



# UNIVERSITÀ DI SIENA 1240

Department of Medical Biotechnology

Doctorate in Genetics, Oncology, and Clinical Medicine  
(GenOMeC)

XXXVII Cycle (2021-2024)

Coordinator: Prof. Ilaria Meloni

## Landscape of *BRAF* transcript variants in human cancer

Scientific disciplinary sector: BIO/11-Molecular Biology

Tutor:

Dr. Laura Poliseno

Institute of Clinical Physiology (IFC),  
National Research Council (CNR)

PhD Candidate:

Maurizio Salvatore Podda

Institute of Clinical Physiology (IFC),  
National Research Council (CNR)

Supervisor:

Dr. Romina D'Aurizio

Institute of Informatics and Telematics (IIT),  
National Research Council (CNR)

Academic year 2023/2024



# Table of contents

|                                                                                               |           |
|-----------------------------------------------------------------------------------------------|-----------|
| <b>Introduction.....</b>                                                                      | <b>12</b> |
| <b>1 Biological Background .....</b>                                                          | <b>12</b> |
| 1.1 The RAF kinase family.....                                                                | 12        |
| 1.2 RAF kinase signaling pathway and structure .....                                          | 12        |
| 1.3 The BRAF Ser-Thr protein kinase.....                                                      | 13        |
| 1.4 Alternative splicing in cancer .....                                                      | 15        |
| 1.5 BRAF splicing variants.....                                                               | 16        |
| 1.6 The role of <i>BRAF-204</i> transcript as possible sponge for miRNAs .....                | 19        |
| 1.7 Kidney Renal Papillary Cell Carcinoma.....                                                | 19        |
| 1.8 Transcription Factor AP-2 Alpha (TFAP2 alpha).....                                        | 20        |
| <b>2 <i>In silico</i> methodologies: The generations of sequencing .....</b>                  | <b>20</b> |
| 2.1 First, second, and third-generation sequencing .....                                      | 22        |
| 2.1.1 First-generation sequencing technology.....                                             | 22        |
| 2.1.2 Second-generation sequencing technology .....                                           | 23        |
| 2.1.3 Third-generation sequencing technology .....                                            | 24        |
| 2.1.4 Comparison between second-generation and third-generation sequencing technologies ..... | 24        |
| 2.2 Transcriptomics .....                                                                     | 25        |
| 2.2.1 Short-reads RNA-Seq Technologies.....                                                   | 25        |
| 2.2.2 Short reads RNA-Seq normalization.....                                                  | 26        |
| 2.2.3 Full-length isoform quantification from short-reads RNA-Seq.....                        | 28        |
| 2.2.4 QuantSeq 3' mRNA sequencing.....                                                        | 29        |
| 2.2.5 Single-Cell Transcriptomics.....                                                        | 30        |
| 2.2.6 Long-reads RNA-Seq Technologies .....                                                   | 31        |
| <b>3 Database resources and info .....</b>                                                    | <b>33</b> |
| 3.1 The Gene Expression Omnibus database.....                                                 | 33        |
| 3.2 The Cancer Genome Atlas (TCGA) .....                                                      | 33        |
| 3.3 The Genotype-Tissue Expression (GTEx).....                                                | 35        |
| 3.4 The Cancer Cell Line Encyclopedia .....                                                   | 36        |
| 3.5 FLIBase: a repository of full-length isoforms .....                                       | 37        |
| 3.6 APPRIS: annotation of principal and alternative splice isoforms database.....             | 37        |
| <b>4 <i>In Silico</i> methodologies: Structural bioinformatics and modeling.....</b>          | <b>38</b> |
| 4.1 AlphaFold .....                                                                           | 38        |
| 4.2 Molecular Dynamics Simulation.....                                                        | 38        |

|                                                                                                                                                                                      |           |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>Aim.....</b>                                                                                                                                                                      | <b>41</b> |
| <b>Materials and Methods .....</b>                                                                                                                                                   | <b>42</b> |
| 1 Overview of RNA-seq and QuantSeq 3' mRNA-Seq Datasets from Cell Lines and Tissue Samples .....                                                                                     | 42        |
| 2 Processing and Quantification of BRAF Expression Using the IsoWorm Pipeline .....                                                                                                  | 42        |
| 3 Quantification of BRAF isoform expression .....                                                                                                                                    | 43        |
| 4 Quantification of BRAF isoform expression using Long Reads and the TCGA and GTEX datasets .....                                                                                    | 44        |
| 5 Clinical analyses of TCGA patient.....                                                                                                                                             | 44        |
| 6 Differential Expression and Gene Set Enrichment Analysis, Gene Set Variation Analysis, and Gene Set Variation Analysis using the Molecular Signatures Database of KIRP genes ..... | 45        |
| 7 miRNA targets cross-reference study.....                                                                                                                                           | 46        |
| 8 Survival Analysis on TFAP2A gene expression.....                                                                                                                                   | 46        |
| 9 Visualization of transcript expression and ratios .....                                                                                                                            | 46        |
| 10 Single cell module .....                                                                                                                                                          | 46        |
| 11 Preparation of the molecular dynamic simulation with AlphaFold.....                                                                                                               | 47        |
| 12 Molecular Dynamics Simulation .....                                                                                                                                               | 47        |
| <b>Results and discussion .....</b>                                                                                                                                                  | <b>49</b> |
| 1 BRAF-ref/220 and BRAF-X1/204 isoforms in publicly available repositories of RNA and protein sequences.....                                                                         | 49        |
| 2 Tools for the quantification of <i>BRAF</i> transcript variants.....                                                                                                               | 51        |
| 3 <i>BRAF-204</i> is the most expressed <i>BRAF</i> transcript variant across cancer cell lines and tissues .....                                                                    | 54        |
| 4 The <i>BRAF-204/BRAF-220</i> ratio has prognostic value in KIRP .....                                                                                                              | 69        |
| 5 Differential expression analysis, GSEA and GSVA of KIRP patients .....                                                                                                             | 74        |
| 6 Exploring the potential role of BRAF-204 in miRNA sponging.....                                                                                                                    | 75        |
| 7 The TFAP2A gene is a prognostic marker on KIRP .....                                                                                                                               | 77        |
| 8 BRAF is not a marker suitable for single cell analysis .....                                                                                                                       | 80        |
| 9 Structural Differences Between BRAF-204 and BRAF-220 .....                                                                                                                         | 83        |
| 9.1 BRAF-204 and BRAF-220 modelling with AlphaFold.....                                                                                                                              | 83        |

|                                                      |           |
|------------------------------------------------------|-----------|
| 9.2 C-terms modelling using AlphaFold .....          | 84        |
| 9.3 Molecular Dynamic Simulations and Analysis ..... | 85        |
| <b>Conclusions.....</b>                              | <b>90</b> |
| <b>References.....</b>                               | <b>91</b> |

## List of figures

|                                                                                                                                                                                                                    |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Figure 1</b> RAF proteins organization and structure, Terrell et al. 2019 [5] .....                                                                                                                             | 13 |
| <b>Figure 2</b> The role of BRAF V600 mutation in cancer [8]. .....                                                                                                                                                | 14 |
| <b>Figure 3</b> Identified Alternative splicing.....                                                                                                                                                               | 16 |
| <b>Figure 4.</b> The evolution of sequencing technologies [40].....                                                                                                                                                | 21 |
| <b>Figure 5</b> Different applications of NGS, including technological platforms, data analysis, and applications [40]. .....                                                                                      | 22 |
| <b>Figure 6</b> NGS technologies with different platforms and principles. [40] .....                                                                                                                               | 23 |
| <b>Figure 7</b> Standard short reads RNAseq procedure .....                                                                                                                                                        | 26 |
| <b>Figure 8</b> Bubble graph of the cell lines deposited on the cancer cell line encyclopedia project divided by tissue [73].....                                                                                  | 36 |
| <b>Figure 9</b> Schematic representation of a molecular dynamics cycle. Illustration taken from Salmaso, 2018.....                                                                                                 | 39 |
| <b>Figure 10</b> Cartoon summarizing the similarities and differences between <i>BRAF-ref/220</i> (red) and <i>BRAF-X1/204</i> (blue) RNA and protein isoforms.....                                                | 49 |
| <b>Figure 11</b> Tracking of BRAF-220 annotation ( <b>a</b> ) and BRAF-204 annotation ( <b>b</b> ) from 2017 [15] to date.....                                                                                     | 50 |
| <b>Figure 12</b> Scheme of IsoWorm. ....                                                                                                                                                                           | 53 |
| <b>Figure 13</b> The 21 <i>BRAF</i> transcripts deposited in Ensembl (release 113) and isoform specific regions used by the custom module of IsoWorm for <i>BRAF-204</i> and <i>BRAF-220</i> quantification. ....  | 54 |
| <b>Figure 14</b> Pie charts showing that BRAF-204 is by far the most abundant of the 21 isoforms of BRAF present in Ensembl (113).....                                                                             | 55 |
| <b>Figure 15</b> Box plots of <i>BRAF-204/BRAF-220</i> ratio (Salmon). ....                                                                                                                                        | 56 |
| <b>Figure 16</b> Box plots of the BRAF-204/BRAF-220 ratio in cancer cell lines, CCLE (custom module) ..                                                                                                            | 57 |
| <b>Figure 17</b> Spearman test to assess the correlation between the quantification of BRAF-204/BRAF-220 ratio obtained with the Salmon module (y axis) and the custom module (x axis) of IsoWorm.                 | 58 |
| <b>Figure 18</b> Box plots of BRAF-204/BRAF-220 ratio tissues, GEO, Salmon .....                                                                                                                                   | 58 |
| <b>Figure 19</b> Box plots of the BRAF-204/BRAF-220 ratio tissues, GEO (custom) .....                                                                                                                              | 59 |
| <b>Figure 20</b> Spearman test to assess the correlation between the quantification of <i>BRAF-204/BRAF-220</i> ratio obtained with the Salmon module (y axis) and the custom module (x axis) of IsoWorm, GEO..... | 59 |

|                                                                                                                                                                                                                     |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Figure 21</b> Box plots of the BRAF-204/BRAF-220 ratio, TCGA.....                                                                                                                                                | 60 |
| <b>Figure 22</b> Box plots of the BRAF-204/BRAF-220 ratio for tcga normal tissue samples. ....                                                                                                                      | 61 |
| <b>Figure 23</b> Box plots of the BRAF-204/BRAF-220 ratio (GTEx).....                                                                                                                                               | 62 |
| <b>Figure 24</b> Box plots of BRAF-220, BRAF-204 that compose the TCGA (n = 9219 samples, 33 cancer types). ....                                                                                                    | 62 |
| <b>Figure 25</b> Box plots of <i>BRAF-220</i> , <i>BRAF-204</i> tissue samples that compose the GTEx (n = 2599 samples, 22 normal tissues). ....                                                                    | 63 |
| <b>Figure 26</b> CCLE Total <i>BRAF</i> levels were calculated using the Salmon module of IsoWorm and the trixmeta R package. They are expressed as $\log_2(\text{TPM} + 0.01)$ and they range between 2 and 4 .... | 64 |
| <b>Figure 27</b> Total <i>BRAF</i> levels were calculated using the Salmon module of IsoWorm and the trixmeta R package. ....                                                                                       | 64 |
| <b>Figure 28</b> TCGA Total <i>BRAF</i> levels were calculated using the recount3 package in R. They are expressed as $\log_2(\text{TPM} + 0.01)$ and they range between 2 and 4.....                               | 64 |
| <b>Figure 29</b> GTEx Total <i>BRAF</i> levels were calculated using the recount3 package in R. They are expressed as $\log_2(\text{TPM} + 0.01)$ and they range between 2 and 4.....                               | 65 |
| <b>Figure 30</b> Pie charts showing the 21 transcript variants of <i>BRAF</i> present in Ensembl, GEO, across 4 leukemia types (111). ....                                                                          | 66 |
| <b>Figure 31</b> Box plots of <i>BRAF-204/BRAF-220</i> ratio, GEO, across 4 leukemia types (111). ....                                                                                                              | 66 |
| <b>Figure 32</b> Total <i>BRAF</i> levels across 4 Leukaemia types, GEO.....                                                                                                                                        | 67 |
| <b>Figure 33</b> Venn diagram of lung cancer patients.....                                                                                                                                                          | 69 |
| <b>Figure 34</b> Correlations matrix.....                                                                                                                                                                           | 77 |
| <b>Figure 35</b> Kaplan-Meier curve that illustrate the survival probability between the two groups on KIRP patients.....                                                                                           | 78 |
| <b>Figure 36</b> UMAP Representation of Single-Cell Analysis Using the IsoWorm Single-Cell Module (Kallisto-Bustools).....                                                                                          | 82 |
| <b>Figure 37</b> Predicted structures of the BRAF-220 and the BRAF-204 complete sequences .....                                                                                                                     | 83 |
| <b>Figure 38</b> Plot of predicted IDDT (Predicted Intrinsic Disorder Index) as a function of position (A BRAF-220, B BRAF-204). ....                                                                               | 84 |
| <b>Figure 39</b> Predicted structures of the C-terminal regions. ....                                                                                                                                               | 85 |
| <b>Figure 40</b> RMSD values of the BRAF 204 C-terms (a) and BRAF-220 C-term (b) with respect to the equilibrated conformations as a function of simulation time. ....                                              | 86 |

**Figure 41** Number of Hydrogen Bonds Over Time of the BRAF 204 C-terms (a) and BRAF-220 C-term (b) with respect to the equilibrated conformations as a function of simulation time (100ns) 87

**Figure 42** Time evolution of the end-to-end distance, showing a structural transition and stabilization at later time points and lower values. a) BRAF-204; b) BRAF-220. .... 88

## List of tables

|                                                                                                                                                 |    |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Table 1</b> Ensembl DB 80 from 2015 containing all the BRAF transcript variants.....                                                         | 18 |
| <b>Table 2</b> Refseq DB 70 from 2015 containing all the BRAF transcript variants .....                                                         | 18 |
| <b>Table 3</b> Different generations of NGS platforms. ....                                                                                     | 21 |
| <b>Table 4</b> Summary of the tumor typology and their numerosity deposited in the TCGA database...                                             | 34 |
| <b>Table 5</b> Sample and data in the GTEx. ....                                                                                                | 35 |
| <b>Table 6</b> Hazard risk table.....                                                                                                           | 70 |
| <b>Table 7</b> Baseline characteristics of the 250 KIRP patients used for hazards regression analyses.....                                      | 72 |
| <b>Table 8</b> Univariate (left) and multivariate (right) Cox proportional hazards regression analyses performed on 250 KIRP patients.....      | 73 |
| <b>Table 9</b> Proportional odds logistic regression for ordered category outcome analysis performed on 250 KIRP patients. ....                 | 74 |
| <b>Table 10</b> Top ten pathways enriched in the GSEA on KIRP patients divided between higher and lower expression of BRAF ratio.....           | 75 |
| <b>Table 11</b> In silico analysis made using RNA22 webserver to test the binding between the mir-320a sequence and the 3'UTR of BRAF-204. .... | 76 |
| <b>Table 12</b> Univariate (left) and multivariate (right) Cox proportional hazards regression analyses performed on 250 KIRP patients.....     | 79 |

## Abstract

The BRAFV600E mutant kinase is widely studied as a cancer driver and therapeutic target. However, the regulation of *BRAF* gene expression remains largely unexplored. Here in this thesis, I investigate how the annotation of the biologically relevant protein-coding BRAF-ref and BRAF-X1 isoforms has evolved in public databases, and I faced the challenges posed by the discrimination and quantification of these *BRAF isoforms*, providing a detailed discussion and comparison of cutting-edge *in silico* methods such as short-read sequencing, long-read sequencing, and single-cell sequencing, which represent the state-of-the-art approaches for isoform identification and quantification. Furthermore, I investigate the clinical implications of the *BRAF-X1/BRAF-ref* transcript ratio, examining how this ratio could influence clinical outcomes. Currently, *BRAF-ref* corresponds to *NM\_004333.6* (NCBI), *ENST00000646891.2/BRAF-220* (Ensembl), MANE select (MANE), PRINCIPAL:4 (APPRIS), and *TCONS\_00952581* (FLIBase). BRAF-X1 corresponds to *NM\_001354609.2* (NCBI), *ENST00000496384.7/BRAF-204* (Ensembl), ALTERNATIVE BRAF isoform (APPRIS) and *TCONS\_00952571* (FLIBase). I built IsoWorm, a bioinformatic pipeline specifically tailored to discriminate and quantify BRAF isoforms and I employed it to analyze more than 500 cancer cell lines and 700 cancer tissue samples. Thanks to the FLIBase database, I also re-analyzed TCGA data on more than 9000 cancer tissue samples and 600 adjacent to tumoral tissue samples, and additionally, I examined more than 2,600 normal tissue samples from the GTEx database.

I integrated also additional modules into IsoWorm, enabling me to analyze different types of short RNA-seq data, such as Quant 3' sequencing and single-cell sequencing data. This permits me a more detailed investigation into the differences between BRAF isoforms.

I consistently found that *BRAF-X1* (now *BRAF-204*) is the most abundant BRAF isoform in human cancer and in normal tissues samples, it is 1.5 to 75 (in cancer) and 4 to 64 times (in normal samples) more expressed than *BRAF-ref* (now *BRAF-220*).

Crucially, I identified how in papillary renal cell carcinoma (KIRP) an extreme *BRAF-204/BRAF-220* ratio, together with the tumor stages and metastasis, higher the risk of the patient's overall survival.

Finally, thanks to the application of *in silico* structural biology techniques such as modeling with AlphaFold and molecular dynamics simulations, I investigated the structural differences between the C-termini of BRAF-220 and BRAF-204. Preliminary data suggest that the C-terminus of

BRAF-220 is more stable compared to that of BRAF-204. This structural stability could differentiate the two isoforms in terms of their biological roles, stability, and functionality.

Together with the experimental characterization undertaken over the years, our *in silico* analyses ultimately establish BRAF as a mix of isoforms, with *BRAF-204* being much more expressed than *BRAF-220*. These findings prompt a systematic benchmarking of the isoforms regarding molecular mechanisms, biological activities, and clinical relevance.

# Introduction

## 1 Biological Background

### 1.1 The RAF kinase family

RAF-protein kinases play different and crucial roles inside cells. The kinase family comprises 518 genes, making it one of the largest protein classes[1]. RAF-family kinases are a central regulatory hub within the signaling network, governing key cellular processes such as proliferation, growth, and differentiation. Based on the nature of the phosphorylated –OH group, these enzymes are classified into three main categories: protein-serine/threonine kinases (385 members), protein-tyrosine kinases (90 members), and tyrosine kinase-like proteins (43 members). The protein-serine/threonine kinases include the RAF family, which comprises A-RAF, B-RAF, and C-RAF. RAF-1, discovered in 1985, is now known as C-RAF, while A-RAF and B-RAF were identified in 1986 and 1988, respectively [2].

### 1.2 RAF kinase signaling pathway and structure

RAF Kinases participate in the RAS-RAF-MEK-ERK signal transduction cascade, also known as the MAPK cascade (mitogen-activated protein kinase) [3]. The RAF kinases have a substrate specificity for MEK1 and MEK2, catalyzing their phosphorylation and activation. MEK1/2 mediates, in turn, the phosphorylation of ERK1 or ERK2, their only physiological substrates [4]. The first group of the cascade includes the RAS family proteins: H-RAS, K-RAS A, K-RAS B, and N-RAS. A-RAF, B-RAF, and C-RAF belong to the second group of components in the MAPK signaling cascade, and within cells, these proteins can exist as both homo- and heterodimers. The third group of the cascade consists of the MEK1 and MEK2 proteins, while the final group includes ERK1 and ERK2. Lastly, the MAPK cascade is not a simple linear pathway but rather a highly branched and interconnected network [2].

The conserved domain structure of RAF kinase family members includes three regions: CR1, CR2, and CR3 (Figure 1). The C1 comprises the RAS-binding domain (RBD) and the cysteine-rich domain (CRD), which are in the N-terminal regulatory region, while the CR3 is the protein kinase domain, near the end of the C-terminal kinase domain. The N-term (regulatory domain) is important in autoinhibition and the RAS-driven activation and membrane association.

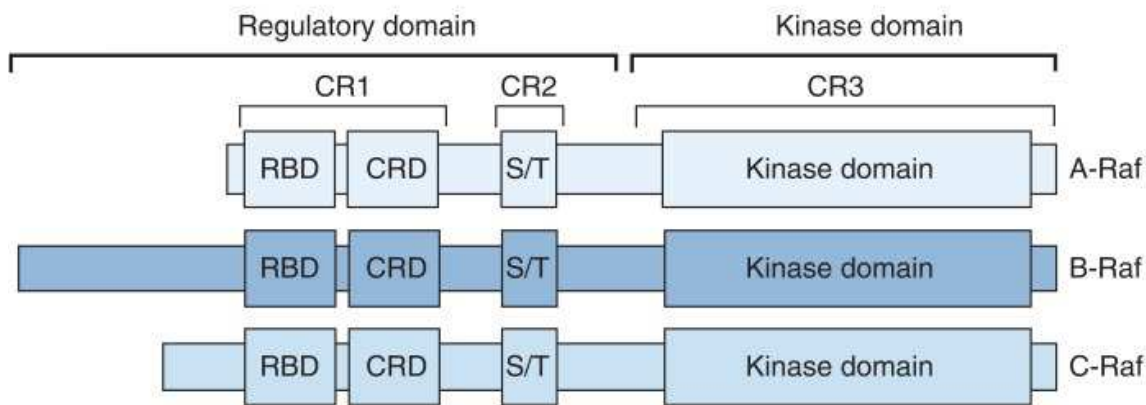


**Figure 1** RAF proteins organization and structure, Terrell et al. 2019 [5]

### 1.3 The BRAF Ser-Thr protein kinase

As mentioned in the previous paragraph, B-Raf proto-oncogene serine/threonine-protein kinase (BRAF) is part of the RAS/RAF/MEK/ERK pathway and acts as an effector of RAS. B-Raf forms dimers through hydrogen bonding and electrostatic interactions between its kinase domains in its active conformation. BRAF phosphorylates and activates MEK, which, after that, by phosphorylation, activates ERK. BRAF protein is essential in mediating various cellular processes, including cell growth, proliferation, differentiation, and survival [6]. The BRAF protein is the most extensively studied member of the RAF kinase family, as it is the predominant RAF kinase altered in human cancers. The most common mutation, accounting for 98% of all BRAF mutations, involves a nucleotide substitution that changes valine at position 600 to glutamic acid (V600E, nucleotide 1799 T > A; codon GTG > GAG; 98% of cases) (Figure 2). Other less frequent substitutions include lysine (V600K, 5–10% of cases) or different amino acids (V600A/D/G/L/M/Q/R, up to 5% of cases). V600 mutation causes BRAF to become independent of RAS activation, resulting in constitutive activity as a monomer. The oncogenic role of mutant BRAFV600E has been demonstrated in several cancers, including melanoma, colorectal cancer, lung cancer, and thyroid cancer [7]. BRAFV600E is involved in a lot mechanisms that cause cancer progression, mostly through the activation of the downstream MEK/ERK pathways, for example BRAFV600E is involved in evasion of senescence and apoptosis, angiogenesis (MEK-dependent activation of HIF-1 $\alpha$  and VEGF), tissue invasion and metastasis, and evasion of the immune response [8].

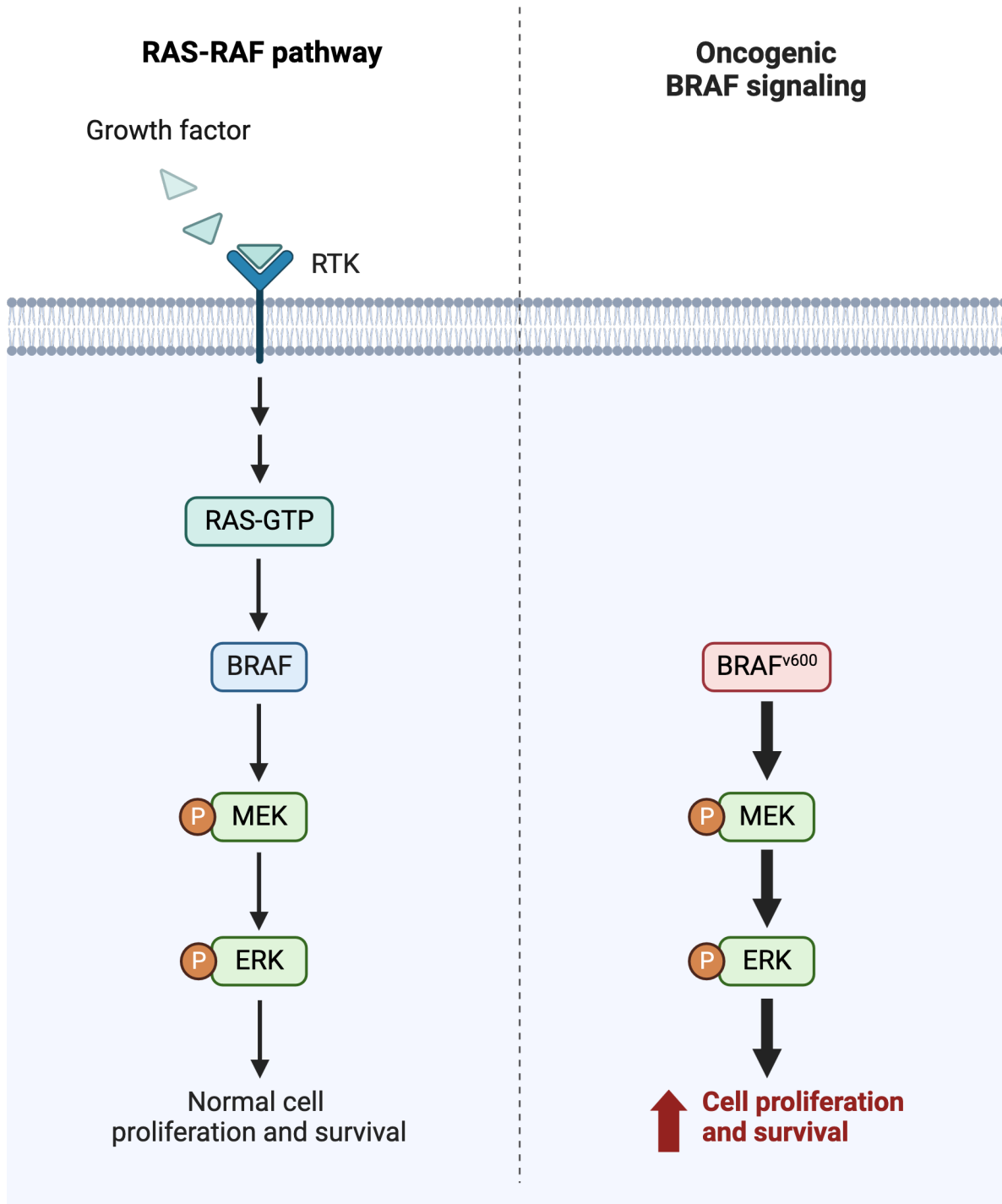


**Figure 2** The role of BRAF V600 mutation in cancer [8].

## 1.4 Alternative splicing in cancer

Alternative splicing is a phenomenon that affects nearly all human protein-coding genes in various ways. It can involve the inclusion or skipping of individual "cassette" exons, switching between alternative 5' and 3' splice sites, differential intron retention, mutually exclusive splicing of adjacent exons, and other, more complex patterns of splice site selection (Figure 3) [9]. Recent discoveries have shown that, on average, a protein-coding gene composed of 11 exons can generate approximately 5.4 distinct mRNA transcripts. An example of extreme alternative splicing is the human TTN gene, which encodes the muscle protein titin. This gene contains 364 coding exons, and, with RNA sequencing, 4,039 distinct splicing events associated with it have been identified [10].

Notably, each protein-coding mRNA molecule that encounters an alternative spliced event is translated into a protein variant (isoform), which could differ in function and structure concerning the reference mRNA molecule.

Alternative splicing can be influenced by different key factors, such as the properties of the splice sites, the effect of enhancing and silencing splicing factors, RNA secondary structures, and chromatin modifications [11].

Disease process can be altered by the activity of splicing factors and variants. In cancer biology, it is well known how gene expression profiles from healthy cells vs tumor cells are different since cancer can be caused by various factors, and alternative splicing can play a role in this complexity.

Growing evidence indicates that dysregulation of alternative splicing results in the production of tumor-associated isoforms, which profoundly affect cellular functions. These isoforms contribute to key cancer hallmarks, including sustained cell proliferation, evasion of apoptosis, metabolic reprogramming, angiogenesis, enhanced invasion, and metastasis, as well as the development of drug resistance [12].

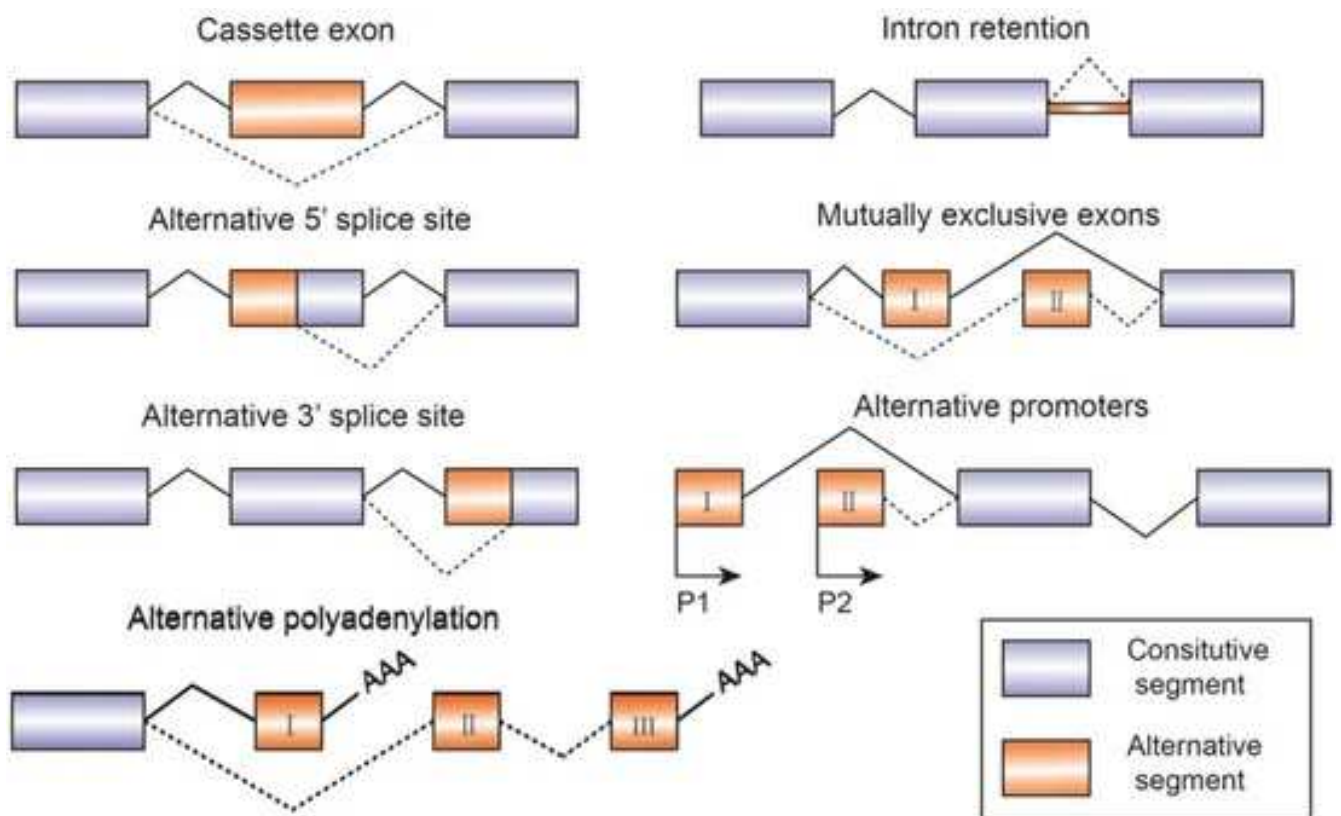


**Figure 3** Identified Alternative splicing

Basic seven types of alternative splicing have been identified: cassette exon, intron retention, alternative 5' splice site, mutually exclusive exons, alternative 3' splice site, alternative promoters, and alternative polyadenylation. In diagrams illustrating alternative splicing, splicing sites are typically connected by dashed lines. Boxes represent exons, while lines between the boxes represent introns [10]

### 1.5 BRAF splicing variants

The regulation of BRAF gene expression and the full spectrum of its biological activities remain largely unexplored, although such knowledge is both valuable per se and instrumental for achieving more precise and effective targeting [13]. In 2017, a study conducted a comprehensive investigation into the landscape of BRAF and its protein isoforms. At the time, in my laboratory we analyzed the 15 annotated and predicted RNA isoforms that were present in the Ensembl (release 80) and NCBI (GRCh38.p1) (table 1-2) databases via RNA-seq data from ~4800 cancer patients (9 cancer types) belonging to the TCGA. By mapping reads on isoform-specific exons or exon-exon junctions, the conclusion was that, in both healthy and cancerous tissues, *BRAF* mRNA invariably exists as a mix of three full-length isoforms, namely, *BRAF-ref* and *BRAF-X1*, which were also previously detected experimentally [14], as well as *BRAF-X2*. These isoforms differ in the C-term of

their coding sequences, as well as in the sequence and length of their 3'UTRs. They also show expression level specificities, with *BRAF-X1* being more highly expressed than *BRAF-ref* and *BRAF-X2* in most cancer cell lines, as well as in some of the cancer types analyzed. In terms of proteins, only BRAF-ref and BRAF-X1 are detectable, whereas BRAF-X2 is unstable (its C-term domain is recognized by the ubiquitin-proteasome pathway, and the protein is degraded) [15].

The ref and X1 isoforms have been further studied, both as mRNAs and as proteins. It has been established that the BRAF-X1 isoform is more phylogenetically conserved than the BRAF-ref isoform [7][16]. Furthermore, it has been identified a group of microRNAs [17] and one mRNA binding protein (PARP1, [18]) that selectively bind the X1 3'UTR, positively or negatively affect BRAF stability/translation and consequently contribute to fine-tuning the output of the ERK signaling pathway. In terms of proteins, both the ref and X1 isoforms are endowed with kinase activity and together account for the known oncogenic features displayed by BRAFV600E in melanoma cells [7] [19]. Interestingly, melanoma-free survival curves generated in vivo in adult zebrafish confirmed that the BRAFV600E-X1 protein is a much weaker melanoma driver than the BRAFV600E-ref protein. Crucially, the presence of the 3'UTR quenches the oncogenicity of the ref protein, underscoring the importance of post-transcriptional regulation and/or coding-independent functions for BRAFV600E biology [19]. The BRAF gene is located on chromosome 7q34 and the *reference transcript* encodes a protein that is composed of 766 amino acids while the *x1 transcript* a protein composed of 767 aa. The two isoforms differ for the last amino acids on the distal part of the C-terms. The last 8 aa for reference are -GYGAFPVH while for x1 -GYGEFAAFK.

| Name     | Transcript ID   | bp   | Protein    | Biotype                 | CCDS     | UniProt | RefSeq                 |
|----------|-----------------|------|------------|-------------------------|----------|---------|------------------------|
| BRAF-001 | ENST00000288602 | 2480 | 766aa      | Protein coding          | CCDS5863 | P15056  | NM_004333<br>NP_004324 |
| BRAF-002 | ENST00000497784 | 2336 | 194aa      | Nonsense mediated decay | -        | H7C5K3  | -                      |
| BRAF-003 | ENST00000496384 | 8294 | 375aa      | Protein coding          | -        | H7C560  | -                      |
| BRAF-004 | ENST00000469930 | 1058 | No protein | Retained intron         | -        | -       | -                      |
| BRAF-005 | ENST00000479537 | 743  | 102aa      | Nonsense mediated decay | -        | H7C4S5  | -                      |

**Table 1** Ensembl DB 80 from 2015 containing all the BRAF transcript variants

| Name           | Transcript ID  | Protein ID     | bp      | Protein |
|----------------|----------------|----------------|---------|---------|
| BRAF reference | NM_004333.4    | NP_004324.2    | 2949 bp | 766 aa  |
| BRAF-X1        | XM_005250045.1 | XP_005250102.1 | 9472 bp | 767 aa  |
| BRAF-X2        | XM_005250046.1 | XP_005250103.1 | 9318 bp | 758 aa  |
| BRAF-X3        | XR_927520.1    | XP_011514831.1 | 2506 bp | 711 aa  |
| BRAF-X4        | XR_927521.1    | XP_011514832.1 | 2588 bp | 593 aa  |
| BRAF-X5        | XM_011516529.1 | NA             | 2221 bp | NA      |
| BRAF-X6        | XR_242190.1    | NA             | 2467 bp | NA      |
| BRAF-X7        | XR_927522.1    | NA             | 2219 bp | NA      |
| BRAF-X8        | XR_927523.1    | NA             | 2301 bp | NA      |
| BRAF-X9        | XM_011516530.1 | NA             | 1892 bp | NA      |

**Table 2** Refseq DB 70 from 2015 containing all the BRAF transcript variants

## 1.6 The role of *BRAF-204* transcript as possible sponge for miRNAs

MicroRNAs (miRNAs) are small, single-stranded non-coding RNAs (21-25 nucleotides) that regulate gene expression by binding primarily to the 3' untranslated region (3'UTR) of target mRNAs [20]. This interaction typically leads to translational repression or mRNA degradation, ultimately reducing protein production. However, certain miRNAs have been shown to enhance gene expression by increasing mRNA stability or promoting translation [21]. In 2015, by employing small RNA sequencing coupled with an optimized analytical workflow (miR-CATCHv2.0), our group identified 20 miRNAs that directly bind to the 3'UTR of *BRAF-204* in melanoma cells [17].

The *BRAF-204* transcript possesses an extended 3'UTR (~7 kb). Notably, *BRAF-204* our group studied, at the time, how the *BRAF-204* is the most abundantly expressed transcript in both cell lines and patient samples, suggesting that it plays a predominant role in miRNA-mediated post-transcriptional regulation of BRAF expression. Between the 20 miRNAs that we identified, some act as negative regulators of gene expression, while others enhance translation or mRNA stability. We further demonstrate that these miRNAs function through distinct mechanisms, variably modulating mRNA stability or translation efficiency [17], [22].

## 1.7 Kidney Renal Papillary Cell Carcinoma

Renal cell carcinoma (RCC) is a common malignancy of the urinary system. The most common pathological type of RCC is clear cell carcinoma (ccRCC), accounting for 70% of renal cell carcinoma, followed by Kidney Renal Papillary Cell Carcinoma (KIRP or pcRCC, 15%) [23]. Due to the low prevalence of KIRP in RCC and the different mechanisms between KIRP and KIRC, patients with KIRP are often excluded from large clinical trials of the current targeted therapies [24].

KIRP is a heterogeneous disease originating in the tubular epithelium of the kidney, and the name derived from the papillary arrangement of tumor cells [25], [26].

KIRP was first divided in 1997 into type 1 and type 2, according to the histopathological characteristics and prognostic differences [27]. It is well known that the histological subtypes have an important prognostic predictor, with type 1 KIRP having a better prognosis than type 2 KIRP [28]. Surgical resection is an effective treatment for early-stage renal cell carcinoma; however, approximately 30% of cases are diagnosed at an advanced stage due to delayed detection.

Furthermore, 10% to 20% of patients experience postoperative recurrence, leading to metastatic kidney cancer, which significantly contributes to mortality [29], [30].

The limited understanding of renal cancer pathogenesis hinders the development of effective therapies for metastatic disease. Consequently, investigating the underlying mechanisms of kidney cancer is crucial for advancing its treatment.

### **1.8 Transcription Factor AP-2 Alpha (TFAP2 alpha)**

The transcription factor AP-2 Alpha is part of the transcription factor family activator protein 2, which is critical for controlling gene expression during embryonic development and differentiation [31]. TFAP2 family members tend to bind to GC-rich cis-regulatory elements when interacting with target genes[32]. The TFAP2A functions have been discussed in the literature; it has been discovered that it has a role in the early development of neural crest cells, the peripheral nervous system, facial and limb mesenchyme, various epithelia, and the extraembryonic trophoblast[31]. The gene also plays a crucial role in melanocyte differentiation by working alongside MITF, and together, they control the expression of genes essential for melanin synthesis. TFAP2A is aberrantly overexpressed in various types of cancer, including basal-squamous bladder cancer [33], pancreatic cancer [34], gastric cancer[35], lung adenocarcinoma[36], nasopharyngeal carcinoma [37] and ovarian cancer [38].

## **2 *In silico* methodologies: The generations of sequencing**

Next-generation sequencing (NGS), **Figure 4**, emerged between 1993 and 1998 [39]. NGS can sequence millions of DNA-RNA fragments at once, providing detailed information about the structure of genomes, genetic and transcript variations, activity, and changes in biological behaviors. Recent advancements in DNA and RNA sequencing have improved speed and accuracy, reduced costs, and enhanced data analysis, revolutionizing research and clinical applications [40]. The advance in NGS platforms (**Table 3**) has enhanced the comprehension of genome sequencing and has allowed scientists to investigate in fields such as transcriptomics, epigenomics, metagenomics, and omics studies (**Figure 5**). The capability of parallel sequencing millions to billions of DNA and RNA fragments has granted new opportunities in different fields of medical science, such as the understanding of genetic variations, gene and transcript expression, disease-causing variants, new drug targets, and study the heterogeneity of tumors and their developmental process [41], [42].

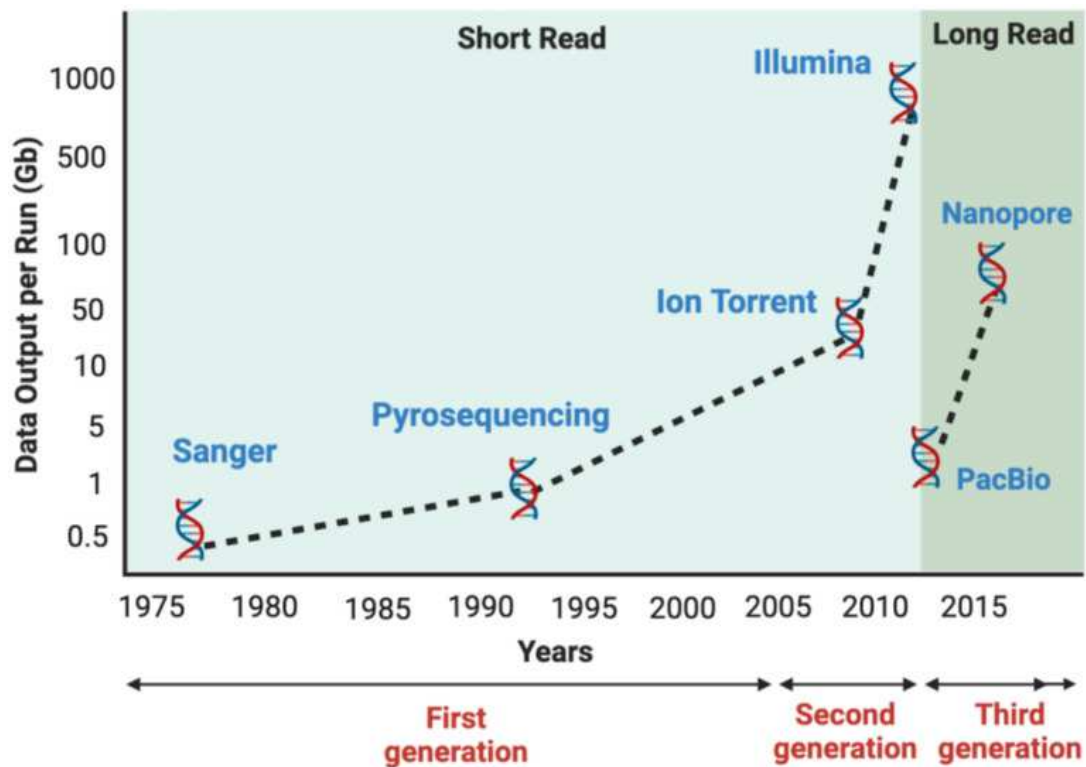


**Figure 4.** The evolution of sequencing technologies [40]

| Platform                                                      | Use                   | Sequencing Technology                           | Amplification Type            | Read Length (bp)      |
|---------------------------------------------------------------|-----------------------|-------------------------------------------------|-------------------------------|-----------------------|
| 454 pyrosequencing                                            | Short read sequencing | Seq by synthesis                                | Emulsion PCR                  | 400–1000              |
| Ion Torrent                                                   | Short read sequencing | Seq by synthesis                                | Emulsion PCR                  | 200–400               |
| Illumina                                                      | Short read sequencing | Seq by synthesis                                | Bridge PCR                    | 36–300                |
| SOLiD                                                         | Short read sequencing | Seq by ligation                                 | Emulsion PCR                  | 75                    |
| DNA nanoball sequencing                                       | Short read sequencing | Seq by ligation                                 | Amplification by Nanoball PCR | 50–150                |
| Helicos single-molecule sequencing                            | Short-read sequencing | Seq by synthesis                                | Without Amplification         | 35                    |
| PacBio Onso system                                            | Short-read sequencing | Seq by binding                                  | Optional PCR                  | 100–200               |
| PacBio Single-molecule real-time sequencing (SMRT) technology | Long-read sequencing  | Seq by synthesis                                | Without PCR                   | average 10,000–25,000 |
| Nanopore DNA sequencing                                       | Long-read sequencing  | Sequence detection through electrical impedance | Without PCR                   | average 10,000–30,000 |

**Table 3** Different generations of NGS platforms.

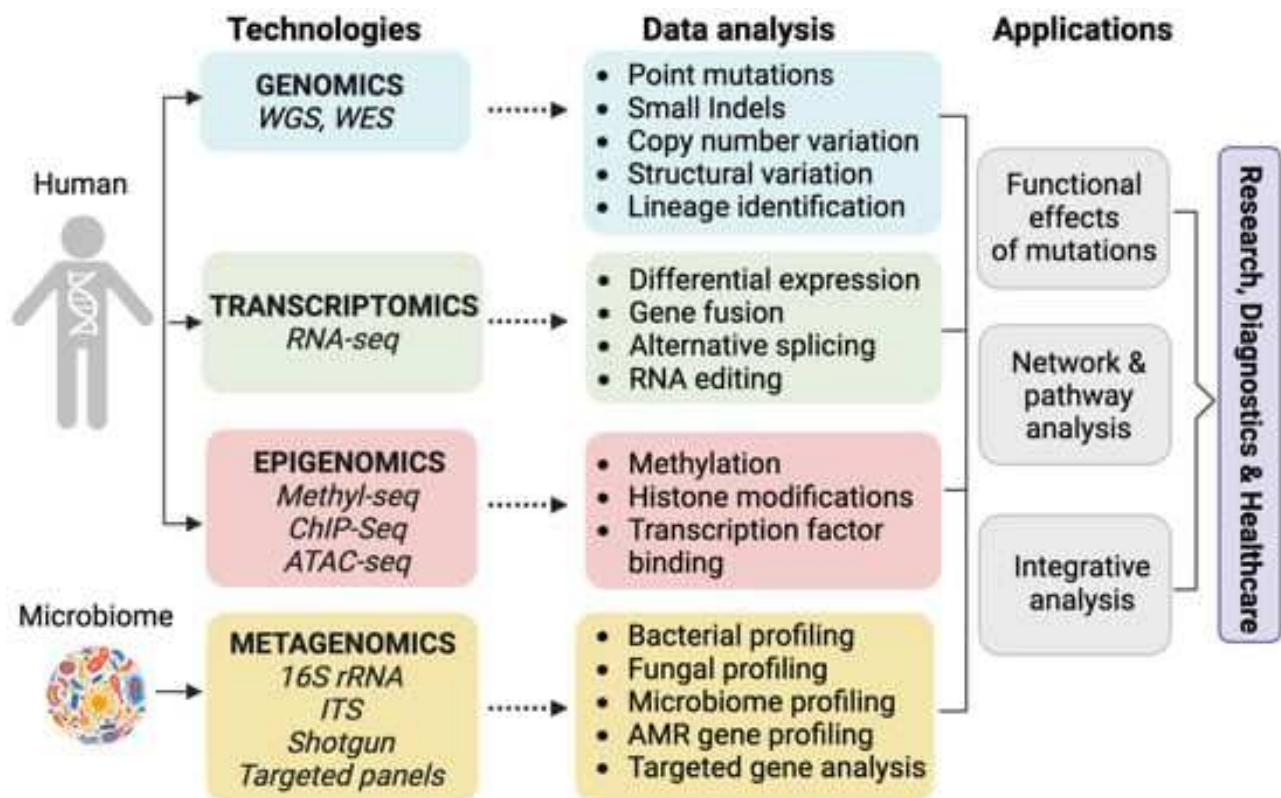


**Figure 5** Different applications of NGS, including technological platforms, data analysis, and applications [40].

## 2.1 First, second, and third-generation sequencing

### 2.1.1 First-generation sequencing technology

The first-generation sequencing technology involves all the tentative made year after year to sequence DNA and RNA using chemical degradation or enzymatic cleavage. The first was made in the 1964 by Robert Holley, which thanks to the use of the ribonuclease of *S. cerevisiae*, was able to obtain the sequence of a nucleic acid [43].

Another important attempt was the work made by Walter Gilbert and Allan Maxam that developed a chemical degradation method that allowed them to retrieve the complete sequence of the PhiX174 bacteriophage [44] however the game-changing technique was developed by Frederick Sanger: with his method that implements the use of dideoxynucleotides, the production of sequence reads was increased up to a few hundred nucleotides in length, revolutionizing the field the methodologies for the sequencing of DNA and RNA [45].

In 1987, the ABI 370 machine was created to automate the Sanger sequencing technique, greatly improving sequencing speed and accuracy. It became the industry standard, with advancements leading to higher-throughput sequencers capable of longer reads.

### 2.1.2 Second-generation sequencing technology

Second-generation approaches have improved the world of sequencing thanks to the simultaneous sequencing of thousands to millions of nucleic acid fragments. These methods generally start with molecule fragmentation followed by end-repair, adapter ligation, surface attachment, and in-situ amplification [46]. Second-generation sequencing technologies have vastly improved DNA sequencing speed and throughput, supporting applications like whole-genome sequencing, transcriptome analysis, and targeted sequencing (**Figure 6**). These advances have driven progress in genetic research, disease studies, and personalized medicine. Short-read sequencing technologies enable the massively parallel sequencing of millions of short DNA fragments simultaneously. However, due to their limited read length, reassembling long DNA sequences can be challenging, particularly in regions with structural variations or low complexity.

The precise identification and quantification of all the transcripts composing the transcriptome varieties can be challenging as well [46].

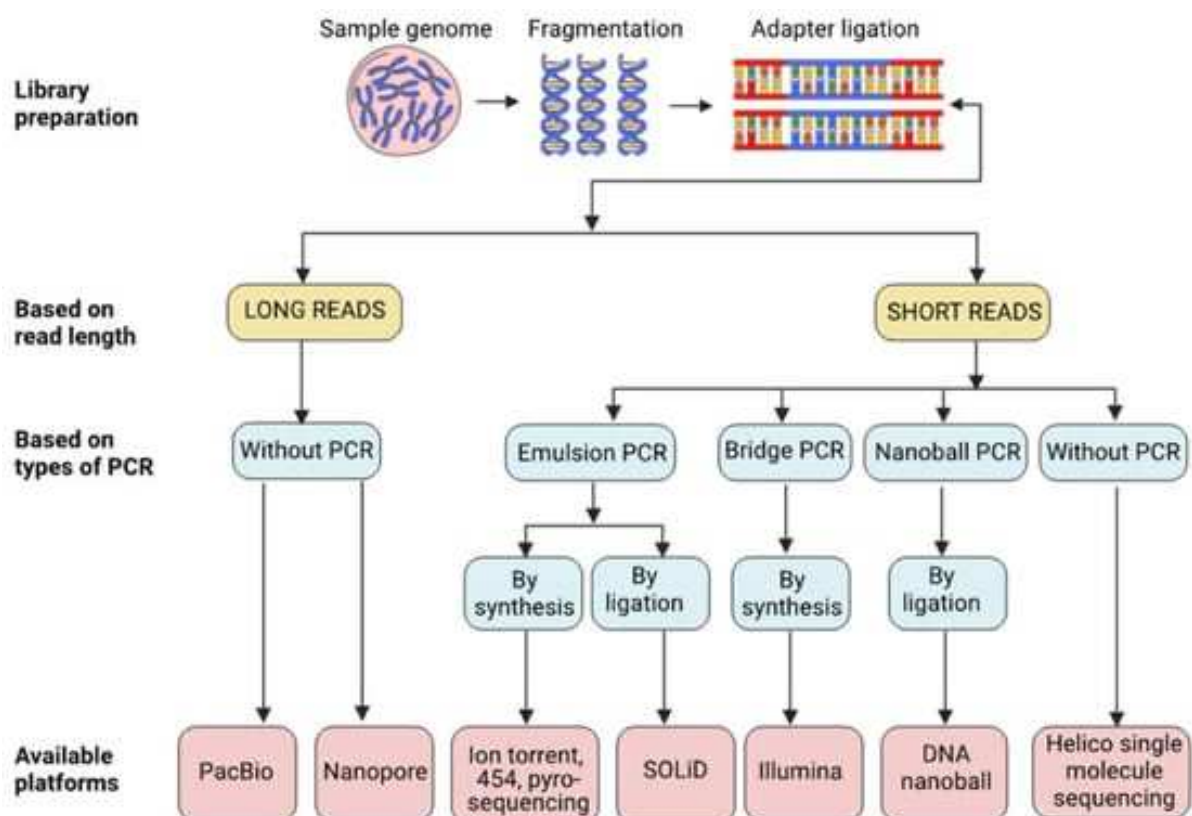


**Figure 6** NGS technologies with different platforms and principles. [40]

### ***2.1.3 Third-generation sequencing technology***

Third-generation sequencing approaches are the last ones added to the roster of the sequencing techniques that offer solutions to overcome the limitations of the previous generations. These approaches, which include platforms such as Pacific Biosciences, PacBio Sequencing, and Oxford Nanopore, can generate a read length of 10kb. The limitations of these technologies are the depth, the number of the reads generated, and the accuracy of the reads.

### ***2.1.4 Comparison between second-generation and third-generation sequencing technologies***

Second-generation sequencing approaches generate, through synthesis-based massive sequencing short-reads up to 250-800bp. In contrast, third-generation sequencing technology generates, through synthesis or by electrical voltage change/impedance, long reads up to 25-30kb. The major advantages of using long reads are on different levels: with long reads, the library preparation is PCR-free, and thanks to the intrinsic nature of the technologies, long reads can easily detect DNA methylation, as well as complex structural variants such as large insertions, deletions, inversions, and duplications[47], [48].

Long-read sequencing technology has also proven its potential for isoform identification, overcoming limitations intrinsically associated with short-read technology. Conversely, the advantages of using short-read technology include capturing most of the genetic variation, determining the abundance of specific sequences, profiling transcript expression, and identifying variants.

This is largely due to the higher number of reads and greater sequencing depth than long reads. In conclusion, both long-reads and short-read methods have pros and cons; therefore, sometimes, the best results are obtained using both types of technologies[49].

## 2.2 Transcriptomics

The transcriptome is the full set of RNA transcripts, both coding and non-coding, produced by an organism or a specific population of cells. The importance of understating the transcriptome is fundamental because analyzing the expression of transcripts across different tissues, conditions, or time points provides essential insights into the functional elements of the genome, the molecular composition of cells and tissues, and the mechanisms underlying development and disease. Transcriptomics aims to catalog all RNA species (mRNAs, non-coding RNAs, and small RNAs), define gene transcriptional structures (start sites, 5' and 3' ends, splicing, and modifications), and quantify transcript expression changes across developmental stages and conditions [50], [51]. From 1991 to 2008, different attempts were made to obtain a partial human transcriptome, and, by 2015, transcriptomes had been published for thousands of individuals and organisms. Now, the two main approaches for studying transcriptomes are the microarrays and the RNA-seq. Microarray technology can measure the abundance of a list of transcripts, which were defined previously, via their hybridization with an array of probes [52]. Some approaches, such as exon-junction or genomic tilling arrays, have been developed as “specialized microarrays” to enhance transcript mapping and isoform detection, but besides these methods’ cost convenience and high throughput, they have big limitations since the arrays are related to the know transcript pool taken in the analysis. Also, they have problems with high background noise, detection range, and the requirement of a complex normalization step due to the challenges in cross-experiment comparisons [51].

The RNA-seq is part of the next-generation sequencing both in the second (short-reads) and third (long-reads) generations, as described below [50].

### 2.2.1 Short-reads RNA-Seq Technologies

Short-reads RNA-Seq is a widely used approach developed for transcriptome profiling using deep-sequencing technologies. Generally, a population of RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends (**Figure 7**). After that, the molecules are sequenced to obtain short sequences (reads) from one end (single-end seq.) or both ends (pair-end seq.). The sequenced reads are from 30 to 400 base pairs depending on the DNA sequencing used. After the sequencing, reads are aligned to a reference genome or transcriptome, or also assembled de novo without any reference, to produce a genome-scale transcription map that accounts for the level of expression of each transcript, and/or for the transcriptional structure. RNA-seq data can be

utilized to identify differential gene expression between conditions, uncover novel transcripts, analyze alternative splicing events, and investigate gene expression dynamics over time or across various tissues and cell types [40], [51]. The quantification of full-length isoforms will be further elaborated and explored in greater depth, as isoform quantification represents a central objective in modern transcriptomic profiling.

This analysis will be instrumental for understanding the role of alternative splicing and isoform switching in BRAF function, shedding light on their potential biological and pathological significance [53].

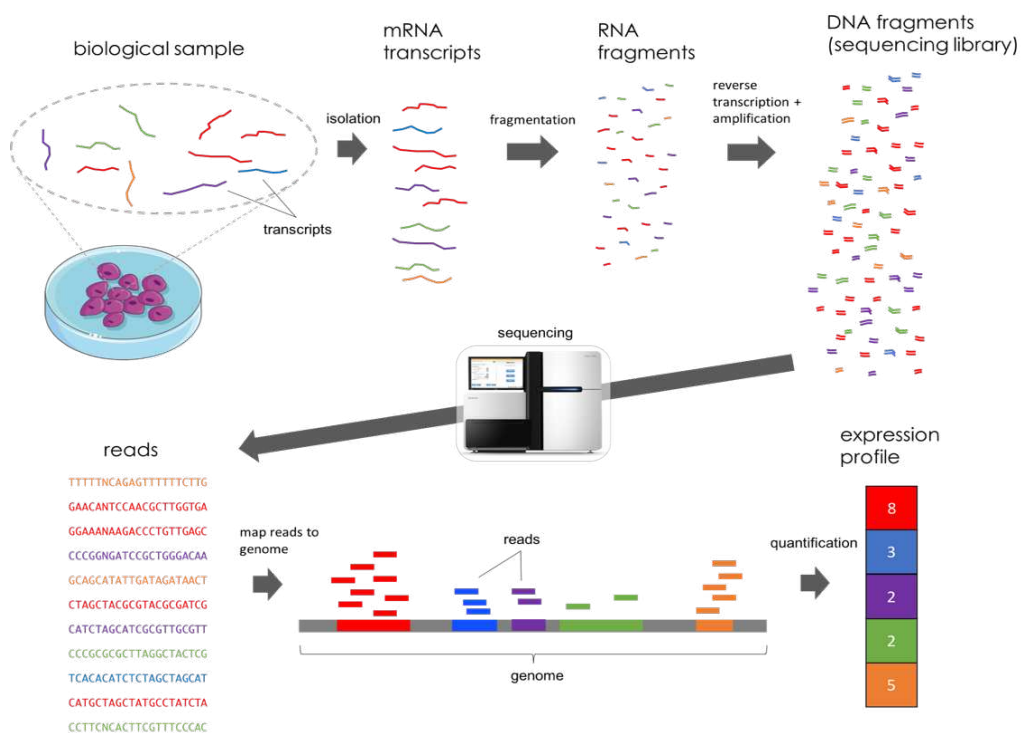


**Figure 7** Standard short reads RNAseq procedure

### 2.2.2 Short reads RNA-Seq normalization

Various factors affect transcript quantification, such as sequencing depth, transcript length, sample-to-sample and batch-to-batch variability [54]. For all these reasons, RNA-seq normalization is essential for RNA-seq data analysis, as well as to make gene expressions directly comparable within and across the sample and to remove batch effects.

Gene expression units such as CPM, FPKM, TPM, TMM are commonly used to normalize RNA-seq:

- **CPM/RPM (Counts/Reads per million):** This metric accounts only for sequencing depth, and for this reason, it is biased in some applications where gene length influences gene expression (RNA-seq). CPM is calculated by dividing the mapped read count by a per-million scaling factor of total mapped reads.

The formula is:

$$CPM_i = \frac{q_i}{\sum_j q_j} * 10^9$$

Where  $q_i$  is the raw fragments, count aligned, and  $\sum_j q_j$  is the total number of mapped fragments.

- **FPKM (Fragments per kilobase of transcript per million mapped fragments):** This method normalizes gene expression units to measure transcript abundance levels in a sample. FPKM (accounts for both the transcript length and the library size (sequencing depth), ensuring more accurate comparisons of gene expression across samples.

The formula is:

$$FPKM_i = \frac{q_i}{l_i * \sum_j q_j} * 10^9$$

Where  $q_i$  is the raw fragments count mapped,  $l_i$  is the respective genomic length, and  $\sum_j q_j$  is the total number of mapped fragments (library size).

- **TPM (Transcripts per million):** This method was proposed as an alternative normalization method to FPKM. The resulting value is divided by the total number of reads mapped to all transcripts, following normalization for transcript length. This is multiplied by one million to simplify downstream analyses [41]. This approach adjusts RNA-seq data for both sequencing depth and transcript length. TPM facilitates comparisons within a single sample but necessitates normalization across the dataset for comparisons between different samples [42].

The formula is:

$$TPM_i = \frac{FPKM_i}{\sum_j (FPKM_j)} * 10^6$$

- **TMM (trimmed mean of M-values):** TMM (Trimmed Mean of M-values) operates under the assumption that most genes exhibit similar expression levels across samples [55]. However, when many genes are uniquely or highly expressed under one experimental condition, it can distort the accurate quantification of other genes. To address this issue, TMM applies scaling factors to adjust library sizes for normalization within a dataset. The process begins by designating one sample as the reference. Fold changes and absolute expression levels of other samples are calculated relative to this reference sample [55]. Subsequently, genes are "trimmed" to exclude those that are differentially expressed based on these values. A trimmed mean of the fold changes is then computed for each sample. Finally, read counts are normalized by scaling them according to this trimmed mean and the total count for each sample.

The formula is:

$$TMM \text{ Scaling Factor for sample } i = \exp \left( \frac{\sum_{j \in non-trimmed} w_j \cdot M_j}{\sum_{j \in non-trimmed} w_j} \right)$$

Where  $M_j$  is the log-ratio of expression levels for gene  $j$  in sample  $i$  compared to a reference sample  $t$ ,  $y_{ij}$  is the raw count for gene  $j$  in sample  $i$ ,  $N_i$  the library size for sample  $i$ ,  $y_{tj}$  the raw count for gene  $j$  in the reference sample  $t$ ,  $N_t$  the library size for the reference sample  $t$ , lastly,  $w_j$  is the inverse variance weigh for gene  $j$ , computed from the raw counts.

### 2.2.3 Full-length isoform quantification from short-reads RNA-Seq

As explained before, alternative splicing and isoform switching play a central role in cell function. Currently, many published algorithms exist, but the effective quantification of full-length isoforms from short-read RNA-Seq remains problematic. The main limitation is that individual short reads do not contain information on long-range interactions that would associate splicing events separated by more than the fragment length. At the same time, the advancements in long-read sequencing technology still lag short-read technology in terms of throughput, exhibit significantly higher per-base error rates, and tend to be more costly. As a result, most RNA-Seq studies rely on short reads if competing technologies will not reach greater maturity.

There are three main approaches [53] to quantify isoform expression using short-read data:

- First: Aligners that are intron-aware that map reads to the genome as STAR [56], the genome alignment info is then used to assign quantified reads to the transcripts
- Second, like the first one, the difference is that these aligners are based on reads aligned to the transcriptome rather than the genome, like RSEM [57]
- Third: approaches that follow the concept of pseudo-alignment, which prioritizes execution performance and does not involve bona fide alignment, like kallisto or Salmon [[58], [59]]

As a matter of fact, the genome aligners are transcriptome-aware, and transcriptome alignments are genome-aware, but today to date they are still distinguished into two groups. The most popular isoform quantification methods are kallisto, Salmon, RSEM, Cufflinks, HTSeq, and featureCounts [53]. Starting from the assumption that HTSeq and featureCounts are not recommended for full-length isoform quantification Salmon marginally outperformed the other methods [53].

RSEM performs as well very well, but Salmon, as a pseudo-aligner, offers the advantage of greater efficiency, skipping the alignment process. Salmon shares many similarities with kallisto; the main difference involves the incorporation of a sample-specific model of fragment GC bias to improve its quantification estimates [58], [59]. Additionally, Salmon includes features to address read start sequence bias (e.g., from random hexamer priming) and positional biases (e.g., 5' or 3' biases) [53].

#### **2.2.4 QuantSeq 3' mRNA sequencing**

Among the RNA sequencing technologies, the QuantSeq 3' mRNA sequencing existed. QuantSeq provides an easy protocol to generate highly strand-specific next-generation sequencing (NGS) libraries close to the 3' end of polyadenylated RNAs within 4.5 h. Only one fragment per transcript is generated, so normalization is not required, directly linking the number of reads mapping to a gene to its expression[60]. The QuantSeq technology targets the 3' ends of mRNA molecules for a quick and cost-effective gene expression analysis by RNA-Seq and other more advanced applications.

QuantSeq offers several versions for different RNA sequencing:

- **QuantSeq FWD V2 with UDI** – Ideal for cost-effective gene expression profiling of eukaryotic samples, this version enables sequencing-ready library preparation from total RNA in under 4.5 hours, with no pre-processing required. Optimized for sample multiplexing.
- **QuantSeq REV (Reverse)** – Designed for poly(A) site identification, it generates Illumina-compatible 3' mRNA-Seq libraries, specifically targeting the 3' ends of polyadenylated transcripts.
- **QuantSeq-Pool** – Perfect for large-scale screening projects, this version streamlines gene expression profiling by incorporating sample barcoding, early pooling, and batch processing.
- **QuantSeq-Flex** – A customizable library preparation protocol for advanced users, allowing the creation of Illumina-compatible libraries from any RNA sample using custom primers.

The QuantSeq REV kit can be used to study alternative polyadenylation sites, map the precise transcript end site, and analyze gene expression.

### ***2.2.5 Single-Cell Transcriptomics***

Single-cell Transcriptomics or single-cell RNA sequencing (scRNA-seq) is a technique that enables the study of gene expression profiles at the individual cell level. It allows the profiling of transcriptomes from individual cells, providing insights into cellular heterogeneity, cell type identification, cell lineage analysis, and gene expression dynamics in complex tissues or developmental processes [46]. The scRNA-seq has also proven invaluable in the identification of rare malignant cells, aiding in the study of cancer biology and personalized medicine [61].

scRNA-seq protocols can be divided in two main categories [62] based on the transcript coverage:

- Full-length transcript sequencing approaches (as Smart-seq2, MATQ-seq)
- 3'-end (Drop-seq, Chromium) or 5'-end (STRT-seq)

Compared to bulk RNA-seq, single-cell RNA-seq (scRNA-seq) generates data that is inherently noisier and more variable. The combination of technical noise and biological variation, such as stochastic transcriptional activity, presents significant challenges for the computational analysis of scRNA-seq datasets. Different methods to analyze scRNA-seq have been developed to overcome these

challenges [63]. Two of the main challenges that require further investigation are cell clustering (essential for analyzing cellular heterogeneity, rare cell types and cell-cell interactions) and cell classification. The latter is still done manually in many cases, and this heavily relies on the domain expertise [63], even if, different methods and computational tools have been developed to annotate cell types, divided in categorized as unsupervised or supervised methods. Another big limit of the technology is the low capture efficiency: for most expressed genes, scRNA-seq data may only capture zero or one copy of the mRNA from the biological samples [64].

A general workflow to analyze single-cell data is composed of different steps. The first step is to remove outlier or low-quality cells and genes (quality control), then, cells are clustered and visualized by a two-dimensional map method as t-distributed stochastic neighbor embedding (tSNE,[65]) or uniform manifold approximation and projection (UMAP, [66]). The cells are also clustered by similar gene expression. The last step is the cell-type annotation and identification. This process is fundamental to understanding and analyzing cellular heterogeneity and facilitates downstream analyses such as cell-cell interactions and data integration [63].

Accurate and efficient annotation of cells or clusters into biologically relevant types, subtypes, and supertypes is a complex and challenging process. Essential elements of the annotation workflow include signature gene databases, robust scoring methods, high-quality reference datasets, and supervised modeling techniques.

### **2.2.6 Long-reads RNA-Seq Technologies**

Third-generation sequencing has revolutionized the study of genomics and transcriptomics. In this paragraph, long-read RNA-Seq technologies will be examined in depth, following the discussion in section 2.3, which has already addressed the general differences between short-read and long-read approaches. Various historical challenges, such as low yield, high error rates, and elevated costs per base, have constrained the widespread use of third-generation sequencing technologies in large-scale sequencing initiatives. In recent years, however, notable advancements have addressed these obstacles, significantly decreased error rates, and marked enhancements in overall performance[67]. Long-reads approaches are divided into 2 key technologies: single-molecule real-time sequencing (PacBio) and the tech that employs nanopores to read long RNA molecules (Oxford Nanopore).

- **Single-molecule real-time sequencing (SMRT):** this technology, developed by PacBio, is divided into two modes of SMRT sequencing: the continuous long-reads (CLR, now obsolete) and the Circular Consensus Sequencing (CCS), which produces HiFi reads [67]. HiFi reads are a typology of reads developed by the PacBio Long Read System. [HiFi reads](#) produce highly accurate [long read sequencing](#) datasets, with an average read length of 10-25 kb and an accuracy of greater than 99.5% [22]. The accuracy of PacBio HiFi sequence data had a significant impact on different fields, from variant discovery to providing access to complex regions of repetitive DNA and discovering new transcript, and all of these reducing the cost of the assembly. However, the high cost has limited the usage of this technology [67].
- **Oxford Nanopore Technologies (ONT):** The technology generates reads of hundreds to thousands of kilobases in length. Two main types of Oxford Nanopore Tech exist: the long-read (10-100kb) and the ultra-long reads (>100kb). ONT platforms, MinION, GridION, and PromethION, differ in flow cell capacity and throughput, with PromethION offering the highest output due to its 2675 nanopore channels, achieving up to 200 Gbases of native genomic DNA reads. Ultra-long reads [68], spanning several megabases, maintain similar accuracy to standard long reads and are crucial for resolving repetitive genomic regions and structural variants and enabling haplotype phasing, surpassing the capabilities of short-read sequencing technologies.

Both PacBio and Oxford Nanopore have also significantly advanced transcriptome analysis since these technologies make it possible to sequence full-length RNA molecules using only short-read approaches (as discussed above). Overall, the main applications of long-reads in transcriptomic are:

- **Reference-Free Transcriptome Assembly:** methods designed for the de novo assembly of long-read transcriptome sequencing data without relying on a reference genome. This approach is particularly beneficial for studying non-model organisms or for uncovering novel transcripts in well-characterized species [69].
- **Reference-Guided Transcriptome Assembly:** methods designed for guided assembly of long-read transcriptome sequencing data using a reference genome [70].

- **Isoform Detection and Quantification:** Long-read sequencing facilitates the identification and quantification of full-length isoforms, providing insights into gene expression regulation and alternative splicing [71].

Despite all the long-read advantages, several historical limiting factors, e.g., limited yield, high error rates, and high cost per base, have hindered the widespread adoption of these third-generation sequencing technologies in large-scale sequencing projects. However, significant progress in recent years has mitigated these limitations, leading to substantial reductions in error rates and overall performance improvements [67].

### 3 Database resources and info

#### 3.1 The Gene Expression Omnibus database

The Gene Expression Omnibus (GEO) is an international public repository managed by the National Center for Biotechnology Information (NCBI) that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. GEO serves as a critical resource for researchers studying gene expression, genotype, and other molecular variations in relation to biological states, diseases, and experimental treatments.

#### 3.2 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) is a groundbreaking initiative in cancer genomics that has molecularly profiled over 20,000 primary cancer samples alongside matched normal tissues, encompassing 33 different cancer types (**Table 4**). Launched in 2006 as a collaborative effort by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), TCGA brought together experts from various fields and institutions to advance our understanding of cancer, with the aim of cataloging and uncovering the key genomic changes of it [66]. The project leveraged advanced genomic analysis technologies, encompassing whole-exome sequencing, whole-transcriptome sequencing, and, in select cases, whole-genome sequencing. Additionally, it employed methods like gene expression profiling, copy number variation (CNV) analysis, SNP genotyping, genome-wide DNA methylation profiling, and microRNA profiling.

| <b>Cancer type</b>                                                      | <b>Abbr.</b> | <b>Tumor samples</b> | <b>Adjacent normal samples</b> |
|-------------------------------------------------------------------------|--------------|----------------------|--------------------------------|
| <b>Adrenocortical Carcinoma</b>                                         | ACC          | 79                   | 0                              |
| <b>Bladder Urothelial Carcinoma</b>                                     | BLCA         | 406                  | 19                             |
| <b>Breast Invasive Carcinoma</b>                                        | BRCA         | 1095                 | 113                            |
| <b>Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma</b> | CESC         | 304                  | 0                              |
| <b>Cholangio Carcinoma</b>                                              | CHOL         | 35                   | 8                              |
| <b>Colon Adenocarcinoma</b>                                             | COAD         | 272                  | 24                             |
| <b>Lymphoid Neoplasm Diffuse Large B-cell Lymphoma</b>                  | DLBC         | 48                   | 0                              |
| <b>Esophageal Carcinoma</b>                                             | ESCA         | 162                  | 0                              |
| <b>Glioblastoma Multiforme</b>                                          | GBM          | 154                  | 0                              |
| <b>Head and Neck Squamous Cell Carcinoma</b>                            | HNSC         | 502                  | 43                             |
| <b>Kidney Chromophobe</b>                                               | KICH         | 65                   | 24                             |
| <b>Kidney Renal Clear Cell Carcinoma</b>                                | KIRC         | 532                  | 72                             |
| <b>Kidney Renal Papillary Cell Carcinoma</b>                            | KIRP         | 290                  | 32                             |
| <b>Acute Myeloid Leukemia</b>                                           | LAML         | 150                  | 0                              |
| <b>Brain Lower Grade Glioma</b>                                         | LGG          | 514                  | 0                              |
| <b>Liver Hepatocellular Carcinoma</b>                                   | LIHC         | 371                  | 50                             |
| <b>Lung Adenocarcinoma</b>                                              | LUAD         | 516                  | 58                             |
| <b>Lung Squamous Cell Carcinoma</b>                                     | LUSC         | 501                  | 49                             |
| <b>Mesothelioma</b>                                                     | MESO         | 87                   | 0                              |
| <b>Ovarian Serous Cystadenocarcinoma</b>                                | OV           | 376                  | 0                              |
| <b>Pancreatic Adenocarcinoma</b>                                        | PAAD         | 178                  | 0                              |
| <b>Pheochromocytoma and Paraganglioma</b>                               | PCPG         | 165                  | 0                              |
| <b>Prostate Adenocarcinoma</b>                                          | PRAD         | 497                  | 52                             |
| <b>Rectum Adenocarcinoma</b>                                            | READ         | 94                   | 0                              |
| <b>Sarcoma</b>                                                          | SARC         | 259                  | 0                              |
| <b>Skin Cutaneous Melanoma</b>                                          | SKCM         | 103                  | 0                              |
| <b>Stomach Adenocarcinoma</b>                                           | STAD         | 375                  | 27                             |
| <b>Testicular Germ Cell Tumors</b>                                      | TGCT         | 150                  | 0                              |
| <b>Thyroid Carcinoma</b>                                                | THCA         | 504                  | 59                             |
| <b>Thymoma</b>                                                          | THYM         | 120                  | 0                              |
| <b>Uterine Corpus Endometrial Carcinoma</b>                             | UCEC         | 178                  | 0                              |
| <b>Uterine Carcinosarcoma</b>                                           | UCS          | 57                   | 0                              |
| <b>Uveal Melanoma</b>                                                   | UVM          | 80                   | 0                              |

**Table 4** Summary of the tumor typology and their numerosity deposited in the TCGA database

### 3.3 The Genotype-Tissue Expression (GTEx)

The Genotype-Tissue Expression (GTEx) Portal is a comprehensive public resource for researchers studying not pathologic tissue and cell-specific gene expression with data from NIH projects.

It's aims is to establish a resource database and associated tissue bank in which to study the relationship between genetic variation and gene expression and other molecular phenotypes in multiple reference tissues (**Table 5**) [72] The GTEx database includes RNA sequencing (RNA-seq) data from over 50 tissue types, it is a very valuable to researcher for analyzing tissue-specific gene expression and understanding the genetic basis of complex traits and diseases.

| Tissue type     | Sample size |
|-----------------|-------------|
| Adrenal gland   | 100         |
| Bladder         | 21          |
| Brain           | 100         |
| Breast          | 100         |
| Cervix          | 19          |
| Colon           | 100         |
| Esophagus       | 100         |
| Heart           | 100         |
| Kidney          | 89          |
| Liver           | 221         |
| Lung            | 100         |
| Muscle          | 100         |
| Ovary           | 180         |
| Pancreas        | 100         |
| Prostate        | 241         |
| Skin            | 100         |
| Small intestine | 187         |
| Spleen          | 241         |
| Stomach         | 100         |
| Testis          | 100         |
| Thyroid         | 100         |
| Uterus          | 100         |

**Table 5** Sample and data in the GTEx.

### 3.4 The Cancer Cell Line Encyclopedia

The Cancer Cell Line Encyclopedia (CCLE) is a comprehensive initiative aimed at producing extensive profiling datasets for nearly 1,000 cell lines (**Figure 8**) derived from various tissue lineages. The initial release featured data on gene expression, DNA copy number variations, and hybrid capture sequencing [73]. Recently, histone profiling, RNASeq, DNA methylation, miRNA profiling, whole genome sequencing, and metabolite profiling were added [74]. Drug and shRNA sensitivity screens increased the richness of data attached to the CCLE [75]. In 2020, the CCLE was expanded to include proteomic profiles for 375 cell lines, obtained through mass spectrometry [76].

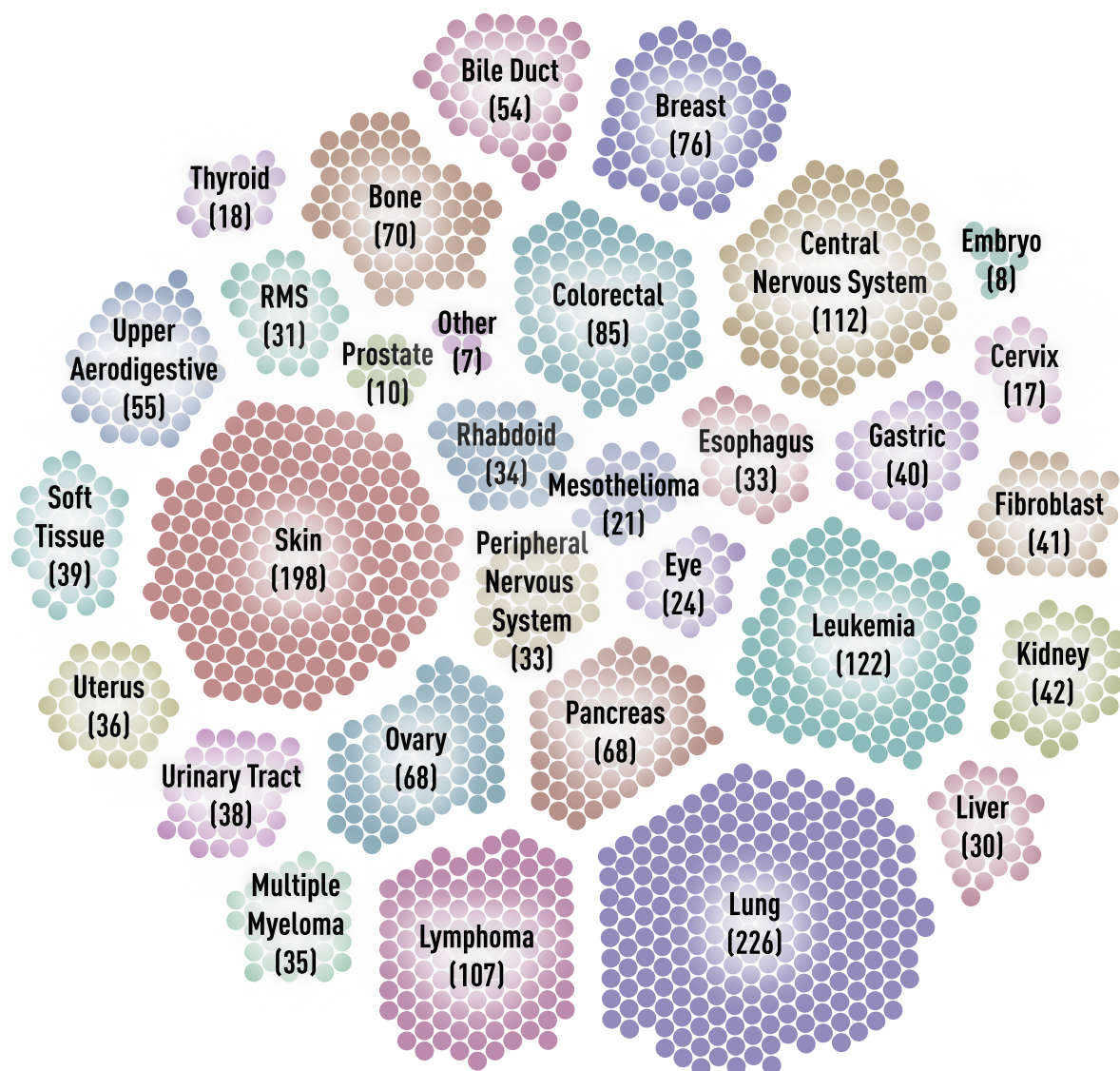


**Figure 8** Bubble graph of the cell lines deposited on the cancer cell line encyclopedia project divided by tissue [73].

### **3.5 FLIBase: a repository of full-length isoforms**

Published in 2023, the FLIBase is a specialized database dedicated to annotating full-length isoforms using advanced long-read sequencing technologies. The database integrates data from 351 long-read RNA sequencing samples and 12,469 short-read RNA sequencing samples derived from a wide range of normal and cancerous human tissues and cells. The current iteration of FLIBase includes 983,789 full-length spliced isoforms, identified through long-read sequencing and validated with short-read exon-exon splice junction data. Among these isoforms, 188,248 have been annotated, while 795,541 remain unannotated.

FLIBase offers a precise and exhaustive representation of full-length transcripts by addressing the inherent limitations of short-read RNA sequencing. This detailed annotation enables researchers to conduct various downstream analyses and investigations. FLIBase has proven highly effective in identifying many previously unannotated isoforms and tumor-specific RNA transcripts [77].

### **3.6 APPRIS: annotation of principal and alternative splice isoforms database**

The APPRIS initiative [78] is online as a database dedicated to annotating human splice isoforms. Its primary purpose is to enhance manual annotations of the human genome by incorporating robust protein structural and functional data. Beyond integrating and analyzing reliable predictions about the impact of splicing, APPRIS designates a single reference sequence for each gene, referred to as the principal isoform. This selection is based on an evaluation of the structural, functional, and conservation attributes of each transcript. APPRIS determines principal isoforms for 85% of the protein-coding genes included in the GENCODE 7 release for ENSEMBL. An analysis of the data indicates that at least 70% of alternative (non-principal) variants lack significant functional or structural features when compared to their principal counterparts.

## 4 *In Silico* methodologies: Structural bioinformatics and modeling

### 4.1 AlphaFold

AlphaFold, developed by DeepMind, is a groundbreaking artificial intelligence system that predicts protein structures with remarkable accuracy. Leveraging deep learning techniques, AlphaFold addresses the "protein folding problem," which has challenged scientists for decades [79]. The system was trained on publicly available protein data and utilizes advanced neural networks to predict the 3D structure of proteins from their amino acid sequences. Its transformative capabilities were demonstrated in the Critical Assessment of Protein Structure Prediction (CASP13) competition, where it achieved unprecedented performance[80]. This innovation has vast implications for biology and medicine, aiding in drug discovery and understanding diseases at the molecular level [81]. As of now, AlphaFold has been used to predict structures for nearly the entire human proteome, providing invaluable resources to researchers worldwide.

### 4.2 Molecular Dynamics Simulation

Molecular dynamics (MD) simulation is a computational technique that simulates the dynamic behaviour of protein or other molecular systems as a function of time. MD is based on a general model of the physics governing interatomic interactions [82]. Through MD simulations that reveal the positions of the atoms at femtosecond temporal resolution, a wide variety of important biomolecular processes can be simulated, like conformational change, ligand binding, and protein folding.

MD allows to describe the interactions between the atoms of the system through the so-called force field. A force field is the sum of functions that depend on the atomic coordinates of the system and a number  $N$  of parameters. The basic force field function is composed of the bonded terms (bond stretching, angle bending, improper dihedrals, dihedrals bond interaction) and the nonbonded terms (van der Waals and electrostatics forces). The force field could be summarized with the following general form used to calculate the total energy:

$$E_{total}=E_{bonded}+E_{nonbonded}$$

:

$$E_{bonded}=E_{stretching}+E_{bending}+E_{torsion} \text{ and } E_{nonbonded}=E_{coulomb}+E_{van\ der\ Waal}$$

There are various typologies of force fields designed by different groups and made up of parameter libraries optimized for various classes of molecules.

Several tools and codes for performing MD simulations of biomolecules, such as AMBER, CHARMM, NAMD, DESMOND, and GROMACS, are currently available. Most of these software supports multiple force fields and perform similar computations: the force exerted on each atom by all the other atoms is computed, and, by using Newton's laws of motion, the spatial position of each atom as a function of time is predicted. Then, the forces on each atom are repeatedly calculated step-by-step and used to update the position and velocity of each atom (**Figure 9**). The resulting trajectory is, in essence, a three-dimensional movie that describes the atomic-level configuration of the system at every point during the simulated time interval.

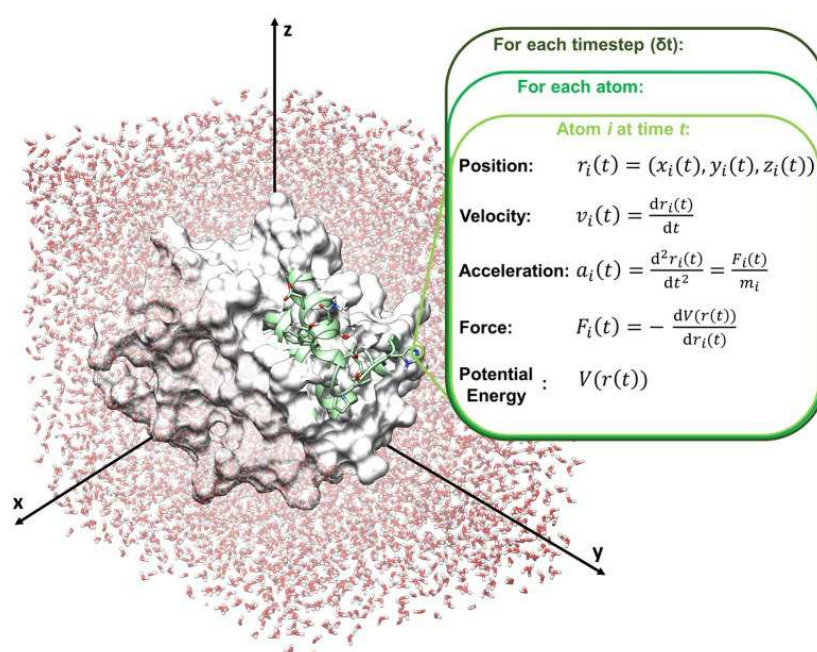


**Figure 9** Schematic representation of a molecular dynamics cycle. Illustration taken from Salmaso, 2018

In a typical MD simulation experiment, the building of the system represents the first step. Usually, the initial model of the system is obtained from experimental structures or homology modelling data. A typical system is composed of the solute (peptides, DNA, membranes, etc...) and the solvent (physiological solution, alcohols or DMSO...). The second step consists of optimizing the molecular geometry of the system. This is done usually by performing an *energy minimization*. For a molecular system composed by  $N$  atoms, energy is a function of the  $3N$  Cartesian coordinates, so it is necessary to find between all the possible conformations with the lowest energy, i.e., the conformation with

more stable geometry. This can be done through algorithms of minimization based on the calculation of the derivatives (*steepest descent* method or *conjugate gradients* method). Once the forces acting on individual atoms are obtained, classical Newton's law of motion is used to calculate accelerations and velocities and to update the atom positions. In MD simulations the thermic energy is provided through a constant average kinetic energy. To start a MD simulation, initial speeds are necessary. These are selected randomly from the gaussian distribution of Maxwell-Boltzmann, that describes the distribution of speeds of an atom relative to its temperature. The integration of movement is done numerically. Among the algorithms used for integrating the atom movement *Verlet* or *Leap Frog* are the most used. To ensure numerical stability, the time steps in a MD simulation must be short, typically only a few femtoseconds ( $10^{-15}$  s) each. Since most of the events of biochemical interest take place on timescales of nanoseconds, microseconds, or longer, a typical simulation involves millions or billions of time steps. This fact, combined with the millions of interatomic interactions typically evaluated during a single time step, causes simulations to be very computationally demanding. The present generation of computers takes benefits of parallelism and accelerators to speed-up the process, therefore larger and complex systems, in conditions ever more realistic, are today allowed to simulate and approaching time scales that begin to be compatible with biological processes. The popularity of MD is due to the growth in the accuracy of the potential functions, which allows to directly compare the computational data with those obtained by laboratory experiments.

## Aim

The objective of my research is to address the challenges of quantifying full-length isoforms from sequencing data, particularly concerning the BRAF gene. This complexity arises from the numerous isoforms annotations that are similar and rapidly changing, as well as the variability in isoform length due to alternative polyA site usage.

To overcome these challenges, I aim to develop a customized, open-source, and reproducible bioinformatics pipeline utilizing the Snakemake management system, with the aim to analyze RNA-seq short-read sequencing data sourced from online datasets, enabling automated handling of large data volumes. Key tools such as STAR, Salmon, and StringTie will be employed to align short reads and quantify mRNA levels specifically for the BRAF-220 and BRAF-204 isoforms, among others.

Additionally, I intend to enhance the pipeline with tools that characterize BRAF expression across various RNA-seq technologies, including single-cell and Quant-seq 3'. Given the critical role of BRAF in cancer biology, I plan to correlate the isoform ratios with patient survival and hazard risk by considering online databases that integrate both short and long-read sequencing data to highlight the importance of these different technologies of reads sequencing on this bioinformatic field.

Furthermore, I aspire to identify isoform-specific signatures to provide a deeper understanding of the genetic landscape related to BRAF transcripts.

Lastly, I will investigate the structural aspects of BRAF isoforms through molecular dynamics simulations and advanced molecular modeling tools such as molecular dynamics simulations. This investigation aims to elucidate the molecular differences between the BRAF-220 and BRAF-204 proteins and analyze their interactions and behavior at the atomic level.

Overall, these efforts are made to enhance our understanding of BRAF isoforms and contribute valuable insights on their molecular mechanisms.

## Materials and Methods

### 1 Overview of RNA-seq and QuantSeq 3' mRNA-Seq Datasets from Cell Lines and Tissue Samples

The raw paired-end RNA-seq FASTQ data of 690 cancer cell lines and 12 diverse tumor types (central nervous system (n = 64), colorectal (n = 56), , haematopoietic and lymphoid (n = 176), kidney (n = 31), liver (n = 23), lung (n = 151), ovary (n = 42), pancreas (n = 41), prostate (n = 7), skin (n = 55), stomach (n = 34), and thyroid (n = 10)) were taken from the Cancer Cell Line Encyclopedia (CCLE) collection. RNA-seq data of CCLE were retrieved under the SRA ID SRP186687. The raw paired-end RNA-seq FASTQ data of 762 cancer tissues were taken from the Gene Expression Omnibus repository; the datasets are from 9 different tumor typologies: Acute Lymphoblastic Leukemia (ALL, n = 173), Acute Myeloid Leukemia (AML, n = 81), Chronic Lymphocytic Leukemia (CLL, n = 29), Chronic Myelogenous Leukemia (CML, n = 26), colorectal (n = 32), glioma (n = 274), lung (n = 42), melanoma (n = 46), and prostate (n = 59). In the order, GSE181157 (ALL), GSE63646 (AML), GSE161711 (CLL), GSE100026 and GSE144119 (CML), GSE146009 (colorectal), GSE48865 (gliomas), GSE140343 (lung), GSE91061 (melanoma), and GSE138503 (prostate). Also, BAM files of aligned RNA-seq reads of 258 cancer tissue samples and 5 different tumor types (esophagus (n = 7), kidney (n = 49), liver (n = 64), ovary (n = 60), and pancreas (n = 78)) were retrieved from the ICGC data portal. They were obtained as sliced BAM relative to the genomic region encompassing the BRAF gene (i.e., chr7:140,718,327 - 140,926,929, GRCh38) using the ICGC API for accessing the data on the web server. Raw single-end Lastly, QuantSeq 3' mRNA-Seq REV FASTQ data and the metadata of 97 lung cancer samples (ID GSE174330) were retrieved from the Gene Expression Omnibus repository.

### 2 Processing and Quantification of BRAF Expression Using the IsoWorm Pipeline

A bioinformatic pipeline called IsoWorm was designed to process FASTQ or BAM files from short-read sequencing and QuantSeq 3' mRNA-Seq REV experiments. IsoWorm is a Snakemake (v.7.0.1) workflow that includes all necessary tools, supportive scripts, and config files to automatically quantify isoforms and specifically BRAF isoforms. In more detail, raw paired-end reads, and metadata were downloaded from the SRA (CCLE and GEO samples) via the "ffq" tool

v.0.2.1, which provided a list of SRA IDs. The FASTQ files were processed with Salmon v.1.10.2 and STAR aligner v.2.7.1 using GRCh38 as the FASTA reference genome/transcriptome and a GTF file with annotations of all annotated transcripts (Ensembl 113). Salmon generates TSV files with transcript quantification data according to the provided annotation, whereas the aligned BAM files produced by STAR were further processed with StringTie v.2.2.1 via a custom GTF file that included only the genomic coordinates for the unique regions distinguishing *BRAF-204* (i.e., chr7:140725145-140719327) and *BRAF-220* (chr7:140732564--140730665). The sliced BAM files (ICGC samples) were downloaded with the ICGC score-client API v.5.8.1 and processed only with StringTie and the custom GTF files. The Quant 3' mRNA-Seq REV FASTQ data were aligned with STAR, and the obtained BAM files were processed through a custom R script via GenomicAlignments (v. 1.34.1) and bedtoolsr (v. 2.30) libraries to detect polyA sites at the 3'UTR. For the detection of the 3'UTR peak region in each sample, the following criteria were considered: a) coverage greater than 5% of the most highly expressed peak [16] and b) a region length of at least 80 bp. All the peaks were compared to each other to select only those that were present and conserved in more than 50% of the samples.

### 3 Quantification of BRAF isoform expression

Through the IsoWorm workflow, we obtained two different isoform estimations: one for all 21 annotated isoforms by Salmon and the second specifically restricted to *BRAF-204* and *BRAF-220* via the GTF custom approach. In the latter approach, we calculated the FPKM for the two regions of interest, exploiting only the information (i.e., length and coverage) available from a sliced BAM, as we did for ICGC samples. We used the following formula:

$$FPKM_i = \frac{q_i}{l_i * \sum_j q_j} * 10^9$$

Where  $q_i$  is the raw fragments count specifically mapping unique genomic regions for *BRAF-220* or *BRAF-204*,  $l_i$  is the respective genomic length, and  $\sum_j q_j$  is the total number of mapped fragments (library size) retrieved from STAR log files. The statistical significance of the differences in the expression levels of the *BRAF* transcripts was checked via the Wilcoxon Mann–Whitney test or the

t-test based on the results of the Shapiro–Wilk test for normality. All the statistical tests were conducted via R v. 4.2.1.

#### **4 Quantification of BRAF isoform expression using Long Reads and the TCGA and GTEx datasets**

Long-read RNA-seq data were obtained from the Full-Length Isoforms in Cancers and Normal Tissues Database (FLIBase, <http://www.flibase.org>) to quantify the *BRAF-220* and *BRAF-204* isoforms. The database annotates full-length splice isoforms by combining long reads for discovery and short reads on exon-exon splice junctions for validation [83], identified both *BRAF-220* and *BRAF-204* as full-length and full-splice isoforms (*TCONS\_00952581* and *TCONS\_00952571*, respectively). With FLIBase, *BRAF-220* (*TCONS\_00952581*) and *BRAF-204* (*TCONS\_00952571*) were quantified in tissue samples belonging to the 33 cancer types that compose the TCGA, 17 cancer types with paired adjacent cancer tissues and 22 not pathologic tissues from the GTEx. Following the guidelines described in [84], we focused on primary tumor samples, for which the annotation of clinical variables is extensive and highly curated (total n = 9219).

#### **5 Clinical analyses of TCGA patient**

In addition to transcript expression levels, the clinical data of TCGA patients were obtained from the FLIBase web server v.1 [83]. We selected the 11 cancer types that are characterized by the highest frequency of *BRAF* alterations [85]: *ACC-adrenocortical carcinoma*, *COAD-colon adenocarcinoma*, *ESCA-esophageal carcinoma*, *GBM-glioblastoma multiforme*, *LUAD-lung adenocarcinoma*, *KIRP-kidney renal papillary cell carcinoma*, *OV-ovarian serous cystadenocarcinoma*, *SKCM-skin cutaneous melanoma*, *TGCT-testicular germ cell tumors*, *THCA-thyroid carcinoma*, and *UCEC-uterine corpus endometrial carcinoma*. For all tumors, patients were grouped according to their *BRAF-204/BRAF-220* ratio, creating 3 groups: Group 1, with the lowest ratios ( $\leq 33$ rd percentile (1Q)); Group 2, with mid-range ratios ( $> 33$ rd percentile and  $\leq 66$ th percentile (2Q)); and Group 3, with the highest ratios ( $> 66$ th percentile (3Q)). We subsequently analyzed these 3 groups via univariate Cox proportional hazards, adding the Benjamini–Hochberg (BH) correction for multiple comparisons.

For KIRP, we also performed multivariate Cox proportional hazards regression analysis, considering tumor stage (TNM I, II, III-IV) and the individual parameters that compose it: size and extent of the primary tumor (T3-T4/T1-T2), involvement of **regional lymph nodes** (N1-N2/N0), and presence of distant metastases (M1/M0).

Univariate and multivariate Cox proportional hazards regression analyses were conducted in R v. 4.2.1 using the *survminer* 0.4.9 and *survival* v. 3.5 Bioconductor packages.

To evaluate whether having a higher ratio increases the probability of being at a higher tumoral stage, we performed a proportional odds logistic regression for ordered category outcome analysis, considering stage I KIRP patients as a reference and using tumor stage (I < II < III-IV) as a response variable. The latter studies were conducted via the *MASS* v. 7.3.60 package in R. The Brant–Walld test was used to evaluate the model.

## **6 Differential Expression and Gene Set Enrichment Analysis, Gene Set Variation Analysis, and Gene Set Variation Analysis using the Molecular Signatures Database of KIRP genes**

The patients were divided again according to their *BRAF-204/BRAF-220* ratio, but into two groups: Group 1, with the lowest ratios (= 25th percentile (1Q)), and Group 3, with the highest ratios ( $\geq$ 75th percentile (2Q)). Subsequently, these two groups were analyzed via differential expression analysis performed in the R environment using Bioconductor's package DESeq2, version 1.44.0. A total of 852 upregulated genes with a log2 fold change  $> 1$  and 469 downregulated genes with a log2 fold change  $< 1$  were identified, all with an adjusted pvalue  $< 0.05$ . The complete list of 34764 genes was used as input for the Gene Set Enrichment Analysis. The same gene list was used for Gene Set Variation Analysis (GSVA) with the Molecular Signatures Database (MSigDB). The MSigDB gene set was downloaded from the C3 repository, specifically the "AP2ALPHA\_01" set, which includes genes regulated by the AP2ALPHA protein.

## 7 miRNA targets cross-reference study

The microRNAs (miRNAs) identified in the miR-CATCHv2.0 study [17] were cross-referenced with miRNAs validated in the miRBase database, specifically focusing on those known to target upregulated genes. Only 1 microRNA, the miR-320a, has been identified as targeting both the *BRAF-204* transcript and one of the upregulated genes, *TFAP2A*, a gene with a log2FC 2,07. To validate that miR-320a binds specifically only to the 3'UTR of *BRAF-204*, RNA22 v2 microRNA target detection was used.

## 8 Survival Analysis on TFAP2A gene expression

The patients were stratified into two groups based on gene expression levels. Group 1 included patients with the lowest gene expression ( $\leq$  25th percentile, 1Q), while Group 2 included those with the highest gene expression ( $\geq$  75th percentile, 3Q). Survival analysis was then conducted to compare these groups with the generated Kaplan-Meier survival curves. The study was performed in the R environment using the survminer (v. 0.4.9) and the survival packages (v. 3.5) from Bioconductor. The expression values were also included as a covariate in the previously presented multivariate model.

## 9 Visualization of transcript expression and ratios

All the plots of the expression profiles of the isoforms and their ratios were calculated and plotted via the R packages: data.table (v. 1.14.2), dplyr (v. 2.2.1), ggplot2 (v. 3.3.6), tidyr (v. 1.2.0), ggsignif (v. 0.6.4), and patchwork (v. 1.1.2). The TPM and FPKM values are represented in the plots as  $\log_2(\text{TPM} + 0.01)$  and  $\log_2(\text{FPKM} + 0.01)$ , respectively.

## 10 Single cell module

A module for single-cell analysis has been integrated into the IsoWorm pipeline to enable gene quantification from short-read RNA-seq data. The raw scRNA-seq FASTQ data of two tumor typologies (SKCM, GSE210963 and KIRP: GSE152938) and generated with the 10X Genomics kit, were retrieved from the Gene Expression Omnibus repository (downloaded with the ffq module of IsoWorm). All the analysis was performed with kallisto | bustools (v 0.51.1). The quality control

steps included filtering cells and genes based on expression thresholds was done, ensuring uniqueness of variable names, and annotation of mitochondrial content. Dimensionality reduction was performed using truncated singular value decomposition (TruncatedSVD) to project the high-dimensional gene expression data into a two-dimensional space. This was visualized using scatter plots generated with matplotlib (v. 3.9.2). Library saturation was assessed by plotting the number of genes detected as a function of unique molecular identifier (UMI) counts. Thresholds for cell inclusion were determined using a knee plot, resulting in the exclusion of low-quality cells. Normalization was performed using scanpy's (v. 1.10.3) `normalize_per_cell` method, followed by log-transformation of the counts. Highly variable genes were identified using the `sc.pp.highly_variable_genes` function with the "seurat" flavor, and data was scaled accordingly. Cells were clustered using the Louvain (v. 0.8.2) algorithm. Principal component analysis (PCA) was conducted to reduce the data to six principal components, which were used for neighborhood graph construction and clustering. The clustering results were visualized using UMAP plots, colored by Louvain clusters.

## **11 Preparation of the molecular dynamic simulation with AlphaFold**

The structure of the human BRAF protein (PDB id: 6UAN, [86] ) were taken from RCSB Protein Data Bank (RCSB PDB) and used as starting coordinates for modelling the C-term of BRAF-204 and BRAF-220. The Structural models of the BRAF-220 and BRAF-204 full sequence, BRAF-220 and BRAF-204 C-terms (-PSLNRAGFQTEDFSLYACASPKTPIQAGGYGAFPVH, -PSLNRAGFQTEDFSLYACASPKTPIQAGGYGEFAAFK) were built using AlphaFold 3 Webserver [87]. Each C-terms peptide was capped with an acetyl group at the N-term (ACE).

## **12 Molecular Dynamics Simulation**

All the molecular dynamics (MD) simulations were performed by using Gromacs2023 [88] with all-atom AMBER03 force field [89] on the Leonardo Cineca High Performance Computing infrastructure [90]. The proteins (C-term BRAF-204 and C-term BRAF-220) were placed in a cubic box and filled with TIP3P water model [91]. All the systems were neutralized using NaCl at the physiological concentration of 150 mM. Periodic boundary conditions were applied to avoid edge effects. The systems were energy minimized by steepest descent (SD) [92] heated and equilibrated

for 100 ps. MD simulations were carried out in the NPT ensemble for 100 ps. The v-rescale algorithm [93] was used to keep the temperature at 310 K. The average pressure was kept at 1 bar ( $\tau_p = 2$  ps) by using the isotropic Parrinello-Rahman barostat [94]. The particle mesh Ewald (PME) method was used to compute the long-range electrostatics and Van der Waals interactions with a cut-off of 1.2 nm. An equal cut-off was used for short-range electrostatics and Van der Waals interactions. Rotation and translational motions of systems were removed, and all bonds were constrained with the LINCS algorithm [95]. Initial velocities were assigned according to Maxwell-Boltzmann distribution.

The trajectories were visualized and analysed with VMD [96] and with tools included in Gromacs. Root Mean Square Deviation (RMSD) of atomic positions in the simulation run have been calculated according to:

$$\Delta R \left[ \frac{1}{N_{at}} \sum_{i=1}^{N_{at}} \langle \{ (\Delta x_i)^2 (\Delta y_i)^2 (\Delta z_i)^2 \} \rangle \right]^2$$

Where  $N_{at}$  is the total number of atoms and the brackets  $\langle \dots \rangle$  represent a time average and  $(\Delta x_i)^2 (\Delta y_i)^2 (\Delta z_i)^2$  are the differences between the instantaneous and the starting atomic coordinates for the atoms.

The *radius of gyration* of the end-to-end distance [97] was calculated with the program gmx gyrate, as follows:

$$R_g = \left( \frac{\sum_i m_i |r_i|^2}{\sum_i m_i} \right)^{\frac{1}{2}}$$

Where  $m_i$  is the mass of atom  $i$  and  $r_i$  is the position of atom  $i$  with respect to the center of the mass of the molecule.

## Results and discussion

### 1 BRAF-ref/220 and BRAF-X1/204 isoforms in publicly available repositories of RNA and protein sequences

As mentioned in the Introduction, the proteins encoded by *BRAF-ref* and *BRAF-X1* differ in the C-term (ref: -GYG**A**FPV**H** vs X1: -GYG**E**F**A**AF**K**), whereas the 3'UTRs differ not only in length but also in sequence. This is because the ref transcript is composed of 18 exons, whereas the X1 transcript is a splicing variant encompassing a shorter version of exon 18 that is joined with a very long downstream 19<sup>th</sup> exon (**Figure 10**) [14] [15] [98]. Interestingly, in melanoma samples, the 19<sup>th</sup> exon is included in *BRAF* gene amplification [99], as well as in the duplicated kinase domain that confers acquired resistance to BRAF/MEK inhibitors [100].

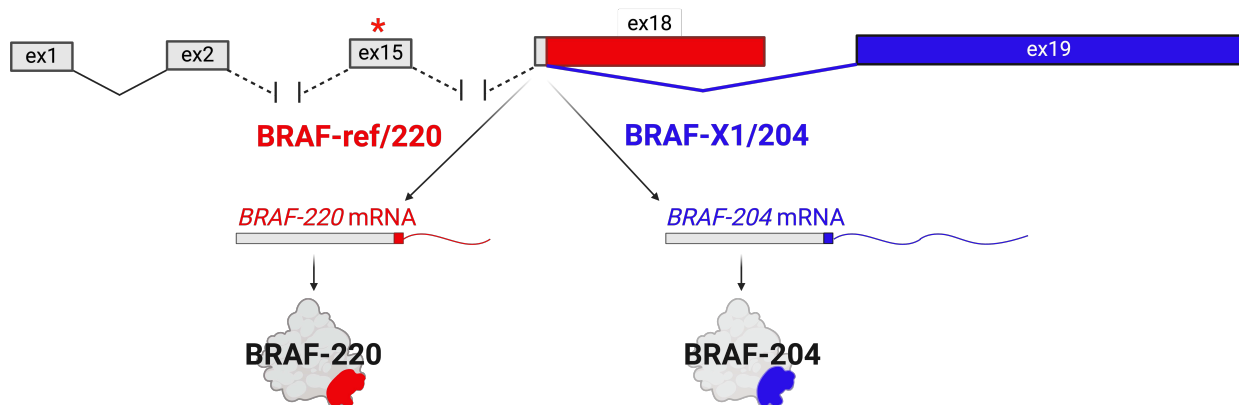


**Figure 10** Cartoon summarizing the similarities and differences between *BRAF-ref/220* (red) and *BRAF-X1/204* (blue) RNA and protein isoforms.

See the text for details.

I tracked IDs and associated sequence details for the *BRAF-ref* and *X1* transcripts from 2017 to the latest release of the NCBI (GRCh38.p14) and Ensembl (113) databases. Currently, NCBI lists 25 *BRAF* transcripts, whereas Ensembl lists 21 *BRAF* transcripts.

*BRAF-ref* corresponds to ID *NM\_004333.6* present in NCBI and ID *ENST00000646891.2/BRAF-220* present in Ensembl. The two sequences are identical and converge into a high-quality annotation transcript. In the MANE (Matched Annotation from the NCBi and EMBL-EBI) database (v. 1.3), the

*BRAF*-220 ID is recognized as the “MANE Select” transcript for the *BRAF* gene [101]. Compared with 2017, *NM\_004333.6* and *ENST00000646891.2/BRAF-220* present longer 5′ and 3′ untranslated regions, such that the nominal length of the transcript exceeded 6 kb (**Figure 11 up**).

*BRAF*-X1 corresponds to “*BRAF* transcript variant 2” present in NCBI with ID *NM\_001354609.2*, while the matching Ensembl ID is *ENST00000496384.7/BRAF-204*. Compared with 2017, these two annotations reveal slightly longer 5′ and 3′UTRs (**Figure 11 down**).

**BRAF-220 ID History**

| NCBI ID (v. GRCh38.p1)<br><i>BRAF-ref</i> ( <i>NM_04333.4</i> )                                                         | Ensembl ID (v. 80)<br><i>BRAF-001</i> ( <i>ENST00000288602</i> )                                                        | NCBI ID (v. GRCh38.p14)<br><i>BRAF</i> transcript variant 1<br>( <i>NM_04333.6</i> )                                      | Ensembl ID (v. 111)<br><i>BRAF-220</i> ( <i>ENST00000646891.2</i> )                                                       |
|-------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| Sequence length: 2480nt;<br>Protein length: 766aa;<br>5'UTR: 61nt;<br>CDS: 2301nt;<br>3'UTR: 121nt;<br>Exons Number: 18 | Sequence length: 2949nt;<br>Protein length: 766aa;<br>5'UTR: 61nt;<br>CDS: 2301nt;<br>3'UTR: 590nt;<br>Exons Number: 18 | Sequence length: 6459nt;<br>Protein length: 766aa;<br>5'UTR: 226nt;<br>CDS: 2301nt;<br>3'UTR: 3932nt;<br>Exons Number: 18 | Sequence length: 6459nt;<br>Protein length: 766aa;<br>5'UTR: 226nt;<br>CDS: 2301nt;<br>3'UTR: 3932nt;<br>Exons Number: 18 |
| 2017 ( <i>Marranci et al.</i> )                                                                                         |                                                                                                                         | 2024 (present)                                                                                                            |                                                                                                                           |

**BRAF-204 ID History**

| NCBI ID (v. GRCh38.p1)<br><i>BRAF-X1</i> ( <i>XM_005250045.1</i> )                                                      | NCBI ID (v. GRCh38.p14)<br><i>BRAF</i> transcript variant 2<br>( <i>NM_001354609.2</i> )                                  | Ensembl ID (v. 111)<br><i>BRAF-204</i> ( <i>ENST00000496384.7</i> )                                                       |
|-------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| Sequence length: 9472nt;<br>Protein length: 767aa;<br>5'UTR: 8nt;<br>CDS: 2304nt;<br>3'UTR: 7160nt;<br>Exons Number: 19 | Sequence length: 9687nt;<br>Protein length: 767aa;<br>5'UTR: 215nt;<br>CDS: 2304nt;<br>3'UTR: 7167nt;<br>Exons Number: 19 | Sequence length: 9578nt;<br>Protein length: 767aa;<br>5'UTR: 107nt;<br>CDS: 2304nt;<br>3'UTR: 7167nt;<br>Exons Number: 19 |
| 2015 ( <i>2017 Marranci et al.</i> )                                                                                    | 2024 (present)                                                                                                            |                                                                                                                           |

**Figure 11** Tracking of *BRAF*-220 annotation (a) and *BRAF*-204 annotation (b) from 2017 [15] to date.

Notably, *BRAF*-X2 (*XM\_005250046.1* in NCBI (GRCh38.p1)) is no longer present (record removed). In compliance with the most updated official nomenclature, from this point onward, I will refer to *BRAF-ref* as *BRAF*-220 and to *BRAF-X1* as *BRAF*-204.

A consistent annotation is not the only issue related to mRNA transcript variants because for many transcripts the evidence of translation is very limited. The APPRIS database [102], which was developed within the GENCODE consortium and is used by GENCODE curators to update gene models, establishes principal isoforms via protein structural information, functional features, and cross-species conservation. Current release of the db (2024\_03v.48 Gencode45/Ensembl113) contains 11 transcript variants for *BRAF*. None of them are selected by the APPRIS core modules as clear principal isoforms, and none are associated with strong proteomics. Nine transcript variants are classified as *MINORS*, with a very low chance of being translated into (functional) proteins. The

remaining two are *BRAF-220*, which is classified as *PRINCIPAL:4* and has the highest TRIFID functional score (= 1), and *BRAF-204*, which is classified as *ALTERNATIVE:1*. This is because it is flagged as TSL = 5 (i.e., without a single transcript that supports the model structure). However, *BRAF-204* has a remarkably high TRIFID score (0.993), which in turn indicates that its likelihood of being functionally relevant and under selective pressure is very high.

The recently published FLIBase (Full Length Isoforms in cancers and normal tissues database, [77]), which annotates full-length splice transcript by combining long reads for discovery and short reads on exon–exon splice junctions for validation [83], identified both *BRAF-220* and *BRAF-204* as full-length and full-splice isoforms (*TCONS\_00952581* and *TCONS\_00952571*, respectively).

Finally, proteomic evidence, curated by CNIO and collected from the APPRIS database [102], confirms that isoform-specific C-terminal peptides were retrieved for both *BRAF-220* (753-767aa -TPIQAGGYGAFPVH (3 times)) and *BRAF-204* (753-767aa -TPIQAGGYGEFAAFK (15 times)). These findings indicate that both *BRAF-220* and *BRAF-204* are translated into proteins.

It is conceivable that *TCONS\_00952571* will provide the experimental support necessary to increase the TSL of *BRAF-204* in the future releases of the NCBI, Ensembl, MANE, and APPRIS databases. In this way, *BRAF* is fully acknowledged as a mix of *BRAF-220* and *BRAF-204* mRNA isoforms, which are both translated into functional proteins and play a biological role.

## 2 Tools for the quantification of *BRAF* transcript variants

Even though the official annotation for *BRAF* transcript variants has been updated (see above), the precomputed quantification reported in various databases (XENA, <https://xenabrowser.net/>; GEPIA2, <http://gepia2.cancer-pku.cn/#index>; SRTdb, <http://www.shengli labs.com/SRTdb/>) is still based on an older Ensembl release (version 40). Specifically, *BRAF-220* is present as the old *ENST00000288602/BRAF-001* transcript, whereas *BRAF-204* is missing.

To quantify the *BRAF-220* and *BRAF-204* transcript variants in the TCGA datasets, I used the abovementioned FLIBase [83]. By combining datasets produced by both short- and long-read technologies, the effort of Shi et al. allows the retrieval of a comprehensive collection of full-length transcripts and their expression levels in both healthy and cancerous tissues. Specifically, a total of 43 *BRAF* transcript variants were reconstructed, most of which are *de novo* and noncoding. As mentioned above, both *BRAF-220* (*TCONS\_00952581*) and *BRAF-204* (*TCONS\_00952571*) are retrieved as full-length and full-splice isoforms.

To quantify *BRAF* transcript variants in additional cancer-related RNA-seq datasets, I focused on the 21 transcripts reported in Ensembl (113) because they are routinely updated within the GENCODE project, and I built an *ad hoc* bioinformatic pipeline called IsoWorm. IsoWorm (**Figure 12**) integrates two distinct methods: Salmon, a *state-of-the-art* tool for transcript variant quantification, and a custom approach designed to specifically distinguish between the *BRAF-220* and *BRAF-204* variants. Salmon quantifies transcripts via a quasimapping approach and probabilistic modeling. Before Salmon was integrated into the pipeline to quantify *BRAF* isoforms from short reads, another tool, Kallisto, was used for the same purpose. However, Kallisto was eventually replaced by Salmon due to its superior performance, as supported by literature [103]. With respect to the custom approach, I take advantage of two manually selected regions that are variant specific. In particular, the region distinctive for *BRAF-220* corresponds to the most downstream ~1kb out of the ~3.9kb long 3'UTR, while the region distinctive for *BRAF-204* corresponds to the most downstream ~4kb out of the ~7kb long 3'UTR (green boxes in **Figure 13**). I then used custom GTF files encompassing the genomic coordinates of those regions. Following the read mapping step, the expression level of full-length transcripts is estimated considering the library size of the dataset under study and the size of the selected regions. The advantages of IsoWorm include easy customization to quantify the transcript variants of interest, versatility in dealing with diverse input file formats from RNAseq experiments, acceptance of raw FASTQ files or alignment bam files, and acquisition from the full experiment or a partial subset. Owing to the implementation of Salmon, and Kallisto before, the expression level of all the known transcripts of the gene of interest can be calculated so that they can be ranked from the highest to the lowest expression. Owing to the implemented custom approach, IsoWorm estimates transcript expression from sliced bam files or partial FASTQ files, favoring faster and lighter quantification of the transcripts of interest. However, transcripts are quantified in FPKM, which allows only intrasample comparisons [104]. IsoWorm is publicly available at <https://github.com/ctglab/isoworm>.

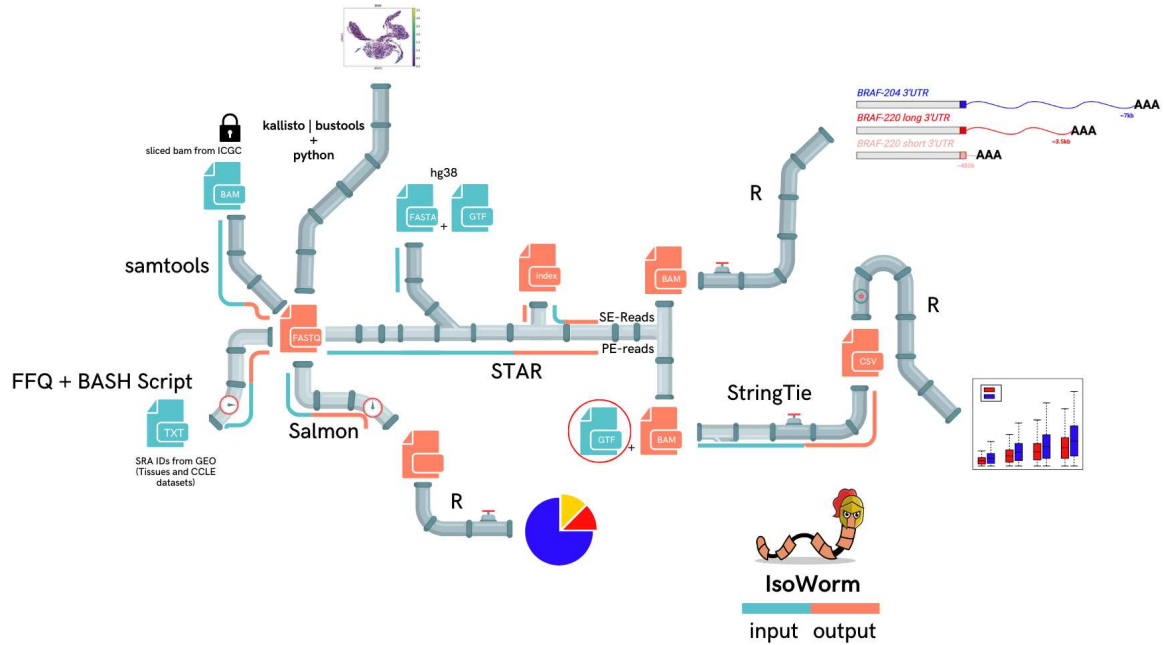


**Figure 12** Scheme of IsoWorm.

The snakemake pipeline that we developed to quantify *BRAF-220* and *BRAF-204* expression levels in large RNA-seq datasets (both single- and paired-end short reads). The pipeline consists of a series of interconnected modules that perform various stages of data analysis. It starts with a .txt file containing SRA IDs, while the indications about the RNAseq library type, or a BAM file, and the reference files (FASTA and GTF) are in the snakemake config file. The custom module of IsoWorm was used to analyze the *BRAF-204* and *BRAF-220* isoforms specifically, and custom gtf files (red circles) were used to quantify isoform-specific genomic regions. The quantification was performed via Stringtie, and all the plots were generated with R language. Conversely, the salmon module of IsoWorm was used to quantify the 21 isoforms annotated in Ensembl db (113). An R script generates pie charts for *BRAF* isoform expression. For single-end reads, a module identifies polyA sites via custom R scripts. The module of single cell was used to calculate and cluster cell line based on gene expression, the plot are made though python while all the filtering and analysis was made with kallisto | bustools.

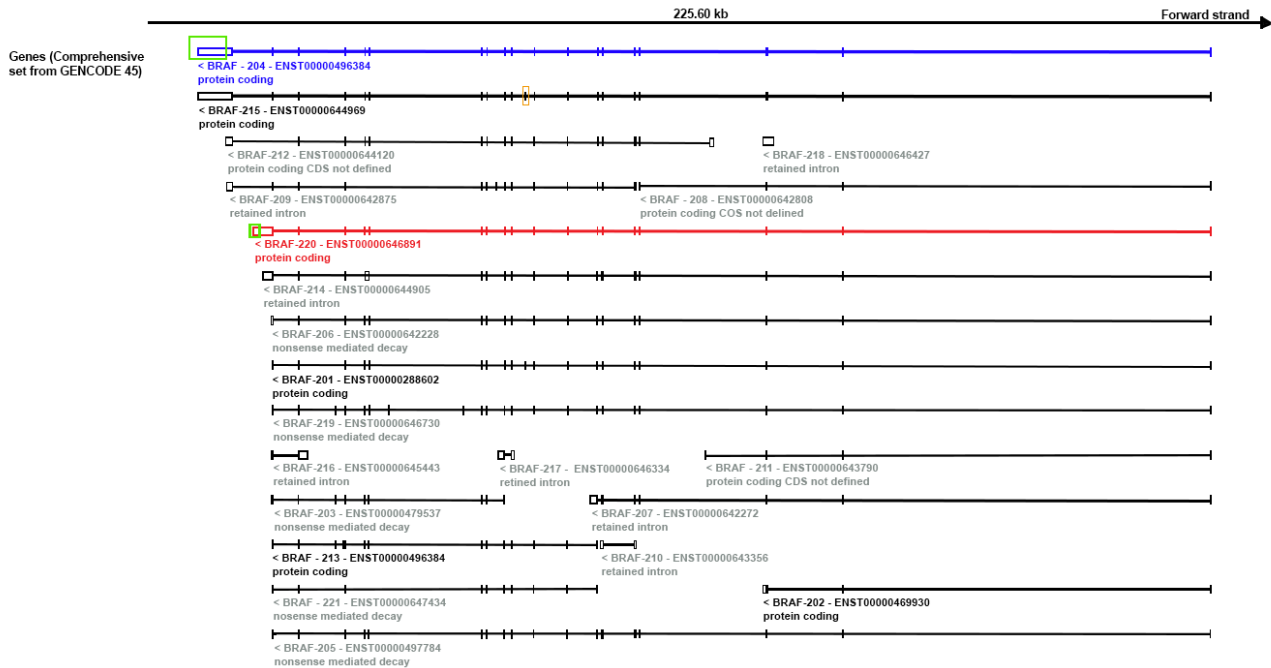


**Figure 13** The 21 *BRAF* transcripts deposited in Ensembl (release 113) and isoform specific regions used by the custom module of IsoWorm for *BRAF-204* and *BRAF-220* quantification.

*BRAF* gene is located on chromosome 7 and is transcribed from the reverse strand. In Ensembl (113) 5 *BRAF* transcripts are deposited as protein-coding, 3 as protein-coding with CDS not defined, 7 as containing retained introns, and 6 as undergoing nonsense-mediated decay. *BRAF-204* is highlighted in blue, while *BRAF-220* is highlighted in red. For the quantification of these two isoforms, the custom module of IsoWorm uses the isoform-specific regions boxed in green. Of note, there is only one exon that distinguishes *BRAF-204* from *BRAF-215* (yellow box). However, according to the Salmon module of IsoWorm, *BRAF-215* is expressed at very low levels in the cancer cell lines and tissues analyzed

### 3 *BRAF-204* is the most expressed *BRAF* transcript variant across cancer cell lines and tissues

IsoWorm was used to quantify *BRAF* transcript variants in 690 CCLE cell lines belonging to 12 cancer types: central nervous system (CNS), colorectal, haematopoietic and lymphoid, kidney, lung, liver, ovary, pancreas, prostate, skin, stomach, and thyroid. I used the Salmon module to calculate the expression levels of all 21 *BRAF* transcript variants reported in Ensembl (release 113). The pie charts of variant expression illustrate that *BRAF-204* is the most highly expressed variant in cancer cell lines, accounting for  $53.6 \pm 4.5\%$  of all *BRAF* transcripts, whereas *BRAF-220* variant accounts for a much smaller fraction ( $2.5 \pm 0.9\%$ , **Figure 14a**). As a direct consequence, *BRAF-204* is 16 to 41 times more highly expressed than *BRAF-220* (**Figure 15**).

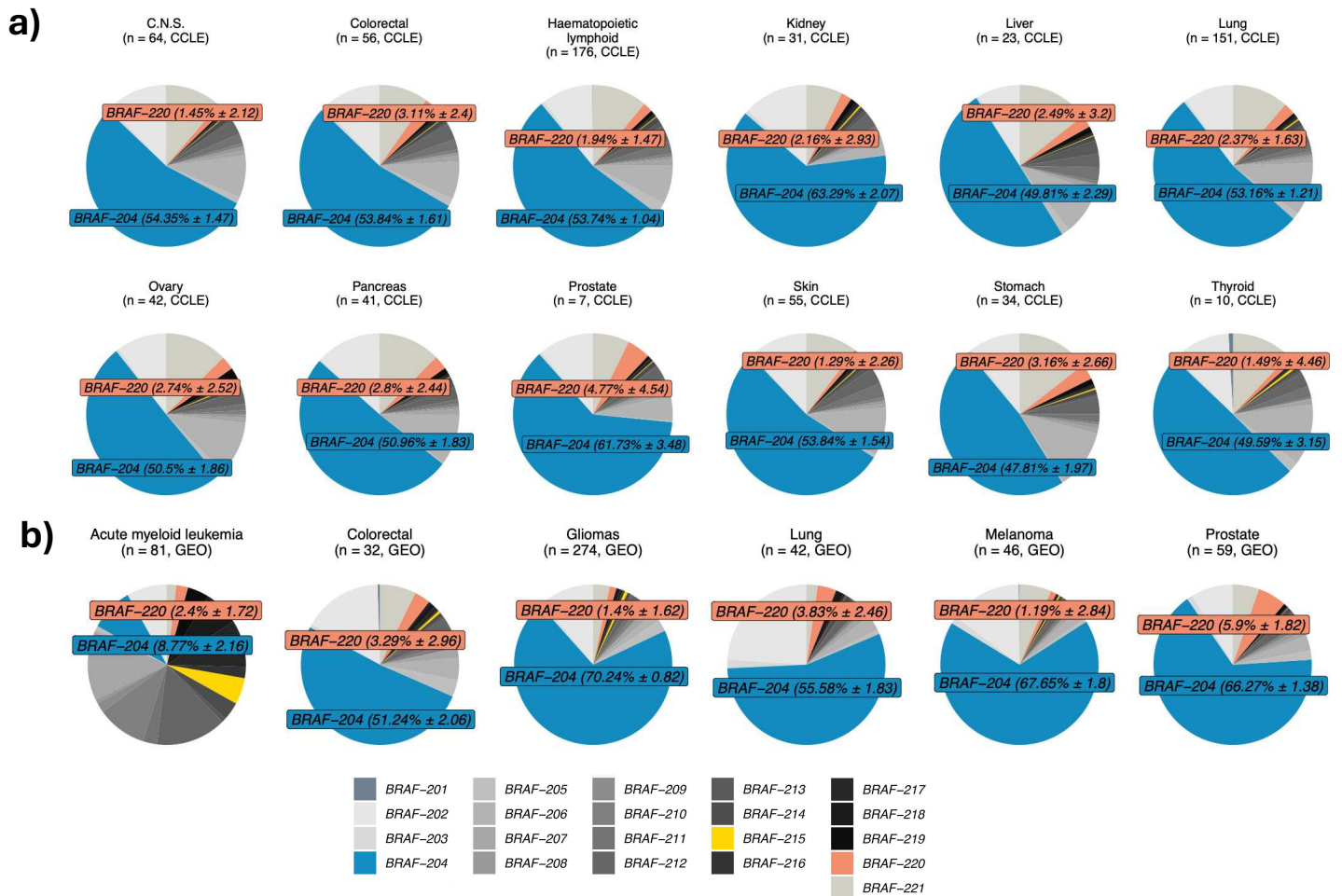


**Figure 14** Pie charts showing that BRAF-204 is by far the most abundant of the 21 isoforms of BRAF present in Ensembl (113)

Pie charts showing the 21 transcript variants of BRAF present in Ensembl (111) in both cell lines (**a**) and tissue samples (**b**). The mean relative abundance (%) of the 21 BRAF variants was estimated via the salmon module of IsoWorm in 690 CCLE cell lines belonging to 12 cancer types and in 534 GEO tissue samples belonging to 6 cancer types. BRAF-204 is highlighted in blue, whereas BRAF-220 is highlighted in red. Mean  $\pm$ SD values are indicated.

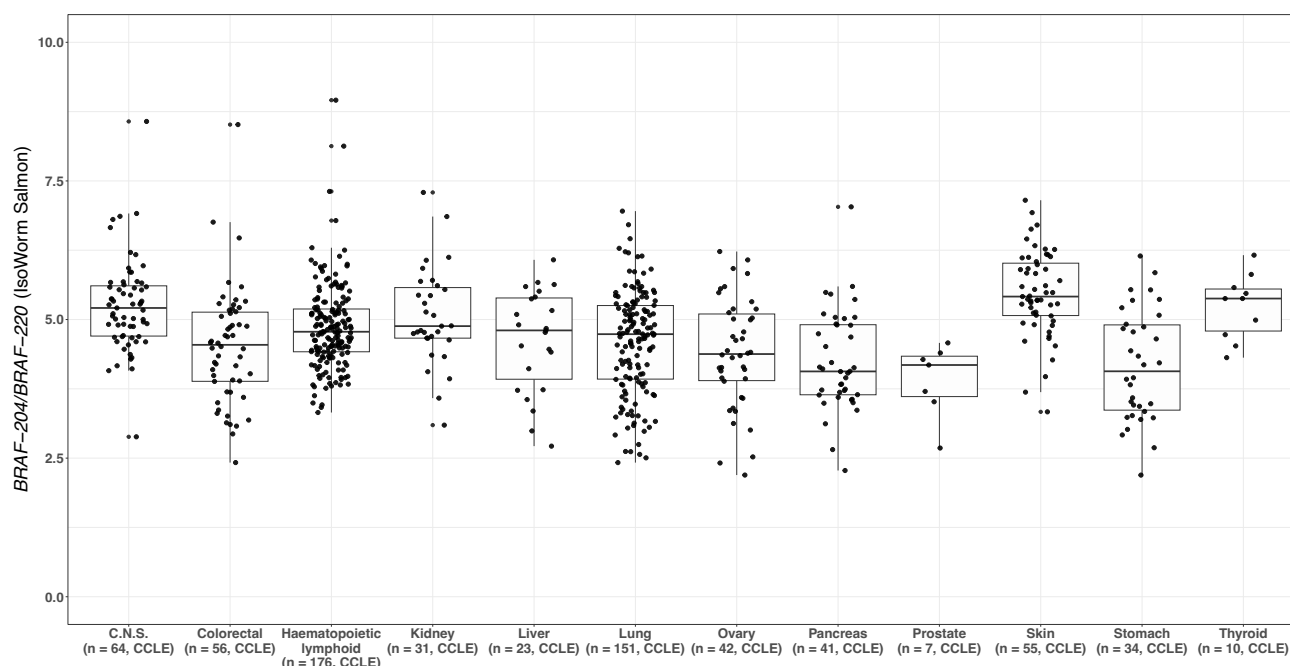


**Figure 15** Box plots of *BRAF-204/BRAF-220* ratio (Salmon).

Quantification of *BRAF-204* and *BRAF-220* isoforms in CCL, followed by the calculation of their ratio, was made using the Salmon module of IsoWorm. Ratios are expressed as  $\log_2(BRAF-204\_TPM + 0.01/BRAF-220\_TPM + 0.01)$ . They are consistently higher than 1

The expression levels of *BRAF-220* and *BRAF-204* were also calculated via the custom module of IsoWorm. We confirmed that the expression of *BRAF-204* is 20–40 times greater than that of *BRAF-220* (**Figure 15**). Overall, the *BRAF-204/BRAF-220* median ratios estimated by the two modules of IsoWorm were highly correlated (Spearman correlation coefficient greater than 0.79 for all datasets, except for thyroid cell lines ( $n = 10$ ), **Figure 16**). This attests to the robustness of the custom module of IsoWorm.

IsoWorm was also used to quantify *BRAF* isoforms in 711 tissue samples belonging to 11 cancer types. They were obtained from GEO (acute myeloid leukemia, colorectal cancer, glioma, lung cancer, melanoma, and prostate cancer) and ICGC (esophageal cancer, kidney cancer, liver cancer, ovarian cancer and pancreatic cancer) databases.

The GEO datasets were analyzed via both quantification modules of IsoWorm. According to the Salmon module, *BRAF-204* is the most expressed isoform in cancer tissues, accounting for accounts for  $53.3 \pm 23\%$  of all *BRAF* transcripts, whereas *BRAF-220* accounts for a smaller fraction ( $3 \pm 1.6\%$ , **Figure 14b**). In terms of the *BRAF-204/BRAF-220* ratio, the expression of *BRAF-204* is 4 to 60 times greater than that of *BRAF-220* (**Figure 18**). Similarly, the custom module of IsoWorm estimates that

the expression of *BRAF-204* is 5-75 times greater than that of *BRAF-220* (**Figure 19, left**). The correlation between the calculations is very high (the Spearman correlation of the median ratios is greater than 0.8 for all the datasets, **Figure 20**).

The ICGC datasets were analyzed only with the custom module of IsoWorm. This is because Salmon requires the size of the entire library and the expression level of each transcript, whereas for the ICGC datasets we could access only the sliced BAM files containing reads from *BRAF* transcripts. Also in this case, *BRAF-204* was expressed at a level that was 6 to 30 times greater than that of *BRAF-220* (**Figure 19, right**).

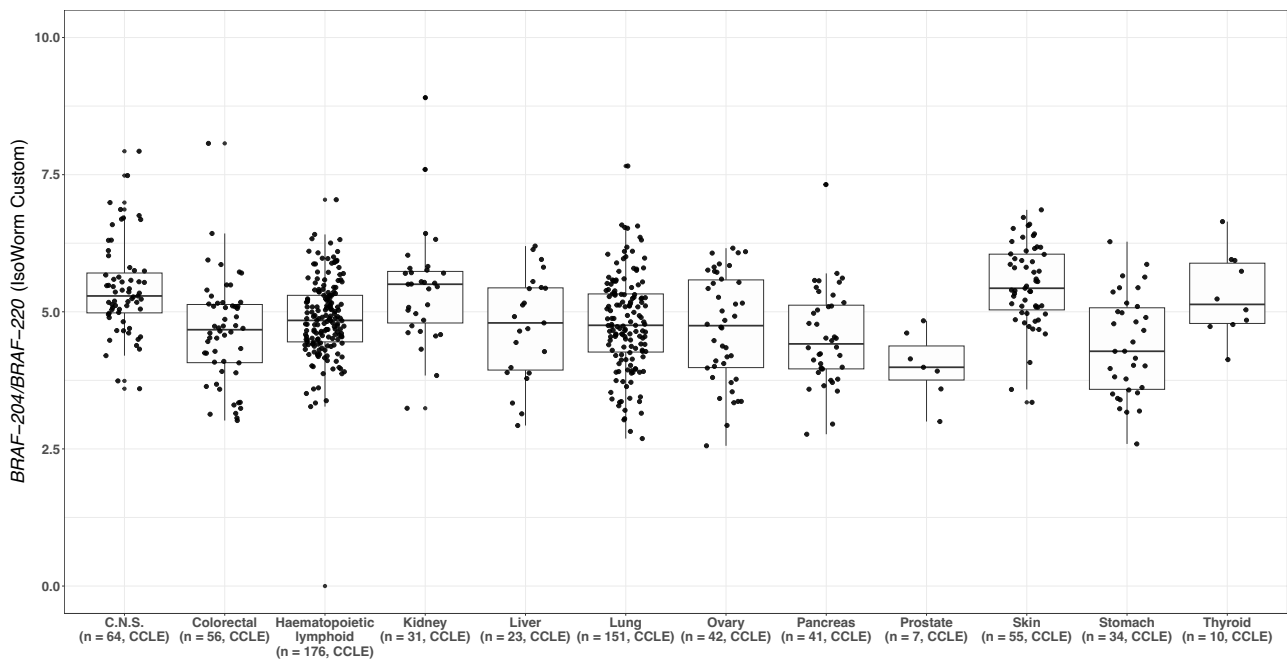


**Figure 16** Box plots of the *BRAF-204*/*BRAF-220* ratio in cancer cell lines, CCLE (custom module)

Ratios are consistently greater than 1. Results obtained from CCLE cell lines are presented (total  $n = 514$ , 11 cancer types). Quantification of *BRAF-204* and *BRAF-220* isoforms and the calculation of their ratio were obtained using the custom module of IsoWorm. The ratios are expressed as  $\log_2 (BRAF-204\_FPKM + 0.01/BRAF-220\_FPKM + 0.01)$ .

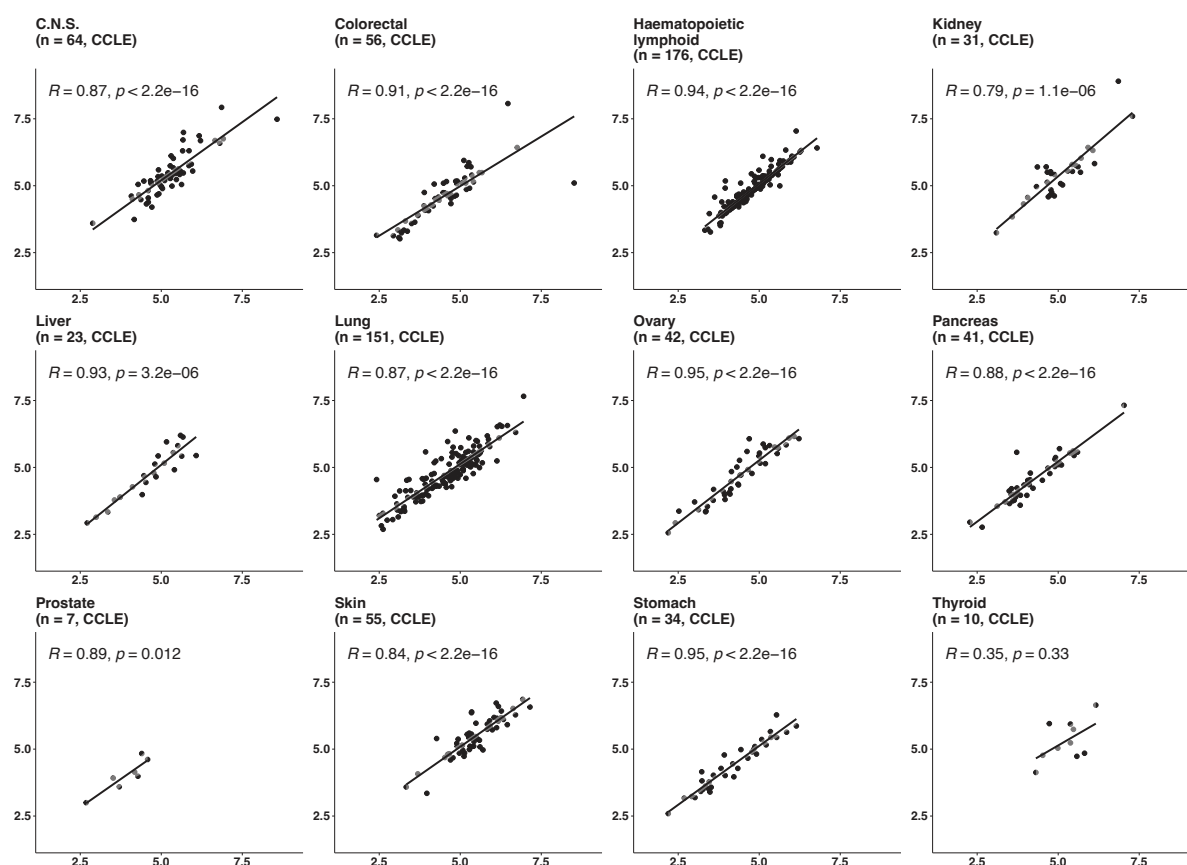


**Figure 17** Spearman test to assess the correlation between the quantification of BRAF-204/BRAF-220 ratio obtained with the Salmon module (y axis) and the custom module (x axis) of IsoWorm.

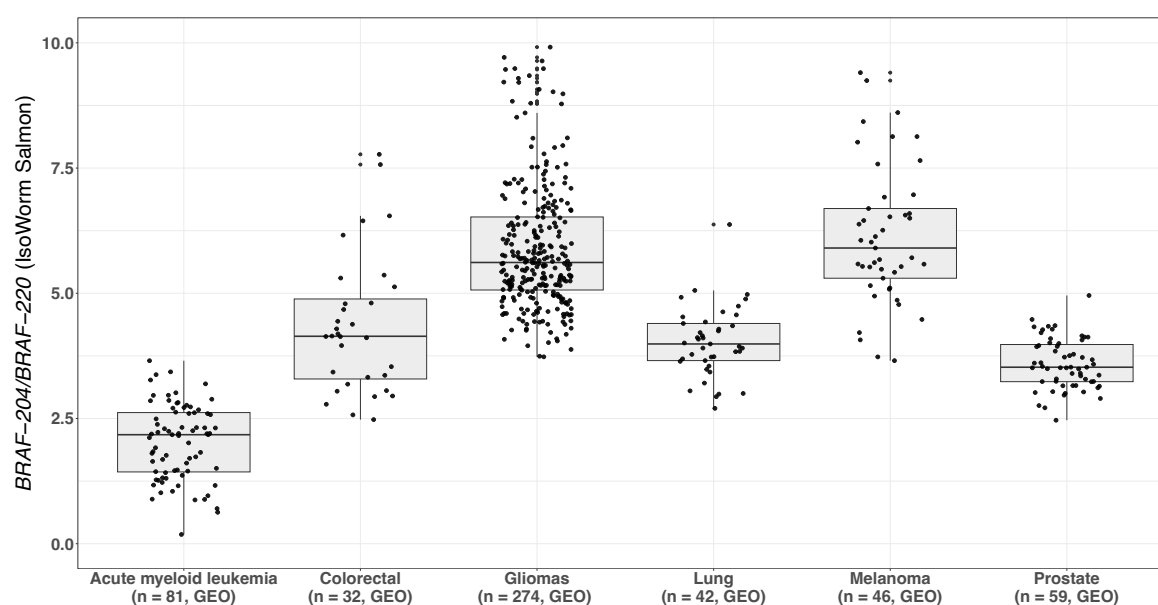


**Figure 18** Box plots of BRAF-204/BRAF-220 ratio tissues, GEO, Salmon

Quantification of *BRAF-204* and *BRAF-220* isoforms, followed by the calculation of their ratio, was made using the Salmon module of IsoWorm. Ratios are expressed as  $\log_2(BRAF-204\_TPM + 0.01/BRAF-220\_TPM + 0.01)$ . They are consistently higher than 1.

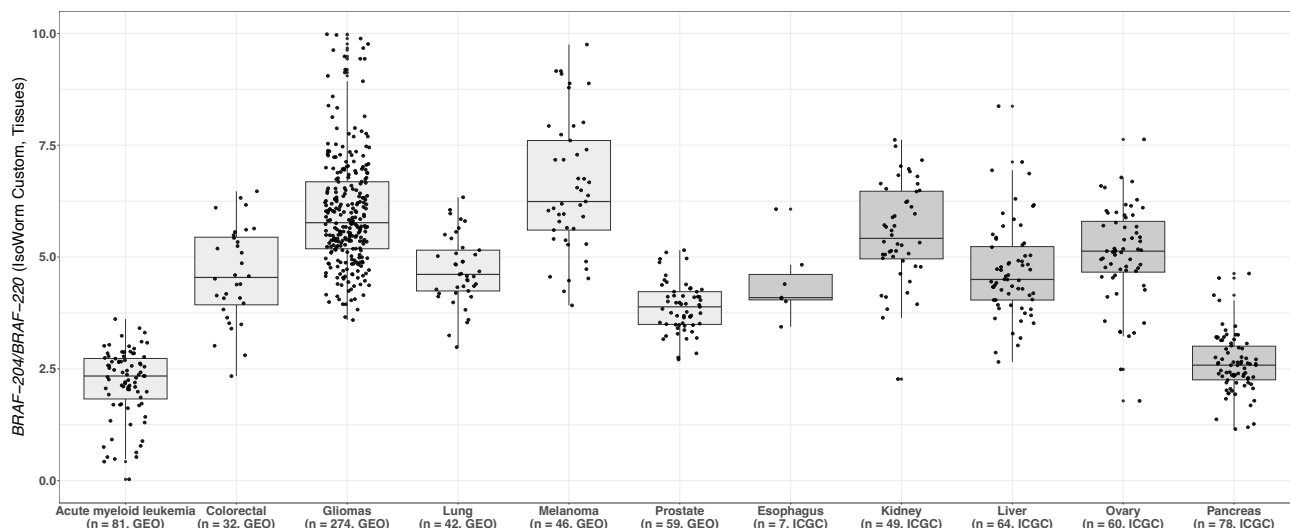


**Figure 19** Box plots of the BRAF-204/BRAF-220 ratio tissues, GEO (custom)

Ratios are consistently greater than 1 in tissue samples. Results obtained from the GEO datasets are presented on the left (total  $n = 536$ , 6 cancer types), whereas the results obtained from the ICGC datasets are presented on the right (total  $n = 258$ , 5 cancer types). Quantification of *BRAF-204* and *BRAF-220* isoforms and the calculation of their ratio were obtained using the custom module of IsoWorm. The ratios are expressed as  $\log_2(BRAF-204\_FPKM + 0.01/BRAF-220\_FPKM + 0.01)$ . These values are consistently greater than 1.

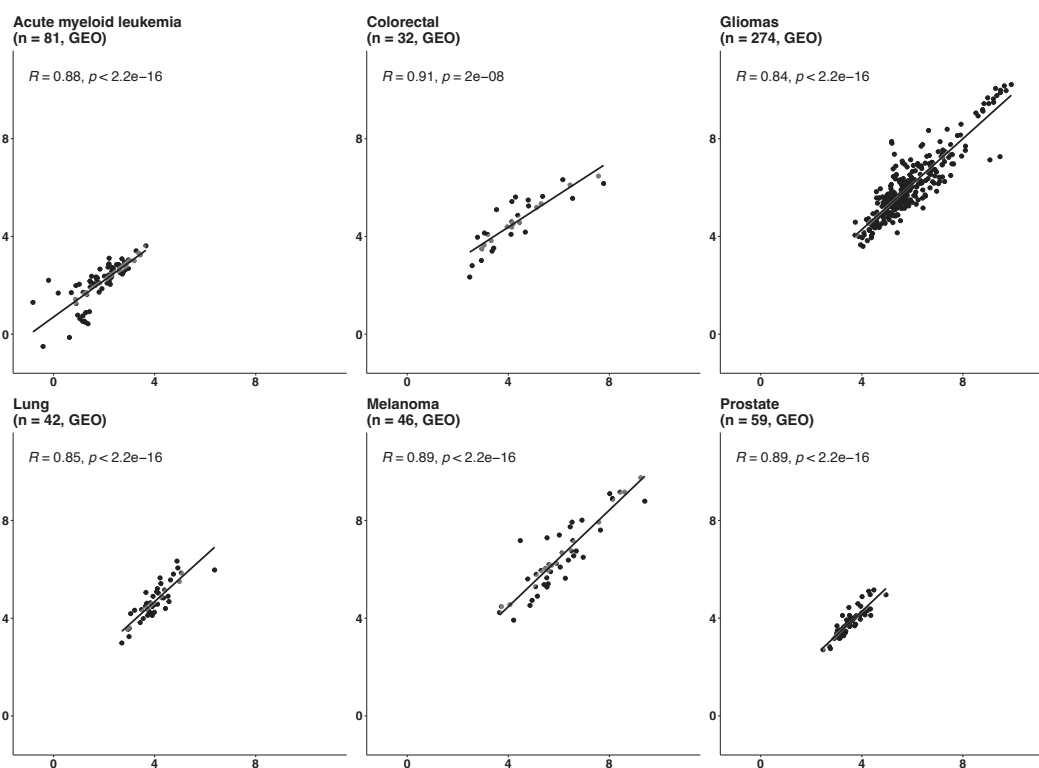


**Figure 20** Spearman test to assess the correlation between the quantification of *BRAF-204/BRAF-220* ratio obtained with the Salmon module (y axis) and the custom module (x axis) of IsoWorm, GEO.

With FLIBase, I quantified *BRAF-220* (TCONS\_00952581) and *BRAF-204* (TCONS\_00952571) in tissue samples belonging to the 33 cancer types that compose the TCGA. Following the guidelines described in [84], we focused on primary tumor samples, for which the annotation of clinical variables is extensive and highly curated (total n = 9219). The results reported in **Figure 21** highlight the widespread predominance of *BRAF-204* over *BRAF-220*, with a median ratio ranging between ~1.5 and ~45.

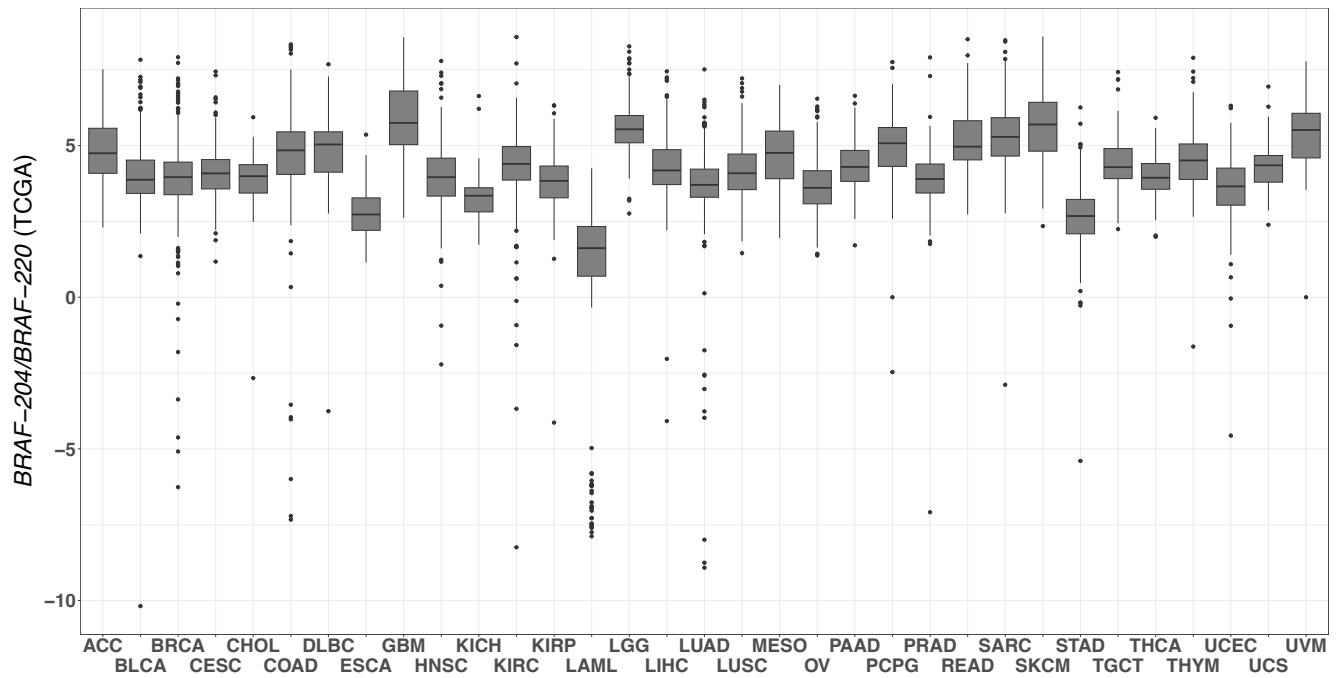


**Figure 21** Box plots of the BRAF-204/BRAF-220 ratio, TCGA.

The results obtained are presented (total n = 9219, 33 cancer types). The quantification of the *BRAF-204* and *BRAF-220* isoforms was obtained via FLIBase, and the *BRAF-204/BRAF-220* ratio was subsequently calculated. The ratios are expressed as  $\log_2(\text{BRAF-204\_TPM} + 0.01/\text{BRAF-220\_TPM} + 0.01)$ . These values are consistently greater than 1.

FLIBase allowed me to quantify BRAF-220 and BRAF-204 ratio in normal tissues as well. We analyzed 2599 samples belonging to 22 different tissues at GTEx and observed that the higher expression levels of BRAF-204 over BRAF-220 characterizes normal tissues too (the median ratio ranges between 9 and 37, Figure 23). Interestingly, the box plots of BRAF-220 and BRAF-204 expression levels in both TCGA (cancer, Figure 24) and GTEx (normal, Figure 25) tissue samples are in agreement with the pie charts in Fig. 14a-b, and they cumulative indicate that BRAF-220/BRAF-204 ratios are dictated by the expression level of the BRAF-204 isoform (in blue), while BRAF-220 levels, besides being much lower, show less variability across tissues (in red).

For the TCGA dataset, I analyzed 630 adjacent normal samples. The analysis revealed that the *BRAF-204* transcript exhibited a higher expression than *BRAF-220*, with the median ratio ranging between ~4 and ~64 (**Figure 22**).

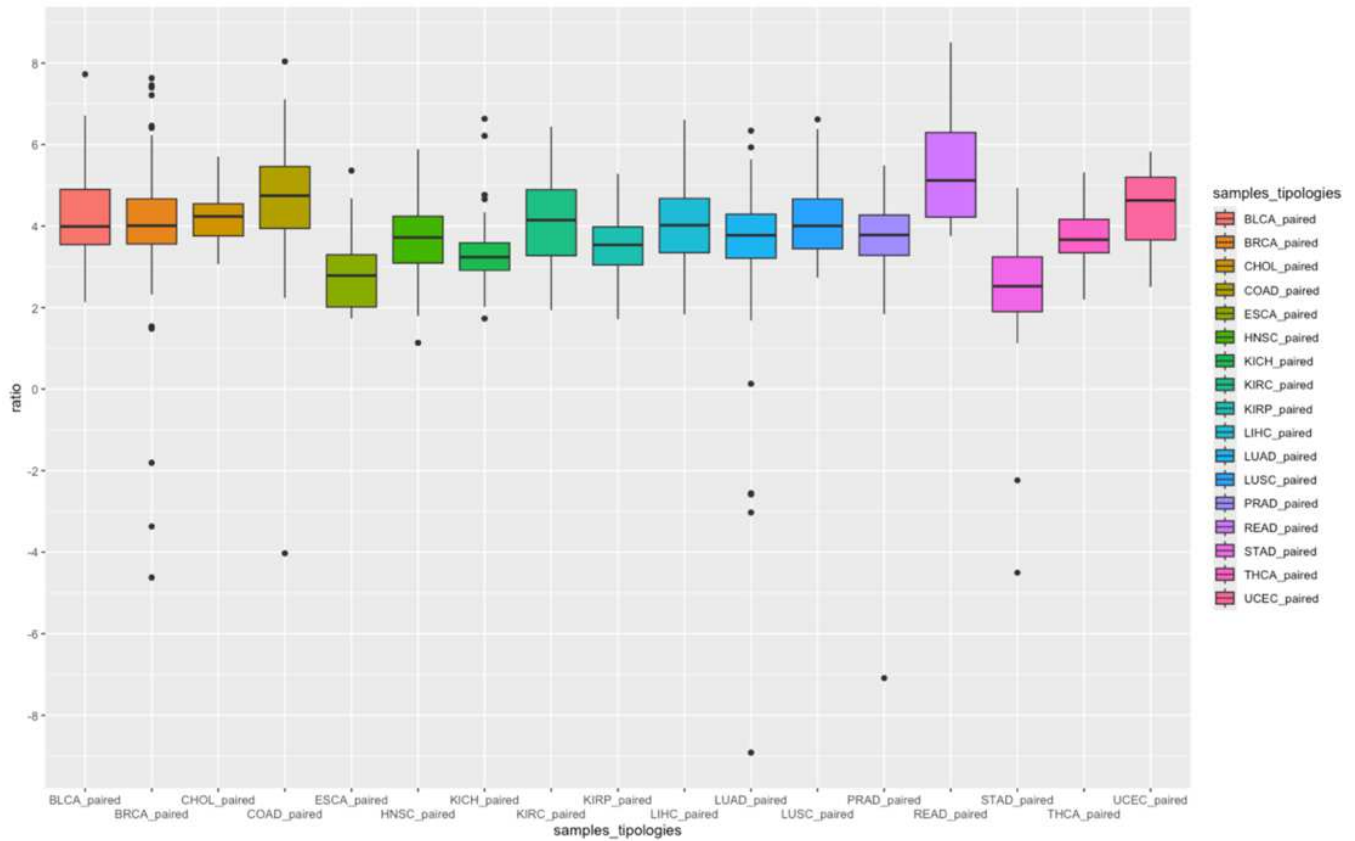


**Figure 22** Box plots of the BRAF-204/BRAF-220 ratio for tcga normal tissue samples.

The results obtained from the TCGA dataset normal are presented (total n = 630, 14 normal tissue types). The quantification of the *BRAF-204* and *BRAF-220* isoforms was obtained via FLIBase, and the *BRAF-204/BRAF-220* ratio was subsequently calculated. The ratios are expressed as  $\log_2(BRAF-204\_TPM + 0.01/BRAF-220\_TPM + 0.01)$ . These values are consistently greater than 1.

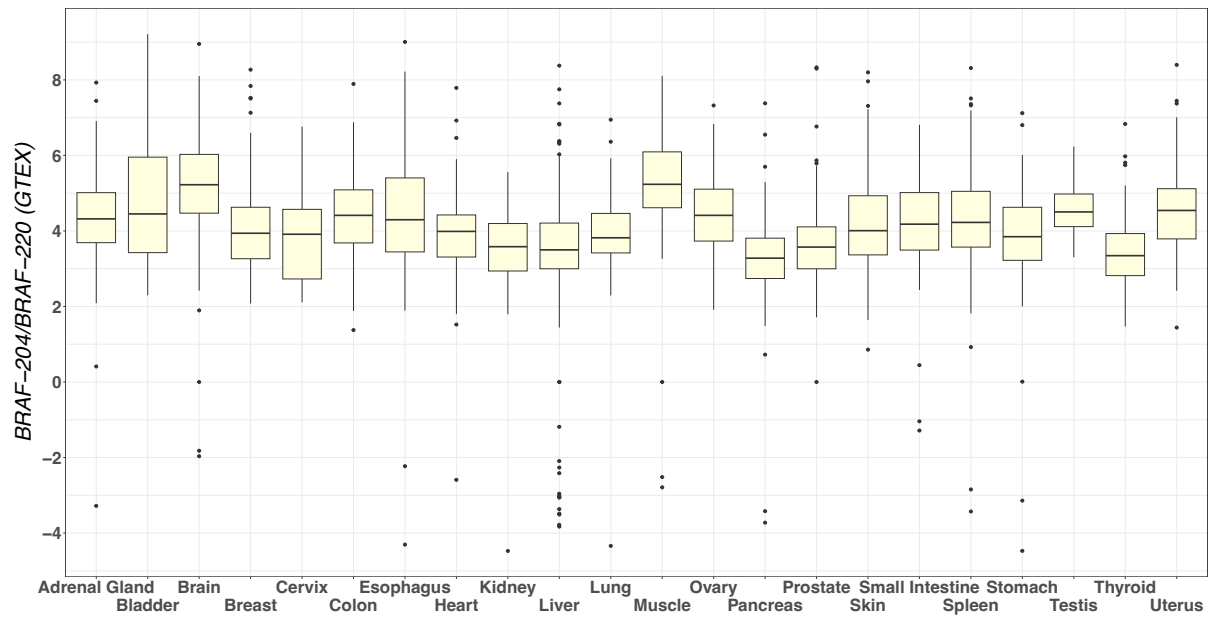


**Figure 23** Box plots of the BRAF-204/BRAF-220 ratio (GTEX)

The results obtained from the GTEX dataset are presented (total  $n = 2599$ , 22 normal tissue types). The quantification of the *BRAF-204* and *BRAF-220* isoforms was obtained via FLIbase, and the *BRAF-204/BRAF-220* ratio was subsequently calculated. The ratios are expressed as  $\log_2(BRAF-204\_TPM + 0.01/BRAF-220\_TPM + 0.01)$ . These values are consistently greater than 1.

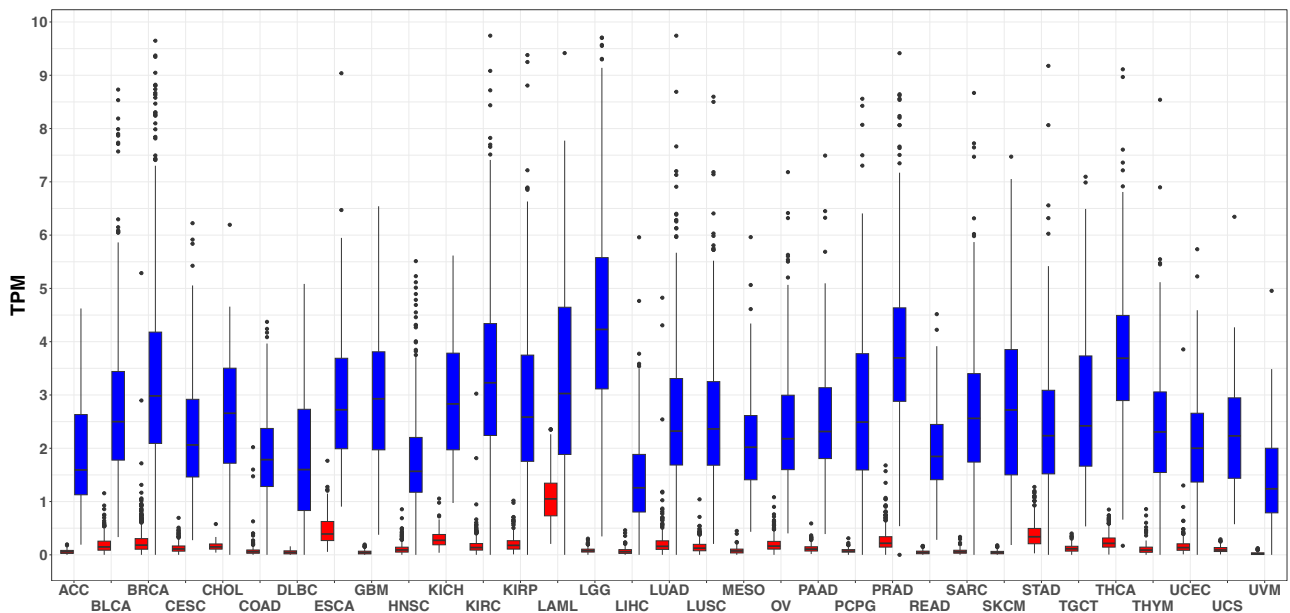


**Figure 24** Box plots of BRAF-220, BRAF-204 that compose the TCGA ( $n = 9219$  samples, 33 cancer types).

The quantification the *BRAF-220* (red) and *BRAF-220* (blue) variants in TPM was obtained via FLIbase

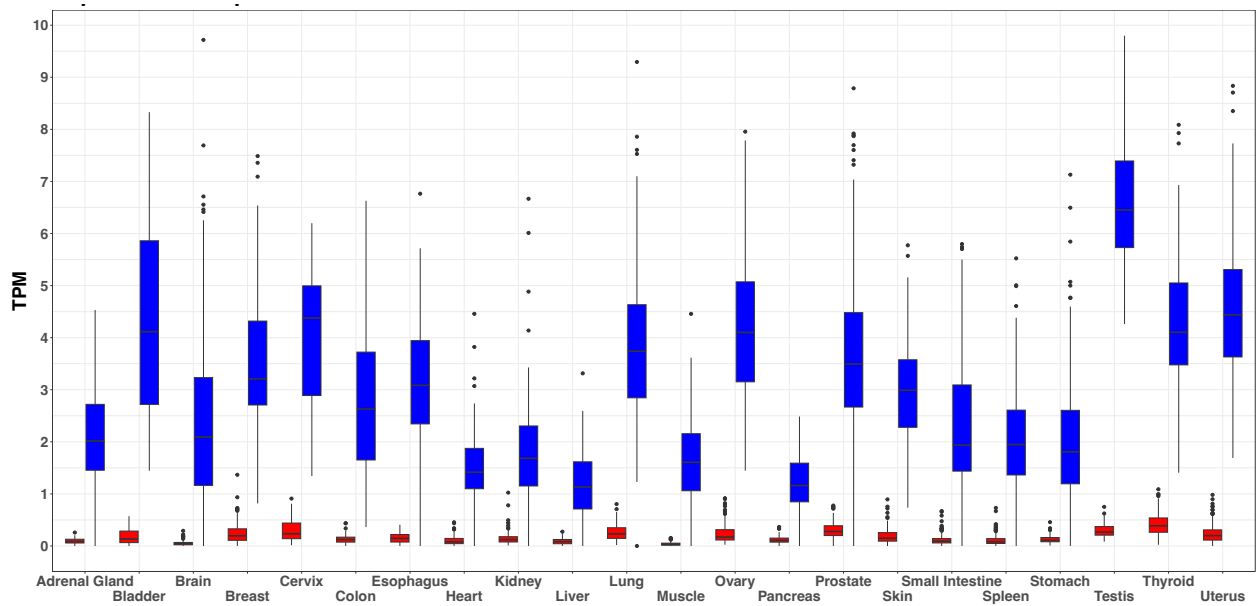


**Figure 25** Box plots of *BRAF-220*, *BRAF-204* tissue samples that compose the GTEX (n = 2599 samples, 22 normal tissues).

The quantification of the *BRAF-220* (red) and *BRAF-220* (blue) variants in TPM was obtained via FLIBase.

I also quantified total BRAF levels. The results are reported in Figure 26 for the CCLE cell lines, Figure 27 for the GEO tissue samples, Figure 28 for the TCGA tissue samples and Figure 29 for the GTEx tissue samples. BRAF-204/*BRAF-220* ratio does not correlate with total BRAF level in any of the datasets

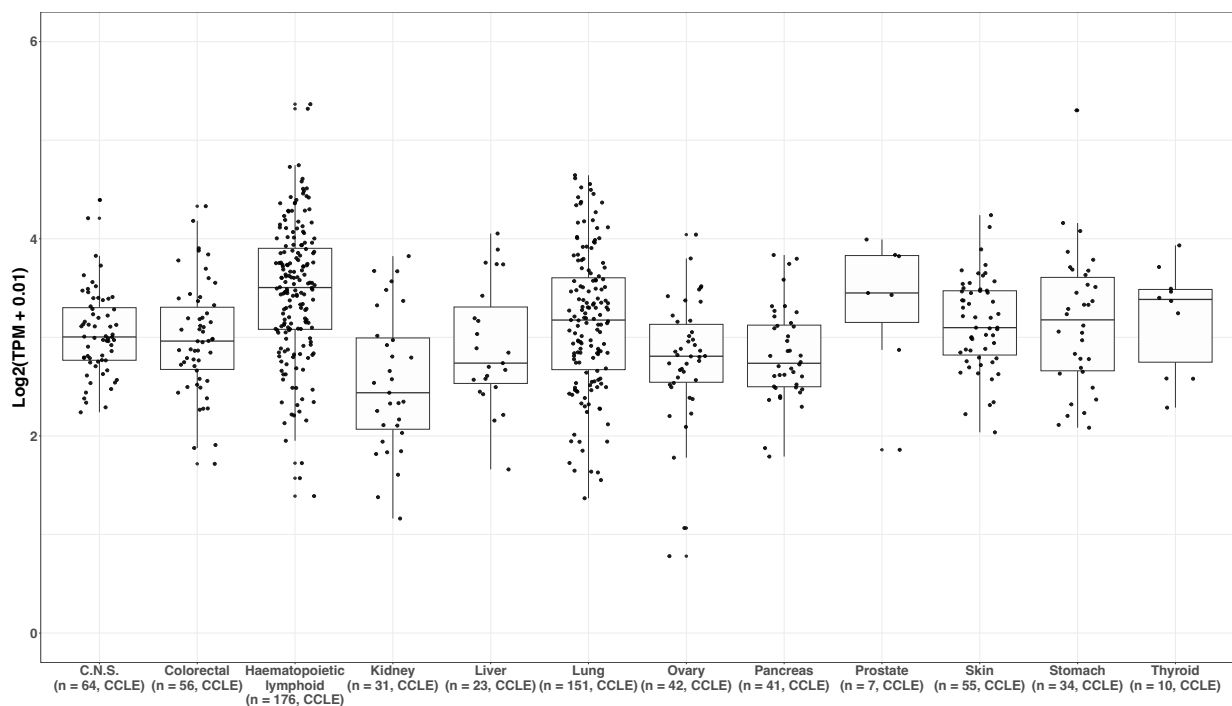


**Figure 26** CCLE Total *BRAF* levels were calculated using the Salmon module of IsoWorm and the trixmeta R package. They are expressed as  $\log_2(\text{TPM} + 0.01)$  and they range between 2 and 4

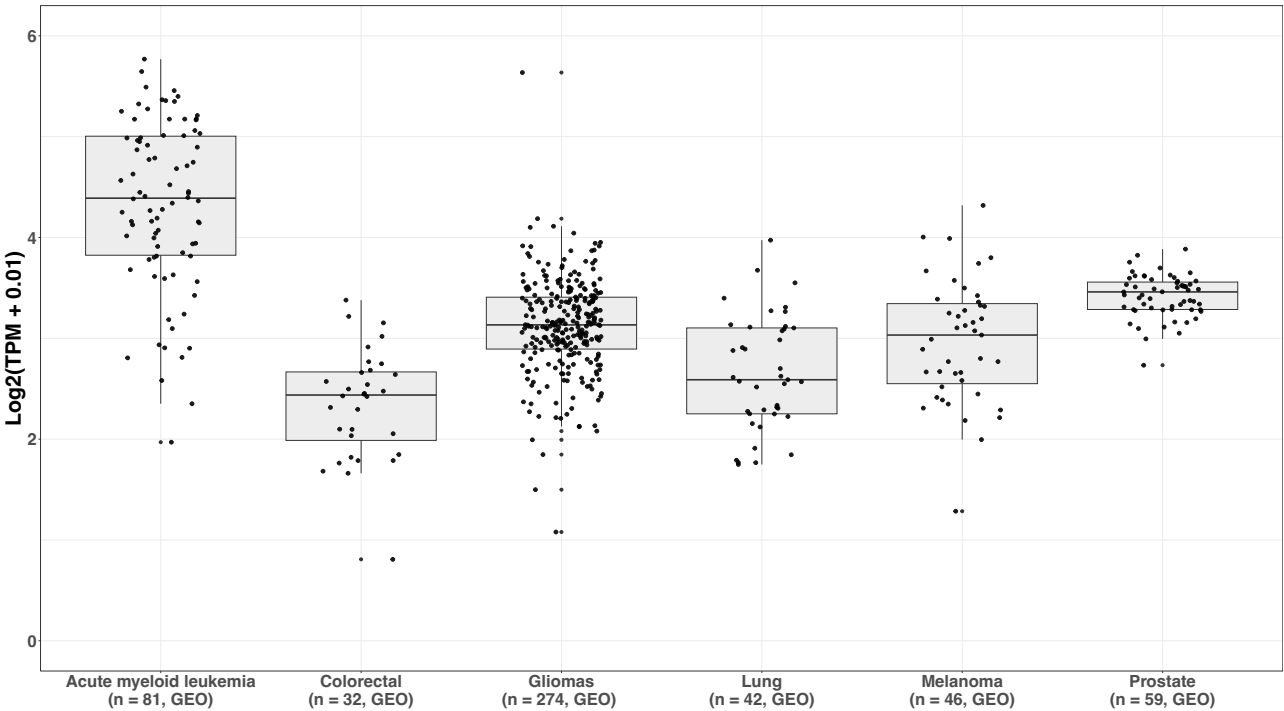


**Figure 27** Total *BRAF* levels were calculated using the Salmon module of IsoWorm and the trixmeta R package. They are expressed as  $\log_2(\text{TPM} + 0.01)$  and they range between 2 and 4.

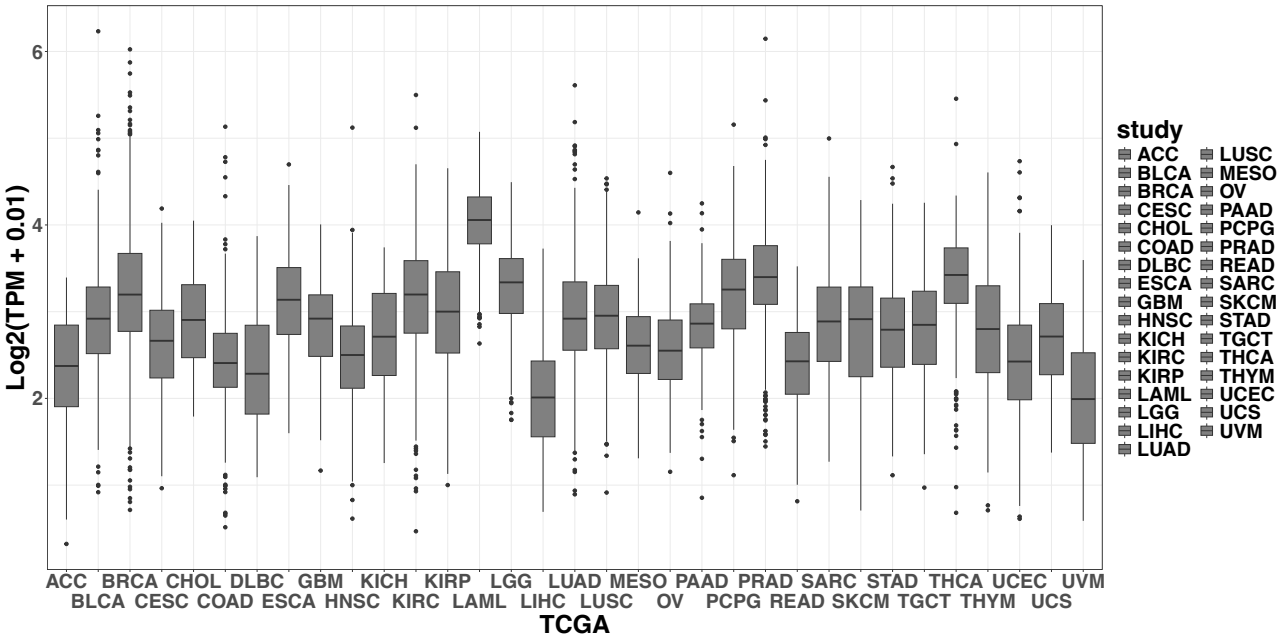


**Figure 28** TCGA Total *BRAF* levels were calculated using the recount3 package in R. They are expressed as  $\log_2(\text{TPM} + 0.01)$  and they range between 2 and 4.

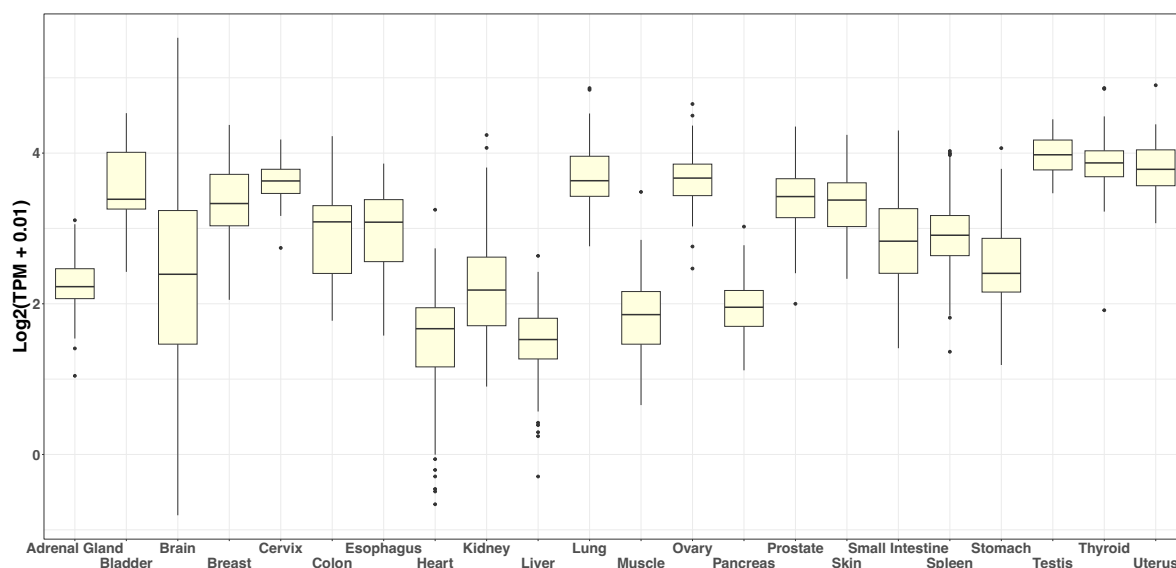


**Figure 29** GTEx Total BRAF levels were calculated using the recount3 package in R. They are expressed as  $\log_2(\text{TPM} + 0.01)$  and they range between 2 and 4.

I want to highlight the extreme case of acute myeloid leukemia (AML), which is characterized by the lowest BRAF-204/BRAF-220 ratio (Fig. 30 and 31). This due to multiple factors, including low expression of BRAF-204, high expression of BRAF-220, and high expression of additional transcript variants (Fig. 14b and Supplementary Fig. 24). AML even shows samples that are completely devoid of BRAF-204 expression (3/81 (4%) in the GEO dataset and 24/150 (16%) in the TCGA dataset). Furthermore, AML is characterized by the highest level of total BRAF (Supplementary Fig. 18 and 27). Interestingly, these features are shared across other leukemia types such as ALL, CLL and CML (Supplementary Fig. 32). Since BRAF mutations have been reported in AML and CLL and are associated with poor prognosis [105], [106], [107], our findings contribute to light up the interest of the scientific community on BRAF gene in the context of these leukemia subtypes. As soon as RNA-seq datasets become available, it will be interesting to study BRAF-204, BRAF-220 and their ratio also in the context of HCL, a leukemia type where the BRAFV600E mutation is present in virtually all cases, and it is well established as cancer driver and therapeutic target [28].

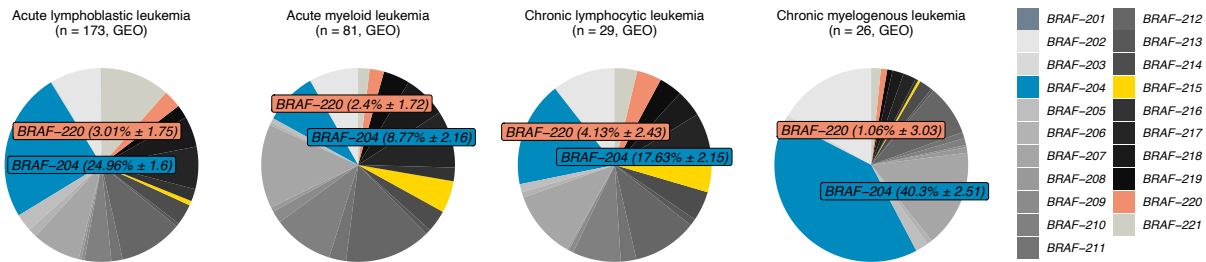


**Figure 30** Pie charts showing the 21 transcript variants of *BRAF* present in Ensembl, GEO, across 4 leukemia types (111).

Their mean relative abundance (%) was estimated via the salmon module of IsoWorm. *BRAF-204* is highlighted in blue, whereas *BRAF-220* is highlighted in red. Mean±SD values are indicated.

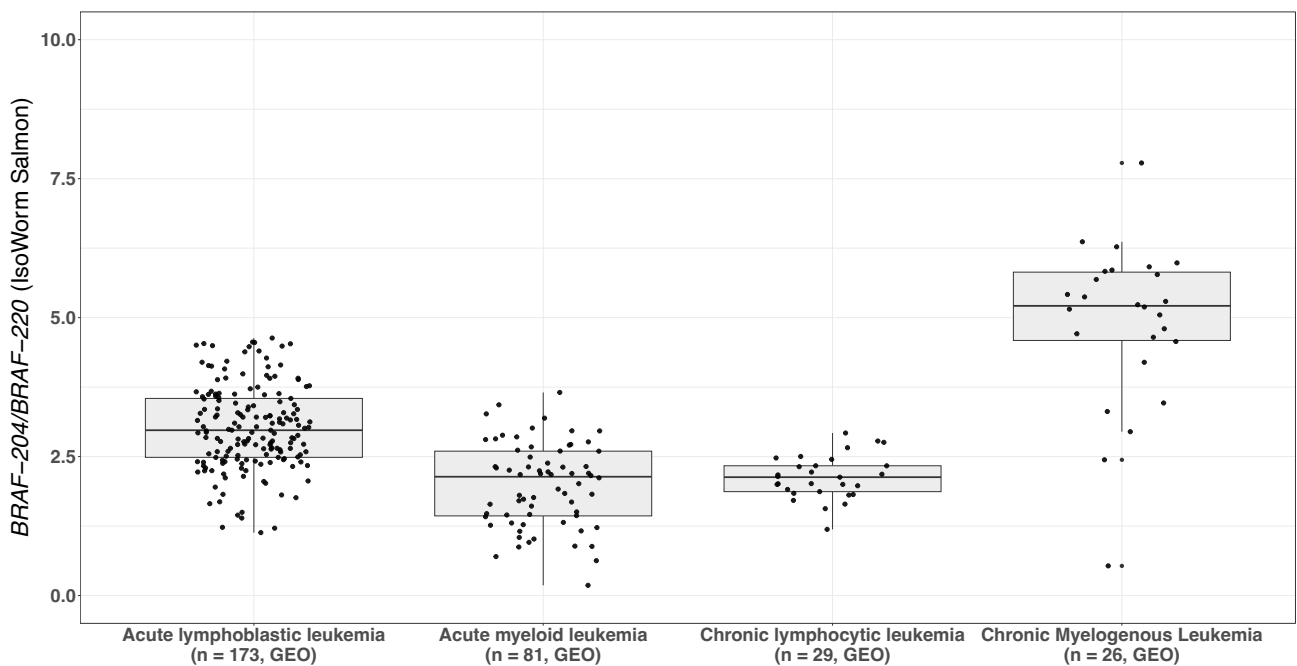


**Figure 31** Box plots of *BRAF-204/BRAF-220* ratio, GEO, across 4 leukemia types (111).

Quantification of *BRAF-204* and *BRAF-220* isoforms, followed by the calculation of their ratio, was made using the Salmon module of IsoWorm. Ratios are expressed as  $\log_2(\text{BRAF-204\_TPM} + 0.01 / \text{BRAF-220\_TPM} + 0.01)$ . They are greater than 1.

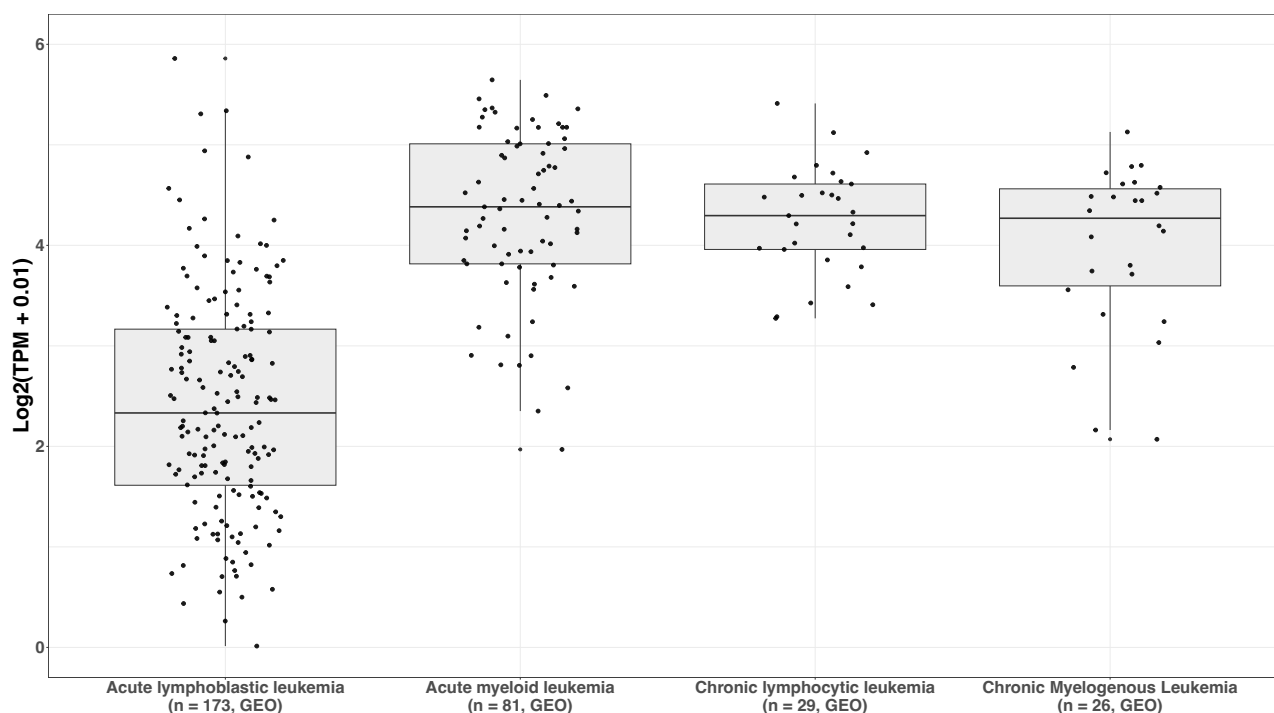


**Figure 32** Total *BRAF* levels across 4 Leukaemia types, GEO

Total *BRAF* levels were calculated using the Salmon module of IsoWorm and the trixmeta R package. They are expressed as  $\log_2(\text{TPM} + 0.01)$ .

The new analysis of the *BRAF-204/BRAF-220* ratio in TCGA samples slightly differs from our previous analysis made in 2017 [15], a discrepancy that is ascribable to both biological and technical factors.

At the biological level, we note that the TCGA dataset is constantly curated, which means that the number, identity, stage and/or subtype of current tissue samples are not the same as those in 2017.

At the technical level, several factors must be considered. First, the strategy for isoform quantification from short-read data inevitably relies on the bulk of isoforms that are present in repositories at the time of analysis. In reference [15], we considered the 15 isoforms that were present in Ensembl (release 80) plus NCBI (GRCh38.p1). Then, we dismissed most of them based on the absence of transcription (no reads on isoform-specific exons or exon–exon junctions). Finally, we established that transcription across the splicing site located in exon 18 was distinctive between the *ref* isoform and the *X1/X2* isoforms. Therefore, we quantified *BRAF-ref* by counting the reads that spanned this splicing site. However, among the *BRAF* isoforms that compose current databases, many share the same splicing site and show some degree of transcription. For example, among the 21 *BRAF* isoforms that are currently present in Ensembl 10 share that splicing site with *BRAF-220*

and are transcribed. Therefore, we conclude that in [15] *BRAF-ref* levels might have been overestimated. Conversely, we note that the quantification through IsoWorm that we describe in this article might suffer from the opposite bias, i.e., underestimation of *BRAF-220* levels. As mentioned above, the current annotation of *ENST00000646891.2/BRAF-220* presents a much longer 3'UTR than the one we considered in 2017 (~100 nt vs ~3.9 kb). The longer 3'UTR allowed us to select an isoform-specific region for *BRAF-220* identification and quantification (most downstream ~1 kb). However, this comes at the price of underestimation of expression levels if there are no reads that map to the UTR until the very 3' end. To address this issue, I resorted to a publicly available dataset in which a QuantSeq 3' mRNA library was built to produce fragments close to the 3' end of polyA sites on RNA extracted from 97 lung cancer patients [108]. Specifically, I inferred the sizes of the *BRAF-220* and *BRAF-204* 3'UTRs by retrieving the positions of the polyA sites. According to our analysis, the *BRAF-204* 3'UTR contains 1 polyA site, which is present in all 97 samples and is located at the end of ~7 kb (chr7: 140726516-140719327) (**Figure 33, in blue**). This experimental evidence matches our previous findings [15] and confirms that *BRAF-X1/204* mRNA is consistently expressed with a very long, homogeneously transcribed 3'UTR. Conversely, the *BRAF-220* 3'UTR contains 2 polyA sites. The first one is located ~400 nt downstream of the stop codon (chr7:140,734,531-140,734,607), and it is present in 81 out of 97 samples. The second one is located ~3.6 kb downstream of the stop codon (chr7:140,731,035-140,731,126), and it is present in 88 out of 97 samples. Interestingly, the short and long 3'UTRs were co-expressed in most samples (77 out of 97) (**Figure 33, in red**). These results indicate that the long version of the *BRAF-220* 3'UTR is indeed expressed and should be selected for normalization of isoform levels. However, these and previous analyses [15] also indicate that the size of the *BRAF-204* 3'UTR is highly variable, both within and across samples. Therefore, the choice of the long 3'UTR is inherently linked with a certain degree of underestimation.

In addition, the precomputed quantification of *BRAF-204* and *BRAF-220* produced through FLIBase on TCGA tissue samples combines the results of short- and long-read sequencing, whereas the quantification we performed in [15] was uniquely based on short reads.

In conclusion, precise isoform quantification via RNA-seq remains problematic despite the efforts of the bioinformatics community in developing increasingly sophisticated methods. The use of short-read and bulk sequencing presents inherent technical limitations and results in very

divergent estimations [109]. Nonetheless, the technical and biological aspects described above prompted us to consider the quantification of *BRAF* isoforms presented here accurate.

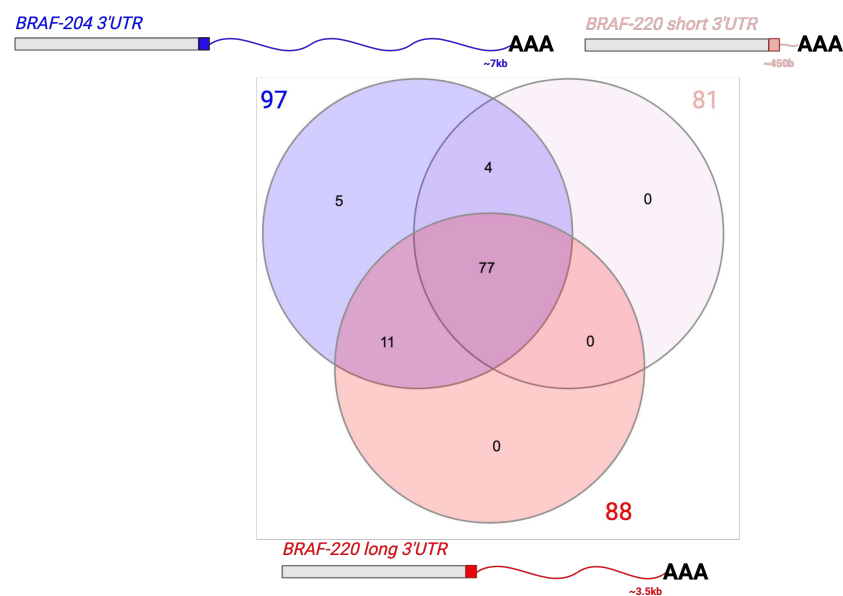


**Figure 33** Venn diagram of lung cancer patients

They are expressing BRAF-220 with a short 3'UTR (upstream polyA site located ~400 nt downstream of the stop codon, light red), BRAF-220 with a long 3'UTR (downstream polyA site located at the end of the ~3.9 kb 3'UTR, dark red), and BRAF-204 with an ~7 kb 3'UTR (terminal polyA site, blue).

#### 4 The *BRAF-204/BRAF-220* ratio has prognostic value in KIRP

I also assessed whether the *BRAF-204/BRAF-220* ratio has an impact on patient prognosis via the quantification of *TCONS\_00952571/BRAF-204* and *TCONS\_00952581/BRAF-220* in TCGA samples, as reported in FLIBase. I selected 11 cancer types that are characterized by the highest frequency of *BRAF* alterations [85]. Five of them presented with *BRAF* mutations in more than 4% of the samples (COAD, LUAD, SKCM, THCA, and UCEC), whereas six presented with *BRAF* copy number variations in more than 60% of the samples (ACC, ESCA, GBM, KIRP, OV, and TGCT). For each of these 11 cancer types, we evaluated the associations between the *BRAF-204/BRAF-220* ratio and patient prognosis (**Table 6**).

| TCGA cancer types with BRAF mutation frequency higher than 4%* |                                       |     | Univariate Cox proportional hazards regression |               |          |                                       |         |               |
|----------------------------------------------------------------|---------------------------------------|-----|------------------------------------------------|---------------|----------|---------------------------------------|---------|---------------|
| Cancer type (abbr)                                             | Cancer type (extended)                | n   | variable                                       | coef          | HR       | se(coef)                              | p-value | p.adjust (BH) |
| COAD                                                           | colon adenocarcinoma                  | 260 | Group2(vsGroup1)                               | 0.00215       | 1.00215  | 0.28969                               | 0.9941  | 0.999         |
|                                                                |                                       |     | Group3(vsGroup1)                               | -0.56388      | 0.56900  | 0.32806                               | 0.0856  | 0.3190        |
| LUAD                                                           | lung adenocarcinoma                   | 500 | Group2(vsGroup1)                               | -0.06831      | 0.93397  | 0.18944                               | 0.718   | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | 0.23412       | 1.26379  | 0.17943                               | 0.192   | 0.5280        |
| SKCM                                                           | skin cutaneous melanoma               | 103 | Group2(vsGroup1)                               | 0.1295        | 1.1382   | 0.4697                                | 0.783   | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | 0.2222        | 1.2489   | 0.4784                                | 0.642   | 0.9152        |
| THCA                                                           | thyroid carcinoma                     | 499 | Group2(vsGroup1)                               | 0.5784        | 17.832   | 0.5894                                | 0.326   | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | -0.1630       | 0.8496   | 0.6712                                | 0.808   | 0.9152        |
| UCEC                                                           | uterine corpus endometrial carcinoma  | 175 | Group2(vsGroup1)                               | -0.9498       | 0.3868   | 0.4847                                | 0.06    | 0.660         |
|                                                                |                                       |     | Group3(vsGroup1)                               | -0.2097       | 0.8109   | 0.3997                                | 0.60    | 0.9152        |
| TCGA cancer types with BRAF CNV frequency higher than 60%*     |                                       |     | Univariate Cox proportional hazards regression |               |          |                                       |         |               |
| Cancer type (abbr)                                             | Cancer type (extended)                | n   | variable                                       | coef          | HR       | se(coef)                              | p-value | p.adjust (BH) |
| ACC                                                            | adrenocortical carcinoma              | 79  | Group2(vsGroup1)                               | -0.2671       | 0.7656   | 0.4287                                | 0.533   | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | -0.8573       | 0.4243   | 0.5009                                | 0.087   | 0.3190        |
| ESCA                                                           | esophageal carcinoma                  | 152 | Group2(vsGroup1)                               | 0.09724       | 1.10213  | 0.33469                               | 0.771   | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | -0.09404      | 0.91025  | 0.32657                               | 0.773   | 0.9152        |
| GBM                                                            | glioblastoma multiforme               | 143 | Group2(vsGroup1)                               | -0.15677      | 0.85490  | 0.23833                               | 0.511   | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | 0.04985       | 1.105111 | 0.23512                               | 0.832   | 0.9152        |
| KIRP                                                           | kidney renal papillary cell carcinoma | 250 | Group2(vsGroup1)                               | 0.5379        | 1.7125   | 0.4655                                | 0.24784 | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | 1.2287        | 3.4169   | 0.4279                                | 0.004** | 0.04*         |
| OV                                                             | ovarian serous cystadenocarcinoma     | 354 | Group2(vsGroup1)                               | -0.10132      | 0.90365  | -0.612                                | 0.541   | 0.957         |
|                                                                |                                       |     | Group3(vsGroup1)                               | -0.04274      | 0.95816  | 0.16422                               | 0.795   | 0.9152        |
| TGCT                                                           | testicular germ cell tumor            | 133 | Group2(vsGroup1)                               | -<br>2,11E+04 | 6,73E-07 | 2,29E+07                              | 0.999   | 0.999         |
|                                                                |                                       |     | Group3(vsGroup1)                               | -<br>4,10E+04 | 1,56E-15 | 3,09E+07                              | 0.999   | 0.9990        |
| *PMID:35910024                                                 |                                       |     |                                                |               |          | p-values: * <0.05, **<0.01, ***<0.001 |         |               |

**Table 6** Hazard risk table

The eleven cancer subtypes where we evaluated the hazard risk for patients Group 1 included patients with the lowest BRAF-204/BRAF-220 ratios (<=33rd percentile (1Q)). Group 2 included patients with mid-range BRAF-204/BRAF-220

ratios (>33rd percentile and ≤66th percentile (2Q)). Group 3 included patients with the highest BRAF-204/BRAF-220 ratios (>66th percentile (3Q)). coef: the coefficient that quantifies the association between the independent variable and the log of the risk of the event. HR: the hazard ratio is the exponential of the coefficient. se(coef): the standard error of the coefficient measures the variability or uncertainty in the estimate of the coefficient.

According to univariate Cox proportional hazards regression analysis, the *BRAF-204/BRAF-220* ratio impacts overall survival in KIRP-*kidney renal papillary cell carcinoma* patients. Specifically, 250 patients were analyzed (**Table 7**), and patients in Group 3 (highest ratios, >66th percentile (3Q)) had an increased risk of worse outcomes than patients in Group 1 did (lowest ratios, ≤33rd percentile (1Q)). As reported in **Table 6** and in **Table 8, on the left**, the hazard ratio (HR) was greater than 3. I then calculated the HR for prognostic factors that are already known to increase the risk of worse outcomes in KIRP patients (see, for example, [110]), namely, the size and extent of the primary tumor (T3-T4/T1-T2), the involvement of regional lymph nodes (N1-N2/N0), the presence of distant metastases (M1/M0), and the overall tumor stage (TNM I, II, III-IV) where group III-IV, the most extreme conditions, were merged due to numerosity reasons.

As expected, I obtained HR values higher than 1 for all of them (**Table 8, left**). Next, I performed multivariate Cox proportional hazards regression analysis, considering all the abovementioned prognostic variables at the same time. As shown in **Table 8, right**, a high *BRAF-204/BRAF-220* ratio was associated with poor survival. Therefore, I can conclude that in KIRP the ratio of *BRAF* isoforms is an independent prognostic factor. In addition, I assessed whether having a higher *BRAF-204/BRAF-220* ratio increases the probability of being in a higher KIRP stage. To this end, I performed a proportional odds logistic regression (POLR) for ordered category outcome analysis. I used stage I as a reference and considered tumor stage as the response variable (I < II < III-IV). Consistent with the hazard regression analyses, I obtained an odds ratio of 2.27 for Group 3, which means that KIRP patients with the highest ratios have higher odds of having a more advanced tumor stage than do patients with the lowest ratios (**Table 9**). Finally, since KIRP is one of the 6 cancer types with *BRAF* copy number variations in more than 60% of samples, I tested whether there was a correlation between copy number status (i.e., 0, 1, 2, 3) and the *BRAF-204/BRAF-220* ratio, but we did not find a correlation (Spearman rho -0.09, p<0.14). To date, several molecular biomarkers have been proposed to improve the prediction of drug responsiveness and disease outcome in KIRP (see, for example, [111] [112] [113] [114] [115]), but the role of *BRAF* has just begun to be explored [85]. Crucially, these findings revealed that high *BRAF-204/BRAF-220* ratios

distinguish patients with worse outcomes. They also spur to fill a gap in public repositories, since we lack an independent KIRP cohort suitable for validating the prognostic value of proposed biomarkers.

I found no impact of the *BRAF*-204/*BRAF*-220 ratio on THCA or SKCM, i.e., the cancer types characterized by the highest percentage of patients carrying *BRAF* mutations. However, given the availability of larger datasets, it might be interesting to assess whether the ratio has prognostic value in the BRAFV600E-negative subgroup, whereas the presence of the mutant protein takes over in the BRAFV600E-positive subgroup.

In more general terms, the results we obtained in KIRP underscore the importance of going beyond the study of the aberrant activity of the mutant kinase, and of including the study of the expression of the *BRAF* gene.

| TCGA KIRP cohort                |                          |
|---------------------------------|--------------------------|
| Variables                       | Total patients (n = 250) |
| <b>Survival Days</b>            |                          |
| <b>Mean <math>\pm</math> SD</b> | 1021.076 $\pm$ 901.3952  |
| <b>Median (Range)</b>           | 744 (0-5925)             |
| <b>Clinical Stage</b>           |                          |
| <b>Stage I</b>                  | 166 (66.4%)              |
| <b>Stage II</b>                 | 21 (8.4%)                |
| <b>Stage III-Stage IV</b>       | 63 (25.2%)               |
| <b>Clinical T</b>               |                          |
| <b>T1-T2</b>                    | 195(78%)                 |
| <b>T3-T4</b>                    | 55(22%)                  |
| <b>Clinical N</b>               |                          |
| <b>N0</b>                       | 223 (89.2%)              |
| <b>N1-N2</b>                    | 27 (10.8%)               |
| <b>Clinical M</b>               |                          |
| <b>M0</b>                       | 241 (94.4%)              |
| <b>M1</b>                       | 9 (3.6%)                 |

**Table 7** Baseline characteristics of the 250 KIRP patients used for hazards regression analyses.

| Variables              | Univariate |         |          |                    |                    | Multivariate |         |          |                    |
|------------------------|------------|---------|----------|--------------------|--------------------|--------------|---------|----------|--------------------|
|                        | coef       | HR      | se(coef) | p-value            | p.adj (BH)         | coef         | HR      | se(coef) | p-value            |
| <b>Group</b>           | 0.5379     | 1.7125  | 0.4655   | 2.24e-01           | 2.89e-01           | 0.6791       | 1.9721  | 0.4906   | 1.66e-01           |
| <b>(Group2/Group1)</b> |            |         |          |                    |                    |              |         |          |                    |
| <b>Group</b>           | 1.2287     | 3.4169  | 0.4279   | <b>4.08e-03**</b>  | <b>5.71e-03**</b>  | 1.4664       | 4.3337  | 0.4643   | <b>1.59e-03**</b>  |
| <b>(Group3/Group1)</b> |            |         |          |                    |                    |              |         |          |                    |
| <b>T (T3–4/T1–2)</b>   | 1.5125     | 4.5383  | 0.3292   | <b>4.33e-06***</b> | <b>7.57-e06***</b> | -            | 0.5299  | 0.5388   | 2.38e-01           |
|                        |            |         |          |                    |                    | 0.6350       |         |          |                    |
| <b>N (N1-N2/N0)</b>    | 2.0616     | 7.8589  | 0.3346   | <b>7.19e-10***</b> | <b>2.51-e09***</b> | 0.7981       | 2.2214  | 0.4569   | 8.00e-02           |
|                        |            |         |          |                    |                    |              |         |          |                    |
| <b>M (M1/M0)</b>       | 4.5773     | 97.2550 | 0.6163   | <b>1.11e-13***</b> | <b>7.77e-13***</b> | 3.7009       | 40.4838 | 0.6840   | <b>6.27e-08***</b> |
|                        |            |         |          |                    |                    |              |         |          |                    |
| <b>TNM stage</b>       | 0.0102     | 1.0103  | 0.7710   | 9.89e-01           | 9.89e-01           | -            | 0.9193  | 0.7723   | 9.13e-01           |
| <b>(II/I)</b>          |            |         |          |                    |                    | 0.0841       |         |          |                    |
| <b>TNM stage</b>       | 1.8503     | 6.3617  | 0.3647   | <b>3.9e-07***</b>  | <b>9.10e-07***</b> | 1.5234       | 4.5877  | 0.6428   | <b>1.77e-02*</b>   |
| <b>(IV-III/I)</b>      |            |         |          |                    |                    |              |         |          |                    |

**Table 8** Univariate (left) and multivariate (right) Cox proportional hazards regression analyses performed on 250 KIRP patients.

The univariate model investigates how each factor (BRAF-204/BRAF-220 ratio, size and extent of the primary tumor (T3-T4/T1-T2), involvement of regional lymph nodes (N1-N2/N0), presence of distant metastases (M1/M0), and tumor stage (TNM I, II, III-IV)) impacts patients' overall survival, whereas the multivariate model considers the impact of all factors simultaneously. Group 1 included patients with the lowest BRAF-204/BRAF-220 ratios ( $\leq$ 33rd percentile (1Q)). Group 2 included patients with mid-range BRAF-204/BRAF-220 ratios ( $>$ 33rd percentile and  $\leq$ 66th percentile (2Q)). Group 3 included patients with the highest BRAF-204/BRAF-220 ratios ( $>$ 66th percentile (3Q)). T1-4: increasing size and extent of the primary tumor. N0/N1-N2: no evidence of a tumor on regional lymph nodes (N0) or evidence of a tumor on regional lymph nodes (N1-N2). M0/M1: no evidence of metastasis (M0), evidence of metastasis (M1). TNM stage I-IV: progressive stage. Stages III and IV were merged.

coef: the coefficient that quantifies the association between the independent variable and the log of the risk of the event. HR: the hazard ratio is the exponential of the coefficient. se(coef): the standard error of the coefficient measures the variability or uncertainty in the estimate of the coefficient. p value: calculated via the z test, the p value indicates the statistical significance of the coefficient. p-adjusted: Benjamini–Hochberg (BH) correction for multiple comparisons. \*  $<0.05$ , \*\* $<0.01$ , \*\*\* $<0.001$ .

|                                  | coefficient | p-value     | OR       |
|----------------------------------|-------------|-------------|----------|
| <b>Group<br/>(Group2/Group1)</b> | 0.0427093   | 8.99e-01    | 1.043634 |
| <b>Group<br/>(Group3/Group1)</b> | 0.8214648   | 1.03e-02*   | 2.273828 |
| <b>1 2</b>                       | 0.9742624   | 6.0e-05***  | 2.649212 |
| <b>2 3</b>                       | 1.3754486   | 4.33e-08*** | 3.956851 |

**Table 9** Proportional odds logistic regression for ordered category outcome analysis performed on 250 KIRP patients.

The response variable was tumor stage (I < II < III-IV), while stage I was used as a reference. Group 1 included patients with the lowest *BRAF-204/BRAF-220* ratios ( $\leq 33$ rd percentile (1Q)). Group 2 included patients with mid-range *BRAF-204/BRAF-220* ratios ( $> 33$ rd percentile and  $\leq 66$ th percentile (2Q)). Group 3 included patients with the highest *BRAF-204/BRAF-220* ratios ( $> 66$ th percentile (3Q)). Coefficient: the coefficient represents the effect of an independent variable on the log odds of being in a higher category for the dependent variable.

OR: odds ratio. p value: \*  $< 0.05$ , \*\*  $< 0.01$ , \*\*\*  $< 0.001$ . 1|2, 2|3: intercepts for the various levels of the outcomes

## 5 Differential expression analysis, GSEA and GSVA of KIRP patients

After stratifying KIRP patients using the methodologies described in the previous paragraph, I performed a differential expression analysis (DE) to identify genes that are upregulated or downregulated between the group with a low or a high *BRAF-204/BRAF-220* ratio. Through this analysis, I identified 852 genes that were upregulated with a log2foldchange greater than 1, and 469 genes that were downregulated with a log2foldchange less than -1. To investigate the molecular signature that could characterized the two groups, I also performed a Gene Set Enriched Analysis (GSEA) using the total gene list in output from the DE analysis. I also performed Gene Set Variation Analysis (GSVA). While GSEA focuses on identifying gene sets enriched in the overall dataset, GSVA provides a continuous enrichment score for each gene set in each sample. Both analyses highlighted rather generic pathways broadly associated with cancer biology, without providing specific functional insights for the context of this study. In GSEA, the most significant pathways included processes related to post-transcriptional regulation (e.g., keratinocyte differentiation, epidermis development, and RNA-mediated gene silencing, **Table 10**). Similarly, GSVA was not able to identify

gene signatures or pathways distinctly associated with the analyzed phenotype. This suggests that the underlying biological processes are widely generic and common in oncological contexts, limiting the ability to draw specific conclusions.

| ID         | Description                                                   | setSize | NES  | p,adjust |
|------------|---------------------------------------------------------------|---------|------|----------|
| GO:0030216 | keratinocyte differentiation                                  | 137     | 1,80 | 0,01     |
| GO:0035195 | miRNA-mediated post-transcriptional gene silencing            | 165     | 1,78 | 0,01     |
| GO:0035194 | regulatory ncRNA-mediated post-transcriptional gene silencing | 178     | 1,77 | 0,01     |
| GO:0016441 | post-transcriptional gene silencing                           | 187     | 1,75 | 0,01     |
| GO:0009913 | epidermal cell differentiation                                | 199     | 1,63 | 0,01     |
| GO:0031047 | regulatory ncRNA-mediated gene silencing                      | 235     | 1,62 | 0,01     |
| GO:0043588 | skin development                                              | 270     | 1,49 | 0,01     |
| GO:0008544 | epidermis development                                         | 333     | 1,48 | 0,01     |
| GO:0007173 | epidermal growth factor receptor signaling pathway            | 106     | 1,43 | 0,04     |
| GO:0098773 | skin epidermis development                                    | 121     | 1,43 | 0,04     |

**Table 10** Top ten pathways enriched in the GSEA on KIRP patients divided between higher and lower expression of BRAF ratio.

## 6 Exploring the potential role of BRAF-204 in miRNA sponging

To date, several molecular biomarkers have been proposed to improve the prediction of drug responsiveness and disease outcome in KIRP (see, for example, [116], [117], [118], [119], [120]), but the role of BRAF has just begun to be explored [16]. Crucially, our findings revealed that high BRAF-204/BRAF-220 ratios distinguish patients with worse outcomes. They also spur to fill a gap in public repositories, since we lack an independent KIRP cohort suitable for validating the prognostic value of proposed biomarkers. Experiments from plants, mammals and prokaryotes clearly demonstrated that endogenous RNAs compete to bind small RNAs (competing endogenous RNA or target mimicry) [121]. In a recent work, we demonstrated how the long 3'UTR of *BRAF-204* can have a role in acting as a sponge for miRNAs [122]. To follow this lead and explore the potential role of *BRAF-204* in miRNA sponging, I searched for genes deregulated in KIRP patients that share binding sites for miRNAs with the *BRAF-204* transcript. To do that, I cross-referenced upregulated genes with validated miRNA targets from the miRBase database, focusing on *BRAF-204* miRNAs identified in the above-mentioned study. I found a transcription factor up regulated (higher in the patients with a higher ratio), *TFAP2A*, that shares miRNA-320A with the 3'UTR of *BRAF-204*. The gene has a log2foldchange of 2.07 with a p.adjusted < 0.05. Subsequently, I confirmed that miR-320A exclusively binds to the 3'UTR of *BRAF-204* using *in silico* methods (**table 12**) via the RNA22 web server[123]. I calculated the binding affinity between miR-320A and the 3'UTRs of both *BRAF-220*

and *BRAF-204* isoforms. The analysis revealed binding exclusively on the 3'UTR of *BRAF-204*, with a folding energy of -13.50 and a p-value of 0.03. This result further validated the specificity of miR-320A for the 3'UTR of *BRAF-204*. It is important to consider that when we investigated the role of miRNAs binding exclusively to *BRAF-204* [122], we identified three distinct classes of miRNAs: those that induce a  $\geq 20\%$  decrease in BRAF mRNA levels, those that cause a  $\geq 20\%$  increase in BRAF mRNA levels, and those with no significant impact, defined as causing a change of less than 20% in either direction. Notably, miR-320A was classified in this last group, demonstrating no significant alteration in *BRAF* mRNA levels. miR-320a functions primarily as a tumor suppressor, inhibiting cell proliferation and promoting apoptosis in cancer cells, thereby helping to prevent tumor formation and growth.

After retrieving the miRNAome of KIRP patients, I analyzed the correlations between the expression levels of *TFAP2A* and *miR-320*, *TFAP2A* and *BRAF-204*, *miR-320* and *BRAF-204*, *TFAP2A* and the *BRAF-204/BRAF-220* ratio, as well as *miR-320* and the *BRAF-204/BRAF-220* ratio. However, no significant correlations were observed in any of these analyses (correlation matrix, **figure 34**). Therefore, the molecular mechanism at the basis of the prognostic value displayed by *BRAF-204/BRAF-220* ratio and *TFP2A* remains to be established.

| mir Name | utr transcript name                      | leftmost position of predicted target site | folding energy (in - Kcal/mol) | heteroduplex                                                      | p value |
|----------|------------------------------------------|--------------------------------------------|--------------------------------|-------------------------------------------------------------------|---------|
| mir320a  | ENST00000496384.7<br>utr3_protein_coding | 3937                                       | -13.50                         | ACCATAGCCAGAGATGGAAGGT<br> :       :    :<br>CCTTCTTGCCCTTCTCTCCG | 6.08E-3 |

**Table 11** In silico analysis made using RNA22 webserver to test the binding between the mir-320a sequence and the 3'UTR of BRAF-204.

The heteroduplex indicates the binding site while the folding energy how strong is the binding, where lower is better.

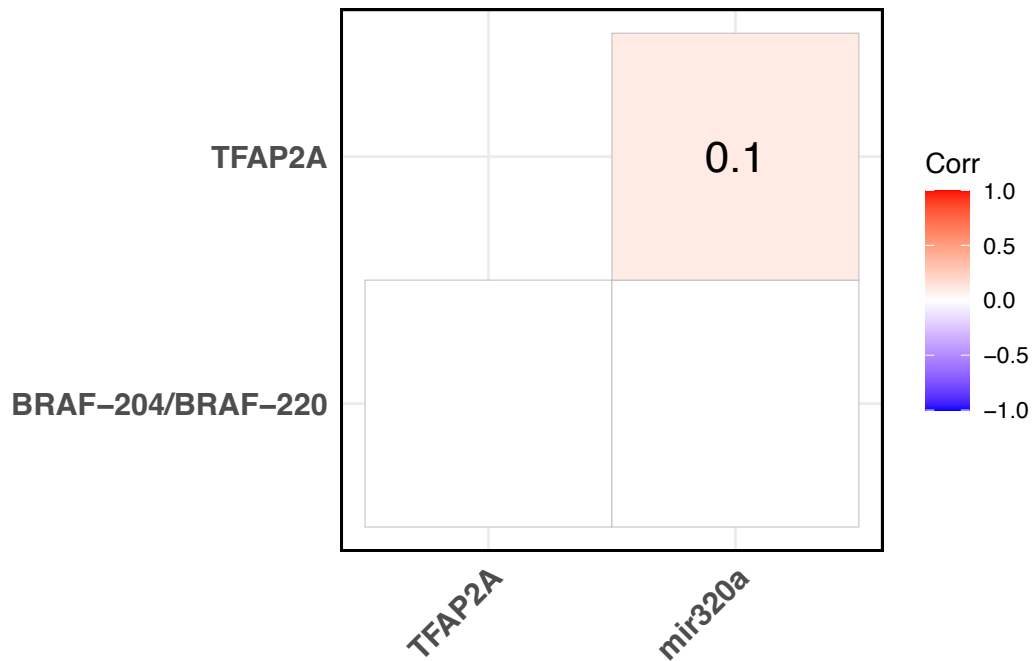


**Figure 34** Correlations matrix

Correlations matrix between the expression levels of TFAP2A and miR-320, TFAP2A and BRAF-204, miR-320 and BRAF-204, TFAP2A and the BRAF-204/BRAF-220 ratio, as well as miR-320 and the BRAF-204/BRAF-220 ratio. The color indicates the correlation coefficient value.

## 7 The TFAP2A gene is a prognostic marker on KIRP

The TFAP2A gene encodes the transcription factor AP-2 alpha (AP-2 $\alpha$ ), a key protein involved in regulating gene expression by binding to specific DNA regions. It plays a crucial role in melanocyte differentiation by working alongside MITF. Together, they control the expression of genes essential for melanin synthesis. It is well known in literature how in KIRP the transcriptional factor has a role in the prognosis of patients, where higher levels of TFAP2A expression are associated with worse prognosis [124]. It has been also proved how the different expression between normal samples and tumoral samples of KIRC and KIRP could suggests a substantial role for the factor in renal tumorigenesis[125]. To confirm the results of the impact on KIRP patients' prognostic, I stratified the KIRP TCGA dataset into two groups: Group 2, consisting of patients with the highest *TFAP2A* expression (>75th percentile, 2Q), and Group 1, comprising patients with the lowest *TFAP2A* expression ( $\leq$ 25th percentile, 1Q). After, I performed a survival analysis and, as shown in the Kaplan-Meier curve on **Figure 35**, it is confirmed how having a higher expression of the transcriptional factor *TFAP2A* is associated with worse prognosis.

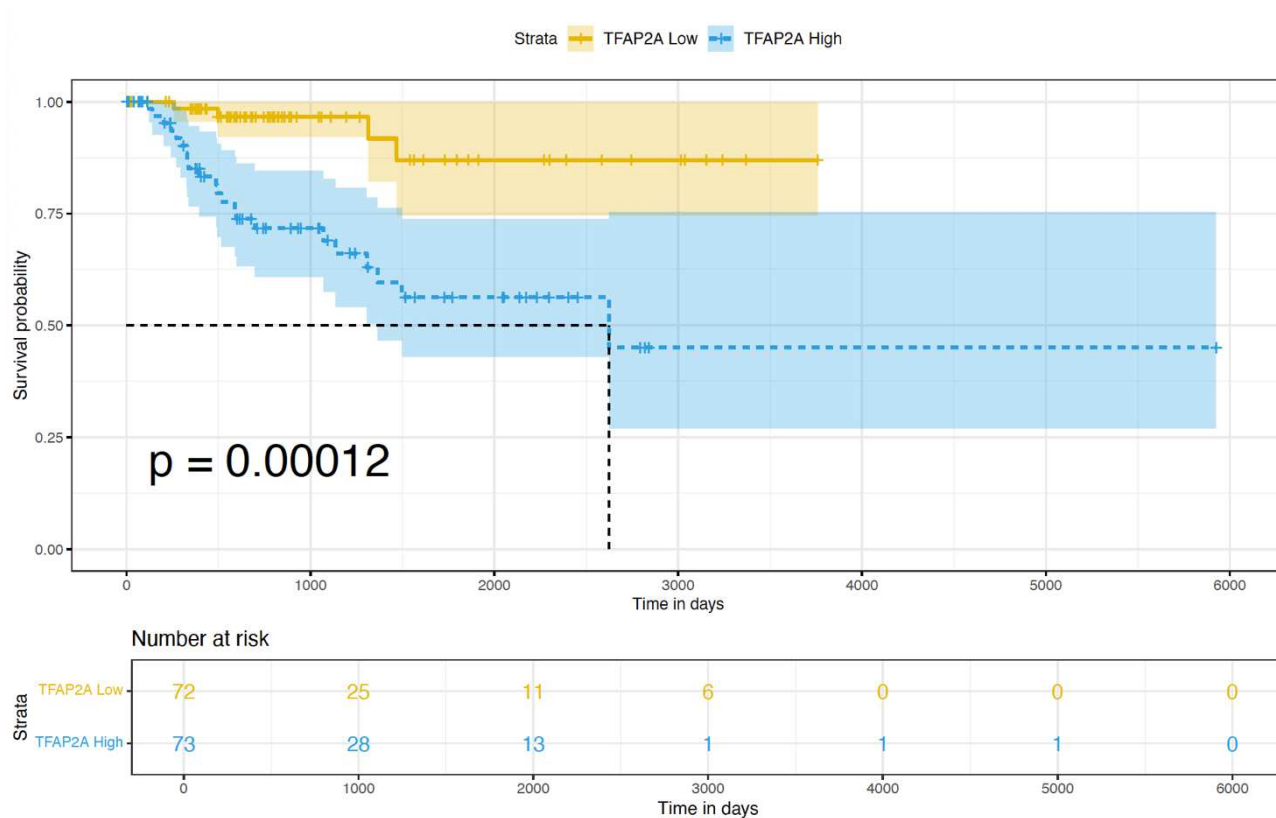


**Figure 35** Kaplan-Meier curve that illustrate the survival probability between the two groups on KIRP patients One having higher (blue one) and one having lower (yellow) *TFAP2A* expression.

To test if the expression of the gene could have an impact also on the hazard risk of the patients, I added the *TFAP2A* expression levels to the multivariate Cox proportional hazards regression model described above and as shown in **table 11**, right, adding this prognostic element increases the hazard risk by 4%.

## KIRP

| Variables                         | Univariate |           |          |       |             | Multivariate |           |          |       |             |
|-----------------------------------|------------|-----------|----------|-------|-------------|--------------|-----------|----------|-------|-------------|
|                                   | coef       | exp(coef) | se(coef) | z     | pvalue      | coef         | exp(coef) | se(coef) | z     | pvalue      |
| <b>Ratio mid(vsRatio Low)</b>     | 0.5379     | 1.7125    | 0.4655   | 1.156 | 0.24784     | 0.53985      | 1.71574   | 0.49014  | 1.101 | 0.27072     |
| <b>Ratio High (vsRatio Low)</b>   | 1.2287     | 3.4169    | 0.4279   | 2.872 | 0.00408**   | 1.45601      | 4.28883   | 0.46597  | 3.125 | 0.00178**   |
| <b>TFAP2A High (vsTFAP2A Low)</b> | 0.06801    | 1.07037   | 0.01437  | 4.731 | 2.23e-06*** | 0.04402      | 1.04500   | 0.01818  | 2.421 | 0.01546*    |
| <b>T (T3–4/T1–2)</b>              | 1.5125     | 4.5383    | 0.3292   | 4.595 | 4.33e-06*** | 1.5125       | 4.5383    | 0.3292   | 4.595 | 0.2388      |
| <b>N (N1/N0)</b>                  | 2.0616     | 7.8589    | 0.3346   | 6.162 | 7.19e-10*** | 2.0616       | 7.8589    | 0.3346   | 6.162 | 0.0806      |
| <b>M (M1/M0)</b>                  | 4.5773     | 97.2550   | 0.6163   | 7.427 | 1.11e-13*** | 4.5773       | 97.2550   | 0.6163   | 7.427 | 6.31e-08*** |
| <b>TNM stage (III–IV/I–II)</b>    | 1.8487     | 6.3517    | 0.3447   | 5.363 | 8.16e-08*** | 1.8487       | 6.3517    | 0.3447   | 5.363 | 0.0152 *    |

**Table 12** Univariate (left) and multivariate (right) Cox proportional hazards regression analyses performed on 250 KIRP patients.

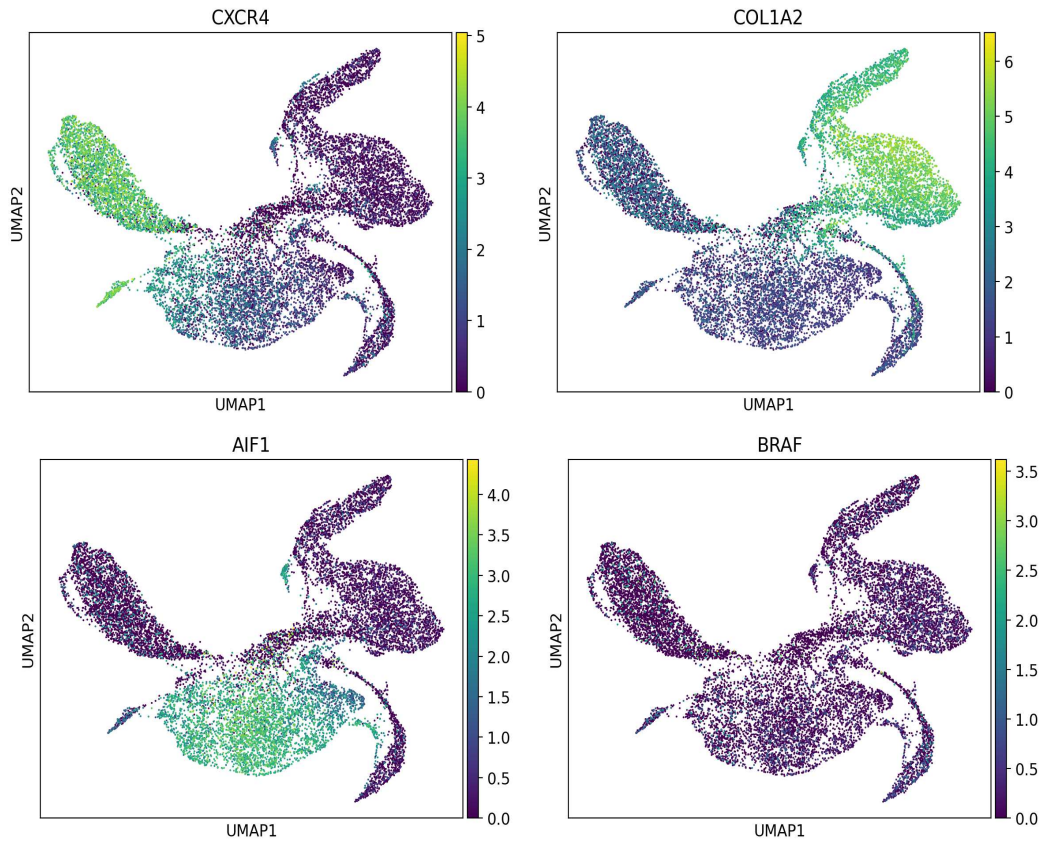
The univariate model investigates how each factor (BRAF-204/BRAF-220 ratio, size and extent of the primary tumor (T3-T4/T1-T2), involvement of regional lymph nodes (N1-N2/N0), presence of distant metastases (M1/M0), and tumor stage (TNM I, II, III-IV)) impacts patients' overall survival, whereas the multivariate model considers the impact of all factors simultaneously. Group 1 included patients with the lowest BRAF-204/BRAF-220 ratios ( $\leq$ 33rd percentile (1Q)). Group 2 included patients with mid-range BRAF-204/BRAF-220 ratios ( $>$ 33rd percentile and  $\leq$ 66th percentile (2Q)). Group 3 included patients with the highest BRAF-204/BRAF-220 ratios ( $>$ 66th percentile (3Q)). T1-4: increasing size and extent of the primary tumor. N0/N1-N2: no evidence of a tumor on regional lymph nodes (N0) or evidence of a tumor on regional lymph nodes (N1-N2). M0/M1: no evidence of metastasis (M0), evidence of metastasis (M1). TNM stage I-IV: progressive stage. Stages III and IV were merged. TFAP2A High (vsTFAP2A Low): higher level of TFAP2A vs lower levels in patients. coef: the coefficient that quantifies the association between the independent variable and the log of the risk of the event. HR: the hazard ratio is the exponential of the coefficient. se(coef): the standard error of the coefficient measures the variability or uncertainty in the estimate of the coefficient. p value: calculated via the z test, the p value indicates the statistical significance of the coefficient. p-adjusted: Benjamini–Hochberg (BH) correction for multiple comparisons. \*  $<0.05$ , \*\* $<0.01$ , \*\*\* $<0.001$ .

## 8 BRAF is not a marker suitable for single cell analysis

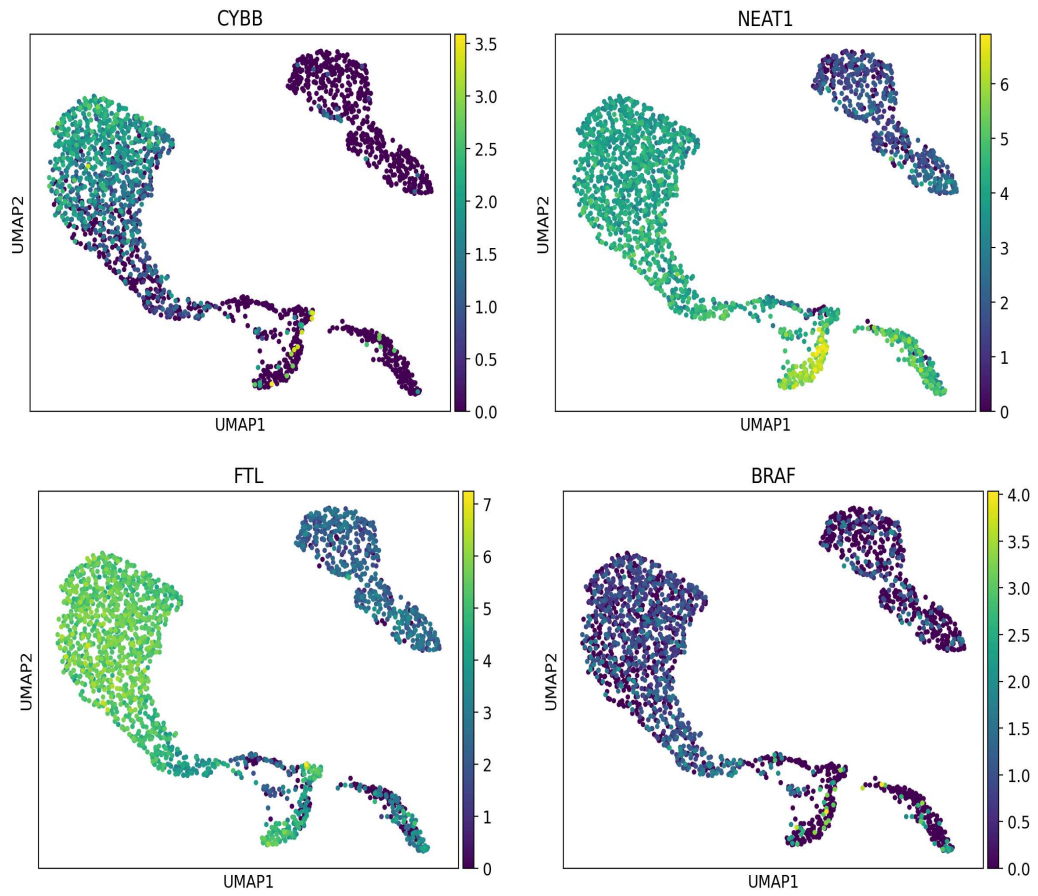
As explained in the introduction, single-cell RNAseq is an advancing technology that could allow me to characterize BRAF expression at a single-cell resolution and reveal subpopulations of cells with high or low BRAF ratio, which could drive tumour progression or therapy resistance.

I have implemented in IsoWorm a module to calculate gene expression at a single cell level using kallisto-bustools, and I analyzed two different melanoma datasets and one KIRP dataset. In all three datasets, I observed that BRAF expression was consistently very low in individual cells, even when compared to markers typically expressed at high levels in specific subpopulations of tumor subtypes (**figures 36a, 36b, 36c**). This low detectability of BRAF could stem from its basal expression levels, which are often overshadowed by the higher transcriptional activity of marker genes. Additionally, the inherent dropout effect in single-cell RNA sequencing disproportionately affects genes with lower expression, making their detection sporadic and more difficult [126], [127].

a)



b)



c)

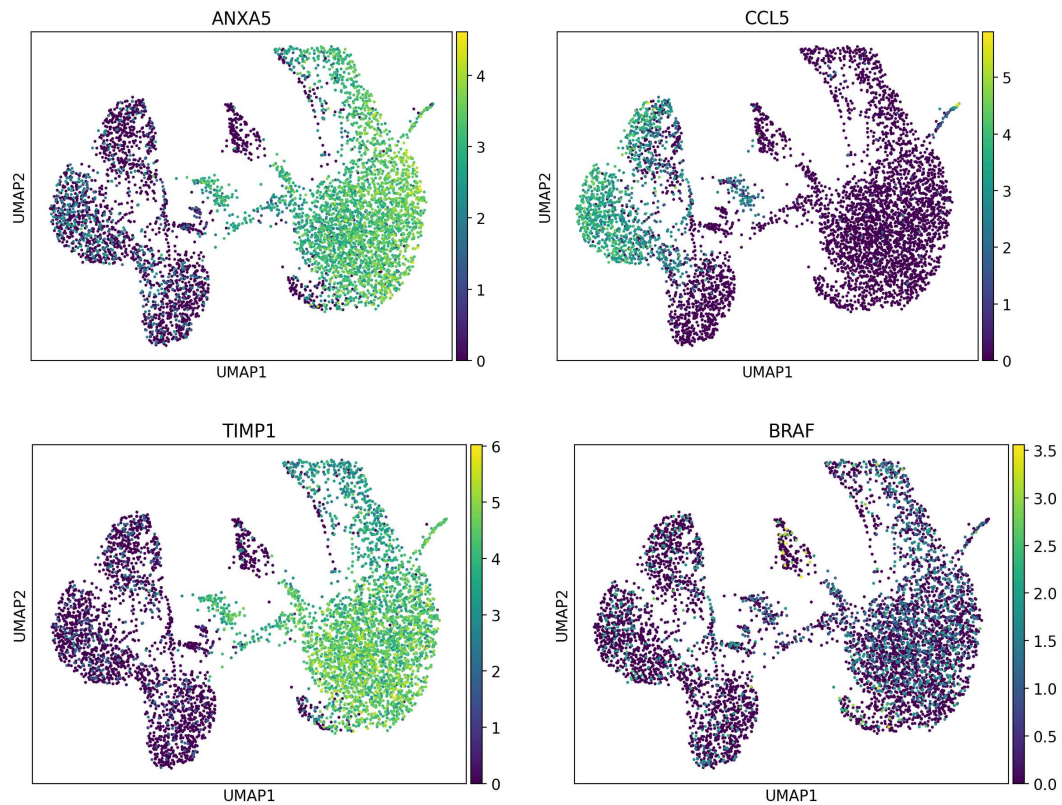


**Figure 36** UMAP Representation of Single-Cell Analysis Using the IsoWorm Single-Cell Module (Kallisto-Bustools)

- a) UMAP visualization for a single KIRP patient, with individual dots representing single cells. The color gradient of the dots indicates the expression level of the marker gene specified in each plot (CXCR4, COL1A2, AIF1, BRAF). The gradient transitions from purple to green, as shown in the legend, and corresponds to expression levels measured as  $\log_2(\text{TPM})$ . Notably, compared to other markers, BRAF is expressed at significantly lower level
- b) UMAP visualization for a single melanoma patient, with individual dots representing single cells. The color gradient of the dots indicates the expression level of the marker gene specified in each plot (CYBB, NEAT2, FTL, BRAF). The gradient transitions from purple to green, as shown in the legend, and corresponds to expression levels measured as  $\log_2(\text{TPM})$ . Notably, compared to other markers, BRAF is expressed at significantly lower level
- c) UMAP visualization for a single melanoma patient, with individual dots representing single cells. The color gradient of the dots indicates the expression level of the marker gene specified in each plot (ANXA5, CCL5, TIMP1, BRAF). The gradient transitions from purple to green, as shown in the legend, and corresponds to expression levels measured as  $\log_2(\text{TPM})$ . Notably, compared to other markers, BRAF is expressed at significantly lower level

## 9 Structural Differences Between BRAF-204 and BRAF-220

### 9.1 BRAF-204 and BRAF-220 modelling with AlphaFold

The primary challenge in studying BRAF from a structural biology perspective is that neither BRAF-204 nor BRAF-220 has been fully crystallized, likely due to the intrinsic properties of this protein [128]. This, combined with the highly flexible nature of the C-terminal regions of both isoforms, poses significant difficulties for using classical homology modeling methods, such as Modeller software. Additionally, the N-terminal region (comprising the first 1–400 amino acids) is highly enriched in loops, and this region has also never been crystallized, further complicating structural analysis. To overcome these challenges, I tried to use AlphaFold 3 webserver to predict both BRAF-220 and BRAF-204 structure using their sequences. The results are in figure 37 A and B. Both proteins show an higher pDDT score (a per-atom confidence estimate on a 0-100 scale where a higher value indicates higher confidence) on the order area of the proteins like the kinase domain, while for both proteins both the N-terms and C-terms show very low pDDT score and this makes these model not suitable for docking or other analysis. The pDDT profile (figure 38) of the two proteins are very similar showing how the BRAF isoforms are conserved with the exception of the C-terms. I will try to the next paragraph to focus my analysis on both C-terms of the two proteins.

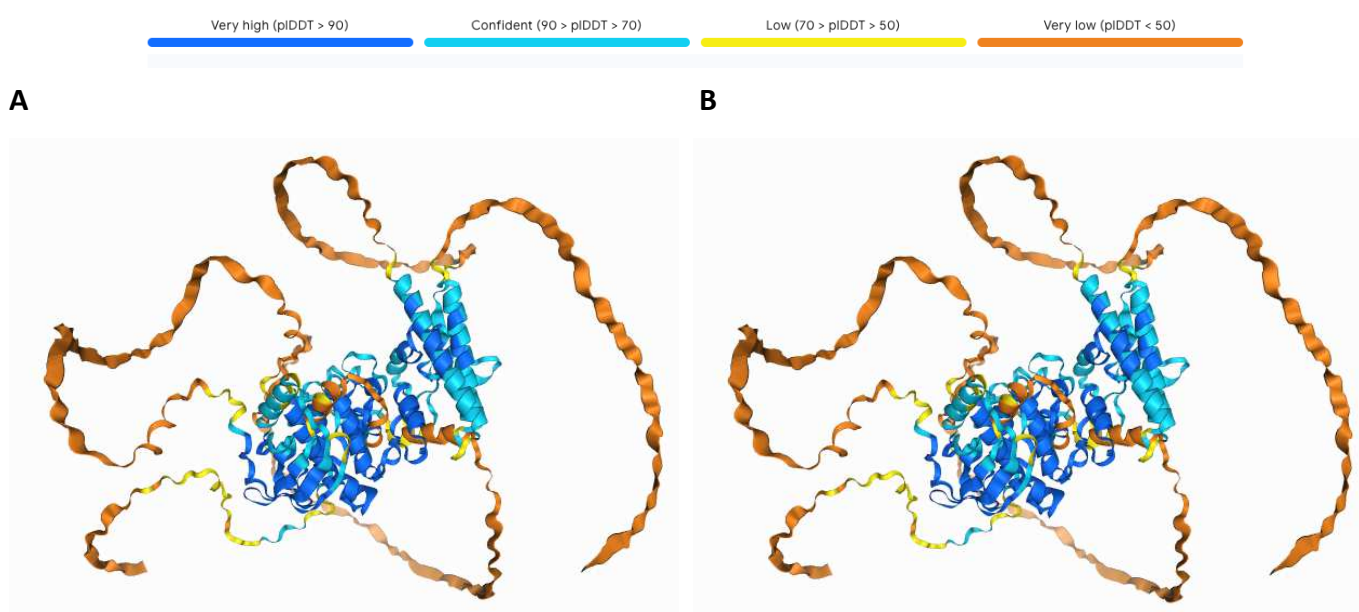


**Figure 37** Predicted structures of the BRAF-220 and the BRAF-204 complete sequences

Obtained using the AlphaFold 2 webserver. In both panels, amino acids with higher prediction accuracy are highlighted in blue and cyan, while those with lower prediction accuracy are highlighted in yellow and red.

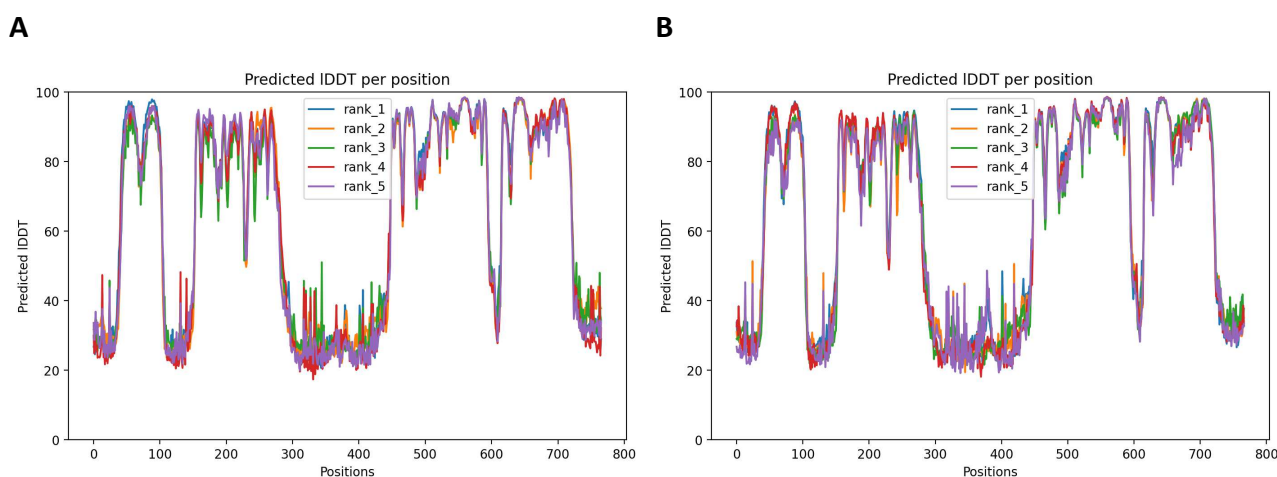


**Figure 38** Plot of predicted IDDT (Predicted Intrinsic Disorder Index) as a function of position (A BRAF-220, B BRAF-204).

The curves represent five different prediction rankings (rank\_1, rank\_2, rank\_3, rank\_4, rank\_5), showing variations in structural stability along the sequence. The X-axis indicates sequence positions, while the Y-axis represents the predicted IDDT value.

## 9.2 C-terms modelling using AlphaFold

To explore how the C-terminal amino acids of these isoforms might differ based on their composition, I decided to use *in silico* technologies to study the molecular properties of BRAF-C terms. [128] I used AlphaFold to predict the structure using only the last 36 and 37 aa of the sequence of BRAF-220 and BRAF-204 C-terms, respectively, and to construct simplified models of the two isoforms. I aimed to focus all the simulations and structural studies on the last part of the C-terms to study how their structure and the presence of an extra amino acid for BRAF-204 can modify the stability of the distal part of the C terms. This is also because the literature indicates that AlphaFold has a higher probability of accurately predicting loop regions when shorter sequences are used. As sequence length increases, the prediction accuracy tends to decrease [129].

The results obtained on the sequences -PSLNRAGFQTEDFSLYACASPKTPIQAGGYGAFPVH for BRAF-220 and PSLNRAGFQTEDFSLYACASPKTPIQAGGYGEFAAFK for BRAF-204 are displayed in **figure 39, right** and **39, left**. The predicted template modeling (pTM) score for BRAF-220 C-term is 0.46 while for BRAF-204 C-term 0.43. The difference in pTM scores (0.46 vs. 0.43) is minor, so in practical terms, this suggests that there is not a substantial difference in structural quality or reliability.

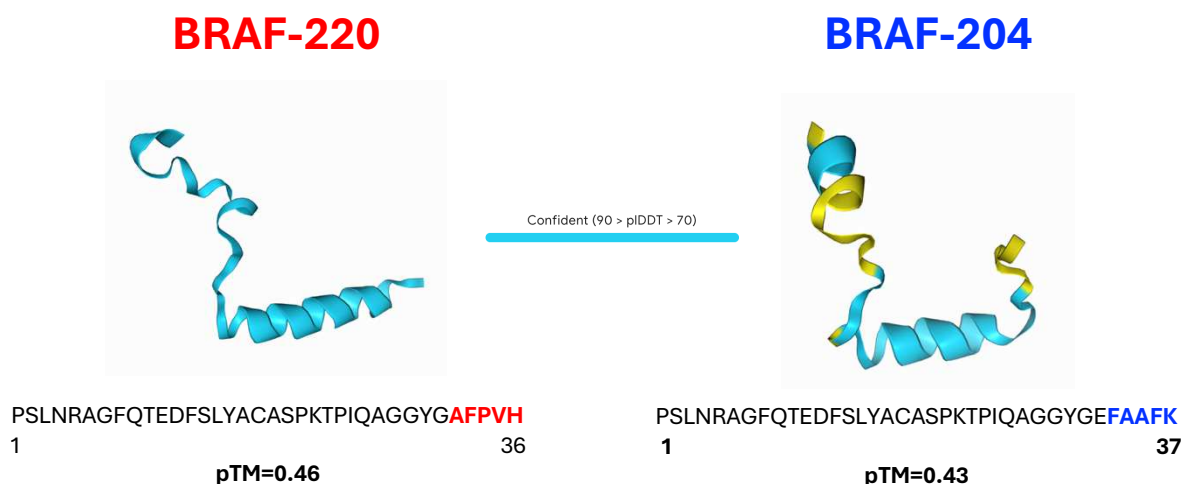


**Figure 39** Predicted structures of the C-terminal regions.

Predicted structure of the BRAF-220 (26aa, left) and the BRAF-204 (37aa, right) C-terminal peptides, obtained using the AlphaFold 2 webserver. In both panels, amino acids with higher prediction accuracy are highlighted in cyan, while those with lower prediction accuracy are highlighted in yellow. The overall predicted template modeling (pTM) score is 0.46 for BRAF-220 and 0.43 for BRAF-204.

### 9.3 Molecular Dynamic Simulations and Analysis

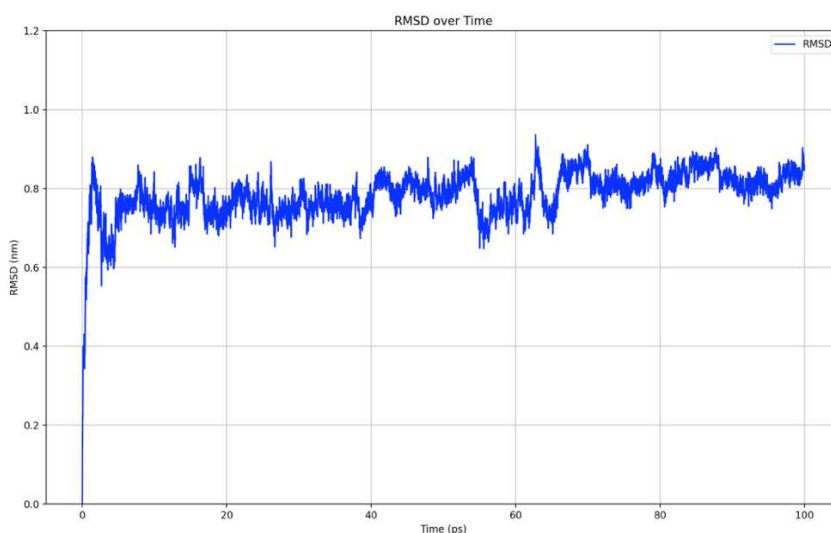
Before performing molecular dynamics simulations on the two peptides, I capped their termini by adding an ACE (acetyl) group to the N-terminus [130]. This modification effectively blocks the first amino acid, which would otherwise connect to the actual protein structure. By doing so, the peptide is prevented from forming loops or unintended interactions during the dynamic's simulation, ensuring a more accurate representation of its behavior.

After that, I performed two molecular dynamic simulations on both peptides using the PDB files generated by AlphaFold plus the ACE groups that has been added. This approach enabled me to analyze their structural dynamics, stability, and potential differences between the two molecules.

Both the peptides were solvated with physiological solution at 150 mM NaCl, heated at 36.85°C (310 K) and followed for 100 ns. The simulations were analyzed to understand the changes occurring at the atomic level in the systems and highlight the stability of the protein. I calculated the root mean square deviations (RMSDs) of the C $\alpha$  atoms of the proteins from their initial values as a function of simulation time, to measure the structural stability and monitor the conformational sampling during the MD simulation. The BRAF-204 C term (**Figure 40 a**) show more pronounced fluctuations around an average RMSD of ~0.8 nm, suggesting greater conformational flexibility or lower stability. In contrast, the BRAF-220 C-term (**Figure 40 b**) shows smoother fluctuations and a

higher average RMSD of  $\sim 1.0$  nm after 0.40 ns, indicating potentially more rigid or stable behavior. Both peptides displayed a rapid increase in RMSD during the initial phase (10ns). The BRAF-204 peptide exhibited sharp peaks in RMSD, indicative of frequent conformational rearrangements, whereas BRAF-220 peptide maintained a more consistent behavior. In general, it seems that the last the peptide that compose the C-term of BRAF-220 is more stable compared to the one of BRAF-204, maybe also because BRAF-204 as an extra amino acid.

a)



b)

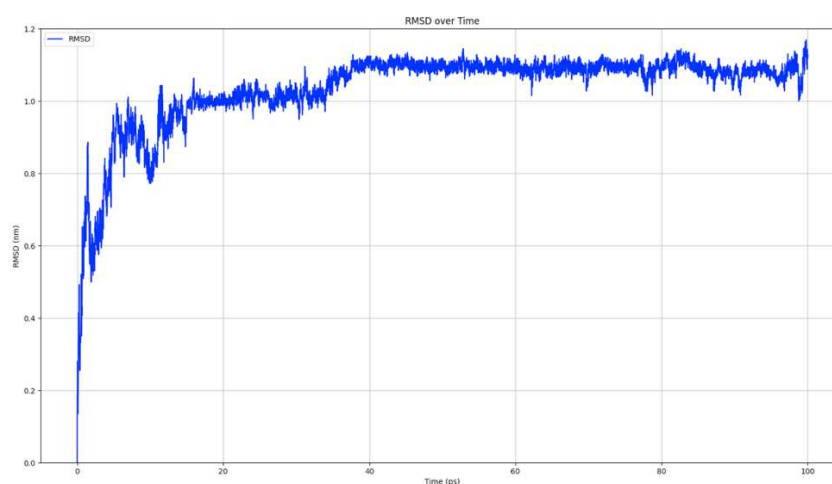
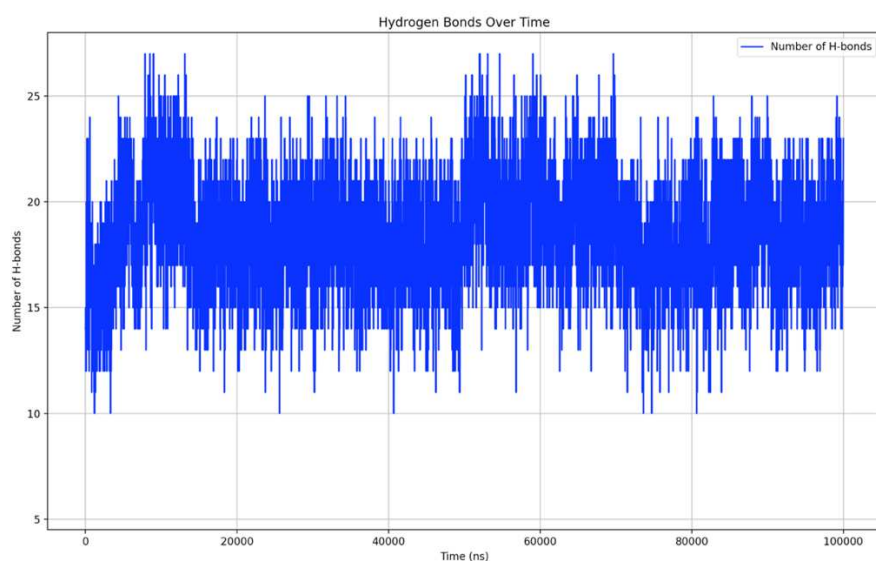


**Figure 40** RMSD values of the BRAF 204 C-terms (a) and BRAF-220 C-term (b) with respect to the equilibrated conformations as a function of simulation time.

I also analyzed hydrogen bond dynamics of the two C-term to identify differences in their behavior over time. BRAF-204 C-term has higher and more variable hydrogen bond number. It ranges between 17 and 25, with frequent rapid fluctuations. BRAF-220 C-term shows stable hydrogen bond dynamics, with fluctuations between 15 and 22 and a dip from 40 to 70 ns before recovery (Figure 41, a and b). Overall, as for the RMSD analysis, BRAF-204 C-term exhibits greater variability, while BRAF-220 C-term demonstrates a more stable and rigid behavior.

a)



b)

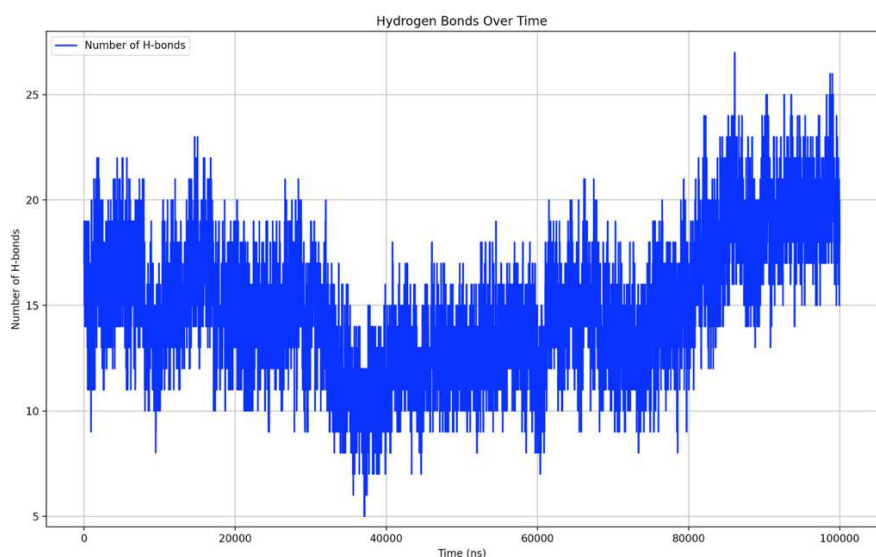
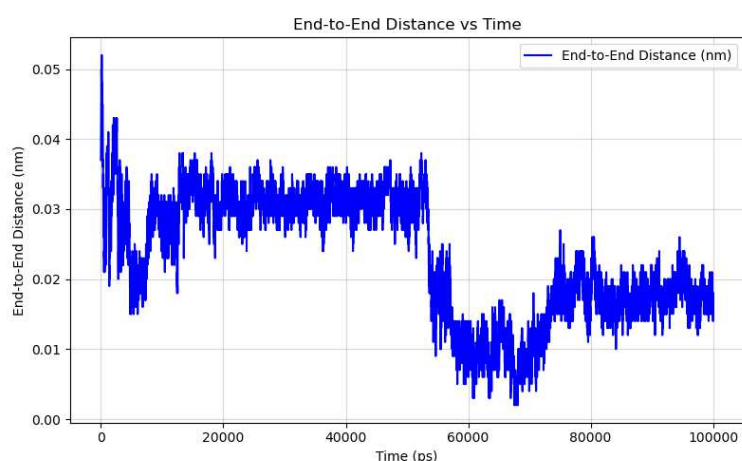


**Figure 41** Number of Hydrogen Bonds Over Time of the BRAF 204 C-terms (a) and BRAF-220 C-term (b) with respect to the equilibrated conformations as a function of simulation time (100ns)

I also analyzed the radius of gyration evolution of the end-to-end distance over time. The radius of gyration indicates the folding behaviour of the protein taken in analysis while the end-to-end distance highlights the molecular extension or compaction, conformational transitions (e.g., folding or collapse), and dynamic structural changes over time. In the BRAF-204 plots, **Figure 42 a**, the distance exhibits a sharp transition around 50,000 ps, decreasing significantly and stabilizing at a lower value ( $\sim 0.015$  nm) in the second half. In contrast, the BRAF-220 plot, **Figure 42 b**, shows a more gradual decrease, stabilizing earlier at an even lower value ( $\sim 0.01$  nm), with fewer pronounced fluctuations. Additionally, the initial range of end-to-end distances is higher in the second plot (up to 0.06 nm) compared to the first plot (up to 0.05 nm). Also, these differences indicate distinct behaviors in the dynamics or folding processes between the BRAF-204 and BRAF-220 proteins.

a)



b)

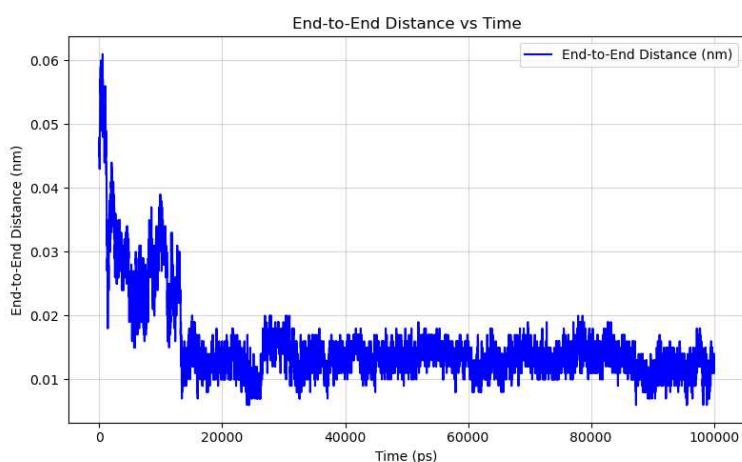


**Figure 42** Time evolution of the end-to-end distance, showing a structural transition and stabilization at later time points and lower values. a) BRAF-204; b) BRAF-220.

To conclude, the *in silico* structural biology analysis reveals some differences and behavioral variations between the C-termini of the two molecules, BRAF-220 and BRAF-204. Without the presence of a cofactor or a known ligand, further analyses, such as docking procedures or interaction studies, cannot be conducted at this stage. However, these initial findings highlight the importance of investigating how these two isoforms behave *in vitro* and *in vivo*.

The dynamic simulations demonstrate that the C-termini of the two proteins exhibit distinct properties: one being more flexible (BRAF-204) while the other remains rigid (BRAF-220). Lastly, the use of techniques and tools such as AlphaFold metainference could help me obtain a more complete and computationally useful structure for further analyses in the future [131].

## Conclusions

Quantifying full-length transcript variants from sequencing data poses significant challenges from both technical and biological points of view. The case of BRAF is particularly complex because the isoforms reported in databases are many, very similar and rapidly changing and because the actual length of each isoform varies within and across samples due to the use of alternative polyA sites. We built IsoWorm, to quantify isoforms starting from short reads RNA-seq data, single-cell data and to identify poly-A sites starting from Quant 3' seq data. Thanks to the analysis of several datasets through the combined use of multiple computational approaches, we established that BRAF is a mix of two full-length, protein-coding and functional isoforms. Crucially, in human cancers, *BRAF-204* mRNA is the predominant isoform compared to *BRAF-220* mRNA, with the notable exception of acute myeloid leukemia, which highlights a potential difference in transcript expression between solid and liquid tumors. Furthermore, the *BRAF-204/BRAF-220* ratio appears to influence patient survival in KIRP, with *BRAF-204* potentially involved in a regulatory network alongside the well-known oncogene TFAP2A. This hypothesis will require further validation in future studies. Lastly, the molecular dynamics simulation and the analysis of the trajectories generated by it provided evidence on the C-term of BRAF-220 having a different behavior compared to the C-term of BRAF-204. Lower RMSD and more stable hydrogen bond dynamics are observed in BRAF-220, in contrast with the higher fluctuation and variability observed in the BRAF-204 dynamics, these findings support the need for a more comprehensive approach to studying BRAF isoforms, emphasizing the importance of developing a complete structural model. A fully detailed model, including the C-terminal regions, is crucial for structural bioinformatics studies, allowing for a deeper understanding of the isoforms' functional and structural differences. Historically, BRAF-220 has been considered the only BRAF isoform and has represented the undivided focus of in vitro and in vivo model systems[132] [133]. Conversely, the *in silico* analyses presented here, together with the experimental characterization described by us and others in [7], [16], [17], [18], [19], [134] introduce a paradigm shift in BRAF-centered studies [135], prompting the benchmarking of *BRAF-204* against *BRAF-220* in terms of expression, function, drug sensitivity and clinical relevance.

## References

- [1] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912–1934, Dec. 2002, doi: 10.1126/SCIENCE.1075762.
- [2] R. Roskoski, "RAF protein-serine/threonine kinases: Structure and regulation," *Biochem Biophys Res Commun*, vol. 399, no. 3, pp. 313–317, Aug. 2010, doi: 10.1016/J.BBRC.2010.07.092.
- [3] E. Park *et al.*, "Cryo-EM structure of a RAS/RAF recruitment complex," *Nature Communications* 2023 14:1, vol. 14, no. 1, pp. 1–11, Jul. 2023, doi: 10.1038/s41467-023-40299-6.
- [4] M. E. Bahar, H. J. Kim, and D. R. Kim, "Targeting the RAS/RAF/MAPK pathway for cancer therapy: from mechanism to clinical studies," *Signal Transduction and Targeted Therapy* 2023 8:1, vol. 8, no. 1, pp. 1–38, Dec. 2023, doi: 10.1038/s41392-023-01705-z.
- [5] E. M. Terrell and D. K. Morrison, "Ras-Mediated Activation of the Raf Family Kinases," *Cold Spring Harb Perspect Med*, vol. 9, no. 1, p. a033746, Jan. 2019, doi: 10.1101/CSHPERSPECT.A033746.
- [6] L. L. Ritterhouse and J. A. Barletta, "BRAF V600E mutation-specific antibody: A review," *Semin Diagn Pathol*, vol. 32, no. 5, pp. 400–408, Sep. 2015, doi: 10.1053/J.SEMDP.2015.02.010.
- [7] A. Marranci *et al.*, "The landscape of BRAF transcript and protein variants in human cancer," *Mol Cancer*, vol. 16, no. 1, Apr. 2017, doi: 10.1186/S12943-017-0645-4.
- [8] P. A. Ascierto *et al.*, "The role of BRAF V600 mutation in melanoma," *J Transl Med*, vol. 10, no. 1, May 2012, doi: 10.1186/1479-5876-10-85.
- [9] J. Ule and B. J. Blencowe, "Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution," *Mol Cell*, vol. 76, no. 2, pp. 329–345, Oct. 2019, doi: 10.1016/J.MOLCEL.2019.09.017/ASSET/78991C78-C127-402B-AEBF-7705933665C4/MAIN.ASSETS/GR4.JPG.
- [10] Q. Liu, L. Fang, and C. Wu, "Alternative Splicing and Isoforms: From Mechanisms to Diseases," *Genes* 2022, Vol. 13, Page 401, vol. 13, no. 3, p. 401, Feb. 2022, doi: 10.3390/GENES13030401.
- [11] P. Jordan, V. Gonçalves, S. Fernandes, T. Marques, M. Pereira, and M. Gama-Carvalho, "Networks of mRNA Processing and Alternative Splicing Regulation in Health and Disease," *Adv Exp Med Biol*, vol. 1157, pp. 1–27, 2019, doi: 10.1007/978-3-030-19966-1\_1.
- [12] S. Pal, R. Gupta, and R. V. Davuluri, "Alternative transcription and alternative splicing in cancer," *Pharmacol Ther*, vol. 136, no. 3, pp. 283–294, Dec. 2012, doi: 10.1016/J.PHARMTHERA.2012.08.005.
- [13] A. E. Erson-Bensan, "RNA-biology ruling cancer progression? Focus on 3'UTRs and splicing," *Cancer Metastasis Rev*, vol. 39, no. 3, pp. 887–901, Sep. 2020, doi: 10.1007/S10555-020-09884-9.
- [14] A. Marranci *et al.*, "Identification of BRAF 3'UTR isoforms in melanoma," *Journal of Investigative Dermatology*, vol. 135, no. 6, 2015, doi: 10.1038/jid.2015.47.
- [15] A. Marranci *et al.*, "The landscape of BRAF transcript and protein variants in human cancer," *Mol Cancer*, vol. 16, no. 1, 2017, doi: 10.1186/s12943-017-0645-4.

- [16] R. De Paolo *et al.*, "Differential impact of BRAFV600E isoforms on tumorigenesis in a zebrafish model of melanoma," *Cell Biosci*, vol. 13, no. 1, pp. 1–5, Dec. 2023, doi: 10.1186/S13578-023-01064-W/FIGURES/1.
- [17] A. Marranci *et al.*, "Systematic evaluation of the microRNAome through miR-CATCHv2.0 identifies positive and negative regulators of BRAF-X1 mRNA," *RNA Biol*, vol. 16, no. 7, pp. 865–878, Jul. 2019, doi: 10.1080/15476286.2019.1600934.
- [18] A. Marranci *et al.*, "PARP1 negatively regulates MAPK signaling by impairing BRAF-X1 translation," *J Hematol Oncol*, vol. 16, no. 1, Dec. 2023, doi: 10.1186/S13045-023-01428-2.
- [19] S. Lubrano *et al.*, "Development of a yeast-based system to identify new hBRAFV600E functional interactors," *Oncogene*, vol. 38, no. 8, pp. 1355–1366, Feb. 2019, doi: 10.1038/S41388-018-0496-5.
- [20] A. V. Orang, R. Safaralizadeh, and M. Kazemzadeh-Bavili, "Mechanisms of miRNA-Mediated Gene Regulation from Common Downregulation to mRNA-Specific Upregulation," *Int J Genomics*, vol. 2014, no. 1, p. 970607, Jan. 2014, doi: 10.1155/2014/970607.
- [21] H. Oki Iwakawa and Y. Tomari, "The Functions of MicroRNAs: mRNA Decay and Translational Repression," *Trends Cell Biol*, vol. 25, no. 11, pp. 651–665, Nov. 2015, doi: 10.1016/J.TCB.2015.07.011.
- [22] A. Rhoads and K. F. Au, "PacBio Sequencing and Its Applications," *Genomics Proteomics Bioinformatics*, vol. 13, no. 5, pp. 278–289, 2015, doi: 10.1016/J.GPB.2015.08.002.
- [23] R. J. Motzer *et al.*, "NCCN Guidelines Insights: Kidney Cancer, Version 2.2020," *J Natl Compr Canc Netw*, vol. 17, no. 11, pp. 1278–1285, 2019, doi: 10.6004/JNCCN.2019.0054.
- [24] S. Dawood, L. Austin, M. C.-O. (Williston, and undefined 2014, "Cancer stem cells: implications for cancer therapy," *showcase.rxprism.com* S Dawood, L Austin, M Cristofanilli *Oncology (Williston Park)*, 2014 • *showcase.rxprism.com*, vol. 15, 2014, Accessed: Dec. 27, 2024. [Online]. Available: [http://showcase.rxprism.com/microsite/eModules/Product\\_Module/contents/Resources/Dawood\\_Oncology\\_2014.pdf](http://showcase.rxprism.com/microsite/eModules/Product_Module/contents/Resources/Dawood_Oncology_2014.pdf)
- [25] K. H. Bai *et al.*, "Identification of cancer stem cell characteristics in liver hepatocellular carcinoma by WGCNA analysis of transcriptome stemness index," *Cancer Med*, vol. 9, no. 12, p. 4290, Jun. 2020, doi: 10.1002/CAM4.3047.
- [26] M. L. Troxell and J. P. Higgins, "Renal cell carcinoma in kidney allografts: histologic types, including biphasic papillary carcinoma," *Hum Pathol*, vol. 57, pp. 28–36, Nov. 2016, doi: 10.1016/J.HUMPATH.2016.06.018.
- [27] B. Delahunt, J. N. Eble, M. R. E. McCredie, P. B. Bethwaite, J. H. Stewart, and A. M. Bilous, "Morphologic typing of papillary renal cell carcinoma: Comparison of growth kinetics and patient survival in 66 cases," *Hum Pathol*, vol. 32, no. 6, pp. 590–595, 2001, doi: 10.1053/hupa.2001.24984.
- [28] Y. Allory, D. Ouazana, E. Boucher, N. Thiounn, and A. Vieillefond, "Papillary renal cell carcinoma. Prognostic value of morphological subtypes in a clinicopathologic study of 43 cases," *Virchows Arch*, vol. 442, no. 4, pp. 336–342, Apr. 2003, doi: 10.1007/S00428-003-0787-1.
- [29] S. McGuire, "World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015," *Adv Nutr*, vol. 7, no. 2, pp. 418–419, 2016, doi: 10.3945/AN.116.012211.
- [30] B. M. Shinder *et al.*, "Surgical Management of Advanced and Metastatic Renal Cell Carcinoma: A Multidisciplinary Approach," *Front Oncol*, vol. 7, no. MAY, May 2017, doi: 10.3389/FONC.2017.00107.

- [31] C. Jin *et al.*, "Crucial role of the transcription factors family activator protein 2 in cancer: current clue and views," *Journal of Translational Medicine* 2023 21:1, vol. 21, no. 1, pp. 1–21, Jun. 2023, doi: 10.1186/S12967-023-04189-1.
- [32] D. Eckert, S. Buhl, S. Weber, R. Jäger, and H. Schorle, "The AP-2 family of transcription factors," *Genome Biol*, vol. 6, no. 13, Dec. 2005, doi: 10.1186/GB-2005-6-13-246.
- [33] H. Yamashita *et al.*, "Repression of transcription factor AP-2 alpha by PPAR $\gamma$  reveals a novel transcriptional circuit in basal-squamous bladder cancer," *Oncogenesis*, vol. 8, no. 12, Dec. 2019, doi: 10.1038/S41389-019-0178-3.
- [34] J. Wu, "Pancreatic Cancer-Derived Exosomes Promote the Proliferation, Invasion, and Metastasis of Pancreatic Cancer by the miR-3960/TFAP2A Axis," *J Oncol*, vol. 2022, 2022, doi: 10.1155/2022/3590326.
- [35] L. Cao *et al.*, "Zinc-finger protein 471 suppresses gastric cancer through transcriptionally repressing downstream oncogenic PLS3 and TFAP2A," *Oncogene*, vol. 37, no. 26, pp. 3601–3616, Jun. 2018, doi: 10.1038/S41388-018-0220-5.
- [36] Z. Guoren *et al.*, "TFAP2A Induced ITPKA Serves as an Oncogene and Interacts with DBN1 in Lung Adenocarcinoma," *Int J Biol Sci*, vol. 16, no. 3, pp. 504–514, 2020, doi: 10.7150/IJBS.40435.
- [37] D. Shi *et al.*, "TFAP2A regulates nasopharyngeal carcinoma growth and survival by targeting HIF-1 $\alpha$  signaling pathway," *Cancer Prev Res (Phila)*, vol. 7, no. 2, pp. 266–277, Feb. 2014, doi: 10.1158/1940-6207.CAPR-13-0271.
- [38] H. Xu, L. Wang, and X. Jiang, "Silencing of lncRNA DLEU1 inhibits tumorigenesis of ovarian cancer via regulating miR-429/TFAP2A axis," *Mol Cell Biochem*, vol. 476, no. 2, pp. 1051–1061, Feb. 2021, doi: 10.1007/S11010-020-03971-9.
- [39] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén, "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release," *Anal Biochem*, vol. 242, no. 1, pp. 84–89, Nov. 1996, doi: 10.1006/ABIO.1996.0432.
- [40] H. Satam *et al.*, "Next-Generation Sequencing Technology: Current Trends and Advancements," *Biology (Basel)*, vol. 12, no. 7, p. 997, Jul. 2023, doi: 10.3390/BIOLOGY12070997.
- [41] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, "Opportunities and challenges in long-read sequencing data analysis," *Genome Biol*, vol. 21, no. 1, Feb. 2020, doi: 10.1186/S13059-020-1935-5.
- [42] "PacBio Sequencing and Its Applications | Enhanced Reader."
- [43] R. W. Holley *et al.*, "STRUCTURE OF A RIBONUCLEIC ACID," *Science*, vol. 147, no. 3664, pp. 1462–1465, 1965, doi: 10.1126/SCIENCE.147.3664.1462.
- [44] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA," *Genomics*, vol. 107, no. 1, pp. 1–8, Jan. 2016, doi: 10.1016/J.YGENO.2015.11.003.
- [45] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc Natl Acad Sci U S A*, vol. 74, no. 12, pp. 5463–5467, 1977, doi: 10.1073/PNAS.74.12.5463.
- [46] T. Hu, N. Chitnis, D. Monos, and A. Dinh, "Next-generation sequencing technologies: An overview," *Hum Immunol*, vol. 82, no. 11, pp. 801–811, Nov. 2021, doi: 10.1016/J.HUMIMM.2021.02.012.
- [47] T. Mantere, S. Kersten, and A. Hoischen, "Long-read sequencing emerging in medical genetics," *Front Genet*, vol. 10, no. MAY, p. 432668, May 2019, doi: 10.3389/FGENE.2019.00426/BIBTEX.

- [48] Y. Sakamoto, S. Sereewattanawoot, and A. Suzuki, "A new era of long-read sequencing for cancer genomics," *J Hum Genet*, vol. 65, no. 1, p. 3, Jan. 2019, doi: 10.1038/S10038-019-0658-5.
- [49] A. S. Kainth, G. A. Haddad, J. M. Hall, and A. J. Ruthenburg, "Merging short and stranded long reads improves transcript assembly," *PLoS Comput Biol*, vol. 19, no. 10, p. e1011576, Oct. 2023, doi: 10.1371/JOURNAL.PCBI.1011576.
- [50] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLoS Comput Biol*, vol. 13, no. 5, p. e1005457, May 2017, doi: 10.1371/JOURNAL.PCBI.1005457.
- [51] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet*, vol. 10, no. 1, p. 57, Jan. 2009, doi: 10.1038/NRG2484.
- [52] M. J. Heller, "DNA microarray technology: devices, systems, and applications," *Annu Rev Biomed Eng*, vol. 4, pp. 129–153, 2002, doi: 10.1146/ANNUREV.BIOENG.4.020702.153438.
- [53] D. Sarantopoulou, T. G. Brooks, S. Nayak, A. Mrčela, N. F. Lahens, and G. R. Grant, "Comparative evaluation of full-length isoform quantification from RNA-Seq," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–24, Dec. 2021, doi: 10.1186/S12859-021-04198-1/FIGURES/12.
- [54] A. Conesa *et al.*, "A survey of best practices for RNA-seq data analysis," *Genome Biol*, vol. 17, no. 1, Jan. 2016, doi: 10.1186/S13059-016-0881-8.
- [55] A. Oshlack, M. D. Robinson, and M. D. Young, "From RNA-seq reads to differential expression results," *Genome Biol*, vol. 11, no. 12, Dec. 2010, doi: 10.1186/GB-2010-11-12-220.
- [56] 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.
- [57] B. Li and C. N. Dewey, "RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–16, Aug. 2011, doi: 10.1186/1471-2105-12-323/TABLES/6.
- [58] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods* 2017 14:4, vol. 14, no. 4, pp. 417–419, Mar. 2017, doi: 10.1038/nmeth.4197.
- [59] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nat Biotechnol*, vol. 34, no. 5, pp. 525–527, 2016, doi: 10.1038/nbt.3519.
- [60] P. Moll, M. Ante, A. Seitz, and T. Reda, "QuantSeq 3' mRNA sequencing for RNA quantification," *Nature Methods* 2014 11:12, vol. 11, no. 12, pp. i–iii, Nov. 2014, doi: 10.1038/nmeth.f.376.
- [61] C. Cheng, W. Chen, H. Jin, and X. Chen, "A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell–Cell Communication," *Cells*, vol. 12, no. 15, p. 1970, Aug. 2023, doi: 10.3390/CELLS12151970.
- [62] G. Chen, B. Ning, and T. Shi, "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis," *Front Genet*, vol. 10, no. APR, p. 317, 2019, doi: 10.3389/FGENE.2019.00317.
- [63] W. A. Awuah *et al.*, "Exploring the current landscape of single-cell RNA sequencing applications in gastric cancer research," *J Cell Mol Med*, vol. 28, no. 7, p. e18159, Apr. 2024, doi: 10.1111/JCMM.18159.
- [64] X. Qiu, A. Hill, J. Packer, D. Lin, Y. A. Ma, and C. Trapnell, "Single-cell mRNA quantification and differential analysis with Census," *Nat Methods*, vol. 14, no. 3, p. 309, Feb. 2017, doi: 10.1038/NMETH.4150.

- [65] L. Van der Maaten, G. H.-J. of machine learning research, and undefined 2008, "Visualizing data using t-SNE.," *jmlr.org* L Van der Maaten, G Hinton *Journal of machine learning research*, 2008•*jmlr.org*, Accessed: Dec. 15, 2024. [Online]. Available: <https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl>
- [66] E. Becht *et al.*, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nat Biotechnol*, vol. 37, no. 1, pp. 38–47, Jan. 2018, doi: 10.1038/NBT.4314.
- [67] E. Espinosa, R. Bautista, R. Larrosa, and O. Plata, "Advancements in long-read genome sequencing technologies and algorithms," *Genomics*, vol. 116, no. 3, p. 110842, May 2024, doi: 10.1016/J.YGENO.2024.110842.
- [68] A. Payne, N. Holmes, V. Rakyar, and M. Loose, "BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files," *Bioinformatics*, vol. 35, no. 13, pp. 2193–2198, Jul. 2019, doi: 10.1093/BIOINFORMATICS/BTY841.
- [69] K. M. Nip *et al.*, "Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2," *Nature Communications* 2023 14:1, vol. 14, no. 1, pp. 1–14, May 2023, doi: 10.1038/s41467-023-38553-y.
- [70] S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, and M. Pertea, "Transcriptome assembly from long-read RNA-seq alignments with StringTie2," *Genome Biol*, vol. 20, no. 1, pp. 1–13, Dec. 2019, doi: 10.1186/S13059-019-1910-1/FIGURES/6.
- [71] A. D. Tang *et al.*, "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns," *Nature Communications* 2020 11:1, vol. 11, no. 1, pp. 1–12, Mar. 2020, doi: 10.1038/s41467-020-15171-6.
- [72] J. Lonsdale *et al.*, "The Genotype-Tissue Expression (GTEx) project," *Nature Genetics* 2013 45:6, vol. 45, no. 6, pp. 580–585, May 2013, doi: 10.1038/ng.2653.
- [73] J. Barretina *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012, doi: 10.1038/NATURE11003.
- [74] M. Ghandi *et al.*, "Next-generation characterization of the Cancer Cell Line Encyclopedia," *Nature*, vol. 569, no. 7757, pp. 503–508, May 2019, doi: 10.1038/S41586-019-1186-3.
- [75] A. Basu *et al.*, "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules," *Cell*, vol. 154, no. 5, Aug. 2013, doi: 10.1016/J.CELL.2013.08.003.
- [76] D. P. Nusinow *et al.*, "Quantitative Proteomics of the Cancer Cell Line Encyclopedia," *Cell*, vol. 180, no. 2, p. 387, Jan. 2020, doi: 10.1016/J.CELL.2019.12.023.
- [77] Q. Shi, X. Li, Y. Liu, Z. Chen, and X. He, "FLIbase: a comprehensive repository of full-length isoforms across human cancers and tissues," *Nucleic Acids Res*, Sep. 2023, doi: 10.1093/NAR/GKAD745.
- [78] J. M. Rodriguez *et al.*, "APPRIS: annotation of principal and alternative splice isoforms", doi: 10.1093/nar/gks1058.
- [79] A. W. Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature* 2020 577:7792, vol. 577, no. 7792, pp. 706–710, Jan. 2020, doi: 10.1038/s41586-019-1923-7.
- [80] M. Alquraishi, "AlphaFold at CASP13," *Bioinformatics*, vol. 35, no. 22, pp. 4862–4865, Nov. 2019, doi: 10.1093/BIOINFORMATICS/BTZ422.
- [81] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature* 2021 596:7873, vol. 596, no. 7873, pp. 583–589, Jul. 2021, doi: 10.1038/s41586-021-03819-2.
- [82] M. Karplus, "Molecular dynamics simulations of biomolecules," *Acc Chem Res*, vol. 35, no. 6, pp. 321–323, 2002, doi: 10.1021/AR020082R.

- [83] Q. Shi, X. Li, Y. Liu, Z. Chen, and X. He, "FLIBase: a comprehensive repository of full-length isoforms across human cancers and tissues," *Nucleic Acids Res*, vol. 52, no. D1, pp. D124–D133, Jan. 2024, doi: 10.1093/nar/gkad745.
- [84] J. Liu *et al.*, "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics," *Cell*, vol. 173, no. 2, pp. 400–416.e11, Apr. 2018, doi: 10.1016/j.cell.2018.02.052.
- [85] Q. Yi *et al.*, "Spectrum of BRAF Aberrations and Its Potential Clinical Implications: Insights From Integrative Pan-Cancer Analysis," *Front Bioeng Biotechnol*, vol. 10, Jul. 2022, doi: 10.3389/fbioe.2022.806851.
- [86] Y. Kondo *et al.*, "Cryo-EM structure of a dimeric B-Raf:14-3-3 complex reveals asymmetry in the active sites of B-Raf kinases," *Science (1979)*, vol. 366, no. 6461, pp. 109–115, Oct. 2019, doi: 10.1126/SCIENCE.AAY0543.
- [87] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature* 2021 596:7873, vol. 596, no. 7873, pp. 583–589, Jul. 2021, doi: 10.1038/s41586-021-03819-2.
- [88] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: fast, flexible, and free," *J Comput Chem*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005, doi: 10.1002/JCC.20291.
- [89] Y. Duan *et al.*, "A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations," *J Comput Chem*, vol. 24, no. 16, pp. 1999–2012, Dec. 2003, doi: 10.1002/JCC.10349.
- [90] M. Turisini, M. Cestari, and G. Amati, "LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI applications," *Journal of large-scale research facilities JLSRF*, vol. 9, no. 1, p. 186, Jan. 2024, doi: 10.17815/JLSRF-8-186.
- [91] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J Chem Phys*, vol. 79, no. 2, pp. 926–935, Jul. 1983, doi: 10.1063/1.445869.
- [92] S. S. Petrova and A. D. Solov'Ev, "The Origin of the Method of Steepest Descent," *Historia Mathematica*, vol. 24, no. 4, pp. 361–375, Nov. 1997, doi: 10.1006/HMAT.1996.2146.
- [93] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J Chem Phys*, vol. 126, no. 1, 2007, doi: 10.1063/1.2408420.
- [94] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J Appl Phys*, vol. 52, no. 12, pp. 7182–7190, 1981, doi: 10.1063/1.328693.
- [95] H. Bekker, "LINCS: A linear constraint solver for molecular simulations," *J Comput Chem*, Jan. 1997, Accessed: Jan. 02, 2025. [Online]. Available: [https://www.academia.edu/17267486/LINCS\\_A\\_linear\\_constraint\\_solver\\_for\\_molecular\\_simulations](https://www.academia.edu/17267486/LINCS_A_linear_constraint_solver_for_molecular_simulations)
- [96] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *J Mol Graph*, vol. 14, no. 1, pp. 33–38, 1996, doi: 10.1016/0263-7855(96)00018-5.
- [97] D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon, "Theoretical Perspectives on Protein Folding," *Annu Rev Biophys*, vol. 39, no. 1, pp. 159–183, Apr. 2010, doi: 10.1146/annurev-biophys-051309-103835.
- [98] F. Aya *et al.*, "Genomic deletions explain the generation of alternative BRAF isoforms conferring resistance to MAPK inhibitors in melanoma," *Cell Rep*, vol. 43, no. 4, p. 114048, Apr. 2024, doi: 10.1016/j.celrep.2024.114048.
- [99] N. K. Hayward *et al.*, "Whole-genome landscapes of major melanoma subtypes," *Nature*, vol. 545, no. 7653, pp. 175–180, 2017, doi: 10.1038/nature22071.

- [100] K. Kemper *et al.*, "BRAF(V600E) Kinase Domain Duplication Identified in Therapy-Refractory Melanoma Patient-Derived Xenografts," *Cell Rep*, vol. 16, no. 1, pp. 263–277, 2016, doi: 10.1016/j.celrep.2016.05.064.
- [101] J. Morales *et al.*, "A joint NCBI and EMBL-EBI transcript set for clinical genomics and research," *Nature*, vol. 604, no. 7905, pp. 310–315, Apr. 2022, doi: 10.1038/s41586-022-04558-8.
- [102] J. M. Rodriguez, F. Pozo, D. Cerdán-Velez, T. Di Domenico, J. Vázquez, and M. L. Tress, "APPRIS: selecting functionally important isoforms," *Nucleic Acids Res*, vol. 50, no. D1, pp. D54–D59, Jan. 2022, doi: 10.1093/NAR/GKAB1058.
- [103] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nat Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017, doi: 10.1038/nmeth.4197.
- [104] Y. Zhao *et al.*, "TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository," *J Transl Med*, vol. 19, no. 1, p. 269, Dec. 2021, doi: 10.1186/s12967-021-02936-w.
- [105] E. Vendramini *et al.*, "KRAS, NRAS, and BRAF mutations are highly enriched in trisomy 12 chronic lymphocytic leukemia and are associated with shorter treatment-free survival," *Leukemia*, vol. 33, no. 8, pp. 2111–2115, Aug. 2019, doi: 10.1038/S41375-019-0444-6.
- [106] N. Giménez *et al.*, "Mutations in the RAS-BRAF-MAPK-ERK pathway define a specific subgroup of patients with adverse clinical features and provide new therapeutic options in chronic lymphocytic leukemia," *Haematologica*, vol. 104, no. 3, pp. 576–586, Feb. 2019, doi: 10.3324/HAEMATOL.2018.196931.
- [107] F. Fei *et al.*, "BRAF Mutations in Patients with Myeloid Neoplasms: A Cancer Center Multigene Next-Generation Sequencing Analysis Experience," *Int J Mol Sci*, vol. 25, no. 10, May 2024, doi: 10.3390/IJMS25105183.
- [108] A. Zingone *et al.*, "A comprehensive map of alternative polyadenylation in African American and European American lung cancer patients," *Nat Commun*, vol. 12, no. 1, p. 5605, Sep. 2021, doi: 10.1038/s41467-021-25763-5.
- [109] D. Sarantopoulou, T. G. Brooks, S. Nayak, A. Mrčela, N. F. Lahens, and G. R. Grant, "Comparative evaluation of full-length isoform quantification from RNA-Seq," *BMC Bioinformatics*, vol. 22, no. 1, p. 266, Dec. 2021, doi: 10.1186/s12859-021-04198-1.
- [110] H. Lan, J. Zeng, G. Chen, and H. Huang, "Survival prediction of kidney renal papillary cell carcinoma by comprehensive LncRNA characterization," *Oncotarget*, vol. 8, no. 67, pp. 110811–110829, Dec. 2017, doi: 10.18632/oncotarget.22732.
- [111] Y. Liu *et al.*, "Identification of prognostic stemness-related genes in kidney renal papillary cell carcinoma," *BMC Med Genomics*, vol. 17, no. 1, p. 121, May 2024, doi: 10.1186/s12920-024-01870-2.
- [112] A. Lim *et al.*, "Designing a predictive Framework: Immune-Related Gene-Based nomogram and prognostic model for kidney renal papillary cell carcinoma," *Int Immunopharmacol*, vol. 131, p. 111878, Apr. 2024, doi: 10.1016/j.intimp.2024.111878.
- [113] X. Li, Y. Liu, L. Zhou, J. Wang, and X. Yang, "Construction and validation of a mutation-related model in papillary renal cell carcinoma and associated immune infiltration," *Kidney Blood Press Res*, May 2024, doi: 10.1159/000539096.
- [114] F. Li, Z. Wu, Z. Du, Q. Ke, Y. Fu, and J. Zhan, "Comprehensive molecular analyses and experimental validation of CDCAs with potential implications in kidney renal papillary cell

- carcinoma prognosis," *Heliyon*, vol. 10, no. 12, p. e33045, Jun. 2024, doi: 10.1016/j.heliyon.2024.e33045.
- [115] X. Gong, Y. Gong, G. Wu, and H. Ke, "Bioinformatics analysis highlights CCNB1 as a potential prognostic biomarker and an anti-kidney renal papillary cell carcinoma drug target," *Medicine*, vol. 103, no. 12, p. e37609, Mar. 2024, doi: 10.1097/MD.00000000000037609.
  - [116] X. Li, Y. Liu, L. Zhou, J. Wang, and X. Yang, "Construction and Validation of a Mutation-Related Model in Papillary Renal Cell Carcinoma and Associated Immune Infiltration," *Kidney Blood Press Res*, vol. 49, no. 1, pp. 355–367, May 2024, doi: 10.1159/000539096.
  - [117] A. Lim *et al.*, "Designing a predictive Framework: Immune-Related Gene-Based nomogram and prognostic model for kidney renal papillary cell carcinoma," *Int Immunopharmacol*, vol. 131, Apr. 2024, doi: 10.1016/J.INTIMP.2024.111878.
  - [118] Y. Liu *et al.*, "Identification of prognostic stemness-related genes in kidney renal papillary cell carcinoma," *BMC Med Genomics*, vol. 17, no. 1, Dec. 2024, doi: 10.1186/S12920-024-01870-2.
  - [119] X. Gong, Y. Gong, G. F. Wu, and H. Ke, "Bioinformatics analysis highlights CCNB1 as a potential prognostic biomarker and an anti-kidney renal papillary cell carcinoma drug target," *Medicine*, vol. 103, no. 12, p. E37609, Mar. 2024, doi: 10.1097/MD.00000000000037609.
  - [120] F. Li, Z. Wu, Z. Du, Q. Ke, Y. Fu, and J. Zhan, "Comprehensive molecular analyses and experimental validation of CDCAs with potential implications in kidney renal papillary cell carcinoma prognosis," *Heliyon*, vol. 10, no. 12, Jun. 2024, doi: 10.1016/J.HELİYON.2024.E33045.
  - [121] M. Migault, E. Donnou-Fournet, M. D. Galibert, and D. Gilot, "Definition and identification of small RNA sponges: Focus on miRNA sequestration," *Methods*, vol. 117, pp. 35–47, Mar. 2017, doi: 10.1016/J.YMETH.2016.11.012.
  - [122] A. Marranci *et al.*, "Systematic evaluation of the microRNAome through miR-CATCHv2.0 identifies positive and negative regulators of BRAF-X1 mRNA," *RNA Biol*, vol. 16, no. 7, p. 865, Jul. 2019, doi: 10.1080/15476286.2019.1600934.
  - [123] K. C. Miranda *et al.*, "A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes," *Cell*, vol. 126, no. 6, pp. 1203–1217, Sep. 2006, doi: 10.1016/J.CELL.2006.07.031.
  - [124] C. Niu *et al.*, "Potential prognosis and immunotherapy predictor TFAP2A in pan-cancer," *Aging (Albany NY)*, vol. 16, no. 2, p. 1021, 2024, doi: 10.18632/AGING.205225.
  - [125] D. Kołat, Ż. Kałuzińska, M. Orzechowska, A. K. Bednarek, and E. Płuciennik, "Functional genomics of AP-2α and AP-2γ in cancers: in silico study," *BMC Med Genomics*, vol. 13, no. 1, pp. 1–18, Dec. 2020, doi: 10.1186/S12920-020-00823-9/FIGURES/11.
  - [126] P. Qiu, "Embracing the dropouts in single-cell RNA-seq analysis," *Nature Communications* 2020 11:1, vol. 11, no. 1, pp. 1–9, Mar. 2020, doi: 10.1038/s41467-020-14976-9.
  - [127] M. Schiffer, "Going single but not solo with podocytes: potentials, limitations, and pitfalls of single-cell analysis," *Kidney Int*, vol. 92, no. 5, pp. 1038–1041, Nov. 2017, doi: 10.1016/j.kint.2017.05.033.
  - [128] Y. Kondo *et al.*, "Cryo-EM structure of a dimeric B-Raf:14-3-3 complex reveals asymmetry in the active sites of B-Raf kinases," *Science*, vol. 366, no. 6461, pp. 109–115, Oct. 2019, doi: 10.1126/SCIENCE.AAY0543.
  - [129] A. O. Stevens and Y. He, "Benchmarking the Accuracy of AlphaFold 2 in Loop Structure Prediction," *Biomolecules*, vol. 12, no. 7, Jul. 2022, doi: 10.3390/BIOM12070985.

- [130] C. Coppa, A. Bazzoli, M. Barkhordari, and A. Contini, "Accelerated Molecular Dynamics for Peptide Folding: Benchmarking Different Combinations of Force Fields and Explicit Solvent Models," *J Chem Inf Model*, vol. 63, no. 10, pp. 3030–3042, May 2023, doi: 10.1021/ACS.JCIM.3C00138/ASSET/IMAGES/LARGE/CI3C00138\_0019.JPEG.
- [131] Z. F. Brotzakis, S. Zhang, M. H. Murtada, and M. Vendruscolo, "AlphaFold-MetaInference: Prediction of Structural Ensembles of Disordered Proteins," *bioRxiv*, p. 2024.11.09.622758, Nov. 2024, doi: 10.1101/2024.11.09.622758.
- [132] Z. Karoulia, E. Gavathiotis, and P. I. Poulikakos, "New perspectives for targeting RAF kinase in human cancer," *Nat Rev Cancer*, vol. 17, no. 11, pp. 676–691, 2017, doi: 10.1038/NRC.2017.79.
- [133] E. E. Patton *et al.*, "Melanoma models for the next generation of therapies," *Cancer Cell*, vol. 39, no. 5, pp. 610–631, May 2021, doi: 10.1016/J.CCELL.2021.01.011.
- [134] A. Marranci *et al.*, "Identification of BRAF 3'UTR Isoforms in Melanoma," *J Invest Dermatol*, vol. 135, no. 6, pp. 1694–1697, Jan. 2015, doi: 10.1038/JID.2015.47.
- [135] A. Honkala, S. V. Malhotra, S. Kummar, and M. R. Junttila, "Harnessing the predictive power of preclinical models for oncology drug development," *Nat Rev Drug Discov*, vol. 21, no. 2, pp. 99–114, Feb. 2022, doi: 10.1038/S41573-021-00301-6.