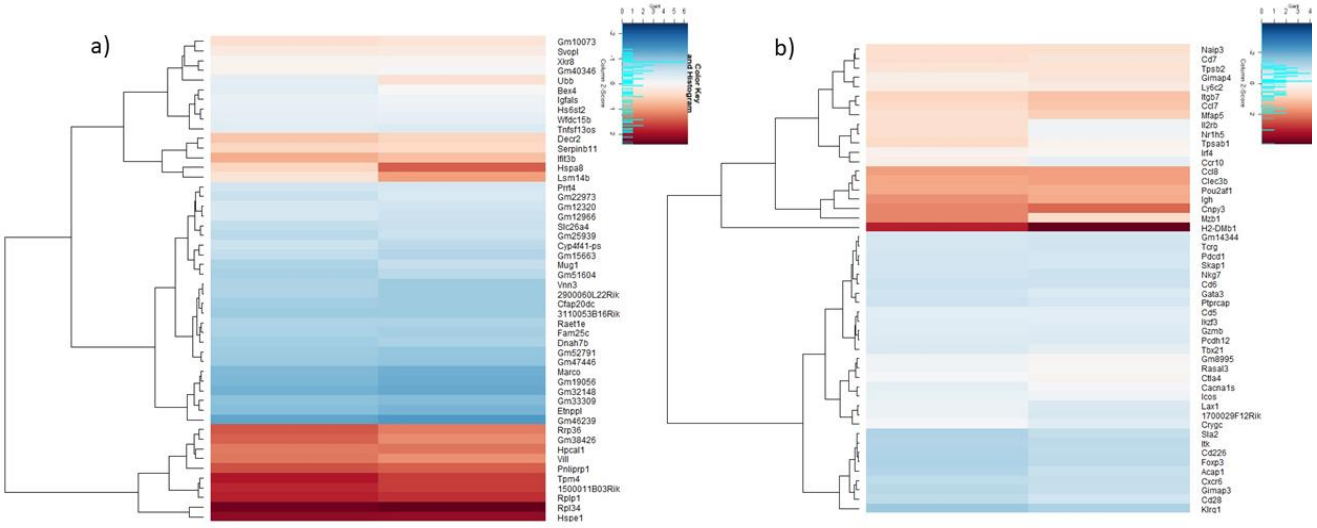


Figure 17- Heatmap with clustering methods to compare immunocompetent mice and immunodeficient mice gene expression. a) The significant upregulated genes. b) The significant downregulated genes. Gene clusters illustrate the major expression patterns and are ordered by fold change and normalized by Z-score.

When comparing ID and IC mice, we observed that different immune backgrounds promote alterations in gene expression. We are currently choosing which genes are better candidates to study in wet lab experiments.

Finally, to comprehend the functions of genes we performed a GO enrichment analysis (PANTHER) for Biological and Molecular functions for the top 10 upregulated and downregulated genes. (Figure 18).

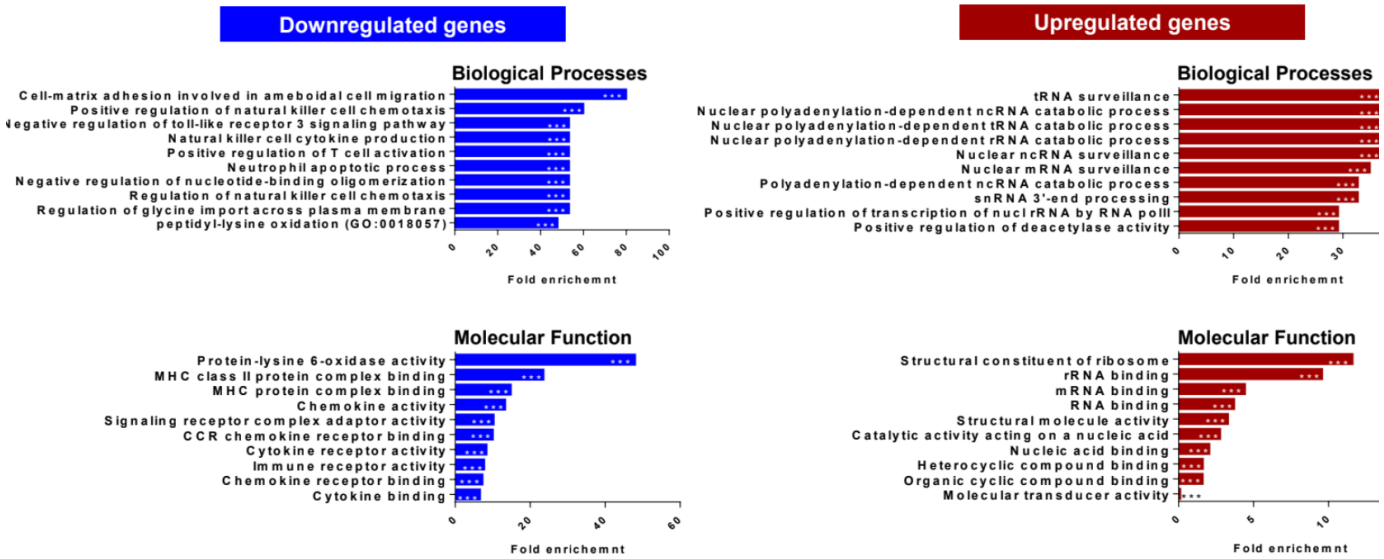


*Figure 18- Top 10 Gene Ontology (GO) terms for Upregulated genes (red) and Downregulated genes (Blue). GO terms selection was preformed based on fold enrichment and p-value < 0.05 using Fisher's test and *** = p<0.0001.*

4.2.2- Mice without immune system show an increase of 3'UTR-APA shortening and IPA events in CRC

Alternative Polyadenylation Analysis

To study 3'UTR-APA and IPA events in intestinal tumours proliferating under an immunocompetent environment and an immunodeficient environment, we developed an R script using APAlyzer to perform the analysis of 3'RNA-seq data. The APAlyzer results consist in volcano plots and tables where is possible to observe the genes that preform significant 3'UTR-APA and IPA alterations.

Our volcano plot, presented on Figure 19, show a prevalence of 3'UTR-APA shortening with 194 events and IPA upregulation with 281 events, comparing immunodeficient mice and immunocompetent mice with CRC (3'UTR-APA DN corresponds to shortening, UP to lengthening; IPA UP corresponds to increased IPA, DN decreased IPA).

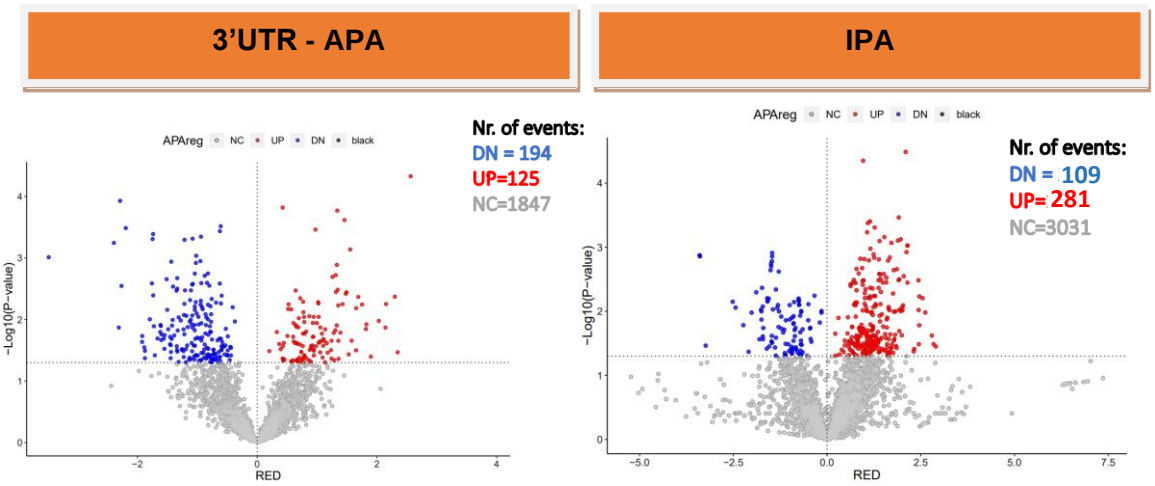


Figure 19- Volcano plots with 3'UTR-APA and IPA APAlyzer results comparing immunodeficient and immunocompetent mice with CRC. In the 3'UTR-APA events red dots indicate 3'UTR-APA lengthening and blue dots 3'UTR-shorteing events. For IPA events blue dots indicate genes that preformed less IPA events and in red genes that preformed more IPA events. The grey dots are the non-significant gene events (less than 5% of difference between two sample groups.)

After observing the volcano plots, the next step was to understand which genes had major alterations in 3'UTR-APA and IPA and comprehend their role. To this end, we ranked the top five genes with 3'UTR-APA and IPA differences (table 3 and 4).

Table 3 – Top 5 most significant genes (p-val < 0.05) for 3'UTR-APA shortening and lengthening events.

3'UTR-APA Shortening Events	3'UTR-APA Lengthening Events
Spred1	*Cnpy4*
Tcf4	*Vcp-kmt*
Rpl31	*Hip1r*
Zfand5	*Zfp106*
Zfp91	*Hmgn1*

Table 4- Top 5 most significant genes (p-val < 0.05) for Down and Upregulated IPA

IPA Downregulated Events	IPA Upregulated Events
March2	*Cul9*
Vangl1	*Nol8*
Poc1b	*Pigs*
Dtd2	*Steap2*
Cxcl12	*Pcbd2*

In 3'UTR-APA lengthening we observe genes such as Spred2 with a development and progression role in colorectal cancer [81] and TCF4 [82] that enhances hepatic metastasis in colorectal cancer enhancing the polarization of anti-inflammatory macrophages (M2). We obtained shorter 3'UTR-APA genes as CNPY4 that works as signalling regulator. When this gene expression is upregulated can inhibit the activity of CRC, promote proliferation and angiogenesis [83] VCP-KMT is a tumour metastasis-promoting gene and is strongly expressed at the protein level in several cancers.

Several genes were investigated on IGV dynamic tool to the visual exploration of our data (Figure 20). The VCP-KMT gene shows a 3'UTR-APA lengthening example (Figure 20a)). It is possible to observe an increase of immunodeficient mice (pink) in the longer isoform and shift of behaviour between immunocompetent mice (in blue) and immunodeficient mice. The Ecil gene was used as example of a shortening 3'UTR-APA example. In immunodeficient mice we are able to observe a prevalence of reads that preferred a shorter PAS (Figure 20b)).

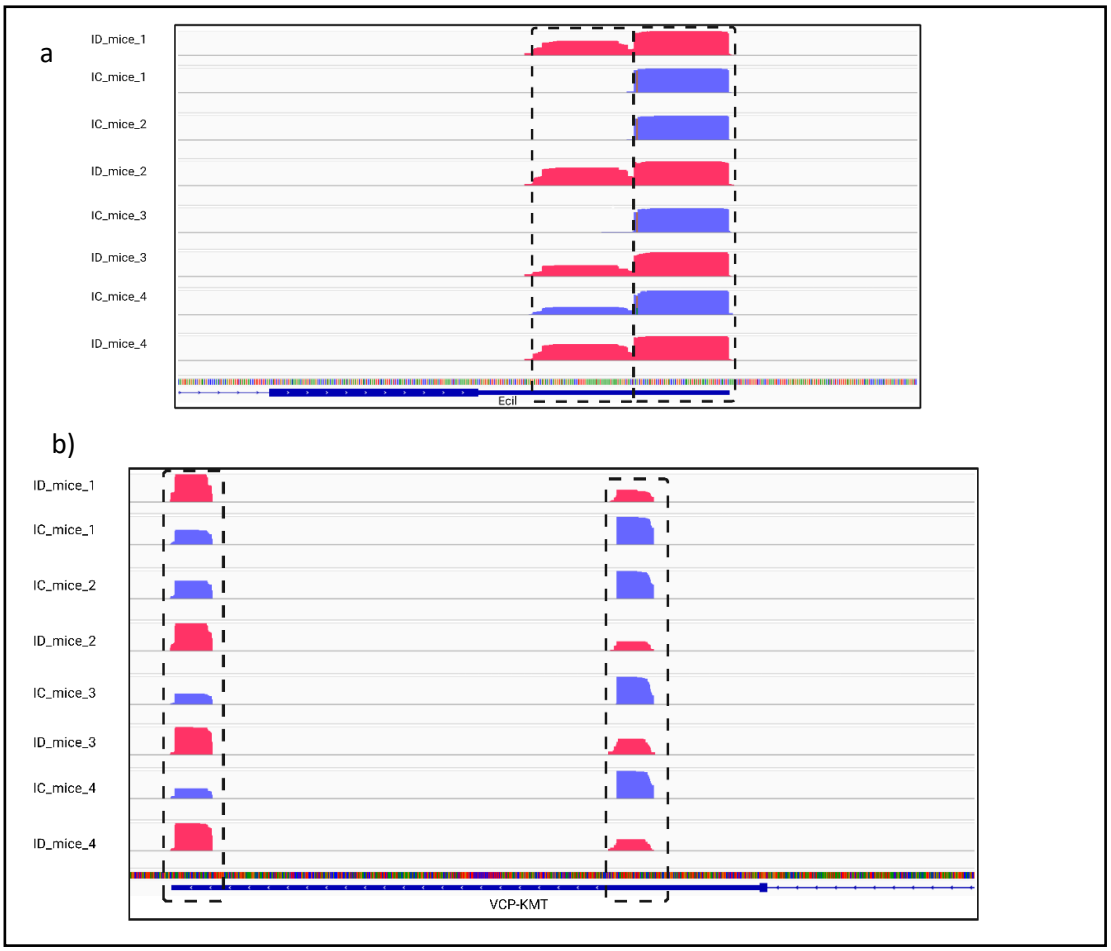


Figure 20- Dynamic exploration of 3'UTR-APA events in the immunocompetent (blue) and immunodeficient (pink) mice. a) Exploration of VCP-KMT gene to illustrate a 3-UTR-APA lengthening event on immunodeficient mice comparing to immunocompetent mice. b) Exploration of Ecil gene to illustrate a 3-UTR-APA shortening event on immunodeficient mice comparing to immunocompetent mice.

When looking to IPA Downregulated events (Table 4), MARCH2 from membrane bound E3 ubiquitin ligases family, duces surface accumulation of several glycoproteins. The knockout of this gene suppresses cell proliferation and promotes autophagy. Cul9 is a Upregulated gene and has a tumour suppressor and promotes apoptosis role.

The genes were also investigated on IGV dynamic tool (Figure 21). The Ppwd1 gene shows an example of an IPA upregulated event (Figure 21a). It is possible to observe an increase of immunodeficient mice (pink) that does not exist on immunocompetent mice (in blue). The Vangl1 gene was used as example of an IPA downregulated example. In immunodeficient mice there are reads in an intronic PAS that is not expressed in immunodeficient mice (Figure 21b).

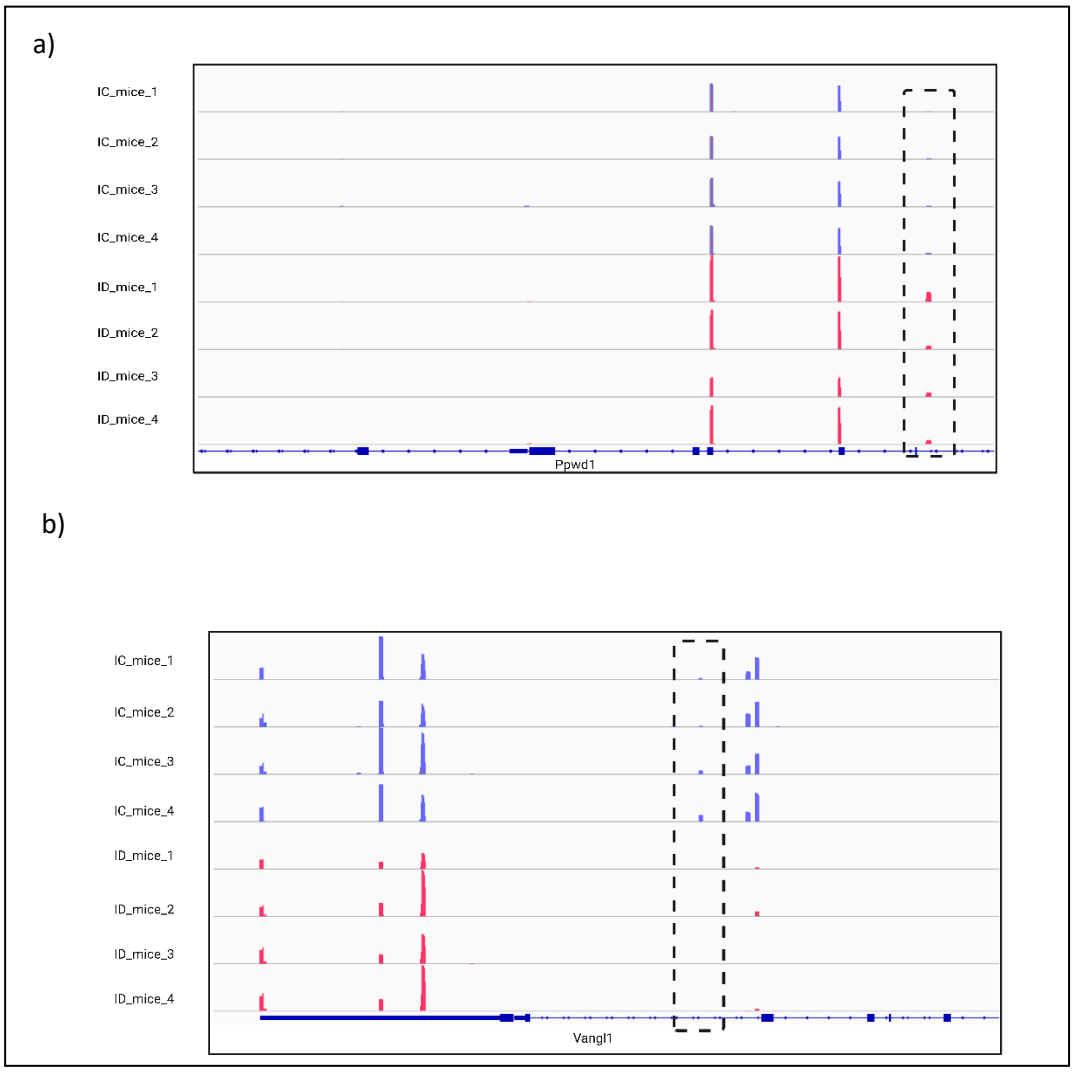


Figure 21- Dynamic exploration of IPA events in the immunocompetent (blue) and immunodeficient (pink) mice. a) Exploration of Ppwd1 gene to illustrate an IPA Upregulated event on immunodeficient mice comparing to immunocompetent mice. b) Exploration of Vangl1 gene to illustrate a IPA downregulated event on immunodeficient mice comparing to immunocompetent mice.

We then analysed the GO terms for each list of genes obtained by APAlyzer (Downregulated IPA, Upregulated IPA, shortened 3'UTR-APA, and lengthened 3'UTR-APA). The genes that were 3'UTR-APA lengthened and genes with downregulation of IPA presented no statistically significant results. The other subsets are presented in Figure 22.

When the GO terms are considered, despite that the GO terms are related to nuclear processes, and cell cycle, the most notorious is "negative regulation of translational initiation ". The fact that these GOs appear in a tumorous environment shows that the immune deficiency is an essential factor to combat the tumour and it affects APA events within the tumour (Figure 22).

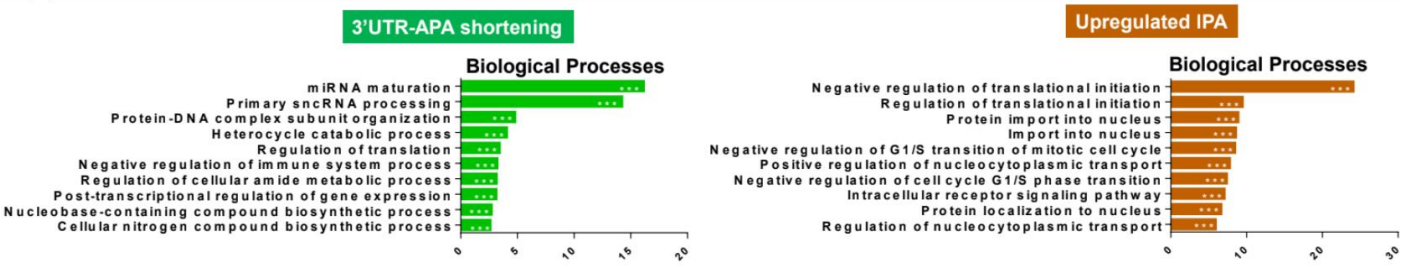


*Figure 22 - Top 10 Gene Ontology (GO) terms for 3'UTR-APA shortening genes (green) and Upregulated IPA genes (Brown). GO terms selection was preformed based on fold enrichment and p-value < 0.05 using Fisher's test and *** = p<0.0001.*

Finally, to identify if there are any correlation between APA events and differential gene expression, we performed a two-tailed Pearson correlation for 3'UTR-APA (Figure 23a)) and IPA (Figure 23b)). No correlation was observed for 3'UTR-APA and DGE but a negative correlation was observed between IPA and DGE. Therefore, upregulated genes have less IPA events and downregulated genes undergoes upregulation in IPA events.

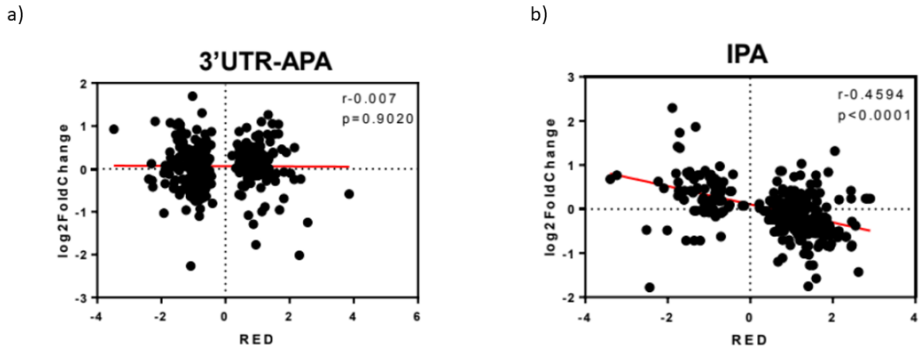


Figure 23 – Correlation between APA events and differential gene expression) comparing immunodeficient mice and immunocompetent mice. Log2Fold change was plotted against relative expression differences (RED) with a two-tailed Pearson correlation for 3'UTR-APA (a) and IPA (b.

4.2.3 – Discussion

Co-transcriptional events such as alternative polyadenylation in the 3'-UTR-APA and IPA, regulate gene expression and generate multiple mRNA isoforms. Immune cells in the tumour microenvironment can either promote or inhibit cancer progression in CRC. Resorting to computational analysis, we unveiled new insights about the immune system control over the CRC transcriptome via 3'UTR-APA and IPA.

Comparing ID mice with IC mice we have found that the immune system change CRC DGE in mice with upregulated genes are associated with polyadenylation and transcription, while mice associated with downregulated genes are associated with immune system. Recent studies suggest that 3'UTR-APA is not correlated with alterations in DGE [84], [85] and upregulation of IPA is correlated with a DGE downregulation [85], [86]. When we tried to find a correlation between APA events and DGE we found a negative correlation in IPA and DGE. Only recently, there is an increasing focus on understanding the correlation between IPA and gene expression. Our observations highlight IPA as an important mechanism with impact in gene expression.

Our APA analysis shows an increasing of 3'UTR-APA shortening events and an upregulation of IPA. Those events are associated with gene expression control, cellular metabolism, and cell cycle. Our findings are product of the first study that displays the relation between immune system, cancer, and APA. These results clarify the differences in CRC gene expression and APA after the immune system depletion. We are currently selecting the most important genes to analyse with more detail.

4.3 – APA signatures in cancers triggered by tobacco smoking

Tobacco smoking habits are one of the top causes of cancer and death worldwide, responsible for more than 15 cancer types and eight million deaths every year [87]. On the other hand, it is the most significant preventable cause of cancer.
It is widely known that cigarette-smoking habits induce alterations at genomic, transcriptomic, and systemic inflammatory response levels [88]. Among all cancers caused by tobacco, bladder, lung and head and neck cancers, are types of cancer that are triggered by smoking habits. However, the effects of smoking habits on global APA mechanisms have not been investigated so far. Hence, we aimed to find possible APA signatures that could be used as biomarkers to fight cancer resistance and to identify possible oncologic lung, bladder, and head and neck patients.

To investigate the 3'UTR-APA events in tobacco cancer we started searching for online and available databases. We have found two online and publicly available databases; TC3A, APAatlas, and two RNA-seq studies PRJN (bladder) and Bio (lung).

4.3.1 – 3'UTR-APA events TCGA and GTex Databases

Data from TC3A (tumour samples) and APAatlas (normal samples) contained RNA-seq data from TCGA and GTEX, respectively, analysed by DaPars. The DaPars algorithm compares normal and tumour pairs of samples using ΔPDUI. However, it is often not possible to use pairs of normal and tumour samples, as normal samples are not available. When this happens, it is common to compare the mean of PDUI normal tissue samples with the mean of PDUI tumour tissues. Due to the lack of normal head and neck samples on the APAatlas database, the analysis proceeded without head and neck data. Thus, we developed an R script to calculate and analyse 3'UTR-APA events for Bladder and Lung cancers. Then we calculated the mean for each APA event and the ΔPDUI values. Hence, only genes found in both APAatlas and TC3A databases were analysed. We first applied the standard threshold, considering shortening if ΔPDUI>-0.2 and lengthening if ΔPDUI>0.2. We repeated the process for ±0.3 thresholds and ±0.4.

Finally, we compared common shortening and lengthening events for bladder and lung tissues. The GTEx/TCGA analysis results shows that 99 genes undergo shortening 3'UTR-APA events and that 36 genes undergo lengthening events for both bladder and lung tissues samples (Figure 24).

4.3.2 – RNA-seq Workflow analysis

The previous analysis required an approximation of ΔPDUI values; once, during the research, we did not have access to tumoral/normal pairs of samples. Hence, to confirm these results, we found two independent studies; one with bladder data (PRJNA356345) [75] and another with lung data (GSE174330) [76] with tumoral/normal pairs of samples that used DaPars to analyse 3'UTR-APA events. First, we analysed the ΔPDUI data from both studies and tried to find common genes between those two studies and our GTEx/TCGA analysis. No common genes were found. We hypothesised that the lack of results could be due to using different methods before DaPars. Hence, we decided to reanalyse the RNA-seq data from scratch, building an RNA-seq workflow.

Of the initial 98 tumour/normal pairs of lung samples, 29 pairs were excluded from the analysis due to the poor quality of the reads. Lastly, we made a similar R script from the one developed in the previous analysis. However, instead of calculating the means for tumour and normal samples and then calculating the ΔPDUI, we could calculate the ΔPDUI for each sample and the mean of ΔPDUIs per each event.

Our results show that despite the number of lung samples, the results from DaPars did not intercept any other dataset. On the other hand, we observed a slight improvement in the Bladder dataset and an interception of genes. From the 920 events found with 3'UTR shortening, four intersect the TCGA/GTEx analysis, and from the 435 events with lengthening 3'UTR-APA, seven events intersect the Bladder dataset (PRJNA356345) with TCGA/GTEx analysis (Figure 25).

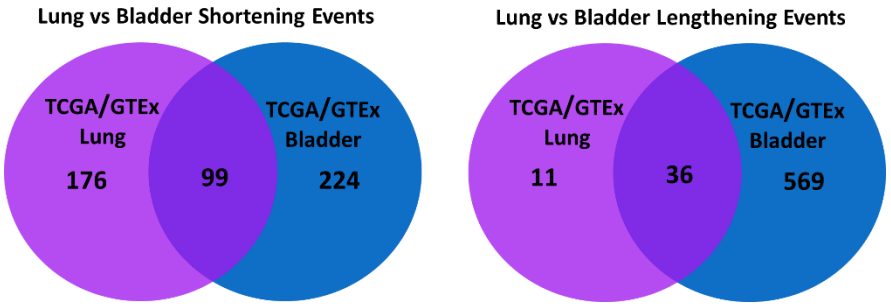


Figure 24- Venn Diagram with common genes for Lung and Bladder cancers from TCGA and GTEX data for 3'UTR-APA shortening and lengthening.

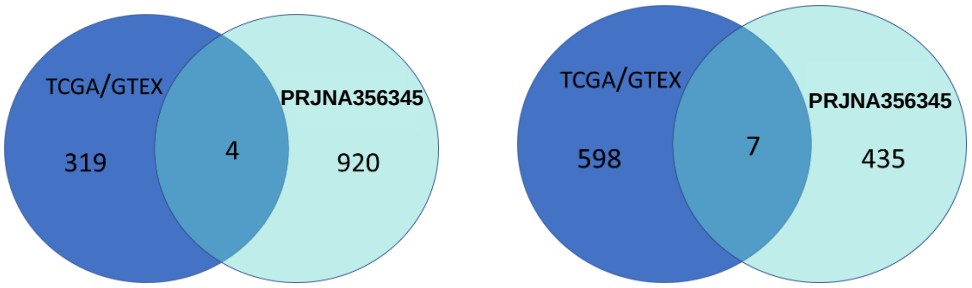


Figure 25- Venn Diagram with common genes for Bladder cancers from TCGA GTEX data and PRJNA356345 data for 3'UTR-APA shortening and lengthening.

4.3.3 – Discussion

It is widely known that APA is a crucial mechanism in gene expression regulation [89]–[91]. Our findings open new doors to target APA signatures therapeutically in the future.

The available databases allow us to verify our hypothesis with the existing RNA-seq data, reducing analysis time and costs. However, the available databases remain incomplete, and due to the variability of bioinformatic tools, there is no consistency in the methods used to process and analyse RNA-seq data.

Here we conducted an exploratory analysis of two cancer types triggered by tobacco smoking: lung and bladder cancers. According to TC3A and APAatlas, we

have found 135 3'UTR-APA events (99 3'UTR-APA shortening and 36 3'UTR-APA events) in common between bladder and lung cancer (Figure 20). However, according to the bibliography and the DaPars ΔPDUI calculation method, these results are an approximation. To increase the robustness of these results, we analysed raw RNA-seq data using Dapars2 to compare with the previous results. Although several samples in the lung dataset presented low-quality reads, we observed an improvement in the bladder analysis.

To increase the robustness of our analyses, we now aim to perform a meta-analysis employing a more diversified collection of high-quality datasets containing both normal and tumoral RNA-seq. Hence, after requiring access to data from dbGAP, we are currently analysing TCGA raw data with normal and tumoral tissues pairs from the same patient from bladder, lung and head and neck samples.

4.4- Evaluation of APAtizer Workflow

Analysing RNA-seq data can be very laborious. The steps between processes can be very time-consuming, such as waiting for a step to finish and preparing for the next step. To make our lives easier, we developed the APAtizer: a comprehensive bioinformatics tool that provides RNA-seq analysis workflow with state-of-the-art tools and analyses for APA events. The workflow was developed in Snakemake, a widely used workflow management system in the bioinformatics community. This tool provides the analysis considering three types of data as input: FASTQ single-end and paired-end data and BAM files.

For FASTQ files, we have a similar workflow. We used Cutadapt to perform Illumina data trimming and STAR to obtain aligned and sorted data. The pipeline is prepared to receive paired-end and single-end data, performing the analysis according to its specificities.

Lastly, we wanted to analyse the samples using Dapars2 or the APAlyzer program to get more robust results considering that the algorithms have different performances.

We tested our pipeline with five pairs of normal/tumour tissue samples for each input type. For the BAM input we selected random head and neck samples from TCGA database. The only criteria we considered was to choose donors that have smoking habits. For Single-end and Paired-end samples we used previous analysed datasets for bladder data (PRJNA356345) and another with lung data (GSE174330) with tumoral/normal pairs of samples.

In table 7, we observe the evaluation of the performance of APAtizer. At first glance, we might think that the input with BAM is the slowest. However, on average, when have a bam file as input the analysis takes 1 h and 40 min for each sample in BAM format, 36 min for each sample in FASTQ SE format and 24 min for each sample in FASTQ PE format. On the other hand, in this experiment, a sample in BAM format was equivalent to about three samples of FASTQ SE and 6.5 samples of FASTQ PE.

Thus, if all samples had the same size, on average, run times of 1h40m for each sample in BAM format, 1h48m for each sample in FASTQ SE format and 2h36m for each sample in FASTQ PE format.

Table 5 - – APAtizer performance evaluation based on runtime and file size of each input

Input Type	Mean sample size	Time to finish the process
BAM	16 GB	~14h
FASTQ Single End	5.5 GB	~6h
FASTQ Paired End	2.5 GB	~4h

These temporal changes are due to the different number of steps needed to analyse each of the samples since BAM files have fewer steps to be performed than files in FASTQ paired-end format. Since the alignment is a time-consuming step and has already been performed in the case of samples in BAM format, it is also natural that this has a faster process than when the input is a FASTQ file.

The limitations of this tool are always based on the memory available to store the data, and it must have at least 27 GB of RAM to run the star aligner.

4.4.1 – APAtizer Results

To our knowledge, APAtizer is the first tool that provides an automated analysis of RNA-seq data specified in 3'UTR-APA and IPA analysis. The APAtizer was used as a control test, FASTQ samples previously analysed in chapter V, section 5.3 - APA signatures in cancers triggered by tobacco smoking as control samples, to ensure consistent results with the previously obtained. The BAM files used to test APAtizer were retrieved from the TCGA database. For the first time, we are analysing the TCGA raw data. We can now provide the first insights into APA signatures in tobacco-related cancers.

Figure 26 represents the DaPars2 results after the APAtizer analysis. The Venn diagrams presenting the 3'UTR-APA events for the intersection of Lung, Bladder and Head&Neck tissues considering APA were lengthening when ΔPDUI > 0.2 and APA shortening when ΔPDUI< -0.2.

The preliminary analyses of 3'UTR-APA shortening events show 51 for Head&Neck Cancer but no intersection with the Lung and Bladder events. The intersection of Lung and Head&Neck presents one gene in common named ACTA1. The product encoded by this gene belongs to the actin family proteins and plays an important role in regulating and progression of Head and Neck cancer [83]. The intersection with Bladder and Head&Neck presents one gene in common named MAVS. This gene encodes mitochondria antiviral signalling proteins. Studies have shown that MAVS that MAVS induce innate immune responses in tumour cells, and mutations in this gene promotes therapy resistance [84].

The 3'UTR-APA Lengthening of Head and Neck cancer show 11 genes, where one is common for the three cancer types. COL1A2 encodes the pro-alpha2 chain of collagen type II [85]. This gene has shown to be upregulated in several cancers such as gastric, hypopharyngeal squamous cell carcinoma, bladder, and lung cancer [85] [86].

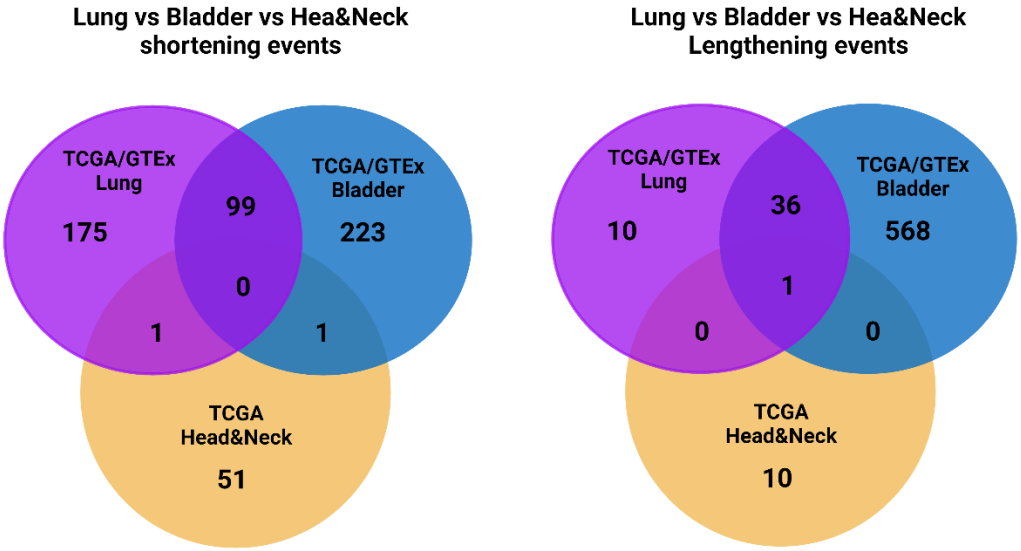


Figure 26 - Venn Diagram after the APAtizer analysis with DaPar2 program with common genes for Bladder cancers from TCGA GTEX and Head&Neck TCGA data.

More studies are needed to verify the results presented on figure 26. The APatyzer tool will be fundamental to analyse the normal and tumour samples from TCGA database for Lung Bladder and Head&Neck tissues.

Overall, the APAtyzer tool makes it easy for any bioinformatician to do their work without interruptions, reducing hands-on time for bioinformaticians and giving experimentalists the possibility to access their data more rapidly.

Chapter V - Conclusion

Conclusion

5.1 Main contribution

Great attention has been put forward into cancer disease and its relationship with the immune system. Our research aimed in identifying new patterns in APA events in immune cells triggered by cancer exposure, in cancer cells developed in immunodeficient mice, with the final aim of better understanding the relation between immune system and cancer.

Our approach was to develop a workflow and analyse RNA-Seq data to find new genes with APA signatures to be tested, aiming to find new therapies for cancer and cancer resistance in the future. We found that the M0-M1 polarization promotes 3'UTR-APA shortening events and more IPA, new results in the field. Moreover, we have found that this pattern is maintained when M1 macrophages are co-cultured with CRC RKO and HCT15 cell lines. Furthermore, we compared CRC cancer in immunocompetent and immunodeficient mice and found that immunodeficient mice genes undergo more 3'UTR-APA shortening events and more IPA events. It is important to refer that this is, to our knowledge, the first time that these analyses were preformed via computational methods. Hence, this work unveiled new scientific knowledge in the mRNA and APA field.

We are currently looking for 3'UTR-APA patterns in cancers associated with tobacco smoking. As we are dealing with many samples after getting access to TCGA data files, and to increase the analysis efficiency, our approach was to create a comprehensive bioinformatic tool, APAtizer. The APAtizer is a RNA-seq automated workflow for APA analysis and, hence, our tool will facilitate further analysis in the mRNA and APA field. This bioinformatic tool developed on this project has the advantage to reduce time and cost to the wet lab. The best gene candidates uncovered by the bioinformatic analysis will be tested and validated by wet lab studies in the group in a collaborative work on-going with IPO.

5.2 Perspectives and future work

The on-going and future steps of the projects developed during this thesis are centred on two major analyses. First, we are currently selecting the candidate genes to continue to study the effect of the immune system on CRC APA events in mice.

In addition, we are still analysing the APA signatures in cancers triggered by tobacco smoking using TCGA database. The analysis will be performed supported by our developed tool, the APAtizer.

At this stage, APAtizer Version 0.2, constitutes a working prototype. In the future we aim to develop and integrate the following features:

- Create a user-friendly interface that does not require prior bioinformatic; knowledge to use the software and add a front-end system for visually displaying the information;
- Include gene expression analysis;
- Support more species genomes to expand the analysis for more organisms.

References

References

[1] S. Bustin, “Molecular Biology of the Cell, Sixth Edition; ISBN: 9780815344643; and Molecular Biology of the Cell, Sixth Edition, The Problems Book; ISBN 9780815344537,” *Int J Mol Sci*, vol. 16, no. 12, pp. 28123–28125, Nov. 2015, doi: 10.3390/ijms161226074.

[2] L. T. MacNeil and A. J. M. Walhout, “Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression,” *Genome Research*, vol. 21, no. 5. Cold Spring Harbor Laboratory Press, pp. 645–657, 2011. doi: 10.1101/gr.097378.109.

[3] P. Cramer, “Organization and regulation of gene transcription,” *Nature*, doi: 10.1038/s41586-019-1517-4.

[4] T. Kujirai and H. Kurumizaka, “Transcription through the nucleosome,” *Curr Opin Struct Biol*, vol. 61, pp. 42–49, Apr. 2020, doi: 10.1016/J.SBI.2019.10.007.

[5] D. L. Bentley, “Coupling mRNA processing with transcription in time and space,” *Nat Rev Genet*, vol. 15, p. 163, 2014, doi: 10.1038/nrg3662.

[6] C. L. Will and R. Lü, “Spliceosome Structure and Function”, doi: 10.1101/cshperspect.a003707.

[7] A. Ramanathan, G. B. Robb, and S. H. Chan, “mRNA capping: Biological functions and applications,” *Nucleic Acids Research*, vol. 44, no. 16. Oxford University Press, pp. 7511–7526, Sep. 19, 2016. doi: 10.1093/nar/gkw551.

[8] S. B. Patel and M. Bellini, “The assembly of a spliceosomal small nuclear ribonucleoprotein particle,” *Nucleic Acids Res*, vol. 36, no. 20, pp. 6482–6493, 2008, doi: 10.1093/nar/gkn658.

[9] J. L. Manley and D. C. di Giammartino, “mRNA Polyadenylation in Eukaryotes,” *Encyclopedia of Biological Chemistry: Second Edition*, pp. 188–193, Jan. 2013, doi: 10.1016/B978-0-12-378630-2.00624-1.

[10] E. de Klerk and P. A. C. ’t Hoen, “Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing,” *Trends in Genetics*, vol. 31, no. 3. Elsevier Ltd, pp. 128–139, Mar. 01, 2015. doi: 10.1016/j.tig.2015.01.001.

[11] I. Pereira-Castro and A. Moreira, “On the function and relevance of alternative 3’-UTRs in gene expression regulation,” *Wiley Interdisciplinary Reviews: RNA*, vol. 12, no. 5. John Wiley and Sons Inc, Sep. 01, 2021. doi: 10.1002/wrna.1653.

[12] B. Tian, Z. Pan, and Y. L. Ju, “Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing,” *Genome Res*, vol. 17, no. 2, pp. 156–165, Feb. 2007, doi: 10.1101/gr.5532707.

[13] N. J. Proudfoot, A. Furger, M. J. Dye Sir, and W. Dunn, “Review Integrating mRNA Processing with Transcription,” 2002.

[14] E. Beaudoing, S. Freier, J. R. Wyatt, J.-M. Claverie, and D. Gautheret, “Patterns of Variant Polyadenylation Signal Usage in Human Genes.” [Online]. Available: www.genome.org

[15] H. S. Yeh and J. Yong, “Alternative polyadenylation of mRNAs: 3’-untranslated region matters in gene expression,” *Molecules and Cells*, vol. 39, no. 4. Korean Society for Molecular and Cellular Biology, pp. 281–285, Apr. 01, 2016. doi: 10.14348/molcells.2016.0035.

[16] A. L. Nicholson and A. E. Pasquinelli, “Tales of Detailed Poly(A) Tails,” *Trends in Cell Biology*, vol. 29, no. 3. Elsevier Ltd, pp. 191–200, Mar. 01, 2019. doi: 10.1016/j.tcb.2018.11.002.

[17] B. Tian, J. Hu, H. Zhang, and C. S. Lutz, “A large-scale analysis of mRNA polyadenylation of human and mouse genes,” *Nucleic Acids Res*, vol. 33, no. 1, pp. 201–212, 2005, doi: 10.1093/nar/gki158.

[18] J. H. Graber, C. R. Cantor, S. C. Mohr, and T. F. Smith, “Genomic detection of new yeast pre-mRNA 3-end-processing signals,” 1999. [Online]. Available: https://academic.oup.com/nar/article/27/3/888/2902413

[19] S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr, “Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression”, doi: 10.1101/gad.229328.113.

[20] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan, “Means to an end: Mechanisms of alternative polyadenylation of messenger RNA precursors,” *Wiley Interdisciplinary Reviews: RNA*, vol. 5, no. 2. pp. 183–196, Mar. 2014. doi: 10.1002/wrna.1206.

[21] F. Ren, N. Zhang, L. Zhang, E. Miller, and J. J. Pu, “Alternative Polyadenylation: a new frontier in post transcriptional regulation,” *Biomarker Research*, vol. 8, no. 1. BioMed Central Ltd, Dec. 01, 2020. doi: 10.1186/s40364-020-00249-6.

[22] L. C. Cheng, D. Zheng, Q. Zhang, A. Guvenek, H. Cheng, and B. Tian, “Alternative 3’ UTRs play a widespread role in translation-independent mRNA association with the endoplasmic reticulum,” *Cell Rep*, vol. 36, no. 3, Jul. 2021, doi: 10.1016/j.celrep.2021.109407.

[23] J. W. Freimer, R. Krishnakumar, M. S. Cook, and R. Blelloch, “Expression of Alternative Ago2 Isoform Associated with Loss of microRNA-Driven Translational Repression in Mouse Oocytes,” *Current Biology*, vol. 28, no. 2, pp. 296-302.e3, Jan. 2018, doi: 10.1016/j.cub.2017.11.067.

[24] J. Brumbaugh *et al.*, “Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling,” *Cell*, vol. 172, no. 1–2, pp. 106-120.e21, Jan. 2018, doi: 10.1016/j.cell.2017.11.023.

[25] D. Burri and M. Zavolan, “Shortening of 3 ’ ’ ’ ’ ’ UTRs in most cell types composing tumor tissues implicates alternative polyadenylation in protein metabolism,” 2021, doi: 10.1261/rna.

[26] A. J. Gruber and M. Zavolan, “Alternative cleavage and polyadenylation in health and disease”, doi: 10.1038/s41576-019-0145-z.

[27] C. Mayr and D. P. Bartel, “Widespread Shortening of 3’UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells,” *Cell*, vol. 138, no. 4, pp. 673–684, Aug. 2009, doi: 10.1016/j.cell.2009.06.016.

[28] C. Mayr, “What are 3’ utrs doing?,” *Cold Spring Harb Perspect Biol*, vol. 11, no. 10, Oct. 2019, doi: 10.1101/cshperspect.a034728.

[29] S. H. Lee, I. Singh, S. Tisdale, O. Abdel-Wahab, C. S. Leslie, and C. Mayr, “Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia,” *Nature*, vol. 561, no. 7721, pp. 127–131, Sep. 2018, doi: 10.1038/s41586-018-0465-8.

[30] World Health Organization, “Colorectal Cancer Awareness Month,” 2022, [Online]. Available: https://www.iarc.who.int/featured-news/colorectal-cancer-awareness-month-2022/

[31] C. Mayr and D. P. Bartel, “Widespread Shortening of 3’UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells,” *Cell*, vol. 138, no. 4, pp. 673–684, Aug. 2009, doi: 10.1016/j.cell.2009.06.016.

[32] S. L. Schuster, A. C. Hsieh, and A. C. Hsieh Fred, “Trends Cancer,” vol. 5, no. 4, pp. 245–262, 2019, doi: 10.1016/j.trecan.2019.02.011.

[33] S. Kreth *et al.*, “In human glioblastomas transcript elongation by alternative polyadenylation and miRNA targeting is a potent mechanism of MGMT silencing”, doi: 10.1007/s00401-013-1081-1.

[34] I. Singh *et al.*, “Widespread intronic polyadenylation diversifies immune cell transcriptomes,” *Nat Commun*, vol. 9, no. 1, Dec. 2018, doi: 10.1038/s41467-018-04112-z.

[35] M. Asnani and A. Thomas-Tikhonenko, “Exons of Leukemia Suppressor Genes: Creative Assembly Required,” *Trends in Cancer*, vol. 4, no. 12. Cell Press, pp. 796–798, Dec. 01, 2018. doi: 10.1016/j.trecan.2018.10.005.

[36] X. Liu *et al.*, “Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in Drosophila melanogaster,” 2017, doi: 10.1261/rna.062661.

[37] R. Vallejos Baier, J. Picao-Osorio, and C. R. Alonso, “Molecular Regulation of Alternative Polyadenylation (APA) within the Drosophila Nervous System,” *J Mol Biol*, vol. 429, no. 21, pp. 3290–3300, Oct. 2017, doi: 10.1016/j.jmb.2017.03.028.

[38] S. Behjati and P. S. Tarpey, “What is next generation sequencing?”, doi: 10.1136/archdischild-2013-304340.

[39] H. Liang and E. Zeng, “RNA-Seq Experiment and Data Analysis,” in *Estrogen Receptors: Methods and Protocols*, K. M. Eyster, Ed. New York, NY: Springer New York, 2016, pp. 99–114. doi: 10.1007/978-1-4939-3127-9_9.

[40] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat Rev Genet*, vol. 10, no. 1, pp. 57–63, 2009, doi: 10.1038/nrg2484.

[41] J. T. Simpson and R. Durbin, “Efficient de novo assembly of large genomes using compressed data structures,” *Genome Res*, vol. 22, no. 3, pp. 549–556, Mar. 2012, doi: 10.1101/gr.126953.111.

[42] B. J. Haas and M. C. Zody, “Advancing RNA-Seq analysis.”

[43] T. Barrett *et al.*, “NCBI GEO: Archive for functional genomics data sets - Update,” *Nucleic Acids Res*, vol. 41, no. D1, Jan. 2013, doi: 10.1093/nar/gks1193.

[44] K. L. Howe *et al.*, “Ensembl 2021,” *Nucleic Acids Res*, vol. 49, no. 2, 2021, doi: 10.1093/nar/gkaa942.

[45] W. J. Kent *et al.*, “The Human Genome Browser at UCSC”, doi: 10.1101/gr.229102.

[46] P. Moll, M. Ante, A. Seitz, and T. Reda, “QuantSeq 3’ mRNA sequencing for RNA quantification,” 2014.

[47] N. Kandhari, C. A. Kraupner-Taylor, P. F. Harrison, D. R. Powell, and T. H. Beilharz, “Molecular Sciences The Detection and Bioinformatic Analysis of Alternative 3 UTR Isoforms as Potential Cancer Biomarkers,” 2021, doi: 10.3390/ijms22105322.

[48] F. Mignone *et al.*, “UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs”, doi: 10.1093/nar/gki021.

[49] R. Wang, R. Nambiar, D. Zheng, and B. Tian, “PolyA DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes,” *Nucleic Acids Res*, vol. 46, pp. 315–319, 2017, doi: 10.1093/nar/gkx1000.

[50] E. R. Gamazon *et al.*, “PACdb: a database for cell-based pharmacogenomics,” 2010, doi: 10.1097/FPC.0b013e328337b8d6.

[51] N. Kandhari, C. A. Kraupner-Taylor, P. F. Harrison, D. R. Powell, and T. H. Beilharz, “Review the detection and bioinformatic analysis of alternative 3’ utr isoforms as potential cancer biomarkers,” *International Journal of Molecular Sciences*, vol. 22, no. 10. MDPI, May 02, 2021. doi: 10.3390/ijms22105322.

[52] X. Feng, L. Li, E. J. Wagner, and W. Li, “TC3A: The Cancer 3’ UTR Atlas,” *Nucleic Acids Res*, vol. 46, no. D1, pp. D1027–D1030, Jan. 2018, doi: 10.1093/nar/gkx892.

[53] W. Hong *et al.*, “APAatlas: Decoding alternative polyadenylation across human tissues,” *Nucleic Acids Res*, vol. 48, no. D1, pp. D34–D39, Jan. 2020, doi: 10.1093/nar/gkz876.

[54] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge,” *Wspolczesna Onkologia*, vol. 1A. Termedia Publishing House Ltd., pp. A68–A77, 2015. doi: 10.5114/wo.2014.47136.

[55] J. Lonsdale *et al.*, “The Genotype-Tissue Expression (GTEx) project,” *Nature Genetics*, vol. 45, no. 6. pp. 580–585, Jun. 2013. doi: 10.1038/ng.2653.

[56] L. You *et al.*, “APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals,” *Nucleic Acids Res*, vol. 43, pp. 59–67, 2015, doi: 10.1093/nar/gku1076.

[57] M. D. Mailman *et al.*, "The NCBI dbGaP database of genotypes and phenotypes." [Online]. Available: www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf.

[58] K. R. Kukurba and S. B. Montgomery, "Topic Introduction RNA Sequencing and Analysis," 2015, doi: 10.1101/pdb.top084970.

[59] L. Li *et al.*, "An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability," *Nat Genet*, vol. 53, no. 7, pp. 994–1005, 2021, doi: 10.1038/s41588-021-00864-5.

[60] R. Wang and B. Tian, "APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms," *Bioinformatics*, vol. 36, no. 12, pp. 3907–3909, Jun. 2020, doi: 10.1093/bioinformatics/btaa266.

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

[62] Z. Xia *et al.*, "ARTICLE Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3 0-UTR landscape across seven tumour types," 2014, doi: 10.1038/ncomms6274.

[63] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[64] "Genomics_in_the_Cloud".

[65] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *BIOINFORMATICS APPLICATIONS NOTE*, vol. 26, no. 6, pp. 841–842, 2010, doi: 10.1093/bioinformatics/btq033.

[66] S. Anders, P. T. Pyl, and W. Huber, "HTSeq-A Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015, doi: 10.1093/bioinformatics/btu638.

[67] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol*, vol. 15, no. 12, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.

[68] Q. Hu, A. Hutson, S. Liu, M. Morgan, and Q. Liu, "Bioconductor toolchain for reproducible bioinformatics pipelines using Rcwl and RcwlPipelines," *Bioinformatics*, vol. 37, no. 19, pp. 3351–3352, Oct. 2021, doi: 10.1093/bioinformatics/btab208.

[69] J. L. Sepulveda, "Using R and Bioconductor in Clinical Genomics and Transcriptomics," *Journal of Molecular Diagnostics*, vol. 22, no. 1. Elsevier B.V., pp. 3–20, Jan. 01, 2020. doi: 10.1016/j.jmoldx.2019.08.006.

[70] J. Köster and S. Rahmann, "Snakemake-a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, 2012, doi: 10.1093/bioinformatics/bts480.

[71] van Rossum, Guido and Drake, and Fred L, *Python 3 Reference Manual*, vol. 10.5555/1593511. Scotts Valley, CA: CreateSpace, 2009.

[72] P. Moll, M. Ante, A. Seitz, and T. Reda, “QuantSeq 3’ mRNA sequencing for RNA quantification,” 2014.

[73] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, “Large-scale gene function analysis with PANTHER Classification System”, doi: 10.1038/nprot.2013.092.

[74] “Gene Ontology Consortium: going forward”, doi: 10.1093/nar/gku1179.

[75] A. Zingone *et al.*, “A comprehensive map of alternative polyadenylation in African American and European American lung cancer patients”, doi: 10.1038/s41467-021-25763-5.

[76] X. Chen *et al.*, “CSTF2-induced shortening of the RAC1 3’UTR promotes the pathogenesis of urothelial carcinoma of the bladder,” *Cancer Res*, vol. 78, no. 20, pp. 5848–5862, Oct. 2018, doi: 10.1158/0008-5472.CAN-18-0822.

[77] M. Martin, “CUTADAPT removes adapter sequences from high-throughput sequencing reads,” *EMBnet J*, vol. 17, Aug. 2011, doi: 10.14806/ej.17.1.200.

[78] J. Wilton, M. Tellier, T. Nojima, A. M. Costa, M. J. Oliveira, and A. Moreira, “Chapter Sixteen - Simultaneous studies of gene expression and alternative polyadenylation in primary human immune cells,” in *Methods in Enzymology*, vol. 655, B. Tian, Ed. Academic Press, 2021, pp. 349–399. doi: https://doi.org/10.1016/bs.mie.2021.04.004.

[79] W. O. Library *et al.*, “Let-7c functions as a metastasis suppressor by targeting MMP11 and PBX3 in colorectal cancer,” *Journal of Pathology J Pathol*, vol. 226, pp. 544–555, 2012, doi: 10.1002/path.3014.

[80] B. Tunca *et al.*, “Overexpression of CK20, MAP3K8 and EIF5A correlates with poor prognosis in early-onset colorectal cancer patients”, doi: 10.1007/s00432-013-1372-x.

[81] H. Wang *et al.*, “Spred2 inhibits epithelial-mesenchymal transition of colorectal cancer cells by impairing ERK signaling,” *Oncol Rep*, vol. 44, no. 1, pp. 174–184, Jul. 2020, doi: 10.3892/or.2020.7586.

[82] W. Tu, J. Gong, Z. Zhou, D. Tian, and Z. Wang, “ARTICLE OPEN TCF4 enhances hepatic metastasis of colorectal cancer by regulating tumor-associated macrophage via CCL2/ CCR2 signaling”, doi: 10.1038/s41419-021-04166-w.

[83] J.-W. Li, Q.-R. Huang, and L.-G. Mo, “Medicine ® CNPY4 is a potential promising prognostic-related biomarker and correlated with immune infiltrates in gliomas,” 2022, doi: 10.1097/MD.0000000000030044.

[84] S. Mitschka and C. Mayr, “Context-specific regulation and function of mRNA alternative polyadenylation,” *Nature Reviews Molecular Cell Biology*. Nature Research, 2022. doi: 10.1038/s41580-022-00507-5.

[85] G. W. Yeo *et al.*, “Alternative polyadenylation mediates genetic regulation of gene expression”, doi: 10.7554/eLife.57492.

[86] R. Wang, D. Zheng, L. Wei, Q. Ding, and B. Tian, “Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes,” *Cell Rep*, vol. 26, no. 10, pp. 2766-2778.e6, Mar. 2019, doi: 10.1016/j.celrep.2019.02.049.

[87] World Health Organization, “Tobacco,” 2022, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tobacco

[88] G. Pintarelli *et al.*, “cigarette smoke alters the transcriptome of non-involved lung tissue in lung adenocarcinoma patients”, doi: 10.1038/s41598-019-49648-2.

[89] S. Guo and S. Lin, “mRNA alternative polyadenylation (APA) in regulation of gene expression and diseases,” *Genes and Diseases*. Chongqing University, 2021. doi: 10.1016/j.gendis.2021.09.005.

[90] S. Venkat, A. A. Alahmari, and M. E. Feigin, “Drivers of gene expression dysregulation in pancreatic cancer”, doi: 10.1016/j.trecan.2021.01.008.

[91] S. Venkat *et al.*, “Alternative polyadenylation drives oncogenic gene expression in pancreatic ductal adenocarcinoma,” 2020, doi: 10.1101/gr.257550.119.