Understanding the functional impact of alternative splicing in human cancer through Proteomics Informed by Transcriptomics (PIT)

Esteban Gea

School of Biological and Chemical Sciences
Queen Mary University of London

Thesis submitted in partial fulfilment
of the requirements of the Degree of

Doctor of Philosophy

2023

This page is intentionally left blank.

Statement of Originality

I, Esteban Gea, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Date: 12/02/2023
digitally signed (Esteban Gea)

Abstract

Recent advances in the fields of next generation sequencing and mass spectrometry have enabled additional depth in sequencing, allowing further developments for the fields of transcriptomics and proteomics. Additionally, software analyses for these data and computing power of machines have also improved over the past years. Yet, even though transcriptomics and proteomics are by essence deeply connected through the central dogma of biology, software pipelines capable of analysing both these data sources together to extract additional biological insights are limited in terms of analysis capabilities and scope of application.

This thesis describes the development of PITsuite, a software suite for integrated analysis and visualisation of transcriptomics and proteomics data. Unlike previous software built for integration of transcriptomics and proteomics data, PITsuite supports a wide range of experimental designs, including quantitative protocols and multi-sample experiments. PITsuite comprises two main components: a customisable analysis pipeline, and a graphical user interface for browsing and visualising results.

The efficacy of PITsuite was evaluated by application to several datasets from different studies. Notably, PITsuite was applied to transcriptomics and proteomics data from prostate cancer cells, as well as publicly available cancer data from TCGA (Tomczak et al., 2015), in order to understand the role of HNRNPA2B1 in prostate cancer, which findings are available in (Foster et al., 2022).

This project led to a new software suite (available at: https://github.com/bezzlab/pitsuite) which provides the first software pipeline to integrate quantitative transcriptomics, proteomics and public data from the raw files up to visualisation of the results through a bespoke graphical user interface.

Acknowledgments

I want to first express my sincere gratitude to Prof. Conrad Bessant, who has been supervising this project and has been a constant source of support and guidance throughout this PhD. Thanks to him, I have considerably expended my knowledge in several areas, such as bioinformatics or software development and I will always be grateful for that. Similarly, I also want to thank Dr. Prabhakar Rajan who has also been supervising this project and has been equally supportive and gave me a chance to experience working in wet lab and collaborating with talented biologists. One of them, Dr. John Foster, particularly deserves gratitude and praises for his passion, his patience and providing such an enriching collaboration.

I also want to thank all the wonderful people who were part of the lab during this project: Nazrath Nawaz, George Elder, Onur Ozcan, Nikhil Branson, Antara Labiba, Magdalena Huebner and Hajar Saihi. I will always remember the support, feedbacks, stimulating discussions and laughters we've had during these years together.

Last but not least, I want to thank my friends and family for their continuous support.

Contents

List of Figures

Chapter 1

Introduction to transcriptomics and proteomics

1.1 Transcriptomics

1.1.1 Introduction to transcriptomics

Transcriptomics is a set of experimental and computational techniques to process and extract information coming from RNA. According to the central dogma of biology, DNA is transcribed into RNA, and for protein-coding genes, is then translated into proteins. Working with RNA offers multiple advantages as it allows identifying which genes are expressed in a sample, which may vary according to the tissue type or other factors, but also to quantify this expression, allowing to, for example, measure the effect of a treatment on a set of genes. Working with RNA transcripts can also enable researchers to identify multiple isoforms of a gene and, similar to gene expression, study the effect of a treatment on alternative splicing. Furthermore, most transcripts do not code for proteins. Amongst the transcripts referenced in the version 38 for humans from the GENCODE project (1.1) (Frankish et al., 2019), protein coding transcripts represent only 37.1% of all transcripts. Other types of transcripts include long non coding RNA or small RNA. Interest for these transcripts has increased in recent years as research has unveiled some of the roles they can play, including in gene

expression regulation (Fang and Fullwood, 2016) (Statello et al., 2020) or mRNA degradation (O'Brien et al., 2018). Since these transcripts do not result in a protein product, transcriptomics remains the only solution to study them.

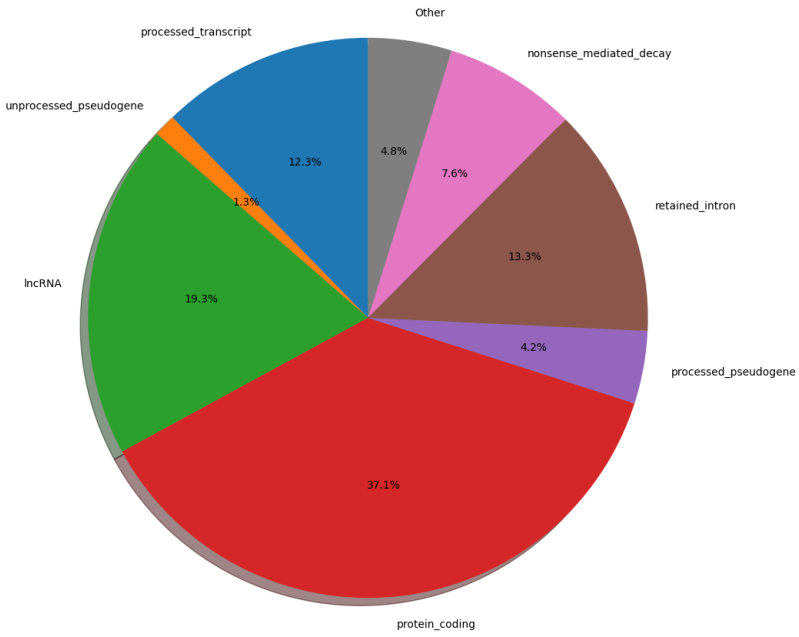

Figure 1.1: Pie chart of transcript types in the GENCODE human annotation version 38

1.1.2 Experimental approaches

1.1.3 Next Generation Sequencing

The most commonly used technique for transcriptome sequencing today is called Next Generation Sequencing (NGS). This technology enables a bigger sequencing depth than other technologies such as micro-arrays, allowing to reconstruct entire genomes or transcriptomes. Behind this term, many techniques and manufacturers exist. We will take the example of Illumina and sequencing by synthesis technology. The first step is sample preparation, where DNA or RNA is isolated. The sequences are then split into fragments and adapters are attached to the end of each fragment, forming a template. Each template is then anchored to a flow cell and for each fragment, the complementary sequence is transcribed using a polymerase, after which the template is removed. The adaptor on the sequence that has just been transcribed then binds to another adaptor on the flow cell, forming a bridge (1.2). A polymerase then binds to the sequence to transcribe the complement, resulting in another sequence, which will also create another bridge with another adaptor. This process called bridge amplification is repeated multiple times in order to end up with a high number of these fragments, in the order of several millions, depending on the sequencing depth chosen. Once the amplification is complete, a polymerase binds to each fragment to generate the complement, but this time adding fluorescent nucleotides. After the complement is made, a light source excites each nucleotide resulting in a fluorescent signal specific to each type of nucleotide, which makes it possible to determine the sequence, called read. Reads can either be single-ended or pair-ended. With single-ended reads, a defined number of nucleotides is read from the start of the fragment. With pair-ended reads, a first read is read from the start of the fragment with the length defined by the machine and a second read is read from the end of the fragment. Since we know the length (also called inner mate distance) between two reads, it is therefore easier to map pair-ended reads to the genome than single-ended reads. Indeed, assuming a read is 75 nucleotides long, it is likely that it might map to multiple locations in the genome, making it difficult to know from which

gene it comes from. On the other hand, if we have a read of 75 nucleotides, separated by a gap of 150 unknown nucleotides, followed by another 75 nucleotides long read, the number of locations on the genome this can map to is much lower. In particular, pair-ended is preferable for repetitive regions or alternative splicing analysis as pair-ended reads can span over a bigger distance, including sometimes multiple exons.

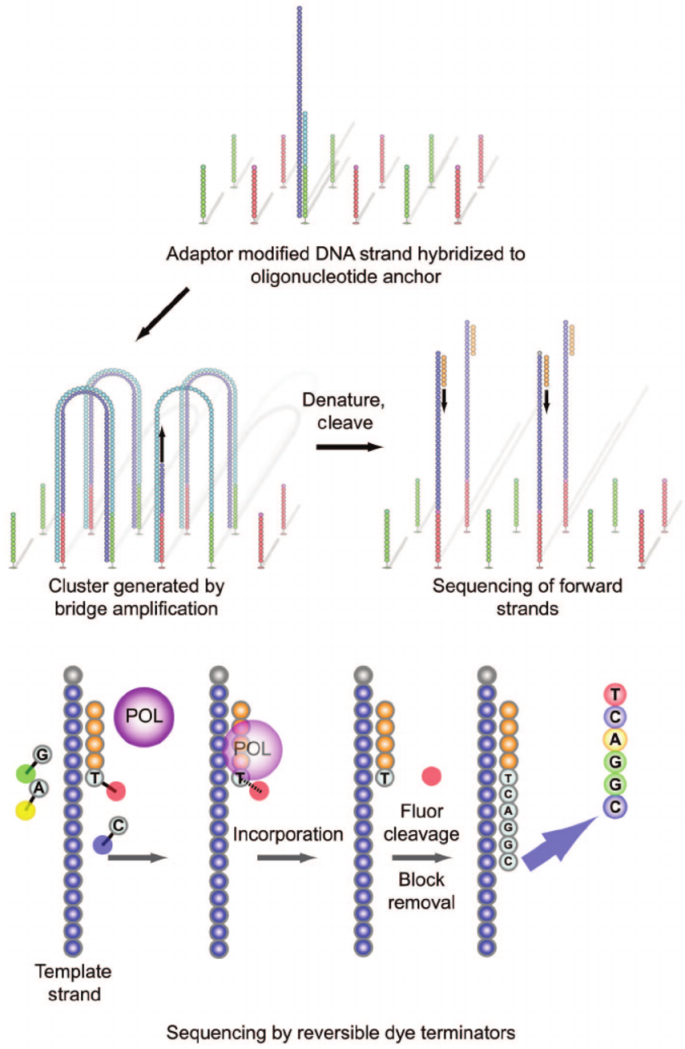

Figure 1.2: Illumina sequencing by synthesis process (Voelkerding et al., 2009).

However, this process is not error free and there are multiple potential sources of error leading to the wrong identification of nucleotides. These

include issues with sample storage or manipulation leading to DNA damage, copy errors by the polymerase, or errors by the software when interpreting the fluorescent signal. The error rate is estimated by Salk et al. to be between 0.1% and 1% and despite increase in coverage over the years thanks to technological improvements, the error rate has remained stable, although Salk et al. mention different practices to reduce it. (Salk et al., 2018). Another way to limit the impact of read errors is to increase coverage, as having a high sequencing coverage over a particular location in the genome will allow building a consensus and thus reduce the impact an error on a read can have during the assembly (1.3).

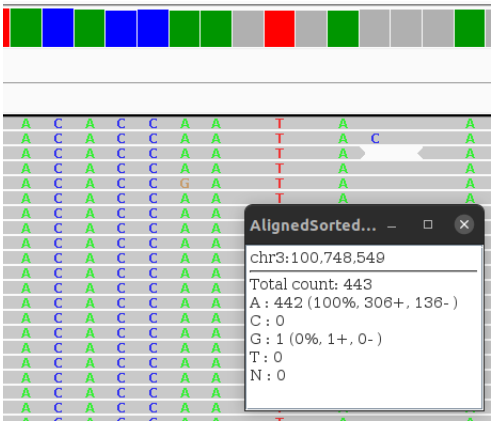

Figure 1.3: Read error visualised in IGV (Thorvaldsdóttir et al., 2013). After reads were mapped to the human genome, one read exhibits a G at position 100,748,549 on chromosome 3, however, 442 other reads mapped to the same location exhibits an A at this position, building a consensus on A.

The output of NGS is a fastq file that contains all the reads identified (1.4). Each read is represented as 4 lines in the files: the first one is the header containing the read ID, the second line the read sequence, the third a "+" and is optionally followed by the same sequence identifier and the fourth line indicates the PHRED quality score Q of each base encoded in ASCII, which indicates the probability P of the base being correct following the equation $Q = -log_{10}P$.

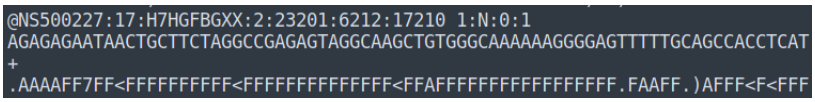

Figure 1.4: Information for a read in a fastq file.

1.1.4 Transcriptomics informatics

Once sequencing has been performed, multiple types of analyses can be performed on transcriptomics data, such as detecting mutations, differential gene expression or differential alternative splicing analysis. However, some of these analyses, such as alternative splicing, cannot be performed directly from the reads obtained from NGS. Instead, it is necessary to assemble the reads in order to identify and quantify which transcripts (expressed sequences) are present in the sample. There are two main ways to perform transcriptome assembly:

- Reference guided: can be used if a reference genome is available
- *De novo*: only uses the raw reads and doesn't require any reference genome or annotation

1.1.4.1 Without a reference genome: *de novo* assembly

De novo assembly is an assembly method that does not require a reference genome. This is particularly useful for orphan organisms which are not well studied and do not possess a reference genome, making a reference guided assembly impossible. Even when a reference genome is available, a *de novo* assembly may still be relevant, as all reference genomes contain parts where the sequence is unknown. It may also be helpful in situations where a genome contains a lot of variations compared to the reference genome, such as in cancer, or for sequencing wild type species, such as for plants. In this case, mapping reads to the reference may prove difficult. Many algorithms and tools exist for *de novo* transcriptome assembly, but we will here take the example of Trinity (Grabherr et al., 2011) as it is well established in the community, is still maintained, supports downstream analysis tools and performs well across multiple species Hölzer and Marz (2019).

1.1.4.1.1 De Bruijn graph and transcript assembly

Trinity is a software that includes several tools used within a workflow. It starts with Inchworm, which creates a dictionary of k-mers (by default k=25) from the reads. This dictionary is then used to create contigs, which are determined by concatenating the most abundant k-mers based on their k-1 overlaps. However, while this step can produce full transcripts, it can only produce one per full length locus, and therefore doesn't provide information about alternative splicing. Therefore, a second tool, Chrysalis is then used. It constructs clusters contigs from Inchworm based on exact k-1 overlaps and uses contigs within each cluster to construct a De Bruijn graph 1.5, that allows to represent which parts are common or different between the contigs of a cluster. Finally, Butterfly iterates through all De Bruijn graphs, and for each, identifies the different paths possible to go through the graph. From each path, Butterfly can extract a transcript, although it can also discard some paths, for example if their read coverage is low.

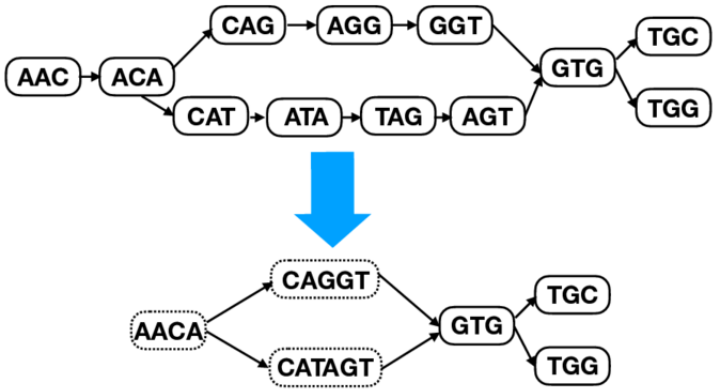

Figure 1.5: Generation of a De Bruijn graph from the k-mers obtained from RNA-Seq reads (Menegaux and Vert, 2020). Different paths can be the result of alternative splicing and will result in different transcripts.

1.1.4.1.2 Quantification

Once transcripts have been assembled, quantification can be performed using Salmon (Patro et al., 2017). Transcript abundance can be estimated by counting the number of reads and normalis-

ing by transcript length. To do so, Salmon aligns the reads to the transcripts assembled by Trinity. However, quantification can be made difficult by reads mapping to different transcripts or GC-content bias. Thus, Salmon uses a two steps quantification process where it first estimates the abundance of each transcript and then uses a statistical model to refine these estimates by taking into account all possible biases. For each transcript, Salmon then returns an abundance in TPM (transcripts per million). TPM is a quantification metric allowing to estimate the relative abundance of a transcript. To calculate it, we count the number of reads mapping to the transcript divided by the transcript length to get its normalised expression. Then the sum of the normalised expression for all transcripts is divided by one million to get a scaling factor. The normalised expression of each transcript is finally divided by this scaling factor to obtain the abundance in TPM. This metric can therefore be interpreted as the number of copies of a given transcript for each million RNA transcripts present in the sample.

1.1.4.2 With a reference genome: reference guided assembly

A reference genome is a file containing the DNA of a species. The most famous example is the human genome project, which in 2001 managed to sequence and assemble the human genome (Craig Venter et al., 2001). Since then, similar assemblies have been performed for many other species. This has traditionally been done using Whole Genome Sequencing (WGS), a type of Next Generation Sequencing that allows to sequence the entire genome. However, in recent years, 3rd generation sequencing technology has emerged, allowing to sequence longer reads compared to NGS, thus making genome assembly easier, especially in regions of the genome that were difficult to assemble with NGS, such as repetitive regions. Therefore, 3rd generation sequencing has now become the preferred method for genome assembly. The sequencing and assembly is usually done on a few individuals of the species from which a consensus sequence is built, which becomes the reference for the species. However, thanks to progress in NGS technology, leading to ever deeper coverage for a lower cost, as well as improvement in assembly

algorithms and computing power, reference genomes are regularly updated, to correct errors or fill regions where the sequence was previously unknown. ENSEMBL is a database built by the European Bioinformatics Institute (EBI) that hosts reference genomes for a number of vertebrates.

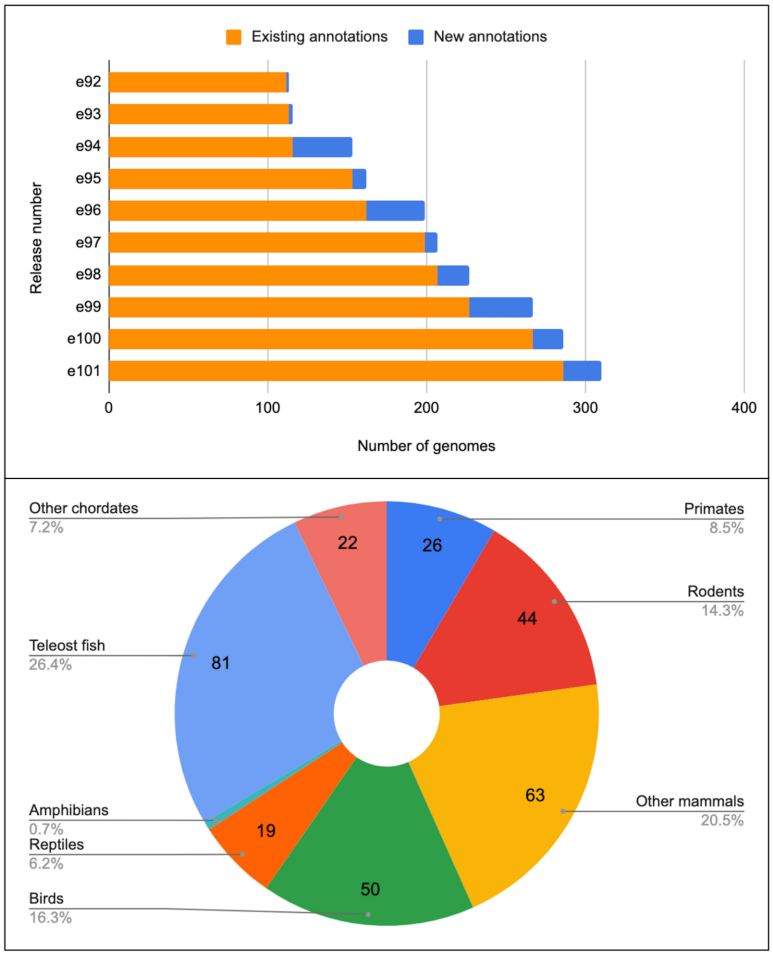

Figure 1.6: Number and types of genomes supported by ENSEMBL in 2020.(Howe et al., 2021)

For each species supported, ENSEMBL provides the reference DNA for each chromosome in fasta format, as well as annotations of the known genes, transcripts and coding sequences with their location on the genome.

1.1.4.2.1 Read Alignment In a reference guided assembly, the first step is to align the reads obtained from RNA-Seq to the reference genome in

order to know which gene and transcript produced it. Multiple alignment tools exist (Musich et al., 2021). Aligning reads to a reference genome is a complicated task for two main reasons. The first one is that the reads may not perfectly map to the region in the reference genome where it comes from. This can be due to nucleotides errors in the read sequence produced by sequencing, but this can also be caused by the fact that the sample used for RNA-Seq does not have exactly the same genome as the reference as mutations can arise, resulting in the substitution, insertion or deletion of nucleotides. The second challenge, which this time is specific to RNA sequencing and not DNA sequencing, is that a read may not align to a continuous sequence on the genome. Indeed, if a read is at the junction of two exons, the start of the read needs to map to the end of the first exon and the end of the read needs to map to the start of the following exon, which may be much further down the genome sequence. To address this issue, some alignment tools are splice aware and are able to split reads to map them to different locations of the genome. This is for example the case of STAR2 (Spliced Transcripts Alignment to a Reference) (Dobin et al., 2012). Read alignment tools output a SAM file (or the binary version BAM) which contains each read's header, sequence and quality, as well as where they align on the reference genome and the quality of the alignment.

1.1.4.2.2 Transcript assembly Once reads have been aligned to the reference genome, one option is to reconstruct the transcripts. Multiple assembly tools are available to perform this task (Hayer et al., 2015), which, unlike *De novo* assemblers, do not work directly on the raw reads but on the aligned reads (BAM files). A popular assembler is Stringtie2 (Kovaka et al., 2019). Stringtie2 can also take as input a GTF file (for example provided by Ensembl) to annotate transcripts. The GTF (Gene Transfer File) is a file format containing annotation of a genome.

A GTF file contains the following columns:

1. Chromosome name

2. Source of the annotation (database or project)

```
chr1    HAVANA  gene    11869   14409   .   +   .   gene_id
"ENSG00000223972.5"; gene_type "transcribed_unprocessed_pseudogene";
gene_name "DDX11L1"; level 2; hgnc_id "HGNC:37102"; havana_gene
"OTTHUMG00000000961.2";
chr1    HAVANA  transcript  11869   14409   .   +   .   gene_id
"ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type
"transcribed_unprocessed_pseudogene"; gene_name "DDX11L1";
transcript_type "processed_transcript"; transcript_name "DDX11L1-202";
level 2; transcript_support_level "1"; hgnc_id "HGNC:37102"; tag "basic";
havana_gene "OTTHUMG00000000961.2"; havana_transcript
"OTTHUMT00000362751.1";
chr1    HAVANA  exon    11869   12227   .   +   .   gene_id
"ENSG00000223972.5"; transcript_id "ENST00000456328.2"; gene_type
"transcribed_unprocessed_pseudogene"; gene_name "DDX11L1";
transcript_type "processed_transcript"; transcript_name "DDX11L1-202";
exon_number 1; exon_id "ENSE00002234944.1"; level 2;
transcript_support_level "1"; hgnc_id "HGNC:37102"; tag "basic";
havana_gene "OTTHUMG00000000961.2"; havana_transcript
"OTTHUMT00000362751.1";
```

Figure 1.7: Part of the GTF file corresponding to version 38 for humans by the GENCODE project (Frankish et al., 2019)

3. Feature - the type of entity, which can be a gene, a transcript, a coding sequence, an exon or other types
4. Start position on the chromosome
5. End position on the chromosome
6. Score, which is an optional number
7. Strand, + or -
8. Frame, can be equal or '0', '1', '0', or '.' to leave empty. It represents the position of the first base in the codon
9. Notes - semicolon-separated list of tag-value pairs giving additional information about each entity. The tags are not pre-defined and are left for the user to specify. In the case of GTF files provided by ENSEMBL, tags contain for example the ENSEMBL ID of a transcript or gene as well as the gene name and the type of transcript (protein coding, lncRNA, ...).

Providing a GTF file to the assembler guides it to give a preference to transcripts that are already referenced in the GTF. In addition, StringTie2

outputs a GFF file and transcripts identified by Stringtie2 that match one of the transcript in the GTF file will contain the ID of this transcript in their notes, allowing users to look for additional information in the ENSEMBL database or using R or Python packages. However, Stringtie2 is also able to identify novel transcripts that are not referenced in the GTF file. These transcripts are given a unique ID starting with "STRG" followed by a number (or "MSTRG" when merging different assemblies).

1.1.4.2.3 Quantification Transcript quantification after reference guided assembly is similar to quantification after *de novo* assembly, except that this time the reads are already aligned and do not need to be aligned to the transcripts. After identification, Stringtie2 also quantifies each transcript in TPM and writes it as a tag/value pair in the last column of the outputted GFF file.

1.1.4.3 Limitations and recent advances

1.1.4.3.1 Limitations While second generation sequencing has been a breakthrough in the field of sequencing, enabling the sequencing of entire genomes or the comparison of RNA profiles between different samples, some bottlenecks remain. These issues include the sequencing price, the error rate of read sequencing, GC content bias, the lack of coverage, especially in the extremities of chromosomes, or sample preparation. However, what might be considered by some to be the main bottleneck is the read length. With second generation sequencing, reads typically have a length between 50 and 200 nucleotides, depending on the experiment and the technology used. Having short reads is an issue as the probability they may map to different locations in the genome remains high. This is particularly true for repetitive regions where it is impossible to map the read exactly if the repetitive motif is longer than the read length, meaning entire genome regions cannot be sequenced with second generation sequencing. In addition, identification and quantification of alternative splicing events requires having reads at the junction between two exons. With short reads, only a small proportion will overlap with an exon junction, making it difficult for assemblers to accurately

quantify each isoform involved in the splicing event. While it is true that using pair-ended reads can solve this issue in some cases by reducing the number of possible mapping locations, it still remains an unsolved issue for second generation sequencing.

1.1.4.3.2 Long read sequencing Third generation sequencing is the latest generation of DNA (and RNA by reverse transcription) sequencing technology. The main advantage compared to the previous generation is that it can produce much longer reads, typically between 10 000 to 30 000 nucleotides long with the latest technologies. (Amarasinghe et al., 2020). This makes third generation sequencing particularly suitable for assembling regions of the genome that cannot be sequenced through second generation sequencing, such as repetitive regions. It is also used for identification of genetic diseases, revealing previous inaccessible mutations in patients (Xiao and Zhou, 2020).

However, and despite considerable progress in recent years, both at a hardware and software level, with the existence of multiple correction algorithms, the error rate remains higher than for second generation sequencing. It is claimed that the error rate for Single Molecule Real time Technology (SMRT) is less than 1% whereas it is less than 5% for Nanopore sequences (Amarasinghe et al., 2020). An additional issue of third generation sequencing is the lower coverage, meaning fewer reads are produced compared to second generation sequencing, making it more difficult to use for differential gene expression or alternative splicing.

A possible solution is to use a hybrid approach, combining both second and third generation sequencing. This approach allows using long reads to assemble genome sequence, including the regions which cannot be assembled from short reads, resulting in a better assembly with longer contigs. Additionally, short reads can also be used in complement of long reads to correct ambiguities resulting from the high error rate of long reads and, at the same time, benefit from the higher coverage to perform quantitative downstream analysis such as differential gene expression or alternative splicing.

Yet, although RNA-Seq can identify and quantify transcripts at the RNA

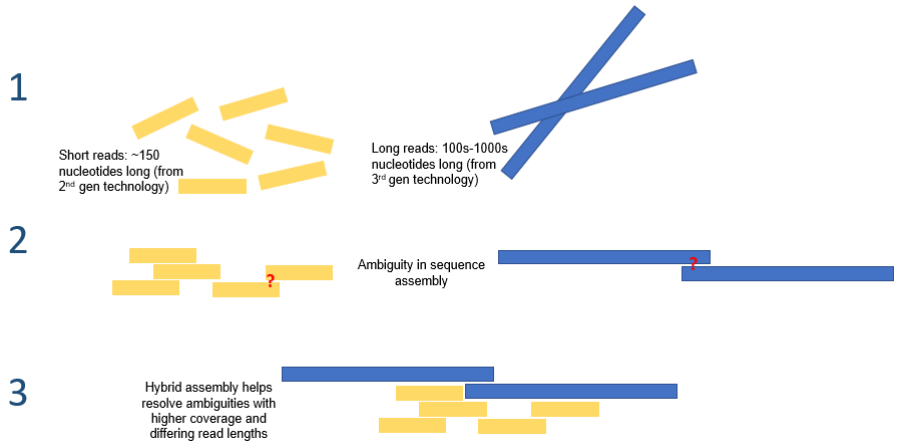

Figure 1.8: Using both second read and third generation sequencing to perform assembly (Thatra)

level, this does not necessarily reflect what will be found at the protein level. This implies the use of other technologies to study the proteome.

1.2 Shotgun proteomics: Mass spectrometry

Proteomics is a field of biology focusing on the study of proteins. Proteins are the molecules responsible for most of the work in a cell and are involved in almost all pathways. Proteins are translated from RNA transcripts by the ribosome, a protein complex (of 80 proteins for humans (Khatter et al., 2015)). Proteins are a string of amino acids that can be deduced from the RNA transcript sequence. Indeed, an mRNA transcript can contain one or several open reading frames (ORF) which are the part of the transcript that will code for a protein. From the start of the sequence, each fragment of three nucleotides will code for one amino acid (1.9). Ribosomes read each codon of the ORF and attach the corresponding amino acid to the end of the backbone of the protein, forming a string of amino acids. Once assembled, the protein folds into its 3D structure (1.10), determined by the physical and chemical properties of its amino acids. It is the 3D structure, as well as the composition of the protein, that determine its function. For example, the protein may exhibit binding sites that will allow it to bind to other molecules, including other proteins.

First base position	Second base position: U				Second base position: C				Second base position: A				Second base position: G				Third base position
U	UUU	[1]P			UCU	S			UAU	Y			UGU	C			U
	UUC				UCC				UAC				UGC				C
	UUA	L			UCA				UAA	Stop			UGA	Stop			A
	UUG				UCG				UAG				UGG	W			G
C	CUU	L			CCU	P			CAU	H			CGU	R			U
	CUC				CCC				CAC				CGC				C
	CUA				CCA				CAA	Q			CGA				A
	CUG				CCG				CAG				CGG				G
A	AUU	I			ACU	T			AAU	N			AGU	S			U
	AUC				ACC				AAC				AGC				C
	AUA				ACA				AAA	K			AGA	R			A
	AUG	M			ACG				AAG				AGG				G
G	GUU	V			GCU	A			GAU	D			GGU	G			U
	GUC				GCC				GAC				GGC				C
	GUA				GCA				GAA	E			GGA				A
	GUG				GCG				GAG				GGG				G

[1]The one letter symbol of amino acids.

Figure 1.9: Codon table, representing which amino acid is coded by each group of 3 nucleotides (Sánchez et al., 2006)

While the estimated number of protein coding genes in humans is around

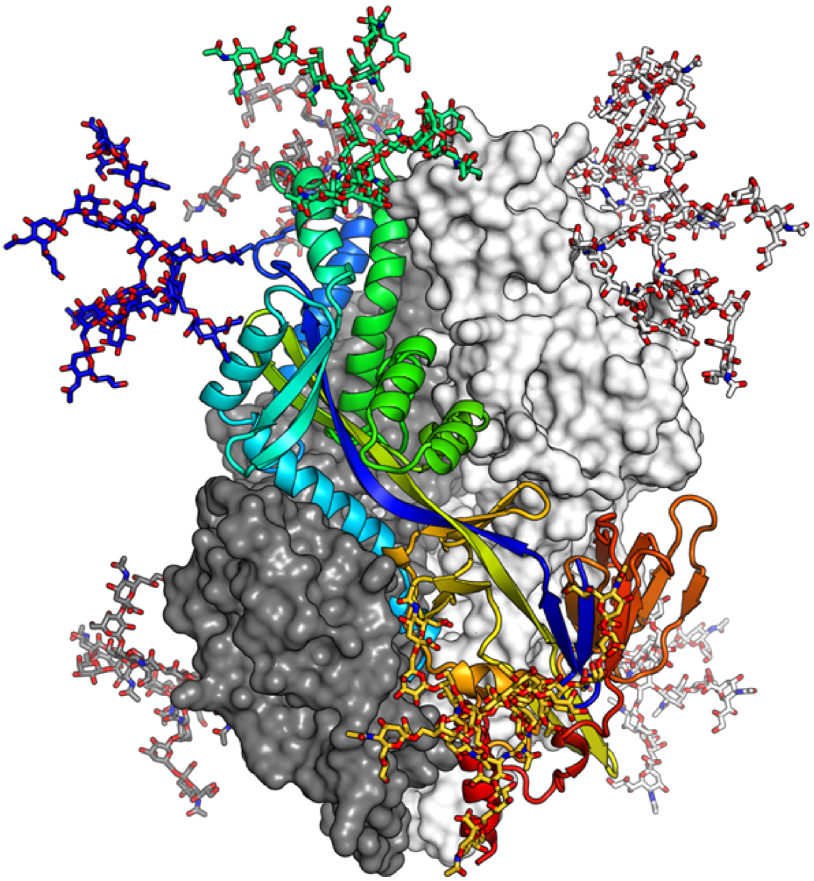

Figure 1.10: 3D structure of the pre-fusion hMPV F trimer protein with different modes of representation (Battles et al., 2017)

20,000 (Piovesan et al., 2019), the actual number of distinct proteins is much higher, as from a single protein coding gene, multiple proteins can be derived. Several phenomena can explain this diversity. The first one is mutations, which can impact the DNA sequence and by extension the RNA sequence and the amino acid sequence. While these mutations may not have an impact on the protein sequence, or at least may not have a harmful impact on the phenotype, others can be the cause of severe genetic diseases. One example is the case of nonsense mutations (Mort et al., 2008). Nonsense mutations are a type of mutations that will affect the RNA transcript, replacing

a codon coding for an amino acid by a stop codon. This early stop codon would result in the ribosome stopping the translation of the protein before its end, resulting in a truncated protein which may not fulfil its function. Many genetic diseases are the result of nonsense mutations, such as cystic fibrosis or haemophilia, as well as certain types of cancer (Benhabiles et al., 2017). Another source of protein diversity is alternative splicing. Indeed, alternative splicing can affect which exons of a gene will be part of a transcript, thus affect the ORF and the amino acid sequence. Finally, post translation modifications can also result in different proteins. While the amino acid sequence of the protein remains the same, the addition or deletion of certain molecules from the protein can affect its function. This is for example the case of phosphorylation, where a kinase attaches a phosphoryl group to a protein, transforming it into its activated state.

Proteomics encompasses various aspects such as the study of protein abundance, protein 3D structures, post translation modifications and others. In this section, we will talk specifically about shotgun proteomics through the use of mass spectrometry, an experimental and computational method allowing the detection and quantification of proteins in a sample.

Mass spectrometry has become the standard technique for identifying and quantifying proteins in a sample. While the first mass spectrometer dates back to 1910, the technology has been constantly evolving over the years (McLafferty, 2011) and today the most commonly used method is liquid chromatography-mass spectrometry (LC-MS/MS). Unlike other types of experiments such as Western Blotting which can only identify and quantify one protein at a time, mass spectrometry is able to identify thousands of proteins in a single run.

An LC-MS/MS experiment includes several stages, from sample preparation, to running the mass spectrometer and ending with the computational analysis of the results (1.11).

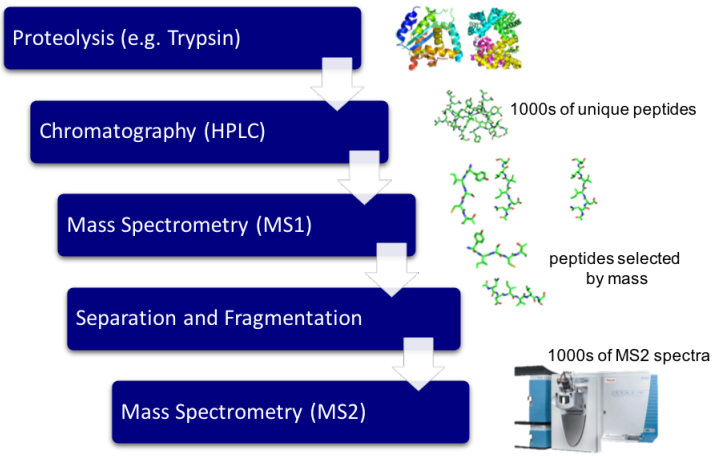

Figure 1.11: Workflow of the different steps of an LC-MS/MS experiment (Bessant, 2017)

1.2.1 Sample preparation

Since proteins tend to be large molecules, often with a mass of tens of thousands of daltons and as such, they cannot be directly detected by a mass spectrometer which usually cannot detect molecules over 4,000 Daltons (Parker et al., 2010). It is therefore necessary to split proteins into smaller fragments called peptides. Enzymes such as Lys-C or trypsin can perform this task. Trypsin is the most commonly used cleavage method as it goes along the amino acid sequence of a protein and cuts after each arganine or lysine unless it is followed by a proline. Considering the distribution of amino acids along a protein sequence, using trypsin allows most peptides to be within a length range compatible with what a mass spectrometer can detect. It is however important to note that depending on the situation, it may be relevant to use another enzyme, for example if we are interested in a particular protein or group or protein and trypsin doesn't allow to generation of peptides of good length for this protein. Sometimes, it is not necessary to use digestion at all. This can for example be the case in mass spectrometry based HLA peptidomics (Gfeller and Bassani-Sternberg, 2018), where mass spectrometry can be used to detect HLA peptides on the surface of the cell. In this case, these peptides are usually already of the right length to be used

in mass spectrometry. On top of this protocol, additional steps can be taken according to the requirement of the study. Enrichment of specific proteins can be performed in order to increase the number of peptides that can be detected for these proteins, thus offering a better coverage of these proteins and a better resolution for quantification. It is also common to enrich samples in order to detect certain post translation modifications. This is for example the case for phosphoproteomics. Phosphoproteins tend to be less abundant than non phosphorylated proteins, yet mass spectrometry identifies more easily peptides that are more abundant, making low abundance peptides more difficult to identify, making it necessary to enrich these peptides to increase their relative abundance (Beltran and Cutillas, 2012).

Finally, labels can be used for relative quantification. Labels allow combining multiple samples in a single mass spectrometry run, by attaching to each peptide a label indicating from which sample the peptide comes from. Multiple labels exist, such as SILAC (Ong et al., 2002), TMT (Thompson et al., 2003) or iTRAQ (Luo and Zhao, 2012). They differ in the molecules used as labels, in the number of labels available (and therefore the number of samples that can be put together in a single run). These labels come as kits that have to be bought separately, and each has a specific protocol to attach the labels to the peptides.

1.2.2 Experimental protocols

Once the sample is prepared, peptides must be separated in order to be able to analyse them all. Multiple techniques allow doing so, but the one commonly used is liquid chromatography. The sample, in liquid form, also called mobile phase in this context, is put through a column and peptides travel through it. Inside this column, absorbent materials, called solid phase, interact with the peptides, causing them to go at different speeds depending on the chemical and physical properties of the peptide, such as hydrophobicity. These properties are determined by the amino acid sequence of the peptide, and post translation modifications if some are present. Therefore, each peptide will stay in the column for a given time, called retention time,

before coming out, allowing peptides to be separated at the end of the column. Peptide are then grouped into batches of peptides with close retention time, and each of these batches is sent separately to the mass spectrometer. Separating peptides into batches thus allows to have multiple spectra that are not too overcrowded due to too many peptides being processed at the same time.

Next, peptides are ionised in order to be able to fly through the mass spectrometer. This process is done through protonation, usually giving them a charge of 1+, 2+ or 3+.

Different types of mass spectrometer exist. In Time Of Flight (ToF) mass spectrometers (Boesl, 2017), a plate charged negatively attracts the positively charged peptides, called ions, giving them kinetic energy. Once the ions have passed the plate, they stop accelerating and move through a vacuum chamber at a speed depending on their weight, with less heavy ions moving faster than heavier ions. At the end of the chamber, ions hit the detector. The detector gives electrons to the ions, according to its charge. This transfer of electrons from the detector to the ion creates a current, allowing to detect the ion. Since we know how long the ion took to travel and the relationship between travel time and ion mass, it is possible to determine the mass of the ion.

Another type of mass spectrometer is the Orbitrap technology. This time, ions enter a vacuum chamber where a coaxial inner spindle-like electrode is present. Its inside is charge negatively, therefore ions are attracted to it. However, the velocity at which ions enter the chamber balance this attraction, putting the ions in orbit around the electrode, like a satellite would orbit around earth. In addition, ions also move back and forth along the electrode axis. It is this movement which is measured, as its frequency is dependent on the ion mass. Using a Fourier Transform, it is possible to obtain the spectrum (1.12).

The name LC-MS/MS comes from the fact that mass spectrometry is run twice. The first time, the whole peptide is sent through the mass spectrometer and its mass to charge ratio (m/z) determined. This step is called MS1. Since multiple peptides are processed at the same time, the result for each batch is a spectrum in which each peak correspond to an ion. These ions

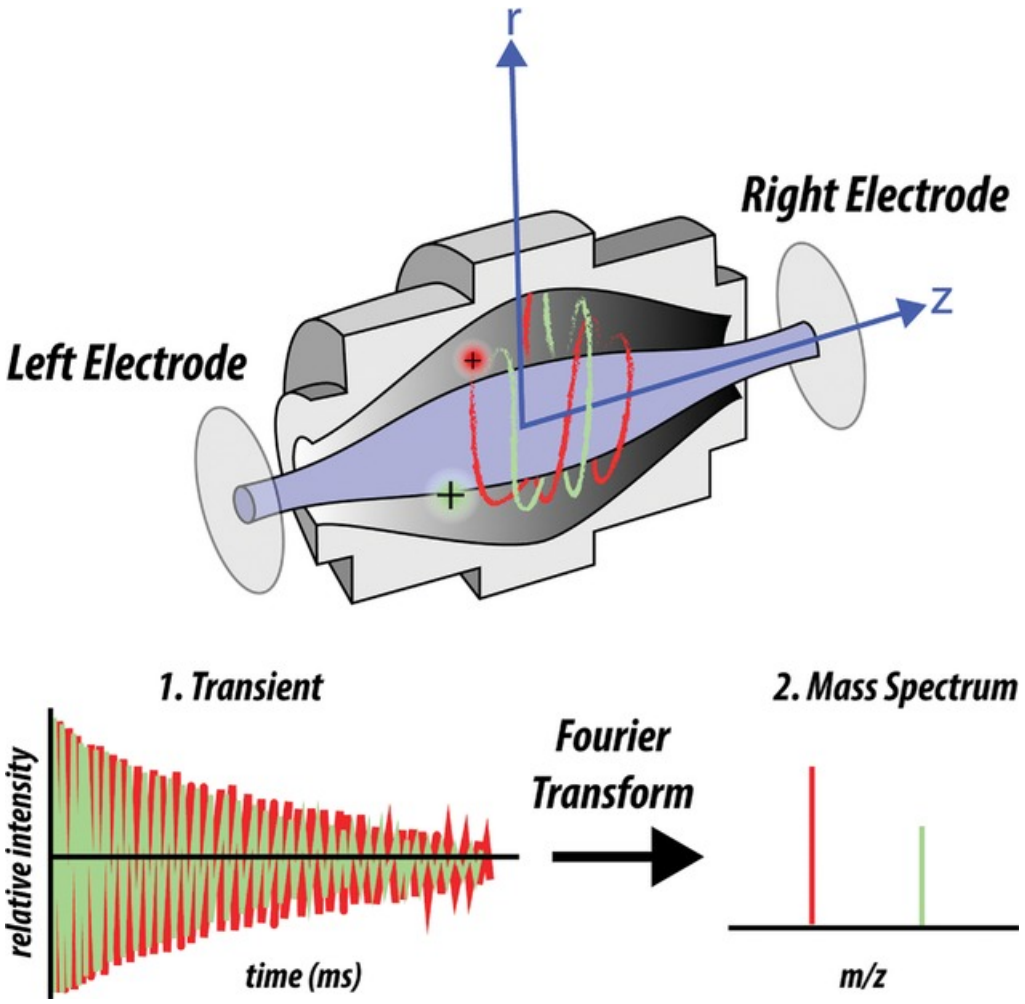

Figure 1.12: The Orbitrap FT mass analyser (Savaryn et al., 2016). As ions move along the electrode axis, the signal detected as they get closer to the left or right electrode gets more intense. The frequency of this signal is used in combination with a Fourier transform to determine the masses of the ions present.

will ideally be peptides, but they can also be contaminants that add noise to the spectrum, hence the importance of careful sample preparation and manipulation. On the spectrum, the x-axis represents the mass to charge ratio whereas the y-axis represents the intensity: the more abundant an ion is, the higher its intensity. For each spectrum, the most abundant ions (corresponding to the most intense peaks) are selected individually for another

run of mass spectrometry, called MS2. This second step is necessary because even though we now know the mass of a peptide and therefore can deduce its amino acid composition by finding which combination of amino acid masses equals the peptide mass, we still cannot determine the peptide sequence. This is true because even though we may know which amino acids are present, we still don't know in which order they are arranged. In MS2, each peptide is broken down into fragments using methods such as collision-induced dissociation. In collision-induced dissociation, ions are accelerated by an electrical field to increase their kinetic energy. They then collide with other molecules present, and the kinetic energy accumulated by the ions causes them to break bonds between atoms of the backbone of the ion. Most of the bonds broken are between the carbon and nitrogen atoms of the ion backbone (1.13) resulting in fragment ions. Such fragments are called b and y ions. Although less frequent (unless programmed on the machine), other bonds can also be broken, for example between two carbon atoms of the backbone, generating other fragment ions (Mitchell Wells and McLuckey, 2005).

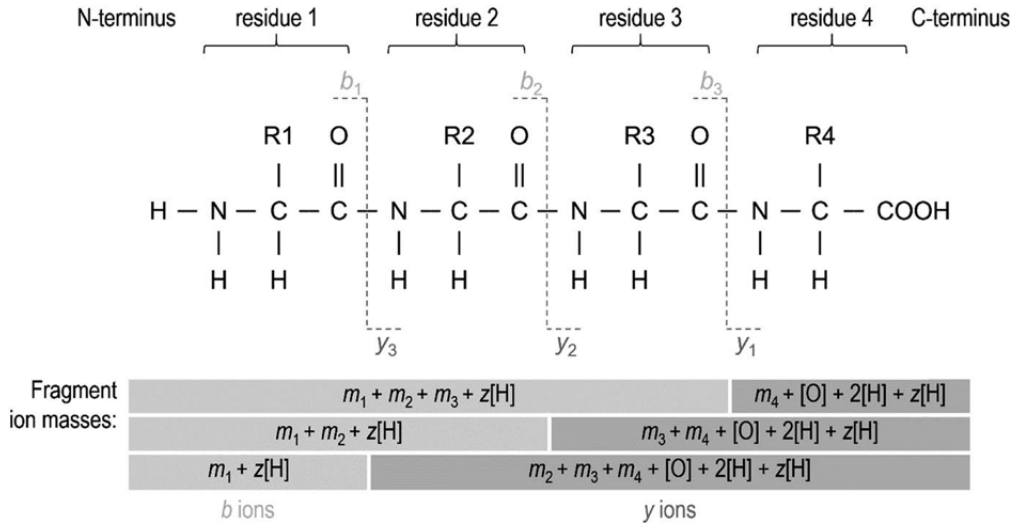

Figure 1.13: Example of fragmentation of a peptide (Bessant, 2017). One fragmentation may happen between the first carbon and nitrogen atoms of the backbone, resulting in two ions: b1 for the left part that contains one amino acid and y3 for the right part that contains three amino acids. Another fragmentation can also happen on the second carbon and nitrogen atoms, resulting in b2 and y2 ions.

After fragmentation, all the fragment ions coming from a peptide are sent together through the mass spectrometer in the same way as previously described. However, this time the spectrum does not display a peak for each peptide ion, but rather a peak for each fragment ion. It is this spectrum which will be used to determine the peptide sequence.

1.2.3 Proteomics informatics

Because the output of the mass spectrometer is a spectrum, it does not give us any direct information about protein identification or quantification, nor post translation modifications (PTMs). To obtain this information, computational methods must be applied. These methods are part of a set of computational techniques called proteome informatics. These computational methods include peptide identification, protein grouping, peptide and protein quantification, and PTM identification. In the section, we will present some of these methods and introduce tools that allow to perform them.

1.2.3.1 Peptide identification

A MS2 spectrum contains information on the mass to charge ratio and intensity of fragment ions coming from a peptide. As previously explained, fragment ions of a same type (b ions, y ions or other types) differ on their amino acid composition. For example, a b1 ion will only contain the first amino acid of a peptide, whereas a b2 ion will contain the first two amino acids of a peptide. A y1 ion will contain the last amino acid of a peptide, a y2 ion, the last two amino acids. Different methods can be used to take advantage of this information to reconstruct the peptide sequence.

1.2.3.1.1 *De novo* peptide sequencing *De novo* peptide sequencing is a computational method allowing to obtain a peptide sequence using only its MS2 spectrum (Dančík et al., 1999). Since the difference of mass of fragment ions is equal to the sum of the masses of the amino acids that differ between these fragments, looking at the difference in mass between fragment ions (how far peaks are on the m/z axis) can tell us which amino acids are

present in the peptide and in which order. *De novo* algorithms are usually applied on the b and y ions, as these are the most abundant ones. These algorithms start looking at the most intense peaks and look at the difference of mass to charge between two peaks and determine whether the difference of mass is equal to the mass of one or several amino acid. It is possible to do this since the mass of each amino acid is known accurately and a mass spectrometer can detect an ion mass with high accuracy (1.14).

Amino acid	code		Avg. mass/ Elenental composition	Monoisotopic mass increment/u
Alanine	A	Ala	C_3H_5NO	71.03711378804
Cysteine	C	Cys	C_3H_5NOS	103.00918447804
Aspartic acid	D	Asp	$C_4H_5NO_3$	115.02694303224
Glutamic acid	E	Glu	$C_5H_7NO_3$	129.04259309652
Phenylalanine	F	Phe	C_9H_9NO	147.0684139166
Glycine	G	Gly	C_2H_3NO	57.02146372376
Histidine	H	His	$C_6H_7N_3O$	137.0589118628
Isoleucine	I	Ile	$C_6H_{11}NO$	113.08406398088
Lysine	K	Lys	$C_6H_{12}N_2O$	128.09496301826
Leucine	L	Leu	$C_6H_{11}NO$	113.08406398088
Methionine	M	Met	C_5H_9NOS	131.0404846066
Asparagine	N	Asn	$C_4H_6N_2O_2$	114.04292744752
Proline	P	Pro	C_5H_7NO	97.05276385232
Glutamine	Q	Gln	$C_5H_8N_2O_2$	128.0585775118
Arginine	R	Arg	$C_6H_{12}N_4O$	156.10111102874
Serine	S	Ser	$C_3H_5NO_2$	87.03202841014
Threonine	T	Thr	$C_4H_7NO_2$	101.04767847442
Valine	V	Val	C_5H_9NO	99.0684139166
Tryptophan	W	Trp	$C_{11}H_{10}N_2O$	186.07931295398
Tyrosine	Y	Tyr	$C_9H_9NO_2$	163.0633285387
p-Serine	pS	pSer	$C_3H_6NO_5P$	166.99835882058
p-Threonine	pT	pThr	$C_4H_8NO_5P$	181.01400888486
p-Tyrosine	pY	pTyr	$C_9H_{10}NO_5P$	243.02965894914

Figure 1.14: Name, composition and mass of each amino acid (Spengler and Hester, 2008).

If the mass shift between two peaks corresponds to the mass of an amino acid, this amino acid is added to the sequence and the algorithm repeats the process to the other peaks. In the case of b ions, the sequence is thus built from start to end, whereas it is built from end to start for y ions. It is important to note that sometimes, identification of some fragments may be missing. For example in 1.15, the y12+ ion was not detected, therefore, the leucine in the peptide sequence cannot be inferred from the y+ fragment ions. However, fragmentation also created the b2 fragment ion which was detected in the spectrum, allowing to identify the leucine. Therefore, b and y ions are

used in complement for identifications, sometimes also with different charges. The other peaks in the spectrum can be the result of other types of fragment ions that were created from bounds being broken in other parts of the peptide backbone during fragmentation, or can be caused by noise in the signal or contaminants.

ELILGSETPSSPR

#1	b^+	b^{2+}	Seq.	y^+	y^{2+}	#2
1	130.04988	65.52858	E			13
2	243.13395	122.07061	L	1256.68455	628.84591	12
3	356.21802	178.61265	I	1143.60048	572.30388	11
4	469.30209	235.15468	L	1030.51641	515.76184	10
5	526.32356	263.66542	G	917.43234	459.21981	9
6	613.35559	307.18143	S	860.41087	430.70907	8
7	742.39819	371.70273	E	773.37884	387.19306	7
8	843.44587	422.22657	T	644.33624	322.67176	6
9	940.49864	470.75296	P	543.28856	272.14792	5
10	1027.53067	514.26897	S	446.23579	223.62153	4
11	1114.56270	557.78499	S	359.20376	180.10552	3
12	1211.61547	606.31137	P	272.17173	136.58950	2
13			R	175.11896	88.06312	1

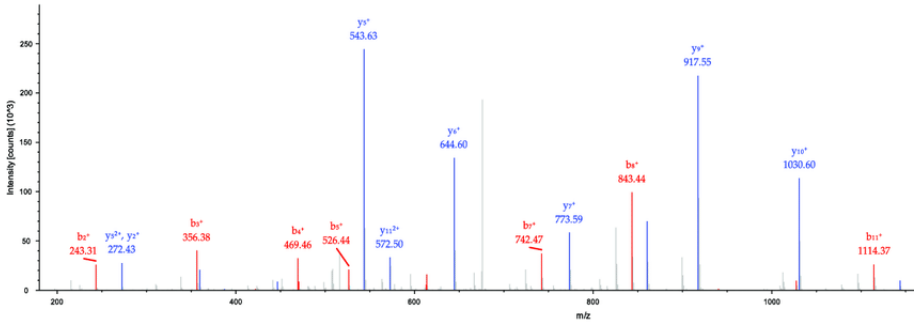

Figure 1.15: An example of *de novo* peptide sequencing (Zhang et al., 2013). Peaks identified as b ions are represented in red, and peaks identified as y ions are represented in blue.

The advantage of *de novo* peptide sequencing is that it does not have any other requirement than the MS2 spectrum and is not dependant of any reference data such as the species proteome, which makes this technique useful in the identification of novel peptides, for example peptides resulting from mutations in the protein sequence. However, it requires a clean spectrum as having a signal-to-noise ratio too high means that many peaks that are the results of noise will be considered as potential ion fragment coming from the peptide and the more of these peaks are present, the higher the risk that

the gap between two peaks comes close to the mass of an amino acid just by chance, which could then lead to the wrong peptide sequence being identified. This is why this technique is commonly not used in shotgun proteomics as it is not the most accurate, although it is still used in certain cases where there are no alternatives, for example for identifying novel peptides. In addition, it is worth noting that in recent years, machine learning models, such as the one developed by Tran et al. combining both a convolution neural network and a LSTM network (Tran et al., 2017), have been developed and claim to have higher accuracy than traditional algorithms.

1.2.3.1.2 Peptide-spectrum matching Peptide-spectrum matching is nowadays the most commonly used technique for peptide identification. Here, a database in fasta format of proteins that are expected to be found is necessary. Such databases can for example be downloaded from Uniprot (Bateman et al., 2017). As of January 2022, Uniprot contains a reference proteome for 20,382 species. The sequences for these proteomes are either determined from proteomics or transcriptomics experiments, or can be derived by homology from other species. A proteome typically contains one sequence per protein, this sequence being built from a consensus coming from the sequences identified in several individuals, and these sequences are regularly updated by Uniprot. For some species, such as model organisms, that are particularly well studied, the proteome can also multiple different sequences for a single gene, corresponding to different protein isoforms.

The database is then digested computationally according to the same rules that is applied for the enzyme used in the mass spectrometry experiment. For example, if trypsin was used in the experiment, an algorithm will take all the proteins present in the database and generate a set of peptides but cutting the protein sequences after a arginine or a lysine, unless followed by a proline. Since we know the mass of each amino acid, it is possible to generate for each peptide a theoretical spectrum showing where the peaks are expected to be on the m/z axis if this peptide was present in the sample.

Once the theoretical spectra have been generated, the experimental spectra are matched to the theoretical ones. In some peptide identification tools,

a spectrum is only compared to a theoretical spectrum if the mass of the precursor ion (the peak corresponding to the peptide in the MS1 spectrum) comes close to the mass of the theoretical peptide. Since no experiment is perfectly accurate and mass spectrometry is no exception, the mass measured for a peptide is not perfectly accurate. It is therefore necessary to allow a tolerance for the matching, between the mass we observe and the mass we expect. Some tools therefore allow the user to set a precursor mass tolerance (PMT), for the matches between the mass of the precursor ion and the mass of the theoretical spectrum. If this tolerance is too low, a spectrum may not match the theoretical spectrum it corresponds to, because the mass error during the measurement is higher than what is allowed, therefore this spectrum would either have no identification or it may be matched to another theoretical spectrum by chance, resulting in a false identification. On the other hand, having a tolerance too high increases the number of theoretical spectral candidates the tools will try to match the observed spectrum to. As this number increases, so does the probability of the observed spectrum matching a wrong theoretical spectrum by chance. In practice, the tolerance is usually set according to the setting of the mass spectrometry experiment, as the error is dependent on the type of mass spectrometer used and other settings.

Once a subset of theoretical spectra have been selected from an experimental spectrum, the search tool tries to determine which one is the best match. Here, another tolerance threshold, the fragment mass tolerance (FMT), can usually be set, depending on the search tool. The fragment mass tolerance indicates how much difference is allowed between the mass of an observed fragment and the mass of the corresponding fragment on the theoretical spectrum. For each theoretical spectrum, the search tool calculates a score indicating how good the match is. The method used to calculate this score depend on the search tool, but all share the same principle that the more peaks can be matched, the higher the score will be. However, each MS2 spectrum contains noise, therefore if all peaks in the experimental spectrum are considered, a high number of false matches will happen just by chance. Therefore, most tools filter peaks to only consider the N most intense peaks,

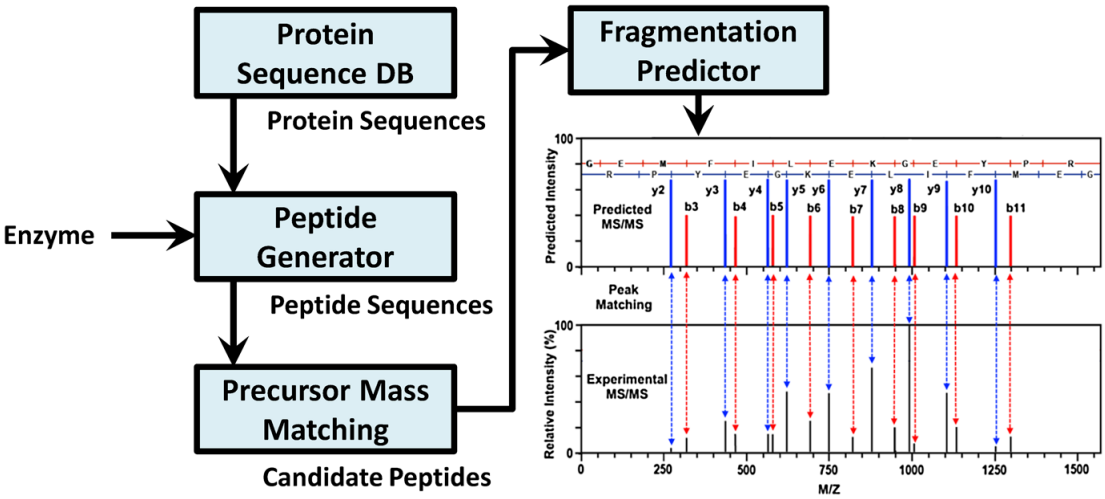

Figure 1.16: The peptide spectrum matching approach tries to match peaks from a theoretical spectrum (blue) to the peaks from the observed experimental spectrum (red) (Bessant, 2017).

where N can be set by the user. Some tools also take the intensity into account in the score: if a peak on the theoretical spectrum matches a peak of high intensity on the experimental spectrum, the increase in score will be higher than if it matches a low intensity peak.

Another important point to consider are post translational modifications (PTM). Indeed, PTMs can add or remove molecules from a protein, therefore modifying its mass. This means a PTM that adds a molecule to a protein would result in the fragment ions containing this PTM to have a higher mass and therefore have their peak shifted to the right on the m/z axis. Search tools distinguish between 2 types of PTM: fixed and variable. A fixed modification is a PTM that is considered to be always present. This is for example the case for carbamidomethylation of cysteine. This PTM is a result of the reduction and alkylation process (Suttapitugsakul et al., 2017). It adds a mass of 57.02 Da to every cysteine residues, therefore all fragment ions will have their peak shifted by 57.02 Da for every cysteine they contain. This means that while generating the theoretical spectra, the search engine needs to take this into consideration and shift the peaks in the theoretical spectra accordingly, otherwise they would not match the observed spectra.

The other type of PTM is variable modification. It differs from fixed modification in the sense that these modifications may or may not be there. This implies that for every residue that may contain a variable modification, it is necessary to generate a theoretical spectrum containing the modification, and therefore add the corresponding mass shift in the peaks, but also one spectrum without the modification in order to match peptides in the sample that do not contain it. A common example is phosphorylation which adds a phosphoryl group of 79.99 Da on a serine, threonine or tyrosine. On a protein, some of these residues can receive a phosphoryl group from a kinase and are called phosphosites. Yet, not all serine, threonine and tyrosine will necessarily receive a phospho group, it is therefore also necessary to include the theoretical spectra not containing it.

Consequently, modifications, whether fixed or variable, need to be specified before running the search so that the search tool can generate the theoretical spectra accordingly by adding the mass shifts corresponding to these modifications. While fixed modifications do not represent a computational problem as they only require shifting the peaks on the spectrum, some challenges can arise with variable modifications. Indeed, since for each residue that may contain a variable modification, it is necessary to generate two theoretical spectra and a single peptide may have several of such residues, the number of theoretical spectra that needs to be generated grows exponentially with the number of variable modifications included in the search. Let's take as an example the sequence QYPMHISDTR. If we only include phosphorylation as variable modification, we already have to generate 4 different theoretical spectra:

- QYPMHISDTR
- QY(Phosphorylation)PMHISDTR
- QYPMHISDT(Phosphorylation)R
- QYP(Phosphorylation)MHISDT(Phosphorylation)R

If we also add oxidation of methionine, another common PTM, we now have to generate 8 different theoretical spectra:

- QYPMHISDTR
- QY(Phosphorylation)PMHISDTR
- QYPMHISDT(Phosphorylation)R
- QYP(Phosphorylation)MHISDT(Phosphorylation)R
- QYPM(oxidation)HISDTR
- QY(Phosphorylation)PM(oxidation)HISDTR
- QYPM(oxidation)HISDT(Phosphorylation)R
- QYP(Phosphorylation)M(oxidation)HISDT(Phosphorylation)R

Having too many theoretical spectra is firstly an issue in terms of computing time, as the processing time grows linearly with the number of theoretical spectra. More importantly, having too many theoretical spectra increases the risks of false identifications, as it is more likely an observed spectrum may match a theoretical spectrum that does not correspond to it just by chance. Therefore, in practice, it is necessary to limit the number of variable modifications included to only a few.

1.2.3.1.3 Spectral matching Finally, another type of peptide identification is spectral matching. It is similar to the peptide spectrum matching approach, except that instead of generating a theoretical spectrum from a peptide sequence obtained from a reference database, in spectral matching, the database contains the reference spectrum already. These spectra are spectra that were previously observed in other experiments and for which the corresponding peptide was identified. Comparing an observed spectrum to another observed spectrum greatly improves computation times and potentially the accuracy of identifications compared to the sequence based theoretical spectrum approach. However, this method suffers from two main issues. First, it is dependent on the availability of reference spectra for this species study. Secondly, spectra are dependent on the mass spectrometer

technology used, the experimental protocol and the manipulation. Therefore, a observed spectrum may not match a spectrum in the database even if they come from the same peptide, because the experiment wasn't performed in the same way.

1.2.3.2 False discovery rate

Regardless of the peptide identification method and tool used, false identifications are inevitable. It is thus necessary to develop a metric to estimate confidence in identifications and be able to filter out identifications that are most at risk of being incorrect. Failing to do so may result in the identification of the wrong proteins, fewer proteins being identified, or less accuracy in protein quantification because of inconsistent results between peptides mapping to a protein. Ideally, we would like to know the false positive rate (FPR), that is the number of false identification divided but the total number of identifications. However, since we do not know the sample composition, we don't know what the false identifications are. Hence, we must estimate the FPR through another metric called false discovery rate (FDR). To do so, a target-decoy approach is used. In addition to the database containing the proteins we can expect to find (for example the Uniprot human proteome), a decoy database is also generated. To generate a decoy database, a common practice is to take the true protein sequences and to reverse them, or to generate a collection of random sequences with the same statistical characteristics as the true protein database. These decoys represent false proteins that are therefore not in the sample, meaning if a spectrum is matched to a decoy peptide, this is a false identification. Each peptide identification tool has its own way of calculating a score indicating how good the match between the observed and theoretical spectrum is or the probability of the peptide identification being correct. After concatenating both target and decoy sequences and running the search, the distribution of identification score usually follows a distribution similar to the one in Figure 1.17.

Target peptides tend to have higher score, since most identifications are correct, although some have lower scores and among them will be most of

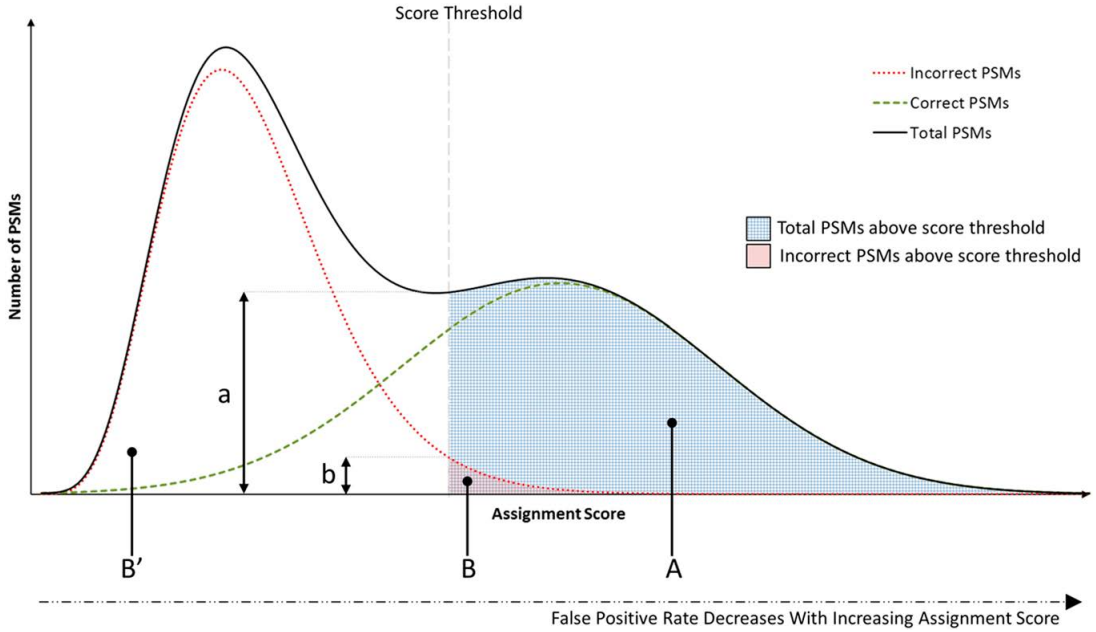

Figure 1.17: Distribution of peptides scores for target peptides (green) and decoys (red). (Bessant, 2017)

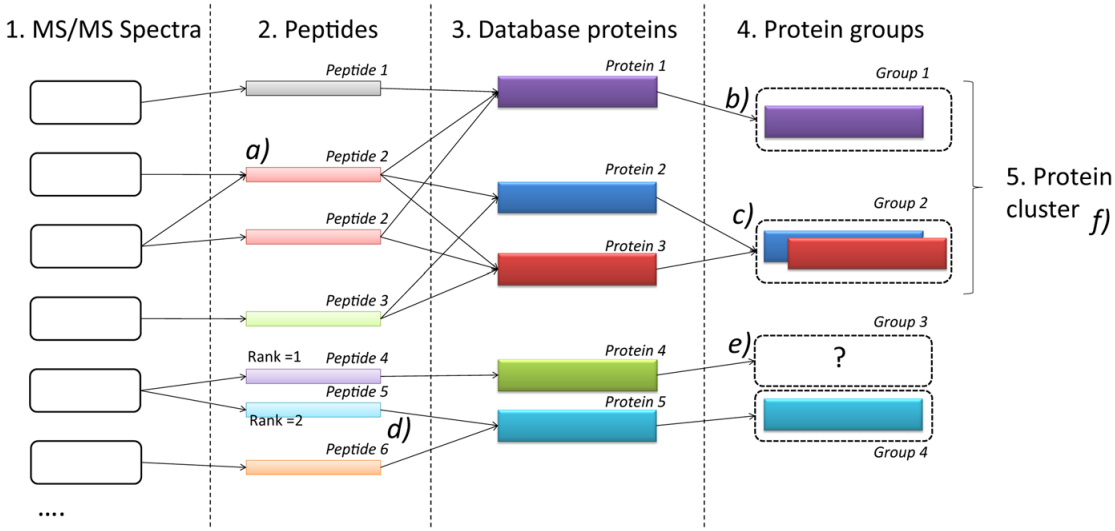

Figure 1.18: Example of how PSM are associated to peptides which are mapped to proteins that are then joined in protein groups based on the peptides they share.

the false identifications, although false identifications with high scores can happen but are less likely. Identifications of decoy peptides, which are by

definition false identifications, except in the rare case where the sequence of a decoy peptide was actually present in the sample but not in the target database, tend to have a lower score as these identifications only happen by chance and therefore are usually not a strong match. We can therefore observe two normal distributions for the target and decoy identifications, with the target distribution shifted higher on the score axis. However, the two distributions will overlap, meaning that some of the decoy identifications will have higher scores than some of the target identifications. The FDR is a value defined by the user indicating the proportion of false identifications we tolerate in our search. In the field of mass spectrometry, this threshold is usually set to 1%, meaning among 100 peptide identifications selected randomly, we would expect one to be incorrect. Once the FDR has been chosen, a score threshold is determined. The threshold corresponds to the value where 1% (in the case of a 1% FDR) of the decoy peptides will have a score higher than the threshold (red area of Figure 1.17). Then, all target peptides with a score below the threshold are removed, as we can expect them to have more than 1% of false identifications. Increasing the FDR increases the threshold score and therefore discards fewer peptides, at the cost of having more false identifications. On the other hand, lowering the FDR will result in fewer false identifications but also fewer true identifications. It is also worth noting that while it would seem intuitive to think that a bigger database would increase the number of peptide identifications, a bigger database will also increase the probability of false identifications as there are more theoretical spectra than can match an observed spectrum by chance. This means the decoy identification scores distribution will tend to be shifted to the right on the score axis compared to a smaller database, meaning that in order to preserve a 1% FDR, the score threshold will need to be set higher, which may result in fewer identifications. Therefore, having a bigger database size may actually decrease the number of identified peptides compared to a smaller one. It is therefore important to try to build a database as specific as possible, that contains everything we can expect to find without adding proteins that cannot be found.

Some tools have more advanced algorithms to estimate which peptide

identifications are correct or incorrect. For example, PeptideProphet (Ma et al., 2012) fits a normal distribution for target peptides and a Gaussian distribution for decoy peptides, using PSM properties, allowing to calculate of probability for each PSM to be correct. This approach has proven to have greater sensitivity than the standard target-decoy FDR approach. In addition, some tools rely on machine learning, such as Percolator (The et al., 2016), in order to increase the number of PSM identified for a given FDR level.

1.2.3.3 Protein inference

Once peptides have been identified, the next step is to map them to their original protein. Indeed, each peptide comes from a protein, but a peptide may sometimes map to multiple proteins, which can be isoforms or share similar function and since peptides are small sequence, it is also possible that a peptide maps to multiple proteins just by chance. Yet, there is no way to know experimentally from which protein a peptide comes from. Therefore, some inference algorithms must be used to try to map peptides to proteins and determine which are the proteins present in the sample. This class of algorithms are called protein grouping algorithms.

Peptide identification tools identify peptide spectrum matches (PSM), predicting what is the peptide sequence corresponding to a given spectrum. Therefore, an identified peptide will have one or more PSM corresponding to it (1.18). A peptide sequence can map to one or more proteins, and a protein can have one or more peptide mapping to it. Ideally, for a protein to be confidently identified, we would like to have multiple peptides mapping to it, as if this protein was not present in the samples these peptides would not have been detected. There is an ongoing debate in the proteomics community about the so called "one hit wonders", that is to say, proteins that have only one unique peptide mapping to them. Some consider that we cannot confidently believe in the presence of a protein with only one peptide supporting it, as this peptide could either be an incorrect match or map to another protein that was not present in the search database. Others argue

that highly confident one hit wonders should still be considered, as this policy shows no decrease in accuracy of the proteins identified while increasing their number (Gupta and Pevzner, 2009).

When multiple proteins share similar peptides, these proteins form what is called a protein ambiguity group (PAG), meaning that one or more of the proteins of this group may be present in the sample, but it may be difficult to know which. Different types of PAG exists and are presented in Figure 1.19

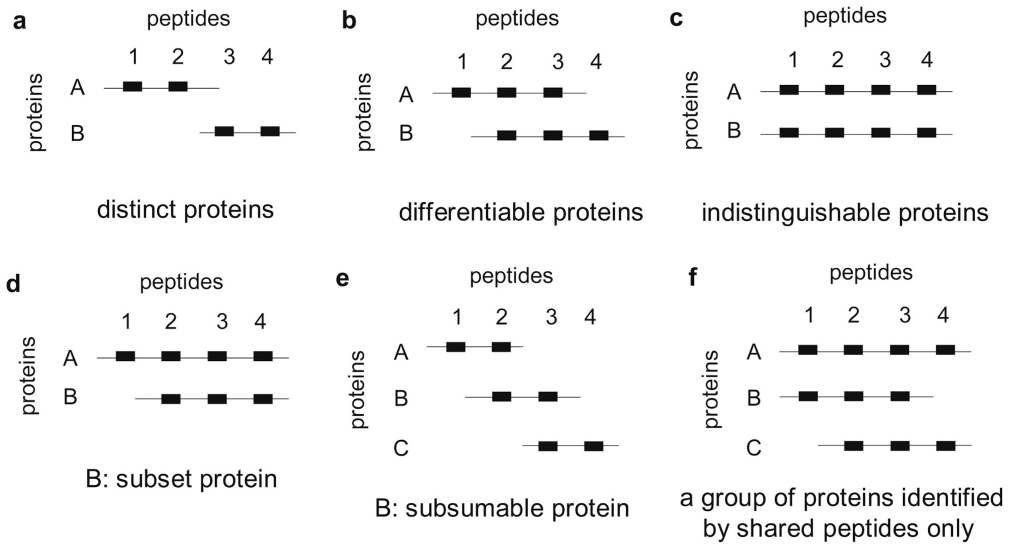

Figure 1.19: Different configurations of protein ambiguity groups (Bessant, 2017)

A commonly used heuristic for dealing with protein ambiguity groups is to use Occam's razor, which states (in one of its many formulations): No more things should be presumed to exist than are absolutely necessary. The fewer assumptions an explanation of a phenomenon depends on, the better the explanation.

This principle can be used for example in case d on Figure 1.19 where the PAG contains proteins A and B. Peptide 1 maps uniquely to protein A, meaning we can be confident in its presence. However, all peptides mapping to protein B also map to protein A, which we know to be present. Therefore, the hypothesis of the presence of protein B requires more assumptions than

simply stating that only protein A is present.

In practice, protein ambiguity groups are often more complex than those presented in Figure 1.19, sometimes containing more than 10 different proteins. In some situations, it is not possible to resolve a PAG, for example in the case of indistinguishable protein (Figure 1.19 c). This happens when several proteins share the exact same set of peptides, making it impossible to know which protein is actually present in the sample. This situation can often happen if protein isoforms are present in the search database. Unless a peptide covers the part which is different between the isoforms, it is not possible to distinguish between them.

Multiple tools, such as ProteinProphet (Ma et al., 2012) exist and use different strategies to resolve PAGs and compute the probabilities of each protein being present in the sample.

1.2.3.3.1 Limitations While LC-MS/MS usually allows the identification of tens of thousands of peptides and thousands of proteins, multiple issues remain. The first one is related to coverage. Indeed, unlike Next Generation Sequencing for transcriptomics, it is rare to have the whole protein sequence covered by peptides and most proteins will only have one or a few peptides mapping to them. This is an issue with regard to protein ambiguity groups, as it sometimes doesn't allow finding which protein is present in the sample. In addition, low coverage mass spectrometry is not always suitable for the study of isoforms, as unless a peptide covers the regions that are different between isoforms, it is not possible to know which one is present in the sample. Protein abundance is also an issue, as mass spectrometry is biased towards more abundant proteins, meaning it is more difficult to detect low abundance proteins.

1.2.3.4 Quantification

Once peptides have been identified, they can also be quantified in order to know their abundance and, from them, infer protein abundance or abundance of a post translation modification. There are two main types of quantification

in mass spectrometry: absolute or relative quantification. In absolute quantification, we want to quantify how much of a peptide or protein is present, whereas in relative quantification, we want to obtain the abundance of a peptide or protein in a sample compared to another sample. There are multiple ways to perform absolute quantification, and it can be useful for example for biomarker measurement (Ankney et al., 2016) but we will here focus on relative quantification as this is the one we used for our pipeline.

1.2.3.4.1 Label-free Label free quantification is a method to quantify without any labels. A first and simple way of doing so is by spectral counting . Since a peptide can have multiple PSMs mapping to it, and under the assumption that the more abundance a peptide is, the more spectra will be found mapping to it, this method offers a way to quantify a peptide. Yet, since not all peptides have the same detectability because of the physical and chemical properties of their peptides, this means the abundance of a peptide can only be compared to the abundance of the same peptide in a different sample, hence the term relative quantification. To obtain the protein abundance, the sum of all the PSMs of all the peptides mapping to the protein is calculated and then divided by the protein length to normalise, since longer proteins tend to have more peptides mapping to them. Other metrics derived from spectral counting exist such as emPAI which takes the ratio of the number of observed peptides by the number of observable peptides (peptides obtained for the protein from in-silico digestion) (Ishihama et al., 2005). While simple to calculate, spectral counting is not as accurate as other methods as the way it is calculated offers a low resolution, and relies on the assumption that the number of spectra or peptides observed is highly correlated to protein abundance.

An important thing to consider, and which applies to all quantification methods, is the case of degenerate peptides (peptides mapping to multiple proteins). Indeed, in this case, different proteins may contribute to the peptide abundance, or the peptide may be used to quantify a protein that is not present in the sample as this peptide also maps to another protein. Some tools completely discard peptides that map to multiple proteins for quantifi-

cation. Others only include unique and razor peptides for quantification. A razor peptide is a peptide mapping to multiple proteins, but where the protein ambiguity group was resolved. For example, in 1.19 d, we determined that protein A was present in the sample and protein B wasn't. Therefore, peptides 2, 3 and 4 and considered to be razor peptides and according to some tools, can therefore be used for quantification. Finally, some tools include degenerate peptides for quantification but weight their intensities by the probability of the protein, to give them less contribution to the overall protein intensity than the unique or razor peptides.

A more complex and accurate way to do label free quantification is to integrate the area under the peak or extracted ion chromatogram (XIC) on the MS1 spectrum (1.20). On this type of spectrum, the retention time is represented on the x-axis and the ion count on the y-axis. Since retention time is dependent on peptide mass, a peak will correspond to a peptide ion. Therefore, by calculating the area under the curve, it is possible to calculate the peptide intensity. However, in a label free experiment, it is necessary to perform one mass spectrometry run per experiment as there is no label to differentiate between peptides coming from different samples. Therefore, since retention times may vary from one experiment to another, the retention times also need to be aligned.

1.2.3.4.2 Labeled Another way of quantifying peptides is by using labels. In this case, these labels are molecules which are added to the samples. An example is Stable Isotope Labeling with Amino acids in Cell culture (SILAC) Ong et al. (2002). SILAC labels allow distinguishing between different samples in a single mass spectrometry run. During cell culture, one sample is fed with normal amino acids that are incorporated to the proteins, whereas another sample can for example be fed with lysine that contains 6 carbon-13 atoms while they normally contain 6 carbon-12 atoms. Therefore, peptides will have their MS1 mass shifted by 6 Da for every lysine they contain, compared to the unlabelled samples. Thus, if this variable modification is included in the search, it is then possible to know from which sample a peptide comes from. It is even possible to

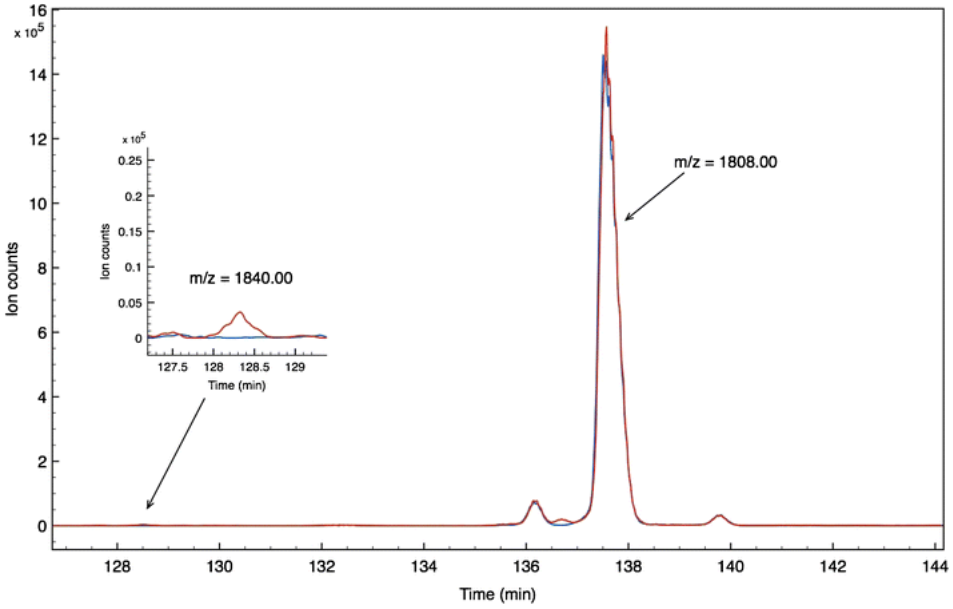

Figure 1.20: Extracted ion chromatograms (XIC) (Bane et al., 2017)

have 3 different samples in the same run, where one is untreated, another is fed with D_4-lysine/$^{13}C_6-$arginine (Lys4/Arg6) and the last one with $^{13}C_6\ ^{15}N_2-$lysine/$^{13}C_6\ ^{15}N_4-$arginine (Lys8/Arg10) (Zhang et al., 2014). The fact these labels are on the same amino acids that trypsin cuts after is no coincidence: by doing so, we ensure that each tryptic peptide (and therefore the huge majority of peptides in the sample, except the last peptide of a protein), will contain a label. SILAC labels are only used to differentiate between different samples, but the peptide intensity is still calculated the same way as it was with label free quantification, by integrating the MS1 spectrum.

Another type of labels commonly used are Tandem Mass Tags (TMT) (Thompson et al., 2003). Here, the labels are isobaric mass tags that are added to the samples after the proteins have been cleaved by an enzyme and which bind to the peptides. Each sample is given a different label in order to later be identified. Multiple TMT kits can be bought, offering up to 16 different labels, meaning up to 16 different samples can be put together in a single mass spectrometry run, resulting in much less time and spent on running the mass spectrometer. During the fragmentation process in the

mass spectrometer, a part of the label is cleaved, resulting in a fragment called the mass reporter, which will have a slightly different mass for each label. Thus, on the MS2 spectrum, a group of peaks close to each other appears on the spectrum (1.21). Each reporter ion corresponds to the intensity of the peptide in the sample that was treated with this label. Since we know the mass of each label and each peak detected, it is easy to know which peaks corresponds to which sample. Then, by comparing the intensity between two peaks, it is possible to calculate the relative abundance of the peptide between these two samples. The advantages of TMT over label free is that it allows combining different samples, saving time on experiments and intensities are more accurately estimated. In addition, since all samples are put together, this remove the variations seen from one mass spectrometry experiment to another and which may affects the results.

1.2.3.4.3 Protein level quantification Relative protein abundance can be determined by the relative abundance of the peptides mapping to it. However, it is important to take into account variations that can happen between experiments or between samples, for example, if during sample preparation, the same quantities of each sample are not added, it will give the impression that most of the proteins of a sample are over expressed compared to the other samples, even though it is not the case. In addition, random effects are to consider, meaning relative changes among peptides mapping to the same proteins will always have some level of variance. However, the more peptides map to a protein, the more accurately it is possible to estimate a protein abundance. Packages such are MSstats (Choi et al., 2014) build statistical models that take all these parameters into account and can predict protein abundance and if proteins are differently expressed between different samples within a FDR level. Tools such as MaxQuant (Cox and Mann, 2008) only use unique or unique and razor peptides for quantification. Other tools (Blein-Nicolas et al., 2012) also use degenerate peptides for quantification, as this allows to have more peptides used for quantification.

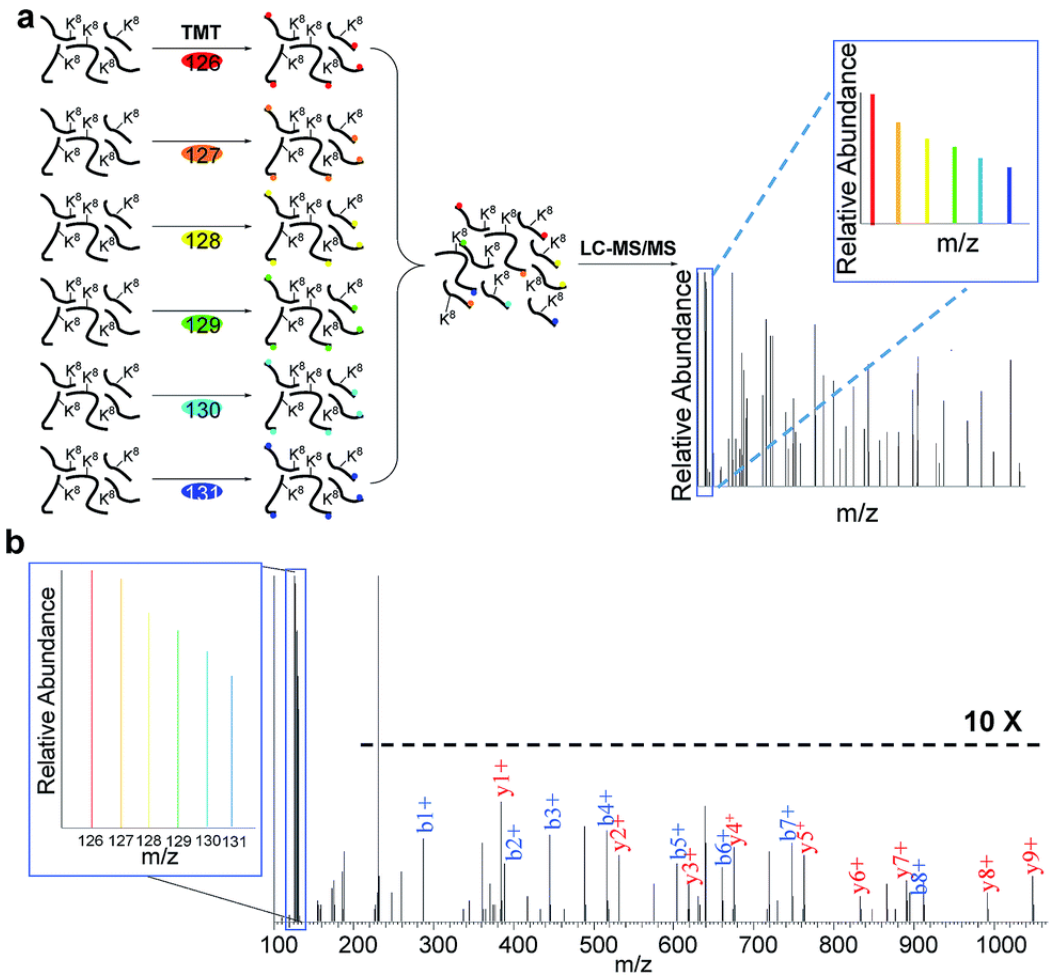

Figure 1.21: An MS2 spectrum with TMT labels(Chen et al., 2016). The peaks on the left are the mass reporters, and their intensity (value on the y-axis) are used to quantify the peptide in each sample.

1.2.3.5 File formats and data standards

Multiple file formats exist in order to manipulate data coming from mass spectrometry experiments or analysis. These files can either be the output of the mass spectrometer that contain the spectra information or the output of the peptide identifications tools that contain the list of identified peptides and additional information linked to them. The Proteomics Standards Initiative (HUPO-PSI) (Orchard et al., 2003) is a consortium of proteomics researchers who define and maintain different file formats.

To store the data coming out of the mass spectrometer, three file formats are commonly used, although others exist (Deutsch, 2012). Files in .raw are files that are outputted by the mass spectrometers from some manufacturers, such as ThermoFisher Scientific. These file contain the raw MS1 and MS2 spectra that can be used for peptide identification. However, this is a proprietary standard in binary format, making it unreadable in text format and also implies that many proteomics tools have little or no support for this format as their input. The Mascot Generic Format (MGF) is a text file format containing MS2 spectra. For each spectrum, a header is written on the first line, with one type of information per line, such as the precursor mass. Following the header, peaks are written, with on each line the mass to charge ratio and intensity for a peak. In recent years, another file standard, mzML (Martens et al., 2011), has gained popularity. It is written in XML which is a flexible file structure using nested tags that can contain attributes in key-value pairs. The mzML format defines a Controlled Vocabulary (CV), which is a list of tags that can be used in mzML files and defined by the HUPO-PSI. Compared to .mgf files, .mzML files are more detailed as they can also contain the MS1 spectrum and additional information. Their flexible format, within the vocabulary allowed, enables them to be used for multiple types of experimental designs, regardless of the mass spectrometer manufacturer. As such, it is now supported and preferred by most proteomics tools (Deutsch, 2012).

For tools that do not support .raw files, ThermoRawFileParser was developed, (Hulstaert et al., 2020) which allows converting .raw files into .mgf or .mzML format.

Peptide identifications tools produce different types of files that can be used for data analysis. Some tools such as MaxQuant (Cox and Mann, 2008) produce multiple files for peptide or PSM identifications, PTMs, protein groups in tabular format. While this format is specific to MaxQuant, it has the advantage of being easy to use. Furthermore, since MaxQuant is a popular peptide identification tool, downstream analysis packages (mostly written in R) have been developed that can directly read MaxQuant output, such as Proteus R (Gierlinski et al., 2018). However, HUPO-PSI standards

exist for peptide identifications file formats such as mzIdentML (Jones et al., 2012) or its simplified, more readable alternative in tabular format, mzTab (Griss et al., 2014). These formats are well established and supported by many downstream proteomics analysis tools, such as the Trans-Proteomics Pipeline (Pedrioli, 2010).

1.3 Research aims

The aim of this project was to develop a software pipeline integrating data coming from RNA-Seq and mass spectrometry. This multi-omic integration allows matching what is observed at the RNA level with what is observed at the protein level, with data coming from the same sample, enabling more sensitive detection of variant proteins such as those produced by alternative splicing. While several tools combining RNA-Seq and mass spectrometry data exist, such as PASS (Wu et al., 2019) or MSProGene (Zickmann and Renard, 2015), none offers a complete integrated pipeline from the raw data to the multiple outputs possible from RNA and protein abundance, to alternative splicing or mutations, or the construction of novel proteomes as well as visualisation of results. In addition, biology is inherently quantitative and research usually involves comparisons of different groups to identify specificities between them, for example comparing a group having received a treatment versus a control group, or a healthy group versus a group with disease. Therefore, our pipeline must be able to compare different conditions, each containing multiples sample or replicates, at both the RNA and protein level. Finally, in order to be usable for a large range of research projects, the pipeline needs to be able to support the multitude of experimental designs and RNA-Seq and mass spectrometry technique used in different contexts.

We therefore aimed to further develop the PIT (Proteomics Informed by Transcriptomics) pipeline, first published by Saha et al. (Saha et al., 2018) and take it further to include all the elements previously mentioned, and make it a publicly available tool for all researchers possessing RNA-Seq and mass spectrometry data.

Once developed, this pipeline was applied to multiple datasets, to provide

biological answers to various research questions. In particular, we used PIT to look at the effect of splicing factor HNRNPA2B1 in prostate cancer in order to understand its mechanisms of action as well as its impact on survival.

Chapter 2

Development of a quantitative proteomics informed by transcriptomics (PIT) pipeline

2.1 Introduction

In this section, we introduce the Proteomics Informed by Transcriptomics (PIT) pipeline. It enables the analysis and aggregation of data coming from RNA-Seq and LC-MS/MS. While previous implementations of such a pipeline were previously released, they were only focused on one specific aspect, such as finding peptide evidence for mutations. Thus, they lacked essential aspects in order for the pipeline to be used in a wide range of applications. This includes adding support for additional layers of analysis such as quantification, alternative splicing, post translation modifications and others. Additionally, as most biological experiments are about understanding the effect of perturbators (drug, virus, bacteria, ...) or genomic elements (RNA, proteins, ...) between different conditions, it is also essential to be able to compare these conditions in order to see what is different between them at the RNA and protein level. Finally, for the software to be widely adopted, a greater focus needs to be dedicated to usability. This means automation, to be able to run the analysis easily, as well as a way of visualising the results that come out of

PIT in a graphical way. Here, we rebuilt the PIT pipeline in order to make it a software suite that can be used by researchers and expended the scope of analyses it provides so that it can show itself more helpful for multi-omics data analyses.

2.2 Initial version and limitations

PIT (Proteomics Informed by Transcriptomics) is a software pipeline first described in 2012 (Evans et al., 2012). The goal of this pipeline was to remove the need for a canonical peptide database, by inferring one coming from RNA-Seq using the same samples for both the RNA-Seq and mass spectrometry. Using such a sample specific database offers several advantages. First, it allows working with non-model organisms. Indeed, while the human and mouse proteomes have been widely studied and offer at least one sequence for most proteins present in these organisms, this isn't the case for most species. In their article, Evans et al. used as an example the Chinese hamster (Cricetulus barabensis griseus). Yet, the most comprehensive protein sequences database, Uniprot, only contains 13 reviewed proteins from this species.

Figure 2.1: The Cricetulus barabensis griseus Uniprot proteomes

The remaining 23,872 were sequences that were predicted computationally by translating in silico RNA sequences referenced by the International Nucleotide Sequence Database Collaboration (INSDC). This consortium collates data from major institutes such as DDBJ, EMBL, GenBank. Yet, these proteins have not been observed experimentally, and may not be genuine gene products of Chinese hamster. On the other hand, some sequences may be present in the sample but not in the database. This can be caused by some genes that are only expressed in certain conditions or by non-synonymous mutations that can change one or several nucleotides in the protein sequences, or even insertions or deletions which can cause a frame shift, resulting in two completely different proteins.
These considerations highlight two different and opposite issues. On the one hand, having sequences present in the database and not in the sample make the database bigger than it needs to be. A bigger database means a higher risk of a spectrum matching to a peptide by chance, even though this peptide is not actually present. This would be seen at the peptide identification level by the decoy peptides distribution shifted towards a higher confidence score. Therefore, in order to maintain a 1% FDR the threshold needs to be put at a higher score, therefore identifying fewer peptides. On the other hand, if the database is not specific enough, some peptides present in the sample, for example resulting from mutation, cannot be detected. Worse, the spectrum from this peptide may match to a different peptide in the database because this is the closest match, potentially resulting in misidentifying a protein. Even for humans, which have a fairly comprehensive proteome as it is a well studied species, some proteins may be irrelevant in a given study, as some of them are present in some tissues but not others, therefore the database contains proteins that will not be present in the sample studied.

In addition, a non-specific database can cause issues for protein quantification. Indeed, in order to quantify a protein abundance, only the peptides that uniquely map to it are used. If the database is not specific, as is the case with a canonical database, a peptide which is thought to be unique may actually also come from another protein. Therefore, when looking at the intensity of this peptide across different channels (in the case of TMT), this

intensity will be influenced by multiple proteins, which may lead to wrong conclusions in regard to the relative abundance of the protein of interest. Therefore, the ideal database is a database which is as specific and exhaustive as possible while also remaining as small as possible.

Evans et al. used Trinity (Grabherr et al., 2011) to perform de novo assembly of RNA-Seq in order to obtain transcripts. From the transcripts, open reading frames (ORF) are predicted in order to get the most likely protein sequences. These sequences are then used as the input database to MaxQuant, which then performs peptide identification and protein grouping. Transcript sequences are aligned to a reference genome using GMAP (Wu and Watanabe, 2005) for annotation, including exon mapping. Finally, the protein sequences can be aligned to known protein databases using BLAST (Altschul et al., 1990) in order to infer their function and differences to other organisms.

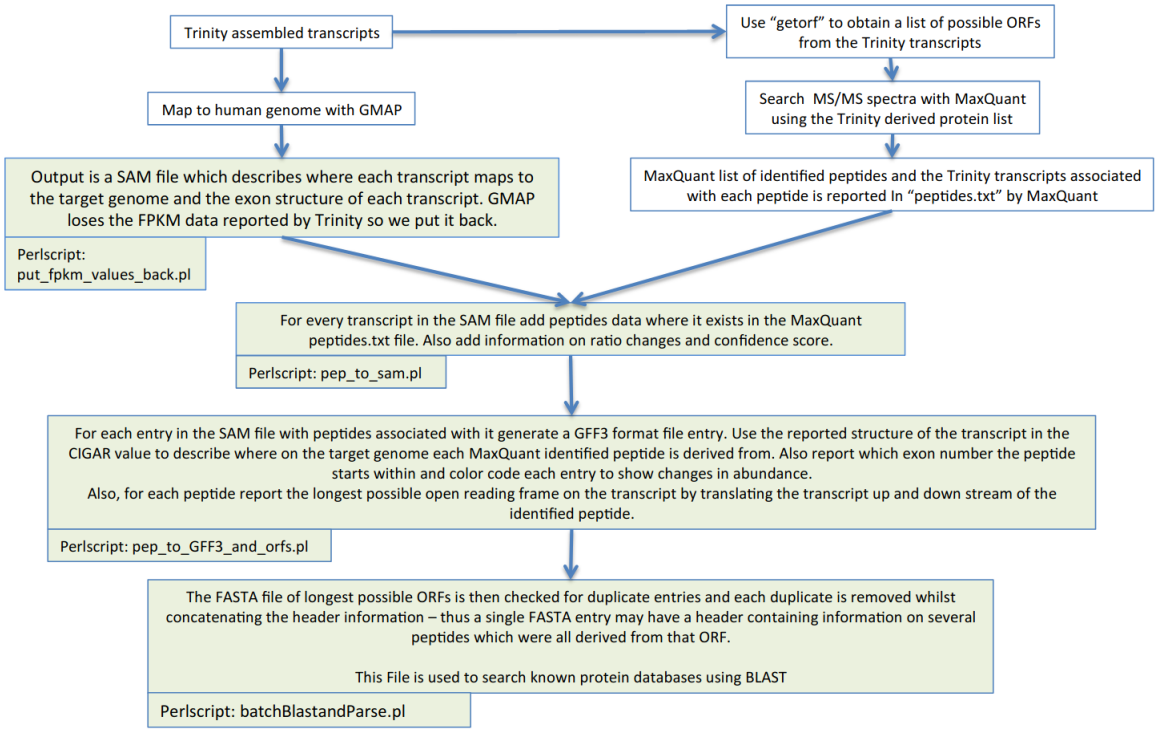

Figure 2.2: Flow chart of the first version of PIT (Evans et al., 2012)

However, this version was not fully automated, as it required to use the program to generate the database, but the user would then have to use MaxQuant manually to run the peptide identification. Additionally, it would not provide any downstream analysis on the peptides that had been identified and was only meant to use MaxQuant with an RNA derived peptide database.

In 2018, a new version of PIT was published. (Saha et al., 2018)

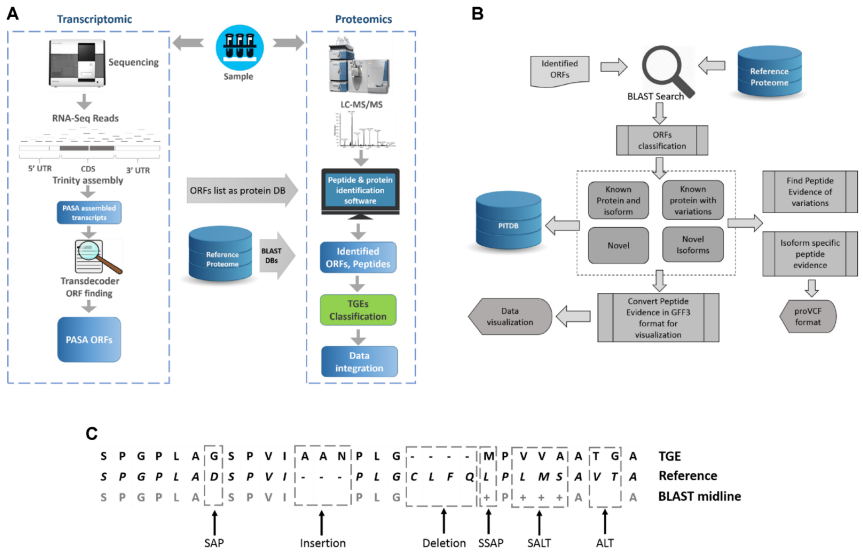

Figure 2.3: Flow chart of the second version of PIT (Saha et al., 2018)

This version of PIT was an extension of the one previously described. Similarly, it relied on an assembly of RNA-Seq reads into transcripts using Trinity and prediction of open reading frames in order to generate a mass spectrometry database. One addition to this version was its full automation. Indeed, once generated, the database was sent to MSGF+ (Kim and Pevzner, 2014) which was run through bash to perform peptide identification. In addition, the pipeline offered some new features. First, it allowed using BLASTp to align the predicted ORF against a reference proteome in order to find variations such as mutations resulting in a different amino acid sequence. Identified variations could then be exported to VCF format.

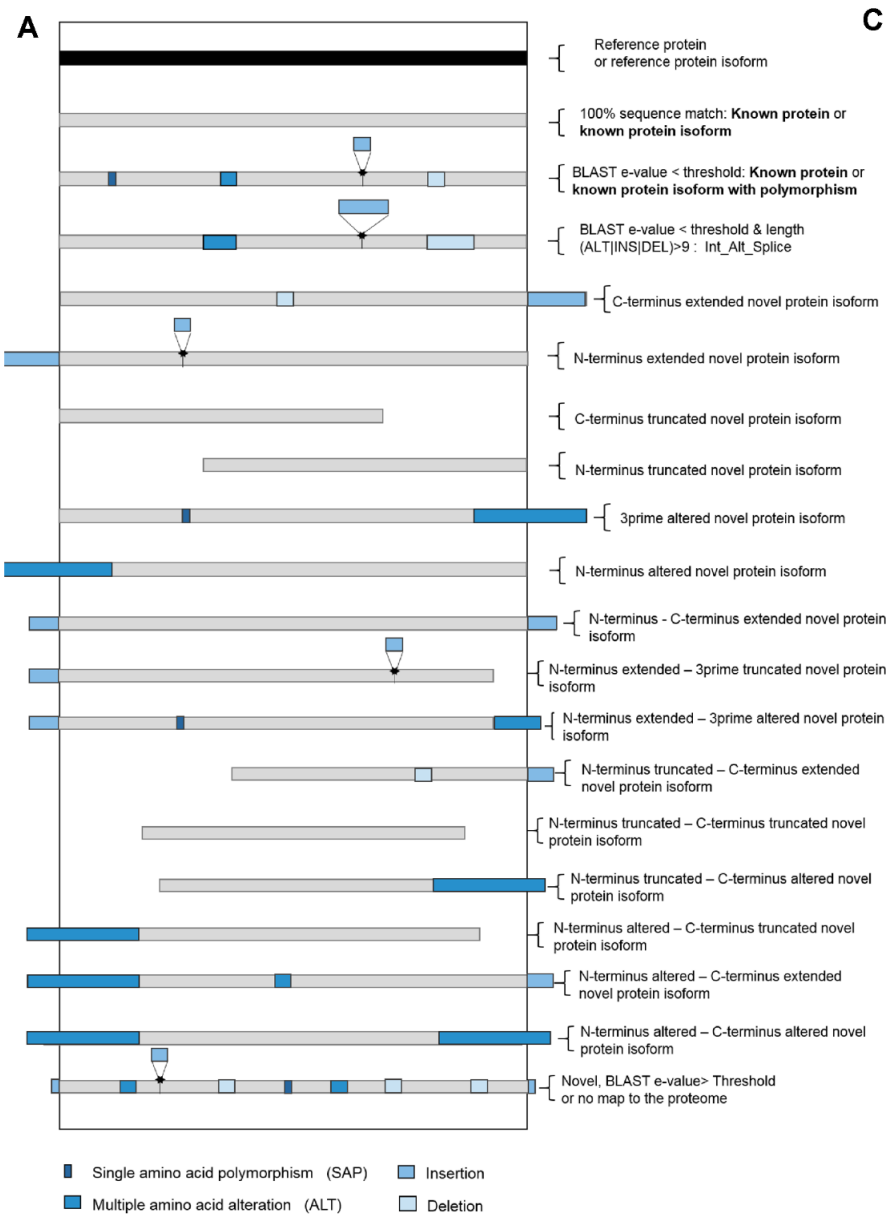

Figure 2.4: Classification of the different ORFs identified by PIT with regard to the BLASTp alignments

Depending on the differences with the best reference sequence found by BLASTp, differences could be identified. A difference of one amino acid could be explained by a non-synonymous mutation, whereas are difference of multiple amino acids (typically more than 7) could be explained as an

alternative event or a frameshift. Additionally, differences of size on the 5' or 3' UTR could also be explained as splicing events, although in case of C or N terminal truncated, it wasn't possible to distinguish between an isoform or a cut in the transcript assembly due to a lack of read coverage or other artefacts.

Yet, the strongest point of this pipeline was its ability to identify peptides from mass spectrometry overlapping with the variations, thus proving their existence at the protein level. This information is particularly relevant, as it may be impossible to infer translation from mRNA only. Indeed, it is known that 5' UTR regulate translation (Araujo et al., 2012), therefore, a splicing event on the 5' UTR, even if it doesn't affect the protein sequence itself, may affect how it is translated and therefore the abundance or even existence at the protein level. Furthermore, because of the nature of transcriptome sequencing, errors can appear in the assembled transcripts. These sources of error are multiple: sequencing errors on the reads resulting in a wrong sequence (particularly with long read sequencing which has a higher error rate than NGS) or the fact that short reads can map to different places in the genome and are known to perform poorly in repetitive region (Tørresen et al., 2019). Having peptide overlapping with these variations provides evidence from an independent experiment that they are indeed correct.

However, while they can already provide useful information, the two versions of PIT described above present some limitations. The first point that can be made is that they are not fully de novo approaches. Indeed, Evans et al. use GMAP to aligns the assembled transcripts to the genome, whereas Saha et al. use BLASTp to align predicted ORFs to a reference proteome in order to find variations in their samples. Either way, there is still a strong dependence on a well annotated reference, implying limited results on non-model organisms.

Furthermore, these pipelines only work for one sample and do not allow the aggregation and comparison of multiples replicates and conditions. This

can be an issue as errors can arise from either sample preparation, the state of the sample at a given time, sequencing or computational analysis. Because of these uncertainties, it may be difficult to know if an observation is real and can be generalised, or is specific to the sample observed. To address this issue or sample variability, multiple replicates are commonly used in biological experiments. This allows the use of statistical methods such as hypothesis testing in order to estimate the probability of a hypothesis being correct. In many cases, users may also want to compare results across multiple conditions in order to identify differences or the impact of a certain procedure.

While an important aspect of mass spectrometry is peptide and protein identification, which was supported by the previously described versions of PIT, sometimes identification is not informative enough as it is a binary state (presence or absence of a protein). Rather, peptide and protein quantification can give a more informative perspective. This is for example the case with alternative splicing. Splicing events often present small changes in inclusion or exclusion of a given exon. Yet, this small change can sometime be enough to cause significant effects on phenotype. In this case, simply identifying peptides will not allow detecting these changes, unless peptide quantification is also used.

Finally, it can be argued that as the pipeline only produces comma separated text files, interpreting results is a challenging task, especially considering it integrates information from multiple levels (RNA and protein).

Thus, all these aspects were taken into consideration in order to develop a third version of PIT that integrates these additional features and offers a data architecture suitable for further developments and analysis of the results by researchers.

2.3 Datasets used

For the development and testing of PIT along its development, three datasets were mainly used. The first one was an experiment silencing the PTEN

gene in prostate cancer cells 2.3.1. Another one was an experiment silencing HNRNA2B1 in PC3 cells 4.5. Experimentally, while these two experiments were both performed using RNA-Seq and LC-MS/MS, they differ by the fact that the former uses TMT labels for LC-MS/MS whereas the latter uses SILAC labels. In addition, the PTEN dataset includes a run enriched for phosphopeptides in addition to a total proteomics run. To develop PIT for orphan organisms, using a de novo assembly, we used data from a paper studying TRAIL-mediated apoptosis sensitisation by Hendra virus in bats (Wynne et al., 2014).

2.3.1 Silencing PTEN in DU145 cells

PTEN is a tumour suppressor gene lost in multiple cancer types, including prostate cancer (PC). The best characterised function of PTEN is at the cell membrane, where it acts as a lipid phosphatase to negatively regulate the PI3K signalling pathway. Other functions of PTEN include as a protein phosphatase, and as a scaffolding protein.

As a protein scaffold, PTEN binds RNA-binding proteins (RBPs) and proteins of the spliceosome complex 1. PTEN undergoes protein arginine methylation 2. Loss of PTEN protein, or manipulation of its methylation, can have significant effects on pre-mRNA splicing.

PTEN loss in patients has been linked with a hypoxic tumour environment (reduced oxygen supply to tumours) in prostate cancer. Hypoxia has a significant impact on pre-mRNA splicing in prostate cancer. We hypothesise that PTEN is protective against splicing changes included in a hypoxic tumour microenvironment. PTEN shuttles into the nucleus upon hypoxic induction and binds RBPs to prevent aberrant splicing. If PTEN is lost in cancer, then this protective mechanism is also lost and hypoxia induced splicing changes are amplified and promote the tumour growth.

Sample preparation and LC-MS/MS was performed by Dr. John Foster. RNA-Seq was performed by the Barts and The London Genome Centre.

2.3.1.1 Sample preparation

1. DU145 (PTEN+/-) PCa cells (2×10^6) were seeded in a 15 cm dish in RPMI +10

2. Next day cells were transfected with non-targeting (NSI) or siPTEN at 40nM using RNAiMax in antibiotic free medium according to the manufacturer's instructions (ThermoFisher 13778075).

3. After for 72 hours cells were washed three times ice-cold PBS and, using a cell scraper, were scraped into

   (a) 1ml 1x SDS solubilization buffer (2% SDS, 100mM Tris-HCl pH 7.5). Samples were homogenised using a probe sonicator and stored at - 80oC before use

   (b) OR 1ml TriReagent for RNA isolation

This resulted in 8 samples named: NS1, NS2, NS3, NS4, siPTEN1, siPTEN2, siPTEN3, siPTEN4.

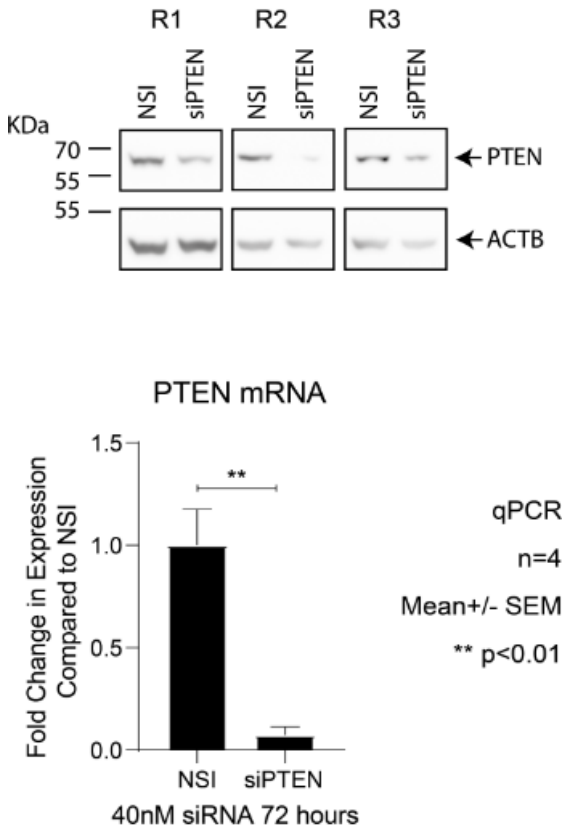

Figure 2.5: Western Blot confirmed repeats 1-3 have PTEN successfully knocked down

RNA-Seq was performed on these 8 samples using Illumina sequencer with 50 millions pair-ended 75bp reads per sample. Finally, for the proteomics, protein concentration was determined by BCA assay according to manufacturer's instructions (Pierce 23225) and 1mg protein from each sample was taken forward for labelling by Tandem Mass Tagging (TMT). Using a TMT10plex™ Isobaric Mass Tag Labeling Kit (Pierce 90113), mass tags were added to each sample. Samples were pooled and mixed, and digested using trypsin by Filter Aided Sample Preparation (FASP). 10% of the total labelled and digested pooled peptides were fractionated into 7 fractions using a High pH Reversed-Phase Peptide Fractionation Kit (Pierce 84868) for detection by mass spectrometry (MS). The remaining 90% of the sample was enriched for phosporylated peptides using a Titansphere Phos-TiO Tip

(Generon 501021307) and pH fractionated into 4 fractions for detection by MS

2.4 Rewriting the pipeline and extending the set of features: PITv3

2.4.1 Pipeline architecture

Because of the important changes in data architecture, features and usability, the decision was made to rewrite PIT from scratch.

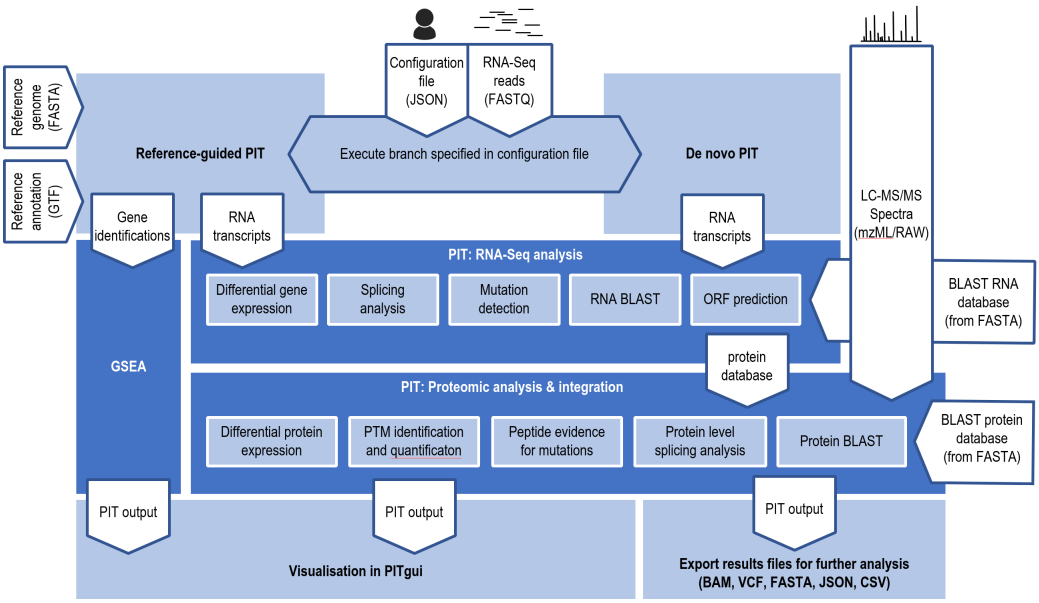

Figure 2.6: Workflow of the new PIT pipeline, from top to down. Depending on the data provided, different parts of the pipeline are run. Eventually, the results can be visualised in PITgui (described in Chapter 3).

Since the aim of this project was to develop a tool that would be helpful and usable by a wide range of researchers, from biologists to bioinformaticians, it was decided to build PIT following a modular approach 2.6. This means adding support for a large panel of experimental designs, but also allowing users to run different functions depending on what the user is interested in. For example, PIT allows the detection and quantification

of alternative splicing at the RNA and protein level. However, if the user is not interested in this part of the pipeline and cares for example about mutations, the alternative splicing part of the pipeline can be skipped. Additionally, users can provide additional data, which will allow PIT to produce additional results. For example, users can provide phosphoproteomics data. In this case, PIT will quantify the different phosphosites in each sample and perform some downstream analysis, such as finding which known kinases phosphorylate this phosphosite. However, if phosphoproteomics data is not available, is it still possible to run PIT and only these results will not be available.

2.4.1.1 Languages and dependencies

PIT was written using a combination of python3 and R. While python was used to write most of the pipeline as it offers a straightforward and flexible way of manipulating data, through data frames and dictionaries for example, while at the same time maintaining decent performance by running C code under the hood for optimal performance, such as in the pandas package. For some tasks requiring the use of specific packages, R was used. Indeed, R contains multiple bioinformatics packages, many of which are stored on the Bioconductor repository (Huber et al., 2015).

2.4.1.2 Data storage

While the version developed by Saha et al. (Saha et al., 2018) used comma separated text files, this format had limitations. Indeed, this format is typically used for tabular data, with a fixed number of columns. However, considering the variety of experimental designs supported, the different classes of evidence that may be present and the integration of multiple types of data, a more flexible format quickly became necessary. The choice was made to use JavaScript Object Notation (JSON) (Pezoa et al.) as it offers the desired flexibility.

JSON is defined as a collection (called object) of key/value pairs. The key is a string of characters and is used to access the value. The later can be

Figure 2.7: The JSON data format (202, 2020)

of different types: numbers (integer or decimal), string of characters, nested JSON objects, JSON arrays (a list of values without keys) and in some cases Null to represent the absence of value, although it isn't supported by all parsers. The main benefit of JSON is the absolute flexibility to define the key/value pairs as we choose, as well as the possibility to nest objects into others without restrictions. Therefore, a parent object representing a concept, such as an RNA transcript, can contain keys and values for storing information such as the position on the gene, as well as other complex child objects that can represent for example the ORF predicted from this transcript or the peptides mapping to it.

Figure 2.8: Example of a JSON file in PIT

In Figure 2.8, one of the files produced by PIT, the parent object represents a transcript, with the key being the transcript ID. Attributes include the transcript sequence, position on the genome, as well as more complex, nested objects giving information about the abundance in transcripts per million (TPM) in each sample as well as the position of each exon on the genome.

However, although JSON represents a convenient solution to manipulate data of the PIT pipeline internally, the files generated by PIT are not suitable for external use as they do not follow a bioinformatics data standard and therefore cannot be parsed by other tools. Therefore, some additional steps must be taken within PIT to be able to extract relevant information from these files and generate an output in conformity with data standards. An example would be mutations, which can be extracted from the JSON file in order to be exported into VCF format. (Danecek et al., 2011)

2.4.1.3 Project configuration

Most bioinformatics software are offered as one or a collection of stand-alone tools that are executed through the command line. However, these commands tend to be particularly verbose as the number of parameters increases and sometimes require running several commands one after the other. However, some tools use a different approach, requiring the user to write a configuration file, with a specific syntax and parameters. This file is then used as input to the command line tool, which then has all the information it needs to proceed with the whole analysis. In the field of mass spectrometry, we can cite MaxQuant (Cox and Mann, 2008) as an example which uses a configuration file based on the XML format. Another tool for peptide identification, Philosopher (da Veiga Leprevost et al., 2020), uses the YAML format to let the user decide which parts of the pipeline to run and with which parameter values.

Similarly, PIT now uses a user generated configuration file in JSON format, in which the experimental design is described, as well as other parameters.

2.4.1.3.1 Supporting multiple experimental designs In order for PIT to be applicable to a wide range of studies, it must be able to support a wide range of possible experimental designs. Some researchers will perform their experiment using only one sample. Others may choose to use technical

replicates in order to account for variation between replicates. Others may also use biological replicates (Bell, 2016), which usually bring more variation than technical replicates, in order to know if the effect seen is something that affects the samples globally or is an off target effect appearing in only one or a few samples. In addition, users may want to compare different conditions to find differences between then. This is commonly done in cases where siRNA (Dana et al., 2017) are used to silence a gene in some samples and look at what changes compared with control samples. In order to take into consideration inter-sample variations, several replicates are often used for each condition.

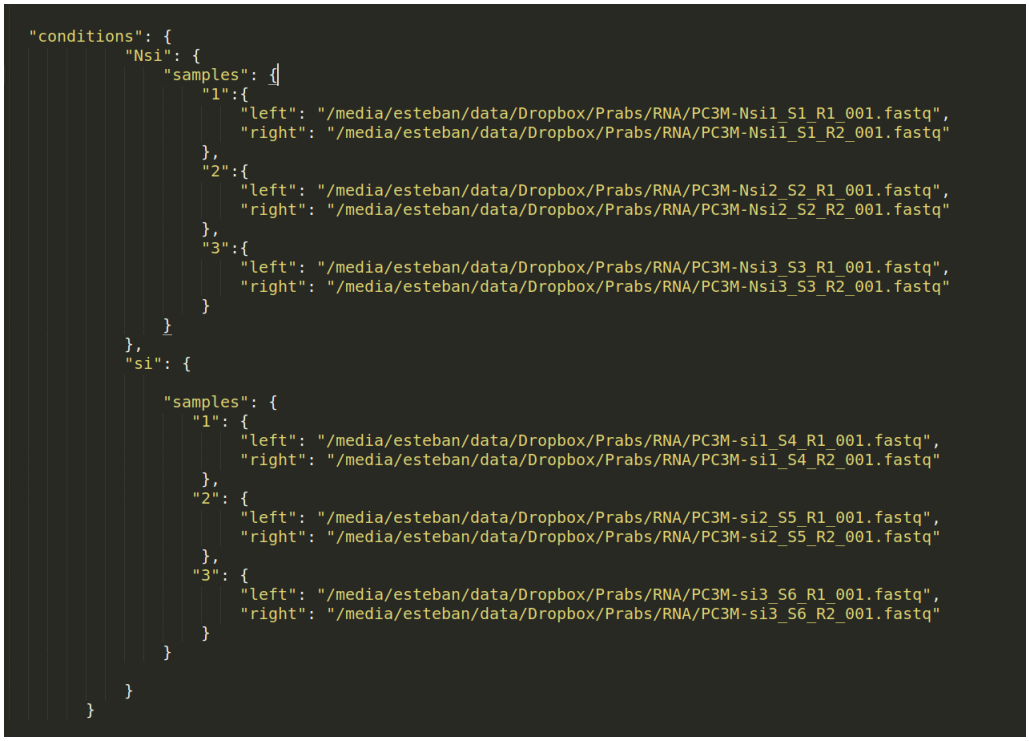

Figure 2.9: Sample definition in the PIT configuration file

At the root of the configuration file JSON (2.9), users put a key called "condition", here "Nsi" (control condition) and "si" (silenced condition). In each of these conditions, the user puts the replicates or samples belonging to these conditions. Here the replicates are named "1", "2" and "3" in each condition. However, it is worth noting that PIT supports any number of con-

ditions and replicates and doesn't limit itself to pairwise comparison. PIT aggregates all replicates together in later stages, such as for differential gene expression and therefore does not distinguish between technical and biological replicates, although it is possible to put the name of the biological replicate in the sample name. Each sample contains keys called "left" and "right", indicating the path of the fastq files containing the left and right reads in the case of pair-ended sequencing. In case of single-ended read sequencing, the user would use the key "single" instead, with the value corresponding to the path to the fastq file.

With regard to the mass spectrometry, there are again multiple experimental designs possible. The mass spectrometry part of the configuration file is centred around the concept of a run. A run corresponds to an experiment done with a mass spectrometer, with a set of parameters, a labelling technique (or no labelling) and which produces one or more output files (2.10).

```
"ms":{
    "runs":{
        "FWD":{
            "files": [      "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_01.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_02.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_03.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_04.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_05.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_06.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_07.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_08.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_09.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_10.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_11.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_12.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_13.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_14.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_15.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_16.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_17.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_18.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_19.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_20.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_FWD_Fxn_21.raw"],
            "modifications": {
                    "fixed": [

                    ],
                    "variable": [

                    ]

                },
            "SILAC": {
                "Nsi": {"label": ["Arg10;Lys8"], "samples": ["1", "2", "3"]},
                "si": {"label": [], "samples": ["1", "2", "3"]}
            }
        },
        "REV":{
            "files": [      "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_01.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_02.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_03.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_04.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_05.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_06.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_07.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_08.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_09.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_10.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_11.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_12.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_13.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_14.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_15.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_16.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_17.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_18.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_19.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_20.raw",
  "/media/esteban/data/maxquant/200415_DAS_OT_JAS_R08_230115_REV_Fxn_21.raw"],
            "modifications": {
                    "fixed": [

                    ],
                    "variable": [

                    ]

                },
            "SILAC": {
                "Nsi": {"label": [], "samples": ["1", "2", "3"]},
                "si":{"label": ["Arg10;Lys8"], "samples": ["1", "2", "3"]}
            }
        }
    },
    "combine": {
        "SILAC": {
            "runs": ["FWD", "REV"],
            "norm": "FWD"
        }
    }

},
```


Figure 2.10: Definition of the mass spectrometry experimental design in the PIT configuration file

A run is defined in the "runs" object, and the key represents the run name. The "files" key represents the mass spectrometry file, which can either be in raw or mzML format (Deutsch, 2010). It can either be the path to a single file, or a list of path in case of fractions. The user can then choose some additional fixed or variable modifications. Finally, users have to enter some information about the post translation modifications. Three quantification techniques are supported in the new PIT:

- TMT
- SILAC
- Label-free

In the case of SILAC, users define which labels (heavy, light or none) are assigned to each condition and which samples were used for the experiment. Similar objects are also defined for the other methods.

As with RNA-Seq, it is a common practice to perform multiples runs in order to account for random variations, biases or off target effects. For example, with SILAC labelling, it is recommended to perform at least two runs: one with the heavy labels on condition A and light labels on condition B and another run where the labels are swapped, light on condition A and heavy on condition B. This avoids any bias due to the presence of labels on the peptides. Therefore, in PIT, multiple runs can be combined as replicates as shown at the bottom of Figure 2.10. Here, the runs FWD and REV defined in 2.10 are combined in a parent run called SILAC and are therefore considered as replicates runs. These FWR and REV runs correspond to two replicate runs where heavy and light SILAC labels were switched between the two different Nsi and si conditions in order to avoid a bias due to the labels. This implies that the peptide intensities between the two runs are normalised, the protein fold change are aggregated and hypothesis testing can be performed.

Other parameters are also used, such as the path to the reference genome and annotation if required, the maximum number of threads that can be used

for the analysis as well as a reference genome or proteome for performing BLAST if desired. With all this information, the whole PIT pipeline can be run through a single command:

```
python Launch.py -c config.json
```

2.4.2 Reference guided PIT

In the previous versions of PIT, de novo assembly was performed using Trinity and ORFs were predicted based on the assembled transcripts. However, while de novo assembly performs best in certain situations, for example in the absence of a good enough reference genome, it also has limitations and for some situations, other alternatives are more suitable. In the presence of a reference genome, reference-guided assembly is considered to perform better, allowing to cover a more important proportion of the genome with fewer errors (Marchant et al., 2016). It is also more convenient to use, as the chromosome and the position a transcript comes from is known. In the presence of an annotated GTF file, for example from ENSEMBL, it is also possible for some of the transcripts to find their corresponding ENSEMBL id and the gene they belong to. This enables additional downstream analysis in PIT, such as Gene Set Enrichment Analysis (GSEA) or system biology analysis. Reference guided assembly is done by mapping reads from each sample to the reference genome chosen by the user using STAR2 (Dobin et al., 2012). STAR2 includes a 2-pass approach that maps the reads to the genome once and then remaps them. This approach provides better mapping for junction reads, which are reads that are on the junction between two exons, thus providing greater accuracy for transcript quantification and alternative splicing analysis.
PIT also supports long-read sequencing, and in this case, performs the alignment using Minimap2 (Li, 2018).
Once the reads have been mapped, transcript assembly is performed using Stringtie2 (Pertea et al., 2015). It outputs a GFF file containing the coordinates of all transcripts identified, as well as the coordinates of the exons of each transcript, the transcript id from the reference annotation transcript id

if the transcript coordinates match, otherwise Stringtie2 gives the transcript a unique id for the sample.

Since Stringtie2 is run individually on each sample, a same transcript found in different samples and not mapping to the reference genome may have different ids. In addition, some transcripts may be different in their coordinates by only a few nucleotides because low read coverage in a sample didn't allow reconstructing the whole transcript. Indeed, since read coverage tends to be lower at the start and end of a transcript, it is therefore not unusual to see a same transcript having slightly different start and end coordinates between different samples. To account for this, Stringtie2 offers a merge function that combines several assemblies and generates a new reference annotation that contains all the transcripts identified across all samples. If some transcripts across different samples are the same or very similar in the coordinates, they can be merged into a single transcript. If this transcript does not match a transcript in the reference annotation, Stringtie2 gives it a unique id starting with MSTRG.

Once the new reference annotation has been generated, Stringtie2 is run again on each simple, this time using the new reference annotation instead of the one used the first time. This ensures that all the transcripts share the same ids and coordinates across all samples.

2.4.3 De novo PIT

In the absence of a reference genome, reference guided assembly is not possible. Even if a reference genome is available, its quality may not be sufficient to assemble numerous transcripts, as it may contain numerous errors or only cover a small fraction of the actual species' genome. This situation can often arise with species that are not well studied. Indeed, while it is estimated that there are around 8.7 millions species populating earth (Sweetlove, 2011), release 105 of ENSEMBL only contains a reference genome for 297 species. Hence, if no reference genome is provided, PIT performs a de novo assembly using Trinity. This time, the RNA-Seq reads from all samples are combined, and a single assembly is performed using Trinity (Grabherr et al., 2011).

Trinity outputs a fasta file containing the sequences of all the transcripts identified, as well as another fasta file containing the sequence of all genes identified. Since transcriptomics RNA-Seq used mRNA as input, transcripts only contains exons without the intronic regions, except for the rare cases of intron retention. Therefore, the genes identified by Trinity are sequences generated from the exons of the transcript or group of transcripts corresponding to this gene. Since reads from all samples are combined before running Trinity, it is not possible to know the presence or absence of an identified transcript in a specific sample. This will be done during the RNA quantification step (2.4.5.1). Since Trinity tends to overestimate the number of transcripts present in the sample, quantification can also be used to filter out transcripts with low TPM, as these are more likely to be assembly artefacts.

2.4.4 Peptide identification

2.4.4.1 Considerations for generating a database

Once transcripts have been identified, the proteins they code for must be determined. This is done in PIT using Transdecoder, which looks at all the Open Reading Frames (ORF) possible in a transcript and trains a statistical model to determine which ORFs are the most likely for a transcript. The goal is to generate a sample as small as possible while making sure the proteins that are potentially present in the sample are included in the database, as using a canonical database presents some issues as mentioned in 2.2. Using RNA derived proteins offers a solution to the database specificity problem. First, since the proteins included in the database are derived from the RNA transcripts identified during RNA-Seq, we have evidence at the RNA level that the gene coding for a protein is expressed in the sample, unlike a canonical database which doesn't rely on any evidence for the presence of a protein and includes all known proteins for the species. This ensures the database is no bigger than it needs to be. On the other hand, since the database is derived from RNA-Seq, we may also find transcripts that code for proteins that are not present in the reference proteome, especially in the case of non-model organism, potentially leading to the discovery of novel

proteins. Finally, in UniProt, there is often one protein per gene, although some isoforms are referenced for model organisms. Using an RNA-derived database allows including different isoforms of the same gene in the database, potentially allowing to find evidence at the peptide level for protein isoforms. In PIT, peptide identification and quantification is performed by MaxQuant (Cox and Mann, 2008).

2.4.5 Quantification

Since changes in gene or protein abundance are rarely binary (presence or absence) and will rather result in quantitative change, it is important to be able to measure the relative change of a gene or a protein, for example in response to a treatment.

2.4.5.1 RNA level quantification

We introduced in 1.1.4.1.2 the TPM metric, which is an unbiased way of measuring the relative abundance of a gene, with regard to other genes in the sample. It is therefore useful to determine which genes are the most abundant at the RNA level, and for each transcript assembled, Stringtie2 returns its abundance in TPM. PIT uses DESEQ2 (Love et al., 2014) to then perform differential gene expression between different groups of samples. However, since it uses a negative binomial distribution, it requires integer values as input to its model. Therefore, the raw read counts, which is the number of RNA-Seq reads mapping to the gene, are used to estimate gene abundance. However, since read coverage varies from one sample to another, read counts must be normalised so that all samples have comparable read counts. DESEQ2 is a package used for differential gene expression. Since variations between replicates are always to be expected and one sample cannot be representative in itself of the whole condition it belongs to, it is important to have multiple replicates in each condition, and statistical methods like the one included in DESEQ2 estimate whether there is a significant change in expression of a gene between two conditions. DESEQ2 is able to do so and returns for every gene the log2 fold change between two conditions as well

as a p-value. Since multiple hypothesis testing is applied for the different genes, DESEQ2 also returns adjusted p-values. PIT includes DESEQ2 and feeds it with the raw read counts obtained from Stringtie2 and stores the output from DESEQ2 in JSON format. In case no replicates are available, PIT calculates the log2 fold change for each gene between two conditions, but no p-value can be calculated.

2.4.5.2 Protein level quantification

Once peptide spectrum matches (PSMs) have been identified by Andromeda, which is the peptide search engine shipped with MaxQuant, they are also quantified by Andromeda. This intensity is calculated differently depending on the type of labels attached to them or their absence. Since PIT supports relative peptide quantification, for each peptide identified, comparing intensities for a given peptide gives an indication of its relative abundance in the different samples. However, as with every experiment, the quantity of sample used isn't exactly the same across different channels, intensities for some samples can be overestimated relatively to the other sample, simply because the total amount of proteins in the sample is higher. Therefore, a normalisation step is required for intensities to be comparable between different samples. Here, sample intensities are normalised so that all samples have the same median intensity across all peptides. This is done using the ProteusR package (Gierlinski et al., 2018). While other packages exist allowing the processing of output from peptide identification search engines, such as MSStats (Choi et al., 2014), ProteusR was chosen as it is developed by the same group as MaxQuant. Therefore, there is a better integration of this package with MaxQuant output.

With regard to protein quantification, the approach chosen in PIT differs slightly from traditional pipelines using a canonical database for peptide identification. Indeed, in such pipelines, the peptides used to quantify a protein are usually either unique or razor peptides. This means that a peptide either maps only to a given protein or that it can also map to another protein for which there is no independent evidence and which is therefore discarded

based on Ockham's razor 1.2.3.3. In the context of PIT, this is different, as the database is derived from translation of transcripts that were observed by RNA-Seq. This implies that we have some level of evidence for these alternative proteins, at least at the RNA level. Therefore, Ockham's razor principles as shown in 1.2.3.3 do not apply here, and only unique peptides are used for protein quantification. Another aspect to consider is how to define whether a peptide is unique, i.e. only maps to one protein. Indeed, since the database is derived from translated RNA-Seq transcripts, they need to be grouped by their gene of origin in order to know if they all come from the same gene and will produce the same protein or group of protein isoforms. In the case of a reference guided assembly, the task is facilitated by the ENSEMBL annotations. Since a GTF annotation is required by PIT for referenced guided assembly, it is often possible to assign a transcript found by Stringtie2 to a transcript existing in the ENSEMBL database, with an identifier starting with ENST. Transcript identifiers are associated with a gene identifier (starting with ENSG) corresponding to the gene producing this transcript 2.11. For some transcripts, the coding sequence produced from their translation is also referenced with an identifier starting with CCDS and if a known protein in UniProt is known to match with this CCDS, it is also referenced.

As seen in 2.12, a gene can produce several transcripts and even several proteins. If a peptide maps to several of these proteins, it is still considered as unique as it comes from the same region in the genome. Therefore, PIT groups coding sequences to which a peptide maps to the ENSEMBL gene identifiers they come from. If several gene identifiers are found for a peptide, it is not considered unique. In some cases, especially when a species doesn't have a well annotated genome, Stringtie2 is unable to map a transcript it has identified to an ENSEMBL transcript identifier. In this case, it is given an STRG identifier. Transcripts with overlapping exons form a cluster, which is given a MSTRG identifier by Stringtie2. This cluster is what Stringtie2 infers to be a gene. Such clusters can be artefacts from sequencing or assembly error, or can correspond to new or non annotated genes. Similarly, if a peptide maps to more than one MSTRG identifiers or a combination of

MSTRG and ENSEMBL identifiers, it is not considered unique.

In the case of de novo assembly, determining peptide uniqueness is more complicated. First, we can't rely on an annotation, since we assume there isn't any. Additionally, since de novo assembly doesn't take advantage of any reference genome to guide it, the coverage of assembly tends to be lower than in reference-guided assembly, with many transcripts being only partially assembled. This also means a weaker ability to identify isoforms of a same gene. While it is still possible, and Trinity also generates clusters from overlapping transcripts, the number of such clusters in human is higher than the total number of human genes. This implies that many of the transcripts are either incorrect or could be grouped together as they are actually part of the same gene. Since peptide uniqueness can only be estimated based on the number of Trinity clusters a peptide maps to, this may lead to believing a peptide is not unique while it actually is, and maps to different clusters that are actually corresponding to the same gene but couldn't be clustered together due to insufficient coverage of the assembly. Having fewer unique peptides means being able to quantify fewer proteins, as some will only be covered by peptides that are considered non-unique. For proteins having multiple unique peptides, it still means less confidence in the accuracy of the estimated abundance as statistically, the more unique peptides map to a protein the more accurately we can estimate its abundance. To ensure uniqueness of a peptide, a solution could be to look at the different clusters a peptide maps to, use BLAST (McGinnis and Madden, 2004) to align the sequences and see if they match to the same genes. However, this process should be performed manually and is subject to interpretation, thus it is not implemented in PIT.

Gene: PTEN ENSG00000171862

Description	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
Gene Synonyms	BZS, MHAM, MMAC1, PTEN1, TEP1
Location	Chromosome 10: 87,862,638-87,971,930 forward strand. GRCh38:CM000672.2
About this gene	This gene has 18 transcripts (splice variants), 137 orthologues, 6 paralogues and is associated with 228 phenotypes.
Transcripts	Hide transcript table

Show/hide columns (1 hidden) — Filter

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000371953.8	PTEN-201	8515	403aa	Protein coding	CCDS31238	P60484-1	NM_000314.8	MANE Select, Ensembl Canonical, GENCODE basic, APPRIS P1, TSL:1
ENST00000693560.1	PTEN-211	8701	576aa	Protein coding		-	-	GENCODE basic
ENST00000700021.1	PTEN-212	3255	388aa	Protein coding		-	-	GENCODE basic
ENST00000688308.1	PTEN-209	3117	403aa	Protein coding	CCDS31238	F6KD01	-	GENCODE basic, APPRIS P1
ENST00000472832.3	PTEN-204	1358	344aa	Protein coding		A0A087X033	-	GENCODE basic, TSL:2
ENST00000700029.1	PTEN-218	806	269aa	Protein coding		-	-	CDS 5' and 3' incomplete
ENST00000688922.2	PTEN-210	2818	73aa	Nonsense mediated decay		-	-	-
ENST00000700022.1	PTEN-213	2794	172aa	Nonsense mediated decay		-	-	-
ENST00000686459.1	PTEN-207	2712	190aa	Nonsense mediated decay		-	-	-
ENST00000688158.2	PTEN-208	8405	No protein	Processed transcript		-	-	-
ENST00000487939.1	PTEN-205	477	No protein	Processed transcript		-	-	TSL:3
ENST00000700024.1	PTEN-215	4002	No protein	Retained intron		-	-	-
ENST00000700025.1	PTEN-216	3127	No protein	Retained intron		-	-	-
ENST00000700023.1	PTEN-214	2428	No protein	Retained intron		-	-	-
ENST00000700026.1	PTEN-217	1597	No protein	Retained intron		-	-	-
ENST00000498703.1	PTEN-206	554	No protein	Retained intron		-	-	TSL:2
ENST00000416679.1	PTEN-202	499	No protein	Retained intron		-	-	TSL:3
ENST00000462694.1	PTEN-203	383	No protein	Retained intron		-	-	TSL:2

Figure 2.11: Table of transcript referenced for the PTEN gene in humans on ENSEMBL (https://www.ensembl.org/Homo$_s apiens/Gene/Summary?db = core; g = ENSG00000171862; r = 10 : 87862638 - 87971930$)

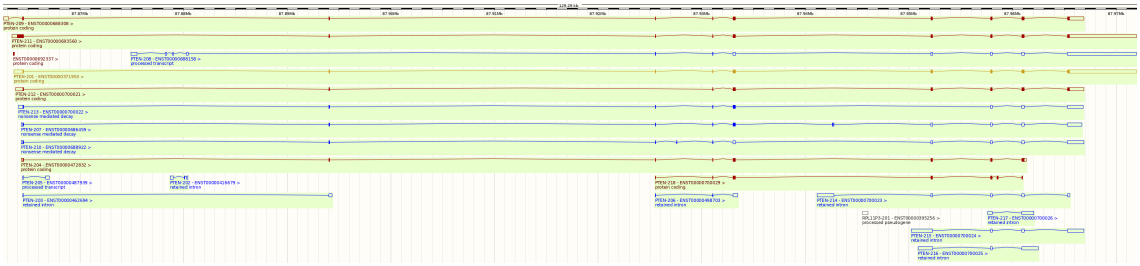

Figure 2.12: Genome browser on ENSEMBL showing gene and transcripts produced in humans for the PTEN gene. Rectangles represent exons and lines represent introns for each transcript. The exons part filled in colour represents fragments of the sequence that are part of the coding sequence. (https://www.ensembl.org/Homo$_s apiens/Gene/Summary?db = core; g = ENSG00000171862; r = 10 : 87862638 - 87971930$).

2.4.5.3 Correlation between RNA and protein abundance

A frequent area of interest in the omics field is the correlation between RNA expression and protein abundance. Indeed, ideally, a strong correlation between RNA expression and protein abundance would allow estimating one from the other, potentially saving time and money on additional experiments. On the other hand, an absence of correlation in some cases can also provide some insights about RNA translation, post translation modifications and transcripts or protein degradation rates. Thus, multiple articles have been published on the subject. The general observation is that there is a significant, yet poor correlation (Maier et al., 2009). Several explanations are put forward. The first one are quantification errors, both at the RNA and protein levels. RNA-Seq and LC-MS/MS are the most commonly used for shotgun transcriptomics and proteomics, however the quantification is only an estimation which can differ from the real value due to experiment inaccuracy and limitations or errors in the algorithms used. Another explanation is translation efficiency (Maier et al., 2009), meaning that some RNA transcripts are translated more than others. Additionally, after translation, proteins have different half-lives and some are degraded faster than others, mostly depending on their function. Indeed, it has been shown that the correlation is stronger for certain categories of genes based on function or localisation in the cell (Gry et al., 2009). Lastly, the lack of correlation can partially be explained by how samples are collected. Indeed, RNA and proteins level of genes constantly change over time and there is also a time gap between the moment RNA is produced and when the corresponding protein is translated. It can therefore be challenging to find the right time gap between the moment the RNA is sampled and the proteins are sampled so that protein levels best reflect RNA levels.

Thanks to the support for RNA and protein quantification in PIT, it can be used to study correlation between these two measurements. We therefore applied PIT to both PTEN and HNRNAPA2B1 datasets to investigate correlation between RNA expression and protein abundance. The RNA expression is calculated as the sum of the TPM of all transcripts for this gene. As far

as protein abundance is concerned, we cannot use the SILAC or TMT labels as they are only used for relative abundance, i.e. calculating a fold change between different samples, but cannot be used for absolute quantification of a protein. We therefore use spectral counting as described in 1.2.3.4.1. We also observe a weak correlation, at 0.38 for the PTEN dataset and 0.32 for the HNRNPA2B1 dataset. Yet, this correlation is significant according to Pearson's test, with a p-value $< 2.2e^{-16}$. This means that transcripts that are more abundant at the RNA level tend to produce more protein, although the weak correlation doesn't allow to accurately predict protein abundance based only on RNA expression. This can be explained by the reasons detailed by (Maier et al., 2009) and (Gry et al., 2009), especially, the fact that spectral counting isn't a very accurate estimator for protein abundance.

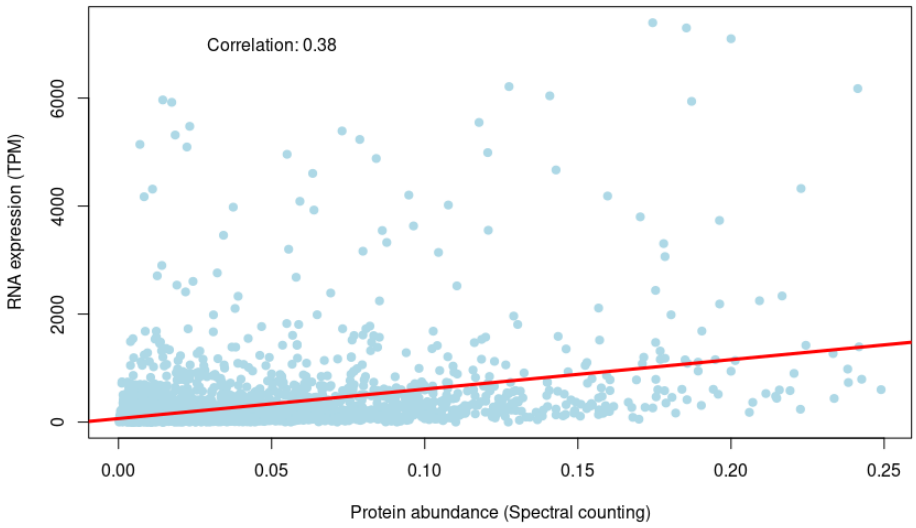

Figure 2.13: Correlation between RNA expression and protein abundance for the PTEN dataset

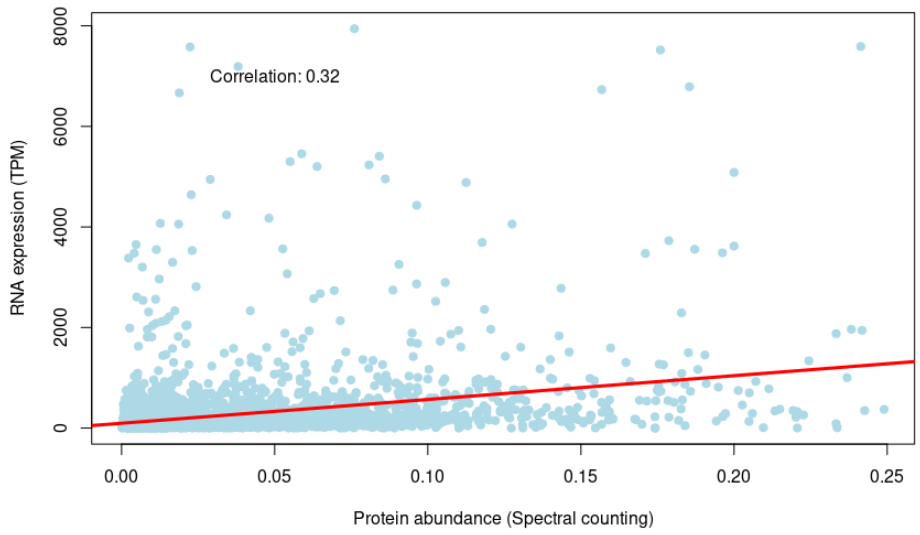

Figure 2.14: Correlation between RNA expression and protein abundance for the HNRNPA2B1 dataset

Another approach is to compare the log2 fold change between RNA and protein between two conditions. Here, we observe a stronger Pearson's correlation, at 0.54 for the PTEN dataset and 0.49 for the HNRNPA2B1 dataset. These numbers are close to those obtained by (Liu et al., 2017) who did a similar analysis but at a transcript level. Fold change correlation analysis can reveal interesting insight. For example, while the slope on the HNRNPA2B1 is close to one, meaning that if we double the amount of RNA expressed, we double the amount of protein translated, this slope is closer to 0.5 for the PTEN dataset. This could potentially indicate that PTEN is having an impact on translation of protein life cycle and degradation. Previous research has indeed already determined that knockdown of PTEN in PTEN positive cells increased ribosome biogenesis (Li et al., 2014). Additionally, what can be more interesting than the correlation itself is the lack of correlation for specific genes. Indeed, we can see that for the vast majority of genes, they are clustered close to the regression line. Genes that are further shown a strong difference of variation at the RNA and protein level, potentially making them

interesting to study.

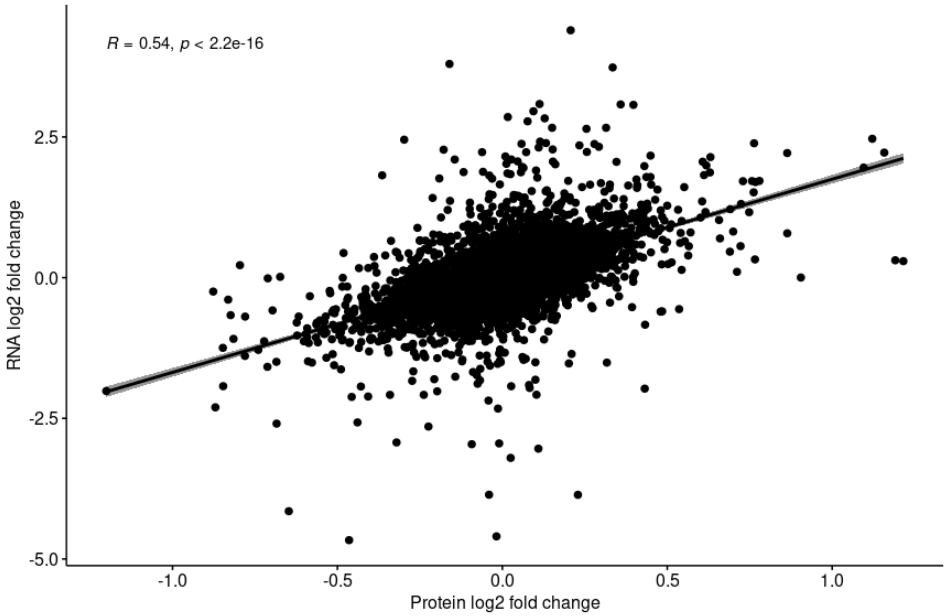

Figure 2.15: Correlation between RNA expression and protein abundance log2 fold change for the PTEN dataset

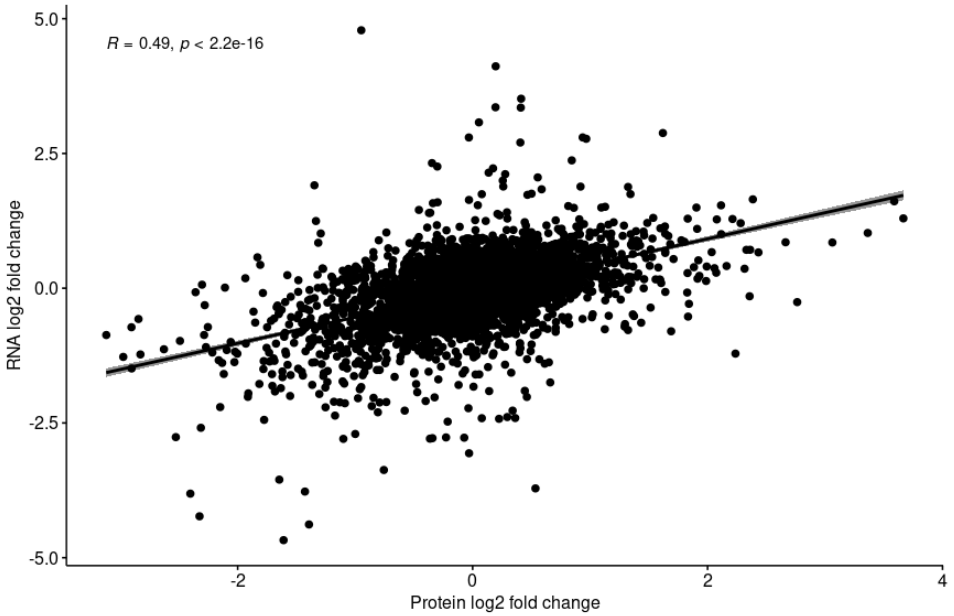

Figure 2.16: Correlation between RNA expression and protein abundance log2 fold change for the HNRNPA2B1 dataset

2.4.6 Mutations

2.4.6.1 The issue of mutation evidence at the protein level

Finding evidence for mutations at the protein level remains a challenge for mass spectrometry. Indeed, the traditional method of doing peptide identification using a canonical database doesn't allow finding any mutation. This is because the sequences in the database are the canonical sequences for the species, and it only contains one sequence per protein (or protein isoform). Because of how peptide spectrum matching algorithms work, a spectrum can only be assigned to a peptide that is present in the database. Therefore, in order to be detected through mass spectrometry, the peptides from mutations we might expect need to be added to the database. There are two main ways to do this.

The first one is to consider known mutations from public databases from databases such as UniProt (Bateman et al., 2017) or dbSNP (Sherry et al., 2001). These mutations are inserted into their respective sequences and the new peptides are generated in silico. This approach is used for example by

(Alfaro et al., 2017). The benefit of this approach is that it relies only on data that is publicly available. However, these databases can contain a considerable number of mutations, resulting in an important increase in peptide database size. Yet, the bigger the database, the more difficult it is to identify peptides since the quality threshold to maintain a 1% FDR is higher since there are more risks of spectra mismatching peptides by chance. Therefore, putting too many mutations in the database can result in fewer peptides being identified overall. On the other hand, some mutation present in the sample but not in the database could still not be detected. This will be particularly the case for non-model organisms that don't have as many mutations recorded in databases.
The other approach is to use RNA-Seq to find mutations at the RNA level, build transcript sequences including these mutations, translate them and find new peptides resulting from these mutations. This has the advantage of being sample specific and therefore only include mutations for which there is some evidence at the RNA level, instead of including all mutations recorded in samples that can be very different. Therefore, the impact on the database size can be smaller. On the other hand, since it is sample specific, this allows to find mutations that are not recorded in any public database. The limitation is that it is sensitive to false positives from RNA-Seq. Indeed, some variations that are considered to be mutations can actually be sequencing or assembly errors. Furthermore, this can include mutations of transcripts in low abundance or that are not actually translated intro proteins.

2.4.6.2 Detecting mutations at the RNA level

To detect mutations at the RNA level, PIT uses GATK (McKenna et al., 2010). The process of mapping reads to the reference genome (such as done with STAR2 (Dobin et al., 2012)) is fault-tolerant, which means a read can be assigned to a region of the genome even if the sequences don't match perfectly. GATK then looks at these alignment disagreements. If for a given position in the reference genome, most reads aligned to this region don't have the same nucleotide at this position as the reference region, this could indicate a

mutation. There are also factors to consider such as the read quality, whether reads can also map to other locations in the genome or allele frequency. For example, in humans, which have diploid cells, one allele can have the same nucleotide at a given locus as the reference genome whereas the other allele may have a different nucleotide. In this case, we would expect about half of the reads mapping to this region to have the reference nucleotide and the other half to have the alternative nucleotide. GATK considers all these factors and returns a VCF file containing all mutations found in the sample, after quality filtering.

2.4.6.3 Finding mutation evidence at the peptide level

VCF is a tabular file format where each row represents a variation and contains information about its location in the genome, as well as optional additional information such as quality or allele frequency. For each variation, PIT finds all the transcripts that have an exon that overlap with the variation location. If the mutation doesn't fall within a CDS that has been predicted by Transdecoder, this transcript is discarded for this mutation, as it will not impact the protein sequence. If the mutation falls within the CDS of the transcript, the part of the transcript that codes for the CDS is extracted, the mutation is inserted into it and this sequence is translated, resulting in a new CDS (unless the mutation is silent, in which case both CDS will be the same and the peptide is therefore not duplicated in the protein database). Then, both old and new CDS are digested in silico and the set of peptides resulting from the new CDS and those that are not present in the old CDS are the peptides that result from this mutation and can therefore be considered as evidence if detected by mass spectrometry.

This list of new peptides resulting from variations are then added to the database fasta file that is used for peptide identification, with a specific identifier that allows PIT to trace which mutation produced this peptide. After peptide identification, if it is found that a spectrum maps uniquely to a peptide coming from a mutation, this acts as evidence (Within the 1% FDR) that this variation is indeed present and produces an alternative

protein compared to the reference genome, as otherwise this peptide could not have been detected.

We used PIT to find peptide evidence compared to the reference genome on the PTEN datasets. A total of 528,758 mutations were identified by GATK in at least one of the eight samples 2.17. Of these, 287 were found to have unique peptide evidence. While this number seems low, several factors can explain it. Firstly, many of these mutations do not take place in the protein coding part of the RNA transcript, thus these changes cannot be seen at a protein level. Secondly, many of these mutations happening in the CDS are silent. Indeed, by looking at the codon table 1.9, we observe that in many cases, a change in the third nucleotide of a codon still produces the same amino acid. Since most mutations in proteins being detrimental, there is a drive from evolution to eliminate such mutations, whereas silent mutations are not under such pressure as they will not affect protein function. As such, we find from the PTEN dataset that out of 528,758 mutations found at the RNA level, there were only 116,698 that resulted in an amino acid change. Then we need to consider the low coverage of mass spectrometry, with about 10% of the sequence of an identified protein being covered by peptides on average for this experiment. Furthermore, even if Transdecoder predicts a CDS from a transcript, this doesn't necessarily mean this CDS will necessarily be translated. And even if it is, it may still not be detected by mass spectrometry, as it only tends to detect peptides from the most abundant proteins. This could be particularly the case of mutation that are in only one allele. Assuming no difference in translation efficiency and protein half-life caused by the mutation, in humans, a peptide arising from a mutation coming from one allele would be only half as abundant as if the mutation was on both alleles. This would make the detection of this peptide more difficult by mass spectrometry. In some cases, a mutation may cause a CDS to no longer be translated or for the resulting protein to be quickly degraded. This is notably the case with mutations producing an early stop codon, resulting in a CDS much shorter than what it normally is. Finally, adding these extra peptides makes the peptide database bigger. In this instance, the number of amino acids present in the database without mutations was 19,223,503. With

mutations, an additional 13,113,760 amino acids are added, meaning amino acids from peptides generated from mutations represent 40% of the amino acid content of the database. This will result in a higher score threshold for PSM in order to respect the 1% FDR and therefore cause fewer peptides to be found overall. To increase the number of mutations for which we find

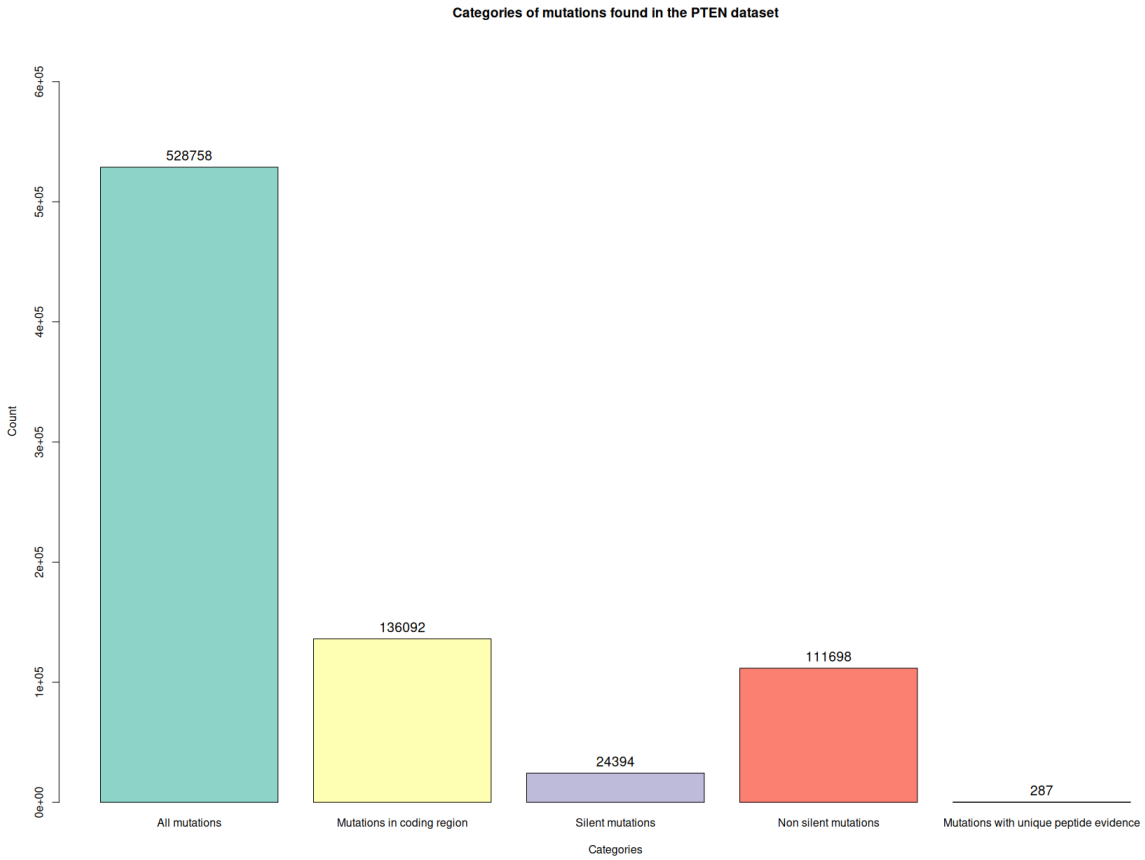

Figure 2.17: Number of mutations and mutated peptides identified or predicted for the PTEN dataset

peptide evidence, several heuristics, which have not yet been implemented in PIT, could be adopted. The goal would be to reduce as much as possible the number of mutated peptides added to the database without loosing peptides that could be detected. This means removed peptides from mutations that we think may be errors or are less likely to be seen in a protein. We could, for example, only keep mutations found in multiple samples when us-

ing replicates. We could also remove transcripts of very low abundance (for example less than 1 TPM) and that do not map to any known ENSEMBL transcript, as they could be assembly artefacts that are not present in the sample. In case users want to study genes or pathways of interest, we could only add to the database peptides from mutations arising in these genes. Finally, the parameters in GATK could be optimised to be more strict while doing the quality filtering of mutations found. Since a bigger database would also reduce the number of non mutated peptides identified, we recommend to only add mutations in case there is a particular interest in them. PIT allows skipping the mutation analysis if users wish to do so.

2.4.7 Alternative splicing

2.4.7.1 Identification and quantification at RNA level

In PIT, alternative splicing events are identified and quantified using SUPPA2 (Trincado et al., 2018). SUPPA2 takes as input a GFF file from transcriptome assembly containing a list of all transcripts identified as well as all their exons and their coordinates on the genome. It then looks for transcripts sharing some exons (or retained introns) and uses this information to extract splicing events, as defined in 2.18. For each event, a Percentage Spliced In (PSI) value is calculated (Schafer et al., 2015). PSI represents the relative abundance of transcripts including an exon over the abundance of all transcripts related to this splicing event (including or excluding this exon). As such, it gives an estimation of how much an exon is included. Since SUPPA2 is able to work on different samples and different conditions, it calculates a PSI for each event in each sample and then calculate a ΔPSI which indicates how much the PSI changes between two conditions for a given splicing event. It then uses statistical testing to calculate a p-value indicating if the difference in ΔPSI is significant.

Both HNRNPA2B1 and PTEN datasets follow a similar distribution of event types, with exon skipping events being most predominant, which we know to be the case for humans (2.19), (2.20).

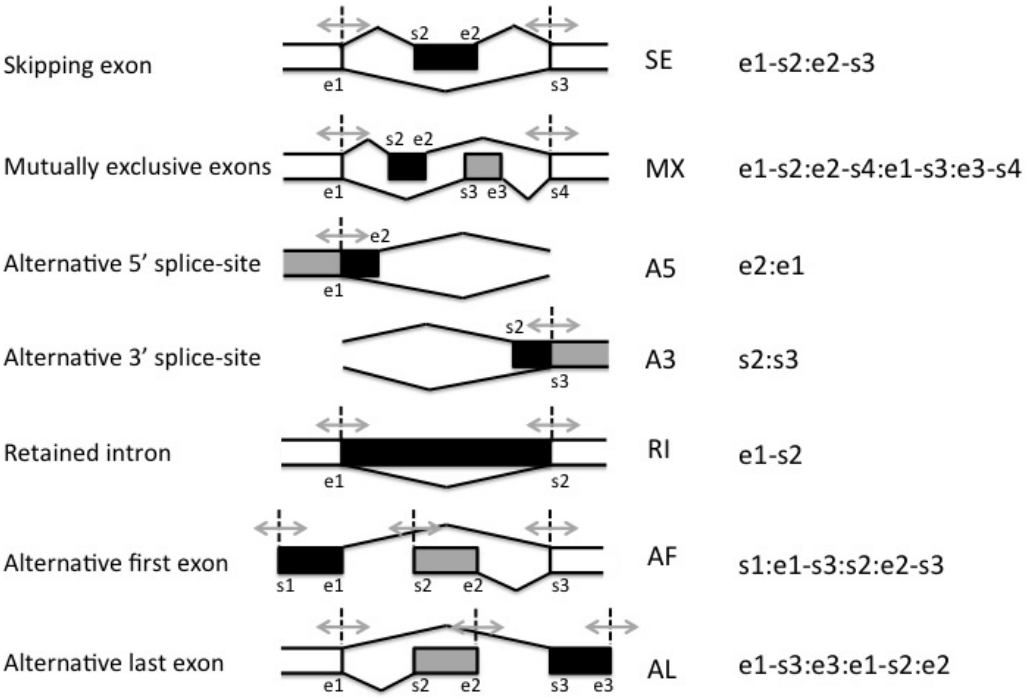

Figure 2.18: Types of alternative splicing events that can be identified by SUPPA2

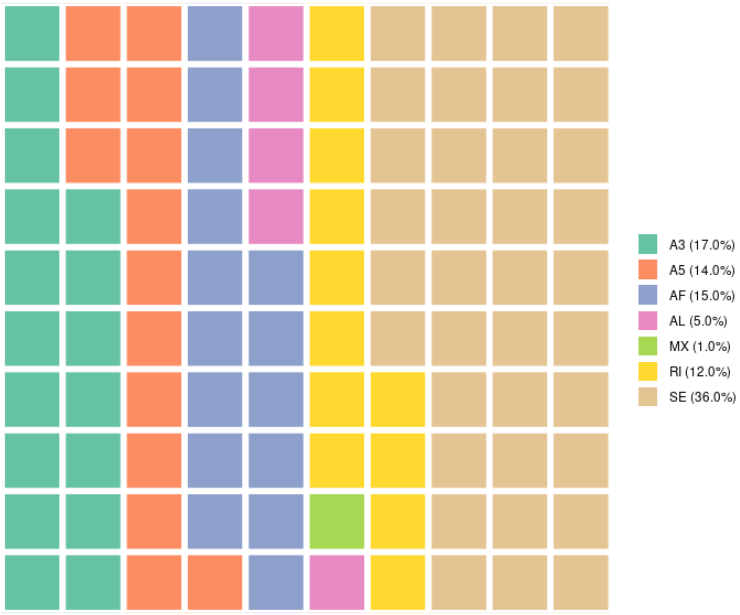

Figure 2.19: Distribution of splicing event types in the HNRNPA2B1 dataset

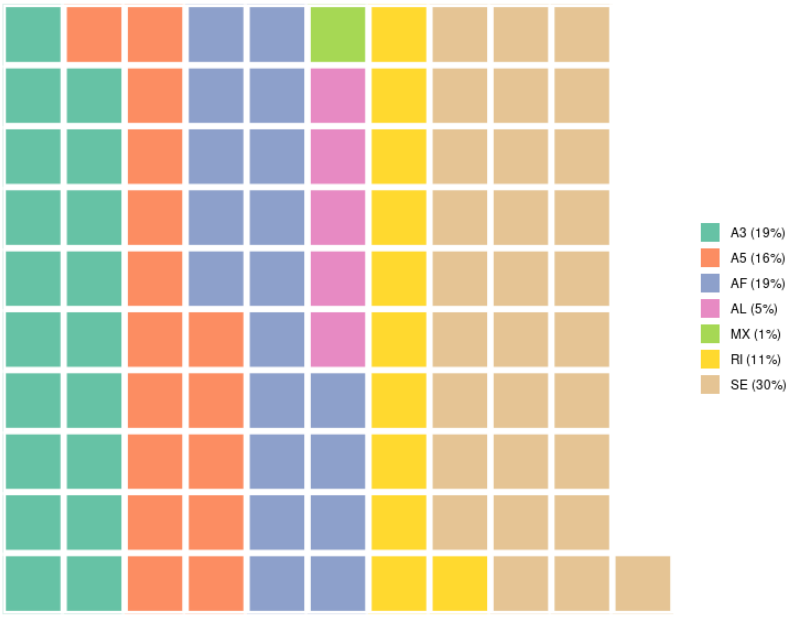

Figure 2.20: Distribution of splicing event types in the PTEN dataset

2.4.7.2 Identification and quantification at protein level

However, splicing events identified at the RNA level are not necessarily reflected at the protein level. For example, some isoforms may have early stop codons and may thus be eliminated by the nonsense-mediated decay (NMC) pathway. Some transcripts may also have splicing events in their five prime untranslated region, which may affect translation efficiency (Hinnebusch et al., 2016).

Therefore, being able to identify and quantify alternative splicing events at the protein level can also provide interesting insights. To this end, LC-MS/MS can be used. Indeed, if peptides overlap with exons involved in the splicing event, it is possible to quantify it, as seen in 2.21. To calculate the differential inclusion of an exon at the protein level between two conditions, we need to look at the intensities of the overlapping peptide(s) in each condition. However, this ratio can be biased by the fact that the overall protein levels are also different between condition. Thus, if a peptide covering a splicing event is less abundant, it could be that the overall protein abundance is lower in this sample, not that this specific isoform is less abundant relatively to the other isoforms. To account for this, we need additional peptides from the protein that do not cover the splicing. These peptides are then used to calculate the overall differential protein abundance between condition. The ratio obtained is then used to normalise the intensities seen on the peptides covering the splicing event, so that the differences in intensities only reflect the effect of differential alternative splicing without the bias or overall differential protein abundance. On the HNRNPA2B1 dataset, a total of 44,767 splicing events were found, 291 of which were significant (<0.05 adjusted p-value for the ΔPSI). Peptides were found to overlap with 5,506 of these alternative splicing events. This number is in line with what we would expect to see as it means that in this case 12.3% of splicing events identified at the RNA found had peptide evidence, and we know that this mass spectrometry experiment has an average of 10-15% coverage of protein sequence. Unlike mutations, since splicing events involve several exons, multiple peptides can usually overlap with them, making them easier to detect. On the

other hand, differential alternative splicing events between two conditions are usually small ($\Delta PSI < 20$), which can make it difficult to see a clear difference at the protein level.

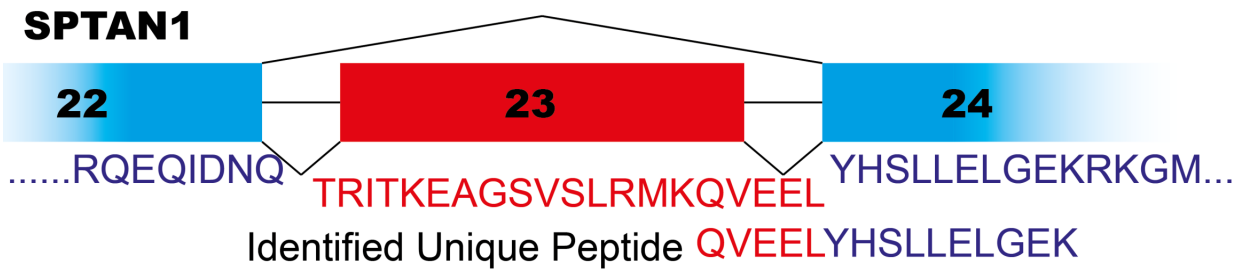

Figure 2.21: Representation of exons 22, 23 and 24 of SPTAN4 with the translated amino acid sequence mapping to these regions. We identified through LC-MS/MS the peptide QVEELYHSLLELGEK which overlap with both exon 23 and 254.

2.4.8 Functional annotation

In the case of the reference-guided arm of PIT, since some transcripts and proteins can be mapped to known sequences, some extra layers of annotations and analysis can be added, which can provide useful insight into samples examined. I present in this section some public annotation tools that have been integrated into PIT and are relevant in this context of integrating transcriptomics and proteomics data.

2.4.8.1 Protein domains: PFAM

PFAM (Mistry et al., 2021) is a database created in 1995 and is managed by the European Bioinformatics Institute (EBI). It groups proteins it references into families based on functions and homology. In addition, it annotates proteins with domains. Domains are parts of a protein that fulfil a specific function, such as binding to another protein or molecule. Thus, for a given sequence, PFAM provides the start and end of each domain on the sequence as well as its function. For unknown sequences, PFAM uses hidden Markov models that compare the sequence to other sequences in the database in order

to determine its family and domains. This information is particularly useful in the context of PIT, since an alteration within a protein domain is more likely to have an impact on protein function than if it happens outside the domain. Thus, mutation, alternative splicing events, differential post translation modifications that happen within a PFAM domain are of particular interest. Hence, PIT integrates PFAM data information in order to allow users to filter results depending on whether an alteration happens within a PFAM domain or not.

2.4.8.2 Gene Ontology

The Gene Ontology Consortium (Ashburner et al., 2000) is a group that manages a controlled vocabulary of terms to describe and classify genes. Each GO (Gene Ontology) term contains an identifier, a name, an ontology and a definition, for example:

Identifier	GO:0048731
Name	system development
Ontology	biological_process
Definition	The process whose specific outcome is the progression of an organismal system over time, from its formation to the mature structure. A system is a regularly interacting or interdependent group of organs or tissues that work together to carry out a given biological process.

Three ontologies are defined (Mistry et al., 2021)(Gen):

1. Molecular function: Molecular-level activities performed by gene products

2. Cellular component: The locations relative to cellular structures in which a gene product performs a function

3. Biological process: The larger processes accomplished by multiple molecular activities

GO terms can be connected through an "is a" relationship. As seen in 2.22, GO terms can be represented as a tree, with a GO term having one or many parents which are more general. For example, according to 2.22, we

can say that hexose biosynthetic process is a hexose metabolic process and is also a monosaccharide biosynthetic process.

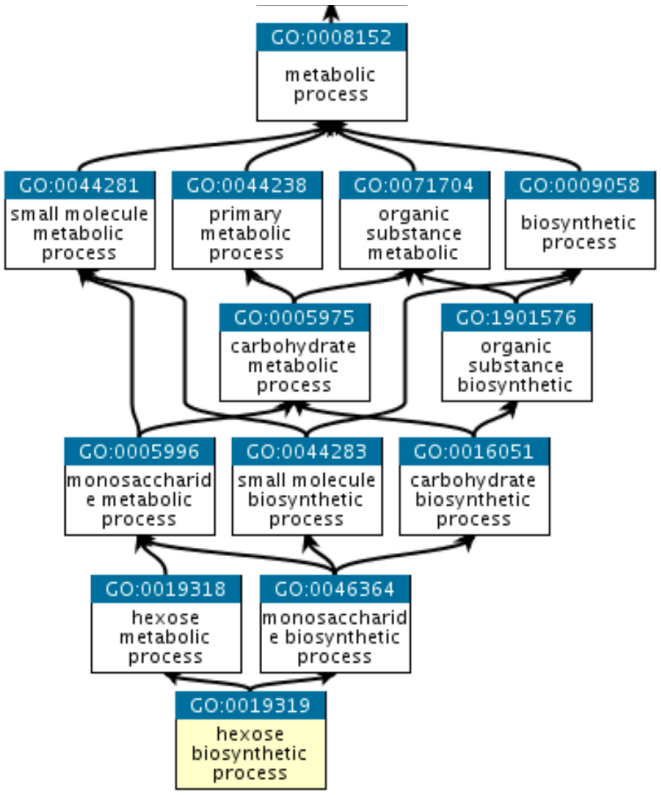

Figure 2.22: Example of GO terms graph. Each node represents a GO term, which is a specialisation of its parent GO term. (Gen)

Using a controlled vocabulary allows grouping genes by GO terms and perform computational analysis such as Gene Set Enrichment Analysis (Subramanian et al., 2005), which would not be possible if gene descriptions were inconsistent. Public datasets such as Ensembl and Uniprot include known GO terms into their genes and protein records, making them easy to access. Thus, PIT includes support for GO terms in order to annotate genes identified, perform filtering as well as Gene Set Enrichment Analysis.

2.4.8.3 KEGG Pathways

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a project started in 1995 at the university of Kyoto under the Human Genome Program (Kane-

hisa and Goto, 2000). It is a database that contains information about biological pathways in various species. The pathways include genes and how they interact with each other, as well as with chemical compounds or small molecules. It also includes pathways for some diseases, such as cancer, to represent known interactions between genes in such diseases 2.23. The pathways are represented using a determined nomenclature, such as the edges between two genes, that are used to represent the type of interaction between those genes. These pathways prove helpful for systems biology. For example, they can be used to see what genes could be affected by the perturbation of another gene or show how a pathway is affected by a change like we did in 4.13. PIT also includes support for KEGG pathways for filtering, Gene Set Enrichment Analysis and visualising changes in a pathway, in order to for example compare what pathways are enriched at the RNA and protein levels.

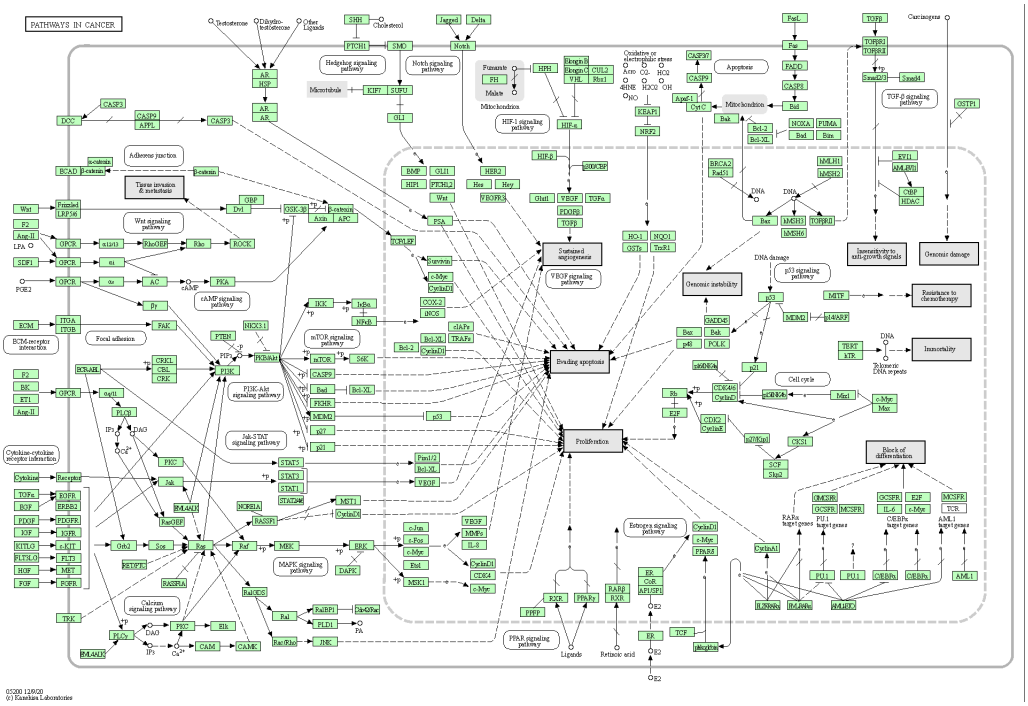

Figure 2.23: KEGG pathway for cancer. This represents multiple genes and pathway involved in cancer through several routes such as proliferation or evading apoptosis. Green rectangles represent genes, white rectangles other pathways, grey rectangles the way the genes contribute to cancer survival and white circles represent other molecules.

2.4.9 Comparing de novo and reference guided assembly

To compare the results between de novo and reference guided assembly, we used the HNRNPA2B1 dataset. The goal was to compare the transcripts obtained from assembly and the impact this has on the peptides identified by LC-MS/MS.

First, it can be observed that the de novo approach produces more transcripts 2.24. As a result, de novo transcripts tend to be less abundant, with a median TPM of 0.38 vs 0.97 for reference guided assembly. De novo transcripts are also shorter with a median length of 389 nucleotides versus 1421 nucleotides for reference guided assembly.

As a result, despite the fact that de novo assembly produces more transcripts than reference guided assembly, these transcripts are shorter. Thus, the peptide database generated from the de novo transcripts contains 35 973 720 nucleotides versus 31 540 343 for the reference guided assembly.

In total, 58 763 peptides were identified for de novo, versus 59 230 for reference guided. Yet, both approaches largely identify the same peptides as there is an overlap of 57 277 peptides between the two (2.25). 1 953 peptides were identified only for reference guided. These can be explained by the fact that de novo assembly fails to assemble certain regions of the transcriptome, therefore the corresponding peptides are not added to the database. However, 1 486 were identified only in de novo. There can be two main explanations for this. The first one could be that some regions are missing in the reference genome and therefore can only be discovered through de novo assembly. However, since this was done on Homo sapiens, there should only be a limited amount of such peptides. The other explanation is that these peptides map to transcripts that are artefacts from de novo assembly and are therefore not actually there.

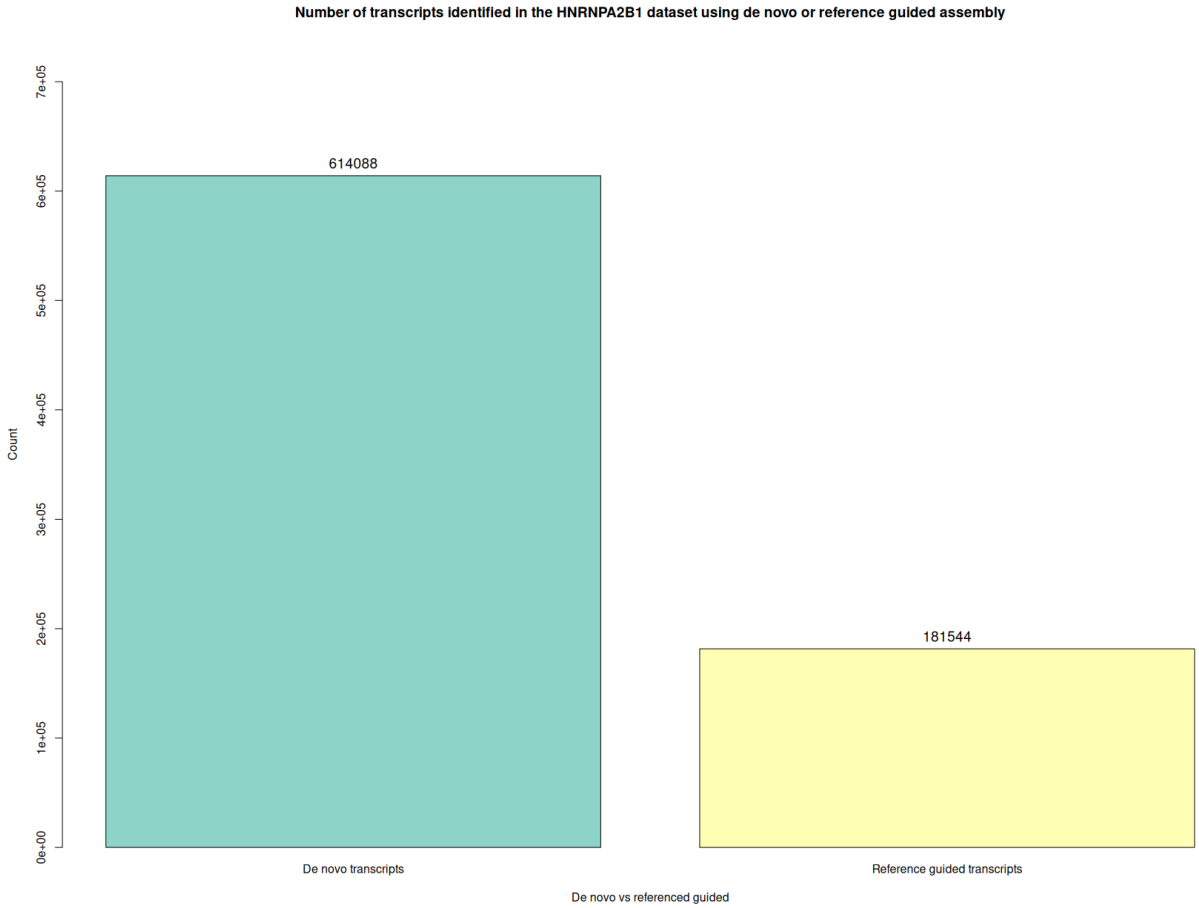

Figure 2.24: Total number of transcripts identified by de novo and reference guided assembly using the HNRNPA2B1 dataset. No transcripts were filtered based on TPM.

As previously said, multiple database optimisation strategies are possible, based on the objectives. This can include removing transcripts below a certain TPM, only keeping transcripts that are known to be protein coding, etc...

It is important to note that depending on the species, these comparisons may vary greatly. For example, in an orphan organism with a very incomplete reference genome, there may be much fewer transcripts and therefore more peptides may be identified through a de novo approach than through a reference guided approach.

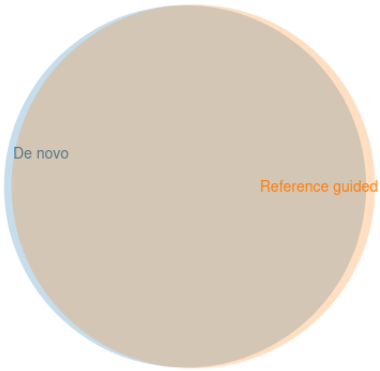

Figure 2.25: Venn diagram of the overlap between peptides identified from the de novo and reference guided assembly.

2.5 Conclusion

In this chapter, we have introduced PIT and shown how it can combine data coming from RNA-Seq, LC-MS/MS and public databases in order to unveil more insights into researchers' data. Indeed, PIT offers a variety of analyses that integrate transcriptomics, proteomics and public data. These analyses can be used for multiple aims such as observing gene and protein expression, alternative splicing, mutations, post translation modification or identifying novel proteins in non-model organisms. The support for quantitative analyses as well as the comparison between samples of different conditions makes multi-omics analyses possible for a wider range of researchers as it fits their experimental design.

For this tool to become widely adopted by the community, several aspects must be considered. The first one is how easy it is to use the pipeline. In this aspect, fitting the image within a docker container makes it much easier to use, as users don't need to install all the dependencies themselves and can run it on any platform. With regard to performance, PIT includes some computationally expensive tasks, such as transcriptome assembly and peptide identification. These tasks, both expensive in time and memory, imply that users must have the computational resources to run PIT. Indeed, even on a 16-core desktop computer with 64 GB of RAM, running PIT can take hours or up to several days depending on the numbers of samples analysed.

While the analyses provided by PIT can already prove useful in a wide variety of contexts, more features could be added to fit particular experiments. PIT was developed in a modular way, with analyses related to gene expression, alternative splicing, mutations which can be performed independently of each other.

Another important aspect is that for analyses supported by PIT, it is important to provide users with parameters to let them define how to run them. As we have seen, in most situation, such as the choice of what to include in the mass spectrometry database, the decisions taken can strongly affect the results. For example, for users looking for peptide evidence for mutation on a specific protein, there is no point adding mutated peptides

for all genes in the database, whereas in other contexts there might. It is therefore important to leave such options open to users, and even potentially run multiple instances of PIT with different parameter sets.

However, PIT produces a large amount of data, spread across multiple files. As such, analysis of the output is difficult, especially for someone without the programming knowledge needed to extract the relevant information. Therefore, it was decided to develop PITgui, a graphical user interface that parses the output produced by PIT and displays the results in a graphical and interactive manner.

Chapter 3

PITgui

3.1 Introduction and rationale for developing a graphical user interface

Over the past decades, the performance of LC-MS/MS and NGS has continuously increased, whereas the costs have decreased considerably. At the same time, computers have also become powerful and cheaper, with Moore's law, stating that the number of transistors on a chip doubles every two years, holding true so far. Hence, it is our belief that the main bottleneck for a much bigger emergence of multi-omics projects is no longer on the hardware side, but rather on the software side.

A recurrent issue in the area of bioinformatics software development is accessibility. A bioinformatics pipeline will typically take some input and output some files. These files can be in text format such as csv or json and therefore be directly read by users. Other files are in binary format and can only be explored after processing by another tool. This is for example the case for bam files that contains read alignments to the reference genome in binary format. Either way, extracting results from such files often proves difficult to do manually due to the quantity of information produced. The more complex the pipeline, the more difficult this process usually is, as the files will be bigger and some information about a specific element may be split into several files. It is thus clear that manually extracting results is not desirable. A more

relevant alternative is to write scripts for post-processing and interpretation of the results. Many packages, such as DESEQ2 (Love et al., 2014) offer functions that can be called from R to interpret the results. The output is also a data frame that can then be interpreted in R as users wish. While this approach offers more flexibility and can make it easier to extract relevant information, it requires programming experience in order to write the code required for handling the data. In addition, users need to understand how to use the package, what format the output is and how to use other packages, for example for generating charts out of the data for graphical representation. This implies that such analyses cannot be performed by everyone, including many of the people who need this data, such as wet lab scientists who don't necessarily have coding experience. This means wet lab biologists often rely on a bioinformatician for data analysis, which can be a serious bottleneck in the project, depending on their availability.

In order to maximise accessibility, especially among wet lab scientists, a solution can be to develop a graphical user interface (GUI). This is a software that is either installed locally or available via a web browser that allows users to visualise their data. This approach is more user-friendly as it doesn't require any coding skill to use. In addition, this provides users with quick access to the data they need, since software can be developed to implement features users will commonly need. For example, MaxQuant (Cox and Mann, 2008) can be run in the command line to perform peptide identification and quantification. Yet, the Max Planck Institute, which develops MaxQuant, also provides Perseus (Tyanova et al., 2016), a graphical user interface developed to be able to import the output from MaxQuant and let users easily manipulate it in a graphical way.

Considering the complexity of the PIT output, which includes dozens of gigabytes of data split across dozens of files, we realised that developing a GUI was also necessary. We wanted wet lab scientists to be able to use this software in order to explorer their data themselves without having to rely on a bioinformatician, at least for the most common analyses. For more specific analyses, the PIT output files can also be used, but this would require programming knowledge.

In this chapter, we will introduce PITgui, a graphical user interface that was developed with the aim of offering users a way to visualise results from PIT without requiring programming experience and in a way that combines transcriptomics and proteomics data.

3.2 Code availability

Source code for PITgui can be found at https://github.com/bezzlab/pitsuite. This repository also contains additional screenshot as well as an introductory video presenting PITgui.

3.3 Software architecture

3.3.1 Languages and libraries

When it comes to developing graphical user interfaces, there are two main approaches. The first one is to develop a web application hosted on a server and available through the web browser. This has the benefit of not requiring any installation on the local machine and being platform independent. Additionally, for future updates, developers can update the web application without imposing any responsibilities on the users. However, in the context of PIT, developing a web application has serious drawbacks. The first one is performance. Web applications run using JavaScript, which is an interpreted language, to provide interactive functionality on the front end. The consequence is that web applications usually cannot reach the performance and scalability of applications installed locally and programmed using compiled languages such as C++ or Java. Since PIT generates a considerable amount of data, performance is critical so that users can have a smooth experience. Similarly, a web application implies that the data is stored on a remote server, which means that when users want to access them, they would have to upload and download the data, which would take much longer than if the data is already available locally. Additionally, centralising biological data on a remote server may cause issue of data confidentiality and safety, especially in the

context of patient data. Finally, a web server means maintaining a server up and running over the years. This server would need enormous capacity, both in terms of storage, to store the data from all the projects analysed, but also in terms of computing power to do all the processing to query and process the data before sending them to the front end. Therefore, in the context of an academic project, this approach can be dangerous in case funding is no longer available in the future.

The other approach is to develop software that is installed locally on the computer, where the data also resides. This solution is better in terms of performance and is able to deal more easily with important quantities of data. Additionally, it is not dependent on a remote server, which means it can always be used in the future. It is therefore this approach we chose, to develop PITgui, a graphical user interface to visualise results from PIT. This software is written in Java, a compiled object-oriented programming language that runs on a Java Virtual Machine (JVM). This JVM is responsible for converting the Java code into instructions that can be executed by the local machine. This means that Java is platform independent, and one program can thus run on either Windows, Linux or macOS. This is unlike C++, another popular programming language which compiles the code directly into instructions that can be understood by the operating system and therefore requires different compilation for every system, making it more difficult to release and install. Another advantage is that many bioinformatics tools are implemented in Java, for example with PeptideShaker (Vaudel et al., 2015). This means that bioinformatics libraries are already implemented and that we can integrate them into our own software. For example, samtools (Li et al., 2009) has developed HTSJDK, a Java library for reading and manipulating BAM and SAM files. We can also cite BioJava (Lafita et al., 2019), a Java library that implements many bioinformatics tools such as multiple sequences alignment and 3D protein structure visualisation.

The third argument for using Java to develop PITgui is that Java has a well established and maintained graphical library, called JavaFX (Jav). This library allows the efficient development of graphical user interfaces by implementing classes that represent graphical components such as buttons,

lists, tables as well as handling user events.
In addition, PITgui also contains some R scripts that are called from PITgui in order to generate plots that are then imported into PITgui.

3.3.2 Architecture

JavaFX is based on the Model View Controller (MVC) architecture pattern 3.1. It involves splitting the code into three main groups of components. The view is where the graphical components of the application are defined. To do so, JavaFX uses FXML, a declarative language similar to HTML, where the graphical components are added to the view by adding tags representing classes implemented by JavaFX. Within the tags, properties can be defined, such as some styles for the graphical components or some event triggers. The models are classes that represent entities that are used in the program. These classes have properties and functions that allow operations to be performed on them. For example, we could have a class called Gene, which has properties such as the chromosome it comes from, the start and end position on the chromosome and the DNA sequence of the gene. It may also implement functions, such as a function to translate the DNA. Finally, the controllers are classes that control the lifecycle of the application, that define what must be displayed on the interface, how to react to user interaction and how to interpret data. Separating components into functional units such as done by the MVC architecture makes the code easier to expend and maintain and is often used for designing graphical user interfaces.

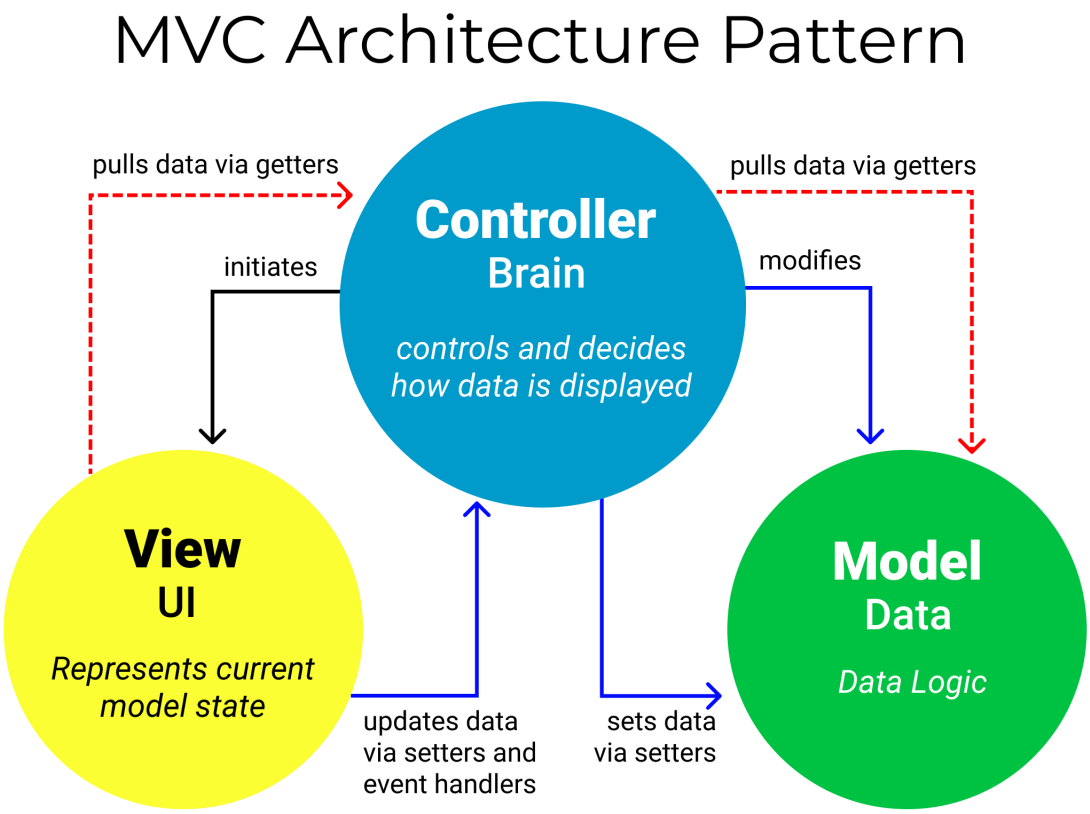

Figure 3.1: Model View Controller (MVC) architecture pattern (The)

3.4 Using PITgui to visualise results from PIT

3.4.1 Importing PIT data

When running PIT for a project, the output is written to a folder chosen by the user. PITgui can then access this folder and automatically import the project and process the files. To do so, PITgui reads the configuration file to get an understanding of the experimental design and to know how to parse files. It then stores the processed data in a NitriteDB database. NitriteDB (Diz) is a NOSQL embedded document store written in Java, so it is possible to integrate it as a library in the Java application. NitriteDB generates a database in a single file on the filesystem, which makes it lightweight and removes the need to install a daemon-based database such as MongoDB. NitriteDB also allows querying and indexing of the database to make it quick to retrieve the data required, allowing low latency for PITgui users.

When launching PITgui, the menu appears. Users can import new projects by selecting the output folder from a PIT analysis. Once a project has been imported, it will appear in the list of imported folder so that the user can reopen the project later on.

Users can also generate a PIT configuration file through PITgui if they do not wish to write it manually 2.4.1.3. Users then see a page with fields asking for information about the RNA-Seq arm of PIT 3.2 and another page for the LC/MS-MS arm of PIT 3.3. Once the configuration file has been written, users can also launch a PIT analysis directly from PITgui. In this case, it will use the configuration file that has been generated to run the Docker image of PIT on the local machine.

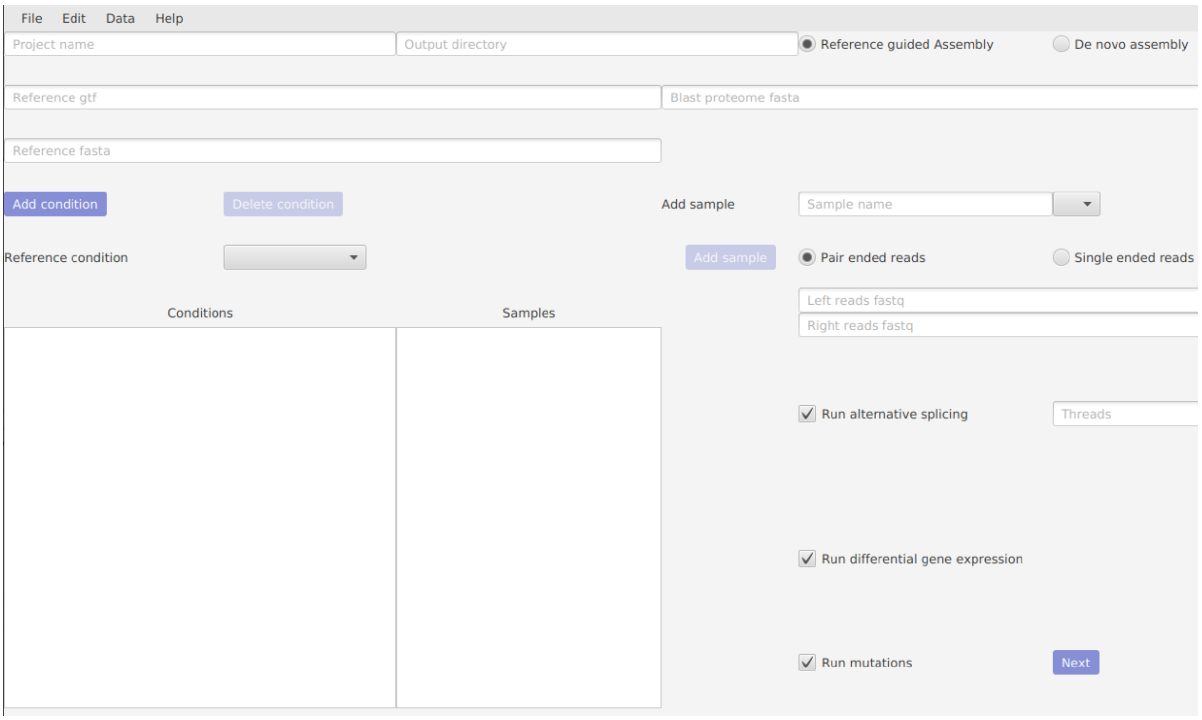

Figure 3.2: Generation of the configuration file in PITgui with the RNA-Seq window. It requires information such as the paths to the fastq files to perform assembly or the path to the reference genome and annotation in the case of the reference-guided PIT. This is also where the different conditions and within them the different samples are defined.

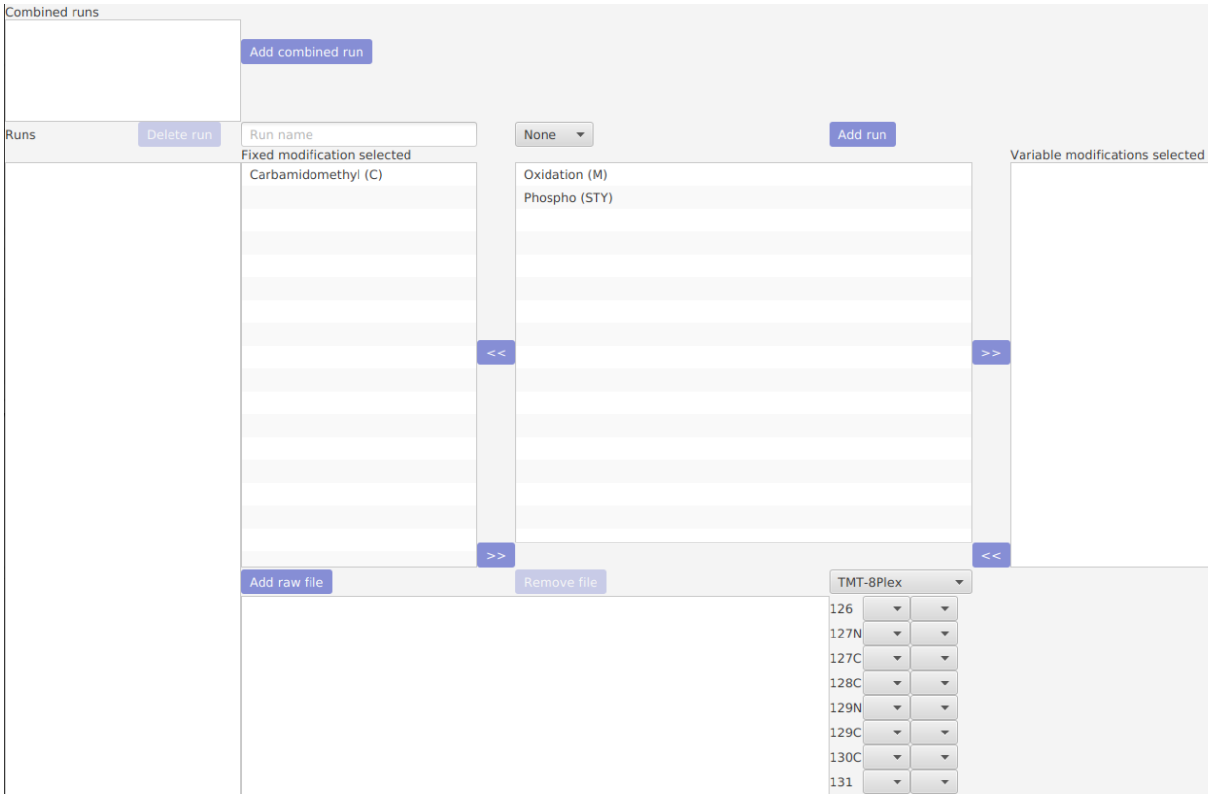

Figure 3.3: Generation of the configuration file in PITgui with the RNA-Seq window. It requires information such as the path to the mass spectrometry raw files and information about how to run MaxQuant, such as labelling information, post translation modifications.

3.4.2 Visualising gene and protein quantification

Once a project has been opened, the PITgui interface is split into several tabs, each corresponding to a type of analysis performed by PIT. By default, when loading a project, the differential gene and protein expression tab is displayed 3.4. This tab shows a tabular summary of genes identified by PIT (3.4.B). For each gene, the log2 fold change is calculated at the RNA level. If multiple replicates are available in each condition, a p-value is also calculated. If mass spectrometry has been performed, a column in the table displays the log2 fold change for the corresponding protein and the p-value if multiple replicates are available. The content of the table can be dynamically changed based on the filters in the table header (3.4.B). When clicking on a row in the table, information about this gene is plotted (3.4 C, D and E). On the right (3.4 F, G and H), scatter plots show how gene and protein expression compare between two conditions. These scatter plots are updated based on the content of the table to only display genes that satisfy the filtering criteria set by the users (3.4. A). PIT and PITgui can work with as many conditions as were defined in the configuration file, but differential gene and protein expression can only be performed pairwise. PIT therefore makes a list of all the pairs of conditions possible based on the conditions entered by the user and performs differential gene and protein expression on all of them. The dropdown menu in 3.4.A then allows the user to select which pair of conditions to compare. 3.4 F, G and H can be replaced by other information by clicking on the tabs on the top right, for example to show results overlaid on to KEGG pathways (3.5. A similar tab exists to filter the table based on GO terms. Finally, Gene Set Enrichment analysis (GSEA) can be performed directly from PITgui, both at the RNA and protein levels 3.6.

Figure 3.4: Differential gene and protein expression tab in PITgui. A. Filtering options for the table. This allows filtering of the table based on log2 fold change or p-value at the RNA or protein level, to select a gene by name, or to only show genes with peptide evidence at the protein level. B. Differential gene and protein expression table. Each gene identified by PIT is represented as a row. The values show are calculated by DESEQ2 (Love et al., 2014) for the RNA level and ProteusR (Gierlinski et al., 2018) for the protein level. C. Differential gene expression for the genes selected in the table. Each bar represents a condition and each dot represents a sample in this condition. 95% confidence intervals are also shown. D. Differential protein expression for the proteins corresponding to the genes selected in the table. Colours represent a condition and each bar represents a sample. Each dot represents the normalised intensity of a peptide found to map uniquely to this protein. E. Normalised intensities for each peptide found to map uniquely to the selected protein. F. Volcano plot showing the distribution of differential gene expression between the two conditions selected. The x-axis represents the log2 fold change and the y-axis represents the -log10 p-value. G. Volcano plot showing the distribution of differential protein expression between the two conditions selected. H. Scatter plot showing the correlation between log2 fold change at the RNA level and log2 fold change at the protein level between the two conditions selected.

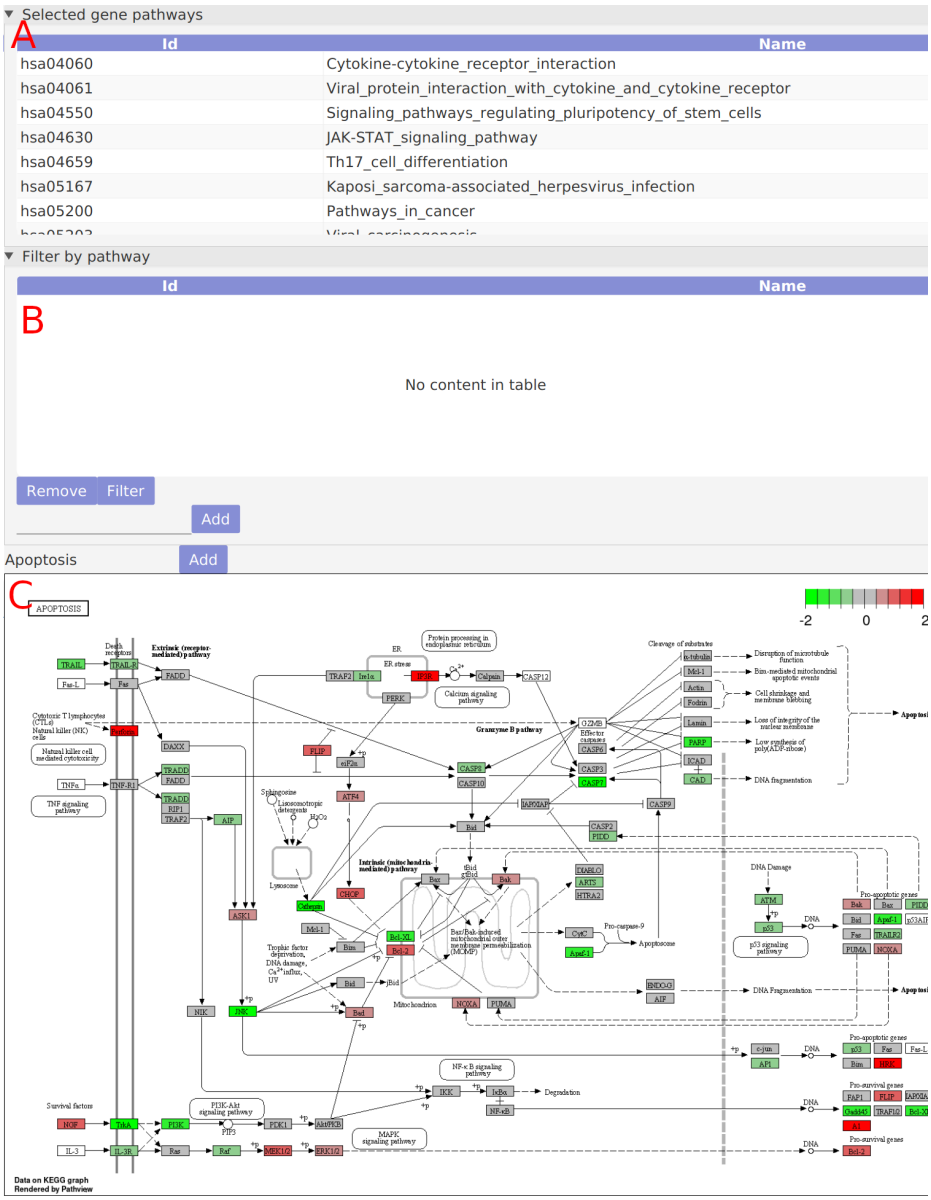

Figure 3.5: KEGG tab in PITgui. A. list of KEGG pathways associated with the gene selected in the table. B. KEGG pathways used to filter the content of the table. C. For a pathway chosen by the user, PITgui can display this pathway while colouring its genes according to the differential gene or protein expression calculated by PIT. This is done using the Pathview R package (Luo and Brouwer, 2013)

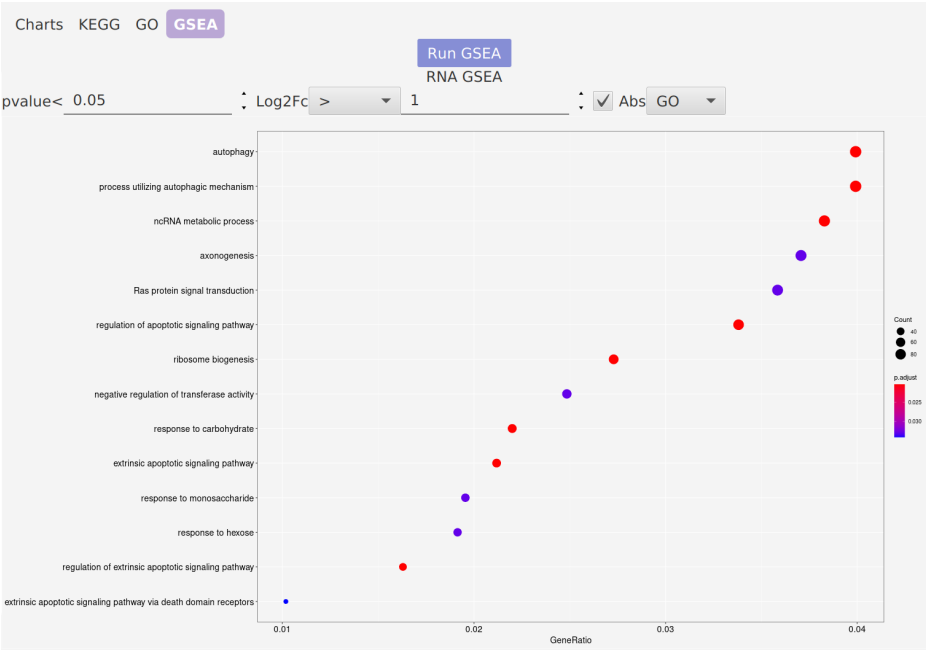

Figure 3.6: Gene Set Enrichment Analysis displayed in PITgui. The enrichment can be performed either on GO terms or KEGG pathways and at the RNA level as well as at the protein level. It is also possible to filter the genes depending on their log2 fold change or p-value. Enrichment is calculated using the ClusterProfiler R package. (Wu et al., 2021)

3.4.3 Bespoke genome browser for PIT

Gene browsers are tools that allow to visualise genomic elements and annotations along a section of the genome. Multiple genome browsers exist, some embedded in a web browser such as JBrowse (Buels et al., 2016). Others are software that can be downloaded and used locally, such as IGV (Thorvaldsdóttir et al., 2013). For PITgui, we decided to implement a new gene browser from scratch, with the help of Daniel Torres González. Building our own gene browser allows us to make it as customisable as we want, which is an important point in a multi-omics where we want to integrate a large variety of information. Additionally, while it would have been possible to integrate a Javascript based gene browser in PITgui, building a new one natively for PITgui offers multiples advantages. Firstly, it is faster. Indeed, integrating a Javascript plugin into a Java application requires using a webview, which has low performance. Considering the amount of data that has to be handled, this would have resulted in very noticeable lags and loading times for users. In addition, it would have made it more difficult to maintain and further develop the code as it would be dependent on two languages (Java and Javascript) and the communication between the webview and the rest of the applications would have been difficult, for example to react to user events. The gene browser is accessible via a dedicated tab to it and by inputting the name of the gene we want to visualise. It can also be accessed by double-clicking a row in a genes table, such as the one in the differential gene and protein expression tab.

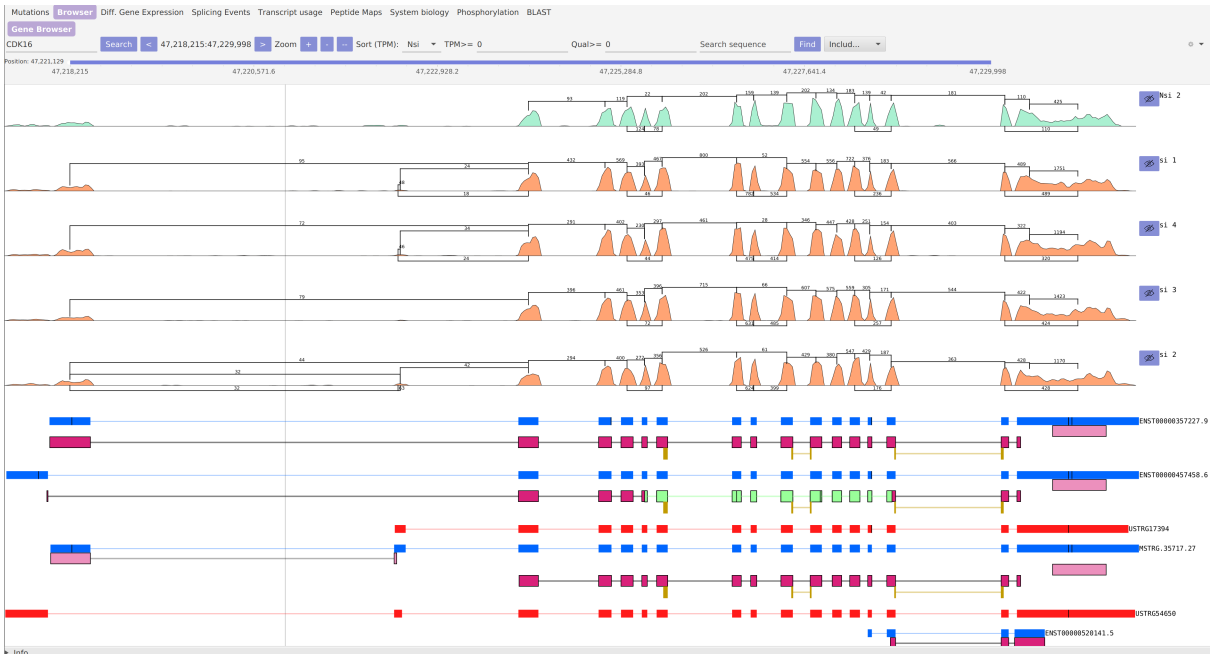

Figure 3.7: PITgui gene browser view of the CDK16 gene with transcripts and proteins identified by PIT

The top view represents RNA-Seq depth in each sample. Higher values mean that more reads were mapped to this region on the gene, hence they align with exons. This also displays a Sashimi plot with the number of reads overlapping two exons. This information can be used to identify alternative splicing events. Below, the different transcripts identified by PIT are represented in blue or red, with the exons as rectangles and the introns as lines. In purple are the coding sequences that were predicted by Transdecoder to be produced by the transcript. Some coding sequences have part of their sequence displayed in light green. In such cases, a PFAM domain has been found and is therefore represented on the sequence. Hovering the mouse over it displays more information about this PFAM domain. Gold rectangles represent peptides that were identified through LC-MS/MS.

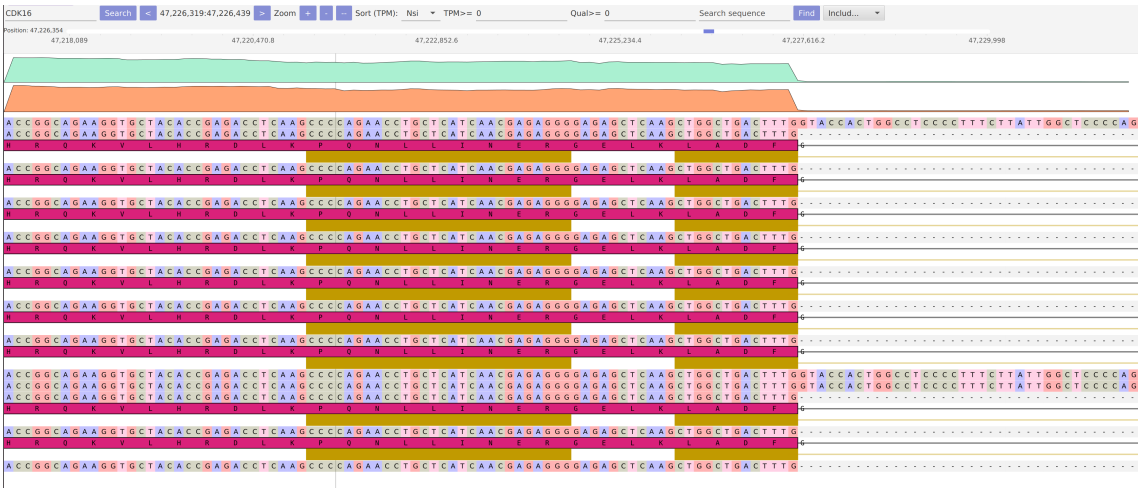

Figure 3.8: Zoomed view of the gene browser view for the CDK16 gene. Each line of nucleotides represents a transcript identified by PIT, with under them in red the translated ORF predicted to be produced by the transcript. When a peptide is identified by mass spectrometry for an ORF, it is displayed as a gold rectangle. Green and oranges areas on top represent RNA-Seq read coverage in two different samples.

The gene browser is interactive and lets users zoom in and out of certain regions. If users zoom enough, they can see the nucleotide and amino acid sequences for this region (3.9).

3.4.4 Mutations

Another feature included in PITgui is the visualisation of mutations identified by PIT. This tab is important as it allows to quickly find relevant mutation based on criteria such as the type of mutation, the gene affected, how it affects the protein sequence, and visualise where the mutation is located in the gene browser.

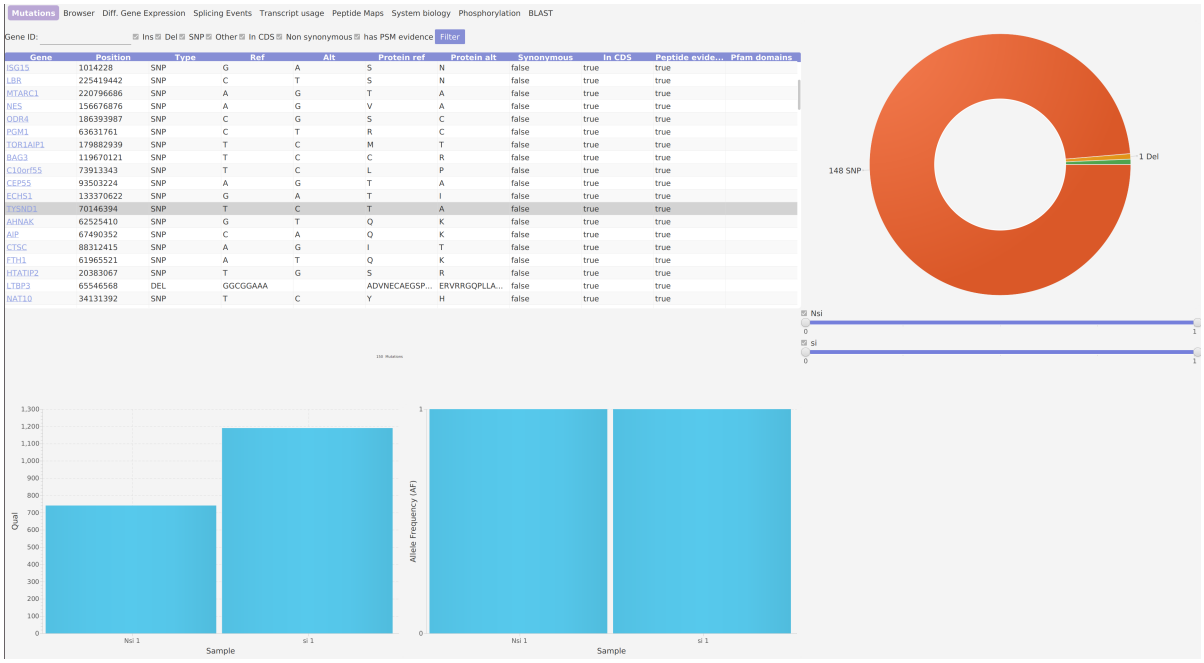

Figure 3.9: Mutations tab in PITgui. It displays a tab containing a table of the mutations found, with filters allowing the user to select a specific gene, the type of mutation (SNP, insertion, deletion), whether the mutation affects the protein sequence and if peptide have been identified providing evidence for this mutation at the protein level. For each condition, a double slider is also included in order to set minimum and maximum values to the number of replicates that must contain this mutation with the given condition.

A double click on a row shows the gene browser zoomed on where the mutation is.

Figure 3.10: Mutation shown in the gene browser. The nucleotide affected is displayed in a darker colour.

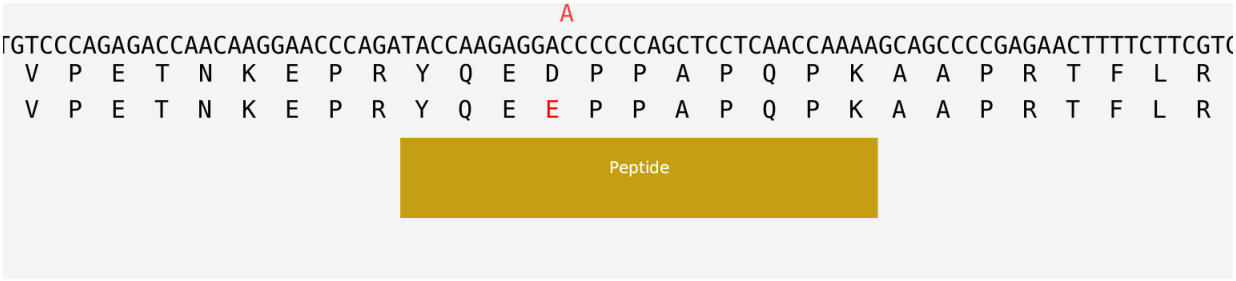

Figure 3.11: Details of a mutation selected in the gene browser. The red nucleotide represent the mutation found in the sample. The RNA sequences represents the canonical sequence for the transcript, with its corresponding amino acid sequence under it. Below is the alternative amino acid sequence we obtain if we consider the mutation identified. The gold rectangle represents the peptide that was identified after taking the mutation into account.

After clicking on a mutation in the gene browser, the nucleotide sequence appears, along with the corresponding protein sequence. If the user selects the mutation, the alternative protein sequence is shown, with the affected amino acid(s) shown in red. If peptide evidence exists, i.e. a spectrum only matches the theoretical peptide generated by this mutation, this peptide is shown as a gold rectangle (3.11).

3.4.5 Alternative splicing

The alternative splicing tab allows visualising alternative splicing events identified by PIT. It is useful in order to find relevant in order to find relevant splicing events according to some criteria such as gene or dPSI and be able to visualise the events in a graphical way in the gene browser. Similarly to the gene and protein expression tab, it includes a table listing the splicing events found, as well as a filter to restrict the number of events displayed in the table. It is also possible to get the GO terms and KEGG pathways for the gene of the selected event. When clicking on an event, a barchart appears, showing the PSI in each condition, as well as a representation of the exons involved in the splicing event (3.12).

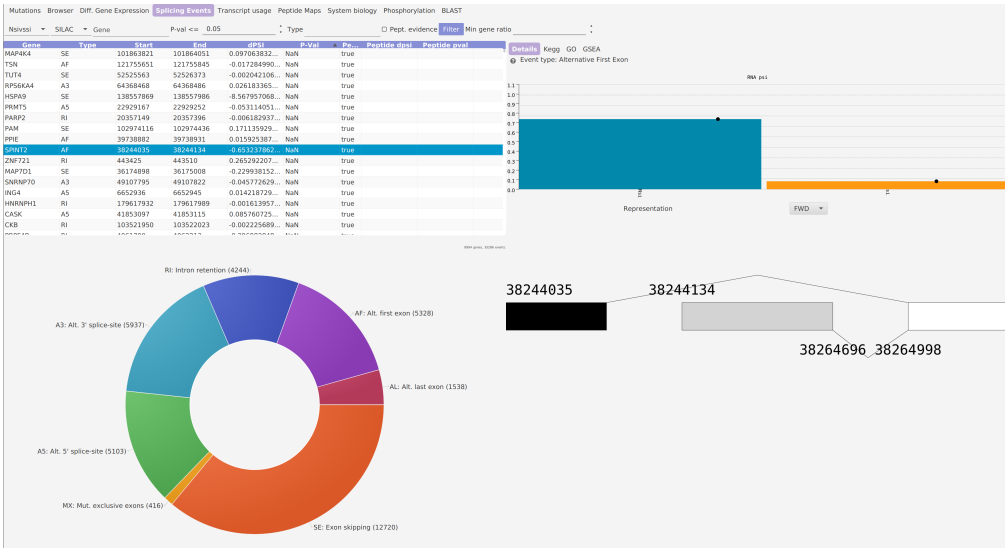

Figure 3.12: Alternative splicing tab in PITgui. The table lists all splicing events found, with their type, gene, coordinates, dPSI and p-value. The donut chart represent the types of all splicing events found in the sample. For a selected event, the barchart represent the PSI in each sample and the carton offers a representation of the splicing event.

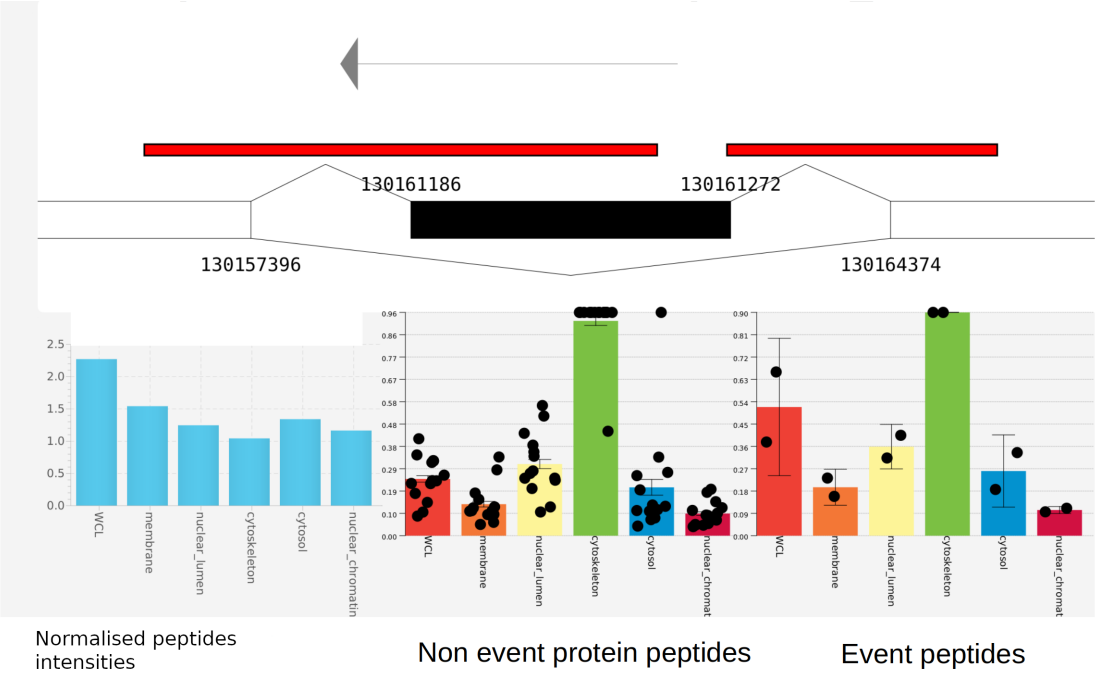

Figure 3.13: Protein view of the alternative splicing tab in PITgui.

On 3.13, the event peptides chart show the relative abundance of peptides that overlap with the splicing event. The non-event protein peptides chart shows the relative abundance of peptide found to uniquely map to this protein, but that do not overlap with the splicing event. These values are used to normalise the event peptides by differences of overall protein abundance. The result is the chart on the left, showing differences in exon inclusion at the protein level between the different conditions.

3.4.6 BLAST

For both reference guided and de novo assembly, users can provide an RNA or protein fasta that will be used to align against the identified transcripts of proteins using BLAST 3.14.

Here, the list on the left shows all proteins identified by PIT. When clicking on one of them, the table on the right is filled with all the hits from the database according to BLAST. Clicking on one row show the alignment between the two sequences. Peptides identified from mass spectrometry are displayed in yellow.

This feature is particularly helpful when working on non-model organisms where a reference is not available. BLAST can be used to align against homologous proteins. Then in PITgui, differences between the two proteins can easily be seen. If a peptide is found on one of these different regions, this provides evidence for this protein at the protein level in addition to the RNA level, which can then be used to construct the species' proteome.

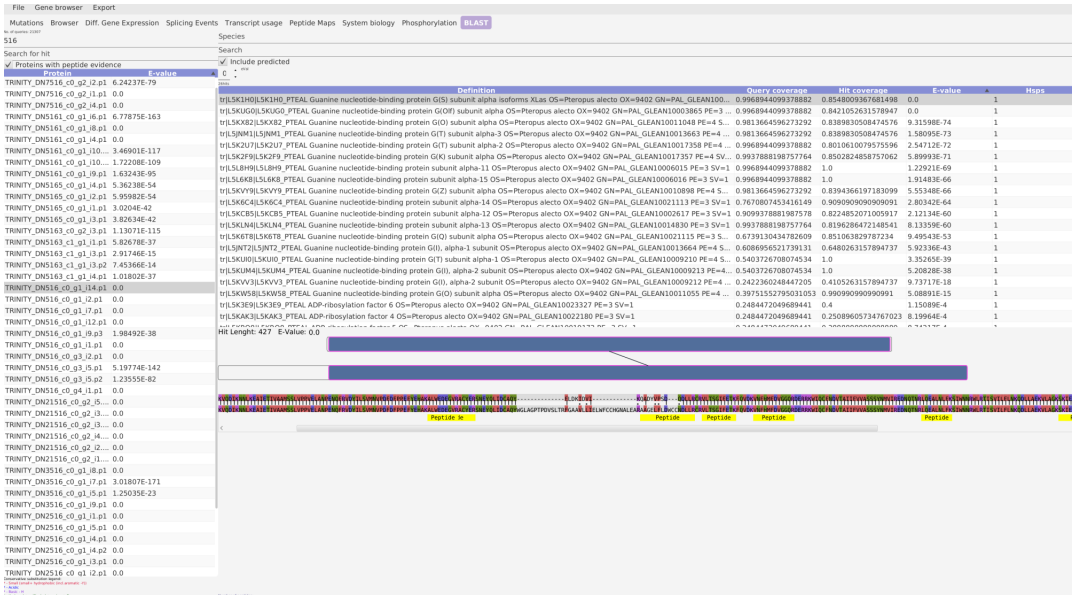

Figure 3.14: BLAST tab in PITgui.

3.5 Discussion

The need for PITgui became apparent after observing that analysing the output of PIT text files was not feasible for most researchers, due to the size of the files and the need to map data between different files. Additionally, the nature of omics data makes it well suited to graphical representation, for example when it comes to aligning sequences or visualising exons on a gene.

While PITgui was initially developed as a way to test and debug PIT's code to see if alignment coordinates were correct, it quickly became a necessary tool for my research and was used across multiple projects, in particular for the work described in chapter 4.

PITgui is also the result of the collaboration with multiple researchers at Queen Mary University of London and particularly the Barts Cancer Institute, who have used PITgui across various projects and have provided valuable user feedback on the interface design and user experience. This collaboration has led to improvements in the user interface, in bug fixes, as well as adding new analyses that were asked for.

Building a graphical user interface is a time-consuming process. Java is a low level language compared to Python. While this additional complexity is necessary in order to have performance good enough to make the software run smoothly and organise the code properly, it makes the codebase more verbose. Furthermore, developing graphical user interfaces is always a complex project as additional considerations come into play such as User Experience (UX) design, anticipating all possible user actions, etc. This also implies that a considerable part of the project must be dedicated to bug fixing and unit testing to ensure stability of the software.

Since this process is more closely related to a software engineering project rather than research, I only built a prototype of PITgui to help me with my research. Further work should be taken to further develop and maintain the software to make it available for the wider public. This process is a continuous one, as even the most developed software, such as MaxQuant, is still regularly updated even though it is relatively mature and benefits from a large development team.

Several ideas for further development can be identified. The first priority would obviously be thorough testing and fixing of identified bugs. This step is important as stability is one of the main criteria for software adoption by the community. A second step would be the ability to connect PITgui to a cloud based platform, as running PIT demands substantial computational resources, which may not be available to everyone. The idea would be to use a cloud platform such as Amazon Web Services (AWS) to run PIT through its docker image. Users would generate the configuration file through PITgui, which would then upload the data to the cloud, run the analysis and push the output back into PITgui once done. The benefits are that it removes the need for computational resources and since the data would be stored in the cloud, it would make it easier to share the data across collaborators. While some implementation of this feature has been done, it wasn't fully completed.

The last area of improvement would be of the features themselves. As explained, PITgui was built in a modular approach, with a list of features that can be used independently of each others. Therefore, maintaining PIT would also mean implementing additional plugins for supporting other types of analysis or even building a software development kit (SDK) to let users build their own plugins that fit their needs. As we have seen, analyses in PITgui often require some level of customisation that will affect the results presented to users. Therefore, within a feature, more work can be done to let users customise their analysis.

Finally, as this software suite is meant to be used by the wider community, some documentation work must be done for it to be adopted by other researchers. This implies writing a tutorial on how to make the best use of PIT and PITgui, and supporting the community, for example through Github issues, running training sessions, etc. This part fell out of the scope of this PhD which was focused on developing a prototype to facilitate research on biological process, but will need to be carried out for to be released to the public.

3.6 Conclusion

In this chapter, we introduced PITgui, a graphical user interface for visualising the output produced by PIT. Due to the size of the files generated and their format, it is difficult to extract the relevant information, especially without writing additional code. PITgui solves this issue by importing directly the PIT output and offering graphical elements to interact with the data, such as filters and tables. In addition, features such as the gene browser are particularly helpful for visualising sequence alignments, which would a tedious task to perform manually from the sequences in text format, especially since it needs to integrate different elements such as transcripts, coding sequences, peptides, read coverage and PFAM domains. Thus, PITgui allows PIT to become usable by a wider range of researchers, by considerably facilitating the interpretation of results.

Chapter 4

Studying the impact of HNRNPA2B1 in prostate cancer

4.1 Acknowledgments

The work presented below is based on (Foster et al., 2022) which I co-authored. Experimental validation and writing of the paper was done by Dr. John Foster. Computational analysis was done by myself. I extend my thanks to Mosammat A Labiba, Chinedu A Anene, Jacqui Stockley, Celine Philippe, Matteo Cereda, Kevin Rouault-Pierre, Hing Leung, Conrad Bessant for their help reviewing this paper. This project was managed under the supervision of Dr. Prabhakar Rajan.

On top of this, some additional computational unpublished work and results are introduced.

4.2 Human cancer

Cancer is a disease consisting of abnormal cell growths, which can potentially spread to different parts of the body. In 2020, it is estimated that 19.3 million new cancer cases were detected worldwide and that 10 millions

deaths occurred due to cancer (Sung et al., 2021). In the United States, cancer is the second cause of death after heart diseases (Ahmad and Anderson, 2021). More than a hundred types of cancer exist (Wha) and usually depend on the location they appear in the body. The most prevalent types of cancer worldwide in 2020 are breast (11.7%), followed by lung (11.4%), colorectal (10.0 %), prostate (7.3%), and stomach (5.6%) cancers, although the distribution differs according to sex. While breast cancer occurs almost exclusively in females, and represents 24.5% of all cancer types for them, the most prevalent cancer for males are lungs (14.3%) and prostate (14.1%). The incidence rate of cancer has been growing over the years, although this could be partly due to ageing of the world population as well as better detection capabilities, both thought technological improvements and better awareness in the population with some countries offering regular screenings to detect potential cancers after a certain age.

Over the years, multiple factors have been identified that can influence the apparition of cancer. Heredity can increase risks of developing a cancer, which is why doctors examine the family history of their patient to look for previous occurrences. A study by Paul Lichtenstein et al. on 44,788 pairs of twins revealed significant associations between heredity and cancer risks, with heredity explaining 42% of the risks of prostate cancer, 35% for colorectal cancer and 27% for breast cancer (Lichtenstein et al., 2009). Other causes are environmental, in particular influenced by our lifestyle. For example, tobacco is estimated to be responsible for 87% of lung cancer deaths, and 30% of all cancer deaths (Furrukh, 2013). Other environmental and lifestyle factors also influence cancer such as diet, in particular with the consumption of ultra-processed foods (Fiolet et al., 2018), certain viruses (Hudnall, 2006), exposure to ultraviolet radiation, certain chemicals and others.

Cancer is caused by mutations that affect the gene (and sometimes protein) sequence, affecting its expression or function. These mutations can be inherited from our ancestry, can be caused by DNA damage by environmental factors such as ultraviolet radiations, or can be the result of random copy errors during cell division. The body normally possesses multiples genes, called tumour suppressors, to prevent cancer from developing or spreading

and which offer several levels of protection. Genes such as BRCA1, BRCA2, and TP53 are genes involved in DNA repair, which are able to repair somatic mutations appearing randomly in our body. Other genes such as RB1 regulates cell division and proliferation. An unwanted cell can also commit apoptosis, resulting in its own death before it can proliferate through uncontrolled cell division (Lowe and Lin, 2000). The immune system is also sometimes able to identify and kill cancerous cells (Janssen et al., 2017).

However, tumours also possess defence mechanisms in order to survive and grow in the body, escaping the body's defences and hijacking its resources. These capabilities are grouped into 6 categories called the hallmarks of cancer (Hanahan and Weinberg, 2011).

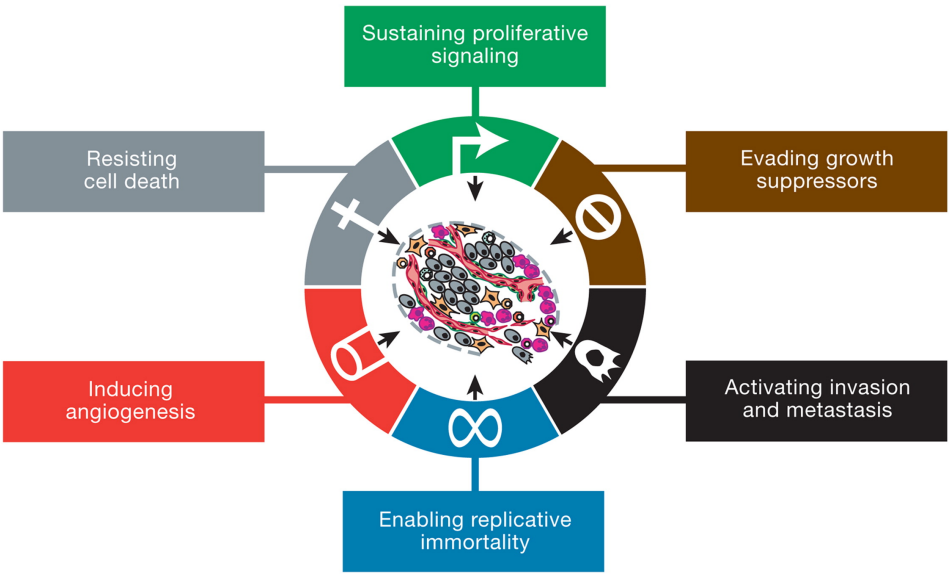

Figure 4.1: The 6 hallmarks of cancer, 6 modes of action a carcinogenic tumour can use in order to survive and proliferate in the body (Hanahan and Weinberg, 2011),

Cancer often appears with a mutation on a tumour suppressor gene, reducing the body's ability to fight cancerous cells. TP53, which is called the guardian of the genome and plays a role in DNA repair, cell division and apoptosis, was found to be mutated in 29% of cancers (Richardson, 2013). Interestingly, other species like elephants rarely suffer from cancer, unlike humans. It has been observed that elephants contain at last 20 different copies

of the TP53 gene (Haupt and Haupt, 2017), whereas humans only have one, meaning that in elephants, even if some copies of TP53 lose their function because of mutations, the other copies can still perform it, whereas this loss can have much more severe consequences for humans.

Another type of cancer related genes are oncogenes. These genes have the potential to cause cancer or promote it. They can for example help a cell escape apoptosis or accelerate cell growth and therefore proliferation. In this category, we for example find transcription factors, which are genes that regulate the expression of other genes. MYC is a family of genes that are over expressed in more than half of human cancers (Gabay et al., 2014). Over expression of this transcription factor leads to over expression of its targets, some of which are involved in cell proliferation, thus promoting cancer. Mutations of MYC or genes regulating its expression can therefore promote cancer.

Since somatic mutations accumulate in the genome over time, risks of a mutation happening on a critical region of the genome increase as we age and thus so do the risks of cancer.

4.2.1 Prostate cancer

The prostate gland is a part of the male reproductive system. This gland grows during puberty and is responsible for producing fluid that mixes with semen during ejaculation to help sperm travel. Prostate cancer is the second most prevalent cancer among men, with 1.4 million new cases and 375,000 deaths worldwide in 2020. (Sung et al., 2021). Møller et al. predict that despite a rise in prostate cancer over the years, by 2030, the mortality rate of prostate cancer in Nordic countries would be about half of the 2000 levels (Møller et al., 2003). This can be explained by changes in lifestyle, better screening to diagnose prostate cancer early, but also a better understanding of this cancer and with it, better treatments are becoming available.

Different types of prostate cancer exist (Typ):

- Adenocarcinoma: this is the most common type of prostate cancer. It is an abnormal growth of the epithelial cells of the prostate gland.

- Transitional cell carcinoma: this type of cancer can start in the urethra or the bladder and spread to the prostate.
- Small cell carcinoma, that develops in small round cells of the neuroendocrine system.
- Squamous cell carcinoma, that starts in the flat cells that cover the prostate glands.
- Prostate sarcoma, which develops outside the prostate glands in the soft tissue—the muscle and nerves—of the prostate

These different types differ by their frequency of occurrence, the speed at which they grow, their aggressiveness as well as potential symptoms.

4.3 Alternative splicing

Splicing is a process happening on pre-mRNA which involves the removal of introns and determines which combination of exons will be selected to make up the mRNA transcript. Splicing if performed by the spliceosome, a complex of RNAs and proteins called small nuclear ribonucleoproteins (snRNP), U1, U2, U4, U5 and U6. The removal of introns starts with the binding of the U1 snRNP to the GU sequence that marks the 5' end of an intron. The 3' end of an intron is marked by the AG sequence. Once the start and end of an intron have been identified, the spliceosome cuts the intron out and links the two exons together. Mutations in the motifs where the spliceosome binds can prevent it from binding the pre-mRNA, causing the intron to remain in the final mRNA sequence and potentially having a strong impact on the corresponding protein sequence.

Alternative splicing is a process allowing a single gene to generate multiple mRNA transcripts and thus, for protein coding genes, multiple proteins. While ENSEMBL (Howe et al., 2021) contains 20,465 protein coding genes, 24,849 non coding genes and 15,217 pseudogenes for humans, it also references 245,000 gene transcripts, which is an average of 4 transcripts per gene. With regard to proteins, ENSEMBL references ∼70,000 of them, or ∼3.5

different proteins per protein coding gene. This process allows an increase in protein diversity, with a single gene being able to produce proteins with completely different sequences, structures and functions as an exon inclusion can cause a frameshift or proteins with a same function but working slightly differently.

Alternative splicing is regulated by promoters, SR proteins which bind exonic splicing enhancer and result in greater inclusion of the exon in the final mRNA transcript. Alternative splicing is also regulated by repressors, hnRNPs, that bind to the exonic splicing silencer elements and reduce inclusion of the corresponding exon (Zhu et al., 2001). This explains why certain isoforms of a gene are observed in some tissues and not others, as the SR proteins and hnRNPs may be differently expressed between tissues.

Different types of alternative splicing can happen and are presented in Figure 4.2

Distribution of the type of splicing events differ greatly between species. While in humans, exon skipping events (called cassette exons in Figure 4.2) are the most common type of event and intron retention are rare, in other species such as volvox, chlamydomonas or arabidopsis, the opposite is observed (4.3).

Multiple diseases are a consequence of defective alternative splicing. Some are the result of mutations that affect the ability of an exon to be retained in the final transcript. This is the case of spinal muscular atrophy in which a mutation on position 6 of exon 7 of SMN2 results in exon 7 being excluded more often, either because the mutation breaks the exonic enhancer to which the splicing promoter SRSF1 normally binds, or by creating a binding site for the splicing repressor hnRNPA1, resulting in a truncated and probably not functional protein (Wirth et al., 2006).

Many research articles have established a link between alternative splicing and cancer (Tazi et al., 2009). Indeed, over expression of some splicing factors has been observed in different types of cancer and may affect the splicing pattern of certain genes. For example, the canonical isoform of survivin (BIRC5) has anti-apoptosis properties whereas isoform survivin-2B has pro-apoptotic properties. Yet, it has been found that the relative abundance

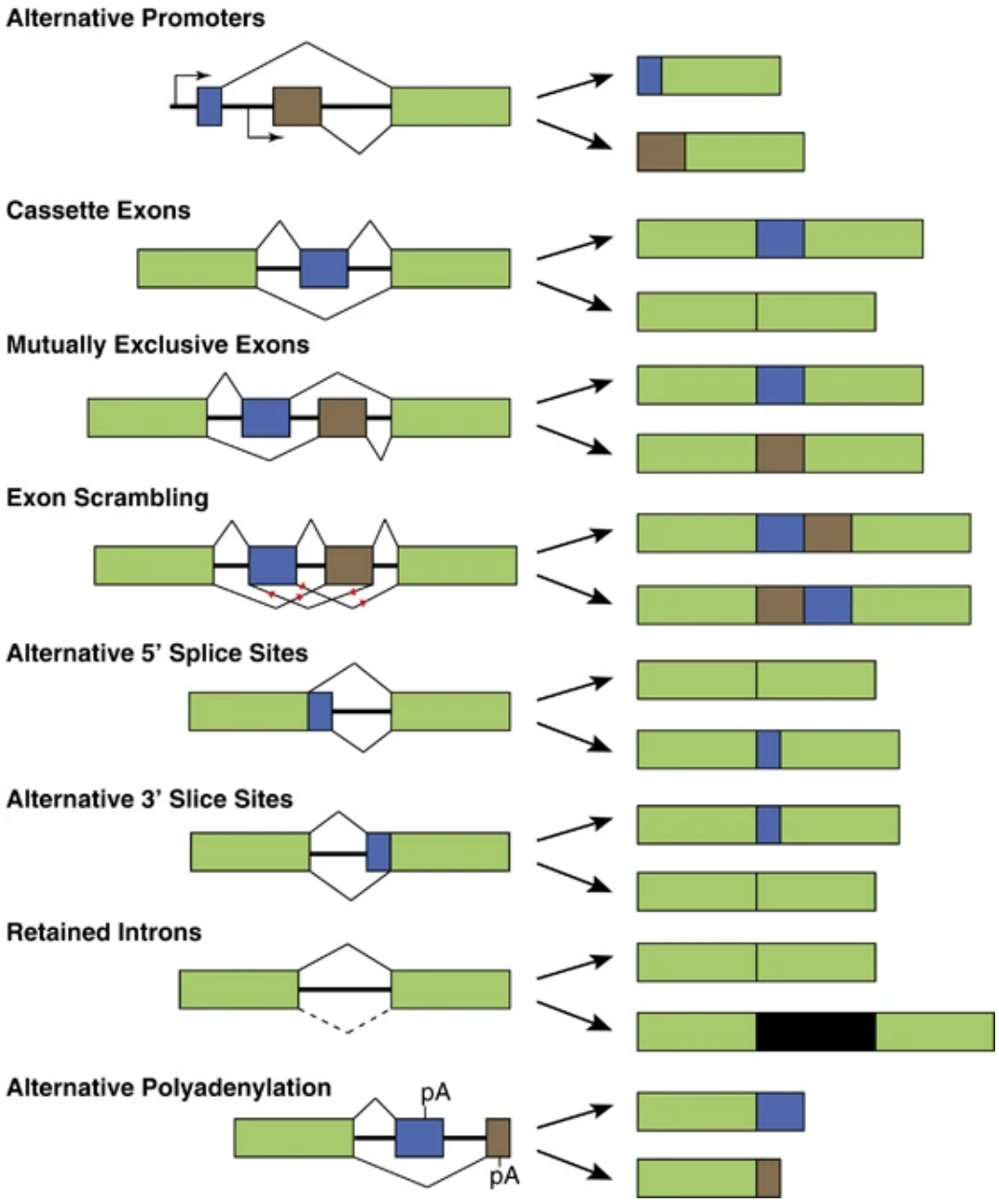

Figure 4.2: The different types of alternative splicing events that can happen. Exons are in green, blue or brown, introns in black(Chen and Weiss, 2014)

of isoform survivin-2B with regard to the abundance of the canonical survivin isoform decreases in advanced gastric carcinoma (Krieg et al., 2002), increasing the anti-apoptosis role of this gene.

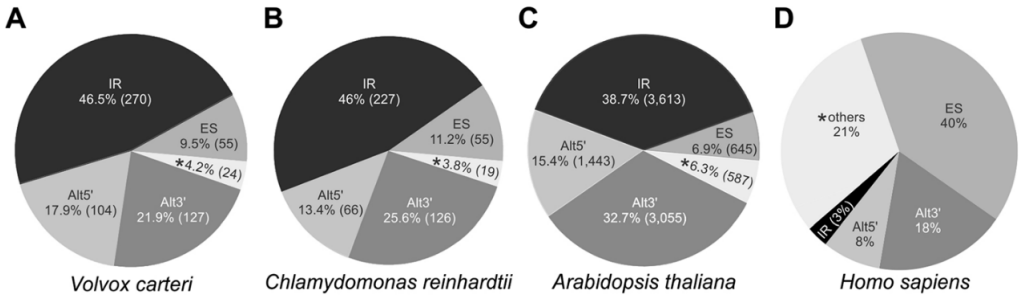

Figure 4.3: Distribution of the type of alternative splicing events in multiple species. This shows that Homo sapiens tend to have a higher proportion of exon skipping events than in Volvox carteri, Chlamydomonas reinhardtii and Arabidopsis, while these last three species have a higher proportion of intron retention. (Kianianmomeni et al., 2014)

4.4 Introduction to HNRNPA2B1

This chapter presents application of PITsuite to study the role of HNRNPA2B1. PIT was therefore used to analyse these data, and PITgui was then used to identify differences between conditions that could explain mechanisms through which HNRNPA2B1 regulates prostate cancer.

The HNRNPA2B1 gene codes for two protein isoforms, A2 and B1, which are members of the heterogeneous nuclear ribonuclear protein (HNRNP) family of RNA-binding proteins (RBPs) (Liu and Shi, 2021). HNRNPA2B1 modulates cellular phenotypes in disease via multiple different RNA processing functions, including alternative pre-mRNA splicing (Li et al., 2017) and mRNA stability (Martinez et al., 2016). In cancer, HNRNPA2B1 can stabilise (Fähling et al., 2006), (Stockley et al., 2014) or destabilise (Zuccotti et al., 2014) mRNAs or control oncogenic splicing switches during tumorigenesis (Clower et al., 2010), (David et al., 2010).

Rapid cellular proliferation during tumorigenesis requires an increased rate of protein synthesis(Lee et al., 2021), however, a limited oxygen and nutrient supply disrupts proteostasis and causes oxidative stress (Bartoszewska and Collawn, 2020). An early cellular response to stress is the stalling of mRNA translation and aggregation of pre-initiation translation complexes

into stress granules (Marcelo et al., 2021) which recruit RBPs including EWSR1, HNRNPA0, HNRNPA1 and HNRNPA2B1 (Jiang et al., 2021), (Wolozin and Ivanov, 2019) Recent studies have identified HNRNPA2B1 cytoplasmic to nuclear translocation in low oxygen conditions, and its association with the polysome, which contains proteins involved in translation, and regulates proteostasis (Ho et al., 2020), (Yao et al., 2013).

Prolonged stress-induced disruption of cellular proteostasis can lead to increased demand on the protein folding machinery of the endoplasmic reticulum (ER) (Rzymski et al., 2010), causing protein re-folding, or destruction of terminally misfolded proteins. ER stress triggers altered unfolded protein response (UPR) gene expression profiles via activation of transcription factor sensors including XBP1, ATF4, and nATF6, which control the three key signalling branches of the UPR (Han and Kaufman, 2017). Sustained UPR activation leads to increased tumorigenicity, metastatic potential, and therapy resistance of cancer cells (Cubillos-Ruiz et al., 2017). In patients, UPR pathway genes are up-regulated (Han and Kaufman, 2017), and the transcriptional targets of XBP1, ATF4 and nATF6 are associated with poor survival (Pällmann et al., 2019), (Sheng et al., 2019).

Prostate cancer (PC) is the commonest male-specific cancer and leading male-specific cause of cancer death (Rebello et al., 2021). In PC, proteostasis is disrupted (Bouchard et al., 2018), and all three branches of the UPR are activated (Pachikov et al., 2021), (Pällmann et al., 2019), (Sheng et al., 2019)). IRE-1-XBP1 activation leads to initiation of c-MYC dependent transcription and is associated with poor patient prognosis (Sheng et al., 2019). In light of evidence implicating HNRNPA2B1 in PC (Stockley et al., 2014) and cellular stress (Ho et al. 2020, Wolozin and Ivanov 2019, Yao et al. 2013), we hypothesised that HNRNPA2B1 may control several stress response pathways including UPR in PC. We reveal for the first time that HNRNPA2B1 regulates expression of UPR pathway genes including IRE1, mediates non-canonical splicing of XBP1 mRNA, and controls a gene signature of IRE1-XBP1 activation that is associated with poor PC patient prognosis.

4.5 Experimental design

NGS and LC-MS/MS data was obtained from PC3 cells. The experimental protocols are described in Appendix A.

Clinical RNA sequencing (RNA-Seq) and microarray data were obtained from cBioPortal (Cerami et al., 2012),(Gao et al., 2013), (Sanchez-Vega et al., 2018). For primary PC (The Cancer Genome Atlas; TCGA, n=491 samples; Memorial Sloan Kettering Cancer Centre; MSKCC, n=179 samples), from Sanchez-Vega et al. ((Sanchez-Vega et al., 2018) for adjacent benign prostate (TCGA, n=52), and cBioPortal (Cerami et al., 2012), (Gao et al., 2013) for metastatic PC (Stand Up to Cancer; SU2C, n=208 samples). Gene expression values were reported for TCGA as RNA-Seq by Expectation-Maximization (RSEM), for SU2C as Fragments per Kilobase of exon Per Million mapped fragments (FPKM) cohorts, or for MSKCC as log2 whole transcript mRNA expression. For comparison of normal (TCGA, n=52) and primary PC tissue (TCGA, n=497) RNA-Seq data were obtained from the Broad Institute Genome Data Analysis Center (GDAC) Firehose database (doi:10.7908/C11G0KM9) (Supplementary Table 1). Cell line RNA-Seq data for LNCaP cells treated with siRNA to XBP1 or and IRE1 inhibitor (MKC8866) were obtained from (Sheng et al., 2019) and gene expression values reported as Log2 Fold Change and adjusted p-value.

4.6 PIT analysis

The RNA-Seq and LC-MS/MS obtained from HNRNPA2B1 samples were analysed through PIT. A first quality control step was performed to make sure the silencing of HNRNPA2B1 was successful.

From the RNA-Seq experiment, we can see (4.4) that HNRNPA2B1 expression is lower in the si samples than in the control NSI sample (Log2 fold change = -4.05, adjusted p-value$<$ 0.001).

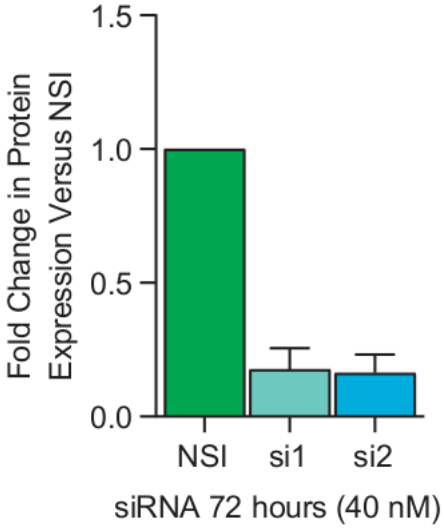

Figure 4.4: Fold change of HNRPA2B1 RNA expression in si1 and si2 samples with regard to the Nsi samples.

In addition, a PCA analysis of gene expression from RNA-Seq shows that replicates of each condition are clustered together on the first principal component axis (4.5). Since this principal component accounts for 33.74% of variance observed in gene expression, it is a good indicator that silencing of HNRNPA2B1 was successful. We can also some clustering on the second principal component, with SI samples being higher than NSI samples on this axis. The exception is Nsi/1 which doesn't cluster with the others NSI samples on the second principal component axis. We suspect some sequencing issues with this sample, as the total read count from RNA-Seq was much lower than in the other samples. Yet, since it clusters well with the other

samples on the first principal component, we decided to keep it for the rest of the analysis.

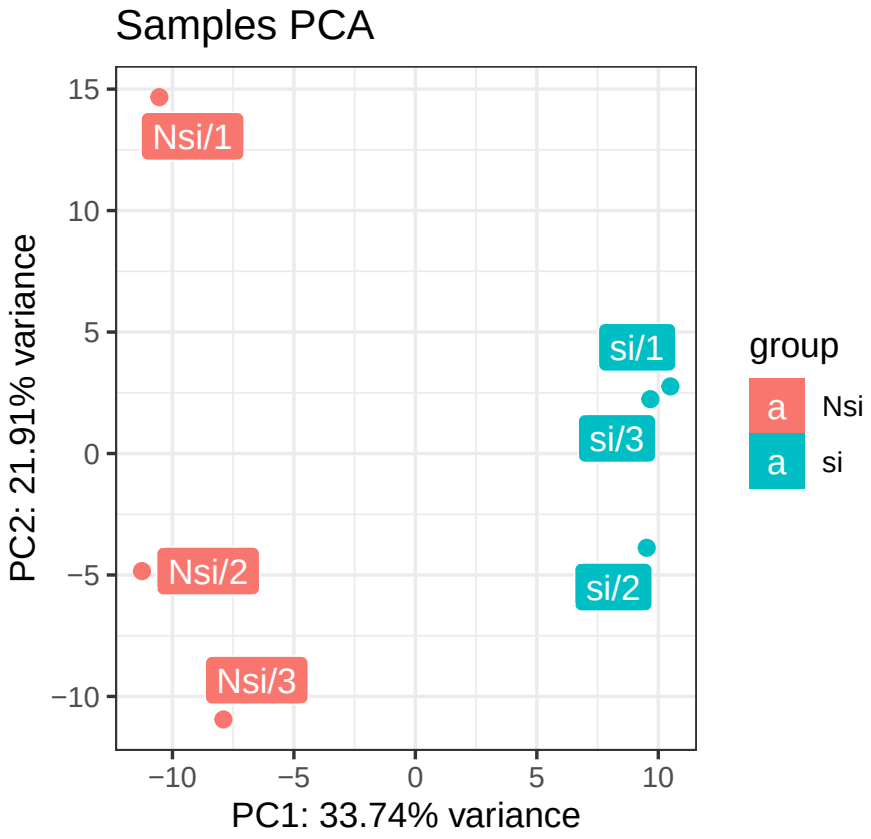

Figure 4.5: Principal component analysis showing distribution of the samples based on the RNA genes expression

4.7 Linking HNRNPA2B1 expression to disease free survival in prostate adenocarcinoma

We have previously shown that HNRNPA2B1 protein expression is specifically up-regulated in patients with aggressive prostate cancer (PC) (Stockley et al., 2014). To validate these findings, we explored HNRNPA2B1 expression in RNA sequencing (RNA-Seq) data from primary prostate tumours (n=491) and adjacent benign prostate tissue (n=52) (Sanchez-Vega et al., 2018). HNRNPA2B1 mRNA expression was significantly higher in tumours compared to adjacent benign prostate tissue (4.6). To determine whether high expression of HNRNPA2B1 is associated with poor patient prognosis, we stratified tumours into two groups based on the normalized expression levels of HNRNPA2B1, with high expression considered the top 25% of the distribution across samples, and the rest of samples considered as low expression. High expression of HNRNPA2B1 was associated with a statistically significant reduction in patient survival, as compared with patients with low HNRNPA2B1 expression (4.7).

To show the monotonous relationship between HNRNPA2B1 expression in prostate cancer patients and disease-free survival, patients were then split into 3 groups:

1. Low: Lowest quartile of HNRNPA2B1 expression
2. Medium: Between 25% and 75% levels of HNRNPA2B1 expression
3. High: Highest quartile of HNRNPA2B1 expression

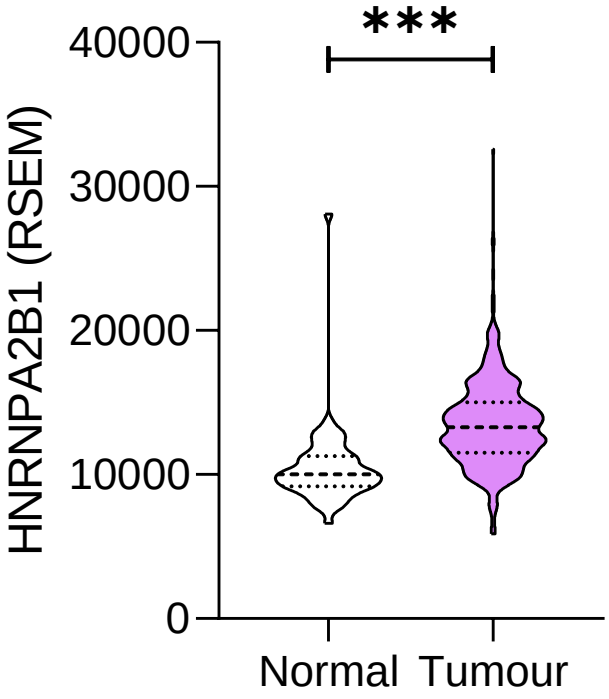

Figure 4.6: Distribution of HNRNPA2B1 expression values reported as RNA-Seq by Expectation-Maximization (RSEM) in primary prostate tumours and benign adjacent tissue from The Cancer Genome Atlas (TCGA) patient cohort. Two-tailed T-test was used to compare treatment groups. *** = $p < 0.001$

Cox proportional hazard ratios were calculated for each category, taking as reference the low expression group using the R survival package (4.8) (Therneau and Grambsch, 2000). This revealed that patients with medium expression of HNRNPA2B1 were already significantly more at risk (p=0.015) than those with low expression, with a hazard ratio of 2.60. Additionally, patients with high expression of HNRNPA2B1 were even more at risk, with a hazard ratio of 4.21 ($p < 0.001$). This indicates that higher levels of HNRNAPA2B1 expression is monotonously linked to higher risk in prostate

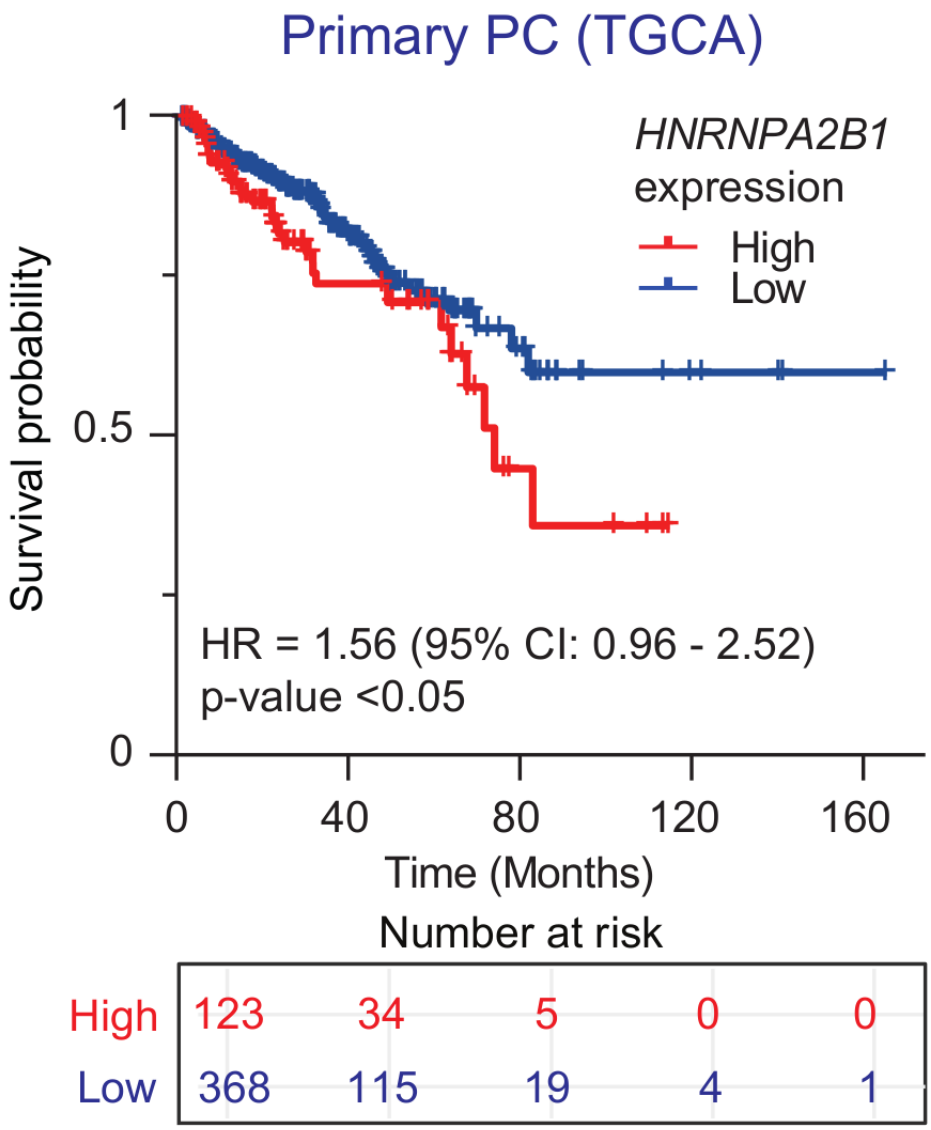

Figure 4.7: Kaplan-Meier plot of disease-free survival for primary PC patients stratified by HNRNPA2B1 expression (low = $< 1st - 3rd$ quartile and high = $> 3rd$ quartile). The number of patients at risk for each group are presented in the table below each X-axis time point. Univariable Cox PH-derived hazard ratios (HR) with 95% confidence intervals (CI) and two-tailed log-rank test p-values are shown

cancer.

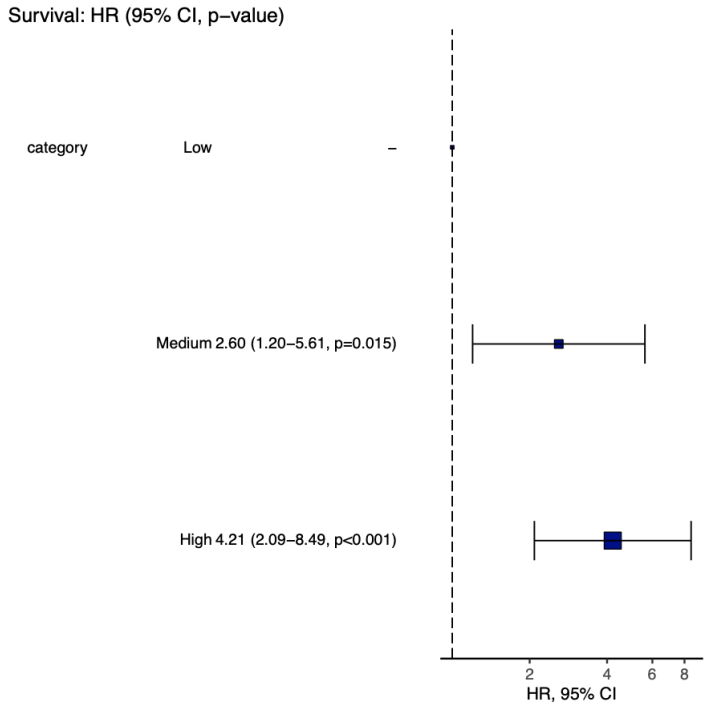

Figure 4.8: Hazard ratios showing risks for patients depending on their HNRNPA2B1 expression

Given the previously established roles for HNRNPA2B1 in the hypoxic response (Ho et al., 2020), (Yao et al., 2013) and stress granule formation (Wolozin and Ivanov, 2019), we wished to determine the most significant cellular stress pathways regulated by HNRNPA2B1 in PC. Firstly, we performed Gene Set Enrichment Class Analysis (GSECA) (Lauria et al., 2020) on RNA-seq datasets from primary (n=491) (Hoadley et al., 2018) and metastatic PC (CRPC) (n=208) (Abida et al., 2019). We compared KEGG stress pathway representation in tumours with high HNRNPA2B1 expression compared with low expression. In primary PC, we found that the top stress pathways associated with high expression of HNRNPA2B1 included the "Proteasome" and "HIF1 signalling pathway" (4.9). In metastatic PC, top pathways associated with high expression of HNRNPA2B1 included "Protein processing in endoplasmic reticulum", "Autophagy", and diseases with a misfolded protein component (4.10).

Subsequently, we performed gene set enrichment analysis (GSEA) using

all KEGG pathways to identify top biological processes enriched upon HNRNPA2B1 depletion on our PC3M cells. This enrichment amongst differentially-expressed genes with log2 fold change of < -0.5 or > 0.5 at $p < 0.05$ significance was performed in R V.4.1.1 using the enrichKEGG function of the clusterProfiler pa(Wu et al., 2021). Consistent with the association of HNRNPA2B1 with cellular stress pathways in PC patients, the KEGG stress pathway "Protein processing in endoplasmic reticulum" was the most significantly enriched pathway 4.11. Within this pathway, HNRNPA2B1 depletion led to down-regulated expression of PERK, ATF6 and IRE1, which encode for the three master ER-stress sensors mediating three key signalling branches of the UPR (Luo and Lee, 2013) 4.12.

Taken together, these data in PC patients and cell lines indicates that HNRNPA2B1 regulates cellular stress pathways, with the most significant pathway being "Protein processing in the endoplasmic reticulum" in PC cells incorporating UPR genes.

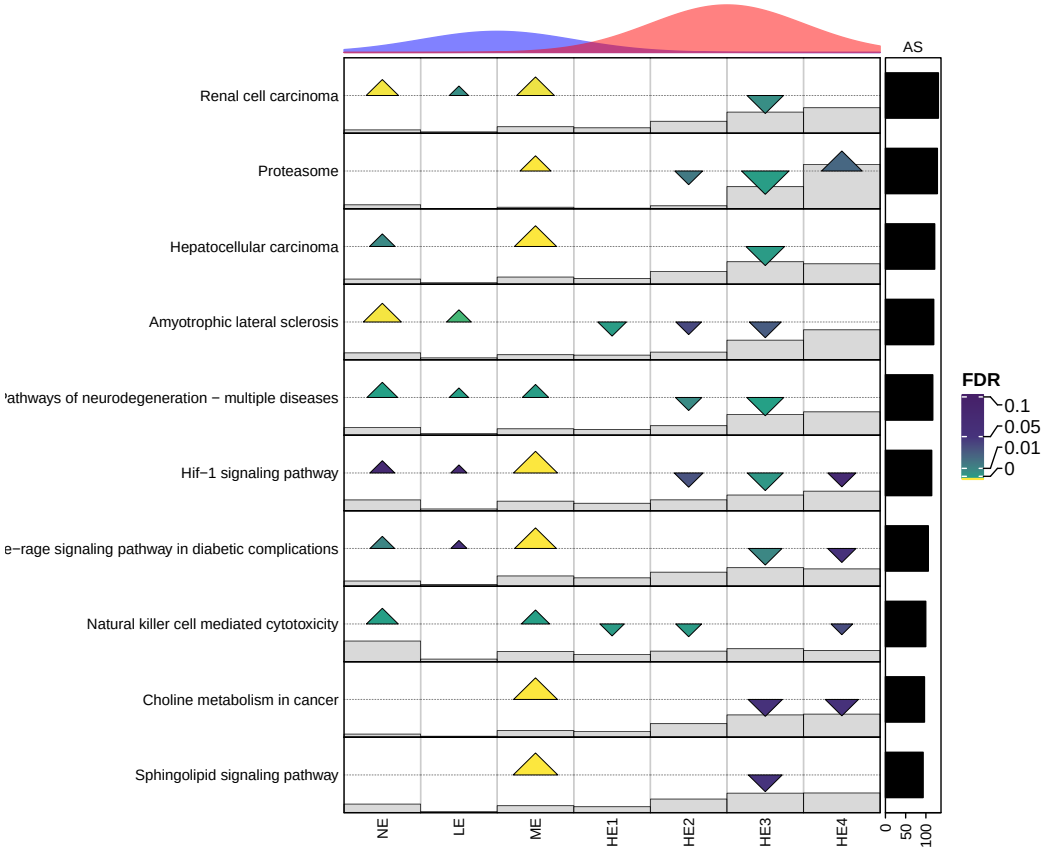

Figure 4.9: GSECA analysis performed on primary PC (TCGA) RNA-Seq dataset by stratification of cohorts based on HNRNPA2B1 expression. Genes in a given Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway are separated into seven expression classes: NE = not expressed, LE= lowly expressed, ME = medium expression, HE1-4 = high expression. Triangles compare the difference in the cumulative proportion of genes in an expression class between HNRNPA2B1 high and low expression groups, and represent the size and enrichment (up) or depletion (down) of genes. AS = association score.

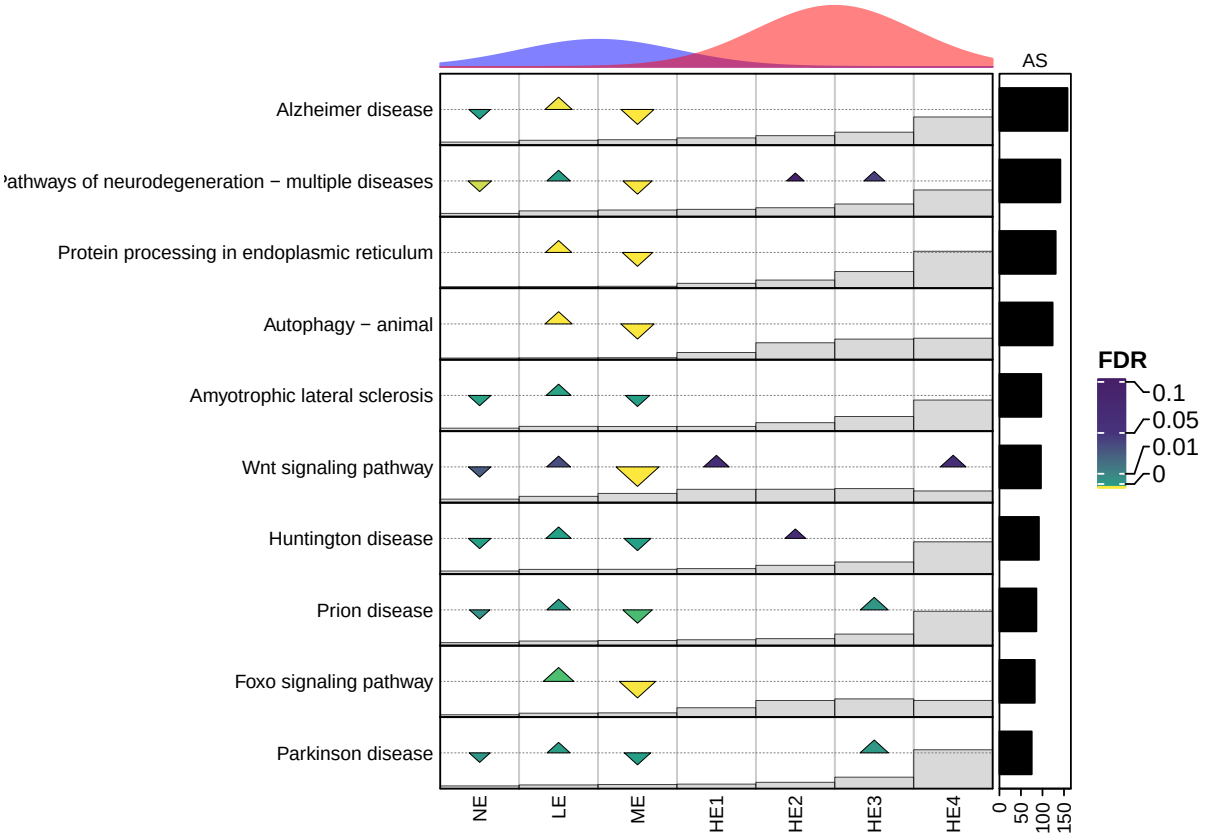

Figure 4.10: GSECA analysis performed on metastatic PC (SU2C) RNA-Seq dataset

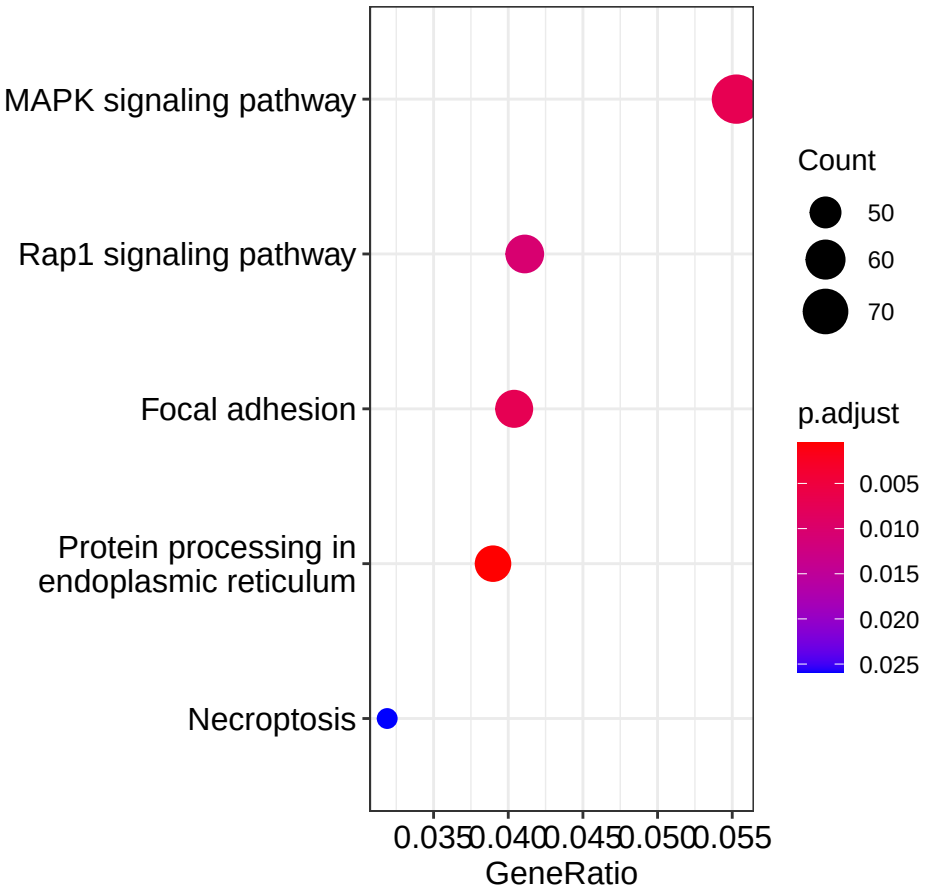

Figure 4.11: KEGG pathway gene enrichment analysis of differentially expressed genes ($p < 0.05$ and absolute log2 fold change > 0.5 or < 0.5) identified by RNA-Seq of PC3M cells upon depletion of HNRNPA2B1 using a single siRNA duplex (si1, 20nM for 72 hours).

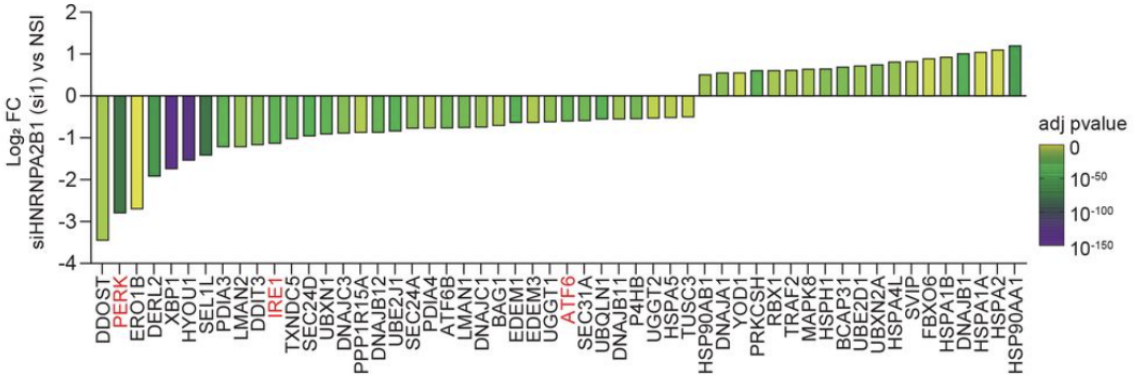

Figure 4.12: Log2 fold change gene expression values for differentially expressed "Protein processing In endoplasmic reticulum" genes upon HNRNPA2B1 depletion in PC3M cells ($p < 0.05$ and absolute log2 fold change > 0.5 or < 0.5). P-values for each gene adjusted using the Benjamini and Hochberg method are represented by the bar colour

Looking further at the protein processing in endoplaplasmic reticulum pathway (4.13), we can see most genes being downregulated after HNRNPA2B1 silencing. This is in particular the case of the 3 stress sensors PERK, ATF6 and IRE1 as well as XBP1. The exception is with ER-associated degradation (ERAD) genes, which are mostly upregulated.

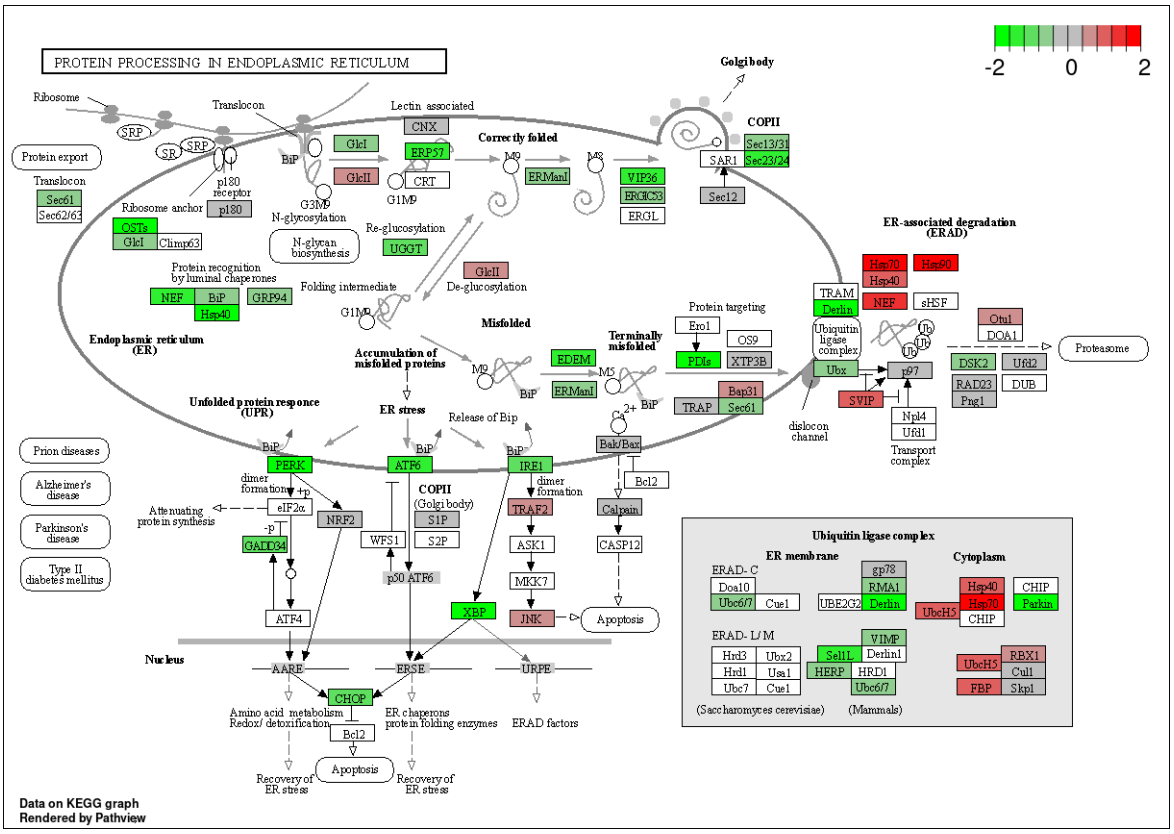

Figure 4.13: KEGG pathway for protein processing in endoplaplasmic reticulum pathway. Genes downregulated after HNRNPA2B1 silencing are shown in green, those upregulated are shown in red.

4.8 HNRNPA2B1 affects processing of IRE1 target mRNAs

To shed light on a putative mechanism of HNRNPA2B1-mediated UPR gene expression, we focussed on the IRE1-XBP1 signalling branch, considering its association with PC disease recurrence (Sheng et al., 2019). XBP1 transcriptional activation requires non-canonical cytoplasmic splicing of XBP1u mRNA to produce the transcriptionally active XBP1s via removal of a variable 26 nucleotide sequence in exon 4 by IRE1 nuclease activity (Calfon et al., 2002), (Uemura et al., 2009) (4.14). We hypothesised that HNRNPA2B1 may regulate UPR genes via XBP1 splicing. To test this, we used established RT-PCR based splicing assays (Savic et al., 2014) to measure the percentage expression of activated XBP1s compared with XBP1u (Fig. 2A). Following treatment of PC3M cells with the UPR activator Thapsigargin (da Silva et al., 2020); we observed a statistically significant increase in XBP1s splicing, compared to controls (4.15). Conversely, following HNRNPA2B1 protein depletion in PC3M cells using two independent siRNA duplexes (4.16); we observed a statistically significant decrease in XBP1s splicing compared with controls (4.17). These data demonstrate that HNRNPA2B1 promotes the non-conventional splicing of XBP1u to XBP1s.

IRE1 also degrades several mRNAs, including the BLOC1S1 mRNA, which encodes a regulator of lysosomal function, as part of the regulated IRE1-dependent decay (RIDD) pathway during ER stress (Chalmers et al., 2019), (Lhomond et al., 2018). We wished to determine whether HNRNPA2B1 could also affect the RIDD pathway by exploring its impact on BLOC1S1 expression. Following treatment of cells with the UPR activator Thapsigargin, we observed a statistically significant reduction in BLOC1S expression (4.18). Concordant with the impact of HNRNPA2B1 on XBP1 splicing, we observed a statistically-significant increase in BLOC1S1 expression upon HNRNPA2B1 depletion (4.19). These data indicate that HNRNPA2B1 may affect multiple IRE1-dependent gene regulatory functions in PC cells.

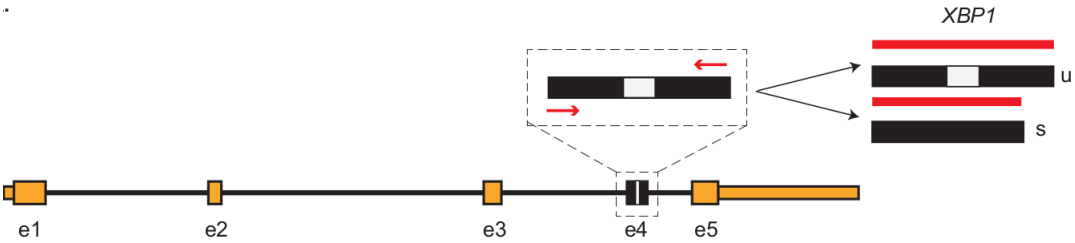

Figure 4.14: Schematic of XBP1 gene. Exons 1-3 and 5 are indicated by yellow boxes, and the non-canonically spliced exon 4 by a black box. XBP1u contains a variable 26-nucleotide region in exon 4 indicated by a white box, the exclusion of which generates the transcriptionally active XBP1s isoform. Red arrows represent RT-PCR primers used to amplify XBP1u and XBP1s products

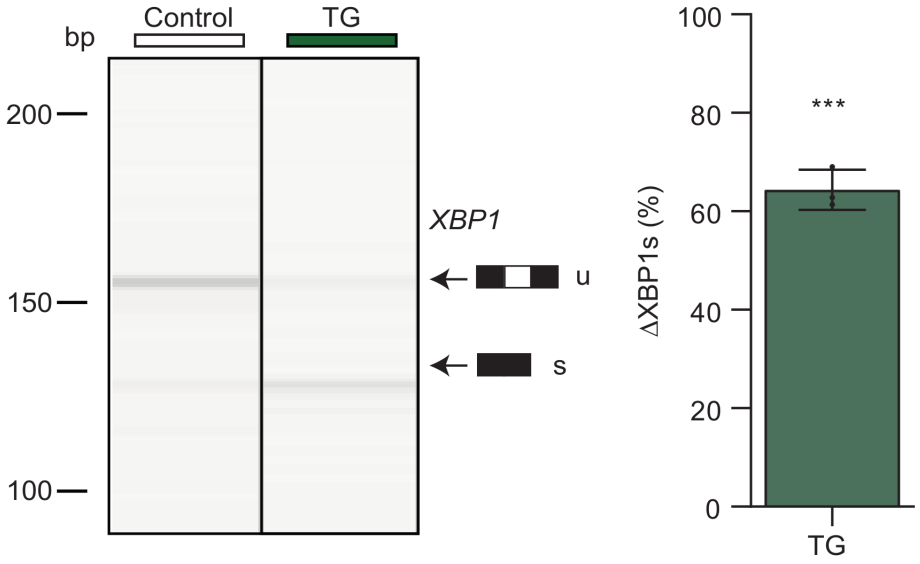

Figure 4.15: (Left panel) PC3M cells were treated with 250 nM Thapsigargin (TG), or vehicle (Control) DMSO for 24 hours and total RNA analysed using XBP1 splicing assays. Representative capillary gel electrophoretogram (QIAxcel) shows two bands representing transcripts with (XBP1u) or without (XBP1s) the exon 4 variable 26-nucleotide region inclusion. (Right panel) Electrophoretograms were quantified to determine the percentage change in XBP1s product expression (XBP1s)

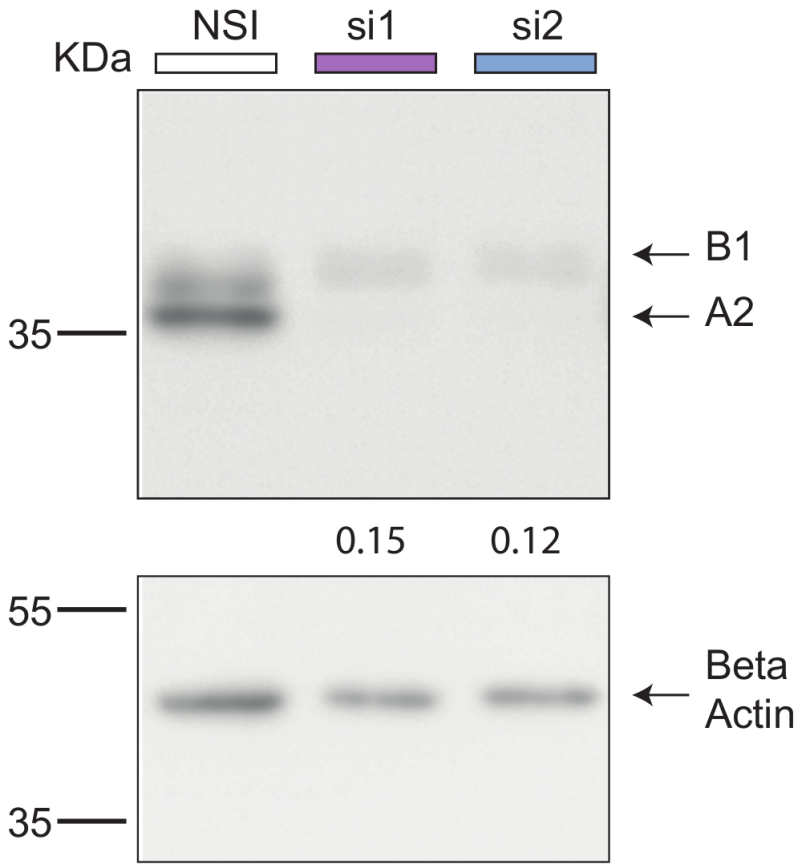

Figure 4.16: PC3M cells were depleted of HNRNPA2B1 expression using two different siRNA duplexes (si1 and si2, 20nM for 72 hours) or non-silencing control (Nsi). Western blot shows HNRNPA2 (major isoform) and B1 (minor isoform) protein expression compared to Beta Actin loading control. The numbers below the HNRNPA2B1 blot indicate the relative reduction in total HNRNPA2B1 protein expression following siRNA depletion compared to Nsi control.

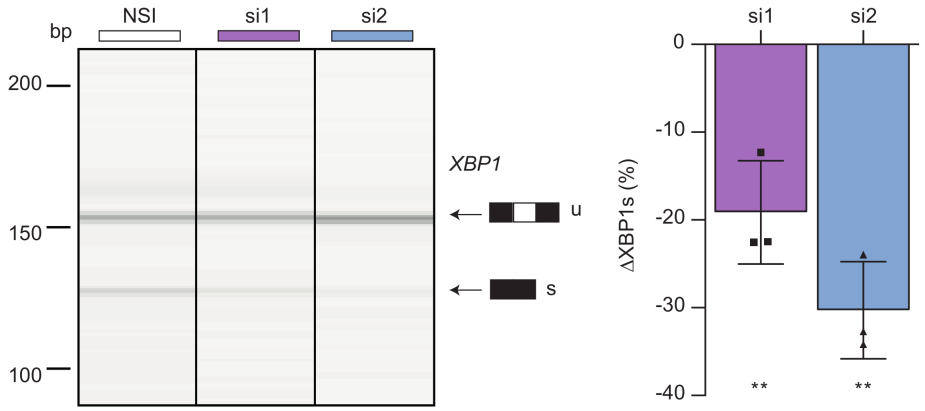

Figure 4.17: (Left panel) Total RNA was analysed using XBP1 splicing assays and representative capillary gel electrophoretogram show two bands representing XBP1u and XBP1s transcripts. (Right panel) Electrophoretograms were quantified to determine the percentage change in XBP1s product expression (XBP1s)

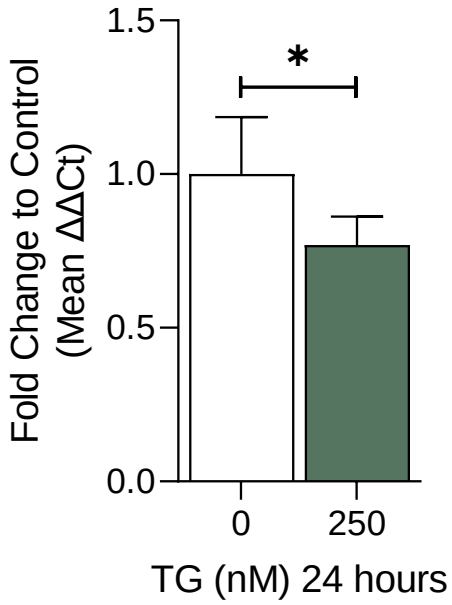

Figure 4.18: Relative change in BLOC1S1 expression to DMSO control measured by qRT-PCR in PC3M cells treated with vehicle (Control) DMSO or Thapsigargin (TG) 250nM for 24 hours.

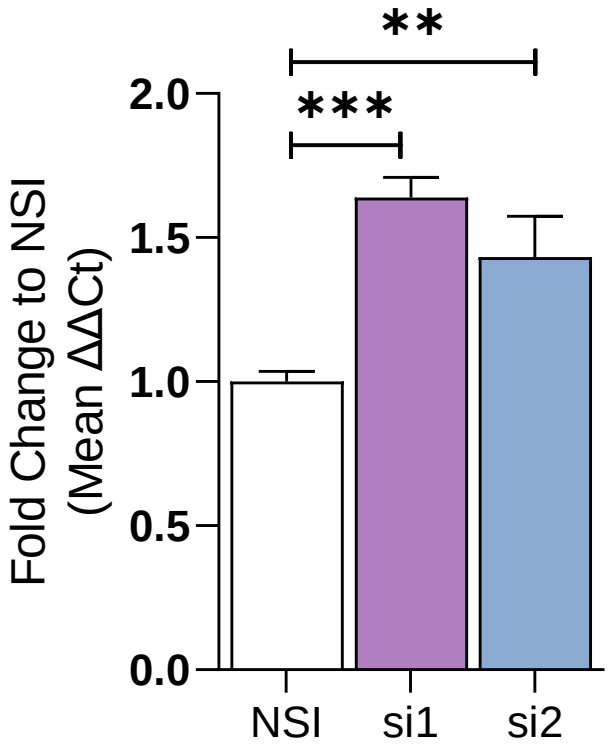

Figure 4.19: Relative change in BLOC1S1 expression to Nsi measured by qRT-PCR in PC3M cells depleted of HNRNPA2B1 expression using two different single siRNA duplexes (si1 and si2, 20nM for 72 hours). At least three biological replicates were used, and Two-tailed T-test was used to compare treatment groups. * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$

4.9 HNRNPA2B1-IRE1-XBP1 co-regulated genes represent a prognostic biomarker signature in primary PC and reveal a potential therapeutic target

Since high HNRNPA2B1 expression is associated with poor PC patient prognosis (Fig. 1B), we hypothesised that this phenomenon may be mediated, in part, by HNRNPA2B1-dependent IRE1-XBP1-related gene expression. To test this, we utilised previously published RNA-Seq data from PC cells depleted of XBP1 or treated with the IRE1 inhibitor MKC8866 (Sheng et al., 2019). To identify protein-coding genes co-regulated by XBP1, IRE1, and HNRNPA2B1, we overlapped lists of differentially expressed protein-coding genes in the three datasets (4.20). We identified a total of 20 HNRNPA2B1-IRE1-XBP1 co-regulated protein-coding genes.

To determine if these 20 genes, or a subset thereof, were associated with disease recurrence, we performed elastic net regression using expression values of these genes in the TCGA cohort and time-to-event data. We applied elastic net selection at lambda with the least mean cross-validation error and coefficient >0.00025 or <-0.00025. We identified four HNRNPA2B1-IRE1-XBP1 (HIX)-regulated genes (FKBP14, TMEM39A, BET1, and CDC6) as the best predictors of disease relapse. Using multivariable Cox regression coefficient-derived patient risk scores for the four genes (see Materials and Methods), we stratified TCGA patients into two risk groups (low risk = <1st-3rd quartile, high risk = >3rd quartile). The high risk group was significantly more likely to relapse compared with the low risk group (4.21, 4.22). Kaplan-Meier plots were generated using time to event data (event = disease recurrence) from patient cohorts using the survfit function of the survminer package in R V.4.1.1 and plotted using ggsurvplot. Univariable analyses were performed using the coxph function of survminer to compare patients with low and high risk scores. To validate the model, risk scores calculated using the coefficients obtained from the derivation cohort were applied to scaled gene expression values from the validation cohort, and Ka-

plan Meier plots generated stratified by risk scores (low = <1st-3rd quartile, high = >3rd quartile).

4.9.1 Limitations

Our study has several limitations: Although the novel link between HNRNPA2B1 and UPR was identified in metastatic PC, the HNRNPA2B1-IRE1-XBP1-controlled prognostic biomarker signature (HIX) was only validated in primary PC patients and based on mRNA expression. HNRNPA2B1 regulated PERK and ATF6 as well as IRE1 expression 4.12, however our validations focussed exclusively on IRE1-XBP1. Hence, we do not know the impact of HNRNPA2B1 on other UPR pathway branches. The precise molecular mechanisms underlying the HNRNPA2B1-mediated regulation of IRE1 and XBP1 remains unclear and warrants further investigation. Future studies using multiple UPR inhibitors in pre-clinical cancer models are required to determine whether targeting one or more UPR branches has therapeutic efficacy for HNRNPA2B1-overexpressing PC patients.

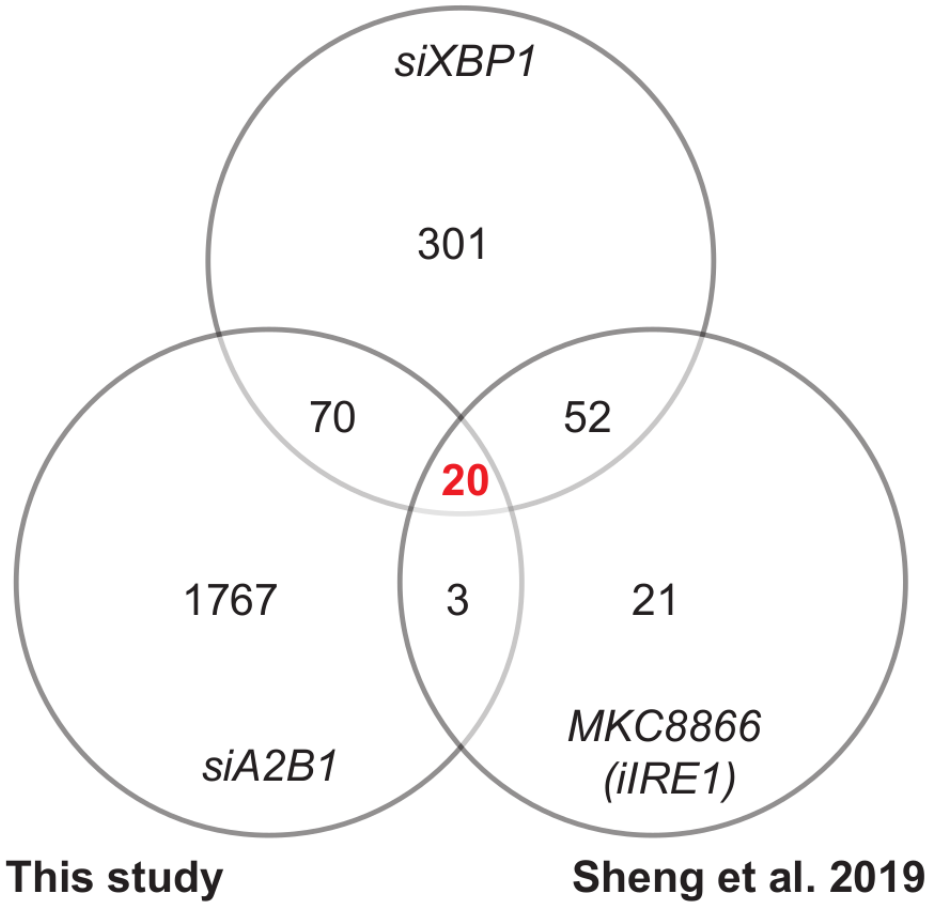

Figure 4.20: Venn diagram of protein-coding genes differentially-expressed and co-regulated by XBP1, IRE1 and HNRNPA2B1 with Log2 fold change <-0.5 and p<0.05 in RNA-Seq datasets from LNCaP cells treated with siRNA to XBP1 or IRE1 inhibitor MKC8866 (Sheng et al., 2019) or PC3M cells treated with siRNA to HNRNPA2B1.

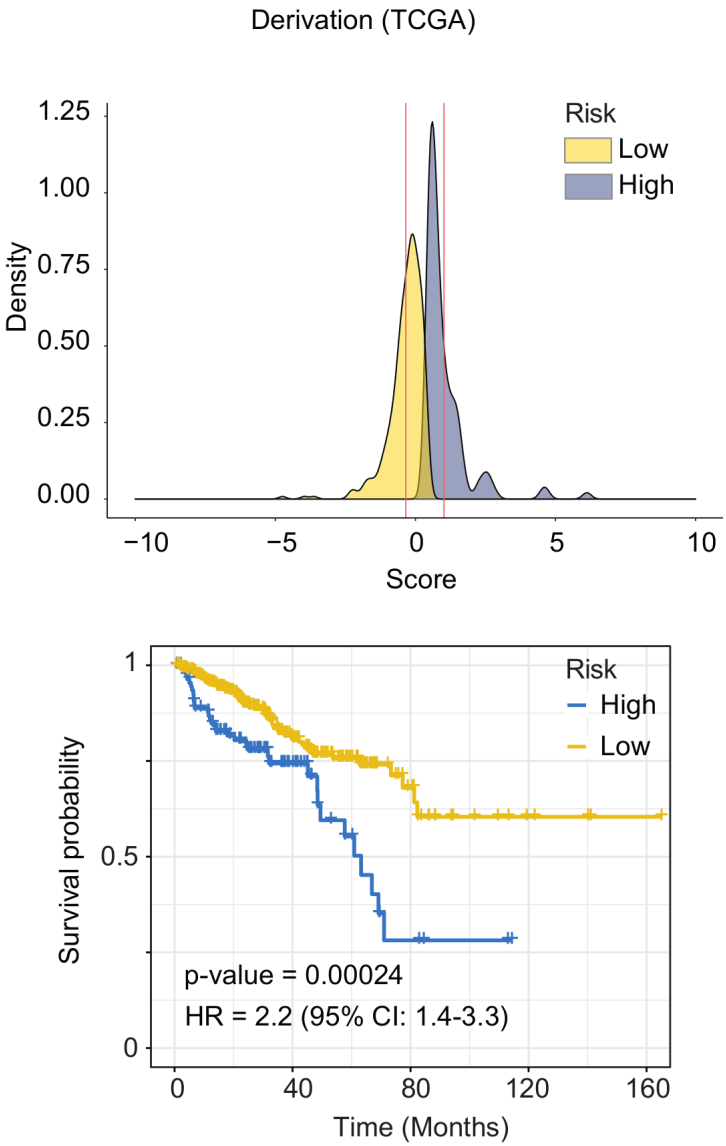

Figure 4.21: (Top panel) Distribution plot of risk scores for derivation (TCGA) cohort. Vertical red lines represent the mean of low and high percentile risk scores. (Bottom panel) Kaplan-Meier plots of disease-free survival probabilities for patients from derivation (TCGA) datasets stratified by risk groups. The number of patients at risk for each group are presented in the table below each X-axis time point. Univariable Cox PH-dervied hazard ratios with 95% confidence intervals (CI) and two-tailed log-rank test p-values are shown.

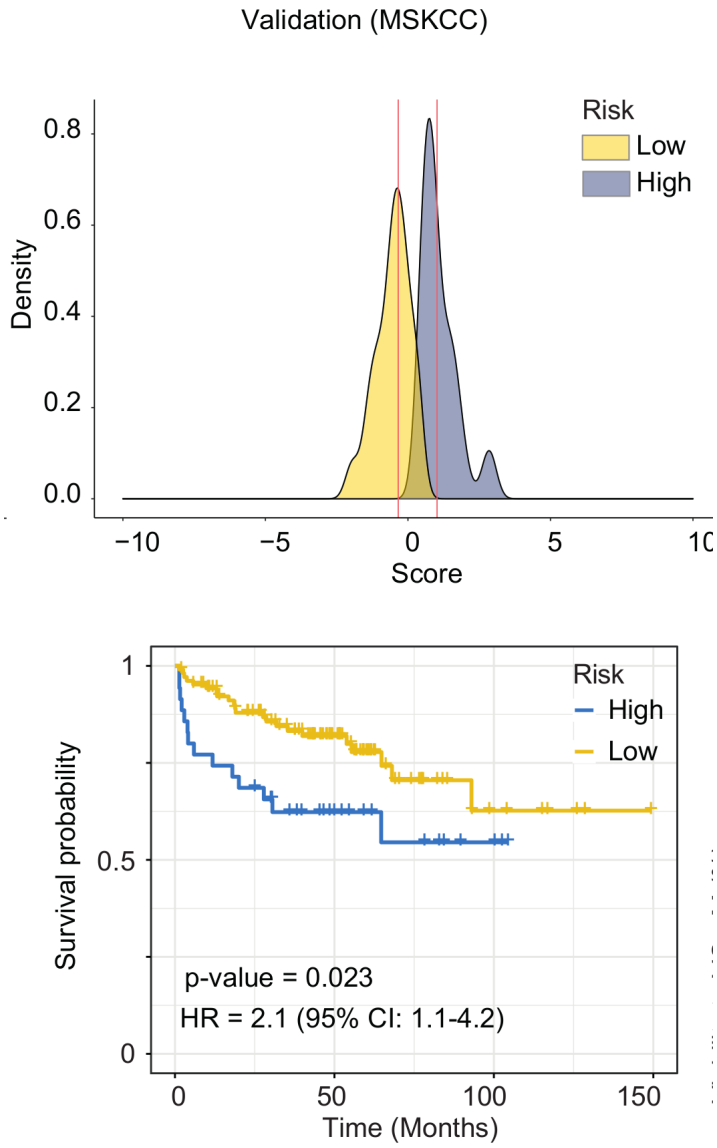

Figure 4.22: (Top panel) Distribution plot of risk scores for validation (MSKCC) cohort. Vertical red lines represent the mean of low and high percentile risk scores. (Bottom panel) Kaplan-Meier plots of disease-free survival probabilities for patients from validation (MSKCC) datasets stratified by risk groups. The number of patients at risk for each group are presented in the table below each X-axis time point. Univariable Cox PH-dervied hazard ratios with 95% confidence intervals (CI) and two-tailed log-rank test p-values are shown.

4.10 Alternative splicing of stress related genes by HNRNAPA2B1

HnRNAPA2B1 being a splicing factor, many alternative splicing events were differentially found between HNRNPA2B1 depleted and control samples using PIT. Among them, several events were found on genes related to stress pathways 4.23.

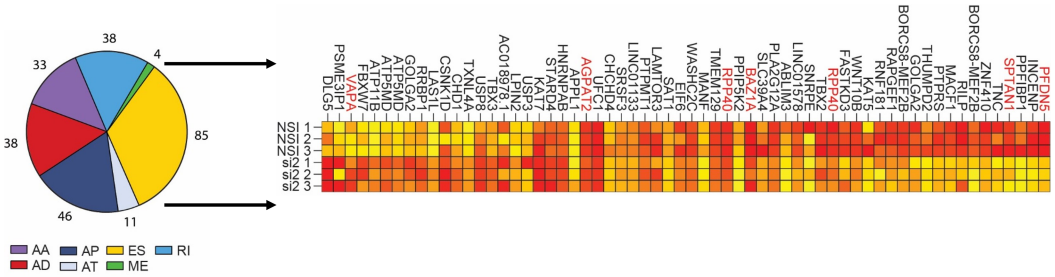

Figure 4.23: List of genes with significantly different exon skipping events between Nsi and si conditions found by PIT. Colour represent PSI in each sample for the selected event with a linear gradient (0: yellow; 100: red)

These results were not included in (Foster et al., 2022) as we focused on the dysregulation of the unfolded protein response pathway by HNRNPA2B1. Yet, we introduce more analysis here as they present some potentially more interesting biological results and a good example of how PIT can be used.

4.10.1 Alternative splicing of SPTAN1

SPTAN1 is a protein involved in cell adhesion, cell-cell contact and apoptosis (Ackermann and Brieger, 2019). It has been shown that its overexpression had tumour promoting effects, while nuclear SPTAN1 can have tumour suppressing effects by enabling DNA repair through interaction with DNA repair proteins (Ackermann and Brieger, 2019). Using PIT, we detected an exon skipping event on exon 23 of SPTAN1.

As seen in 4.25, exon 23 is more retained on SPTAN1 in the Nsi samples than in the si samples. Using SUPPA2, we can also observe this difference,

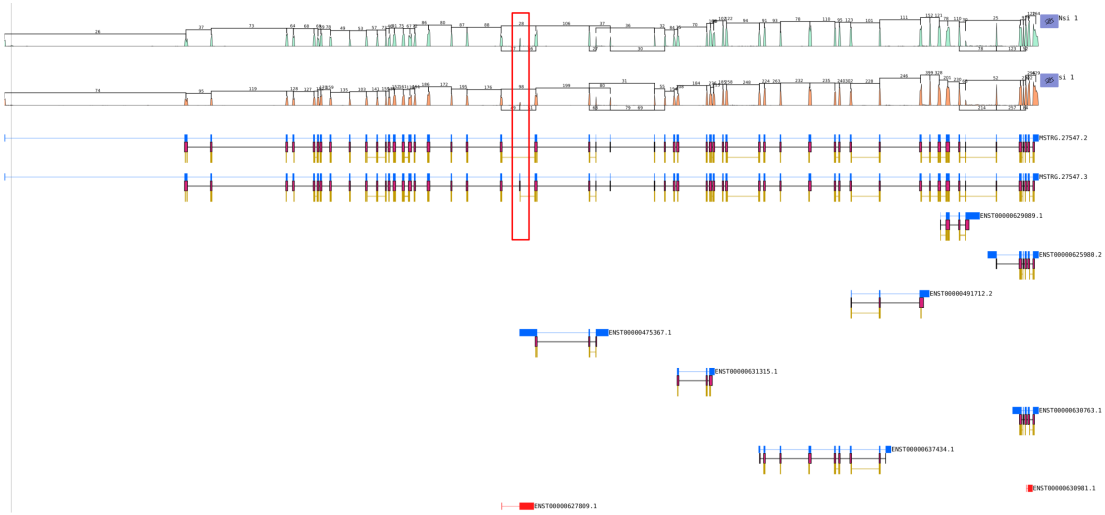

Figure 4.24: SPTAN1 gene represented in the PITgui gene browser. The red rectangle shows exon 23 where the splicing event is taking place.

Figure 4.25: Zoom on SPTAN1 exon 23 in the PITgui gene browser. Read coverage shows the exon is more included in the Nsi samples (green) than in the si samples (orange)

with the Nsi sample having a higher PSI than the si samples (4.26) with a ΔPSI of 0.64. We confirmed this with a PCR 4.27 which shows a clear switch in isoforms between Nsi and si samples.

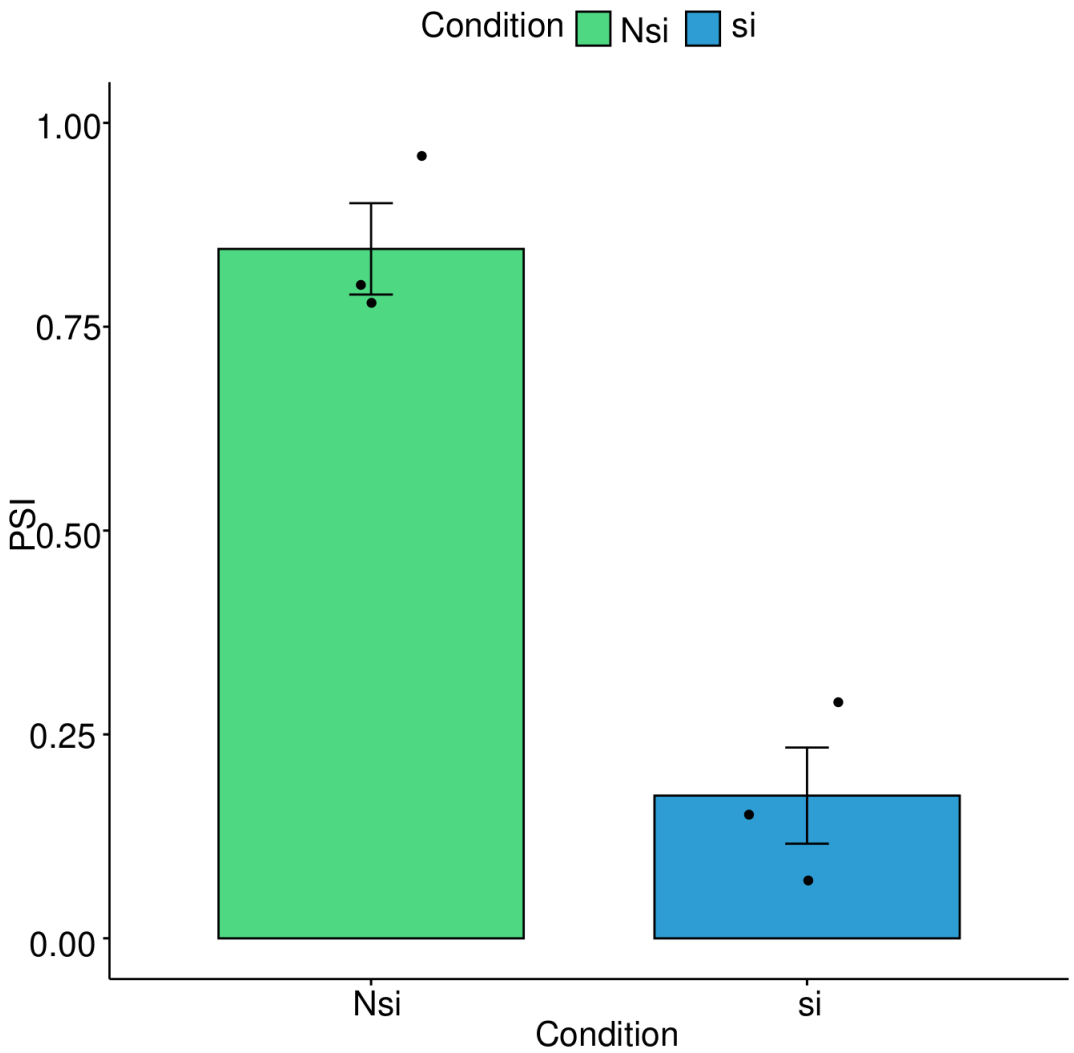

Figure 4.26: Percent spliced in (PSI) for exon 23 of SPTAN1 for Nsi and si samples

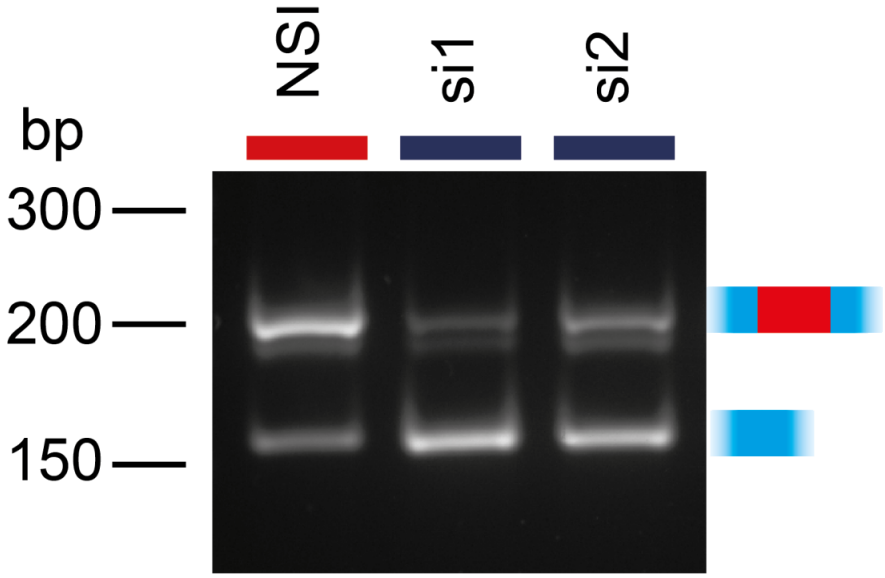

Figure 4.27: PCR showing inclusion of exon 23 of SPTAN1 in Nsi and si samples.

Additionally, it can be seen from 4.25 that a peptide was identified through mass spectrometry that uniquely maps to this event. The peptide has the sequence QVEELYHSLLELGEK and overlaps with exons 23 and 24 (4.28). Based on the methodology described in 2.4.7.2, we can determine if the change in inclusion of this exon between the two conditions is also reflected at the protein level. We then find that at the protein level, inclusion of exon 23 is about half in the si condition than it is in the Nsi condition (4.29), which is consistent we what we observed at the RNA level.

Figure 4.28: Representation of exons 22, 23 and 24 of SPTAN4 with the translated amino acid sequence mapping to these regions. We identified through LC-MS/MS the peptide QVEELYHSLLELGEK which overlap with both exon 23 and 254.

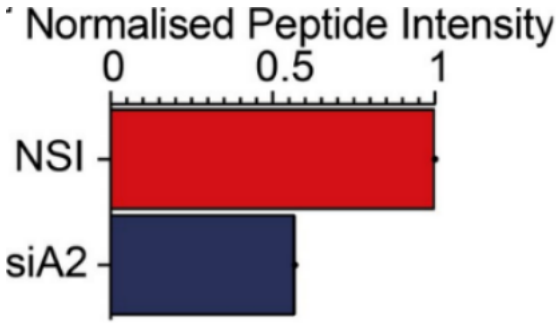

Figure 4.29: Normalised peptide intensity for the peptide QVEELYHSLLELGEK which overlaps with exon 23 and 24 of SPTAN1, showing less inclusion of exon 23 in the si condition.)

4.11 HNRNPA2B1 regulates gene expression through UTR binding

Untranslated regions (UTR) play a critical role is gene expression regulation as they are involved in controlling the translation, degradation and localisation of mRNA transcripts (Mignone et al., 2002). Multiple articles have been published showing that HNRNAPA2B1 binds to the 3' UTR of certain genes and thus controls the stability of their mRNA transcripts (Yin et al., 2021) (Liu et al., 2020)(Stockley et al., 2014). Using bulk RNA-Sequencing data and analysing it through PIT, we can observe some patterns in how gene expression is deregulated by HNRNPA2B1 silencing. By looking at the volcano plot (Kev) from the differential gene expression, it appears that amongst the genes that are significantly deregulated (adjusted p-value < 0.05), there are more genes downregulated than upregulated 4.30. By looking into details, we observe that 65% of significantly differentially expressed genes are downregulated in si versus 35% that are upregulated (4.34).

To test the hypothesis that HNRNAPA2B1 was regulating these genes through UTR binding, we used eCLIP (Van Nostrand et al., 2016) data performed on HNRNPA2B1. After peaks quality filtering, HNRNPA2B1 was found to bind to 37 293 locations in the genome. We then selected all the genes with significant differential gene expression (adjusted p-value $<$ 0.05) and grouped them into three categories:

1. HNRNPA2B1 binds to the 3' UTR
2. HNRNPA2B1 binds to the 5' UTR
3. HNRNPA2B1 doesn't bind to either UTR

A χ^2 test was performed to test if there were differences in frequencies in these categories 4.32. We found that genes that had HNRNPA2B1 binding to one of their UTR, especially the 5' UTR, were more likely to display a differential expression after silencing HNRNPA2B1. In line, with what has been found in previous studies on specific genes (Yin et al., 2021) (Liu et al.,

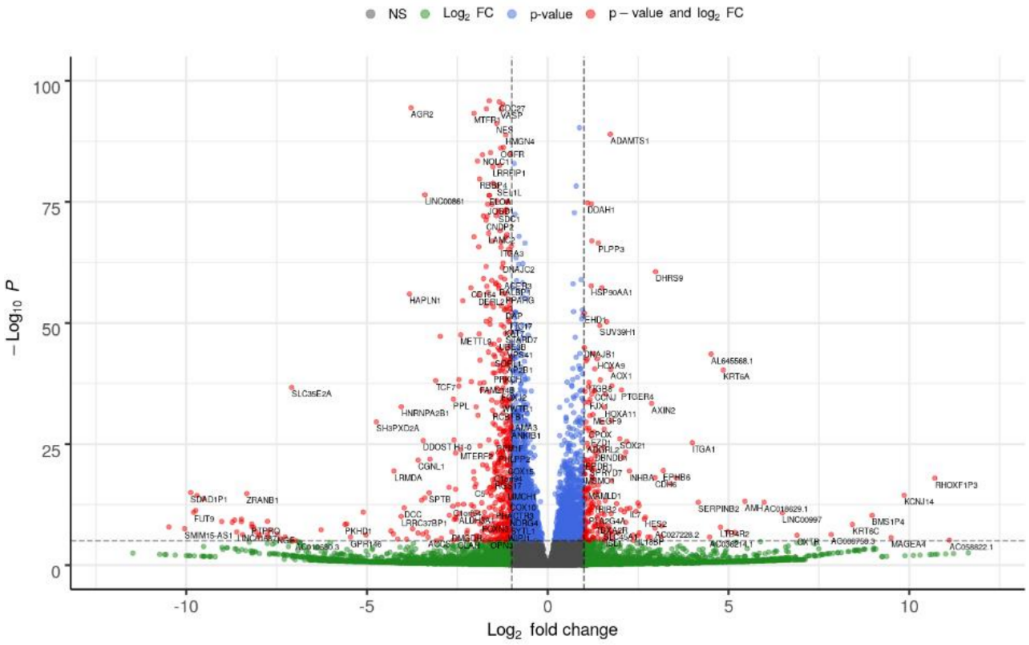

Figure 4.30: Volcano plot of the differential gene expression observed after silencing HNRPA2B1. The horizontal dashed line represents a 0.05 adjusted p-value and the vertical dashed lines represent a log2 fold change of -1 and 1

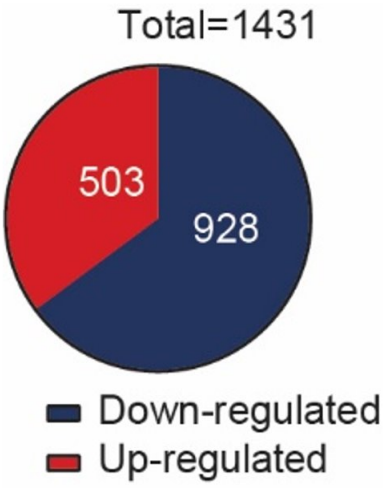

Figure 4.31: Pie chart of the distribution of deregulated genes after HNRNPA2B1 silencing. We observe that more genes are downregulated than upregulated.)

2020)(Stockley et al., 2014), this confirms that HNRNPA2B1 can regulate gene expression.

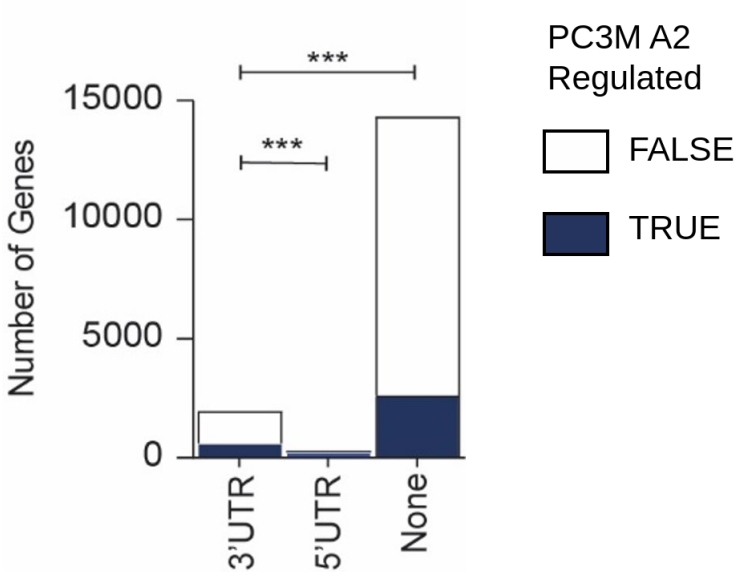

Figure 4.32: χ^2 test showing the proportion of genes that were deregulated or not by silencing HNRNPA2B1 depending on whether HNRNPA2B1 binds to one of their UTR. ***: p-value < 0.05

We then performed GSEA on the set of genes that were significantly differentially expressed and had HNRNPA2B1 binding on at least one of their UTRs. We found that many of the top enriched pathways are related to cancer such as base excision repair, cell cycle or microRNAs in cancer. Interestingly, we also find that the protein processing in ER pathway mentioned in 4.8 is also enriched. Furthermore, eCLIP data shows that HNRNPA2B1 binds to XBP1 in two locations: near the exon4 which get spliced but also on the 3' UTR. As we saw that XBP1 was also differentially expression following HNRNPA2B1 silencing 4.13, this suggests that HNRNPA2B1 might be regulating the unfolded protein response in multiple ways, through splicing of exon 4 of XBP1 but also through regulation of mRNA stability of several genes of this pathway, including XBP1.

While the focus of (Foster et al., 2022) was on how HNRNPA2B1 regulates the unfolded protein response pathway, these additional data show other

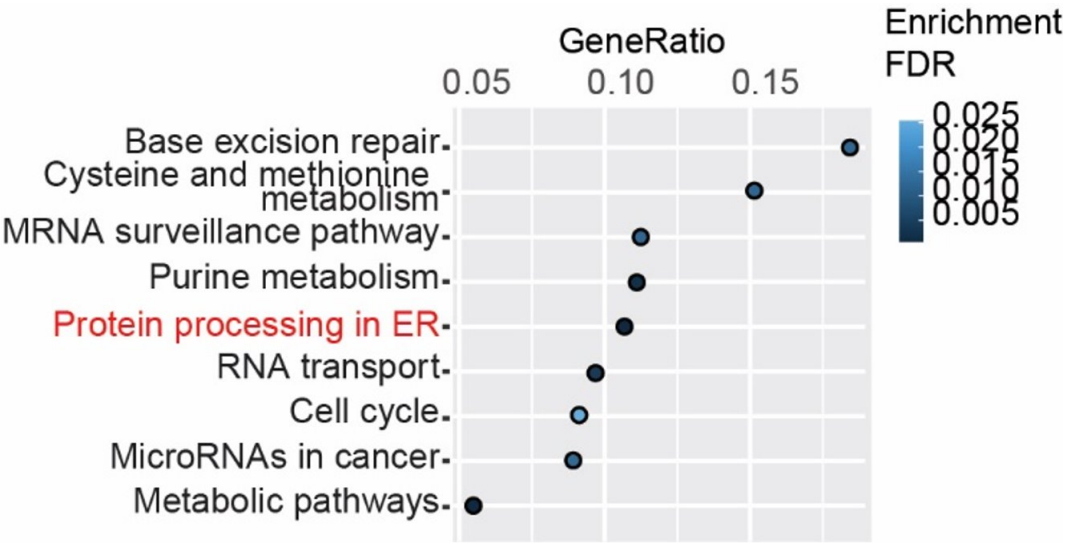

Figure 4.33: Gene Set Enrichment Analysis of KEGG pathways for significantly differentially expressed genes with HNRNPA2B1 binding on at least one of their UTR.

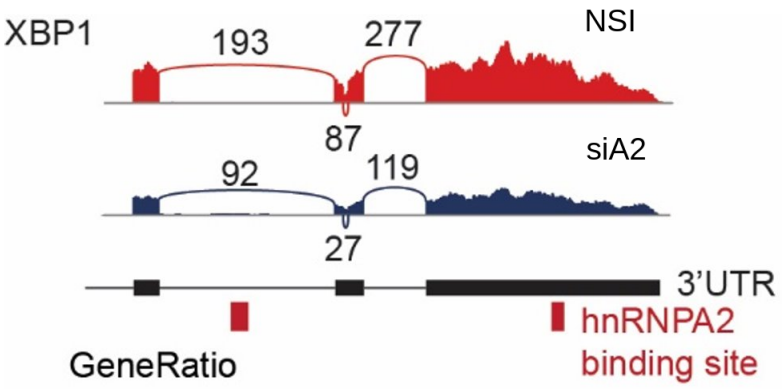

Figure 4.34: Binding sites of HNRNPA2B1 on XBP1.

splicing events and gene regulations controlled by HNRNPA2B1. Since we have shown the detrimental role of high HNRNPA2B1 in prostate cancer, it is likely that this effect occurs through different pathways and mechanisms, and we provide hereby a list of genes that could be potential targets to investigate further in relation to HNRNPA2B1.

4.12 Conclusion

In this chapter, we detailed how PITsuite was used to understand the impact of HNRNPA2B1 in cancer. While the RNA-Seq data for this project had been analysed previously, it didn't lead to any significant new findings. However, by integrating the RNA-Seq data with mass spectrometry and public data through PIT, we identified pathways that were affected by silencing of HNRNAPA2B1. Indeed, the protein processing in the endoplasmic reticulum pathway was flagged as affected both at the RNA and protein level as well as in the public datasets, which allowed us to focus on this particular pathway and try to understand through which mechanisms this was happening. For this, PITgui also considerably facilitated the research, as it allowed meaningful interaction with the output in order to find results of interest without having to write code. In addition, rendering the results in a graphical way proved helpful, for example in the case of the XBP1 splicing of exon 4 4.14 that was easy to see in the gene browser, or the rendering of differentially expressed genes of the protein processing in endoplasmic reticulum pathway 4.13. This made it obvious that the three stress sensors of this pathway (PERK, ATF6 and IRE1) were differentially expressed, which led our investigation in this direction.

Chapter 5

Conclusion

The aim of this project was to extend the PIT methodology and implement it as a prototype software suite capable of integrating transcriptomics, proteomics and analysing it within the context of public data. A secondary objective was to apply this pipeline to biological research projects.

In chapter 2, we introduced a substantially improved methodology for Proteomics Informed by Transcriptomics (PIT). We showed how this pipeline integrates transcriptomics, proteomics and public data in a modular fashion, allowing users to customise the analyses to their experimental design and research question. This integration of multiple layers of biological information allows findings that would not be possible by using only one layer. For example, using a canonical database for mass spectrometry doesn't allow the discovery of novel proteins, whereas this becomes possible when using a sample specific RNA derived database. Compared to previous versions of PIT, major new capabilities were added such as support for RNA and protein quantification, moving beyond PIT's initial ability to only return information about the presence or absence of a transcript or protein. Other major features included support for post translation analysis, alternative splicing or proteome reconstruction. A particular emphasis was also put on broadening the applications in which PIT can prove helpful. For this, we added support for the comparison between samples of different conditions, in order to reveal

differences between them at the RNA or protein level.

However, while applying PIT, we also noticed some limitations. The main one is without a doubt the coverage of mass spectrometry. This was particularly obvious was looking for peptide evidence for mutation, where only a few such instances were found. One of the reasons is that mass spectrometry usually only detects a minority of peptides produced from each protein. This has several causes, such as protein abundance or the physiochemical properties of the peptides. However, as mass spectrometry technology gets better over time, we can expect to see peptide coverage increasing in the future, in which case multi-omics analyses such as PIT will become increasingly relevant.

Additionally, these issues can be mitigated by adopting a different strategy in PIT. For example, we say in chapter 2 in the example of mutations that enriching specific proteins for mass spectrometry or building the peptide database in a more targeted way could increase the chances of finding peptides evidence for mutations on a set of proteins of interest.

In chapter 3, we started with the observation that PIT, like most complex bioinformatics pipelines, produces complex output in text or binary format and is therefore difficult to use directly for users, especially those without programming experience. Based on this observation, we developed PITgui, a graphical user interface that ingests the output of a PIT analyses and displays the results in a user-friendly way. This piece of software, developed in collaboration with biologists who provided feedbacks about the design and functionalities, is installed on the user's computer, is interactive, allowing users to select the results they want through menus, filters and other options. In addition, it provides graphical tools such as charts or a gene browser to display information in a graphical manner, making it easier for users to interpret than observing raw numbers or sequences.

In chapter 4, we used PIT and PITgui to investigate the role of splicing factor HNRNPA2B1 in prostate cancer. Using transcriptomics and proteomics data coming from PC3 cells, as well as clinical data on prostate cancer coming from TCGA, we uncovered how HNRNPA2B1 regulates the unfolded protein response pathway and linked this effect to disease free sur-

vival in prostate cancer patients. During this project, PIT proved itself to be critical to the analysis as it was able to integrate the different layers of information available. For example, visualising the results of the PIT analysis through PITgui that revealed that the UPR pathway was regulated by HNRNPA2B1 at the RNA and protein level.

During this project, in parallel to PITsuite development, it was also used across various projects in collaboration with wet lab scientists. This revealed how PIT could be useful in a wide range of contexts, providing additional insight in researchers' data, with its results being included in a published research article about characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer(Gadaleta et al., 2020).

These collaborations also provided helpful feedback on what was expected for PIT, in terms of analysis features, usability, results visualisation. This back and forth process with biologists contributed in making PITsuite a better tool for the scientific community.

These projects have shown there was great interest in multi-omics analyses thanks to progress in NGS and LC-MS/MS technologies. Yet, the bottleneck was mostly on the data analysis side, with no software existing that would answer common requirements such as comparing samples of different conditions, integration of NGS, LC-MS/MS and public data and visualisation of results in a graphical user interface. PIT was therefore designed to fill this gap and provide these much required capabilities. As such, it was welcomed positively, with many research groups willing to integrate it in their research projects.

Appendix A

HNRNPA2B1 controls an unfolded protein response-related prognostic gene signature in prostate cancer

HNRNPA2B1 controls an unfolded protein response-related prognostic gene signature in prostate cancer

John G Foster[1], Esteban Gea[2,3,$], Mosammat A Labiba[1,2], Chinedu A Anene[3,4], Jacqui Stockley[5,6,#], Celine Philippe[7], Matteo Cereda[8,9], Kevin Rouault-Pierre[7], Hing Leung[5,6] Conrad Bessant[2,10], Prabhakar Rajan[1,10,11,12,13]

1. Centre for Cancer Cell and Molecular Biology, Barts Cancer Institute, Cancer Research UK Barts Centre, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK.
2. School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK.
3. Centre for Cancer Genomics and Computational Biology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK
4. Centre for Cancer Biology and Therapy, School of Applied Science, London South Bank University, 103 Borough Rd, London SE1 0AA, UK
5. Cancer Research UK Beatson Institute, Garscube Estate, Switchback Road, Glasgow, G61 1BD, UK.
6. Institute of Cancer Sciences, University of Glasgow, Garscube Estate, Switchback Road, Glasgow, G61 1BD, UK.
7. Centre for Haemato-Oncology, Barts Cancer Institute, Cancer Research UK Barts Centre, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK.
8. Department of Biosciences, University of Milan, Milan, Italy
9. Italian Institute for Genomic Medicine, c/o IRCCS, Str. Prov.le 142, km 3.95, 10060 Candiolo (TO), Italy
10. Division of Surgery and Interventional Science, University College London, Charles Bell House, 3rd floor, 43-45 Foley Street, London, W1W 7TS
11. The Alan Turing Institute, The British Library, 96 Euston Rd, London, NW1 2DB, UK

12. Department of Urology, Barts Health NHS Trust, The Royal London Hospital, Whitechapel Road, London, E1 1BB, UK

13. Department of Uro-oncology, University College London NHS Foundation Trust, 47 Wimpole Street, London, W1G 8SE, UK

Current Address: Arquer Diagnostics Ltd, North East Business Innovation Centre, Wearfield, Sunderland, SR5 2TA, UK

$ Current Address: Alloy Therapeutics, Inc, 44 Hartwell Ave, Lexington, Massachusetts 02421, USA

Correspondence to: Dr. Prabhakar Rajan, Centre for Cancer Cell and Molecular Biology, Barts Cancer Institute, Cancer Research UK Barts Centre, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, UK. Email: p.rajan@qmul.ac.uk

Running Title: HNRNPA2B1 controls a UPR-related prognostic gene signature

Abstract

HNRNPA2B1 is associated with prostate cancer (PC) disease aggressiveness and underlies pro-tumourigenic cellular stress responses. By analysing >500 PC transcriptomes, we reveal that *HNRNPA2B1* over-expression is associated with poor patient prognosis and stress response pathways. These include the "*protein processing in the endoplasmic reticulum*" (ER) pathway, which incorporates the unfolded protein response (UPR). By RNA-sequencing of HNRNPA2B1-depleted cells PC cells, we identified HNRNPA2B1-mediated down-regulation of UPR genes including the master ER-stress sensor *IRE1*, which induces ER proteostasis. Consistent with *IRE1* down-regulation in HNRNPA2B1-depleted cells, we observed reduced splicing of the IRE1-target and key UPR effector XBP1s. Furthermore, HNRNPA2B1 depletion up-regulates expression of the IRE1-dependent decay (RIDD) target gene *BLOC1S1*, which is degraded by activated IRE1. We identify a HNRNPA2B1-IRE1-XBP1-controlled four gene prognostic biomarker signature (HIX) which classifies a subgroup of primary PC patients at high risk of disease relapse. Pharmacological targeting of IRE1 attenuated HNRNAPA2-driven PC cell growth. Taken together, our data reveal a putative novel mechanism of UPR activation in PC by HNRNPA2B1, which may promote an IRE1-dependent yet potentially-targetable recurrent disease phenotype.

Keywords: prostate cancer, HNRNPA2B1, UPR, XBP1, IRE1

Introduction

The *HNRNPA2B1* gene codes for two protein isoforms, A2 and B1, which are members of the heterogeneous nuclear ribonuclear protein (HNRNP) family of RNA-binding proteins (RBPs) (Liu & Shi 2021). HNRNPA2B1 modulates cellular phenotypes in disease via multiple different RNA processing functions including alternative pre-mRNA splicing (Li et al 2017) and mRNA stability (Martinez et al 2016). In cancer, HNRNPA2B1 can stabilise (Fahling et al 2006, Stockley et al 2014) or destabilise (Zuccotti et al 2014) mRNAs or control oncogenic splicing switches during tumorigenesis (Clower et al 2010, David et al 2010).

Rapid cellular proliferation during tumorigenesis requires an increased rate of protein synthesis (Lee et al 2021), however a limited oxygen and nutrient supply disrupts proteostasis and causes oxidative stress (Bartoszewska & Collawn 2020). An early cellular response to stress is the stalling of mRNA translation and aggregation of pre-initiation translation complexes into stress granules (Marcelo et al 2021) which recruit RBPs including EWSR1, HNRNPA0, HNRNPA1 and HNRNPA2B1 (Jiang et al 2021, Wolozin & Ivanov 2019). Recent studies have identified HNRNPA2B1 cytoplasmic to nuclear translocation in low oxygen conditions, and its association with the polysome, which contains proteins involved in translation, and regulates proteostasis (Ho et al 2020, Yao et al 2013).

Prolonged stress-induced disruption of cellular proteostasis can lead to increased demand on the protein folding machinery of the endoplasmic reticulum (ER) (Rzymski et al 2010), causing protein re-folding, or destruction of terminally misfolded proteins. ER stress triggers altered unfolded protein response (UPR) gene expression profiles via activation of transcription factor sensors including XBP1, ATF4, and nATF6, which control the three key signalling branches of the UPR (Han & Kaufman 2017). Sustained UPR activation leads to increased tumorgenicity, metastatic potential, and therapy resistance of cancer cells (Cubillos-Ruiz et al 2017). In patients, UPR pathway genes are up-regulated (Han & Kaufman 2017), and the transcriptional targets of XBP1, ATF4 and nATF6 are associated with poor survival (Pallmann et al 2019, Sheng et al 2019).

Prostate cancer (PC) is the commonest male-specific cancer and leading male-specific cause of cancer death (Rebello et al 2021). In PC, proteostasis is disrupted (Bouchard et al 2018), and all three branches of the UPR are activated (Pachikov et al 2021, Pallmann et al 2019, Sheng et al 2019). IRE-1-XBP1 activation leads to initiation of c-MYC dependent transcription and is associated with poor patient prognosis (Sheng et al 2019). In light of evidence implicating HNRNPA2B1 in PC (Stockley et al 2014) and cellular stress (Ho et al 2020, Wolozin & Ivanov 2019, Yao et al 2013), we hypothesised that HNRNPA2B1 may control several stress response pathways including UPR in PC. We reveal for the first time that HNRNPA2B1 regulates expression of UPR pathway genes including *IRE1*, mediates non-canonical splicing of XBP1 mRNA, and controls a gene signature of IRE1-XBP1 activation that is associated with poor PC patient prognosis.

Results

HNRNPA2B1 overexpression is associated with poor patient prognosis and cellular stress pathways in prostate cancer

We have previously shown that HNRNPA2B1 protein expression is specifically up-regulated in patients with aggressive prostate cancer (PC) (Stockley et al 2014). To validate these findings, we explored *HNRNPA2B1* expression in RNA sequencing (RNA-Seq) data from primary prostate tumours (n=491) and adjacent benign prostate tissue (n=52) (Sanchez-Vega et al 2018). *HNRNPA2B1* mRNA expression was significantly higher in tumours compared to adjacent benign prostate tissue (Fig. 1A). To determine whether high expression of *HNRNPA2B1* is associated with poor patient prognosis, we stratified tumours into two groups based on the normalized expression levels of *HNRNPA2B1*, with high expression considered the top 25% of the distribution across samples, and the rest of samples considered as low expression. High expression of *HNRNPA2B1* was associated with a statistically significant reduction in patient survival, as compared with patients with low *HNRNPA2B1* expression (Fig. 1B).

Given the previously established roles for *HNRNPA2B1* in the hypoxic response (Ho et al 2020, Yao et al 2013) and stress granule formation (Wolozin & Ivanov 2019), we wished to determine the most significant cellular stress pathways regulated by HNRNPA2B1 in PC. Firstly, we performed Gene Set Enrichment Class Analysis (GSECA) (Lauria et al 2020) on RNA-seq datasets from primary (n=491) (Hoadley et al 2018) and metastatic PC (CRPC) (n=208) (Abida et al 2019). We compared KEGG stress pathway representation in tumours with high *HNRNPA2B1* expression compared with low expression. In primary PC, we found that the top stress pathways associated with high expression of *HNRNPA2B1* included the "*Proteasome"* and "*HIF1 signaling pathway"* (Fig. 1C). In metastatic PC, top pathways associated with high expression of *HNRNPA2B1* included "*Protein processing in endoplasmic reticulum*", "*Autophagy*", and diseases with a misfolded protein component (Fig. 1D).

To validate these findings, we performed RNA-Seq of PC3M cells treated with either with a single siRNA duplex targeting *HNRNPA2B1* or a non-targeting control. We observed a statistically-significant reduction in *HNRNPA2B1* gene expression following siRNA treatment as compared with the control (Log_2 fold change = -4.05 adjusted p-value<0.001, Supplementary Table 5). Subsequently, we performed gene set enrichment analysis (GSEA) using all KEGG pathways to identify top biological processes enriched upon *HNRNPA2B1* depletion. Consistent with the association of HNRNPA2B1 with cellular stress pathways in PC patients, the KEGG stress pathway "*Protein processing in endoplasmic reticulum*" was the most significantly enriched pathway (Fig. 1E). Within this pathway, *HNRNPA2B1* depletion led to down-regulated expression of *PERK, ATF6* and *IRE1*, which encode for the three master ER-stress sensors mediating three key signaling branches of the UPR (Luo & Lee 2013) (Fig. 1F).

Taken together, these data in PC patients and cell lines indicates that HNRNPA2B1 regulates cellular stress pathways, with the most significant pathway being "*Protein processing in the endoplasmic reticulum*" in PC cells incorporating UPR genes.

HNRNPA2B1 affects processing of IRE1 target mRNAs

To shed light on a putative mechanism of HNRNPA2B1-mediated UPR gene expression, we focussed on the IRE1-XBP1 signalling branch, considering its association with PC disease recurrence (Sheng et al 2019). XBP1 transcriptional activation requires non-canonical cytoplasmic splicing of XBP1u mRNA to produce the transcriptionally active XBP1s via removal of a variable 26 nucleotide sequence in exon 4 by IRE1 nuclease activity (Calfon et al 2002, Uemura et al 2009) (Fig. 2A). We hypothesised that HNRNPA2B1 may regulate UPR genes via XBP1 splicing. To test this, we used established RT-PCR based splicing assays (Savic et al 2014) to measure the percentage expression of activated XBP1s compared with XBP1u (Fig. 2A). Following treatment of PC3M cells with the UPR activator Thapsigargin (da Silva et al 2020); we observed a statistically significant increase in XBP1s splicing, compared to controls (Fig. 2B). Conversely, following

HNRNPA2B1 protein depletion in PC3M cells using two independent siRNA duplexes (Fig. 2C); we observed a statistically significant decrease in XBP1s splicing compared with controls (Fig. 2D). These data demonstrate that HNRNPA2B1 promotes the non-conventional splicing of XBP1u to XBP1s.

IRE1 also degrades several mRNAs, including the BLOC1S1 mRNA, which encodes a regulator of lysosomal function, as part of the regulated IRE1-dependent decay (RIDD) pathway during ER stress (Chalmers et al 2019, Lhomond et al 2018). We wished to determine whether HNRNPA2B1 could also affect the RIDD pathway by exploring its impact on *BLOC1S1* expression. Following treatment of cells with the UPR activator Thapsigargin, we observed a statistically significant reduction in *BLOC1S* expression (Fig. 2E). Concordant with the impact of HNRNPA2B1 on XBP1 splicing, we observed a statistically-significant increase in *BLOC1S1* expression upon HNRNPA2B1 depletion (Fig. 2F). These data indicate that HNRNPA2B1 may affect multiple IRE1-dependent gene regulatory functions in PC cells.

HNRNPA2B1-IRE1-XBP1 co-regulated genes represent a prognostic biomarker signature in primary PC and reveal a potential therapeutic target

Since high *HNRNPA2B1* expression is associated with poor PC patient prognosis (Fig. 1B), we hypothesised that this phenomenon may be mediated, in part, by HNRNPA2B1-dependent IRE1-XBP1-related gene expression. To test this, we utilised previously published RNA-Seq data from PC cells depleted of XBP1 or treated with the IRE1 inhibitor MKC8866 (Sheng et al 2019). To identify protein-coding genes co-regulated by XBP1, IRE1, and HNRNPA2B1, we overlapped lists of differentially expressed protein-coding genes in the three datasets (Fig. 3A). We identified a total of 20 HNRNPA2B1-IRE1-XBP1 co-regulated protein-coding genes.

To determine if these 20 genes, or a subset thereof, were associated with disease recurrence, we performed elastic net regression using expression values of these genes in the TCGA cohort and time-to-event data (Fig. 3B). We applied elastic net selection (Fig. 3B, left panel, Supplementary

Table 8) at lambda with the least mean cross-validation error and coefficient >0.00025 or <-0.00025. We identified four HNRNPA2B1-IRE1-XBP1 (HIX)-regulated genes (*FKBP14*, *TMEM39A*, *BET1*, and *CDC6*) as the best predictors of disease relapse (Fig 3B, right panel). Using multivariable Cox regression coefficient-derived patient risk scores for the four genes (see Materials and Methods), we stratified TCGA patients into two risk groups (low risk = <1st-3rd quartile, high risk = >3rd quartile) (Fig. 3C, top panel). The high risk group was significantly more likely to relapse compared with the low risk group (Fig. 3C, bottom panel).

To validate these findings, we applied risk score calculations to an independent microarray-derived dataset (MSKCC) (Fig. 3D, top panel). Consistently, the high risk group was significantly more likely to relapse compared with the low risk group (Fig. 3D, bottom panel). Taken together, these data indicate that IRE1-XBP1-mediated gene activation may underlie the recurrent disease phenotype associated with HNRNPA2B1.

To determine whether the IRE-XBP1 signalling branch of the UPR might represent a potential therapeutic target for HNRNPA2B1 over-expressing PC, we firstly transiently ectopically expressed HNRNPA2, the predominant protein isoform encoded by *HNRNPA2B1* (Fig. 2C) in PC3M cells (Fig. 3E, top panel). Consistent with previously published data (Stockley et al 2014), we observed a statistically significant increase in cell growth following ectopic HNRNPA2 expression compared with controls (Fig. 3E, bottom panel). Subsequently, we treated HNRNPA2 overexpressing cells or controls with the IRE1 inhibitor STF083010 (Dong et al 2021). Following STF083010 treatment at 50 and 100μM doses, the effect of ectopic HNRNPA2B1 expression on cell growth was attenuated (Fig. 3E, bottom panel). These data suggest that IRE1 may be a potential therapeutic target in HNRNPA2B1 overexpressing PC tumours.

Discussion

In this study, we reveal that high expression of HNRNPA2B1 in primary PC is associated with early disease recurrence. Our data indicate that this effect may mediated by HNRNPA2B1-controlled unfolded protein response (UPR) pathway-related genes via the major ER stress sensor IRE1. We show that HNRNPA2B1 controls IRE1-dependent XBP1 splicing and a subset of IRE1-XBP1 co-regulated genes classifies a subgroup of PC patients at high risk of disease relapse. Finally, we reveal that treatment with an IRE1 inhibitor attenuates HNRNPA2-driven PC cell growth, highlighting a novel line of therapy.

HNRNPA2B1 is known to play an important role in the formation of stress granules (Jiang et al 2021), and hypoxic adaptation (Ho et al 2020, Yao et al 2013). Here, we identify a link between *HNRNPA2B1* expression and several stress response pathways in primary and metastatic PC patients. In primary PC, we find that *HNRNPA2B1* largely is associated with metabolic stress pathways, whereas in metastatic PC it is associated with proteostasis stress such as "*Protein processing in endoplasmic reticulum*". In tumourigenesis, sustained metabolic stresses, such as those caused by hypoxia, can disrupt proteostasis, induce ER stress, and activate the UPR (Ottens et al 2021). Hence, the association of *HNRNPA2B1* with "*Protein processing in endoplasmic reticulum*" in late-stage metastatic PC may be as a result of disrupted proteostasis acquired early in the disease course in a subset of patients with aggressive primary tumours over-expressing HNRNPA2B1.

Next, we reveal that HNRNPA2B1 regulates UPR gene expression including the master ER-stress sensor *IRE1*. Specifically, our findings implicates HNRNPA2B1 in IRE1-dependent processes of XBP1 splicing and RIDD activation. These two processes are mechanistically distinct, requiring dimerization or oligomerization of a phosphorylated version of the ribonuclease IRE1, respectively (Coelho & Domingos 2014). Given that depletion of HNRNPA2B1 increased expression of both *XBP1u* and *BLOC1S1*, we might speculate that HNRNPA2B1 may act downstream of IRE1 to regulate these mutually-exclusive events. Based on its known role in mRNA processing (Fahling et al 2006, Stockley et al 2014), it is possible that HNRNPA2B1 is either stabilises and/or facilitates transport of XBP1 and BLOC1S1 mRNAs to IRE1 at the ER membrane.

To identify a HNRNPA2B1-IRE1-XBP1-controlled prognostic biomarker signature (HIX) in PC patients, we initially used previously published RNA-seq datasets from PC cells treated with either the IRE1 inhibitor MKC8866 or depleted of XBP1 expression (Sheng et al 2019). Interestingly, both XBP1 siRNA and IRE1 inhibition regulate MYC protein expression and induce expression of several MYC target genes (Sheng et al 2019). Since MYC promotes the transcription of *HNRNPA2B1* (David et al 2010), we might speculate that HNRNPA2B1 is a component of the MYC-driven UPR activation.

XBP1 underpins several cancer hallmarks: XBP1 increases the key fatty acid metabolic enzyme SCD1 expression in MYC-driven cancers (Xie et al 2018). XBP1-mediated transcription of SNAI1, SNAI2, ZEB2, and TCF3 can mediate epithelial to mesenchymal transition and invasion (Cuevas et al 2017). By formation of a co-transcriptional complex with HIF1, XBP1 can controls angiogenesis (Chen et al 2014). Moreover, inhibitors of the IRE1-XBP1 pathway reduce tumour growth and sensitize cells to chemotherapy in pre-clinical models (Logue et al 2018, Sheng et al 2019). Here, we show that the known impact of XBP1 on PC cell growth and disease recurrence (Sheng et al 2019) is influenced by HNRNPA2B1.

Our study has several limitations: Although the novel link between HNRNPA2B1 and UPR was identified in metastatic PC, the HNRNPA2B1-IRE1-XBP1-controlled prognostic biomarker signature (HIX) was only validated in primary PC patients and based on mRNA expression. HNRNPA2B1 regulated *PERK* and *ATF6* as well as *IRE1* expression (Fig. 1F), however our validations focussed exclusively on IRE1-XBP1. Hence, we do not know the impact of HNRNPA2B1 on other UPR pathway branches. The precise molecular mechanisms underlying the HNRNPA2B1-mediated regulation of IRE1 and XBP1 remains unclear and warrants further investigation. Future studies using multiple UPR inhibitors in pre-clinical cancer models are required to determine whether targeting one or more UPR branches has therapeutic efficacy for HNRNPA2B1-overexpressing PC patients.

Materials and Methods

Transcriptomic datasets

Clinical RNA sequencing (RNA-Seq) and microarray data were obtained from cBioPortal (Cerami et al 2012, Gao et al 2013, Sanchez-Vega et al 2018). For primary PC (The Cancer Genome Atlas; TCGA, n=491 samples; Memorial Sloan Kettering Cancer Centre; MSKCC, n=179 samples), from Sanchez-Vega *et al*. (Sanchez-Vega et al 2018) for adjacent benign prostate (TCGA, n=52), and cBioPortal (Cerami et al 2012, Gao et al 2013) for metastatic PC (Stand Up to Cancer; SU2C, n=208 samples). Gene expression values were reported for TCGA as RNA-Seq by Expectation-Maximization (RSEM), for SU2C as Fragments per Kilobase of exon Per Million mapped fragments (FPKM) cohorts, or for MSKCC as $\log_2$ whole transcript mRNA expression. For comparison of normal (TCGA, n=52) and primary PC tissue (TCGA, n=497) RNA-Seq data were obtained from the Broad Institute Genome Data Analysis Center (GDAC) Firehose database (doi:10.7908/C11G0KM9) (Supplementary Table 1). Cell line RNA-Seq data for LNCaP cells treated with siRNA to XBP1 or and IRE1 inhibitor (MKC8866) were obtained from Sheng *et al*. (Sheng et al 2019) and gene expression values reported as Log_2 Fold Change and adjusted p-value.

Survival analysis

Patient samples were stratified into two groups by mRNA expression as follows: low = <1st-3rd quartile and high = >3rd quartile (Supplementary table 1). Kaplan-Meier plots were generated using time to event data (event = disease recurrence) from patient cohorts (TCGA; 487 out of 491 patients) using the *survfit* function of the *survminer* package in R V.4.1.1 and plotted using *ggsurvplot*. Univariable analyses were performed using the *coxph* function of *survminer*.

Gene set enrichment class analysis (GSECA)

A list of 35 gene sets representing stress associated pathways was obtained from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway database https://www.genome.jp/kegg/pathway.html (Supplementary table 2). Patient samples were stratified into two groups by mRNA expression as follows: low = <1st-3rd quartile and high = >3rd quartile. GSECA was performed in R V.4.1.1 as previously described on stratified samples (Lauria et al 2020) considering the 35 KEGG stress associated pathway gene sets. An independent Monte Carlo simulation (1,000 iterations) was performed to determine the success rate (SR) of the association between the two cohorts. (Lauria et al 2020). Gene sets with GSECA association score (GAS) ≤0.05, adjusted p-value≤0.05 and success rate (SR)≥0.7 were considered as significant (Supplementary Table 2).

Cell lines, transfections, and drug treatments

The PC3M cell line was generated as previously described (Pettaway et al 1996) and Short Tandem Repeat (STR) profiling (DDC Medical) used to confirm identity. Cells were maintained at sub-confluency, in RPMI-1640 medium (21875-034, Gibco) containing 2 mM L-glutamine, supplemented with 10% foetal calf serum (FCS) (Gibco), 100 units/ml penicillin and 100 µg/ml streptomycin (15140-122, Gibco), and incubated at 37^{o}C, 5% CO_2 in a humidified incubator. Cells were regularly screened for contamination with mycoplasma. DNA and siRNA transfections were performed as detailed in the figure legends using ViaFect (E4981, Promega) and RNAiMax (13778-075, Thermo Fisher Scientific), respectively, according to the manufacturers' instructions (Supplementary Table 3). Cells were treated with IRE1 inhibitor (STF083010), at concentrations indicated in the figure legends, or vehicle control (DMSO).

Antibodies, plasmids, and oligonucleotides

The plasmid pCAGPM-HA-hnRNPA2 (Katoh et al 2011) was a gift from Dr Y. Matsuura (Osaka University, Japan), and pcDNA3.1-HA was a gift from Professor T. Sharp (Queen Mary University of London, UK). The following antibodies were used: anti-HNRNPA2B1 (ab31645, Abcam), anti-actin (A1978, Sigma), anti-mouse IgG HRP-linked (P044701-2, Dako), and anti-rabbit IgG HRP-linked (P044801-2, Dako). The IRE1 inhibitor STF083010 was purchased from Merck Life Science, UK (SML0409). Sequences used to generate siRNA duplexes are as previously described (Stockley et al 2014) or commercially-designed (ON-TARGETPlus, Dharmacon Horizon Discovery) (Supplementary Table 3). Primers for PCR were designed using the National Center for Biotechnology Information (NCBI) Primer-BLAST tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast) with the Ensembl (http://www.ensembl.org) Transcript ID for the principal mRNA isoform and synthesised by Integrated DNA Technologies (Supplementary Table 3). Primers for *in vitro* splicing analysis are as previously published (Savic et al 2014) (Supplementary Table 3).

SDS-PAGE and Western blotting

Whole cell lysis was performed in RIPA (Radio-Immunoprecipitation Assay) buffer for 30 minutes at 4^{o}C. Protein concentration was determined by Bicinchoninic acid (BCA) assay (10678484, Thermo Fisher Scientific) and samples adjusted to the same total protein concentration. Proteins were separated by size by SDS-Polyacrylamide Gel Electrophoresis (SDS–PAGE) on 10% w/v gels and electroblotted onto a Polyvinylidene fluoride or polyvinylidene difluoride (PVDF) membrane (3010040001, Sigma). Luminata Crescendo Western HRP substrate (10776189, Thermo Fisher Scientific) was used for signal detection, and protein bands were visualised on a Chemidoc system (Amersham Imager 600, Amersham). Antibody concentrations were as follows: anti-HNRNPA2B1 (1:1 000), anti-actin (1:100 000); HRP-linked secondaries (1:5 000). Where indicated, densitometry assessments of protein bands were performed using Image Studio Lite

v.5.2 (LI-COR), and signal intensities used to calculate relative normalised fold-change in protein expression (Supplementary Table 4).

RNA-Seq and gene set enrichment analysis

Total RNA was extracted from cells using the QIAgen RNeasy mini kit (74004, QIAgen), and treated with DNAse I (AMPD1, Sigma) to exclude genomic contamination. Libraries were generated using the TruSeq RNA Library Prep Kit v2 (RS-122-2001, Illumina) and 75bp paired end sequencing performed to 30M read depth using the NextSeq 500 (Illumina). Reads were aligned to the genome (hg38) using STAR v2.7.3a in dual pass mode. Transcripts were assembled and quantified in Transcripts Per Million (TPM) using Stringtie v2.1.1 (Pertea et al 2015). Read normalisation and differential gene expression analysis was performed using DESeq2 v1.34.0 in R V.4.1.1 (Supplementary Table 5). Enrichment of KEGG pathways amongst differentially-expressed genes with $\log_2$ fold change of <-0.5 or >0.5 at $p<0.05$ significance was performed in R V.4.1.1 using the *enrichKEGG* function of the *clusterProfiler* package (Wu et al 2021) in R and plotted with *dotplot* (Supplementary Table 5). Raw data have been deposited at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE198261, and all details are Minimum Information About a Microarray Experiment (MIAME) compliant.

Quantitative Reverse Transcription PCR

Total RNA was extracted from cells using TRI Reagent Solution (M9738, Thermo Fisher Scientific), and reverse transcribed to cDNA using the High Capacity cDNA Reverse Transcription Kit (4368814, Applied Biosystems). cDNA (20ng per condition) was combined with forward and reverse primers (Supplementary Table 3) and the Luna Universal qPCR Master Mix master mix (M3003, NEB) containing SYBR green and ROX passive dye to a final 10ul reaction volume. Binding of SYBR green to DNA was analysed in a QuantStudio 5 Real-Time PCR system (Thermo

Fisher Scientific). Reaction conditions were as follows: Initial denaturation at 95°C for 10 minutes, 40 cycles of denaturation for 15 seconds at 95°C, plus annealing, extension, and signal capture at 60°C for 1 minute. The 2-ΔΔCT method was used to determine relative gene expression using the geometric mean expression of two validated endogenous control genes (*ACTB* and *B2M*) (Supplementary Table 6).

XBP1 Splicing Assays

Total RNA was extracted from cells using TRI Reagent Solution (M9738, Thermo Fisher Scientific), and reverse transcribed to cDNA using the High Capacity cDNA Reverse Transcription Kit (4368814, Applied Biosystems). cDNAs (20ng per condition) were combined with primers flanking the variable exonic region of XBP1 (Savic et al 2014) (Supplementary Table 3), dNTPs and Taq Polymerase (NEB, M0273) in standard reaction buffer to a final 10ul reaction volume. Reactions were performed in a ProFlex thermocycler (Applied Biosystems) as follows: Initial denaturation at 95°C for 30 seconds, 30 cycles of denaturation for 15 seconds at 95°C, plus annealing at 52°C for 30 seconds, and extension at 68°C for 1 minute; followed by a final extension at 68°C for 5 minutes. PCR products were resolved, detected and quantified by capillary gel electrophoresis (QIAxcel, QIAgen) (Supplementary Table 7).

Derivation and validation of a prognostic biomarker panel

To identify the combination of genes which are the strongest predictors of PC recurrence, the *glmnet* package (Friedman et al 2010) in R V.4.1.1 was used to fit gene expression to time-to-event data in the TCGA (derivation) cohort using cox regression with an $\alpha = 0.2$using a coefficient cut off of >0.00025 or <-0.00025 at λ minimum. To obtain coefficients representing the relative contributions of the selected genes to the prognostic value of the signature, multivariable analysis was performed using time-to event data and grouped expression of each of the four signature

genes (low = <1st-3rd quartile and high = >3rd quartile), by the *coxph* function of *survminer* package. Coefficients for each gene were obtained from the high expression group (Supplementary table 8). Next, a risk score (i) for each patient was derived from the coefficients of the multivariable Cox PH model as follows: $(i) = \sum_{j=1}^{n} \alpha j * ej,$

where αj is the scaled j gene expression value with ej coefficient in the derivation multivariable model (Royston & Altman 2013). Risk group cut-offs were defined based upon quartiles of gene signature score in TCGA data (low = <1st-3rd quartile, high = >3rd quartile).

Kaplan-Meier plots were generated using time to event data (event = disease recurrence) from patient cohorts using the *survfit* function of the *survminer* package in R V.4.1.1 and plotted using *ggsurvplot* (Supplementary table 8). Univariable analyses were performed using the *coxph* function of *survminer* to compare patients with low and high risk scores. To validate the model, risk scores calculated using the coefficients obtained from the derivation cohort were applied to scaled gene expression values from the validation cohort, and Kaplan Meier plots generated stratified by risk scores (low = <1st-3rd quartile, high = >3rd quartile).

Cell growth assay

Cells (n = 2000) were seeded into each well of a 96-well plate and grown to ~20–30% confluence prior to transfection with DNA as indicated in the figure legends. After 72 hours, (3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyltetrazolium Bromide) (MTT) (M6494, Thermo Fisher Scientific) was added to each well to a final concentration of 0.67 mg/ml and incubated at 37oC, 5% CO2 in a humidified incubator for 2 h. MTT reagent was then removed, and 100µl dimethyl sulfoxide (DMSO) (10213810, Thermo Fisher Scientific) added to each well, and the plate was agitated at room temperature for 15 minutes. Absorbance was measured at 560nm and 630nm (SpectraMax Plus384 Absorbance Microplate Reader, Molecular Devices), and normalized by subtracting the 630nm value from the 560nm value. Percentage viability (%) was calculated as: the treatment

absorbance divided by the DMSO control absorbance. All data were normalized to a vector only control (Supplementary Table 9).

Data Availability

RNA-Seq data from this publication have been deposited to Gene Expression Omnibus and assigned the identifier accession number GSE198261.

Acknowledgements

We would like to thank Y. Matsuura (RIMD, Japan) and T. Sharp (QMUL, UK) for providing plasmid DNA vectors used in this study. We are grateful to the P. Herzyk, J. Galbraith, G. Hamilton, and M. Mudaliar (University of Glasgow Polyomics, UK) as well as A. Hedley and G. Kalna (CR-UK Beatson Institute, UK) for assistance with RNA-seq and bioinformatics. We would also like to thank P. Grevitt (QMUL, UK) and P. Baptista-Ribeiro (QMUL, UK) for their critical appraisal of earlier versions of the manuscript. The research performed in this study was funded by the Royal College of Surgeons of England/Cancer Research UK Clinician Scientist Fellowship in Surgery (C19198/A15339 to PR), The Urology Foundation and John Black Charitable Foundation (to PR), Barts Charity (MGU0533 to PR) and Orchid Charity (to PR).

Conflict of Interest

The authors declare no conflicts of interest

References

Abida W, Cyrta J, Heller G, Prandi D, Armenia J, Coleman I, Cieslik M, Benelli M, Robinson D, Van Allen EM, et al. 2019. Genomic correlates of clinical outcome in advanced prostate cancer. Proc Natl Acad Sci U S A. 116(23):11428-11436. doi:10.1073/pnas.1902651116

Bartoszewska S, Collawn JF. 2020. Unfolded protein response (upr) integrated signaling networks determine cell fate during hypoxia. Cell Mol Biol Lett. 25:18. doi:10.1186/s11658-020-00212-1

Bouchard JJ, Otero JH, Scott DC, Szulc E, Martin EW, Sabri N, Granata D, Marzahn MR, Lindorff-Larsen K, Salvatella X, et al. 2018. Cancer mutations of the tumor suppressor spop disrupt the formation of active, phase-separated compartments. Mol Cell. 72(1):19-36 e18. doi:10.1016/j.molcel.2018.08.027

Calfon M, Zeng H, Urano F, Till JH, Hubbard SR, Harding HP, Clark SG, Ron D. 2002. Ire1 couples endoplasmic reticulum load to secretory capacity by processing the xbp-1 mrna. Nature. 415(6867):92-96. doi:10.1038/415092a

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. 2012. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2(5):401-404. doi:10.1158/2159-8290.CD-12-0095

Chalmers F, Mogre S, Son J, Blazanin N, Glick AB. 2019. The multiple roles of the unfolded protein response regulator ire1alpha in cancer. Mol Carcinog. 58(9):1623-1630. doi:10.1002/mc.23031

Chen X, Iliopoulos D, Zhang Q, Tang Q, Greenblatt MB, Hatziapostolou M, Lim E, Tam WL, Ni M, Chen Y, et al. 2014. Xbp1 promotes triple-negative breast cancer by controlling the hif1alpha pathway. Nature. 508(7494):103-107. doi:10.1038/nature13119

Clower CV, Chatterjee D, Wang Z, Cantley LC, Vander Heiden MG, Krainer AR. 2010. The alternative splicing repressors hnrnp a1/a2 and ptb influence pyruvate kinase isoform expression and cell metabolism. Proc Natl Acad Sci U S A. 107(5):1894-1899. doi:10.1073/pnas.0914845107

Coelho DS, Domingos PM. 2014. Physiological roles of regulated ire1 dependent decay. Front Genet. 5:76. doi:10.3389/fgene.2014.00076

Cubillos-Ruiz JR, Bettigole SE, Glimcher LH. 2017. Tumorigenic and immunosuppressive effects of endoplasmic reticulum stress in cancer. Cell. 168(4):692-706. doi:10.1016/j.cell.2016.12.004

Cuevas EP, Eraso P, Mazon MJ, Santos V, Moreno-Bueno G, Cano A, Portillo F. 2017. Loxl2 drives epithelial-mesenchymal transition via activation of ire1-xbp1 signalling pathway. Sci Rep. 7:44988. doi:10.1038/srep44988

da Silva DC, Valentao P, Andrade PB, Pereira DM. 2020. Endoplasmic reticulum stress signaling in cancer and neurodegenerative disorders: Tools and strategies to understand its complexity. Pharmacol Res. 155:104702. doi:10.1016/j.phrs.2020.104702

David CJ, Chen M, Assanah M, Canoll P, Manley JL. 2010. Hnrnp proteins controlled by c-myc deregulate pyruvate kinase mrna splicing in cancer. Nature. 463(7279):364-368. doi:10.1038/nature08697

Dong L, Tan CW, Feng PJ, Liu FB, Liu DX, Zhou JJ, Chen Y, Yang XX, Zhu YH, Zhu ZQ. 2021. Activation of trem-1 induces endoplasmic reticulum stress through ire-1alpha/xbp-1s pathway in murine macrophages. Mol Immunol. 135:294-303. doi:10.1016/j.molimm.2021.04.023

Fahling M, Mrowka R, Steege A, Martinka P, Persson PB, Thiele BJ. 2006. Heterogeneous nuclear ribonucleoprotein-a2/b1 modulate collagen prolyl 4-hydroxylase, alpha (i) mrna stability. J Biol Chem. 281(14):9279-9286. doi:10.1074/jbc.M510925200

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 33(1):1-22.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Sci Signal. 6(269):pl1. doi:10.1126/scisignal.2004088

Han J, Kaufman RJ. 2017. Physiological/pathological ramifications of transcription factors in the unfolded protein response. Genes Dev. 31(14):1417-1438. doi:10.1101/gad.297374.117

Ho JJD, Balukoff NC, Theodoridis PR, Wang M, Krieger JR, Schatz JH, Lee S. 2020. A network of rna-binding proteins controls translation efficiency to activate anaerobic metabolism. Nat Commun. 11(1):2677. doi:10.1038/s41467-020-16504-1

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. 2018. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 173(2):291-304 e296. doi:10.1016/j.cell.2018.03.022

Jiang L, Lin W, Zhang C, Ash PEA, Verma M, Kwan J, van Vliet E, Yang Z, Cruz AL, Boudeau S, et al. 2021. Interaction of tau with hnrnpa2b1 and n(6)-methyladenosine rna mediates the progression of tauopathy. Mol Cell. doi:10.1016/j.molcel.2021.07.038

Katoh H, Mori Y, Kambara H, Abe T, Fukuhara T, Morita E, Moriishi K, Kamitani W, Matsuura Y. 2011. Heterogeneous nuclear ribonucleoprotein a2 participates in the replication of japanese encephalitis virus through an interaction with viral proteins and rna. J Virol. 85(21):10976-10988. doi:10.1128/JVI.00846-11

Lauria A, Peirone S, Giudice MD, Priante F, Rajan P, Caselle M, Oliviero S, Cereda M. 2020. Identification of altered biological processes in heterogeneous rna-sequencing data by discretization of expression profiles. Nucleic Acids Res. 48(4):1730-1747. doi:10.1093/nar/gkz1208

Lee LJ, Papadopoli D, Jewer M, Del Rincon S, Topisirovic I, Lawrence MG, Postovit LM. 2021. Cancer plasticity: The role of mrna translation. Trends Cancer. 7(2):134-145. doi:10.1016/j.trecan.2020.09.005

Lhomond S, Avril T, Dejeans N, Voutetakis K, Doultsinos D, McMahon M, Pineau R, Obacz J, Papadodima O, Jouan F, et al. 2018. Dual ire1 rnase functions dictate glioblastoma development. EMBO Mol Med. 10(3) doi:10.15252/emmm.201707929

Li Y, Sahni N, Pancsa R, McGrail DJ, Xu J, Hua X, Coulombe-Huntington J, Ryan M, Tychhon B, Sudhakar D, et al. 2017. Revealing the determinants of widespread alternative splicing perturbation in cancer. Cell Rep. 21(3):798-812. doi:10.1016/j.celrep.2017.09.071

Liu Y, Shi SL. 2021. The roles of hnrnp a2/b1 in rna biology and disease. Wiley Interdiscip Rev RNA. 12(2):e1612. doi:10.1002/wrna.1612

Logue SE, McGrath EP, Cleary P, Greene S, Mnich K, Almanza A, Chevet E, Dwyer RM, Oommen A, Legembre P, et al. 2018. Inhibition of ire1 rnase activity modulates the tumor cell secretome and enhances response to chemotherapy. Nat Commun. 9(1):3267. doi:10.1038/s41467-018-05763-8

Luo B, Lee AS. 2013. The critical roles of endoplasmic reticulum chaperones and unfolded protein response in tumorigenesis and anticancer therapies. Oncogene. 32(7):805-818. doi:10.1038/onc.2012.130

Marcelo A, Koppenol R, de Almeida LP, Matos CA, Nobrega C. 2021. Stress granules, rna-binding proteins and polyglutamine diseases: Too much aggregation? Cell Death Dis. 12(6):592. doi:10.1038/s41419-021-03873-8

Martinez FJ, Pratt GA, Van Nostrand EL, Batra R, Huelga SC, Kapeli K, Freese P, Chun SJ, Ling K, Gelboin-Burkhart C, et al. 2016. Protein-rna networks regulated by normal and als-associated mutant hnrnpa2b1 in the nervous system. Neuron. 92(4):780-795. doi:10.1016/j.neuron.2016.09.050

Ottens F, Franz A, Hoppe T. 2021. Build-ups and break-downs: Metabolism impacts on proteostasis and aging. Cell Death Differ. 28(2):505-521. doi:10.1038/s41418-020-00682-y

Pachikov AN, Gough RR, Christy CE, Morris ME, Casey CA, LaGrange CA, Bhat G, Kubyshkin AV, Fomochkina, II, Zyablitskaya EY, et al. 2021. The non-canonical mechanism of er stress-mediated progression of prostate cancer. J Exp Clin Cancer Res. 40(1):289. doi:10.1186/s13046-021-02066-7

Pallmann N, Livgard M, Tesikova M, Zeynep Nenseth H, Akkus E, Sikkeland J, Jin Y, Koc D, Kuzu OF, Pradhan M, et al. 2019. Regulation of the unfolded protein response through atf4 and fam129a in prostate cancer. Oncogene. 38(35):6301-6318. doi:10.1038/s41388-019-0879-2

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. Nat Biotechnol. 33(3):290-295. doi:10.1038/nbt.3122

Pettaway CA, Pathak S, Greene G, Ramirez E, Wilson MR, Killion JJ, Fidler IJ. 1996. Selection of highly metastatic variants of different human prostatic carcinomas using orthotopic implantation in nude mice. Clin Cancer Res. 2(9):1627-1636.

Rebello RJ, Oing C, Knudsen KE, Loeb S, Johnson DC, Reiter RE, Gillessen S, Van der Kwast T, Bristow RG. 2021. Prostate cancer. Nat Rev Dis Primers. 7(1):9. doi:10.1038/s41572-020-00243-0

Royston P, Altman DG. 2013. External validation of a cox prognostic model: Principles and methods. BMC Med Res Methodol. 13:33. doi:10.1186/1471-2288-13-33

Rzymski T, Milani M, Pike L, Buffa F, Mellor HR, Winchester L, Pires I, Hammond E, Ragoussis I, Harris AL. 2010. Regulation of autophagy by atf4 in response to severe hypoxia. Oncogene. 29(31):4424-4435. doi:10.1038/onc.2010.191

Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. 2018. Oncogenic signaling pathways in the cancer genome atlas. Cell. 173(2):321-337 e310. doi:10.1016/j.cell.2018.03.035

Savic S, Ouboussad L, Dickie LJ, Geiler J, Wong C, Doody GM, Churchman SM, Ponchel F, Emery P, Cook GP, et al. 2014. Tlr dependent xbp-1 activation induces an autocrine loop in rheumatoid arthritis synoviocytes. J Autoimmun. 50:59-66. doi:10.1016/j.jaut.2013.11.002

Sheng X, Nenseth HZ, Qu S, Kuzu OF, Frahnow T, Simon L, Greene S, Zeng Q, Fazli L, Rennie PS, et al. 2019. Ire1alpha-xbp1s pathway promotes prostate cancer by activating c-myc signaling. Nat Commun. 10(1):323. doi:10.1038/s41467-018-08152-3

Stockley J, Villasevil ME, Nixon C, Ahmad I, Leung HY, Rajan P. 2014. The rna-binding protein hnrnpa2 regulates beta-catenin protein expression and is overexpressed in prostate cancer. RNA Biol. 11(6):755-765. doi:10.4161/rna.28800

Uemura A, Oku M, Mori K, Yoshida H. 2009. Unconventional splicing of xbp1 mrna occurs in the cytoplasm during the mammalian unfolded protein response. J Cell Sci. 122(Pt 16):2877-2886. doi:10.1242/jcs.040584

Wolozin B, Ivanov P. 2019. Stress granules and neurodegeneration. Nat Rev Neurosci. 20(11):649-666. doi:10.1038/s41583-019-0222-5

Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. 2021. Clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (N Y). 2(3):100141. doi:10.1016/j.xinn.2021.100141

Xie H, Tang CH, Song JH, Mancuso A, Del Valle JR, Cao J, Xiang Y, Dang CV, Lan R, Sanchez DJ, et al. 2018. Ire1alpha rnase-dependent lipid homeostasis promotes survival in myc-transformed cancers. J Clin Invest. 128(4):1300-1316. doi:10.1172/JCI95864

Yao P, Potdar AA, Ray PS, Eswarappa SM, Flagg AC, Willard B, Fox PL. 2013. The hilda complex coordinates a conditional switch in the 3'-untranslated region of the vegfa mrna. PLoS Biol. 11(8):e1001635. doi:10.1371/journal.pbio.1001635

1 Zuccotti P, Colombrita C, Moncini S, Barbieri A, Lunghi M, Gelfi C, De Palma S, Nicolin A, Ratti A,

2 Venturin M, et al. 2014. Hnrnpa2/b1 and nelav proteins bind to a specific u-rich element in cdk5r1

3 3'-utr and oppositely regulate its expression. Biochim Biophys Acta. 1839(6):506-516.

4 doi:10.1016/j.bbagrm.2014.04.018

Figure Legends

Figure 1. HNRNPA2B1 overexpression is associated with poor patient prognosis and cellular stress pathways in primary prostate cancer.

(A) Distribution of *HNRNPA2B1* expression values reported as RNA-Seq by Expectation-Maximization (RSEM) in primary prostate tumours and benign adjacent tissue from The Cancer Genome Atlas (TCGA) patient cohort. Two-tailed T-test was used to compare treatment groups. *** = p<0.001 **(B)** Kaplan-Meier plot of disease-free survival for primary PC patients stratified by *HNRNPA2B1* expression (low = <1st-3rd quartile and high = >3rd quartile). The number of patients at risk for each group are presented in the table below each X-axis time point. Univariable Cox PH-derived hazard ratios (HR) with 95% confidence intervals (CI) and two-tailed log-rank test p-values are shown. **(C-D)** GSECA analysis performed on **(C)** primary PC (TCGA) and **(D)** metastatic PC (SU2C) RNA-Seq datasets by stratification of cohorts based on *HNRNPA2B1* expression. Genes in a given Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway are separated into seven expression classes: NE = not expressed, LE= lowly expressed, ME = medium expression, HE1-4 = high expression. Triangles compare the difference in the cumulative proportion of genes in an expression class between *HNRNPA2B1* high and low expression groups, and represent the size and enrichment (up) or depletion (down) of genes. AS = association score. **(E)** KEGG pathway gene enrichment analysis of differentially expressed genes (p<0.05 and absolute $\log_2$ fold change > 0.5 or < -0.5) identified by RNA-Seq of PC3M cells upon depletion of HNRNPA2B1 using a single siRNA duplex (si1, 20nM for 72 hours). **(F)** Log_2 fold change gene expression values for differentially expressed "*Protein processing In endoplasmic reticulum*" genes upon HNRNPA2B1 depletion in PC3M cells (p<0.05 and absolute $\log_2$ fold change > 0.5 or < -0.5). P-values for each gene adjusted using the Benjamini and Hochberg method are represented by the bar colour – see key.

Figure 2. HNRNPA2B1 regulates processing of IRE1 target mRNAs.

(A) Schematic of XBP1 gene. Exons 1-3 and 5 are indicated by yellow boxes and the non-canonically spliced exon 4 by a black box. XBP1u contains a variable 26-nucleotide region in exon

4 indicated by a white box, the exclusion of which generates the transcriptionally active XBP1s isoform. Red arrows represent RT-PCR primers used to amplify XBP1u and XBP1s products. **(B, left panel)** PC3M cells were treated with 250 nM Thapsigargin (TG), or vehicle (Control) DMSO for 24 hours and total RNA analysed using XBP1 splicing assays. Representative capillary gel electrophoretogram (QIAxcel) shows two bands representing transcripts with (XBP1u) or without (XBP1s) the exon 4 variable 26-nucleotide region inclusion**. (B, right panel)** Electrophoretograms were quantified to determine the percentage change in XBP1s product expression (ΔXBP1s). **(C)** PC3M cells were depleted of HNRNPA2B1 expression using two different siRNA duplexes (si1 and si2, 20nM for 72 hours) or non-silencing control (Nsi). Western blot shows HNRNPA2 (major isoform) and B1 (minor isoform) protein expression compared to Beta Actin loading control. The numbers below the HNRNPA2B1 blot indicate the relative reduction in total HNRNPA2B1 protein expression following siRNA depletion compared to Nsi control. **(D, left panel)** Total RNA was analysed using XBP1 splicing assays and representative capillary gel electrophoretogram show two bands representing XBP1u and XBP1s transcripts. **(D, right panel)** Electrophoretograms were quantified to determine the percentage change in XBP1s product expression (ΔXBP1s). **(E)** Relative change in *BLOC1S1* expression to DMSO control measured by qRT-PCR in PC3M cells treated with vehicle (Control) DMSO or Thapsigargin (TG) 250nM for 24 hours. **(F)** Relative change in *BLOC1S1* expression to Nsi measured by qRT-PCR in PC3M cells depleted of HNRNPA2B1 expression using two different single siRNA duplexes (si1 and si2, 20nM for 72 hours). At least three biological replicates were used, and Two-tailed T-test was used to compare treatment groups. * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$

Figure 3. HNRNPA2B1-IRE1-XBP1 co-regulated genes represent a prognostic biomarker signature in primary PC and reveal a potential therapeutic target

(A) Venn diagram of protein-coding genes differentially-expressed and co-regulated by XBP1, IRE1 and HNRNPA2B1 with Log_2 fold change <-0.5 and $p<0.05$ in RNA-Seq datasets from LNCaP cells treated with siRNA to XBP1 or IRE1 inhibitor MKC8866 (Sheng et al 2019) or PC3M cells treated with siRNA to HNRNPA2B1. **(B)** Derivation of prognostic biomarker panel by elastic net selection of 20 HNRNPA2B1, IRE1, and XBP1 co-regulated protein-coding genes in the TCGA

cohort to generate a single four gene panel as the best predictors of disease relapse. **(B, left panel)** Cross-validation curve (red dots) with standard deviation. Left vertical dashed line is the value of λ that gives minimum mean cross-validated error (lambda.min), right vertical dashed line is the value of λ that gives the most regularized model such that the cross-validated error is within one standard error of the minimum (lambda.1se). **(B, right panel)** Heat map displaying the $\log_2$ fold change expression of the four HIX signature genes following treatment of LNCaPs with the IRE1 inhibitor MKC886, or XBP1 or HNRNPA2B1 depletion in LNCaP and PC3M cells respectively. **(C-D, top panels)** Distribution plot of risk scores for **(C)** derivation (TCGA) and **(D)** validation (MSKCC) cohorts. Vertical red lines represent mean of low and high percentile risk scores. **(C-D, bottom panels)** Kaplan-Meier plots of disease-free survival probabilities for patients from **(C)** derivation (TCGA) and **(D)** validation (MSKCC) datasets stratified by risk groups. The number of patients at risk for each group are presented in the table below each X-axis time point. Univariable Cox PH-dervied hazard ratios with 95% confidence intervals (CI) and two-tailed log-rank test p-values are shown. **(E)** PC3M cells were transfected with 3µg plasmid DNA (72 hours) encoding HNRNPA2 or vector only (VO) control. **(E, top panel)** Western blot shows HNRNPA2 protein expression compared to Beta Actin loading control. **(E, bottom panel)** PC3M cell viability was measured by MTT assay following transfection with 300 ng of plasmid DNA vector encoding HNRNPA2 or VO control. Cells were simultaneously treated with either 50 or 100 µM STF083010 or vehicle control (DMSO). Three biological replicates were used, and Two-tailed T-test was used to compare treatment groups. * = $p<0.05$

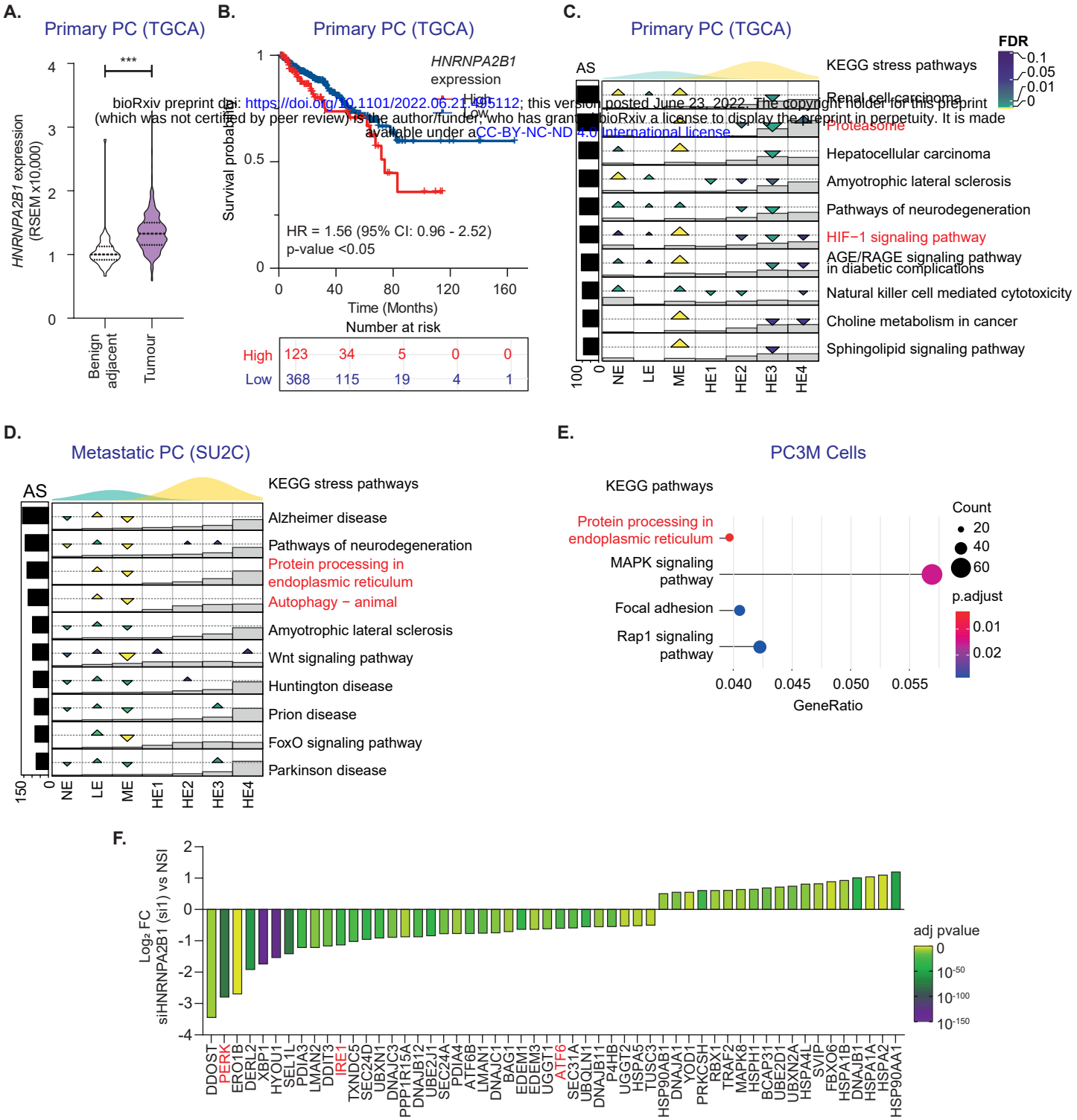

Figure 1

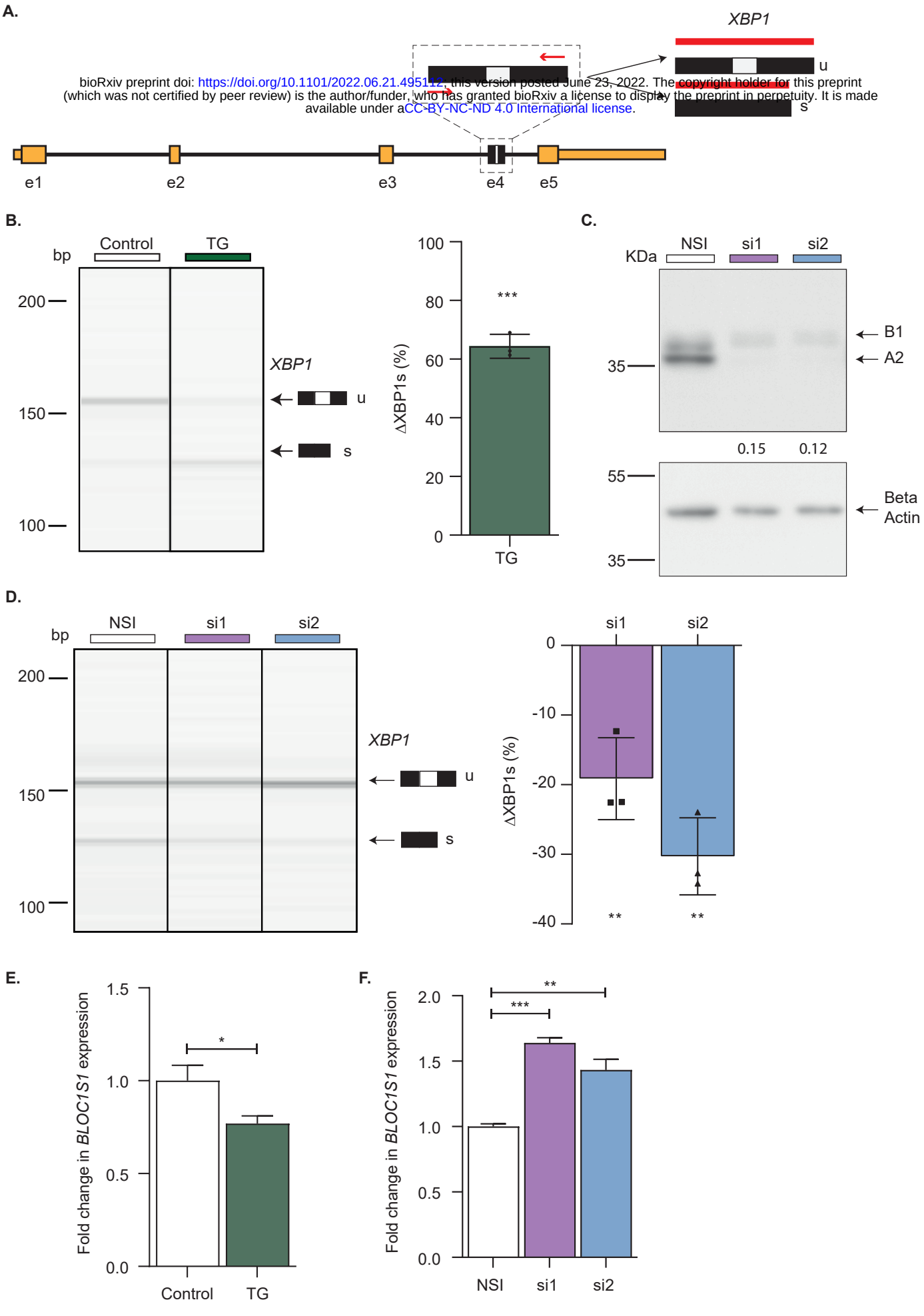

Figure 2

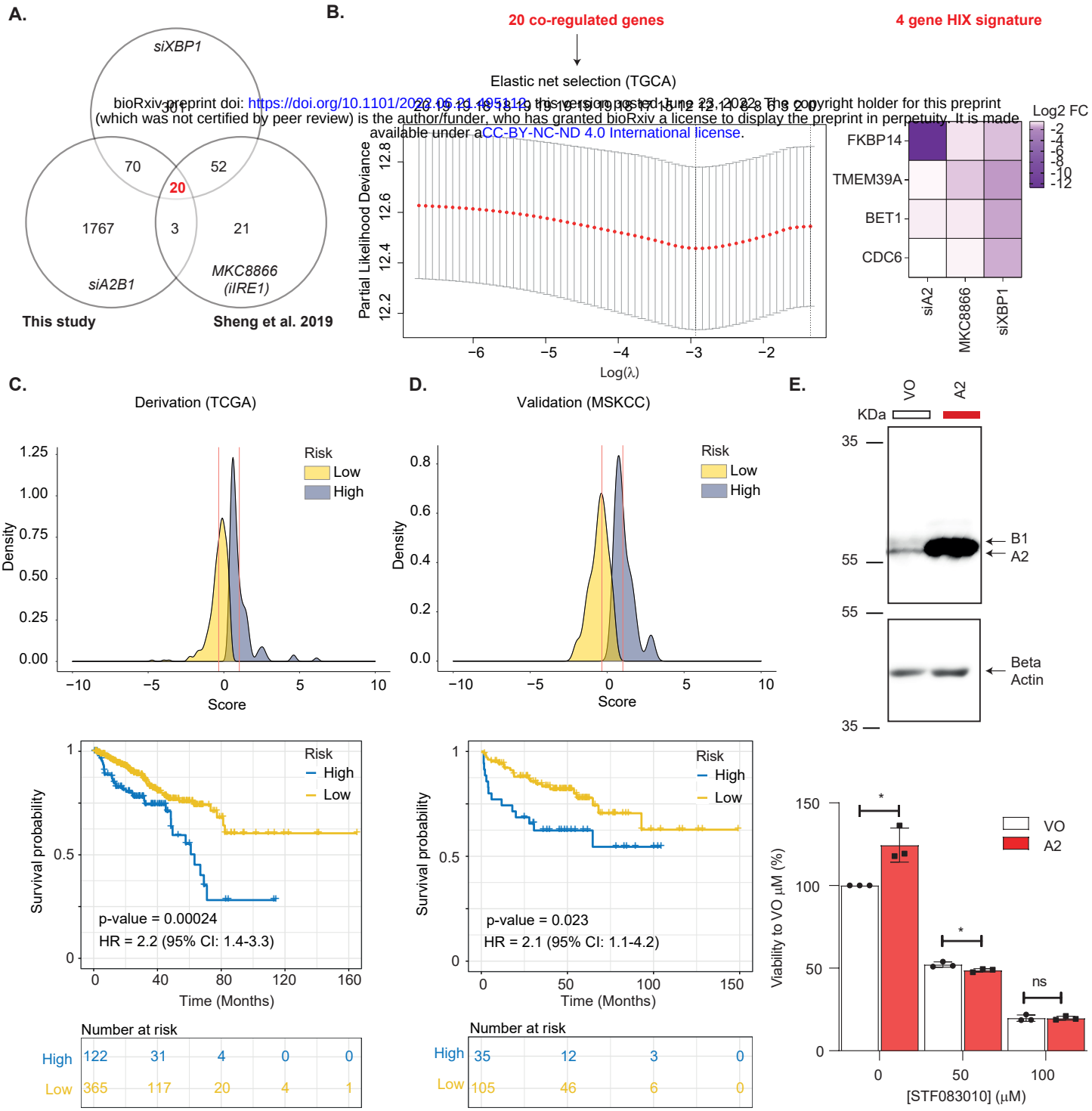

Figure 3

A.0.1 Running and deploying

A.0.1.1 Application containers

For complex pipelines which include multiple other packages or tools, maintenance and deployment is a major consideration. Indeed, many tools are platform dependent. For example, binary programs written in C or C++ need to be compiled differently on Windows, Linux and Mac. Many R or Python packages are dependent on specific system libraries that need to be installed. Therefore, depending on the type of the user's operating system, developers need to provide different versions of their pipelines. Yet, this does not entirely solve the issue, as some libraries also need to be installed on the computer. This mean developers also need to provide a complex set of instructions detailing the requirements to be able to run the pipeline. For users without advanced computational skills or who have limited permissions in terms of software installation on their computer, this can be challenging or even unsolvable.

An additional issue is that packages used by a pipeline may be regularly updated. These updates may not be retro-compatible, or some functions may be deprecated, which could break the pipeline. One solution could be to always maintain the pipeline up to date to make it compatible with the latest packages, but this requires to always having people being able to react quickly to changes. Another solution could be to only install specific versions of the packages that are known to be compatible with the pipeline. While this solution would work for Python and R packages that usually keep a history of previous versions, the situation becomes more complicated for system libraries, for example on Ubuntu, or third-party software, where older versions may no longer be provided.

Over the past few years, a solution has gained popularity in many fields, including bioinformatics. It involves creating an image of the pipeline frozen in time within a container that contains its own image of an operating system and anything else developers have decided to include in it. While multiple software exist around this concept, the most commonly used is Docker (Doc).

A.0.1.2 Docker

Docker is a virtualisation platform developed in 2013 by Docker. Inc. It allows running isolated software containers on any infrastructure. It is therefore possible to, for example, run Linux software on a Windows machine. Docker is also used extensively on cloud infrastructure such as AWS as it allows an easier deployment of applications.

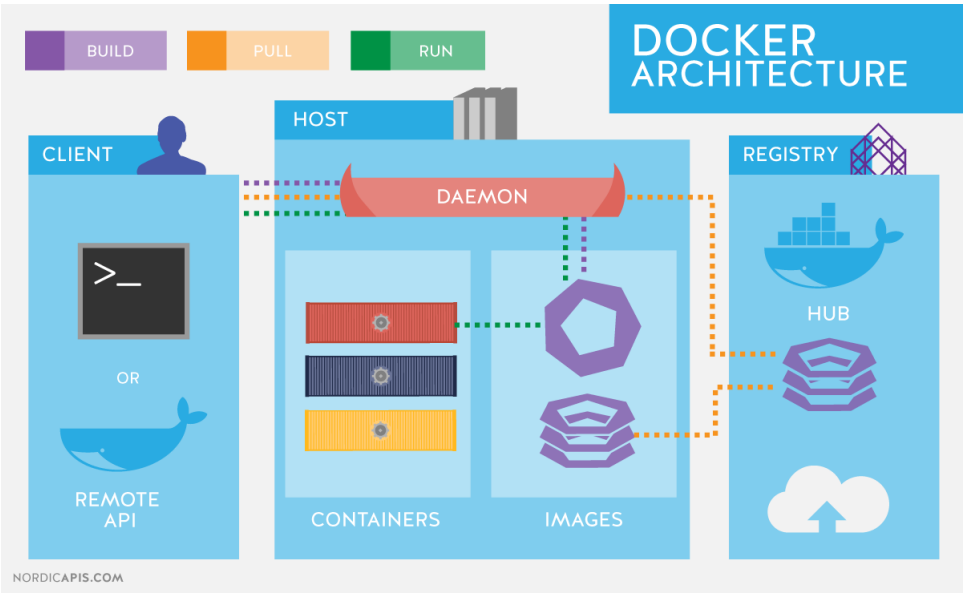

Figure A.1: Docker architecture

Docker is centred around the concept of images and containers. An image is equivalent to a frozen operating system on which some packages and applications are installed. A container is an instance of an image, that will run an application present in the image. Several containers can be instantiated from a single image and can run in parallel and isolated from each other.

In order to build an image, it is necessary to write a Dockerfile which specifies the base image to use, which packages to install, what additional commands to run, and which command to run when instantiating a new container.

The Dockerfile below is used to create the PIT docker image.

```
FROM ubuntu:bionic
```

```
ENV TZ=Europe/Minsk
RUN apt-get update
RUN apt-get install -y dirmngr gnupg apt-transport-https ca-certificates soft

RUN apt-get update  && apt-get upgrade -y
RUN ln -snf /usr/share/zoneinfo/$TZ /etc/localtime && echo $TZ > /etc/timezon
RUN apt install software-properties-common -y

RUN  apt install -y \
    python3 \
    samtools \
    cmake \
    python3-pip \
    libxml2-dev \
    cpanminus \
    libssl-dev \
    libcurl4-openssl-dev  \
    openjdk-11-jdk \
    libfontconfig1-dev libharfbuzz-dev libfribidi-dev libfreetype6-dev libpng
  && pip3 install virtualenv

RUN apt-key adv --keyserver keyserver.ubuntu.com --recv-keys 51716619E084DAB9
RUN add-apt-repository 'deb https://cloud.r-project.org/bin/linux/ubuntu bion
RUN apt install -y r-base r-base-core r-recommended r-base-dev

RUN apt update --allow-unauthenticated && apt upgrade -y

RUN apt-key adv --keyserver keyserver.ubuntu.com --recv-keys A6A19B38D3D831EF
RUN apt-add-repository 'deb https://download.mono-project.com/repo/ubuntu sta
RUN apt-get update
RUN apt-get install -f -y  mono-complete

RUN wget -qO- https://packages.microsoft.com/keys/microsoft.asc | gpg --dearm
```

```
RUN mv microsoft.asc.gpg /etc/apt/trusted.gpg.d/
RUN wget -q https://packages.microsoft.com/config/ubuntu/20.04/prod.list
RUN mv prod.list /etc/apt/sources.list.d/microsoft-prod.list
RUN chown root:root /etc/apt/trusted.gpg.d/microsoft.asc.gpg
RUN chown root:root /etc/apt/sources.list.d/microsoft-prod.list
RUN apt update
RUN apt-get install -y aspnetcore-runtime-2.1 dotnet-sdk-2.1
RUN cpanm URI::Escape
RUN mkdir /project
WORKDIR /usr/src/app
COPY . .
RUN pip3 install --no-cache-dir -r requirements.txt
RUN Rscript installPackages.R

CMD python3 /usr/src/app/LaunchDocker.py
```

It uses an Ubuntu bionic image as base operating system, then proceeds to install all system packages and software required by PIT. The COPY . . instruction copies all the PIT code and files from the host computer into the image. Finally, CMD python3 /usr/src/app/LaunchDocker.py means that upon creating a new container from this image, the LaunchDocker.py script will be called in order to start the PIT pipeline.

Once created, an image can be uploaded to an online repository. The largest one is Dockerhub (https://hub.docker.com/) but others exist, in particular for specific platform such as Amazon Web Services (AWS) which has an Elastic Cloud Repository (ECR) that can host docker images for them to be used by other AWS services.

Users can then pull an image from a repository to their computer using the "docker pull" command and start a new container from this image using the "docker run" command. The main benefit of doing so is that users can then run PIT regardless of their operating system and without having to install anything apart from Docker. This approach is becoming more and more common within bioinformatics, with other pipelines including Trinity

(Grabherr et al., 2011) or Immcantation (Gupta et al., 2015) that offer docker images to run their pipelines. As for PIT, a Docker image is also available at https://hub.docker.com/repository/docker/nirkoty/pit.

A.0.1.3 Apptainer

While docker is particularly popular on personal computers and cloud infrastructure, it has limitations that do not make it the most suitable solution for a High Performance Computer (HPC) such as those commonly used by universities or other academic institutions. The main limitation is security. Indeed, Docker is managed through a daemon, which means that in order to let the HPC users start Docker containers, it is necessary to give them access to the daemon. This is a security hazard in a multi-user environment, where giving someone access to the daemon would also give them access to the containers of other users. In addition, docker container are started with root authorisation, while it may be preferable to give users a more limited level of permissions.

For these reasons, a preferred alternative to Docker on HPC is Apptainer (Kurtzer et al., 2017). The concept is similar to Docker, and it is possible to convert a Docker image into an Apptainer image. One of the differences lies in the fact that Apptainer doesn't work with a daemon. An Apptainer image is just a single file, from which we can create container which run in their own process. Thus, it is not possible to access other containers through the use of a daemon, which makes Apptainer more secure for an HPC environment. Since running PIT requires substantive computing power and is targeted at, among others, the academic community, HPCs are a very suitable place to run PIT, therefore we also built a Singularity image for PIT.

Bibliography

Dizitart — Nitrite. URL https://www.dizitart.org/nitrite-database.html.

Docker: Lightweight Linux Containers for Consistent Development and Deployment — Linux Journal. URL https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment.

Gene Ontology overview. URL http://geneontology.org/docs/ontology-documentation/.

JavaFX. URL https://openjfx.io/.

kevinblighe/EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. URL https://github.com/kevinblighe/EnhancedVolcano.

The Model View Controller Pattern – MVC Architecture and Frameworks Explained. URL https://www.freecodecamp.org/news/the-model-view-controller-pattern-mvc-architecture-and-frameworks-explained/.

Types of Prostate Cancer: Prostatic Adenocarcinoma & Other Forms — CTCA. URL https://www.cancercenter.com/cancer-types/prostate-cancer/types.

What Is Cancer? - National Cancer Institute. URL https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

JSON format, 2020. URL https://livebook.manning.com/book/get-programming-with-haskell/chapter-40/.

W. Abida, J. Cyrta, G. Heller, D. Prandi, J. Armenia, I. Coleman, M. Cieslik, M. Benelli, D. Robinson, E. M. Van Allen, A. Sboner, T. Fedrizzi, J. M. Mosquera, B. D. Robinson, N. De Sarkar, L. P. Kunju, S. Tomlins, Y. M. Wu, D. N. Rodrigues, M. Loda, A. Gopalan, V. E. Reuter, C. C. Pritchard, J. Mateo, D. Bianchini, S. Miranda, S. Carreira, P. Rescigno, J. Filipenko, J. Vinson, R. B. Montgomery, H. Beltran, E. I. Heath, H. I. Scher, P. W. Kantoff, M. E. Taplin, N. Schultz, J. S. Debono, F. Demichelis, P. S. Nelson, M. A. Rubin, A. M. Chinnaiyan, and C. L. Sawyers. Genomic correlates of clinical outcome in advanced prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 166 (23):11428–11436, 2019. ISSN 10916490. doi: 10.1073/pnas.1902651116.

A. Ackermann and A. Brieger. The Role of Nonerythroid Spectrin αII in Cancer, 2019. ISSN 16878469.

F. B. Ahmad and R. N. Anderson. The Leading Causes of Death in the US for 2020. *JAMA*, 325(18):1829–1830, 5 2021. ISSN 0098-7484. doi: 10.1001/JAMA.2021.5469. URL https://jamanetwork.com/journals/jama/fullarticle/2778234.

J. A. Alfaro, A. Ignatchenko, V. Ignatchenko, A. Sinha, P. C. Boutros, and T. Kislinger. Detecting protein variants by mass spectrometry: A comprehensive study in cancer cell-lines. *Genome Medicine*, 9(1):62, 12 2017. ISSN 1756994X. doi: 10.1186/s13073-017-0454-9. URL http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0454-9.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 00222836. doi: 10.1016/S0022-2836(05)80360-2. URL https://pubmed.ncbi.nlm.nih.gov/2231712/.

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 Biology 2020 21:1*, 21(1):1–16, 2 2020. ISSN 1474-760X. doi: 10.1186/S13059-020-1935-5. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1935-5.

J. A. Ankney, A. Muneer, and X. Chen. Relative and Absolute Quantitation in Mass Spectrometry–Based Proteomics. *https://doi.org/10.1146/annurev-anchem-061516-045357*, 11:49–77, 6 2016. ISSN 19361335. doi: 10.1146/ANNUREV-ANCHEM-061516-045357. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-anchem-061516-045357.

P. R. Araujo, K. Yoon, D. Ko, A. D. Smith, M. Qiao, U. Suresh, S. C. Burns, and L. O. F. Penalva. Before It Gets Started: Regulating Translation at the 5 UTR. *Comparative and Functional Genomics*, 2012:1–8, 2012. ISSN 1531-6912. doi: 10.1155/2012/475731. URL http://www.hindawi.com/journals/ijg/2012/475731/.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: Tool for the unification of biology, 5 2000. ISSN 10614036.

J. Bane, O. Mozziconacci, L. Yi, Y. J. Wang, A. Sreedhara, and C. Schöneich. Photo-oxidation of IgG1 and Model Peptides: Detection and Analysis of Triply Oxidized His and Trp Side Chain Cleavage Products. *Pharmaceutical Research*, 34(1):229–242, 1 2017. ISSN 1573904X. doi: 10.1007/S11095-016-2058-2/FIGURES/11. URL https://link.springer.com/article/10.1007/s11095-016-2058-2.

S. Bartoszewska and J. F. Collawn. Unfolded protein response (UPR) integrated signaling networks determine cell fate during hypoxia, 3 2020. ISSN 16891392.

A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, and J. Zhang. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45 (D1):D158–D169, 1 2017. ISSN 0305-1048. doi: 10.1093/NAR/GKW1099. URL `https://academic.oup.com/nar/article/45/D1/D158/2605721`.

M. B. Battles, V. Más, E. Olmedillas, O. Cano, M. Vázquez, L. Rodríguez, J. A. Melero, and J. S. McLellan. Structure and immunogenicity of pre-fusion-stabilized human metapneumovirus F glycoprotein. *Nature Communications*, 8(1), 12 2017. ISSN 20411723. doi: 10.1038/S41467-017-01708-9. URL `https://www.bnl.gov/newsroom/news.php?a=213084`.

G. Bell. Replicates and repeats, 4 2016. ISSN 17417007. URL `http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-016-0254-5`.

L. Beltran and P. R. Cutillas. Advances in phosphopeptide enrichment techniques for phosphoproteomics. *Amino Acids*, 43(3):1009–1024, 9 2012. ISSN 09394451. doi: 10.1007/S00726-012-1288-9/FIGURES/2. URL https://link.springer.com/article/10.1007/s00726-012-1288-9.

H. Benhabiles, S. Gonzalez-Hilarion, S. Amand, C. Bailly, A. Prévotat, P. Reix, D. Hubert, E. Adriaenssens, S. Rebuffat, D. Tulasne, and F. Lejeune. Optimized approach for the identification of highly efficient correctors of nonsense mutations in human diseases. *PLOS ONE*, 12(11): e0187930, 11 2017. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0187930. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187930.

C. Bessant. *Proteome Informatics*. New Developments in Mass Spectrometry. The Royal Society of Chemistry, 2017. ISBN 978-1-78262-428-8. doi: 10.1039/9781782626732. URL http://dx.doi.org/10.1039/9781782626732.

M. Blein-Nicolas, H. Xu, D. de Vienne, C. Giraud, S. Huet, and M. Zivy. Including shared peptides for estimating protein abundances: a significant improvement for quantitative proteomics. *Proteomics*, 12(18):2797–2801, 9 2012. ISSN 1615-9861. doi: 10.1002/PMIC.201100660. URL https://pubmed.ncbi.nlm.nih.gov/22833229/.

U. Boesl. Time-of-flight mass spectrometry: Introduction to the basics. *Mass Spectrometry Reviews*, 36(1):86–109, 1 2017. ISSN 1098-2787. doi: 10.1002/MAS.21520. URL https://onlinelibrary.wiley.com/doi/full/10.1002/mas.21520https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.21520https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/mas.21520.

J. J. Bouchard, J. H. Otero, D. C. Scott, E. Szulc, E. W. Martin, N. Sabri, D. Granata, M. R. Marzahn, K. Lindorff-Larsen, X. Salvatella, B. A. Schulman, and T. Mittag. Cancer Mutations of the Tumor Suppressor SPOP Disrupt the Formation of Active, Phase-Separated Compart-

ments. *Molecular Cell*, 72(1):19–36, 10 2018. ISSN 10974164. doi: 10.1016/j.molcel.2018.08.027.

R. Buels, E. Yao, C. M. Diesh, R. D. Hayes, M. Munoz-Torres, G. Helt, D. M. Goodstein, C. G. Elsik, S. E. Lewis, L. Stein, and I. H. Holmes. JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biology*, 17(1), 4 2016. ISSN 1474760X. doi: 10.1186/s13059-016-0924-1.

M. Calfon, H. Zeng, F. Urano, J. H. Till, S. R. Hubbard, H. P. Harding, S. G. Clark, and D. Ron. IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. *Nature*, 415(6867): 92–96, 1 2002. ISSN 00280836. doi: 10.1038/415092a.

E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 5 2012. ISSN 21598274. doi: 10.1158/2159-8290.CD-12-0095. URL `/pmc/articles/PMC3956037/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956037/`.

F. Chalmers, S. Mogre, J. Son, N. Blazanin, and A. B. Glick. The multiple roles of the unfolded protein response regulator IRE1α in cancer. *Molecular Carcinogenesis*, 58(9):1623–1630, 9 2019. ISSN 10982744. doi: 10.1002/mc.23031.

J. Chen and W. A. Weiss. Alternative splicing in cancer: implications for biology and therapy. *Oncogene 2015 34:1*, 34(1):1–14, 1 2014. ISSN 1476-5594. doi: 10.1038/onc.2013.570. URL `https://www.nature.com/articles/onc2013570`.

W. Chen, J. M. Smeekens, and R. Wu. Systematic study of the dynamics and half-lives of newly synthesized proteins in human cells. *Chemical Science*, 7(2):1393–1400, 1 2016. ISSN 20416539. doi: 10.1039/C5SC03826J. URL `https://pubs.`

rsc.org/en/content/articlehtml/2016/sc/c5sc03826jhttps://pubs.rsc.org/en/content/articlelanding/2016/sc/c5sc03826j.

M. Choi, C. Y. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean, and O. Vitek. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17): 2524–2526, 9 2014. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTU305. URL https://academic.oup.com/bioinformatics/article/30/17/2524/2748156.

C. V. Clower, D. Chatterjee, Z. Wang, L. C. Cantley, M. G. Heidena, and A. R. Krainer. The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 107(5):1894–1899, 2 2010. ISSN 00278424. doi: 10.1073/pnas.0914845107.

J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, 12 2008. ISSN 10870156. doi: 10.1038/nbt.1511. URL http://www.nature.com/naturebiotechnology.

J. Craig Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian,

W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M. H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Ni Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu,

A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2 2001. ISSN 00368075. doi: 10.1126/SCIENCE.1058040/SUPPL{_}FILE/1058040S3-3{_}MED.GIF. URL https://www.science.org/doi/abs/10.1126/science.1058040.

J. R. Cubillos-Ruiz, S. E. Bettigole, and L. H. Glimcher. Tumorigenic and Immunosuppressive Effects of Endoplasmic Reticulum Stress in Cancer, 2 2017. ISSN 10974172.

D. C. da Silva, P. Valentão, P. B. Andrade, and D. M. Pereira. Endoplasmic reticulum stress signaling in cancer and neurodegenerative disorders: Tools and strategies to understand its complexity, 5 2020. ISSN 10961186.

F. da Veiga Leprevost, S. E. Haynes, D. M. Avtonomov, H. Y. Chang, A. K. Shanmugam, D. Mellacheruvu, A. T. Kong, and A. I. Nesvizhskii. Philosopher: a versatile toolkit for shotgun proteomics data analysis, 9 2020. ISSN 15487105. URL https://www.nature.com/articles/s41592-020-0912-y.

H. Dana, G. M. Chalbatani, H. Mahmoodzadeh, R. Karimloo, O. Rezaiean, A. Moradzadeh, N. Mehmandoost, F. Moazzen, A. Mazraeh, V. Marmari, M. Ebrahimi, M. M. Rashno, S. J. Abadi, and E. Gharagouzlo. Molecular Mechanisms and Biological Functions of siRNA. *International journal of biomedical science : IJBS*, 13(2):48–57, 6 2017. ISSN 1550-9702. URL http://www.ncbi.nlm.nih.gov/pubmed/28824341http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5542916.

V. Dančík, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology : a journal of computational molecular cell biology*, 6(3-4):327–342, 9 1999. ISSN 1066-5277. doi: 10.1089/106652799318300. URL https://pubmed.ncbi.nlm.nih.gov/10582570/.

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. The variant call format and VCFtools. *Bioinformatics*, 27

(15):2156–2158, 8 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr330. URL /pmc/articles/PMC3137218/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/.

C. J. David, M. Chen, M. Assanah, P. Canoll, and J. L. Manley. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature*, 463(7279):364–368, 1 2010. ISSN 00280836. doi: 10.1038/nature08697. URL https://pubmed.ncbi.nlm.nih.gov/20010808/.

E. W. Deutsch. Mass spectrometer output file format mzML. *Methods in molecular biology (Clifton, N.J.)*, 604:319–331, 2010. ISSN 19406029. doi: 10.1007/978-1-60761-444-9{_}22. URL /pmc/articles/PMC3073315/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3073315/.

E. W. Deutsch. File Formats Commonly Used in Mass Spectrometry Proteomics. *Molecular & Cellular Proteomics : MCP*, 11(12):1612, 12 2012. ISSN 15359476. doi: 10.1074/MCP.R112.019695. URL /pmc/articles/PMC3518119//pmc/articles/PMC3518119/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518119/.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 10 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts635. URL https://doi.org/10.1093/bioinformatics/bts635.

V. C. Evans, G. Barker, K. J. Heesom, J. Fan, C. Bessant, and D. A. Matthews. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature Methods*, 9 (12):1207–1211, 12 2012. ISSN 15487091. doi: 10.1038/nmeth.2227. URL /pmc/articles/PMC3581816/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3581816/.

M. Fähling, R. Mrowka, A. Steege, P. Martinka, P. B. Persson, and B. J. Thiele. Heterogeneous nuclear ribonucleoprotein-A2/B1 modulate colla-

gen prolyl 4-hydroxylase, α (I) mRNA stability. *Journal of Biological Chemistry*, 281(14):9279–9286, 4 2006. ISSN 00219258. doi: 10.1074/jbc.M510925200.

Y. Fang and M. J. Fullwood. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, Proteomics & Bioinformatics*, 14 (1):42–54, 2 2016. ISSN 1672-0229. doi: 10.1016/J.GPB.2015.09.006.

T. Fiolet, B. Srour, L. Sellem, E. Kesse-Guyot, B. Allès, C. Méjean, M. Deschasaux, P. Fassier, P. Latino-Martel, M. Beslay, S. Hercberg, C. Lavalette, C. A. Monteiro, C. Julia, and M. Touvier. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. *BMJ*, 360:322, 2 2018. ISSN 0959-8138. doi: 10.1136/BMJ.K322. URL `https://www.bmj.com/content/360/bmj.k322https://www.bmj.com/content/360/bmj.k322.abstract`.

J. G. Foster, E. Gea, M. A. Labiba, C. A. Anene, J. Stockley, C. Philippe, M. Cereda, K. Rouault-Pierre, H. Leung, C. Bessant, and P. Rajan. HNRNPA2B1 controls an unfolded protein response-related prognostic gene signature in prostate cancer. *bioRxiv*, page 2022.06.21.495112, 6 2022. doi: 10.1101/2022.06.21.495112. URL `https://www.biorxiv.org/content/10.1101/2022.06.21.495112v1https://www.biorxiv.org/content/10.1101/2022.06.21.495112v1.abstract`.

A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids*

research, 47(D1):D766–D773, 1 2019. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gky955.

M. Furrukh. Tobacco Smoking and Lung Cancer: Perception-changing facts. *Sultan Qaboos University Medical Journal*, 13(3):345, 2013. ISSN 20750528. doi: 10.12816/0003255. URL /pmc/articles/PMC3749017//pmc/articles/PMC3749017/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3749017/.

M. Gabay, Y. Li, and D. W. Felsher. MYC Activation Is a Hallmark of Cancer Initiation and Maintenance. *Cold Spring Harbor Perspectives in Medicine*, 4(6), 2014. ISSN 21571422. doi: 10.1101/CSHPERSPECT.A014241. URL /pmc/articles/PMC4031954//pmc/articles/PMC4031954/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4031954/.

E. Gadaleta, P. Fourgoux, S. Pirró, G. J. Thorn, R. Nelan, A. Ironside, V. Rajeeve, P. R. Cutillas, A. E. Lobley, J. Wang, E. Gea, H. Ross-Adams, C. Bessant, N. R. Lemoine, L. J. Jones, and C. Chelala. Characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer. *npj Breast Cancer 2020 6:1*, 6(1):1–12, 8 2020. ISSN 2374-4677. doi: 10.1038/s41523-020-00182-9. URL https://www.nature.com/articles/s41523-020-00182-9.

J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269), 4 2013. ISSN 19450877. doi: 10.1126/scisignal.2004088.

D. Gfeller and M. Bassani-Sternberg. Predicting antigen presentation-What could we learn from a million peptides? *Frontiers in Immunology*, 9(JUL): 1716, 7 2018. ISSN 16643224. doi: 10.3389/FIMMU.2018.01716/BIBTEX.

M. Gierlinski, F. Gastaldello, C. Cole, and G. J. Barton. Proteus: an R package for downstream analysis of MaxQuant output.

bioRxiv, page 416511, 9 2018. ISSN 2692-8205. doi: 10.1101/416511. URL https://www.biorxiv.org/content/10.1101/416511v2https://www.biorxiv.org/content/10.1101/416511v2.abstract.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 7 2011. ISSN 10870156. doi: 10.1038/nbt.1883. URL https://www.nature.com/articles/nbt.1883.

J. Griss, A. R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G. G. Thallinger, R. M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, S. Neumann, J. Fan, F. Reisinger, Q. W. Xu, N. Del Toro, Y. Pérez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaíno, and H. Hermjakob. The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Molecular & Cellular Proteomics : MCP*, 13(10):2765, 10 2014. ISSN 15359484. doi: 10.1074/MCP.O113.036681. URL /pmc/articles/PMC4189001//pmc/articles/PMC4189001/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4189001/.

M. Gry, R. Rimini, S. Strömberg, A. Asplund, F. Pontén, M. Uhlén, and P. Nilsson. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics*, 10(1):1–14, 8 2009. ISSN 14712164. doi: 10.1186/1471-2164-10-365/FIGURES/5. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-365.

N. Gupta and P. A. Pevzner. False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule. *Journal of Proteome Research*, 8(9):4173, 2009. ISSN 15353893. doi: 10.1021/PR9004794. URL /pmc/articles/PMC3398614//pmc/articles/PMC3398614/

?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398614/.

N. T. Gupta, J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H. Kleinstein. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data: Table 1. *Bioinformatics*, 31(20):3356–3358, 10 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv359. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv359.

J. Han and R. J. Kaufman. Physiological/pathological ramifications of transcription factors in the unfolded protein response, 2017. ISSN 15495477.

D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation, 3 2011. ISSN 00928674. URL http://www.cell.com/article/S0092867411001279/fulltexthttp://www.cell.com/article/S0092867411001279/abstracthttps://www.cell.com/cell/abstract/S0092-8674(11)00127-9.

S. Haupt and Y. Haupt. P53 at the start of the 21st century: lessons from elephants. *F1000Research*, 6, 2017. ISSN 1759796X. doi: 10.12688/F1000RESEARCH.12682.1. URL /pmc/articles/PMC5701437//pmc/articles/PMC5701437/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5701437/.

K. E. Hayer, A. Pizarro, N. F. Lahens, J. B. Hogenesch, and G. R. Grant. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, 31(24):3938, 7 2015. ISSN 14602059. doi: 10.1093/BIOINFORMATICS/BTV488. URL /pmc/articles/PMC4673975//pmc/articles/PMC4673975/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4673975/.

A. G. Hinnebusch, I. P. Ivanov, and N. Sonenberg. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science*

(New York, N.Y.), 352(6292):1413, 6 2016. ISSN 10959203. doi: 10.1126/SCIENCE.AAD9868. URL /pmc/articles/PMC7422601//pmc/articles/PMC7422601/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7422601/.

J. J. Ho, N. C. Balukoff, P. R. Theodoridis, M. Wang, J. R. Krieger, J. H. Schatz, and S. Lee. A network of RNA-binding proteins controls translation efficiency to activate anaerobic metabolism. *Nature Communications*, 11 (1), 12 2020. ISSN 20411723. doi: 10.1038/s41467-020-16504-1.

K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A. M. Taylor, A. D. Cherniack, V. Thorsson, R. Akbani, R. Bowlby, C. K. Wong, M. Wiznerowicz, F. Sanchez-Vega, A. G. Robertson, B. G. Schneider, M. S. Lawrence, H. Noushmehr, T. M. Malta, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K. S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. de Bruijn, J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, J. S. Parker, M. D. Wilkerson, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani,

A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. Van Den Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. V. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T. Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. De Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca, B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. Têtu, A. Bergeron,

M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M. H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca, S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. Van Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. De Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. Giné, A. L. Guillermo, N. Van Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino, J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J. P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, o. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shannon, A. Spillane, J. Stretch, M. Synott, J. Thompson, J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drwiega, J. Miller, J. Smith, H. Zaren, J. W. Park, N. P. Hung, E. Kebebew,

W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A. P. Gimenez-Roqueplo, A. Piché, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K. F. Becker, K. P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussioutas, D. Beer, T. Giordano, A. M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman, F. Valdivieso, R. Dhir, J. Luketich, E. M. Pinero, M. Quintero-Aguilo,

C. G. Carlotti, J. S. Dos Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. Van Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, A. Mariamidze, J. M. Stuart, C. C. Benz, and P. W. Laird. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304, 4 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.03.022.

M. Hölzer and M. Marz. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, 8(5):1–16, 5 2019. ISSN 2047217X. doi: 10.1093/GIGASCIENCE/GIZ039. URL `https://academic.oup.com/gigascience/article/8/5/giz039/5488105`.

K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. Ridwan Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. G. Giron, T. Grego, C. Guijarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O. G. Izuogu, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez, J. C. Marugán, T. Maurel, A. C. McMahon, S. Mohanan, B. Moore, M. Muffato, D. N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. de Silva, B. Flint, A. Frankish, S. E. Hunt, G. R. IIsley, N. Langridge, J. E. Loveland, F. J. Martin, J. M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, F. Cunningham, A. D. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–

D891, 1 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAA942. URL https://academic.oup.com/nar/article/49/D1/D884/5952199.

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MaCdonald, V. Obenchain, A. K. Oles, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12 (2):115–121, 1 2015. ISSN 15487105. doi: 10.1038/nmeth.3252. URL https://support.bioconductor.org.

S. D. Hudnall. Cancer issue: Viruses and Human Cancer. *The Yale Journal of Biology and Medicine*, 79(3-4):115, 3 2006. ISSN 1551-4056. doi: 10.1007/978-1-4939-0870-7. URL /pmc/articles/PMC1994798/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1994798/.

N. Hulstaert, J. Shofstahl, T. Sachsenberg, M. Walzer, H. Barsnes, L. Martens, and Y. Perez-Riverol. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *Journal of Proteome Research*, 19(1):537–542, 1 2020. ISSN 15353907. doi: 10.1021/ACS.JPROTEOME.9B00328/SUPPL{_}FILE/PR9B00328{_}SI{_}001.PDF. URL https://pubs.acs.org/doi/abs/10.1021/acs.jproteome.9b00328.

Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Molecular & Cellular Proteomics*, 4(9):1265–1272, 9 2005. ISSN 1535-9476. doi: 10.1074/MCP.M500061-MCP200.

L. M. Janssen, E. E. Ramsay, C. D. Logsdon, and W. W. Overwijk. The immune system in cancer metastasis: friend or foe? *Journal for ImmunoTherapy of Cancer*, 5(1):79, 12 2017. ISSN 2051-1426. doi:

10.1186/S40425-017-0283-9. URL https://jitc.bmj.com/content/5/1/79https://jitc.bmj.com/content/5/1/79.abstract.

L. Jiang, W. Lin, C. Zhang, P. E. Ash, M. Verma, J. Kwan, E. van Vliet, Z. Yang, A. L. Cruz, S. Boudeau, B. F. Maziuk, S. Lei, J. Song, V. E. Alvarez, S. Hovde, J. F. Abisambra, M. H. Kuo, N. Kanaan, M. E. Murray, J. F. Crary, J. Zhao, J. X. Cheng, L. Petrucelli, H. Li, A. Emili, and B. Wolozin. Interaction of tau with HNRNPA2B1 and N6-methyladenosine RNA mediates the progression of tauopathy. *Molecular Cell*, 81(20):4209–4227, 10 2021. ISSN 10974164. doi: 10.1016/j.molcel.2021.07.038.

A. R. Jones, M. Eisenacher, G. Mayer, O. Kohlbacher, J. Siepen, S. J. Hubbard, J. N. Selley, B. C. Searle, J. Shofstahl, S. L. Seymour, R. Julian, P. A. Binz, E. W. Deutsch, H. Hermjakob, F. Reisinger, J. Griss, J. A. Vizcaíno, M. Chambers, A. Pizarro, and D. Creasy. The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Molecular & Cellular Proteomics : MCP*, 11(7), 7 2012. ISSN 15359476. doi: 10.1074/MCP.M111.014381. URL /pmc/articles/PMC3394945//pmc/articles/PMC3394945/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3394945/.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes, 1 2000. ISSN 03051048.

H. Khatter, A. G. Myasnikov, S. K. Natchiar, and B. P. Klaholz. Structure of the human 80S ribosome. *Nature 2015 520:7549*, 520(7549):640–645, 4 2015. ISSN 1476-4687. doi: 10.1038/nature14427. URL https://www.nature.com/articles/nature14427.

A. Kianianmomeni, C. S. Ong, G. Rätsch, and A. Hallmann. Genome-wide analysis of alternative splicing in Volvox carteri. *BMC Genomics*, 15(1):1–21, 12 2014. ISSN 14712164. doi: 10.1186/1471-2164-15-1117/FIGURES/7. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-1117.

S. Kim and P. A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5(1):1–10, 10 2014. ISSN 20411723. doi: 10.1038/ncomms6277. URL www.nature.com/naturecommunications.

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 Biology*, 20(1):1–13, 12 2019. ISSN 1474760X. doi: 10.1186/S13059-019-1910-1/FIGURES/6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1910-1.

A. Krieg, C. Mahotka, T. Krieg, H. Grabsch, W. Müller, S. Takeno, C. V. Suschek, M. Heydthausen, H. E. Gabbert, and C. D. Gerharz. Expression of different survivin variants in gastric carcinomas: first clues to a role of survivin-2B in tumour progression. *British Journal of Cancer*, 86(5):737, 3 2002. ISSN 00070920. doi: 10.1038/SJ.BJC.6600153. URL /pmc/articles/PMC2375298//pmc/articles/PMC2375298/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2375298/.

G. M. Kurtzer, V. Sochat, and M. W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459, 5 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177459. URL https://dx.plos.org/10.1371/journal.pone.0177459.

A. Lafita, S. Bliven, A. Prlić, D. Guzenko, P. W. Rose, A. Bradley, P. Pavan, D. Myers-Turnbull, Y. Valasatava, M. Heuer, M. Larson, S. K. Burley, and J. M. Duarte. BioJava 5: A community driven open-source bioinformatics library. *PLOS Computational Biology*, 15(2):e1006791, 2 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006791. URL https://dx.plos.org/10.1371/journal.pcbi.1006791.

A. Lauria, S. Peirone, M. D. Giudice, F. Priante, P. Rajan, M. Caselle, S. Oliviero, and M. Cereda. Identification of altered biological processes in

heterogeneous RNA-sequencing data by discretization of expression profiles. *Nucleic Acids Research*, 48(4):1730–1747, 2 2020. ISSN 13624962. doi: 10.1093/nar/gkz1208.

L. J. Lee, D. Papadopoli, M. Jewer, S. del Rincon, I. Topisirovic, M. G. Lawrence, and L. M. Postovit. Cancer Plasticity: The Role of mRNA Translation, 2 2021. ISSN 24058033.

S. Lhomond, T. Avril, N. Dejeans, K. Voutetakis, D. Doultsinos, M. McMahon, R. Pineau, J. Obacz, O. Papadodima, F. Jouan, H. Bourien, M. Logotheti, G. Jégou, N. Pallares-Lupon, K. Schmit, P. Le Reste, A. Etcheverry, J. Mosser, K. Barroso, E. Vauléon, M. Maurel, A. Samali, J. B. Patterson, O. Pluquet, C. Hetz, V. Quillien, A. Chatziioannou, and E. Chevet. Dual IRE 1 RN ase functions dictate glioblastoma development . *EMBO Molecular Medicine*, 10(3), 3 2018. ISSN 1757-4676. doi: 10.15252/emmm.201707929.

H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 9 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL `https://academic.oup.com/bioinformatics/article/34/18/3094/4994778`.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 8 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352`.

P. Li, D. Wang, H. Li, Z. Yu, X. Chen, and J. Fang. Identification of nucleolus-localized PTEN and its function in regulating ribosome biogenesis. *Molecular Biology Reports*, 41(10):6383–6390, 9 2014. ISSN 15734978. doi: 10.1007/s11033-014-3518-6.

Y. Li, N. Sahni, R. Pancsa, D. J. McGrail, J. Xu, X. Hua, J. Coulombe-Huntington, M. Ryan, B. Tychhon, D. Sudhakar, L. Hu, M. Tyers,

X. Jiang, S. Y. Lin, M. M. Babu, and S. Yi. Revealing the Determinants of Widespread Alternative Splicing Perturbation in Cancer. *Cell Reports*, 21(3):798–812, 10 2017. ISSN 22111247. doi: 10.1016/j.celrep.2017.09.071.

P. Lichtenstein, I. V. H. Olm, I. K. V. Erkasalo, A. Nastasia, I. Liadou, J. Aakko, K. Aprio, M. Arkku, K. Oskenvuo, E. Ero, P. Ukkala, A. Xel, S. Kytthe, K. Ari, and H. Emminki. Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *http://dx.doi.org/10.1056/NEJM200007133430201*, 343(2):78–85, 8 2009. ISSN 0028-4793. doi: 10.1056/NEJM200007133430201. URL `https://www.nejm.org/doi/10.1056/NEJM200007133430201`.

Y. Liu and S. L. Shi. The roles of hnRNP A2/B1 in RNA biology and disease, 3 2021. ISSN 17577012.

Y. Liu, M. Gonzàlez-Porta, S. Santos, A. Brazma, J. C. Marioni, R. Aebersold, A. R. Venkitaraman, and V. O. Wickramasinghe. Impact of Alternative Splicing on the Human Proteome. *Cell reports*, 20(5):1229–1241, 8 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2017.07.025. URL `https://linkinghub.elsevier.com/retrieve/pii/S2211124717309919http://www.ncbi.nlm.nih.gov/pubmed/28768205http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5554779`.

Y. Liu, H. Li, F. Liu, L. B. Gao, R. Han, C. Chen, X. Ding, S. Li, K. Lu, L. Yang, H. M. Tian, B. B. Chen, X. Li, D. H. Xu, X. L. Deng, and S. L. Shi. Heterogeneous nuclear ribonucleoprotein A2/B1 is a negative regulator of human breast cancer metastasis by maintaining the balance of multiple genes and pathways. *EBioMedicine*, 51, 1 2020. ISSN 23523964. doi: 10.1016/j.ebiom.2019.11.044.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15 (12):550, 12 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL `http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8`.

S. W. Lowe and A. W. Lin. Apoptosis in cancer. *Carcinogenesis*, 21(3): 485–495, 3 2000. ISSN 0143-3334. doi: 10.1093/CARCIN/21.3.485. URL https://academic.oup.com/carcin/article/21/3/485/2365672.

B. Luo and A. S. Lee. The critical roles of endoplasmic reticulum chaperones and unfolded protein response in tumorigenesis and anticancer therapies, 2 2013. ISSN 09509232.

R. Luo and H. Zhao. Protein quantitation using iTRAQ: Review on the sources of variations and analysis of nonrandom missingness. *Statistics and its interface*, 5(1):99, 1 2012. ISSN 1938-7989. doi: 10.4310/sii.2012.v5.n1.a9. URL /pmc/articles/PMC3719432//pmc/articles/PMC3719432/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3719432/.

W. Luo and C. Brouwer. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14): 1830–1831, 7 2013. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt285. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt285.

K. Ma, O. Vitek, and A. I. Nesvizhskii. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC bioinformatics*, 13 Suppl 16(16):1–17, 11 2012. ISSN 14712105. doi: 10.1186/1471-2105-13-S16-S1/FIGURES/13. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S16-S1.

T. Maier, M. Güell, and L. Serrano. Correlation of mRNA and protein in complex biological samples, 12 2009. ISSN 00145793.

A. Marcelo, R. Koppenol, L. P. de Almeida, C. A. Matos, and C. Nóbrega. Stress granules, RNA-binding proteins and polyglutamine diseases: too much aggregation?, 6 2021. ISSN 20414889.

A. Marchant, F. Mougel, V. Mendonça, M. Quartier, E. Jacquin-Joly, J. A. da Rosa, E. Petit, and M. Harry. Comparing de novo and reference-based

transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus. *Insect Biochemistry and Molecular Biology*, 69: 25–33, 2 2016. ISSN 0965-1748. doi: 10.1016/J.IBMB.2015.05.009.

L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Römpp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz, and E. W. Deutschh. mzML—a Community Standard for Mass Spectrometry Data *. *Molecular & Cellular Proteomics*, 10(1):R110.000133, 1 2011. ISSN 1535-9476. doi: 10.1074/MCP.R110.000133. URL `http://www.mcponline.org/article/S1535947620313876/fulltexthttp://www.mcponline.org/article/S1535947620313876/abstracthttps://www.mcponline.org/article/S1535-9476(20)31387-6/abstract`.

F. J. Martinez, G. A. Pratt, E. L. Van Nostrand, R. Batra, S. C. Huelga, K. Kapeli, P. Freese, S. J. Chun, K. Ling, C. Gelboin-Burkhart, L. Fijany, H. C. Wang, J. K. Nussbacher, S. M. Broski, H. J. Kim, R. Lardelli, B. Sundararaman, J. P. Donohue, A. Javaherian, J. Lykke-Andersen, S. Finkbeiner, C. F. Bennett, M. Ares, C. B. Burge, J. P. Taylor, F. Rigo, and G. W. Yeo. Protein-RNA Networks Regulated by Normal and ALS-Associated Mutant HNRNPA2B1 in the Nervous System. *Neuron*, 92(4): 780–795, 11 2016. ISSN 10974199. doi: 10.1016/j.neuron.2016.09.050.

S. McGinnis and T. L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(Web Server issue):W20, 7 2004. ISSN 03051048. doi: 10.1093/NAR/GKH435. URL `/pmc/articles/PMC441573//pmc/articles/PMC441573/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC441573/`.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297, 9 2010. ISSN 10889051. doi: 10.1101/GR.107524.110. URL `/pmc/articles/`

PMC2928508//pmc/articles/PMC2928508/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928508/.

F. W. McLafferty. A Century of Progress in Molecular Mass Spectrometry. *http://dx.doi.org/10.1146/annurev-anchem-061010-114018*, 4:1–22, 6 2011. ISSN 19361327. doi: 10.1146/ANNUREV-ANCHEM-061010-114018. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-anchem-061010-114018.

R. Menegaux and J.-P. Vert. Embedding the de Bruijn graph, and applications to metagenomics. *bioRxiv*, page 2020.03.06.980979, 3 2020. ISSN 2692-8205. doi: 10.1101/2020.03.06.980979. URL https://www.biorxiv.org/content/10.1101/2020.03.06.980979v1https://www.biorxiv.org/content/10.1101/2020.03.06.980979v1.abstract.

F. Mignone, C. Gissi, S. Liuni, and G. Pesole. Untranslated regions of mRNAs, 2 2002. ISSN 14656906. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-3-reviews0004.

J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. Sonnhammer, S. C. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 1 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAA913. URL https://academic.oup.com/nar/article/49/D1/D412/5943818.

J. Mitchell Wells and S. A. McLuckey. Collision-Induced Dissociation (CID) of Peptides and Proteins. *Methods in Enzymology*, 402:148–185, 1 2005. ISSN 0076-6879. doi: 10.1016/S0076-6879(05)02005-7.

B. Møller, H. Fekjær, T. Hakulinen, H. Sigvaldason, H. H. Storm, M. Talbäck, and T. Haldorsen. Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. *Statistics in medicine*, 22 (17):2751–2766, 9 2003. ISSN 0277-6715. doi: 10.1002/SIM.1481. URL https://pubmed.ncbi.nlm.nih.gov/12939784/.

M. Mort, D. Ivanov, D. N. Cooper, and N. A. Chuzhanova. A meta-analysis of nonsense mutations causing human genetic disease. *Human mutation*, 29(8):1037–1047, 8 2008. ISSN 1098-1004. doi: 10.1002/HUMU.20763. URL https://pubmed.ncbi.nlm.nih.gov/18454449/.

R. Musich, L. Cadle-Davidson, and M. V. Osier. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Frontiers in Plant Science*, 12:692, 4 2021. ISSN 1664462X. doi: 10.3389/FPLS.2021.657240/BIBTEX.

J. O'Brien, H. Hayder, Y. Zayed, and C. Peng. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, 9(AUG):402, 8 2018. ISSN 16642392. doi: 10.3389/FENDO.2018.00402/BIBTEX.

S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics : MCP*, 1(5):376–386, 2002. ISSN 1535-9476. doi: 10.1074/MCP.M200025-MCP200. URL https://pubmed.ncbi.nlm.nih.gov/12118079/.

S. Orchard, W. Zhu, R. K. Julian, H. Hermjakob, and R. Apweiler. Further advances in the development of a data interchange standard for proteomics data. *PROTEOMICS*, 3(10):2065–2066, 10 2003. ISSN 1615-9861. doi: 10.1002/PMIC.200300588. URL https://onlinelibrary.wiley.com/doi/full/10.1002/pmic.200300588https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.200300588https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.200300588.

A. N. Pachikov, R. R. Gough, C. E. Christy, M. E. Morris, C. A. Casey, C. A. LaGrange, G. Bhat, A. V. Kubyshkin, I. I. Fomochkina, E. Y. Zyablitskaya, T. P. Makalish, E. P. Golubinskaya, K. A. Davydenko, S. N. Eremenko, J. J. M. Riethoven, A. S. Maroli, T. S. Payne, R. Powers, A. Y. Lushnikov,

A. J. Macke, and A. Petrosyan. The non-canonical mechanism of ER stress-mediated progression of prostate cancer. *Journal of Experimental and Clinical Cancer Research*, 40(1), 12 2021. ISSN 17569966. doi: 10.1186/s13046-021-02066-7.

N. Pällmann, M. Livgård, M. Tesikova, H. Zeynep Nenseth, E. Akkus, J. Sikkeland, Y. Jin, D. Koc, O. F. Kuzu, M. Pradhan, H. E. Danielsen, N. Kahraman, H. M. Mokhlis, B. Ozpolat, P. P. Banerjee, A. Uren, L. Fazli, P. S. Rennie, Y. Jin, and F. Saatcioglu. Regulation of the unfolded protein response through ATF4 and FAM129A in prostate cancer. *Oncogene*, 38 (35):6301–6318, 8 2019. ISSN 14765594. doi: 10.1038/s41388-019-0879-2.

C. E. Parker, M. R. Warren, and V. Mocanu. Mass Spectrometry for Proteomics. *Neuroproteomics*, pages 71–91, 1 2010. doi: 10.1533/9781908818058.171. URL `https://www.ncbi.nlm.nih.gov/books/NBK56011/`.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature methods*, 14(4):417, 2017. ISSN 15487105. doi: 10.1038/NMETH.4197. URL `/pmc/articles/PMC5600148//pmc/articles/PMC5600148/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5600148/`.

P. G. Pedrioli. Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods in molecular biology (Clifton, N.J.)*, 604:213–238, 2010. ISSN 1940-6029. doi: 10.1007/978-1-60761-444-9{_}15. URL `https://pubmed.ncbi.nlm.nih.gov/20013374/`.

M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015. ISSN 15461696. doi: 10.1038/nbt.3122. URL `/pmc/articles/PMC4643835/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4643835/`.

F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč. Foundations of JSON Schema. doi: 10.1145/2872427.2883029. URL http://dx.doi.org/10.1145/2872427.2883029.

A. Piovesan, F. Antonaros, L. Vitale, P. Strippoli, M. C. Pelleri, and M. Caracausi. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, 12(1):1–5, 6 2019. ISSN 17560500. doi: 10.1186/S13104-019-4343-8/TABLES/2. URL https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-019-4343-8.

R. J. Rebello, C. Oing, K. E. Knudsen, S. Loeb, D. C. Johnson, R. E. Reiter, S. Gillessen, T. Van der Kwast, and R. G. Bristow. Prostate cancer. *Nature Reviews Disease Primers*, 7(1), 12 2021. ISSN 2056676X. doi: 10.1038/s41572-020-00243-0.

R. B. Richardson. p53 mutations associated with aging-related rise in cancer incidence rates. *Cell Cycle*, 12(15):2468, 8 2013. ISSN 15514005. doi: 10.4161/CC.25494. URL /pmc/articles/PMC3841325//pmc/articles/PMC3841325/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841325/.

T. Rzymski, M. Milani, L. Pike, F. Buffa, H. R. Mellor, L. Winchester, I. Pires, E. Hammond, I. Ragoussis, and A. L. Harris. Regulation of autophagy by ATF4 in response to severe hypoxia. *Oncogene*, 29(31): 4424–4435, 8 2010. ISSN 09509232. doi: 10.1038/onc.2010.191.

S. Saha, D. A. Matthews, and C. Bessant. High throughput discovery of protein variants using proteomics informed by transcriptomics. *Nucleic Acids Research*, 46(10):4893–4902, 6 2018. ISSN 13624962. doi: 10.1093/nar/gky295. URL https://academic.oup.com/nar/article/46/10/4893/4990017.

J. J. Salk, M. W. Schmitt, and L. A. Loeb. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature Reviews Genetics 2018 19:5*, 19(5):269–285, 3 2018. ISSN 1471-0064.

doi: 10.1038/nrg.2017.117. URL https://www.nature.com/articles/nrg.2017.117.

R. Sánchez, R. Grau, and E. Morgado. A novel Lie algebra of the genetic code over the Galois field of four DNA bases. *Mathematical Biosciences*, 202(1):156–174, 7 2006. ISSN 00255564. doi: 10.1016/J.MBS.2006.03.017.

F. Sanchez-Vega, M. Mina, J. Armenia, W. K. Chatila, A. Luna, K. C. La, S. Dimitriadoy, D. L. Liu, H. S. Kantheti, S. Saghafinia, D. Chakravarty, F. Daian, Q. Gao, M. H. Bailey, W. W. Liang, S. M. Foltz, I. Shmulevich, L. Ding, Z. Heins, A. Ochoa, B. Gross, J. Gao, H. Zhang, R. Kundra, C. Kandoth, I. Bahceci, L. Dervishi, U. Dogrusoz, W. Zhou, H. Shen, P. W. Laird, G. P. Way, C. S. Greene, H. Liang, Y. Xiao, C. Wang, A. Iavarone, A. H. Berger, T. G. Bivona, A. J. Lazar, G. D. Hammer, T. Giordano, L. N. Kwong, G. McArthur, C. Huang, A. D. Tward, M. J. Frederick, F. McCormick, M. Meyerson, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K. S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, I. de Bruijn, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, J. S. Parker, M. D. Wilkerson, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall,

K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. Van Den Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. V. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T. Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. De Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca,

B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. Têtu, A. Bergeron, M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M. H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca, S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. Van Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. De Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. Giné, A. L. Guillermo, N. Van Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino, J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J. P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, J. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shannon, A. Spillane, J. Stretch, M. Synott, J. Thompson, J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drw-

iega, J. Miller, J. Smith, H. Zaren, J. W. Park, N. P. Hung, E. Kebebew, W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A. P. Gimenez-Roqueplo, A. Piché, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K. F. Becker, K. P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussioutas, D. Beer, T. Giordano, A. M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman,

F. Valdivieso, R. Dhir, J. Luketich, E. M. Pinero, M. Quintero-Aguilo, C. G. Carlotti, J. S. Dos Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. Van Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, A. Mariamidze, E. M. Van Allen, A. D. Cherniack, G. Ciriello, C. Sander, and N. Schultz. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, 173(2):321–337, 4 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.03.035.

J. P. Savaryn, T. K. Toby, and N. L. Kelleher. A researcher's guide to mass spectrometry-based proteomics. *PROTEOMICS*, 16(18):2435–2443, 9 2016. ISSN 1615-9861. doi: 10.1002/PMIC.201600113. URL `https://onlinelibrary.wiley.com/doi/full/10.1002/pmic.201600113https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201600113https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.201600113`.

S. Savic, L. Ouboussad, L. J. Dickie, J. Geiler, C. Wong, G. M. Doody, S. M. Churchman, F. Ponchel, P. Emery, G. P. Cook, M. H. Buch, R. M. Tooze, and M. F. McDermott. TLR dependent XBP-1 activation induces an autocrine loop in rheumatoid arthritis synoviocytes. *Journal of Autoimmunity*, 50:59–66, 2014. ISSN 10959157. doi: 10.1016/j.jaut.2013.11.002.

S. Schafer, K. Miao, C. C. Benson, M. Heinig, S. A. Cook, and N. Hubner. Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). *Current Protocols in Human Genetics*, 87(1):1–11, 10 2015. ISSN 1934-8266. doi: 10.1002/0471142905.hg1116s87. URL `https://onlinelibrary.wiley.com/doi/10.1002/0471142905.hg1116s87`.

X. Sheng, H. Z. Nenseth, S. Qu, O. F. Kuzu, T. Frahnow, L. Simon, S. Greene, Q. Zeng, L. Fazli, P. S. Rennie, I. G. Mills, H. Danielsen,

F. Theis, J. B. Patterson, Y. Jin, and F. Saatcioglu. IRE1α-XBP1s pathway promotes prostate cancer by activating c-MYC signaling. *Nature Communications*, 10(1), 12 2019. ISSN 20411723. doi: 10.1038/s41467-018-08152-3.

S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 1 2001. ISSN 03051048. doi: 10.1093/nar/29.1.308.

B. Spengler and A. Hester. Mass-based classification (MBC) of peptides: Highly accurate precursor ion mass values can be used to directly recognize peptide phosphorylation. *undefined*, 19(12):1808–1812, 12 2008. ISSN 10440305. doi: 10.1016/J.JASMS.2008.08.005.

L. Statello, C. J. Guo, L. L. Chen, and M. Huarte. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology 2020 22:2*, 22(2):96–118, 12 2020. ISSN 1471-0080. doi: 10.1038/s41580-020-00315-9. URL `https://www.nature.com/articles/s41580-020-00315-9`.

J. Stockley, M. E. M. Villasevil, C. Nixon, I. Ahmad, H. Y. Leung, and P. Rajan. The RNA-binding protein hnRNPA2 regulates β-catenin protein expression and is overexpressed in prostate cancer. *RNA Biology*, 11(6), 2014. ISSN 15558584. doi: 10.4161/rna.28800. URL `https://pubmed.ncbi.nlm.nih.gov/24823909/`.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 10 2005. ISSN 00278424. doi: 10.1073/PNAS.0506580102/SUPPL{_}FILE/06580FIG7.JPG. URL `www.pnas.orgcgidoi10.1073pnas.0506580102`.

H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 5 2021. ISSN 1542-4863. doi: 10.3322/CAAC.21660. URL https://pubmed.ncbi.nlm.nih.gov/33538338/.

S. Suttapitugsakul, H. Xiao, J. Smeekens, and R. Wu. Evaluation and optimization of reduction and alkylation methods to maximize peptide identification with MS-based proteomics. *Molecular bioSystems*, 13(12):2574, 2017. ISSN 17422051. doi: 10.1039/C7MB00393E. URL /pmc/articles/PMC5698164//pmc/articles/PMC5698164/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5698164/.

L. Sweetlove. Number of species on Earth tagged at 8.7 million. *Nature*, 8 2011. ISSN 0028-0836. doi: 10.1038/NEWS.2011.498.

J. Tazi, N. Bakkour, and S. Stamm. Alternative splicing and disease, 1 2009. ISSN 09254439. URL /pmc/articles/PMC5632948/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5632948/.

N. Thatra. HybridAssembly.

M. The, M. J. MacCoss, W. S. Noble, and L. Käll. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of the American Society for Mass Spectrometry*, 27 (11):1719–1727, 11 2016. ISSN 18791123. doi: 10.1007/s13361-016-1460-7.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer New York, New York, NY, 2000. ISBN 978-1-4419-3161-0. doi: 10.1007/978-1-4757-3294-8. URL http://link.springer.com/10.1007/978-1-4757-3294-8.

A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by

MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 4 2003. ISSN 00032700. doi: 10.1021/AC0262560. URL https://pubs.acs.org/doi/abs/10.1021/ac0262560.

H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178, 3 2013. ISSN 14675463. doi: 10.1093/BIB/BBS017. URL /pmc/articles/PMC3603213//pmc/articles/PMC3603213/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3603213/.

K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge, 2015. ISSN 14282526.

O. K. Tørresen, B. Star, P. Mier, M. A. Andrade-Navarro, A. Bateman, P. Jarnot, A. Gruca, M. Grynberg, A. V. Kajava, V. J. Promponas, M. Anisimova, K. S. Jakobsen, and D. Linke. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases, 12 2019. ISSN 13624962. URL https://academic.oup.com/nar/article/47/21/10994/5580909.

N. H. Tran, X. Zhang, L. Xin, B. Shan, and M. Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 114(31):8247–8252, 8 2017. ISSN 10916490. doi: 10.1073/PNAS.1705691114/-/DCSUPPLEMENTAL. URL https://www.pnas.org/content/114/31/8247https://www.pnas.org/content/114/31/8247.abstract.

J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyras. SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):40, 3 2018. ISSN 1474760X. doi: 10.1186/s13059-018-1417-1. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1417-1.

S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox. The Perseus computational platform for comprehensive analysis of (prote)omics data, 8 2016. ISSN 15487105.

A. Uemura, M. Oku, K. Mori, and H. Yoshida. Unconventional splicing of XBP1 mRNA occurs in the cytoplasm during the mammalian unfolded protein response. *Journal of Cell Science*, 122(16):2877–2886, 8 2009. ISSN 00219533. doi: 10.1242/jcs.040584.

E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, 6 2016. ISSN 15487105. doi: 10.1038/nmeth.3810.

M. Vaudel, J. M. Burkhart, R. P. Zahedi, E. Oveland, F. S. Berven, A. Sickmann, L. Martens, and H. Barsnes. PeptideShaker enables reanalysis of MS-derived proteomics data sets: To the editor, 1 2015. ISSN 15461696.

K. V. Voelkerding, S. A. Dames, and J. D. Durtschi. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, 55(4):641–658, 4 2009. ISSN 0009-9147. doi: 10.1373/CLINCHEM.2008.112789. URL `https://academic.oup.com/clinchem/article/55/4/641/5629392`.

B. Wirth, L. Brichta, and E. Hahnen. Spinal muscular atrophy and therapeutic prospects. *Progress in molecular and subcellular biology*, 44:109–132, 2006. ISSN 0079-6484. doi: 10.1007/978-3-540-34449-0{_}6. URL `https://pubmed.ncbi.nlm.nih.gov/17076267/`.

B. Wolozin and P. Ivanov. Stress granules and neurodegeneration, 11 2019. ISSN 14710048.

P. Wu, L. Pu, B. Deng, Y. Li, Z. Chen, and W. Liu. PASS: A Proteomics Alternative Splicing Screening Pipeline. *PROTEOMICS*, 19(13):

1900041, 7 2019. ISSN 1615-9861. doi: 10.1002/PMIC.201900041. URL https://onlinelibrary.wiley.com/doi/full/10.1002/pmic.201900041https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201900041https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.201900041.

T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo, and G. Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 8 2021. ISSN 2666-6758. doi: 10.1016/J.XINN.2021.100141.

T. D. Wu and C. K. Watanabe. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 5 2005. ISSN 13674803. doi: 10.1093/bioinformatics/bti310. URL https://pubmed.ncbi.nlm.nih.gov/15728110/.

J. W. Wynne, B. J. Shiell, G. A. Marsh, V. Boyd, J. A. Harper, K. Heesom, P. Monaghan, P. Zhou, J. Payne, R. Klein, S. Todd, L. Mok, D. Green, J. Bingham, M. Tachedjian, M. L. Baker, D. Matthews, and L. F. Wang. Proteomics informed by transcriptomics reveals Hendra virus sensitizes bat cells to TRAIL-mediated apoptosis. *Genome Biology*, 15(11):532, 2014. ISSN 1474760X. doi: 10.1186/S13059-014-0532-X. URL /pmc/articles/PMC4269970//pmc/articles/PMC4269970/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4269970/.

T. Xiao and W. Zhou. The third generation sequencing: the advanced approach to genetic diseases. *Translational Pediatrics*, 9(2):163, 4 2020. ISSN 22244344. doi: 10.21037/TP.2020.03.06. URL /pmc/articles/PMC7237973//pmc/articles/PMC7237973/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7237973/.

P. Yao, A. A. Potdar, P. S. Ray, S. M. Eswarappa, A. C. Flagg, B. Willard, and P. L. Fox. The HILDA Complex Coordinates a Conditional Switch in the 3-Untranslated Region of the VEGFA mRNA. *PLoS Biology*, 11(8), 8 2013. ISSN 15449173. doi: 10.1371/journal.pbio.1001635.

M. Yin, M. Cheng, C. Liu, K. Wu, W. Xiong, J. Fang, Y. Li, and B. Zhang. HNRNPA2B1 as a trigger of RNA switch modulates the miRNA-mediated regulation of CDK6. *iScience*, 24(11):103345, 11 2021. ISSN 25890042. doi: 10.1016/j.isci.2021.103345.

G. Zhang, K. Deinhardt, and T. A. Neubert. Stable Isotope Labeling by Amino Acids in Cultured Primary Neurons. *Methods in molecular biology (Clifton, N.J.)*, 1188:57, 2014. ISSN 10643745. doi: 10.1007/978-1-4939-1142-4{_}5. URL /pmc/articles/PMC4212509//pmc/articles/PMC4212509/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4212509/.

Y. Zhang, L. Y. Chen, X. Han, W. Xie, H. Kim, D. Yang, D. Liu, and Z. Songyang. Phosphorylation of TPP1 regulates cell cycle-dependent telomerase recruitment. *Proceedings of the National Academy of Sciences*, 110(14):5457–5462, 4 2013. ISSN 0027-8424. doi: 10.1073/PNAS.1217733110. URL https://www.pnas.org/content/110/14/5457https://www.pnas.org/content/110/14/5457.abstract.

J. Zhu, A. Mayeda, and A. R. Krainer. Exon Identity Established through Differential Antagonism between Exonic Splicing Silencer-Bound hnRNP A1 and Enhancer-Bound SR Proteins. *Molecular Cell*, 8(6):1351–1361, 12 2001. ISSN 1097-2765. doi: 10.1016/S1097-2765(01)00409-9.

F. Zickmann and B. Y. Renard. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics*, 31(12):i106, 6 2015. ISSN 14602059. doi: 10.1093/BIOINFORMATICS/BTV236. URL /pmc/articles/PMC4765881//pmc/articles/PMC4765881/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765881/.

P. Zuccotti, C. Colombrita, S. Moncini, A. Barbieri, M. Lunghi, C. Gelfi, S. De Palma, A. Nicolin, A. Ratti, M. Venturin, and P. Riva. HnRNPA2/B1 and nELAV proteins bind to a specific U-rich element in CDK5R1 3'-UTR and oppositely regulate its expression. *Biochimica et*

Biophysica Acta - Gene Regulatory Mechanisms, 1839(6):506–516, 2014. ISSN 18764320. doi: 10.1016/j.bbagrm.2014.04.018.