

Effects of Trastuzumab on Alternative Splicing in HER2+ BT474 Breast Cancer Cells

by

Benjamin D. Piticar

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Biology

Waterloo, Ontario, Canada, 2024

© Benjamin D. Piticar 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Alternative splicing (AS) is a fundamental process that enhances transcriptomic diversity and protein isoform complexity in eukaryotic cells. Deregulation of AS plays a critical role in various diseases, particularly cancer, where erroneous splicing patterns contribute to oncogenesis, tumor progression, and therapeutic resistance. Human epidermal growth factor receptor (HER)2, overexpressed in more than 20% of breast cancers, has clinical relevance in AS. The humanized monoclonal antibody (mAb) Trastuzumab has been used for decades to fight HER2+ breast cancer with positive results. However, there are many unknowns surrounding the mechanism of this drug and the effects it has on alternative splicing. Next generation long read sequencing technologies like Oxford Nanopore allow researchers to sequence full length transcripts. The ability to sequence long intron/exon spans and repeating regions enables long read sequencing technology to provide new insights into the AS patterns of genes.

Identifying AS events is important for understanding changes in function and pathways affected by differing conditions (e.g., drug treatment, diseased cells). Software tools tailored to long read data like Long-read Isoform Quantification and Analysis (LIQA), Full-Length Alternative Isoform Analysis of RNA (FLAIR), FLAME, and Nanosplicer enhance data analysis abilities to detect AS events. They analyze sequencing data by mapping reads to genomes, identifying splice junctions, and clustering AS events. These tools then compare AS events between conditions (e.g., treated vs. untreated cells) to pinpoint significant AS variations.

Accurate transcriptome sequencing is vital for research in drug development and diagnostics. The optimization of sequencing methods is an ongoing endeavour that requires continuous development. A chapter of this thesis compares the impact of omitting polyadeny-

lation (poly(A)) enrichment in Oxford Nanopore Technologies (ONT)’s direct complementary deoxyribonucleic acid (cDNA) sequencing protocol to improve library prep efficiency and sequencing accuracy. Analysis indicated that excluding poly(A) selection does not negatively affect sequencing metrics but enhances read length and enables the sequencing of histone messenger ribonucleic acid (mRNA). Further, significant differences in poly(A) tail lengths between selected and unselected samples were found, suggesting a bias against shorter tails with enrichment. Gene composition and unique gene identification remain consistent across conditions. These findings support the potential advantages of omitting poly(A) enrichment in transcriptome sequencing while advocating for further validation.

With optimized methods, the impact of Trastuzumab on AS profiles in BT474 and SKBR3 cell lines was tested. Using ONT long-read sequencing, significant AS events were identified in response to Trastuzumab treatment. The findings reveal differential isoform expression in genes involved in cellular signaling, RNA processing, and stress response pathways. Notably, nuclear paraspeckle assembly transcript 1 (NEAT1) and pre-mRNA processing factor 38B (PRPF38B) exhibited distinct AS patterns across multiple cell lines, suggesting HER2-mediated regulatory mechanisms. This study underscores the utility of ONT sequencing and high throughput data processing as an effective and efficient means for explaining complex AS landscapes affected by therapeutic treatments.

Acknowledgements

While many individuals helped me along the way, some need to be mentioned for their impact. First, I would like to thank my wife Jessica for her love and support that made this all possible.

Further, I would like to thank Dr. Brendan McConkey for his advice and support, and for providing me with the opportunity to perform this work. I would also like to thank my committee members, Dr. Andrew Doxey and Dr. Dale Martin, for sharing their advice and knowledge.

I would also like to thank the other members of the McConkey Lab - Desmond Singh, Philip Cotra, Karstin Rinas, and Yifan Yang - for their support, insight, and advice.

Finally, I would like to thank Dr. Barbara Moffatt for her continued support and enthusiasm for my work, constantly pushing me to be a better scientist.

Dedication

This is dedicated to my loving wife Jessica and baby Arthur, born August 11, 2024, who waited two days after my defence to be born.

Table of Contents

Author’s Declaration	ii
Abstract	iii
Acknowledgements	v
Dedication	vi
List of Figures	xi
List of Tables	xiv
List of Abbreviations	xvi
1 Introduction	1
1.1 Alternative Splicing in Breast Cancer	1
1.1.1 mRNA	1
1.1.2 Alternative Splicing in mRNA	2

1.1.3	AS in Cancer	4
1.1.4	Human Epidermal Growth Factor Receptor 2 in Breast Cancer . . .	7
1.2	Treatment of HER2+ Tumors	10
1.2.1	Monoclonal Antibodies	10
1.2.2	Trastuzumab: Mechanisms and Effects	10
1.2.3	Trastuzumab and HER2 Isoforms	14
1.3	Transcriptomics	14
1.3.1	Third-Generation Sequencing Methods	14
1.3.2	Common Practices in NGS Data Processing	16
1.3.3	Gene Expression Analysis	18
1.3.4	Differential Alternative Splicing Analysis	18
1.3.5	High Throughput Data Analysis	19
1.4	Conclusion	20
2	Differential Alternative Splicing Software Review	21
2.1	Introduction	21
2.2	Algorithm Overview	22
2.2.1	Mapping Reads to Reference Genome	22
2.2.2	Identify Splice Junctions	24
2.2.3	Cluster Splice Junctions into Splice Events	25
2.2.4	Identify differential splicing events between Samples	25

2.3	Examples	25
2.3.1	LIQA	26
2.3.2	FLAIR	28
2.3.3	Flame	31
2.3.4	Nanosplicer	33
2.4	Conclusion	34
3	Optimization of Sequencing Methods	36
3.1	Introduction	36
3.2	Methods	40
3.3	Results and Discussion	42
3.4	Conclusions	54
4	Testing Affects of Trastuzumab	56
4.1	Introduction	56
4.2	Methods	58
4.3	Results and Discussion	63
4.4	Conclusions	90
5	Conclusion and Future Directions	92
	References	94

APPENDICES	119
A Supplementary Figures	120

List of Figures

1.1	Types of alternative splicing	3
1.2	Diagram of spliceosome formation and splicing	5
1.3	Diagram of HER2 pathways	8
1.4	Diagram of signaling abnormalities resulting from HER2 overexpression . .	9
2.1	Example of a heatmap	28
3.1	Protocol for poly(A) ribonucleic acid (RNA) enrichment	38
3.2	Protocol for cDNA synthesis	39
3.3	Workflow overview for poly(A) selection experiment	43
3.4	Overall raw read length histogram for each condition.	45
3.5	Average read length scatterplot	46
3.6	Transcripts per million scatterplot	48
3.7	PolyA tail length boxplot for each sequenced library	50
3.8	Gene make-up plot for selected and unselected samples	52
3.9	Upset plot for all samples	53

4.1	Workflow overview of AS experiment	59
4.2	Mycoplasma test results agarose gel	64
4.3	Agarose gels for RNA extraction of all replicates	65
4.4	Read length histogram for all adaptor trimmed samples with binned stacked Phred scores	68
4.5	Boxplot of average Phred score per read for each sample	69
4.6	Boxplot of read lengths for each sample	70
4.7	Histogram of raw Phred scores across all samples	71
4.8	Histogram of processed Phred scores across all samples	72
4.9	Scatterplot of read length vs. average Phred score per read for all raw samples	73
4.10	Scatterplot of read length vs. average Phred score per read for all processed samples	74
A.1	Sashimi plot of AS isoforms in the NEAT1 gene, cell line BT474	121
A.2	Sashimi plot of AS isoforms in the NEAT1 gene, cell line MCF7	122
A.3	Sashimi plot of AS isoforms in NEAT1 gene between cell lines BT474 and MCF7	123
A.4	Sashimi plot of AS isoforms in the PRPF38B gene, cell line SKBR3	124
A.5	Sashimi plot of AS isoforms in the PRPF38B gene, cell line MCF7	125
A.6	Sashimi plot of AS isoforms in PRPF38B gene between cell lines BT474 and MCF7	126

A.7 Sashimi plot of AS isoforms in PRPF38B gene between cell lines BT474 and SKBR3	127
A.8 Sashimi plot of AS isoforms in the ADP ribosylation factor 4 (ARF4) gene, cell line BT474	128
A.9 Sashimi plot of AS isoforms in the ARF4 gene, cell line SKBR3	129
A.10 Sashimi plot of AS isoforms in ARF4 gene between cell lines BT474 and MCF7	130
A.11 Sashimi plot of AS isoforms in ARF4 gene between cell lines BT474 and SKBR3	131
A.12 Sashimi plot of AS isoforms in the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) gene, cell line SKBR3	132
A.13 Sashimi plot of AS isoforms in the peptidyl-prolyl cis-trans isomerase (PPIA) gene, cell line BT474	133

List of Tables

3.1	Experiment RNA and read amounts overview	44
3.2	Genes in unselected group with the largest log difference in transcripts per million (TPM)	47
3.3	Genes in selected group with the largest log difference in TPM	49
3.4	Multiple comparison of means - Tukey HSD	51
4.1	Cell lines used and their gene expression types	63
4.2	AS sequencing overview	66
4.3	Genes expressing significant alternative splicing between cell line BT474 treatment and control groups	77
4.4	Genes expressing significant alternative splicing between cell line SKBR3 treatment and control groups	79
4.5	Genes expressing significant alternative splicing between cell line MCF7 treatment and control groups	80
4.6	Genes expressing significant alternative splicing between cell line BT474 and SKBR3 control groups	82

4.7	Genes expressing significant alternative splicing between cell line BT474 and MCF7 control groups	84
4.8	Genes expressing significant alternative splicing between cell line MCF7 and SKBR3 control groups	86
4.9	gene ontology (GO)/Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis results for cell line SKBR3	89
4.10	GO/KEGG enrichment analysis results for cell line MCF7	89
4.11	GO/KEGG enrichment analysis results for cell line BT474 vs SKBR3	89
4.12	GO/KEGG enrichment analysis results for cell line BT474 vs MCF7	89
4.13	GO/KEGG enrichment analysis results for cell line MCF7 vs SKBR3	90

List of Abbreviations

AKT protein kinase B.

ALS Amyotrophic Lateral Sclerosis.

ARF4 ADP ribosylation factor 4.

AS alternative splicing.

BAM Binary Alignment/Map.

BED browser extensible data.

cDNA complementary deoxyribonucleic acid.

CDS protein-coding sequences.

CIGAR Compact Idiosyncratic Gapped Alignment Report.

DAS differential alternative splicing.

DNA deoxyribonucleic acid.

dsDNA double stranded deoxyribonucleic acid.

EGFR epidermal growth factor receptor.

ER estrogen receptor gene.

FLAIR Full-Length Alternative Isoform Analysis of RNA.

GAPDH glyceraldehyde 3-phosphate dehydrogenase.

GC guanine-cytosine.

GO gene ontology.

GTF gene transfer format.

HER human epidermal growth factor receptor.

HnRNP heterogeneous nuclear ribonucleoproteins.

IIS isoform identification software.

KEGG Kyoto Encyclopedia of Genes and Genomes.

LIQA Long-read Isoform Quantification and Analysis.

mAb monoclonal antibody.

MAPK mitogen-activated protein kinases.

MDM2 murine double minute 2.

mRNA messenger ribonucleic acid.

NEAT1 nuclear paraspeckle assembly transcript 1.

NGS next generation sequencing.

ONT Oxford Nanopore Technologies.

PCR polymerase chain reaction.

PI3K phosphoinositide 3-kinase.

poly(A) polyadenylation.

PPIA peptidyl-prolyl cis-trans isomerase.

PR progesterone.

PRPF38B pre-mRNA processing factor 38B.

RNA ribonucleic acid.

rRNA ribosomal ribonucleic acid.

SAM Sequence Alignment/Map.

snRNP small nuclear ribonucleoproteins.

SRS short-read sequencing.

TPM transcripts per million.

Chapter 1

Introduction

1.1 Alternative Splicing in Breast Cancer

1.1.1 mRNA

In eukaryotic cells, genetic information stored in deoxyribonucleic acid (DNA) is transcribed into RNA by RNA polymerase 1 [1]. The maturation of most mRNA molecules occur co-transcriptionally. The addition of the poly(A) tail to newly synthesized RNA molecules involves multiple proteins and sequence elements, including a polyadenylation signal and downstream GU- or U-rich sequences. Cleavage and polyadenylation occur co-transcriptionally, with poly(A) polymerase adding adenosines to form the tail, typically 200–250 nucleotides long. However, some transcripts have shorter tails as a result of poly(A) limiting elements. [2]. AS, discussed in the next section, involves the removal of introns and splicing together the remaining exons. Finally, a 7-methyl guanosine cap is added to the 5' end of the processed mRNA molecule. The mature mRNA molecule

is then translated into an amino acid sequence by ribosomes, and the resulting proteins then serve various functional and structural roles in the cell, possibly causing additional post-translational modifications to the protein. This is known as the central dogma of molecular biology [1].

This energy-intensive process occurs continuously within the cell, and mRNA molecules are translated into multiple proteins, and the mRNA molecule is quickly degraded as part of the regulatory mechanisms for protein synthesis [3]. To be more efficient, systems are in place at each phase of the central dogma to regulate the expression level of the proteins produced. This thesis will discuss the regulatory pathway AS, which occurs during the post-transcription modification phase of mRNA.

1.1.2 Alternative Splicing in mRNA

Within the human genome, there are estimated 20,000–25,000 protein-encoding genes [4]. However, the human proteome is estimated to have an upper limit of 6.13 million unique protein isoforms [4]. Three mechanisms are attributed to the disparity between the number of genes and proteins- amino acid polymorphisms, post-translational modifications, and AS [4].

More than 90% of multi-exon human genes undergo AS which is a regulatory mechanism that allows multiple different mRNA isoforms to be produced from a single gene, increasing the diversity of the transcriptome and proteome of a cell. AS is the mechanism that determines which introns are removed and which exons are joined to form the final mRNA molecule [5].

Mechanisms of AS have been categorized into several types, shown in Figure 1.1. Many genes exhibit intricate transcriptional control due to the use of alternative promoters.

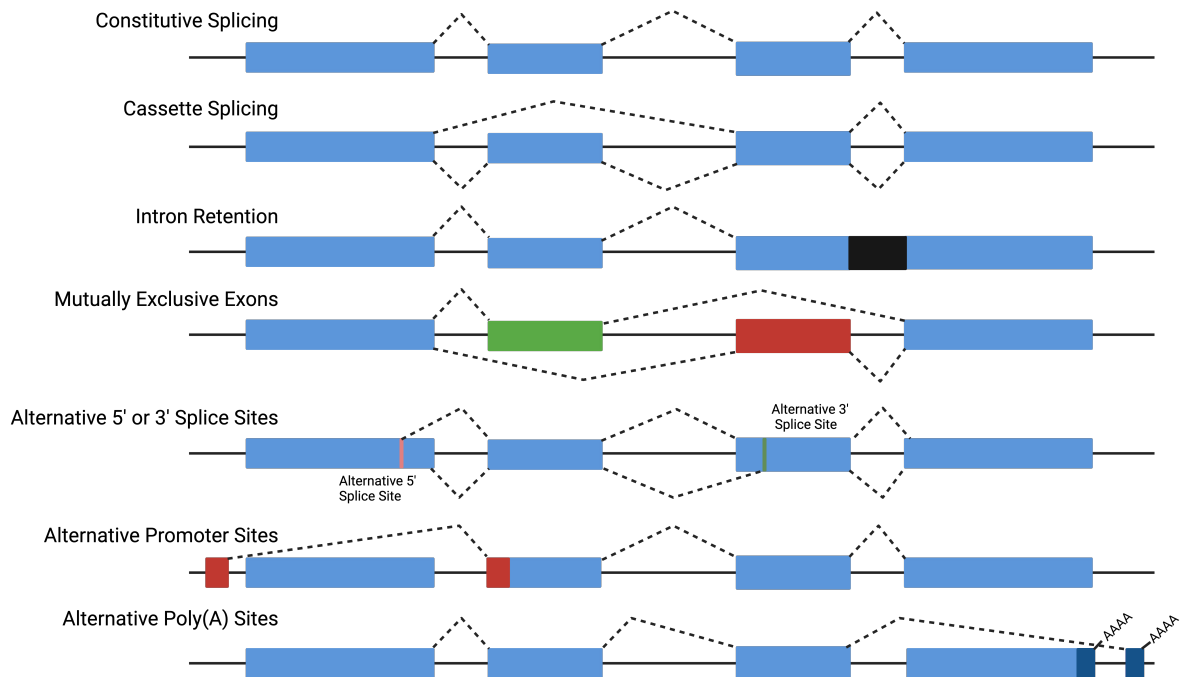


Figure 1.1: **Types of alternative splicing.** Constitutive splicing, the default mechanism producing a single, unvarying transcript without AS events. Cassette splicing or exon skipping, the complete exclusion of one or more exons. Intron retention, the retention of a complete intron within the mature mRNA transcript [6]. Mutually exclusive exons, when two or more exons are mutually exclusive in the mature mRNA. Alternative 5' or 3' splice sites, partial intron retention or deletion by alternative splice sites at the ends of introns. Alternative promoter splicing/usage, isoforms are transcribed using different promoters, which include alternative first exons [6]. Alternative poly(A) splicing, mRNA isoforms with variable lengths of the 3' untranslated region by choosing different poly(A) sites, which affects the stability, localization, and translation efficiency of the mRNA [7, 8, 6]. Light blue boxes represent exons and the thin line connecting them represent introns. The red and green box in the mutually exclusive exon represent different exons that would not be found in the same isoform. Created with BioRender.com.

Approximately 30–50% of human genes and nearly half of mouse genes possess multiple alternative promoters, spanning genomic distances of up to thousands of kilobases [6].

As illustrated in Figure 1.2, the splice event is split into 2 major steps. First, a spliceosome is formed that is mainly composed of five small nuclear ribonucleoproteins (snRNP) U1, U2, U4, U5, and U6, named for their uridine rich sequences. It is responsible for identifying the 5' splice site, branch point sequence, and 3' splice site. Additionally, DExD/H-type RNA-dependent ATPases/helicases play a role in spliceosome assembly and splicing execution [6, 9]. AS events are determined by the interaction between cis-acting elements (exonic splicing enhancers, intronic splicing enhancers, exonic splicing silencers and intronic splicing silencers) and transacting factors, including serine/arginine rich proteins and heterogeneous nuclear ribonucleoproteins (HnRNP) [6, 9]. Serine/arginine rich proteins interact with enhancer elements and spliceosomal components, influencing splice site selection, while HnRNPs bind to intronic splicing silencers to inhibit the actions of serine/arginine rich proteins, affecting the choices of splice sites. Furthermore, other factors such as HnRNP H, HnRNP F, HnRNP M, and HnRNP L play a diverse role in AS regulation, including alteration of the isoform ratio and responses to cellular signals [6, 9]. External stimuli such as; receptor stimulation, cellular stress (pH, temperature, metal ion and osmotic conditions), and changes in neuronal activity have been observed to influence AS, even during cell development and differentiation [10].

1.1.3 AS in Cancer

Stringent regulation of AS is vital for normal tissue function, as irregular splicing is found in many diseases. The SpliceDisease database, includes 2,337 entries of splicing mutations linked to 303 genes and 370 diseases. These associations have been experimentally vali-

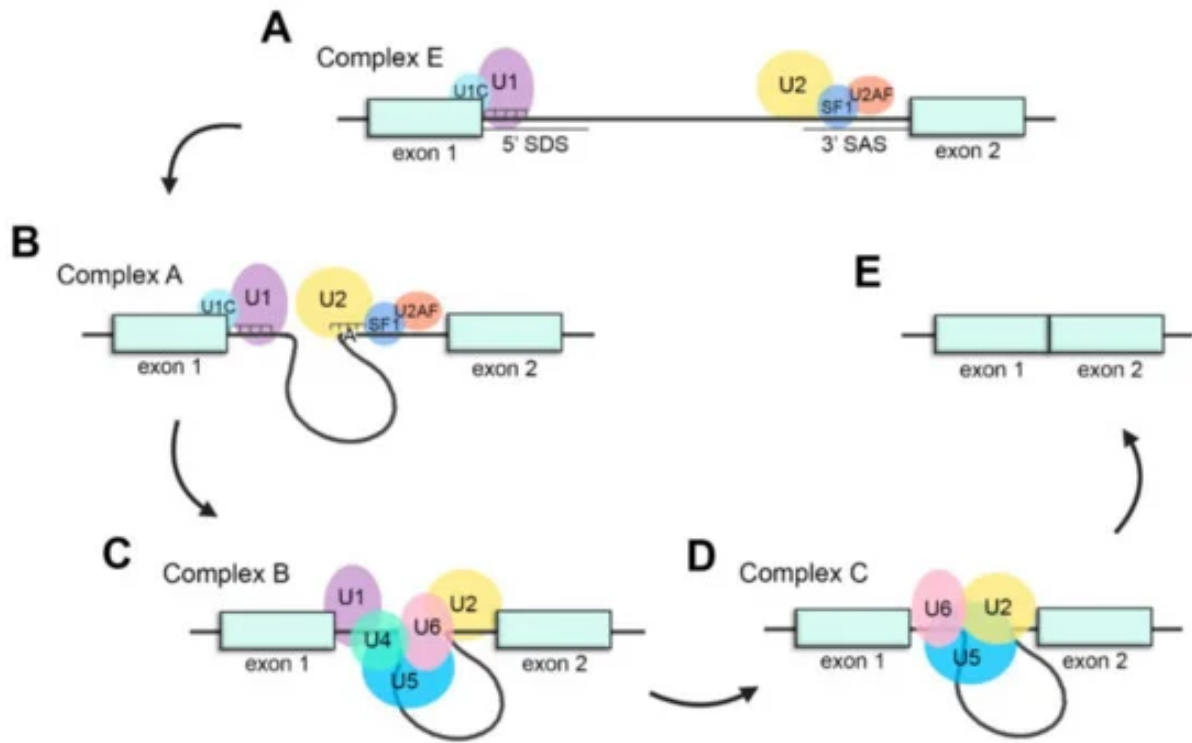


Figure 1.2: **Diagram of spliceosome formation and Splicing.** (A), the spliceosome assembly begins with the recognition of the splice donor site by U1 snRNP and the splice acceptor site by U2 snRNP, leading to the formation of complex E. (B), U2 snRNP identifies the adenosine at the branch-site, transitioning to complex A. (C), the U4/U6·U5 tri-snRNP integrates into the spliceosome, resulting in complex B formation. (D), U4 and U1 dissociate, and U6 replaces U1 to recognize the splice donor site. This interaction with U2 forms complex C, catalyzing the splicing reaction. (E), exons are ligated, and both intronic pre-mRNA and spliceosomal snRNPs are released, completing the splicing process. *Source:* [11]. Licensed under CC BY 4.0.

dated in a total of 898 publications [12]. As stated above, irregular splicing can result from changes in the concentration, composition, localization, or activity of regulatory splicing factors, as well as mutations in the components of the core splicing machinery [6]. This disruption leads to altered splice site recognition and produces irregular protein isoforms, contributing to various disease states, including cancer. A study examined 8,705 patients' RNA and whole-exome sequencing data and found that tumors have up to 30% more AS events than normal samples [13]. AS has been shown to play an important role in cancer development, influencing characteristics such as cell proliferation, migration, changes in DNA methylation, and resistance to apoptosis and chemotherapy [14]. Although a complete understanding has not yet been achieved, it is becoming more evident that alterations in AS due to mutations or fluctuations in splicing factors can give rise to isoforms whose functions play a role in the onset and advancement of cancer [14]. Understanding the fundamental mechanisms of splicing in cancer progression holds promise for advancing cancer biology and developing diagnostic, prognostic, and minimally toxic therapeutic approaches [6]. For example, the murine double minute 2 (MDM2) gene is an oncoprotein that inhibits the transcriptional transactivation of the p53 tumor suppressor, relocating p53 from the nucleus to the cytoplasm, and inducing polyubiquitylation of p53 [15]. MDM2 is both a biomarker for cancer and a target for therapy, since alternative spliced forms that lack a p53 binding region are found more frequently in poorly differentiated and late-stage tumors [16, 17, 18].

1.1.4 Human Epidermal Growth Factor Receptor 2 in Breast Cancer

The HER2 oncogene is found in the 17q12 chromosome region. It is one of four membrane receptor tyrosine kinases (RTKs) [19]. Figure 1.3 shows the structure and pathway of the HER2 gene.

More than 20% of breast cancer tumors overexpress the HER2 gene (HER2+). A study used *in situ* hybridization and found that the cell lines BT474 and SKBR3, which are known to be HER2+, had ≈ 45 copies of the gene, representing an 8-11 fold increase, compared to the normal human skin fibroblast cell lines [21]. The overexpression of HER2 impacts many pathways, illustrated in Figure 1.4. Another gene found to be overexpressed in $\approx 80\%$ of breast cancer is the estrogen receptor gene (ER) [22]. This gene along with progesterone (PR) makes up what is known as hormone-positive breast cancer [23]. Breast cancers that do not overexpress ER, HER2, and PR are classified as “triple negative” and have been more closely associated with the BRCA1 gene [24]. A study in 2021 that looked at over 34,000 patients with breast cancer found that there were significant differences in survival rates between ER-/PR-/HER2+ and ER-/PR+/HER2+ breast cancer, compared to ER+/PR+/HER2+ breast cancers [25]. ER and HER2 expression levels can be used to determine whether hormonal therapy will be effective against the cancer tumor [23].

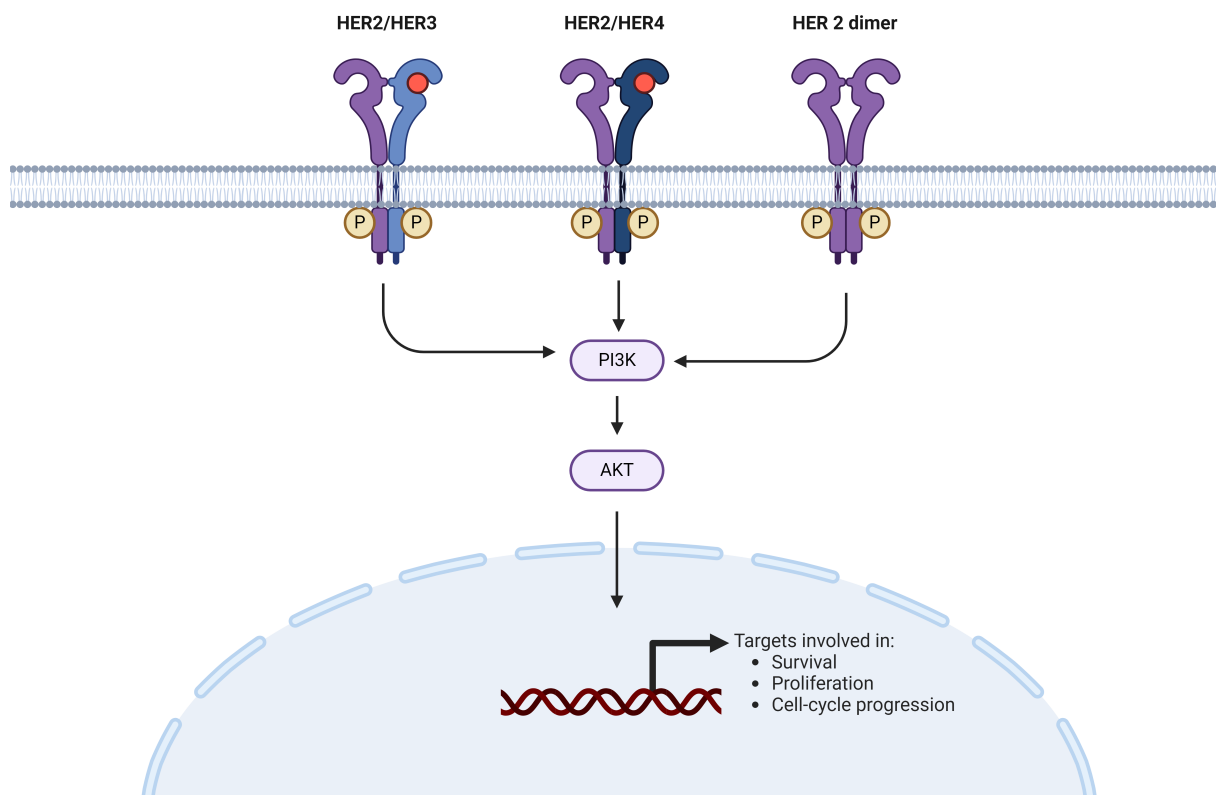


Figure 1.3: **Diagram of HER2 pathways.** The HER2 protein consists of an extracellular ligand binding domain, a transmembrane domain, and an intracellular catalytic domain of tyrosine kinase. When the ligand is activated, the receptors dimerize to form homodimers and/or heterodimers which leads to transphosphorylation that impacts the phosphoinositide 3-kinase (PI3K)/protein kinase B (AKT) pathway. This pathway has a role in cell metabolism, growth, proliferation, and survival [19, 20]. Created with BioRender.com.

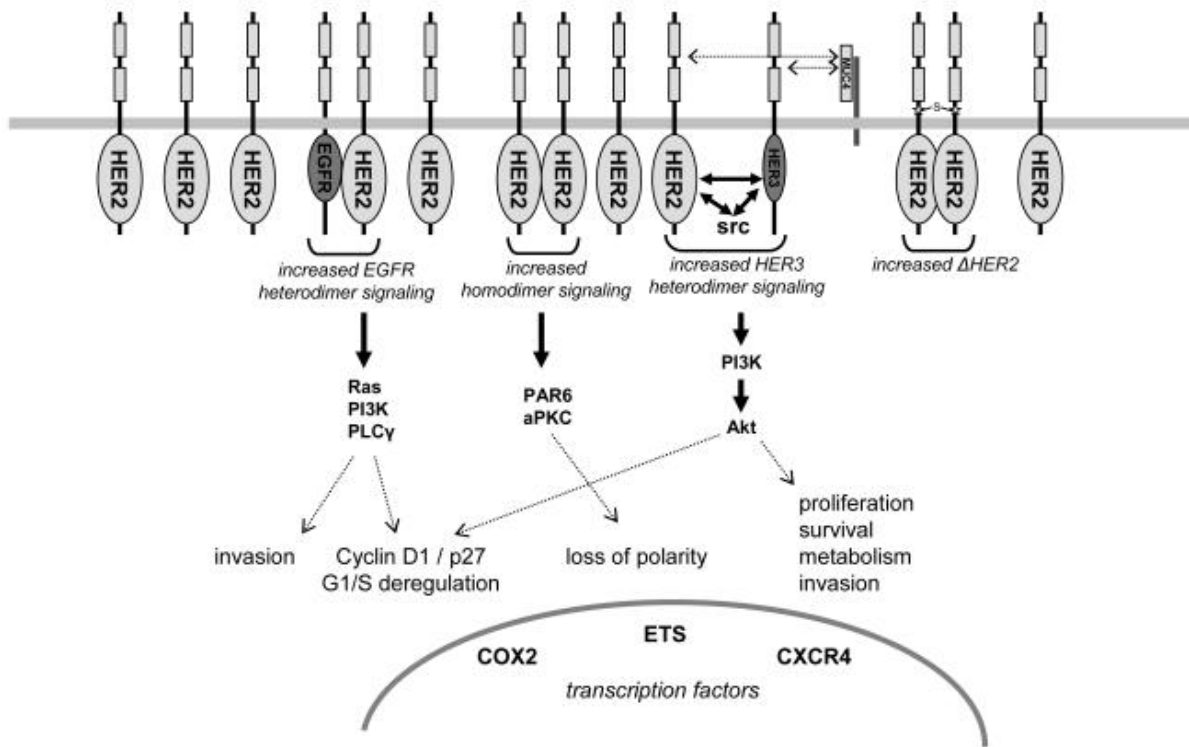


Figure 1.4: **Diagram of signaling abnormalities resulting from HER2 overexpression.** When HER2 is overexpressed, there is an increased formation of various dimers that contain HER2. This includes an increase in HER2-epidermal growth factor receptor (EGFR) dimers, which promote cell proliferation and invasion. Additionally, elevated levels of HER2 homodimers disrupt cell polarity. In addition, HER2-HER3 dimers are increased, leading to enhanced proliferation, cell survival, invasion, and metabolic functions. Enhanced expression of HER2 also leads to an increase in the Δ HER2 isoform, which exhibits more potent signaling characteristics. Furthermore, in cells overexpressing HER2, numerous transcription factors are activated, resulting in a wide range of changes in gene expression [26]. Reproduced with permission from Springer Nature (<https://www.nature.com/articles/1210477>).

1.2 Treatment of HER2+ Tumors

1.2.1 Monoclonal Antibodies

In the last 25 years, mAbs have become a predominant form of cancer treatment. One driving factor of the explosion in antibody therapeutics is fewer adverse effects due to their high specificity [27]. In 2020, human, humanized, chimeric, and murine antibodies accounted for 51, 34.7, 12.5, and 2.8% of all mAbs in clinical use, respectively [27]. Human and humanized mAbs are mainly used as a result of the human antimouse antibody response that can develop in patients. Human antimouse antibody's clear mouse mAbs faster than other mAbs and can lead to patient allergic responses and a lower tumor penetration by the mAb [27]. Humanized mAbs do not share these challenges because of their increased sequence similarity to human antibodies. Complementarity determining region grafting is one of the most popular techniques in the production of humanized mAbs and was developed in 1986 by Gregory P. Smith [27]. This method involves transplanting non-human complementarity determining region sequences into human antibody framework sequences, allowing the designed antibody to continue to have a high affinity for the target antigen [27].

1.2.2 Trastuzumab: Mechanisms and Effects

Trastuzumab (Herceptin[®]) is a humanized mAb that has been used in treatments against HER2+ metastatic breast cancer since its approval in 1998 [28]. Trastuzumab interacts with three specific regions within domain IV through electrostatic and hydrophobic binding [28]. Initially, trastuzumab was thought to block tumor growth by preventing homodimerization of HER2. However, research has yielded conflicting results regarding trastuzumab's

impact on HER2 homodimerization and heterodimerization. Studies have shown that trastuzumab does not significantly affect HER2 heterodimerization when specific ligands are present. However, its influence on ligand-independent pairing remains contentious [28]. For example, while some studies have shown that trastuzumab effectively blocks certain dimerization interactions, others have found the results inconclusive. Additional research using reversible cross-linkers has revealed that trastuzumab disrupts non-ligand-dependent dimerization while leaving ligand-dependent dimer interactions unaffected. To investigate the impact of trastuzumab on HER2 homodimerization, one study used Chinese hamster ovary cells engineered to overexpress human HER2 [28]. These cells are ideal for studying homodimerization because of their minimal expression of other HER receptors. Surprisingly, the findings indicated that trastuzumab actually promotes HER2 homodimerization in a manner dependent on dose and duration of exposure [28]. These results align with previous observations of increased HER2 homodimerization in HER2+ breast cancer cell lines after trastuzumab treatment. However, the mechanism by which trastuzumab induces both HER2 homodimerization and inhibits non-ligand-dependent heterodimerization remains unclear [28]. Overall, these findings suggest that trastuzumab's effects on HER2 dimerization are somewhat limited and complex. Beyond dimerization, the specific binding of trastuzumab to HER2 contributes to its therapeutic efficacy by affecting downstream signaling pathways and cellular processes [29].

The phosphorylation dynamics induced by trastuzumab add an unexpected layer to its effects. Contrary to expectations, the antibody triggers HER2 phosphorylation which is linked to increased homodimerization and tyrosine kinase activity [28]. Although HER2 phosphorylation is expected to up-regulate ADAM17 and other targets, promoting ligand-dependent heterodimerization of HER2, studies show minimal impact on this process [30]. In particular, increased HER2 phosphorylation on tyrosine 1248 correlates with a positive

prognosis in HER2+ breast cancers under trastuzumab therapy [31]. However, limited research exists on trastuzumab's effect on EGFR and HER4 phosphorylation, underscoring the need for further exploration [31, 32]. The dual nature of trastuzumab in simultaneously inducing HER2 phosphorylation and inhibiting cancerous characteristics in HER2+ breast cancer cells remains a fascinating area for further investigation.

The impact of trastuzumab extends beyond its influence on HER2 dimerization and phosphorylation, with connections to other cellular processes such as the mitogen-activated protein kinases (MAPK) pathway [28, 33]. Deregulation of this pathway has been linked to various cancers and is associated with HER2 overexpression in HER2+ breast cancer [34]. Although trastuzumab is involved in MAPK signaling, several studies have presented conflicting results. Short-term exposure increases Extracellular signal-regulated kinases 1 and 2 phosphorylation, but inhibitory effects vary between experimental conditions and cell lines [35]. Interestingly, even with complete suppression of Extracellular signal-regulated kinases 1 and 2 phosphorylation by Mitogen-activated protein kinase kinase inhibitors, trastuzumab sensitive breast cancer cell lines maintain viability, suggesting a nuanced relationship between trastuzumab and the MAPK pathway [28]. This underscores the need for further investigation under consistent experimental conditions to delineate the precise impact of trastuzumab on MAPK signaling and its potential implications for cell viability [28].

In terms of intracellular signaling, the influence of trastuzumab on the PI3K/AKT pathway, frequently activated in cancer, is of paramount importance [36]. In HER2+ breast cancer, where over-expression of HER2 leads to HER2-HER3 heterodimer formation, and activation of the PI3K/AKT signaling pathway, trastuzumab plays a crucial role [28]. Trastuzumab's impact on HER3 phosphorylation and PI3K activity in trastuzumab sensitive breast cancer cell lines presents conflicting findings despite its established role

in suppressing AKT phosphorylation, a pivotal component within this signaling cascade [28]. Rapid reduction in AKT phosphorylation by trastuzumab appears to involve a pathway independent of PI3K activity and HER2-mediated HER3 phosphorylation, suggesting a complex and multifaceted mechanism of action. Understanding these complex interactions is essential to refine treatment strategies and maximize the therapeutic potential of trastuzumab [28].

Ongoing research endeavors are dedicated to unraveling the complexities surrounding the effects of trastuzumab on HER2 endocytosis and degradation. Studies present conflicting findings, with some indicating HER2 downregulation with trastuzumab, while others report no significant effect [28]. Long-term use generally down-regulates HER2, yet the precise mechanism remains elusive. Investigation of HER2 endocytosis reveals a nuanced picture, one study suggesting no significant effect and another indicating internalization and degradation after prolonged treatment [37, 38]. The potential involvement of Caveolin-1 in stabilizing HER2 at the cell membrane, enhancing trastuzumab-induced cytotoxicity, adds another layer of complexity to the ongoing research efforts [39]. The involvement of truncated HER2 fragments, especially p95HER2, has emerged as an important area of study to understand resistance to trastuzumab [40].

In conclusion, the ongoing research landscape surrounding trastuzumab reflects the dynamic and multifaceted nature of its mechanisms and effects. The continued exploration of these complexes is the key to refine treatment strategies, improve therapeutic outcomes, and address challenges such as resistance to trastuzumab. As we dive deeper, we uncover layers of complexity, paving the way for more targeted and effective interventions in the realm of HER2+ breast cancer.

1.2.3 Trastuzumab and HER2 Isoforms

Despite the widespread clinical use of trastuzumab, as HER2-targeted therapy, patients with HER2-positive breast tumors still experience the lowest disease-specific survival rates, and a considerable proportion do not benefit from trastuzumab treatment [41]. The presence of the HER2 isoform, HER2 Δ 16, is observed in 90% of patients with tumor expression, correlating with metastatic disease. Moreover, HER2 Δ 16 expression in breast tumor cells contributes to trastuzumab resistance [41].

In addition to the isoforms shown to inhibit the effects of Trastuzumab, little is known about the effects Trastuzumab has on AS in HER2+ breast cancer. By performing experiments to identify genes and pathways that undergo AS events due to Trastuzumab treatment, the elusive mechanism of action of Trastuzumab may come closer to view.

1.3 Transcriptomics

Transcriptomics is the study of the complete set (transcriptome) of all mRNA molecules (transcripts) in a cell, tissue, or organism [42]. The main areas of study range from transcription and expression levels to functions, locations, trafficking, degradation of RNA, and AS. Identifying AS events falls within the scope of transcriptomics [42].

1.3.1 Third-Generation Sequencing Methods

Third-generation sequencing or next generation sequencing (NGS) methods represent a significant advancement in DNA sequencing technology and allow for high throughput, in-depth studying of the transcriptome [43]. Both short and long-read sequencing meth-

ods contribute to more comprehensive and accurate genomic and transcriptomic analysis. Short-read technologies, such as Illumina sequencing, using a Sanger sequencing-by-synthesis approach, involve fragmentation of DNA to <300 bases long, adding adapters and performing bridge amplification to create clusters of identical DNA fragments in a flow cell [43]. In the sequencing step, fluorescent-labeled nucleotides are added, and emitted signals are captured by a camera, enabling the determination of DNA sequence [43]. This highly parallel process, capable of sequencing millions of clusters simultaneously, offers high accuracy and cost-effectiveness. Renowned for its efficiency in generating large-scale genomic data, Illumina sequencing has become a fundamental technology in genomics research [43]. However, short-read sequencing (SRS) technology has its drawbacks. Especially in structural variant detection and genome assembly, fragmented reads are too short to detect over 70% of the structural variation of the human genome [44]. Another potential issue with Illumina sequencing is how to read and interpret sequences with high levels of guanine-cytosine (GC) repeating regions [44]. This is due to the shorter reads not spanning the beginning and end of the region which makes it difficult to accurately determine the number of repeats that occur.

In contrast, long-read technologies, such as those offered by Pacific Biosciences and ONT, generate significantly longer sequencing reads [43]. These longer reads can contribute to improved genome assembly, better detection of structural variations, and better understanding of complex genomic regions [44]. Pacific Biosciences uses single-molecule real-time sequencing, while ONT uses nanopore sequencing, allowing direct sequencing of native DNA molecules without the need for amplification [44].

More specifically, ONT sequencing employs nanopores embedded in a synthetic membrane that pass through a single-stranded DNA molecule through the pore [44]. In tandem with the movement of the DNA molecule, there is an ionic current passing through the

pore. As each nucleotide passes through the pore, the current is disrupted. This current disruption is measured in real time and saved in fast5/pod5 files, which can be converted to raw reads through a process called basecalling [45].

1.3.2 Common Practices in NGS Data Processing

The median error rates of ONT sequencing are relatively high at 6% [46]. However, ONT's new Q20+ chemistry boasts an accuracy of $\geq 99\%$. Error rates and other factors such as reads mapping to multiple genes and gene mutations make proper processing of the raw sequenced data critical for accurate assembly and analysis. There are several tools in the bioinformatics toolkit to allow for filtering, trimming and alignment and annotation. Porechop is a program developed for ONT raw reads to remove adaptor sequences from reads and split reads that contain an adaptor in the middle of said read [47]. This error occurs when the Minknow software, which reads the current disruptions in real time, cannot distinguish between the end of a read and the beginning of another [47, 48]. The output from the basecalling software is a FASTQ file that contains read sequences and meta-information about each target read. One important metadata that is collected is the Phred Scores for each base in a given read. The Phred score is a value assigned to each base in a read and represents the probability of error for that nucleotide. The equation to calculate the basecalling accuracy is where Q is the Phred score and P is the probability of error:

$$Q = -10 \log_{10} P \text{ or } P = 10^{-Q/10}$$

Common practices are to trim the first 150 bases of any given genomic read and remove reads with an average Phred score of less than 20 which would correspond to a 1% average

error rate for the filtered read. Programs like Chopper are designed not only to filter based on the average Phred score of a given read, but also to trim the ends of the said read based on the input parameters [49].

Once reads are mapped to a reference in a program like minimap2, a long-read tailored aligner, each alignment is given a few metrics related to quality that can be used to further filter data. Minimap2 assigns a MAPQ score ranging from 0 to 60, scaled to be similar to Phred scores, that represents the likelihood that an alignment is inaccurate. A score of 60 indicates high confidence that the read is accurately assigned to a specific part of the reference sequence. In contrast, lower scores indicate that the read shares similarities with multiple regions of the reference sequence, making it difficult to confidently assign it to a single region [50]. Samtools allows for filtering based on MAPQ score and primary/secondary alignments with parameters `-F 260 -q 60` which will only return primary mapped reads with a MAPQ score of 60 [51]. These parameters are very stringent and lower MAPQ scores like 40 could be used for noisy error-prone reads.

Following the processing of reads, the filtered BAM file is ready for annotation. Bedtools is a program with many bioinformatic analysis features. In the context of gene annotation, the aligned reads are converted to a browser extensible data (BED) file. With the *intersect* function and parameters `-f 0.95 -r`, the mapped reads are annotated on a reference BED file with the gene that has reciprocal overlap 95% in terms of the location of the chromosome. This means that both the mapped read and the given gene must contain at least a 95% overlap of each other for the annotation to be in the final output [52].

1.3.3 Gene Expression Analysis

Gene expression analysis involves studying the activity of genes to understand how they function and contribute to cellular and biological processes. It provides insights into which genes are turned on or off in specific conditions, tissues, or diseases. This analysis is crucial for understanding normal biological processes, development, and various diseases, including cancer. The two main methods for gene expression analysis are microarray technology and RNA sequencing.

Microarrays allow simultaneous measurement of expression levels for thousands of genes, while RNA sequencing provides a more detailed and accurate transcriptome analysis by directly sequencing RNA molecules with the sequencing methods discussed above [53]. Gene expression analysis is used in cancer research to identify genes associated with tumor development, progression, and response to treatment [54]. In drug testing, this analysis helps to evaluate the impact of pharmaceuticals on gene expression patterns, helping to identify potential therapeutic targets and understand drug mechanisms [54]. This information is crucial for developing targeted therapies, improving treatment efficacy, and minimizing adverse effects in cancer and other diseases [54].

1.3.4 Differential Alternative Splicing Analysis

Differential alternative splicing (DAS) analysis is used to study splice variations that are used to develop diagnostics, prognosis, and treatment tools with limited negative effects in cancer [6]. It involves detection, statistical comparison, and effect prediction of splicing patterns under different conditions, aiding in the identification of specific splice variants linked to cancer progression or treatment response [55]. Understanding the role of DAS in cancer treatments may have important implications for personalized medicine, reveal-

ing potential therapeutic targets and diagnostic markers [56]. Monitoring AS patterns during treatments can provide information on the efficacy and potential drug resistance mechanisms [56]. High-throughput sequencing enables detailed analyses of AS landscapes, offering intricate details of splicing alterations in cancer cells. Integrating this knowledge into treatment strategies holds promise for developing more precise interventions based on the unique splicing profiles of patients [56]. Accurate isoform characterization allows researchers to uncover isoform-specific expression patterns, identify novel isoforms, and explore their roles in biological processes.

1.3.5 High Throughput Data Analysis

Whole transcriptomic sequencing outputs reach into gigabytes worth of data and it is imperative to have access to analysis pipelines that can handle multiplexing, dynamic computational power allocation and, parallel processing. While some programs offer multi-threaded processing, most lack the flexibility and compatibility to handle multiple sample inputs and seamless interaction with high performance computer clusters. Nextflow is a powerful and scalable workflow management system designed specifically for bioinformatics and scientific data analysis [57]. It simplifies the creation, execution, and maintenance of complex computational pipelines, allowing researchers to seamlessly integrate diverse tools and processes into a cohesive workflow. Nextflow utilizes a domain-specific language that facilitates the definition of workflows in a portable and reproducible manner. Its key strength lies in its ability to manage dependencies, handle parallelization, and efficiently distribute computing tasks across various computing environments, such as local machines, clusters, or cloud platforms [57].

1.4 Conclusion

Investigation of significant AS events in HER2+ breast cancer when treated with the mAb Trastuzumab could uncover novel isoforms, changes in isoform expression, and biological processes and pathways affected by AS events. This deeper insight into Trastuzumab's mechanism of action will allow for more targeted therapeutics and understanding into the complexities of breast cancer's response to mAb's at the transcriptional level. We set out to first optimize the methods used in Direct cDNA sequencing with Oxford Nanopore's long read technology, with a focus on finding efficiencies and testing the effects on sequencing outcome. We then performed a literature review on five AS identification software designed for long-read sequencing. Finally, we developed an automated high-throughput pipeline, integrating DAS software, to investigate the significant AS events in HER2+ breast cancer cells, when treated with Trastuzumab.

Chapter 2

Differential Alternative Splicing Software Review

2.1 Introduction

Isoform identification plays a crucial role in transcriptomic research as it provides a deeper understanding of the complexity and diversity of gene expression. Traditional gene expression analysis focuses on the overall expression levels of genes. However, genes can produce multiple isoforms by AS, leading to significantly different protein structures and functions. More than 95% of human multiexon genes are subject to AS [58]. Researchers can uncover isoform-specific expression patterns by accurately characterizing isoforms, identifying novel isoforms, and exploring their roles in various biological processes.

In the field of disease and medicine, dysregulation of AS has been associated with numerous diseases, including cancer, neurodegenerative disorders, and cardiovascular diseases [59, 60, 61]. Studying isoforms can provide insights into disease-specific isoform profiles that

have the potential to guide the development of targeted therapies and precision medicine approaches [59].

Advances in long-read sequencing technologies, such as Oxford Nanopore platforms, can allow researchers to accurately identify AS events [58, 62]. Full-length transcripts are sequenced and compared with a reference genome using isoform identification software (IIS) tailored for long-read transcripts to identify and annotate AS events. In this chapter, the algorithms used in long read AS identification will be discussed as well as how long read IIS implements these algorithms. The IIS software that will be examined is LIQA, Flames, FLAIR, and Nanosplicer [63, 64, 65, 66].

2.2 Algorithm Overview

The steps that are typically consistently found in the IIS examined is as are mapping reads, identifying splice junctions, clustering splice junctions, and identifying differential splicing events. These are described in Sections 2.2.1 to 2.2.4 below.

2.2.1 Mapping Reads to Reference Genome

Sequencing output (FASTA or FASTQ format) is mapped to a reference genome using alignment/mapping software. For example, minimap2 reads a FASTA/FASTQ file that contains multiple reads and aligns them with a reference genome [50]. The reference file is a (gene transfer format (GTF)) file that contains detailed annotations of genomic features such as genes, exons, and transcripts, specifying their positions and types on a reference genome. Each entry includes information like sequence name, feature type, start and end positions, strand, and additional attributes like gene and transcript IDs. First, an index

of the reference genome is made using a data structure called a minimizer index. Indexing enables fast retrieval of candidate regions that are likely to align with the query reads [50]. Minimap2 follows a seed-chain-align process. A set of short sub-sequences called seeds is extracted from the read. These “seeds” serve as potential starting points for alignments and are selected based on their uniqueness and abundance in both the read and the reference genomes. Generated seeds are then used to quickly identify candidate regions in the reference genome that potentially match the read. The seeds are then aligned to the candidate regions of the reference genome using an efficient alignment algorithm. Minimap2 then employs a chaining algorithm to link the aligned seed regions, forming a chain of potential alignment regions [50]. This helps to identify the most likely alignment path and reduce the number of false positives. A dynamic programming-based algorithm is then used to score and more precisely align the reads within each chain region, considering alignment parameters such as gaps and mismatches. Finally secondary alignments can be removed, and low quality or redundant alignments filtered out based on alignment scores or alignment overlap with a tool samtools [50, 51].

The following parameters are recommended for long-read alignment to a reference genome for AS detection with minimap2 [50]:

```
minimap2 -ax splice --junc-bed <anno.bed> <ref.fa> <query.fa> > <aln.sam>
```

-ax splice	generate a Compact Idiosyncratic Gapped Alignment Report (CIGAR) string and output alignments in the Sequence Alignment Map (SAM) file format. Also, it enables the splice mode, where long deletions are taken as introns and shown as 'N' in the CIGAR string. A CIGAR is a concise way to show how a sequence aligns to a reference genome. They use symbols like I (for insertions), D (for deletions), and M (for matches) to represent how each base in the query aligns with the reference.
-junc-bed anno.bed	Prefers splicing in annotations, which are found in the BED file that contains genomic coordinate data, such as gene annotations, transcription factor binding sites, chromatin accessibility regions, or variant positions.
ref.fa and query.fa	the reference genome file in FASTA format and the FASTA file that is being aligned to the reference genome
aln.sam	SAM output file contains the aligned reads, CIGAR strings and other information such as mapping score and tags.

2.2.2 Identify Splice Junctions

Splice junctions are regions of the mRNA sequence in which an intron was removed and two exons joined [67]. Reads that span exon-intron boundaries and have proportions of their sequence aligning to different exons in the reference genome are used to identify the most probable splice junction locations [68]. Identified splice junctions are reported in the alignment output files (SAM or BAM). Once a splice junction is identified, it is classified as canonical (following the GT-AG rule), non-canonical, or AS events (see below), based on characteristics such as junction coverage, alignment quality, and supporting reads. The

identified reads are then filtered and refined to remove junctions with low confidence and annotated by associated genes, transcripts, and protein function [68].

2.2.3 Cluster Splice Junctions into Splice Events

Splice events occur when a group of splice junctions have common acceptor and donor sites but include or exclude different exons. Splice junctions that share genomic coordinates and patterns are grouped together. These clusters are then used to create splice events by separating according to the location of the splice junction in the genome, supporting reads, and statistical measures of similarity. The splice events are then classified into categories such as AS events or constitutive splicing events [63, 64, 65, 66].

2.2.4 Identify differential splicing events between Samples

Finally, splice events are compared between samples to determine AS events that occurred due to the introduced variable. This can be a novel drug or a diseased cell compared to a healthy cell. Depending on the program used, a statistical model, such as the Gaussian copula regression model or the Dirichlet multinomial model, will be used to determine which splicing events differ significantly between the samples in question [63, 64, 65, 66].

2.3 Examples

The following is a detailed overview of LIQA, Flame, FLAIR, and Nanosplicer [63, 65]. LIQA and FLAIR are commonly used comprehensive DAS programs. FLAME is an isoform

identification program that is used in DAS analyses [64]. Nanosplicer is a long read alignment tool that harnesses the electrical current output file (fast5/pod5) from the nanopore to enhance the identification of the splice junction [66].

2.3.1 LIQA

Like most software, the IIS listed above are bound by the computer's memory and storage capacity. However, LIQA is not able to run parallel processing which can affect performance on large datasets [63]. The LIQA workflow is broken down into 3 'tasks.' The first step is preprocessing of the reference file:

```
liqa -task refgene -ref <example.gtf> -format gtf -out <example.refgene>
```

-task refgene	Enables preprocessing of reference file.
-ref example.gtf	The reference file used is a GTF file.
-format gtf	Sets the format of the reference file as GTF.
-out example.refgene	Output file name.

The program takes a reference file in GTF format and creates a matrix file for the next step. The second step is to quantify the expression of the isoform [63]. This step is broken into 2 parts. The first is to filter the mapped reads (from minimap2 in sorted BAM format) on the basis of quality and length. Then, following script can be used to quantify isoform expression [63].

```
samtools view <bam_file> -F 2308 -q 50 -O BAM -o <bam_filtered>
```

```
liqa -task quantify -refgene <example.refgene> -bam <example.bam> -out \
    isoform_expression_estimates -max_distance 20 -f_weight 1
```

view bam_file	Enables user to view the Binary Alignment Map (BAM) file
-F 2308	Sets a filter to exclude reads based on their bitwise FLAG values. In this case supplementary alignments, secondary alignments and unmapped reads are excluded.
-q 50	Sets a filter to exclude reads that have a q score less than the input value. In this case, a MAPQ score of 50 is the threshold.
-O BAM	Sets the output format to BAM.
-o bam_filtered	sets the output file name to bam_filtered.
-task quantify	
-max_distance 20	Sets the maximum length of an alignment error at the exon boundary and is recommended to remain at 20
-f_weight 1	Sets the weight for bias correction in isoform usage and is recommended that t remain at 1

The previous 2 steps would be conducted on 2 datasets, collected from different conditions, such as a treatment and control group of cells. (3) during the last 'task', LIQA performs differential splicing detection. Below is an example command line input [63].

```
liqa -task diff \
-condition_1 <list_of_isoform_expression_estimation_file_for_condition1> \
-condition_2 <list_of_isoform_expression_estimation_file_for_condition2> \
-out <test_results_file>
```

-task diff	Enables the Differential analysis program.
-condition_*	Sets the 2 different lists of condition files that will be compared.

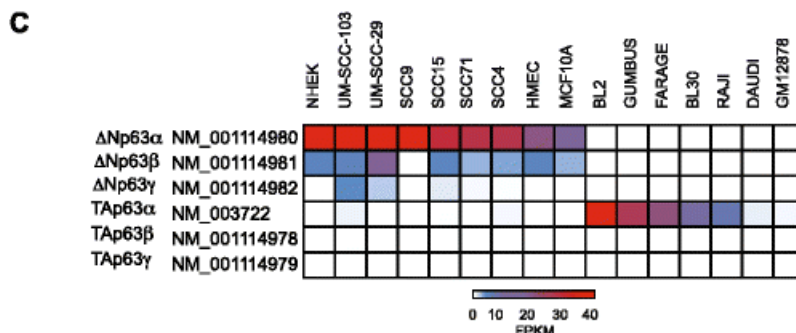


Figure 2.1: **Example of a heatmap.** Expression levels of p63 isoforms, averaged across replicates, in FPKM (fragments per kilobase of transcript per million) across representative cell lines that express p63. *Source:* [72]. Licensed under CC BY 4.0.

The output file is a Gene-based results file (.csv) that contains the gene name, test P value, and Hellinger distance [63]. This measures the splicing difference between two conditions/inputs for a gene in terms of isoform relative abundance. The resulting data can be visualized in many ways to illustrate the observations made, for example, a heatmap can show differential AS events across different groups or conditions. The columns are unique conditions, and the rows are differential AS events. The color and intensity of each cell on the heat map would represent the level of expression of that AS event. Furthermore, a GO/KEGG term enrichment analysis can be performed to determine GO/KEGG terms that are overrepresented in the list of genes that express significant AS. This would allow researchers to determine gene pathways that are significantly affected by AS events [69, 70, 71]

2.3.2 FLAIR

FLAIR is another comprehensive AS analysis tool for long read data. It differs from the other programs by only considering splice sites that are identified from a curated set of

splice junctions and present in an input annotation file [65]. However, if a splice site is within a window size of 10bp, it is replaced with a correct splice site. FLAIR is divided into 3 main processing steps (align, correct, and collapse), followed by various analysis pathways, including differential splicing analysis (FLAIR diffSplice) and Fisher’s exact test (diff_iso_usage) to determine significant alternative splicing [65]. FLAIR align uses minimap2 as the alignment method. Below is the FLAIR alignment command:

```
flair align -g ref.fa -r query.fa [OPTIONS]
```

FLAIR’s GitHub page described option arguments including multi threads, quality thresholds, and additional annotation [65]. The output files consist of the alignment file (BAM and index) and the annotations file (BED) are used in the correction step, along with a GTF reference file. As mentioned above, the correction step replaces the incorrect splice sites within 10bp of the correct splice sites with the corrected site and may be called by the following command [65].

```
flair correct -q query.bed12 -g ref.fa [OPTIONS]
```

-q query.bed12	The output from FLAIR align in BED12 format, if reads are already aligned in a BAM file, bam2Bed12 can be used to convert to proper format.
----------------	---

The FLAIR collapse step generates a high-confidence set of expressed isoforms for further analysis [65]. First, FLAIR creates a preliminary nanopore isoform transcriptome from only the reads that have been fully corrected for splice errors, organizing reads into unique groups based on identical splice junction chains [65]. Within each group, FLAIR identifies one or more representative isoforms by determining confident transcription start

and end sites, using the density of read start and end positions for guidance. The process then includes reassigning all reads to a first-pass isoform, even those reads that could not be fully corrected for splice errors but have minimap2-aligned splice sites [65]. This involves realigning reads to the spliced sequences of the first-pass isoform set and assigning each read to the isoform it aligns with best, requiring a minimum MAPQ score of 1 [65]. This step is essential to address misalignment due to splice site errors by restricting reads to align only with supported splice junction chains, which also helps to identify splice site variations more accurately [65]. In the final step, FLAIR eliminates any first-pass isoforms with less than three supporting reads, resulting in a final set of nanopore isoforms that are backed by sufficient coverage and thus considered high-confidence [65].

```
flair collapse -g ref.fa -q <query.bed> -r <quary.fa> [OPTIONS]
```

-q query.bed	Bed file of aligned/corrected reads.
--------------	--------------------------------------

Isoform quantification is performed using FLAIR quantify. The output isoform FASTA file from the “FLAIR collapse” step is read, and only alignments that have quality scores of 1 or higher are considered. To normalize the isoform counts within each sample, each count is divided by the upper quartile (75th percentile) of the read counts from protein-coding genes. For this analysis, only genes identified as protein coding in the GENCODE v24 annotation were considered. The output is a counts file isoform by sample that is used for the FLAIR diffSplice (Differential Splice Analysis) program [65].

```
flair quantify -r reads_manifest.tsv -i isoforms.fa [OPTIONS]
```

-r reads_manifest.tsv	
-i isoforms.fa	Fasta of FLAIR collapsed isoforms.

Custom scripts are available for FLAIR and were written to identify alternative acceptor, alternative donor, cassette exon skipping, and intron retention events (FLAIR diffSplice) [65]. DRIMSeq is used to perform statistical analyzes that consider batch sequencing and RNA batch variations, factoring in the batch numbers associated with each sample [65]. The final product is a text file that contains information such as the gene name, isoform name, p-value, sample isoform count, and alternative sample isoforms for the gene count. This data can be visualized similar to the outputs from LIQA and Flames. In addition, FLAIR provides a program called `plot_isoform_usage` that allows for isoform structure viewing and percentage usage [65].

```
diff_iso_usage counts_matrix colname1 colname2 diff_isos.txt
```

```
plot_isoform_usage <isoforms.psl>|<isoforms.bed> counts_matrix.tsv \
    gene_name
```

2.3.3 Flame

The FLAME software has 2 main functions. The first is gene specific FLAME which performs isoform identification on a query aligned transcriptome for a specific gene of interest [64]. The other function that FLAME can perform is FLAME-GLOW, which is global isoform identification. FLAME is a Python-based program that is executable via the command line, as shown below. Similarly to LIQA, FLAME does not allow multi threaded analysis, which can affect performance on larger datasets [64]. The input file is a BED file that contains genomic coordinate data, such as gene annotations, transcription factor binding sites, chromatin accessibility regions, or variant positions [64] This file is created using an alignment software like minimap2 to align the transcript data to a reference

genome [50]. The resulting Sequence Alignment/Map (SAM) file is converted into a BAM file with a samtools and then a BED file with bedtools [51, 52].

```
samtools view <input.sam> -b BAM -o <output.bam>
```

```
bedtools bamtobed [OPTIONS] -i <output.bam> > <anno.bed>
```

```
./FLAME.py -I <INPUT.bed> -GTF <anno.gtf> -G <Gene>
```

bamtobed	converts a Binary Alignment/Map (BAM) file to a BED file
-G Gene	Enables user to set a target gene. If whole transcriptome used, omit this parameter.

The “Annotation.gtf” file contains information about gene structures, including exons, introns, coding sequences, and untranslated regions, for the target genome. The file can be downloaded from the Ensembl website [73, 74]. Finally, the desired gene can be listed under the annotation option -G [64]. FLAME-GLOW is called by omitting the -G option in the command line above. FLAME-GLOW first categorizes long reads according to their alignment with a reference annotation in GTF format [64]. Next, it untangles the exonic regions of each read, identifying conflicting ones that differ in start position, end position, or length from the reference. Identical exon arrangements are collapsed and quantified to provide a quantitative description of splice variation patterns at the single transcript level. It also detects long-range exon coupling resulting from primer artifacts [64]. The exon/intron boundaries are cataloged to detect novel splice sites and exons, and the results are compiled into an output file for manual assessment. The output file provides details such as unique isoform IDs, transcript information that includes genomic coordinates and lengths, exon structures, identified splicing events, and their expression levels [64].

2.3.4 Nanosplicer

Nanosplicer is an alignment tool designed for enhancing long read splice results. Nanosplicer has the option of increasing the number of threads allocated to the task, which can increase speeds and performance [66]. Nanosplicer takes information from a SAM file, a BED file (converted from a GTF file), and the file that contains the raw electrical signal data (FAST5/POD5 file) known as squiggles [66]. First, the junctions are identified within the reads. Sequences with a pair of 5' and 3' splice sites, which are the boundaries between introns and exons, are identified by mapping the alignment reads to the reference genome, similar to how LIQA performs [66]. Reads containing junctions are aligned with their corresponding squiggle (from the FAST5 files) using the “resquiggle” tool in Tombo [75]. The section of the squiggle aligned with the junction is extracted, and each current measurement is assigned to each base of the read. Data are normalized to remove systematic differences for an accurate comparison with candidate squiggles [66]. The candidate squiggles are obtained by constructing a list of candidate splice junctions. These include the junction supported by the read, nearby canonical splice junctions, annotated splice junctions, user-defined candidate splice junctions, and nearby splice junctions supported by other reads. The candidate junction motifs are assembled by connecting sequences on each side of the candidate splice junction [66]. All candidates must have the same nucleotide sequence and squiggle signal at the beginning and end. A dynamic time distortion method is used to align each junction squiggle with each candidate squiggle [66]. Current measurements are then assigned in the junction squiggle to each mean and standard deviation in the candidate squiggle. This alignment allows for comparison and measurement of similarity between the squiggles. The Nanosplicer model is then used to identify the splice junctions, and the junction squiggles are split into segments and normalized [66]. A mixture model is then built using the candidate squiggles and alignments. Each junc-

tion squiggle is assigned a variable that indicates which candidate the squiggle originated from [66]. The probability of each splice junction is then calculated and used to confirm candidate sites with the support of the junction squiggles. A squiggle information quality step is also implemented to filter out junctions that do not have high-quality alignments to any candidate squiggle. Below is an example of calling the Nanosplicer program from the command line [66].

```
python3 JWR_checker.py [OPTIONS] <BAM file> <output hdf5 file>
```

```
python3 NanoSplicer.py -i <example.bam> -f <fast5s> -r <ref.fa> \  
    <output hdf5 file>
```

The output generated by Nanosplicer includes a probability table and a BED file. The probability table provides information about each splice junction, such as: read id, reference name, initial junction, junction alignment quality, candidates (all candidate splice junctions), candidate sequence motif preference (this shows preferential use of specific intron sequence patterns near splice junctions), squiggle information quality and best_prob (the value of the best probability among all candidates). The BED file contains corrected splice junctions which can be visualized in a genome browser like IGV (Integrative Genomics Viewer) [66]. The output can be used to enhance splice identification for DAS programs like FLAIR and LIQA.

2.4 Conclusion

While there are many other long read isoform identification software that could be reviewed, we focused on the 4 discussed above [76, 77]. All four programs have been used in peer-

reviewed literature and have unique capabilities while maintaining the overall method of isoform identification and analysis. In chapter 4, LIQA is selected to identify significant AS events in HER2+ breast cancer cells due to its simple, yet effective design and the ability to be implemented into a high-throughput Nextflow script.

Chapter 3

Optimization of Sequencing Methods

3.1 Introduction

Accurate sequencing and analyzing the transcriptome of an organism can be used in many research areas, such as novel drug testing and diagnostics. NGS marks a significant advancement in DNA sequencing technology that enables thorough examination of the transcriptome. Short- and long-read sequencing methods contribute to more comprehensive and accurate genomic analysis [43]. Short-read technologies such as Illumina sequencing use a Sanger sequencing-by-synthesis approach, which involves DNA fragmentation, adapter addition, and bridge amplification to create clusters of identical DNA fragments in a flow cell. During sequencing, fluorescent-labeled nucleotides are incorporated and emitted signals are captured by a camera to determine the DNA sequence [43]. This highly parallel process offers high accuracy, making Illumina sequencing a dominant technology in genomics research. However, SRS has limitations, particularly in detecting structural variants and assembling genomes due to the short length of fragmented reads [44]. Additionally, in-

interpreting sequences with high levels of GC repeating regions poses challenges. Long-read technologies, such as ONT, generate substantially longer sequencing reads. These extended reads offer enhanced capabilities in genome assembly, structural variation detection, and comprehension of intricate genomic regions [44]. In ONT sequencing, nanopores within a synthetic membrane facilitate the passage of single-stranded DNA molecules through the pore. Concurrently, an ionic current flows through the pore, which is disrupted as each nucleotide passes through. This disruption is measured in real time and recorded in fast5/pod5 files, which can be converted into raw reads through a process known as base-calling [44]. Long-read ONT sequencing methods can effectively analyze the whole human transcriptome by sequencing full-length reads [78]. ONT transcriptomic sequencing workflows often consist of sequencing an mRNA molecule using a single cDNA strand or double stranded deoxyribonucleic acid (dsDNA) [79].

In order to create the cDNA, many ONT standard procedures call for isolating mRNA through poly(A) enrichment, shown in Figure 3.1. While this protocol is effective in isolating mRNA, there are some limitations. A recent study found that poly(A) selection caused a poly(A) tail length bias and an increased starting material requirement for the ONT direct RNA sequencing kit (SQK-RNA002, ONT; version of the protocol as released: September 13th, 2021) [80].

Included in the library preparation for ONT sequencing, double stranded cDNA is synthesized from mRNA by reverse transcription, shown in Figure 3.2. In the case of ONT protocols, the cDNA undergoes end-prep, adaptor ligation, and then sequencing [82].

ONT Direct cDNA sequencing kit (SQK-DCS-109) only instructs for mRNA as the starting material. Although this kit was recently placed on legacy, the recently released Direct cDNA Sequencing Kit (SQK-LSK114) lists total RNA as the starting material. One objective of this experiment was to determine if using total RNA as the starting material

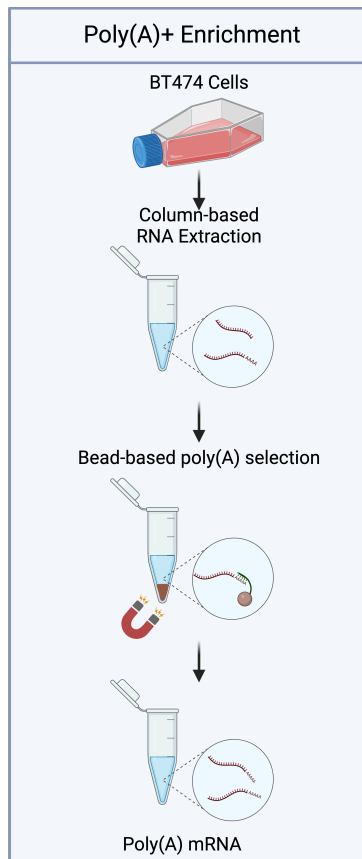


Figure 3.1: **Protocol for poly(A) RNA enrichment.** First, total RNA is collected and poly(A) selection is performed to enrich the sample for mRNA. The poly(A) tail of the mRNA molecule binds to Oligo (dT) metal beads and the supernatant is removed [81]. After a wash with 70% ethanol, the bound mRNA is eluted into a final solution [81]. Created with BioRender.com.

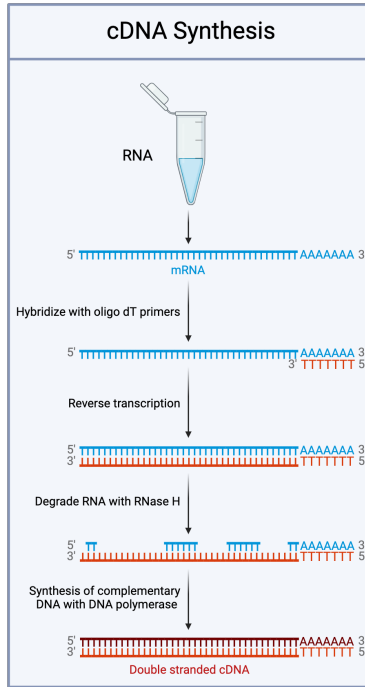


Figure 3.2: **Protocol for cDNA synthesis.** The poly(A) primer anneals to the 3' poly(A) tail of the mRNA [82]. The reverse transcriptase enzyme then synthesizes the cDNA strand using the mRNA as a template and the poly(A) primer as a starting point [82]. The RNA strand is degraded using RNase H or other RNA-specific enzymes. DNA polymerase synthesizes a complementary strand to the cDNA template. The resulting double-stranded cDNA is then purified using magnetic beads, which selectively bind the cDNA fragments based on their size, removing unwanted impurities such as primer dimers, unincorporated dNTPs, and other contaminants [82]. The purified cDNA can then be used for various downstream applications such as polymerase chain reaction (PCR), sequencing, or cloning. Created with BioRender.com.

for the direct cDNA sequencing protocol would yield similar results to those found with the SQK-RNA002 when using total RNA as the starting material. It was hypothesized that omitting the poly(A) selection step from the workflow would remove poly(A) tail selection bias and would not have a significant negative impact on sequencing yield, read length and coverage, gene make-up and unique gene capture. It was also hypothesized that poly(A) selection would induce a selection bias for longer poly(A) tailed mRNA as seen in other studies [80]. This hypothesis is based on the understanding that the use of poly(A) primers and DNA binding beads act as an enrichment step.

3.2 Methods

Cell Culturing and RNA collection

Figure 3.3 depicts the experimental workflow for this experiment. RNA used for all the prepped libraries was extracted from the BT474 cell line. Cells were grown in Dulbecco's Modified Eagle Medium with 10% Fetal Bovine Serum and 1% PenStrep at 37°C with 5% CO₂. Cells were subcultured once 80-90% confluency was reached (every 3 to 4 days) and the RNA was collected with the Quick-RNA™ Miniprep Kit from Zymo Research, 3 days after subculture. The quality of total RNA was analyzed with the Qubit 4 Fluorometer RNA broad range kit. The quality of the total RNA was confirmed to verify that there was minimal contamination with a 1% agarose gel.

Poly(A) selection

The poly(A) mRNA Magnetic Isolation Module from NEBNext® was used to select poly(A) mRNA. The manufacturer’s protocol was doubled to obtain the required amount of mRNA needed for cDNA synthesis and subsequent sequencing and 10 µL nuclease-free water was used for the final solution as a substitute for 17 µL Tris buffer. The concentration of the selected sample was determined with the Qubit 4 Fluorometer RNA broad range kit.

Library Preparation

Direct cDNA sequencing libraries were prepared with the ONT (SQK-DCS109, ONT; version of the protocol as released: Aug 2019). The amount of AMPure XP beads was increased to the recommended AMPure XP ratio of 1.8 µL of beads to 1 µL of the sample [83]. No other modifications were made to the direct cDNA sequencing protocol.

Nanopore Sequencing and base calling

Raw voltage traces were collected as Fast5 files using the MinKNOW software of ONT (version 4.1.22). Basecalling was conducted with Guppy and subsequent analysis was completed on the high performance computer Narval from the Digital Research Alliance of Canada [45].

Post-processing Analysis

A Nextflow workflow script was made to eliminate variability in the analysis. The scripts are available here:

https://github.com/benpitaru/Nextflow-read_analysis

Basecalled reads were aligned to the Human transcriptome (GRCh38) using MiniMap2 (version 2.24-r1122) with parameters *-ax splice* which enables the splice alignment mode and interprets long deletions as introns [84, 50]. Using Bedtools, the output BAM file was converted to a BED file and intersected with the Human Genome (GRCh38) BED file with the following parameters *-f 0.90 -r* to identify mapped reads that had a 90% overlap with a known annotated gene [52]. Primary mapped reads and bases were compiled with Samtools (version 1.17) [51]. Individual reads were labeled with a gene type, according to the corresponding gene name found in the GRCh38 BED reference [84]. Nanopolish (version 0.14.0) was used with the default parameters to identify the poly(A) tail lengths in all sequenced libraries.

Plotting and Visualizations

All statistical tests and visualizations were conducted with custom Python scripts.

3.3 Results and Discussion

The objective of this experiment was to determine if there was a significant negative impact on sequencing yield, read length and coverage, gene make-up, and unique gene capture when poly(A) enrichment was omitted from the library prep protocol. Eleven replicates of the experiment were processed; however, only replicates 6 and 11 had sufficient amounts of mRNA after poly(A) enrichment, so only samples 6 and 11 were sequenced. This speaks to a significant potential drawback of conducting poly(A) enrichment, as the extra step

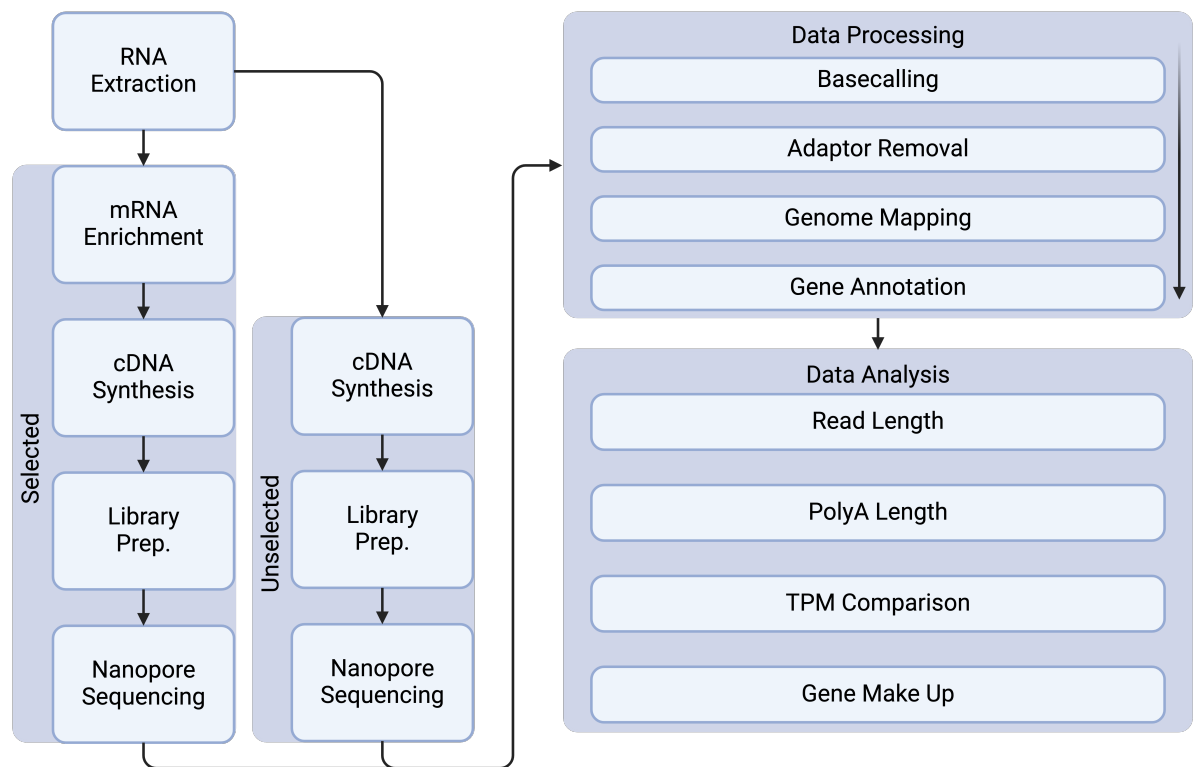


Figure 3.3: Workflow overview for poly(A) selection experiment.

in the protocol further decreases the material and yield variability. Table 3.1 shows input RNA levels, read counts, primary aligned reads, reads aligned to protein-coding sequences (CDS), and bases mapped to CDS. When comparing the average bases mapped to CDS per ng of input RNA, A higher degree of variability was found between the selected samples, compared to the unselected samples. Although not have enough samples were used to provide statistical significance, the dramatic increase in variability shows that omitting poly(A) enrichment may allow for more consistent results.

Table 3.1: Experiment RNA and read amounts overview.

Replicate	Condition	Total RNA (ng)	Library Prep RNA (ng)	Sequencing Time	Read counts	Primary Alignments	CDS Alignments	Bases Mapped to CDS	Bases mapped per ng
11	Selected	75000	733	72h	5.66M	1038.4K	1012.1K	1379.2M	18389.9
6	Selected	83200	150	72h	2.37M	296.1K	287.9K	335.5M	4032.4
11	Unselected	-	5625	72h	3.43M	88.5K	73.9K	110.8M	19696.8
6	Unselected	-	6240	72h	4.97M	139.7K	119.2K	162.6M	26057.7

Looking between the selected and unselected conditions, a small but significant increase was found in the median read length from 714 bases in the selected group to 805 bases in the unselected group ($s = 3.60e^{13}$, $p = 0.0$), depicted in figure 3.4.

When comparing the average read length for each gene identified under both conditions, there was a slight increase in average read length in the unselected group compared to the selected (Regression Line: $y = 0.83x + 1.31$; Spearman R: 0.87), as seen in figure 3.5. The slope coefficient of 0.83 demonstrates a positive correlation in read length between selected and unselected groups. However, unselected average read lengths are being sequenced at slightly longer lengths. Further, the Spearman R correlation of 0.87 shows that there is little change in average read length for a given gene between the unselected and selected conditions.

Therefore, omitting the poly(A) enrichment step in the library prep for direct cDNA sequencing with nanopore has been shown to not negatively affect read lengths, but rather possibly increase read lengths.

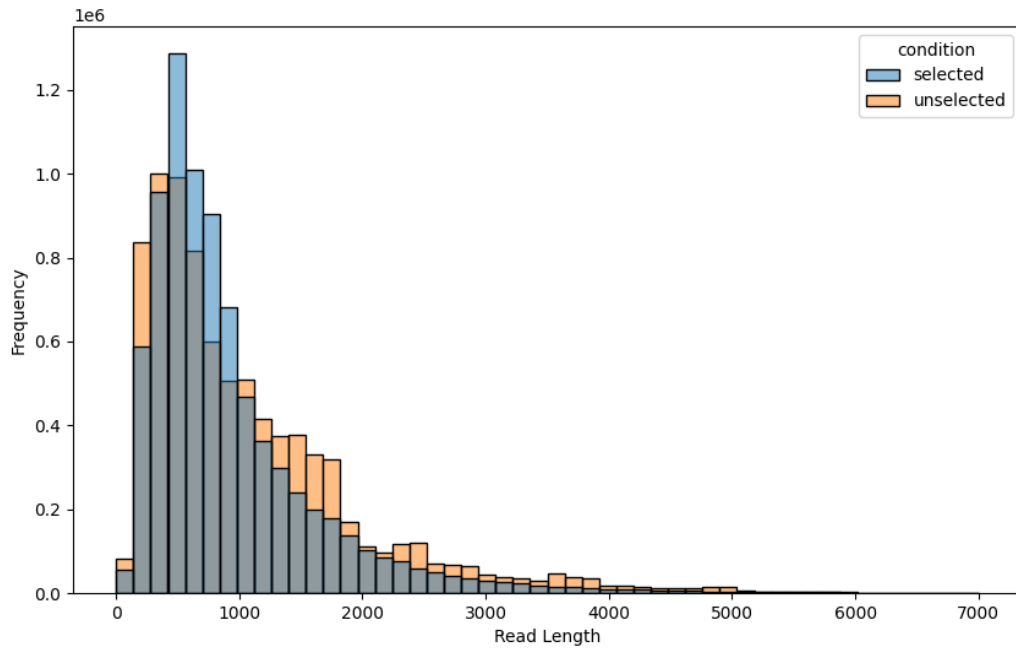


Figure 3.4: **Overall raw read length histogram for each condition.** Statistical analysis used Mann-Whitney U Test ($s = 3.60e^{13}$, $p = 0.0$) $\mu_{selected} = 714.0$, $\mu_{unselected} = 805.0$.

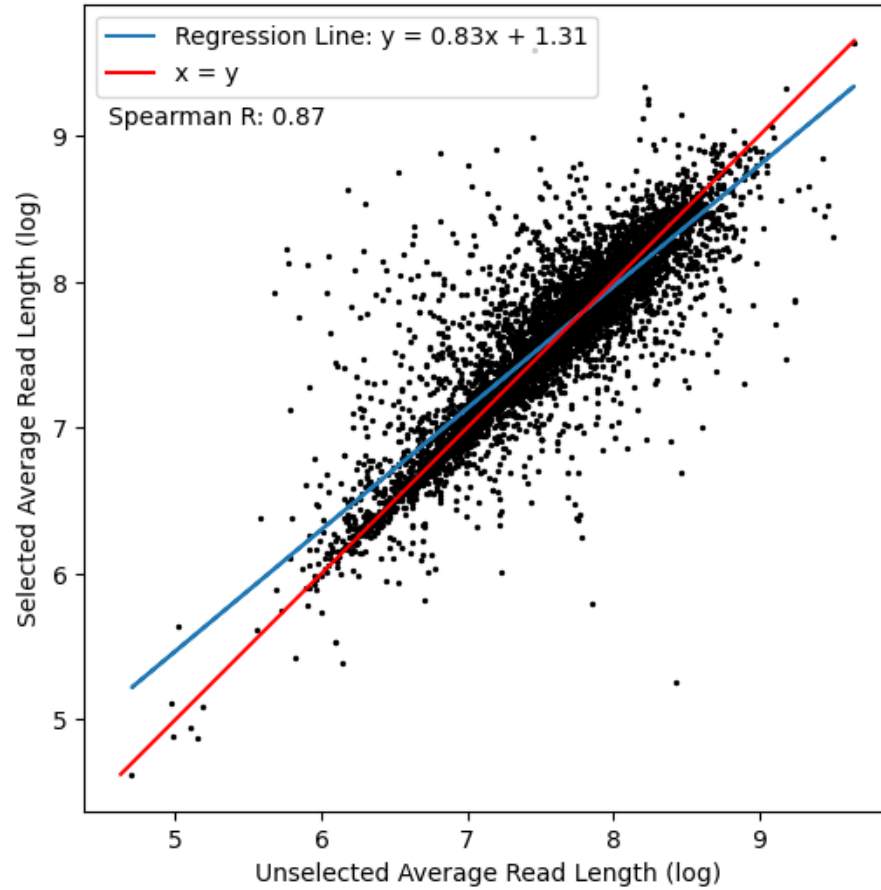


Figure 3.5: **Average read length scatterplot.** Each point represents the average length of reads that mapped to specific gene. Only genes that were found in both the selected and unselected groups are represented in this scatterplot.

Figure 3.6 shows the TPM corrected read counts for each gene found in both selected and unselected libraries. A minimum of five reads was used as a threshold for in each condition. Omitting the poly(A) enrichment step did not significantly affect gene coverage, as the TPM counts spearman R correlation was a notable 0.97. Several genes were found in more abundance in the unselected group. Upon further investigation, of the 10 genes with the largest log difference where the unselected group was larger than the selected group (Table 3.2), 5 of them were clustered histones. A possible explanation for this is that while most mRNA molecules in eukaryotes have a poly(A) tail at the 3' end, the mRNA of histones has a 3' stem loop [85]. The lack of the poly(A) tail causes the histone mRNA to be removed from the selected sample as a result of lack of binding to the oligo-dT beads. However, the reverse transcriptase can use the stem loop as a primer to perform reverse transcription, converting the mRNA to cDNA, allowing it to be sequenced. The top 20 genes where the log difference between the selected to unselected groups are listed in Table 3.3.

Table 3.2: Genes in unselected group with the largest log difference in TPM, compared to the selected group.

Gene ID	Gene Biotype	Selected	Unselected	Log Difference	Gene Name
ENSG00000197061	protein coding	4.15	9.44	5.30	H4 clustered histone 11
ENSG00000197409	protein coding	3.53	8.03	4.51	H3 clustered histone 12
ENSG00000185130	protein coding	3.45	7.79	4.34	H2B clustered histone 13
ENSG00000211459	Mt_rRNA	7.55	11.10	3.55	mitochondrially encoded 12S rRNA
ENSG00000277209	ribozyme	3.76	6.74	2.98	ribonuclease P RNA component H1
ENSG00000277075	protein coding	4.60	7.07	2.47	H2A clustered histone 8
ENSG00000210082	Mt_rRNA	9.61	12.04	2.43	mitochondrially encoded 16S rRNA
ENSG00000275713	protein coding	4.38	6.72	2.33	H2B clustered histone 9
ENSG00000169710	protein coding	1.38	3.27	1.89	fatty acid synthase
ENSG00000113300	protein coding	-1.59	0.29	1.88	CCR4-NOT transcription complex subunit 6
ENSG00000156052	protein coding	-3.10	-1.26	1.84	G protein subunit alpha q
ENSG00000139324	protein coding	-1.40	0.44	1.84	transmembrane O-mannosyltransferase targeting cadherins 3
ENSG00000082175	protein coding	-1.26	0.58	1.84	progesterone receptor
ENSG00000126821	protein coding	-0.79	1.03	1.82	sphingosine-1-phosphate phosphatase 1
ENSG00000182827	protein coding	-0.50	1.32	1.81	acyl-CoA binding domain containing 3
ENSG00000134086	protein coding	0.36	2.14	1.78	von Hippel-Lindau tumor suppressor
ENSG00000163125	protein coding	-1.49	0.25	1.74	regulation of nuclear pre-mRNA domain containing 2
ENSG00000238227	protein coding	1.99	3.67	1.68	transmembrane protein 250
ENSG00000148943	protein coding	0.32	1.98	1.66	lin-7 homolog C, crumbs cell polarity complex component
ENSG00000183943	protein coding	-1.86	-0.20	1.66	protein kinase cAMP-dependent X-linked catalytic subunit

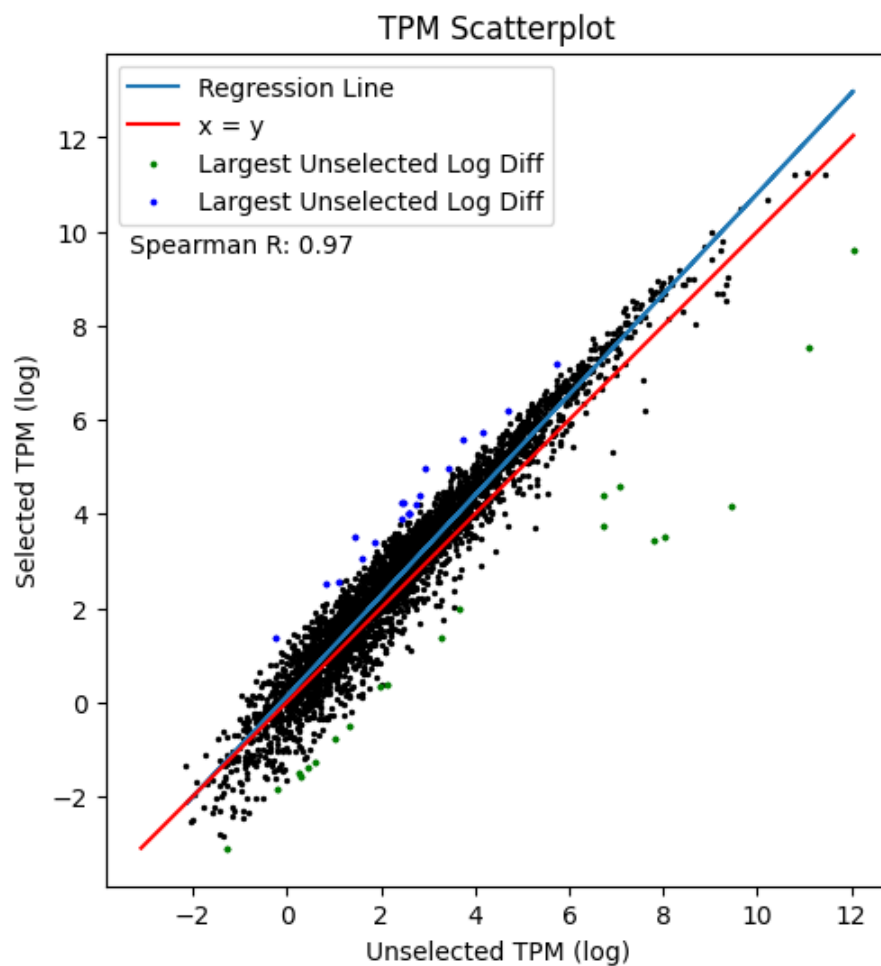


Figure 3.6: **Transcripts per million scatterplot.** The points highlighted in green are the genes where the unselected group has the greatest log difference. The points highlighted in blue are the genes where the selected group has the greatest log difference.

Table 3.3: Genes in unselected group with the largest log difference in TPM, compared to the unselected group.

Gene ID	Gene Biotype	Selected	Unselected	Log Difference	Gene Name
ENSG00000119335	protein coding	3.53	1.46	2.07	SET nuclear proto-oncogene
ENSG00000258315	protein coding	4.97	2.92	2.05	chromosome 17 open reading frame 49
ENSG00000283041	processed pseudogene	5.58	3.75	1.82	eukaryotic translation elongation factor 1 gamma
ENSG00000156990	protein coding	4.24	2.45	1.79	RNA pseudouridine synthase D3
ENSG00000100836	protein coding	4.26	2.48	1.78	poly(A) binding protein nuclear 1
ENSG00000148399	protein coding	2.51	0.83	1.68	diphthamide biosynthesis 7
ENSG00000088930	protein coding	1.36	-0.25	1.61	5'-3' exoribonuclease 2
ENSG00000258232	lncRNA	5.75	4.15	1.60	novel transcript, antisense to TUBA1C
ENSG00000171224	protein coding	4.41	2.82	1.59	family with sequence similarity 241 member B
ENSG00000189060	protein coding	4.97	3.42	1.56	H1.0 linker histone
ENSG00000182952	protein coding	3.40	1.88	1.52	high mobility group nucleosomal binding domain 4
ENSG00000243364	protein coding	4.21	2.74	1.48	ephrin A4
ENSG00000260260	lncRNA	6.18	4.70	1.48	small nucleolar RNA host gene 19
ENSG00000198901	protein coding	2.57	1.11	1.46	protein regulator of cytokinesis 1
ENSG00000159231	protein coding	3.05	1.59	1.46	carbonyl reductase 3
ENSG00000267321	lncRNA	3.90	2.44	1.46	small nucleolar RNA host gene 30
ENSG00000124614	protein coding	7.19	5.74	1.45	ribosomal protein S10
ENSG00000213983	protein coding	4.03	2.58	1.45	adaptor related protein complex 1 subunit gamma 2
ENSG00000100814	protein coding	4.03	2.59	1.44	cyclin B1 interacting protein 1
ENSG00000239382	protein coding	4.01	2.58	1.43	alkB homolog 6

Other studies showed that the poly(A) enrichment step induced a bias against mRNA reads with shorter poly(A) tails. However, this was yet to be tested for the direct cDNA sequencing protocol, as previous studies used direct RNA protocols [80]. Figure 3.7 shows the spread of poly(A) tail lengths for each sample. An analysis of variance (ANOVA) test revealed a highly significant difference between the groups ($p < 0.05$), with a p-value of $2.0456e-17$. This indicates strong evidence to reject the null hypothesis of equal poly(A) tail-length means between the sequenced libraries. A Post-hoc comparisons test using the Tukey Honestly Significant Difference (HSD) test results are summarized in Table 3.4 and further reinforce the claim that a slight bias is introduced against shorter poly(A)-tailed reads. When comparing the mean poly(A) tail lengths, all of the unselected libraries had significantly smaller mean tail lengths than the selected ones. The only comparison that did not have a significant difference in tail length mean was between the unselected groups.

Due to the mechanism of PolyA enrichment, one concern was that by omitting this

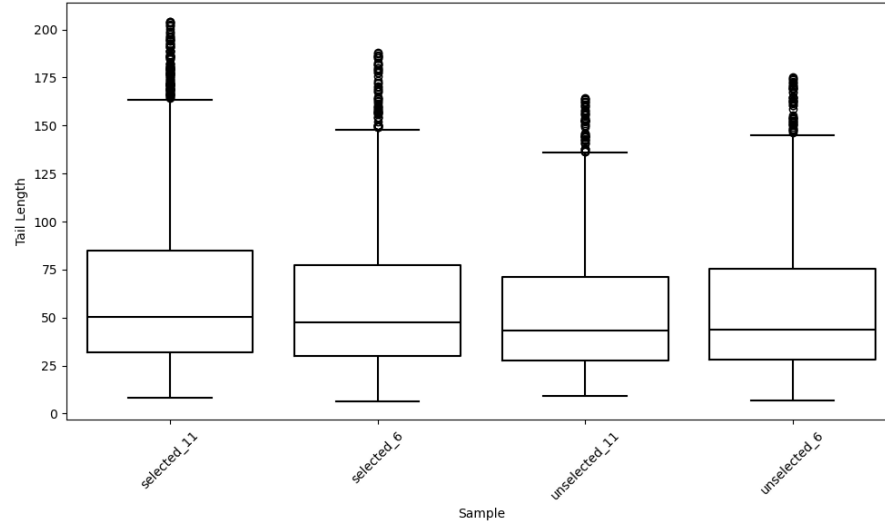


Figure 3.7: **PolyA tail length boxplot for each sequenced library.** The median values for each group are as follows: selected_11 (50.63), selected_6 (47.71), unselected_11 (43.625), and unselected_6 (43.87). The interquartile ranges, represented by the boxes, span from the 25th percentile (Q1) to the 75th percentile (Q3) with the following values: selected_11 (Q1: 32, Q3: 85.47), selected_6 (Q1: 30.17, Q3: 78.4), unselected_11 (Q1: 28, Q3: 71.75), and unselected_6 (Q1: 28.29, Q3: 75.43). The whiskers extend to the range of the data, and each group's sample size is as follows: selected_11 (n=3566), selected_6 (n=1307), unselected_11 (n=1346), and unselected_6 (n=1701).

Table 3.4: **Multiple comparison of means - Tukey HSD.**

Group 1	Group 2	Mean Difference	p-adj	Lower Bound	Upper Bound	Reject Null
selected_11	selected_6	-3.546	0.027	-6.805	-0.287	True
selected_11	unselected_11	-9.806	0.000	-13.041	-6.571	True
selected_11	unselected_6	-7.673	0.000	-10.653	-4.694	True
selected_6	unselected_11	-6.260	0.000	-10.185	-2.336	True
selected_6	unselected_6	-4.127	0.023	-7.844	-0.410	True
unselected_11	unselected_6	2.133	0.448	-1.562	5.828	False

selective step, a large portion of the sequenced library would be lost due to reads that were not of interest. mRNA only constitutes approximately 5% of total RNA and although cDNA synthesis in theory selects mRNA using oligo-dT primers, the other RNA molecules are still present in the library and could find their way to the sequencing step. To determine the effects that omitting the poly(A) enrichment step had on gene makeup and capture, The proportion of gene types in each sequenced library is shown in figure 3.8. Reads mapping to protein coding genes made up on average 97.36% of the selected libraries and 84.46% of the unselected libraries. Of the remaining gene types, the notable difference between the unselected and selected libraries was found to be mitochondrial ribosomal ribonucleic acid (rRNA), which comprised an average of 12.9% of the unselected libraries. There is some evidence that rRNA can have poly(A) tails, however, this does not explain the disparity in gene type composition between unselected and selected groups [86]. While these gene type compositions between conditions are large, the 12.9% loss in protein coding reads is mitigated by the apparent decrease variability in average bases mapped to CDS per ng of RNA when poly(A) enrichment is omitted and lower starting material requirement.

The number of unique genes found in each library and between libraries was obtained and expressed in Figure 3.9. If there was no affect on gene composition, the majority of uniquely identified genes would be found in all datasets. Figure 3.9 shows that 4215 genes

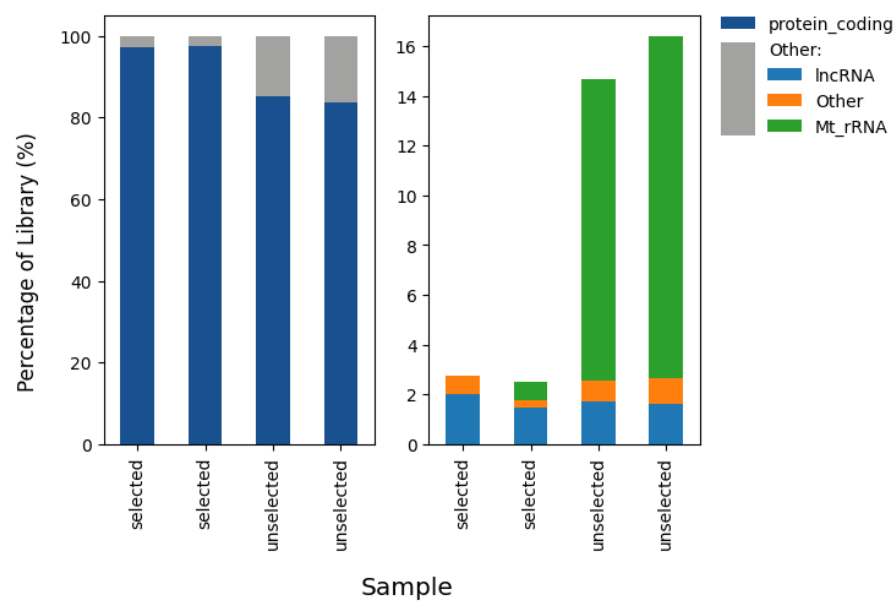


Figure 3.8: **Gene make-up plot for selected and unselected samples.** The plot on the right is a magnified bar graph of the gray “Other” category in the left plot.

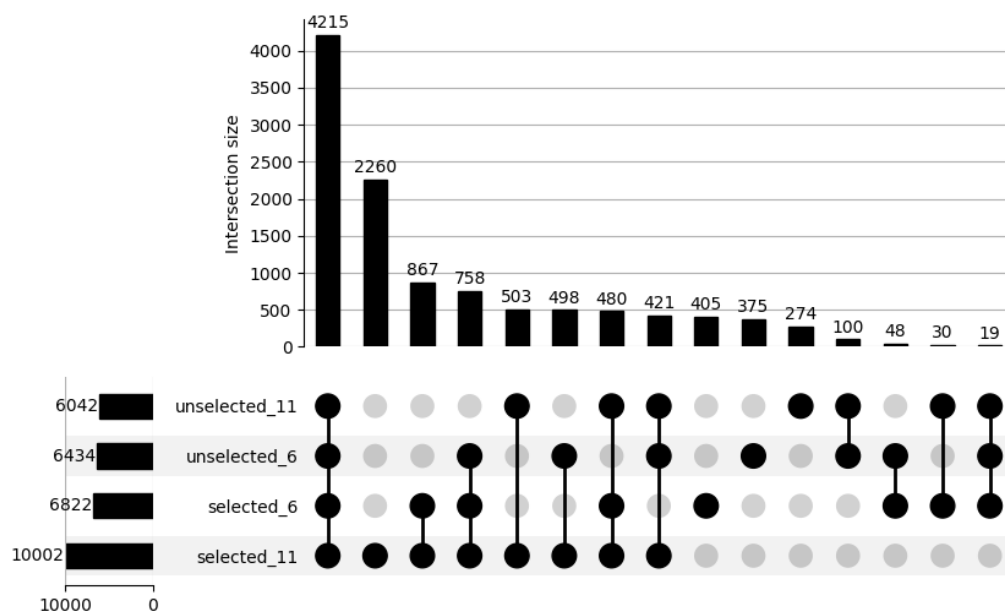


Figure 3.9: **Upset plot for all samples.** The bar graph on the top depicts the number of unique genes that were found between different combinations of samples. For example, 867 genes were found to be in only the selected samples. The bar graph on bottom left shows the total number of genes identified in each sample.

were found in all datasets with the second highest number of unique genes being replicate 11, selected. The interpretation of these results is that the high levels of unique genes found in the selected 11 dataset was due to the high level of sequenced data (between 3.26x and 0.85x more primary aligned reads, and 1.06x to 1.47x more unique identified genes, compared to all other sequenced libraries). Other notable observations in figure 3.9 is that the number of unique genes found in either condition is not drastically different, when considering the affects that the large selected 11 library has on the number of unique genes found in each condition. While more data is required to solidify the findings, the current data shows some support that omitting the poly(A) enrichment step does not negatively change gene make up and capture.

3.4 Conclusions

The purpose of this experiment was to determine whether omitting the poly(A) enrichment step prior to cDNA synthesis for direct cDNA sequencing with ONT long-read sequencing would have negative effects on the data collected. Sequencing yields, read lengths and coverage, poly(A) tail lengths, gene make-up, and capture were tested. In all cases, omitting the poly(A) selection step did not negatively impact the outcomes of the experiment and in the cases of sequencing yields, read lengths and coverage and poly(A) tail lengths, there was a potential positive impact on experiment outcomes. Further, omitting the poly(A) enrichment step allowed for a greater sequencing depth of histone mRNA molecules.

Cell lines tend to be more repeatable than population samples. You could say that the experiments were conducted with data from a single cell line and additional confirmation with other samples, cell types, and organisms such as yeast would provide additional confirmation.

Furthermore, the experiments were conducted with data from a single cell line and additional confirmation with other samples, cell types, and organisms such as yeast would provide additional confirmation. Finally, the results can be confirmed with more recently developed kits. For example, the Ligation Sequencing kit (SQK-LSK114) from ONT, used with the Direct cDNA sequencing protocol or the Native Barcoding kit (SQK-NBD114.24) which includes barcoding, used with a double stranded cDNA synthesis protocol. These kits are currently in use and therefore, the results would be more relevant when considering library preparation methods.

Chapter 4

Testing Affects of Trastuzumab

4.1 Introduction

AS is a biological mechanism in eukaryotic cells in which many different functional and structural variants can be produced from a single gene by removing specific introns and joining specific exons [87, 58, 88]. AS is a fundamental mechanism and may have important implications to fully understand when dealing with clinical drugs, as AS events occur in over 90% of human genes with multiple exons [87, 5]. Identifying AS events in breast cancer has many clinical implications, such as breast cancer patient survival time [89]. One study found that AS events occurred in groups of genes linked to oxidative stress, induced cell death, and DNA damage response, affecting cancer development [89]. In approximately 20-25% of breast cancer diagnoses, the (HER2) protein is found in high concentrations [90]. HER2 is a surface membrane protein involved in signal transduction, cell proliferation, survival, differentiation, angiogenesis, and metastasis [90]. Tumors overexpressing the HER2 gene have higher proliferation rates, a greater frequency of aneuploidy, and

poorer patient prognosis [91]. However, advances in targeted therapies, such as Herceptin (Trastuzumab), have significantly improved the prognosis for individuals with HER2+ tumors, making early detection and personalized treatment crucial in managing this sub-type of breast cancer. Trastuzumab is a monoclonal antibody that binds to the HER2 domain outside the cell and, among other interactions, inhibits homodimerization [92]. In binding of Trastuzumab, the HER2 signaling pathway is altered, affecting transcription factors that regulate many genes [90]. When given the drug in the early stages of breast cancer development, the patient survival rate increases dramatically [90]. Although Trastuzumab has been used clinically since 1998, not every aspect of the biological mechanisms of the drug are understood. This paper aims to shed light on the effects that Trastuzumab has on AS in HER2+ breast cancer cells.

The Illumina SRS technology platform has been the dominant sequencing platform for the past decade. However, short-read sequencing technology has its drawbacks. Especially in the detection of structural variants and genome assembly, the reads are too short to detect over 70% of the structural variation of the human genome [44]. Another issue with Illumina sequencing is how to read and interpret sequences with high levels of GC repeating regions [44]. The third-generation long-read sequencing methods, like ONT, allow AS to be observed by sequencing full-length cDNA transcripts and comparing these to a reference genome [44, 62]. ONT has been shown to produce better results than SRS due to its ability to sequence strands up to 2.3Mb in length, although at the cost of a higher error rate, which allows the method to address the issues SRS faces by sequencing at lengths across repeating regions [44]. Therefore, ONT can detect novel mRNA isoforms with greater sensitivity by sequencing full-length mRNA or cDNA.

A suitable computational pipeline is required for the analysis of high-throughput sequence data. One such system is Nextflow [63, 57], which can be used to create a pipeline

for long read analysis using the LIQA software discussed in Chapter 2. Nextflow is a powerful and scalable workflow management system designed specifically for bioinformatics and scientific data analysis. It simplifies the creation, execution, and maintenance of complex computational pipelines, allowing researchers to seamlessly integrate diverse tools and processes into a cohesive workflow. Nextflow utilizes a domain-specific language that facilitates the definition of workflows in a portable and reproducible manner. Its key strength lies in its ability to manage dependencies, handle parallelization, and efficiently distribute computing tasks across various computing environments, such as local machines, clusters, or cloud platforms [57]. The analysis pipeline was then used to perform a comparative analysis on ONT long read sequenced data to test the effects that Trastuzumab has on AS in HER2+ breast cancer.

4.2 Methods

Cell Lines and Culturing

Figure 4.1 depicts the workflow for the following methods. Human cell lines BT474, SKBR3, and MCF-7 cells were purchased from the American Type Culture Collection and cultured according to their instructions. Human cell line BT474 over-expresses HER2 and ER [93, 94]. The second is the human cell line SKBR3 which over-expresses HER2 [94]. The final cell line is MCF7 that does not overexpress HER2 or ER [94]. MCF7 will act as a negative control as treatment with Trastuzumab should have little effect as neither HER2 nor ER is over-expressed. BT474 and SKBR3 will act as experiment groups for HER2/ER and only HER2, respectively. BT474, SKBR3 and MCF7 will grow for up to 4-5 weeks after thawing from cryo-preservation [95].

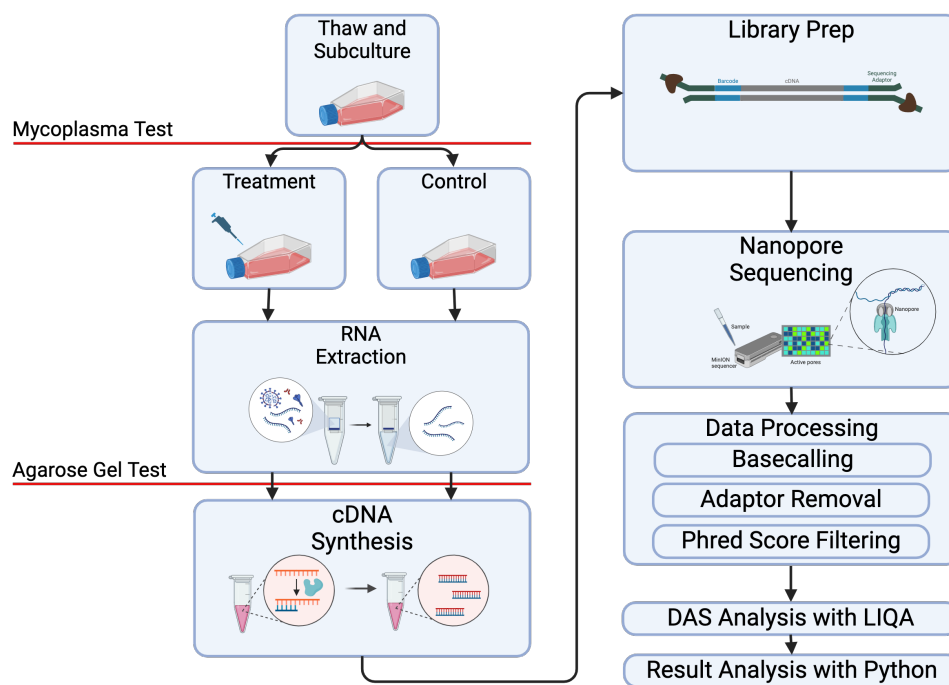


Figure 4.1: **Workflow overview of AS experiment.** Created with BioRender.com.

BT474, SKBR3 and MCF7 cells were grown in Dulbecco's Modified Eagle Medium with 10% Fetal Bovine Serum and 1% Pen-strep for up to 4-5 weeks after thawing from cryo-preservation [95]. Subculture occurred once the flask coverage reached 80-90%, following the standard protocol for each cell line.

Mycoplasma presence was tested using the EZ-PC™ Mycoplasma Detection Kit from FroggBio. The manufacturer protocol was followed, and all cell lines tested negative for mycoplasma.

Cell Treatment and RNA Extraction

Cell cultures were grown to 80-90% confluency over 7 days, then was split into treatment and control groups. Treatment groups received 22 µg of Trastuzumab on days 1 and 4 after subculture by pipetting 5 µL into new media. Control cultures also had media changed on days 1 and 4, without treatment of Trastuzumab. On day 7, treatment and control cells were lysed and their RNA was extracted using the Quick-RNA™, Miniprep kit from Zymo Research. The standard protocol was followed except for using 2 columns for each sample and combining the elution of each column. The quality and quantity of total RNA isolated from each sample were tested using a GelRed electrophoresis gel (4.3) and a Qubit fluorometer Broad-Range RNA kit, respectively.

cDNA synthesis

First strand synthesis of cDNA was performed with the Maxima H Minus First Strand cDNA Synthesis Kit (#K1652). One replicate of each cell line used 1 µL of M-MLV from the RT-PCR kit (Cat. #R1011, 20 reactions). The Invitrogen second-strand cDNA synthesis

kit and protocol were followed for second-strand synthesis. 180 μ L of AMPure XP beads were used in each sample for bead purification with a binding time of 10 minutes at room temperature (20 C). 80% ethanol was used to wash the beads twice. The purified dsDNA was eluted in 14 μ L of nuclease-free water and stored at -20 C.

Library Prep and Minion Sequencing

Library prep was performed with the ONT Native Barcoding Kit 24 V14 (SQK-NBD114.24) and sequenced on the MinION Mk1B with the FLO-MIN114 flowcell. Three sequencing runs with differing parameters were performed. Runs 1 and 2 were normalized to 130ng of dsDNA per sample as input, while run 3 had between 669.2 ng and 2583.0 ng, which was all the remaining sample. The rationale was that sequencing depth would be increased in run 3. Furthermore, run 1 had all 24 samples in the sequenced pool, while runs 2 and 3 only had 18 due to lack of material. Sequencing runs 2 and 3 ran for 72 hours, while run 1 ran for 13 hours.

Post-Sequencing Processing

All code used for this experiment is available on Github at:

<https://github.com/benpiticar/Nextflow-Liqa>

A custom Nextflow script was developed and used to perform the post-sequencing processing [57]. Pod5 files were basecalled and demultiplexed using the Guppy base call software (v6.4.6 + ae70e8f) [45]. Adapters were removed with Porechop (v0.2.4) and reads were filtered according to length and quality with Chopper (v0.6.0) [96, 49]. The processed

reads were then mapped to the NCBI Homo Sapiens GRCh38 DNA Primary Assembly using Minimap2 (v2.17-r941) with the -ax splice parameters. Mapped reads were then filtered and sorted with Samtools (v1.9) [51, 50]. Filtering parameters with read length ≥ 200 , Phred score ≥ 10 , and MAPQ score ≥ 60 were established. The score threshold was determined by testing the number of reads after processing versus the number of reads mapped and determining which threshold would produce high quality results with sufficient data for analysis.

Differential AS Analysis

LIQA (v1.1.16) was used to perform isoform identification and differential AS identification [63]. The read thresholds were optimized to obtain the maximum amount of data without compromising the results. A false discovery rate correction was used to determine statistical significance for genes showing potential AS.

Gene Ontology/KEGG Enrichment Analysis

The GO and Kyoto Encyclopedia of KEGG terms for genes that showed significant AS were identified with the Python module Gprofiler (v1.2.2) [97]. Another false discovery rate correction was performed to determine the GO and KEGG pathways that were significantly overrepresented to display significant AS in each cell line and between cell lines [69, 70, 71]. Individual gene splicing was visualized using ggshashimi (v1.1.5) [98].

4.3 Results and Discussion

The human cell lines BT474, SKBR3, and MCF7 were used to conduct the experiments. According to liturgy and manufacturer documentation, the overexpression of HER2 and ER is shown in table 4.1 [99].

Table 4.1: **Cell lines used and their gene expression types.**

Cell Line	HER2+	ER+
BT474	+	+
SKBR3	+	-
MCF7	-	-

Each cell line was tested for Mycoplasma contamination. Mycoplasma contamination in cell cultures poses significant challenges by altering cell characteristics, such as gene expression which can interfere with experimental results, and impacting cell viability through nutrient depletion. These small wall-less bacteria evade detection and filtration methods, leading to unreliable data and potentially compromising the safety and efficacy of biopharmaceutical products. Their resistance to many antibiotics due to the lack of a cell wall makes eradication difficult, necessitating regular testing and strict sterile techniques to prevent and manage contamination. A polymerase chain reaction (PCR) was performed to amplify mycoplasma DNA found in the cell culture media. Gel electrophoresis was used to detect the presence of amplified mycoplasma DNA. Mycoplasma was not detected in all cell lines, as seen in Figure 4.2, confirming the negative result.

On day 7, the RNA was extracted from both groups of Trastuzumab treated and untreated cells and electrophoresis was performed on agarose gel to determine the purity of the RNA and genomic contamination, shown in the Figure 4.3. Two distinct bands were expected to be found, one for 18S and one for 28S. All samples showed distinct bands with



Figure 4.2: **Mycoplasma test results agarose gel.** Lanes consist of DNA ladder, Positive Control, Negative Control, BT474, SKBR3, and MCF7 from left to right.

no smearing or genomic material in the input well. Some smaller RNA material separate from the distinct bands was also observed. This was probably the mRNA and other RNA molecules that were to be sequenced.

After converting the RNA to cDNA, library prep and sequencing was performed with ONT's SQK-NBD114.24 kit and the Minion FLO-MIN114 flowcell (R10.4.1). An overview of the sequencing portion is shown on table 4.2. Different sequencing times and amounts of input cDNA were used for each run to determine when the majority of sequencing occurred and if an increase in starting material correlated with an increase in sequenced reads. Most sequencing occurs in the first 13 hours and an increase in the input cDNA sample did not improve the sequenced read count. A possible cause of the low read count is a limiting reagent in the library prep kit that is causing the low sequencing yield. SQK-NBD114.24 is typically intended for genomic and amplicon sequencing, and it is possible that it is not

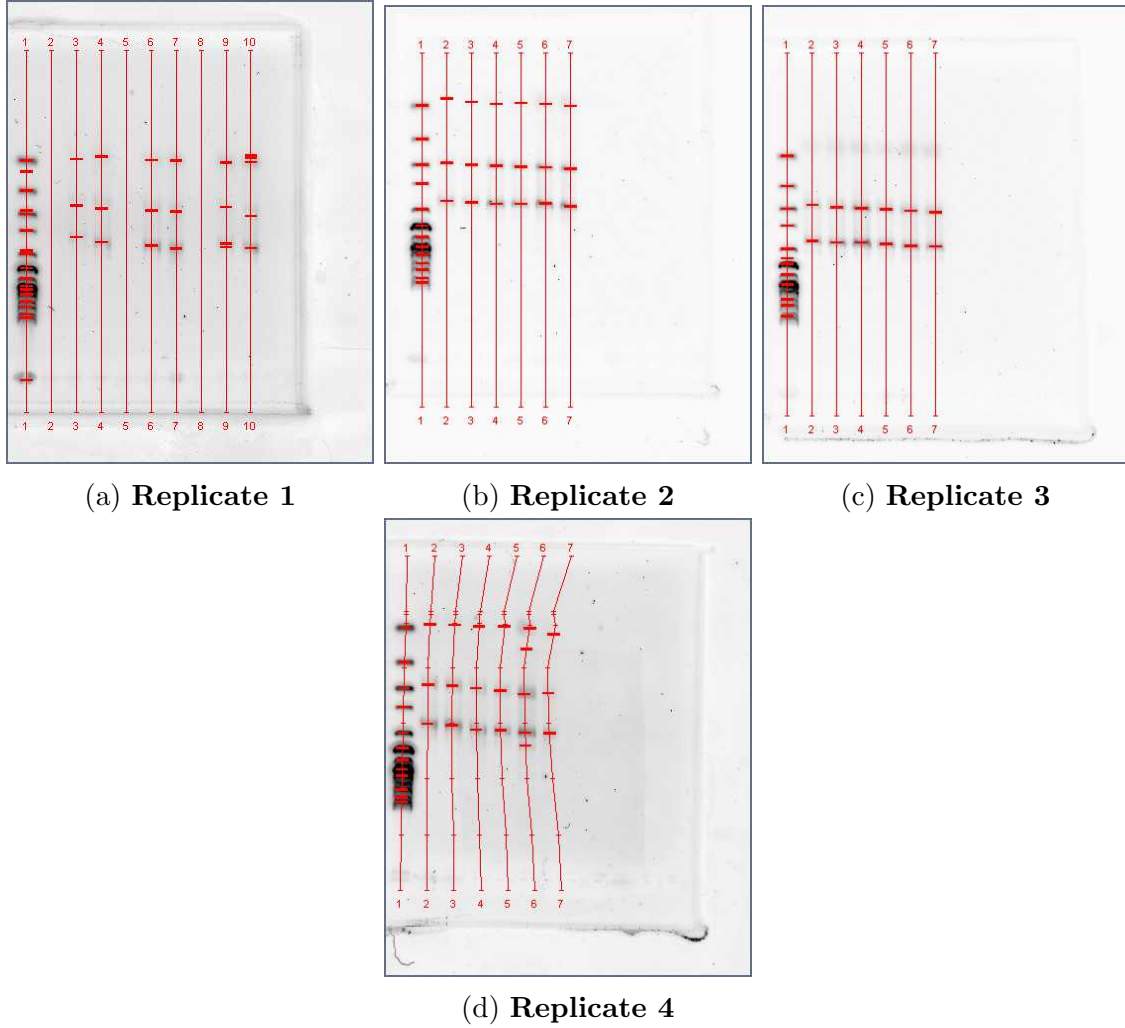


Figure 4.3: **Agarose gels for RNA extraction of all replicates.** For replicates 2,3, and 4, the lanes contain DNA ladder (lane 1), BT474 Treatment (lane 2), BT474 Control (lane 3), SKBR3 Treatment (lane 4), SKBR3 Control (lane 5), MCF7 Treatment (lane 6), and MCF7 Control (lane 7). (a) Replicate 1 lanes contain DNA ladder (lane 1), BT474 Treatment (lane 3), BT474 Control (lane 4), SKBR3 Treatment (lane 6), SKBR3 Control (lane 7), MCF7 Treatment (lane 9), and MCF7 Control (lane 10). 18S and 28S bands can be seen in all samples and no long genomic material near the wells.

optimized to work with transcriptome sequencing. For future experiments involving DAS analysis of ONT sequenced data, it is recommend increasing the amount of adaptor used per barcoded sample or using another protocol that is specifically for full transcriptome cDNA.

Table 4.2: **AS sequencing overview.**

Sequencing Run	Samples Used	cDNA Used/Sample	Time Sequenced	Reads Sequenced
1	18 + 6*	130ng	13 hours	1.84 M
2	18	130ng	72 hours	4.89 M
3	18	Max amount	72 hours	2.17 M

*6 additional replicates were added to this run.

Once all sequencing runs were completed, the Nextflow pipeline was run with various Phred scores and MAPQ thresholds. Phred scores of 0, 10, 12, and 15 were tested, which are accuracy thresholds of 0% (no error threshold), 90.0%, 93.7%, and 96.8%, respectively. The goal was to determine a suitable cutoff threshold to maintain the quality of the result while retaining most reads. The rationale for this approach was that the structural variation of long reads was being examined, which allowed for a higher mutation/error rate because the focus was on the presence of intron/exons rather than single-nucleotide polymorphisms (SNPs).

MAPQ score, which is a metric to measure alignment accuracy, ranges from 0 to 60 for the minimap2 output [50, 100]. MAPQ scores of 0, 10, 20, 50 and 60 were tested, combined with each Phred score threshold. Increasing the MAPQ threshold had minimal impact on the number of reads mapped. Based on the results in Figure 4.4, a read-average Phred score of 10 and a MAPQ score of 60 provide sufficient read filtering while maintaining most of the collected data. It is worth noting that if sequencing depth was increased, the thresholds

would be able to be increased. However, with the lower sequencing depth obtained, a conservative filtering approach was used as more sequence information is retained.

Once Phred and MAPQ thresholds were determined for filtering, the Nextflow pipeline was rerun. Figure 4.4 shows the distribution of read lengths for all adapter-trimmed reads in samples and the average Phred scores for the samples at each read length. Most reads had a Phred score ≥ 20 with spikes ≤ 20 around read length of 500 and 1000 bases. In terms of data processing, the pipeline removed adaptors from sequenced reads, and filtered reads; less than an average Phred score of 10 (10% error rate), and MAPQ score of 60. The median average Phred score for raw reads in each sample was similar, between 22-25.5, with no apparent pattern compared to samples with higher spreads in Figure 4.5. The read length for the raw reads in each sample was similar, with some samples exhibiting a larger variance at random than others, as seen in the Figure 4.6. Due to similarities in the Phred score and read lengths between samples, biological replicates were pooled for each cell line and condition to increase the sample size for analysis. The mean Phred score of the pooled samples increased due to read processing ($M = 22.67$, $SD = 2.97$ to $M = 23.73$, $SD = 2.91$), seen in Figures 4.7 and 4.8. Finally, the mean and variance in the read length decreased ($M = 1385.50$, $SD = 1187.46$ to $M = 1241.43$, $SD = 1102.27$) and the variance in the average Phred score increased, as seen in Figs. 4.9 and 4.10.

In Figure 4.9, 91,651 raw reads had lengths greater than 5Kb (representing 1.4% of the total sequenced reads), which is high, as the length of mRNA in mammals follows a lognormal distribution with a median of 1.4Kb and a deviation of 2 [101]. Only 33,832 reads with

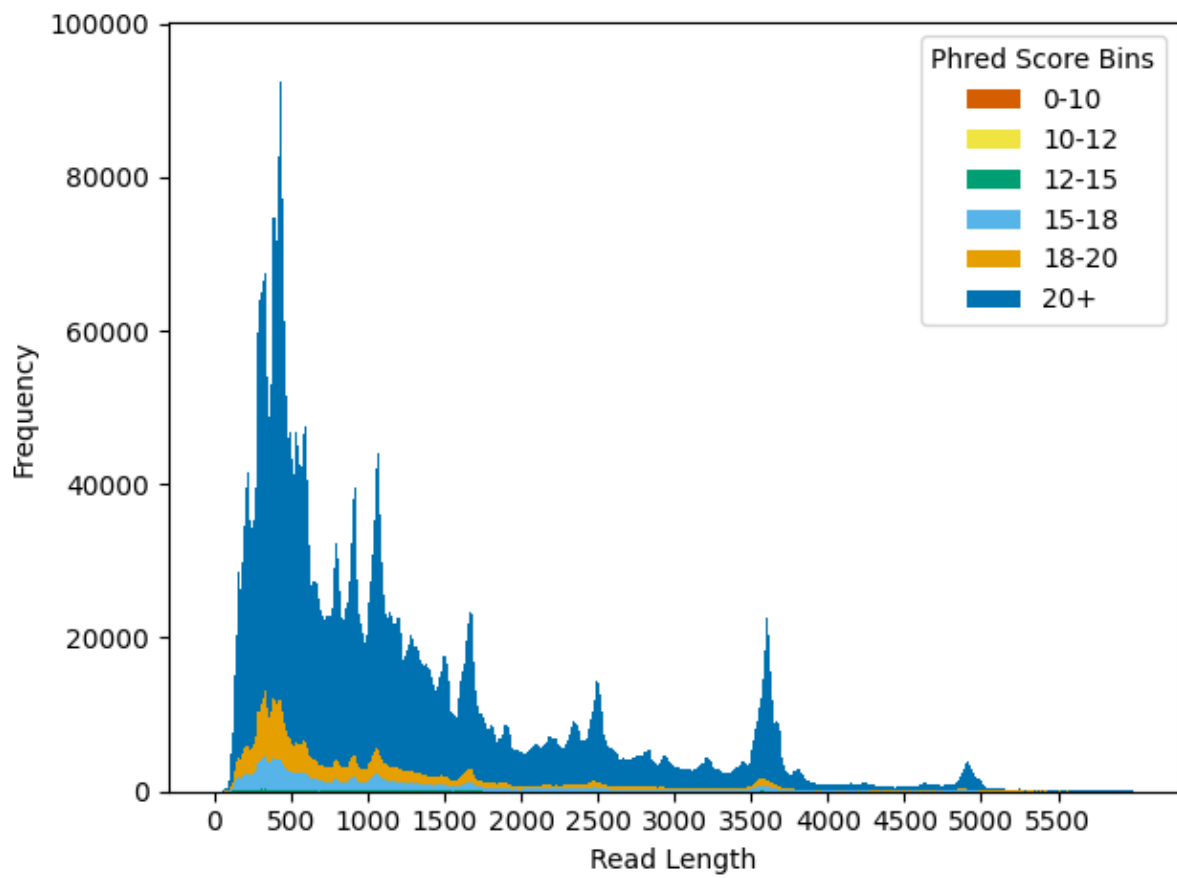


Figure 4.4: Read length histogram for all adaptor trimmed samples with binned stacked Phred scores.

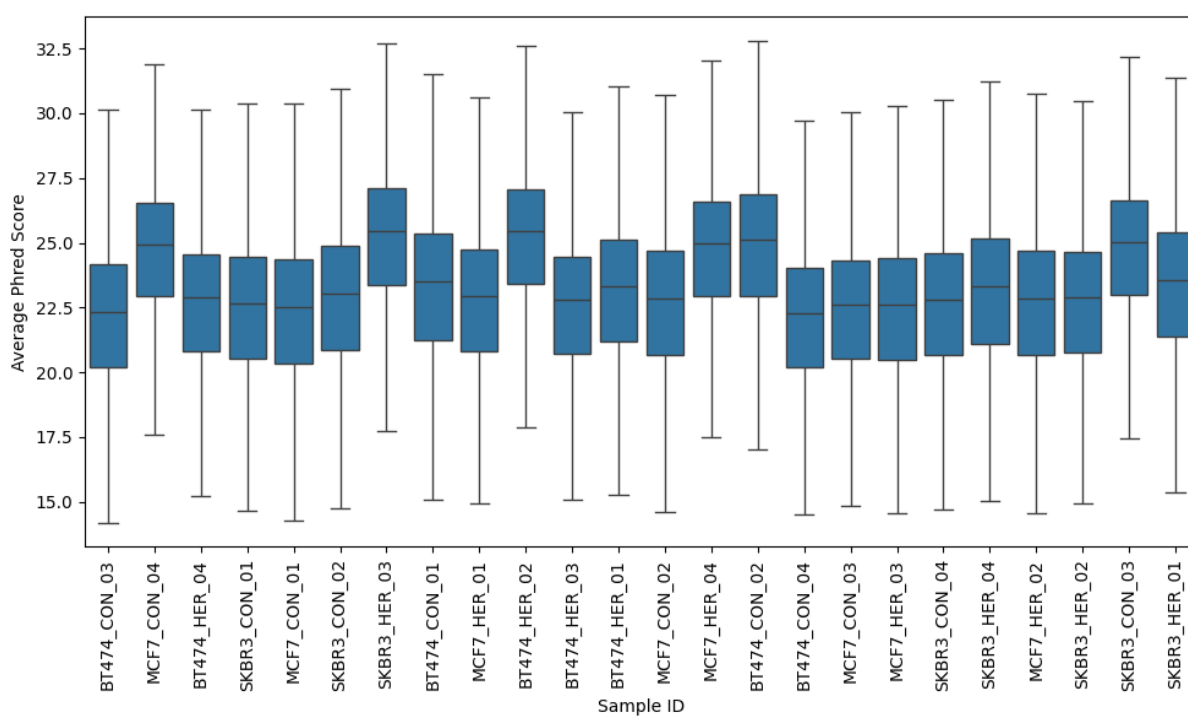


Figure 4.5: Boxplot of average Phred score per read for each sample.

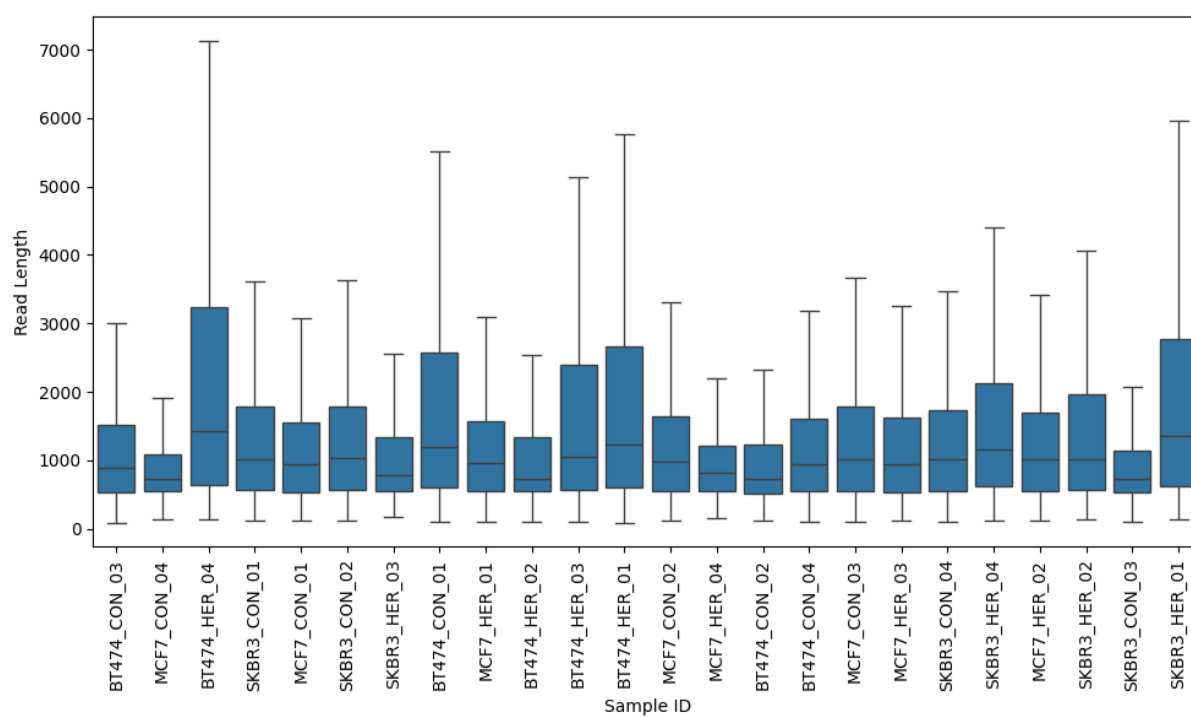


Figure 4.6: Boxplot of read lengths for each sample.

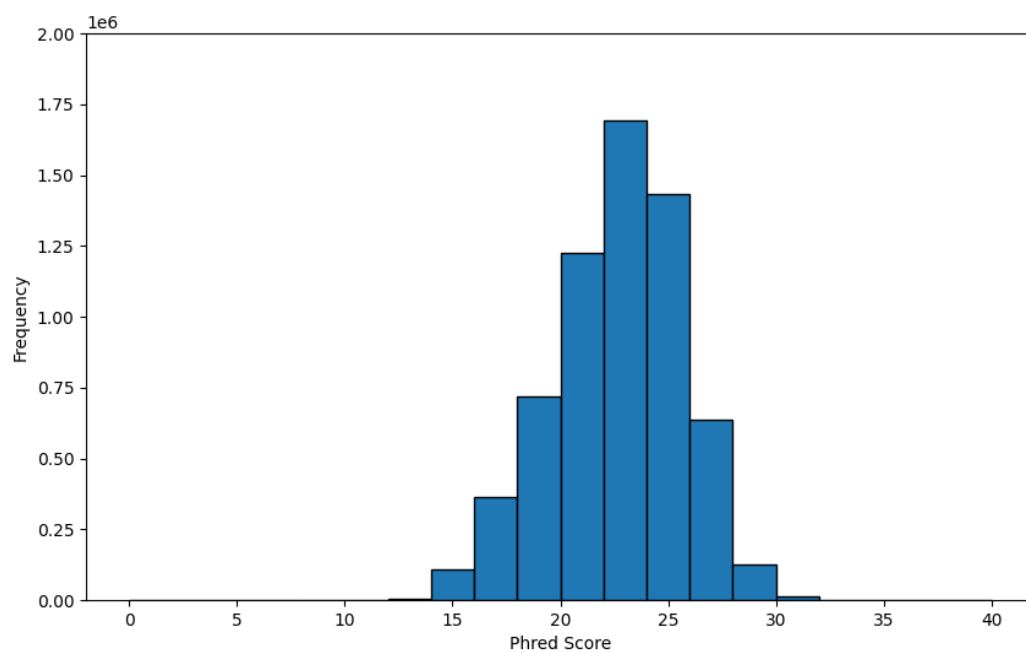


Figure 4.7: **Histogram of raw Phred scores across all samples.**

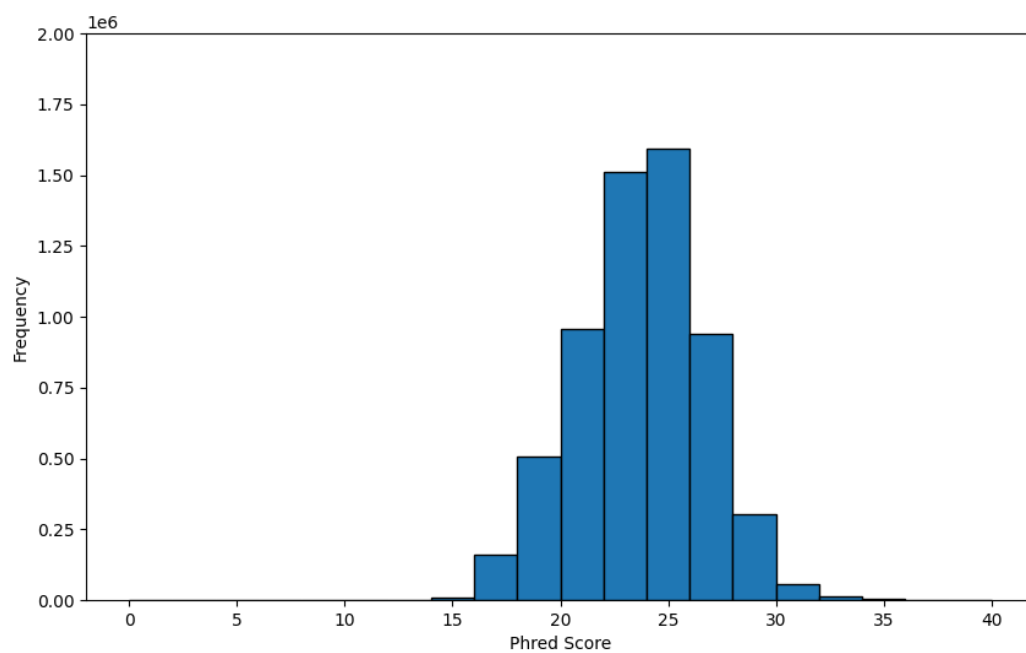


Figure 4.8: **Histogram of processed Phred scores across all samples.**

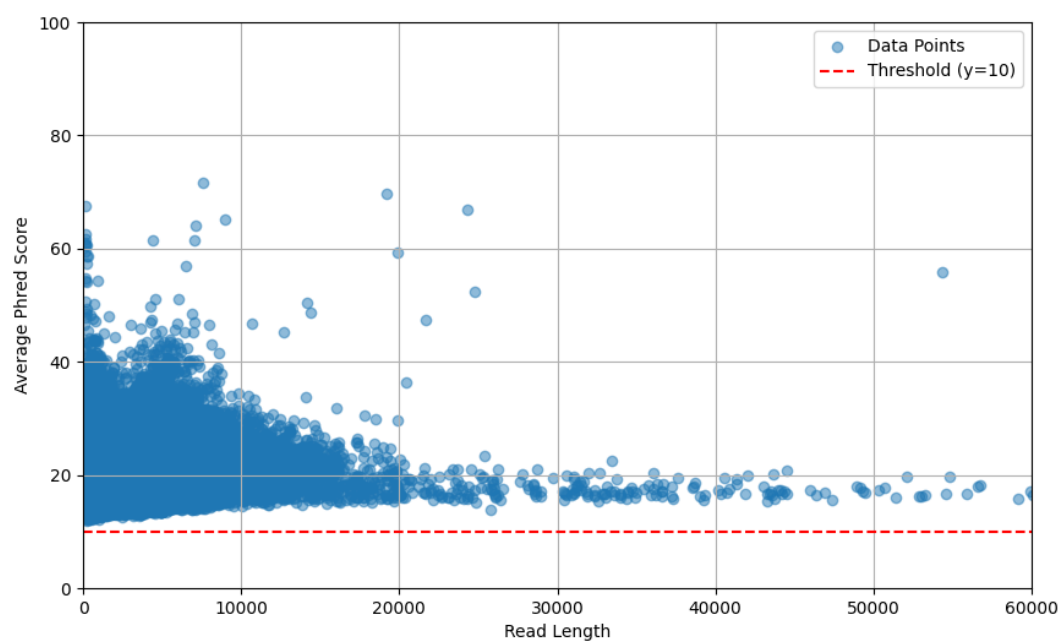


Figure 4.9: Scatterplot of read length vs. average Phred score per read for all raw samples.

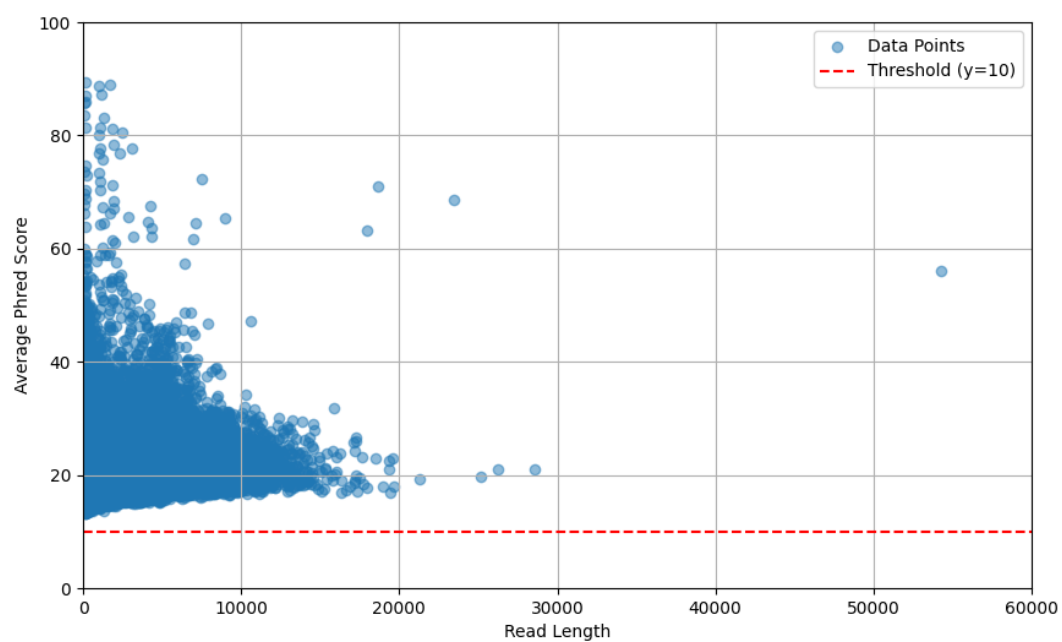


Figure 4.10: Scatterplot of read length vs. average Phred score per read for all processed samples.

lengths greater than 5Kb were found in the processed reads. Porechop (v0.2.4) was used to determine the number of reads that contained an adaptor in the middle of the read and considered the read to be a sequencing error in which the Minion recording software could not differentiate between the end of one read and the beginning of another [47]. Adaptors range from 20 to 60 bases in length, plus the 24 bases for the barcode sequence. Porechop identified 30,670 reads containing adaptors in the middle of the reads. Of the remaining 60,981 reads greater than 5Kb in length, minimap2 was used with the parameters *-ax splice* and a custom script, employing samtools and bedtools to determine the composition of the gene type. After correction of reads per million, 96% of the remaining reads greater than 5Kb mapped to protein coding sequences and the remaining 4% mitochondrial rRNA. The reasoning for using reads per million and not TPM stems from the purpose of determining which type of genes make up these ultra long reads and not a gene expression analysis which commonly uses TPM. Although identification of the type of gene to which these ultra-long reads mapped, there are many questions that remain unanswered. For example, there is a possibility that these reads are genomic contamination. While unlikely due to the stringent filtering with samtools parameters *-F 260 -q 60* which filters out any read that is not a primary mapped read and uniquely mapped, and bedtools requiring a 95% overlap in chromosome location to be considered a match, it is theoretically possible for a DNA strand to escape RNA purification methods. In Figure 4.10, these ultra-long reads are not present as they have been split into smaller reads in the adaptor removal process.

At the end of the DAS analysis, Table 4.3 shows the 27 genes that were found that showed significant AS in the cell line BT474. Eight of the 27 genes had multiple isoforms in the treatment group. NEAT1 expressed significant AS in cell line BT474. NEAT1 is a long non-coding RNA that forms the main component of the paraspeckle sub-organelles. Some research suggests it also acts as a transcriptional regulator for multiple genes involved in

cancer progression [102]. NEAT1 also had significant AS in cell line MCF7 but with a different isoform between the MCF7 and BT474 treatment groups. Upon examining the isoform pileups in Figures A.1, A.2, and A.3, the presence of Trastuzumab caused drastic change in the number of exons and overall size of the isoforms for NEAT1. Interestingly, Trastuzumab caused similar changes in splicing patterns in both cell line BT474, which overexpresses HER2 and ER, and in cell line MCF7 which does not overexpress either HER2 or ER. The exact cause needs further study to confirm that overexpression of HER2 and ER does not play a role in the different isoforms that are expressed in the NEAT1 gene due to Trastuzumab treatment. PRPF38B expressed significant AS in cell lines SKBR3 and MCF7, shown in Figures A.4 and A.5 respectively. The reads in the control group had different isoform structures as seen in Figure A.6 and A.7 but none were found in the treatment groups, possibly due to the presence of Trastuzumab. However, only 2 reads in BT474 and MCF7 mapped to the PRPF38B gene. The localization of PRPF38B can be used as a biomarker to predict the survival benefits of trastuzumab in patients with HER2 + breast cancer [103]. Deeper sequencing depth of PRPF38B in treated BT474 and MCF7 cell lines will illuminate whether Trastuzumab is causing significant AS or changes in gene expression. ARF4 was also found to express significant AS in more than one cell line, BT474 and SKBR3, shown in Figure A.8 and A.9 respectively. ARF4 also showed significant AS between BT474-MCF7 control groups (Figure A.10) and BT474-SKBR3 control groups (Figure A.11). Unregulated ARF4 has been linked to cell migration, invasion and proliferation in several types of cancer [104]. Interestingly, the isoform found in the SKBR3 treatment group (Figure A.9) was the same found in the MCF7 control group (Figure A.10). One interpretation of this finding is that Trastuzumab is causing genes in cells that overexpress HER2 to revert to regular splicing patterns in cells that express regular levels of HER2. Fifteen genes were found in the cell line SKBR3, listed in

Table 4.4, two expressed multiple isoforms in the treatment group. Figure A.12 depicts GAPDH's large change in isoform counts, going from 76 reads assigned to the gene in the control group of SKBR3 to 29 in the treatment group. However, only small changes in different isoform counts were found. GAPDH is a prominent housekeeping protein that is overexpressed in many types of cancer [105]. In a study from 2012 and 2023, both used the GAPDH mRNA levels as a normalization for the data [106, 107]. While the low sample size is a factor, based on the findings in this experiment, if GAPDH undergoes significant AS, it may not be the best choice for mRNA normalization, rather a potential target for more research and therapies. PPIA is another gene that expressed significant AS in cell line BT474. Figure A.13 shows multiple isoform structures between the treatment (HER) and control (CON) groups. The PPIA protein is responsible for catalyzing the cis-trans isomerization of proline imidic peptide bonds that regulate biological processes including apoptosis, transcription, inflammation, and intracellular signalling [108, 109, 110, 111]. Twenty-eight genes were found in the cell line MCF7, 8 expressing multiple isoforms in the treatment group, shown in Table 4.5. Between the control groups of BT474 and MCF7, 36 genes showed significant AS (Table 4.7). There were 44 genes with significant differentially spliced genes between BT474 and SKBR3 (Table 4.6) and 25 between SKBR3 and MCF7 4.8.

Table 4.3: Genes expressing significant alternative splicing between cell line BT474 treatment and control groups.

Gene Name	adj p-value	Description
PPIA	0.000000	peptidylprolyl isomerase A
FLNA	0.000000	Filamin A
Continued on next page		

Genes expressing significant alternative splicing between cell line BT474 treatment and control groups.

Gene Name	adj p-value	Description
VTA1	0.000000	vesicle trafficking 1
SMC1A	0.000000	structural maintenance of chromosomes 1A
ATP6V1A	0.000000	H(+)-transporting two-sector ATPase
SUPT6H	0.000000	Transcription elongation factor spt6
ANGEL1	0.000000	Angel homolog 1
ARF4	0.000008	ADP-ribosylation factor
RBM3	0.000008	RNA binding motif protein 3
H2AC6	0.000134	Histone H2A
PEX26	0.000192	Peroxisome assembly protein 26
MTHFD2	0.000192	methenyltetrahydrofolate cyclohydrolase
RSRC2	0.001662	arginine and serine rich coiled-coil 2
NEAT1	0.001894	nuclear paraspeckle assembly transcript 1
ZNF652	0.002657	Zinc finger protein 652
CLPTM1	0.002657	Cleft lip and palate transmembrane protein 1
PFDN4	0.002921	Prefoldin subunit 4
C6orf62	0.002921	Chromosome 6 open reading frame 62
SSR1	0.004263	signal sequence receptor subunit 1
NIBAN2	0.010516	Niban apoptosis regulator 2
MICOS10	0.017827	MICOS complex subunit MIC10
TOB1	0.024012	Protein Tob1

Continued on next page

Genes expressing significant alternative splicing between cell line BT474 treatment and control groups.

Gene Name	adj p-value	Description
HMGB2	0.027383	High-mobility group box 2
MPI	0.044999	mannose phosphate isomerase
VEZF1	0.046871	vascular endothelial zinc finger 1
CCNG2	0.046954	Cyclin-G2
C8orf82	0.048743	Chromosome 8 open reading frame 82

Table 4.4: **Genes expressing significant alternative splicing between cell line SKBR3 treatment and control groups.**

Gene Name	adj p-value	Description
BCAS4	0.000000	Breast carcinoma amplified sequence 4
AHR	0.000005	Aryl hydrocarbon receptor
PRPF38B	0.000023	Pre-mRNA-splicing factor 38B
HSD17B10	0.000030	3-hydroxyacyl-CoA dehydrogenase type-2
SDF4	0.000054	stromal cell derived factor 4
APPL1	0.000114	DCC-interacting protein 13-alpha
TMEM109	0.001033	Transmembrane protein 109
CALR	0.005584	Calreticulin
HIPK1	0.006876	Homeodomain interacting protein kinase 1
ARF4	0.006876	ADP-ribosylation factor

Continued on next page

Genes expressing significant alternative splicing between cell line SKBR3 treatment and control groups.

Gene Name	adj p-value	Description
MCL1	0.007858	MCL1 protein
MRPS16	0.012981	mitochondrial ribosomal protein S16
GAPDH	0.016266	Glyceraldehyde-3-phosphate dehydrogenase
FAM83H	0.045744	family with sequence similarity 83 member H
AP1S1	0.049220	AP complex subunit sigma

Table 4.5: **Genes expressing significant alternative splicing between cell line MCF7 treatment and control groups.**

Gene Name	adj p-value	Description
HDGF	0.000000	heparin binding growth factor
HSP90AA1	0.000000	Heat shock protein HSP 90-alpha
ENSG00000124593	0.000000	novel protein
ATIC	0.000000	Bifunctional purine biosynthesis protein ATIC
QPRT	0.000000	quinolinate phosphoribosyltransferase
SEC62	0.000000	Translocation protein SEC62
LSM5	0.000000	U6 snRNA-associated Sm-like protein LSm5
RTN4	0.000000	Reticulon
GNA12	0.000000	G protein subunit alpha 12
TSC22D1	0.000001	TSC22 domain family member 1

Continued on next page

Genes expressing significant alternative splicing between cell line MCF7 treatment and control groups.

Gene Name	adj p-value	Description
UQCRQ	0.000003	Cytochrome b-c1 complex subunit 8
RPN1	0.000004	ribophorin I
SCAND1	0.000008	SCAN domain containing 1
ATN1	0.000015	Atrophin 1
CD2BP2	0.000020	CD2 cytoplasmic tail binding protein 2
PRPF38B	0.000127	Pre-mRNA-splicing factor 38B
UBE2K	0.001384	E2 ubiquitin-conjugating enzyme
PDXK	0.001808	Pyridoxal kinase
CSE1L	0.002676	Exportin-2
SCAF8	0.002676	SR-related CTD associated factor 8
UBE2Z	0.003408	Ubiquitin-conjugating enzyme E2 Z
PACSIN2	0.005653	protein kinase C and casein kinase sub. in neurons 2
NEAT1	0.008693	nuclear paraspeckle assembly transcript 1
CREBBP	0.009021	CREB binding protein
ATP6V0E1	0.012804	V-type proton ATPase subunit
PRRC2A	0.020351	Proline-rich coiled-coil 2A
ABI2	0.033503	Abl-interactor 2
BASP1	0.048726	Brain abundant membrane attached signal protein 1

Table 4.6: Genes expressing significant alternative splicing between cell line BT474 and SKBR3 control groups.

Gene Name	adj p-value	Description
UQCRQ	0.000000	Cytochrome b-c1 complex subunit 8
NUTF2	0.000000	Nuclear transport factor 2
CMBL	0.000000	Carboxymethylenebutenolidase homolog
HSD17B10	0.000000	3-hydroxyacyl-CoA dehydrogenase type-2
MEPCE	0.000000	RNA methyltransferase
TMPO	0.000000	thymopoietin
ARF4	0.000002	ADP-ribosylation factor
NEDD8	0.000002	NEDD8
RPL23A	0.000006	ribosomal protein L23a
NHERF2	0.000009	NHERF family PDZ scaffold protein 2
PAQR4	0.000015	progesterin and adiponQ receptor family member 4
EIF3D	0.000054	Eukaryotic translation initiation factor 3 subunit D
CNN3	0.000115	Calponin
DCTPP1	0.000210	dCTP pyrophosphatase 1
RBM42	0.000330	RNA-binding protein 42
SELENOF	0.000330	Selenoprotein F
MTA2	0.001165	Mta2 protein
QSOX1	0.001411	Sulfhydryl oxidase
MRPS16	0.001703	mitochondrial ribosomal protein S16
PRPF38B	0.001779	Pre-mRNA-splicing factor 38B

Continued on next page

Genes expressing significant alternative splicing between cell line BT474 and SKBR3 control groups.

Gene Name	adj p-value	Description
MICOS10	0.002706	MICOS complex subunit MIC10
STARD7	0.002706	StAR related lipid transfer domain containing 7
OTUD6B-AS1	0.003045	OTUD6B antisense RNA 1 (head to head)
RXRA	0.005189	retinoid X receptor alpha
PCBP4	0.005189	Poly(RC)-binding protein 4
H2AC6	0.005375	Histone H2A
DDX21	0.007388	RNA helicase
UBQLN1	0.007443	Ubiquilin-1
CSNK2A2	0.010382	casein kinase 2 alpha 2
TMEM167A	0.012547	Protein kish
NME1-NME2	0.012602	NME1-NME2 readthrough
ARCN1	0.014300	archain 1
SLC7A5	0.023441	Solute carrier family 7 member 5
ATP9A	0.024450	ATP synthase subunit 9, mitochondrial
PSMC2	0.024450	proteasome 26S subunit, ATPase 2
ENSG00000288534	0.025171	TMX2-CTNND1 readthrough (NMD candidate)
MRPS14	0.027020	mitochondrial ribosomal protein S14
MED1	0.032321	mediator complex subunit 1
MPI	0.038142	mannose phosphate isomerase
SNRNP200	0.041480	small nuclear ribonucleoprotein U5 subunit 200

Continued on next page

Genes expressing significant alternative splicing between cell line BT474 and SKBR3 control groups.

Gene Name	adj p-value	Description
BCL9L	0.041539	BCL9 like
YEATS2	0.043772	YEATS domain containing 2
KPNA4	0.047423	Importin subunit alpha
ATN1	0.048678	Atrophin 1

Table 4.7: Genes expressing significant alternative splicing between cell line BT474 and MCF7 control groups.

Gene Name	adj p-value	Description
STARD7	0.000000	StAR-related lipid transfer protein 7, mitochondrial
CREBBP	0.000000	histone acetyltransferase
PACSIN2	0.000000	Protein kinase C and casein kinase substrate in neurons protein 2 isoform X1
UQCRQ	0.000000	Cytochrome b-c1 complex subunit 8
CLNS1A	0.000000	Methylosome subunit pICln
DDX46	0.000000	RNA helicase
NHP2	0.000000	H/ACA ribonucleoprotein complex subunit 2
ELP2	0.000000	Elongator complex protein 2
ARF4	0.000000	ADP-ribosylation factor
TSEN34	0.000000	tRNA-splicing endonuclease subunit Sen34

Continued on next page

Genes expressing significant alternative splicing between cell line BT474 and MCF7 control groups.

Gene Name	adj p-value	Description
NEAT1	0.000000	nuclear paraspeckle assembly transcript 1
OTUD6B-AS1	0.000000	OTUD6B antisense RNA 1 (head to head)
SMC1A	0.000000	Structural maintenance of chromosomes protein
SUPT6H	0.000000	Transcription elongation factor spt6
PSMA1	0.000000	Proteasome subunit alpha type
RPL10A	0.000000	Ribosomal protein
SEC62	0.000001	Translocation protein SEC62
BCL9L	0.000007	BCL9 like
HMGB2	0.000012	High-mobility group box 2
LANCL1	0.000014	Glutathione S-transferase LANCL1
TAB2	0.000267	TGF-beta-activated kinase 1 and MAP3K7-binding protein 2
ATP9A	0.000372	ATP synthase subunit 9, mitochondrial
ZCRB1	0.000726	Zinc finger CCHC-type and RNA-binding motif-containing protein 1
TXNDC5	0.001225	protein disulfide-isomerase
PRPF38B	0.001345	Pre-mRNA-splicing factor 38B
MSI2	0.001376	Musashi RNA binding protein 2
ENSG00000258466	0.001886	novel protein
DSCAM-AS1	0.002185	DSCAM antisense RNA 1
Continued on next page		

Genes expressing significant alternative splicing between cell line BT474 and MCF7 control groups.

Gene Name	adj p-value	Description
SCAND1	0.009361	SCAN domain-containing protein 1
MRPL51	0.009721	Large ribosomal subunit protein mL51
SPAG7	0.011720	Sperm-associated antigen 7 homolog
CCNI	0.020537	Cyclin-I
AK6	0.020537	Adenylate kinase isoenzyme 6
SRP14	0.032246	Signal recognition particle 14 kDa protein
H2AC6	0.032246	Histone H2A
ENSG00000124593	0.032246	novel protein

Table 4.8: Genes expressing significant alternative splicing between cell line MCF7 and SKBR3 control groups.

Gene Name	adj p-value	Description
UBE2Z	0.000000	Ubiquitin-conjugating enzyme E2 Z
GNA12	0.000000	G protein subunit alpha 12
PABPN1	0.000000	Polyadenylate-binding protein 2
NUTF2	0.000000	Nuclear transport factor 2
SLC25A44	0.000000	Solute carrier family 25 member 44
SPPL2A	0.000000	Signal peptide peptidase-like 2A
DNAJC5	0.000000	DnaJ homolog subfamily C member 5

Continued on next page

Genes expressing significant alternative splicing between cell line MCF7 and SKBR3 control groups.

Gene Name	adj p-value	Description
GOLGB1	0.000000	golgin B1
PIEZO1	0.000000	Piezo-type mechanosensitive ion channel component 1
AKT1S1	0.000000	AKT1 substrate 1
PNKD	0.000000	PNKD metallo-beta-lactamase domain containing
NPEPL1	0.000009	aminopeptidase like 1
PSMA1	0.000009	Proteasome subunit alpha type
STARD7	0.000022	StAR related lipid transfer domain containing 7
HIF1A	0.000023	Hypoxia-inducible factor 1 alpha
UBE2L3	0.000058	E2 ubiquitin-conjugating enzyme
DDX21	0.000740	RNA helicase
RAP1A	0.002312	small monomeric GTPase
SLC38A2	0.006063	solute carrier family 38 member 2
FAM83H	0.006320	family with sequence similarity 83 member H
DDIT4	0.019521	DNA damage inducible transcript 4
CALU	0.019521	calumenin
KDELRL1	0.030372	ER lumen protein-retaining receptor
AK6	0.035540	adenylate kinase 3
CISD3	0.042890	CDGSH iron sulfur domain 3

A GO enrichment analysis was performed to determine which GO or KEGG terms were overrepresented in the list of significant AS genes for each cell line and between cell line

control groups. Intrinsic apoptotic signaling pathway in response to DNA damage was identified as a biological process significantly overrepresented in genes that expressed AS between the treatment and control groups in cell line SKBR3, shown in Table 4.9. MCF7 had only two GO terms significantly over representation and both were cellular component in function, shown in Table 4.10. However, between cell line control groups MCF7 and SKBR3, in Table 4.13, the largest number of GO terms were significantly over represented with 5 of the 6 GO terms being biological process in function. Tables 4.12 and 4.11 list the GO terms found between cell lines BT474 and MCF7, and BT474 and SKBR3 respectively. Furthermore, when the filtering Phred score threshold was changed, genes that expressed significant AS and significantly enriched GO terms changed drastically. This led to further investigation of the read coverage of each AS event and found that of the genes that expressed significant AS, less than 77 reads were assigned to each isoform, some as low as a count of 1 to 2.

Table 4.9: **GO/KEGG enrichment analysis results for cell line SKBR3.**

Source	Native	Name	p-value	Term Size	Query Size	Precision	Recall
GO:BP	GO:0008630	intrinsic apoptotic signaling pathway in response to DNA damage	0.036790	106	12	0.250000	0.028302

Table 4.10: **GO/KEGG enrichment analysis results for cell line MCF7.**

Source	Native	Name	p-value	Term Size	Query Size	Precision	Recall
GO:CC	GO:0005654	nucleoplasm	0.008664	4220	25	0.560000	0.003318
GO:CC	GO:0032433	filopodium tip	0.040720	19	25	0.080000	0.105263

Table 4.11: **GO/KEGG enrichment analysis results for cell line BT474 vs SKBR3.**

Source	Native	Name	p-value	Term Size	Query Size	Precision	Recall
GO:MF	GO:0050693	LBD domain binding	0.036224	7	40	0.050000	0.285714

Table 4.12: **GO/KEGG enrichment analysis results for cell line BT474 vs MCF7.**

Source	Native	Name	p-value	Term Size	Query Size	Precision	Recall
GO:CC	GO:0043231	intracellular membrane-bounded organelle	0.009867	14605	31	0.967742	0.002054
GO:CC	GO:0005654	nucleoplasm	0.010714	4220	31	0.516129	0.003791
GO:CC	GO:0005681	spliceosomal complex	0.041922	206	31	0.129032	0.019417

Table 4.13: **GO/KEGG enrichment analysis results for cell line MCF7 vs SKBR3.**

Source	Native	Name	p-value	Term Size	Query Size	Precision	Recall
GO:CC	GO:0005737	cytoplasm	0.000093	12345	25	1.000000	0.002025
GO:BP	GO:0038179	neurotrophin signaling pathway	0.021634	40	24	0.125000	0.075000
GO:BP	GO:0007269	neurotransmitter secretion	0.035673	145	24	0.166667	0.027586
GO:BP	GO:0099643	signal release from synapse	0.035673	145	24	0.166667	0.027586
GO:BP	GO:0032006	regulation of TOR signaling	0.037647	147	24	0.166667	0.027211

Several of the genes that expressed AS showed reads that corresponded to isoforms only under one condition. For example, in cell line BT474, gene High-mobility group protein B2 (HMGB2) had 9 reads mapping to 6 different isoforms in the control group but 0 in the treatment group. This raised concern that the significant change observed in AS was due to changes in gene expression. A gene expression analysis was performed for each cell line independently with a custom bash and Python script, implementing samtools, bedtools, and pyDeseq2. At the end of the analysis, no genes were found showing significant changes in gene expression due to Trastuzumab treatment. Although this is likely inaccurate due to the low sample size, the lack of differences in expression level suggests that changes in expression are unlikely to be producing false positive results for AS.

4.4 Conclusions

The development of this high-throughput pipeline and sequencing protocol shows promise. Notable AS events were identified due to the presence of Trastuzumab, causing different splice patterns between HER2+ and control cell lines. Trastuzumab caused ARF4 in HER2+ cell lines to express isoforms similar to the control group of regular expressing HER2. With more sequencing, a larger dataset would add more significance power to support or provide more information on the genes and biological pathways impacted by

Trastuzumab. The protocol used can further be optimized by enabling adaptive sampling to increase protein coding representation. However, this comes with its own set of challenges and limitations. The pipeline that was developed shows that a multistep complex RNA sequencing pipeline can be transformed into a scalable and comprehensive workflow which could be used in different comparative studies. For example, the analysis pipeline could be used to find significant AS events between the collected libraries and the normal breast cell library to determine AS at the start of the study to see if trastuzumab is causing any pathways to return to "normal" splicing.

Chapter 5

Conclusion and Future Directions

This thesis describes the use of transcriptomics and high-throughput data analysis to investigate AS changes in HER2+ breast cancer cells, treated with the monoclonal antibody Trastuzumab. Different methods and tools for sequencing library preparation and AS analysis were compared to determine the best method for accurate, high throughput alternative splicing analysis. The optimized method was then used to sequence and analyze Trastuzumab-treated BT474, SKBR3, and MCF7 cultured cells. The development of the high-throughput Nextflow pipeline and sequencing workflow revealed a great potential for quick and efficient AS detection.

Chapter 2 investigated several AS identification software and compared the differences in methods used. The AS identification software LIQA was determined to be best suited for the AS identification experiment because of its simplicity and ability to be built out into a high throughput Nextflow script. Chapter 2 investigates the methods used in ONT sequencing library preparation. Comparison between prepared libraries that underwent poly(A) enrichment and those that did not revealed that the omission of this step allowed

for reduced RNA sample requirements, without sacrificing the quality of data collected. Chapter 4 implements the strategies from chapter 2 and 3 to determine the significant changes in AS due to Trastuzumab in HER2+ breast cancer cells. Multiple genes were identified in all 3 cell lines that displayed significant AS due to Trastuzumab, some indicating a connection to the drug and overexpression of HER2. While results were obtained, low sequencing yields limited the significance of the findings.

The research conducted in this thesis lays a solid foundation to be built upon. Further testing of the affects that omitting poly(A) selection has on sequencing outcomes will help researchers save time and resources by removing a non-essential step from library preparation protocols. The DAS analysis sequencing workflow and pipeline developed in chapter 4 provides researchers with an efficient and reproducible workflow for any DAS comparative analysis with ONT sequenced data. For example, using the workflow to identify cryptic exons in Amyotrophic Lateral Sclerosis (ALS) tissues. Cryptic exons are non-conserved, non-annotated intronic sequences that can be spliced into the mature mRNA due to mutations that cause new splice sites [112]. These novel mRNA isoforms can lead to novel protein structures and functions that can be targeted for gene-therapy treatments and biomarkers. By sequencing the transcriptome of a healthy mouse tissue and tissue from a mouse with ALS, genes that undergo significant AS can be quickly and efficiently identified. This workflow can also be used in combination with amlicon sequencing and adaptive sequencing on the ONT Flongle flowcell to identify cryptic exons for specific genes that are known to cause ALS. In conclusion, while further develop of the workflows in this can be made, there are many new experiments that can be done with the methods and workflows discussed in this thesis.

References

- [1] G W Li and X S Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308–315, 7 2011.
- [2] Angela L Nicholson and Amy E Pasquinelli. Tales of detailed poly(A) tails. *Trends in Cell Biology*, 2018.
- [3] J A Bernstein, A B Khodursky, P H Lin, S Lin-Chao, and S N Cohen. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15):9697–9702, 7 2002.
- [4] E A Ponomarenko, E V Poverennaya, E V Ilgisonis, M A Pyatnitskiy, A T Kopylov, V G Zgoda, A V Lisitsa, and A I Archakov. The size of the human proteome: The width and depth. *International journal of analytical chemistry*, 2016:7436849, 2016.
- [5] N L Barbosa-Morais, M Irimia, Q Pan, H Y Xiong, S Gueroussov, L J Lee, V Slobodeniuc, C Kutter, S Watt, R Colak, T Kim, C M Misquitta-Ali, M D Wilson, P M Kim, D T Odom, B J Frey, and B J Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)*, 338(6114):1587–1593, 12 2012.

- [6] Yan Wang, Jing Liu, B O Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, Jing Zhang, Qing-Hua Min, Wei-Ming Yang, and Xiao-Zhong Wang. Mechanism of alternative splicing and its regulation. *Biomedical reports*, 3(2):152–158, 3 2015.
- [7] Ramana V. Davuluri, Yutaka Suzuki, Sumio Sugano, Christoph Plass, and Tim H.M. Huang. The functional consequences of alternative promoter use in mammalian genomes. *Trends in genetics : TIG*, 24(4):167–177, 4 2008.
- [8] Maarten M.G. Van Den Hoogenhof, Yigal M. Pinto, and Esther E. Creemers. RNA Splicing. *Circulation Research*, 118(3):454–468, 2016.
- [9] Chuangye Yan, Ruixue Wan, and Yigong Shi. Molecular Mechanisms of pre-mRNA Splicing through Structural Biology of the Spliceosome. *Cold Spring Harbor Perspectives in Biology*, 11(1), 1 2019.
- [10] Stefan Stamm, Shani Ben-Ari, Ilona Rafalska, Yesheng Tang, Zhaiyi Zhang, Debra Toiber, T A Thanaraj, and Hermona Soreq. Function of alternative splicing. *Gene*, 344:1–20, 1 2005.
- [11] Izarbe Aísa-Marín, Rocío García-Arroyo, Serena Mirra, Gemma Marfany, Tamar Ben-Yosef, and Susanne Roosing. Retinal Genes in Health and Disease. *Int. J. Mol. Sci*, page 22, 2021.
- [12] Juan Wang, Jie Zhang, Kaibo Li, Wei Zhao, and Qinghua Cui. SpliceDisease database: linking RNA splicing and disease. *Nucleic acids research*, 40(Database issue), 1 2012.
- [13] André Kahles, Kjong Van Lehmann, Nora C. Toussaint, Matthias Hüser, Stefan G. Stark, Timo Sachsenberg, Oliver Stegle, Oliver Kohlbacher, Chris Sander, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L.

Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson,

Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Belair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, Harsha Vardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandath, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lightenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann,

David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliente, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej

Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn
 Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalka-
 nis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta
 Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang,
 Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason,
 Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Ak-
 eredolu, Sergei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein,
 Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Cas-
 tle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas
 Flotte, Nasra Giamma, Thai Ho, Michael Kendrick, Jean Pierre Kocher, Karla Kopp,
 Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar
 Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Small-
 ridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson,
 Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonza-
 lez, Carmen Behrens, Jolanta Bondaruk, Russell Broadbudd, Bogdan Czerniak, Bitu
 Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar,
 Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta,
 David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso,
 Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Va-
 lerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Gra-
 ham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew
 Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hik-
 mat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Dou-
 glas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien,
 Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jen-

nifer Smith, Howard Zaren, Joong Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl Friedrich Becker, Klaus Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce,

Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Degin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffrey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, and Gunnar Rätsch. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*, 34(2):211–224, 8 2018.

- [14] Lulzim Shkreta, Brendan Bell, Timothée Revil, Julian P. Venable, Panagiotis Prinios, Sherif Abou Elela, and Benoit Chabot. Cancer-Associated Perturbations in Alternative Pre-messenger RNA Splicing. *Cancer treatment and research*, 158:41–94, 2013.
- [15] Michael Mendoza, Garni Mandani, and Jamil Momand. The MDM2 gene family. *Biomolecular concepts*, 5(1):9, 3 2014.
- [16] Jialiang Zheng, Fenglin Miao, Zhao Wang, Yuan Ma, Zhenhang Lin, Yaqin Chen, Xu Kong, Yue Wang, Aobo Zhuang, Ting Wu, and Wengang Li. Identification of MDM2 as a prognostic and immunotherapeutic biomarker in a comprehensive pan-cancer analysis: A promising target for breast cancer, bladder cancer and ovarian cancer immunotherapy. *Life Sciences*, 327:121832, 8 2023.
- [17] Ali M. Alaseem. Advancements in MDM2 inhibition: Clinical and pre-clinical investigations of combination therapeutic regimens. *Saudi pharmaceutical journal : SPJ : the official publication of the Saudi Pharmaceutical Society*, 31(10), 10 2023.
- [18] Karolina I Pellot Ortiz, Julian S Rechberger, Leo F Nonnenbroich, David J Daniels, and Jann N Sarkaria. MDM2 Inhibition in the Treatment of Glioblastoma: From Concept to Clinical Investigation. *Biomedicines*, 11(7), 7 2023.
- [19] Uma Krishnamurti and Jan F Silverman. HER2 in breast cancer: a review and update. *Advances in Anatomic Pathology*, 21(2):100–107, 3 2014.
- [20] Brian A. Hemmings and David F. Restuccia. PI3K-PKB/Akt Pathway. *Cold Spring Harbor Perspectives in Biology*, 4(9), 9 2012.
- [21] Olli P. Kallioniemi, Anne Kallioniemi, Wayne Kurisu, Ann Thor, Ling Chun Chen, Helene S. Smith, Frederic M. Waldman, Dan Pinkel, and Joe W. Gray. ERBB2 am-

- plification in breast cancer analyzed by fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, 89(12):5321, 6 1992.
- [22] Franco Lumachi, Davide A Santeufemia, and Stefano Mm Basso. Current medical treatment of estrogen receptor-positive breast cancer. *World journal of biological chemistry*, 6(3):231–239, 8 2015.
- [23] F Lumachi, A Brunello, M Maruzzo, U Basso, and S M M Basso. Treatment of estrogen receptor-positive breast cancer. *Current medicinal chemistry*, 20(5):596–604, 2013.
- [24] Kartik Aysola, Akshata Desai, Crystal Welch, Jingyao Xu, Yunlong Qin, Vaishali Reddy, Roland Matthews, Charlotte Owens, Joel Okoli, Derrick J Beech, Chandrika J Piyathilake, Shyam P Reddy, and Veena N Rao. Triple Negative Breast Cancer - An Overview. *Hereditary genetics : current research*, 2013(Suppl 2):2161–001, 2013.
- [25] Y Fan, Y Wang, L He, S Imani, and Q Wen. Clinical features of patients with HER2-positive breast cancer and development of a nomogram for predicting survival. *ESMO open*, 6(4):100232, 8 2021.
- [26] M. M. Moasser. The oncogene HER2; Its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene*, 26(45):6469, 10 2007.
- [27] Ruei-Min Lu, Yu-Chyi Hwang, I-J Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science*, 27(1):1, 2020.
- [28] Hamid Maadi, Mohammad Hasan Soheilifar, Won-Shik Choi, Abdolvahab Moshaghian, and Zhixiang Wang. Trastuzumab Mechanism of Action; 20 Years of Re-

- search to Unravel a Dilemma. *Cancers*, 13(14):3540. doi: 10.3390/cancers13143540, 7 2021.
- [29] Hamid Maadi, Babak Nami, Junfeng Tong, Gina Li, and Zhixiang Wang. The effects of trastuzumab on HER2-mediated cell signaling in CHO cells expressing human HER2. *BMC cancer*, 18(1):238–x, 3 2018.
- [30] Carl P Blobel. ADAMs: key components in EGFR signalling and development. *Nature reviews.Molecular cell biology*, 6(1):32–43, 1 2005.
- [31] Milos Dokmanovic, Yun Wu, Yi Shen, Jieqing Chen, Dianne S Hirsch, and Wen Jin Wu. Trastuzumab-induced recruitment of Csk-homologous kinase (CHK) to ErbB2 receptor is associated with ErbB2-Y1248 phosphorylation and ErbB2 degradation to mediate cell growth inhibition. *Cancer biology & therapy*, 15(8):1029–1041, 8 2014.
- [32] Merel Gijzen, Peter King, Tim Perera, Peter J Parker, Adrian L Harris, Banafshé Larijani, and Anthony Kong. HER2 phosphorylation is maintained by a PKB negative feedback loop in response to anti-HER2 herceptin in breast cancer. *PLoS biology*, 8(12):e1000563, 12 2010.
- [33] Claudia Korzeniecki and Ronny Priefer. Targeting KRAS mutant cancers by preventing signaling transduction in the MAPK pathway. *European journal of medicinal chemistry*, 211:113006, 2 2021.
- [34] Cornelia Braicu, Mihail Buse, Constantin Busuioc, Rares Drula, Diana Gulei, Lajos Raduly, Alexandru Rusu, Alexandru Irimie, Atanas G Atanasov, Ondrej Slaby, Calin Ionescu, and Ioana Berindan-Neagoe. A Comprehensive Review on MAPK: A Promising Therapeutic Target in Cancer. *Cancers*, 11(10):1618. doi: 10.3390/cancers11101618, 10 2019.

- [35] Yutaka Kato, Richard I. Tapping, Shuang Huang, Mark H. Watson, Richard J. Ulevitch, and Jiing Dwan Lee. Bmk1/Erk5 is required for cell proliferation induced by epidermal growth factor. *Nature* 1998 395:6703, 395(6703):713–716, 10 1998.
- [36] Babak Nami, Hamid Maadi, and Zhixiang Wang. Mechanisms Underlying the Action and Synergism of Trastuzumab and Pertuzumab in Targeting HER2-Positive Breast Cancer. *Cancers*, 10(10):342. doi: 10.3390/cancers10100342, 9 2018.
- [37] Milos Dokmanovic, Dianne S Hirsch, Yi Shen, and Wen Jin Wu. Rac1 contributes to trastuzumab resistance of breast cancer cells: Rac1 as a potential therapeutic target for the treatment of trastuzumab-resistant breast cancer. *Molecular cancer therapeutics*, 8(6):1557–1569, 6 2009.
- [38] Cary D Austin, Ann M De Mazière, Paul I Pisacane, Suzanne M van Dijk, Charles Eigenbrot, Mark X Sliwkowski, Judith Klumperman, and Richard H Scheller. Endocytosis and sorting of ErbB2 and the site of action of cancer therapeutics trastuzumab and geldanamycin. *Molecular biology of the cell*, 15(12):5268–5282, 12 2004.
- [39] Patrícia M R Pereira, Sai Kiran Sharma, Lukas M Carter, Kimberly J Edwards, Jacob Pourat, Ashwin Ragupathi, Yelena Y Janjigian, Jeremy C Durack, and Jason S Lewis. Caveolin-1 mediates cellular distribution of HER2 and affects trastuzumab binding and therapeutic efficacy. *Nature communications*, 9(1):5137–w, 12 2018.
- [40] Maurizio Scaltriti, Federico Rojo, Alberto Ocaña, Judit Anido, Marta Guzman, Javier Cortes, Serena Di Cosimo, Xavier Matias-Guiu, Santiago Ramon y Cajal, Joaquin Arribas, and José Baselga. Expression of p95HER2, a truncated form of the HER2 receptor, and response to anti-HER2 therapies in breast cancer. *Journal of the National Cancer Institute*, 99(8):628–638, 4 2007.

- [41] Felicia C Huynh and Frank E Jones. MicroRNA-7 inhibits multiple oncogenic pathways to suppress HER2 Δ 16 mediated breast tumorigenesis and reverse trastuzumab resistance. *PloS one*, 9(12):e114419, 12 2014.
- [42] David Skerrett-Byrne Anthony, Chen Jiang Chen, Brett Nixon, and Hubert Hondermarck. Transcriptomics. *Encyclopedia of Cell Biology: Volume 1-6, Second Edition*, 6:363–371, 1 2022.
- [43] Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, 122(1):e59, 4 2018.
- [44] G A Logsdon, M R Vollger, and E E Eichler. Long-read human genome sequencing and its applications. *Nature reviews.Genetics*, 21(10):597–614, 10 2020.
- [45] R R Wick, L M Judd, and K E Holt. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome biology*, 20(1):129–y, 6 2019.
- [46] Clara Delahaye and Jacques Nicolas. Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*, 16(10), 10 2021.
- [47] Quentin Bonenfant, Laurent No  , and H  l  ne Touzet. Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. *Bioinformatics Advances*, 3(1):vbac085, 2022.
- [48] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology 2021 39:11*, 39(11):1348–1365, 11 2021.

- [49] Wouter De Coster and Rosa Rademakers. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*, 39(5):btad311, 2023.
- [50] H Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, 34(18):3094–3100, 9 2018.
- [51] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 2021.
- [52] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.
- [53] Ignacio San Segundo-Val and Catalina S. Sanz-Lozano. Introduction to the gene expression analysis. *Methods in Molecular Biology*, 1434:29–43, 2016.
- [54] Jin Min Xue, Yi Liu, Ling Hong Wan, and Yu Xi Zhu. Comprehensive Analysis of Differential Gene Expression to Identify Common Gene Signatures in Multiple Cancers. *Medical Science Monitor*, 26:0–0, 2 2020.
- [55] Rebecca F Halperin, Apurva Hegde, Jessica D Lang, Elizabeth A Raupach, Research Group, Christophe Legendre, Winnie S Liang, Patricia M Lorusso, Aleksandar Sekulic, Jeffrey A Sosman, Jeffrey M Trent, Sampathkumar Rangasamy, Patrick Pirrotte, and Nicholas J Schork. Improved methods for RNAseq-based alternative splicing analysis. *Scientific Reports* —, 11:10740, 123.
- [56] Jiawei Ouyang, Yijie Zhang, Fang Xiong, Shanshan Zhang, Zhaojian Gong, Qijia Yan, Yi He, Wenling Zhang, Ming Zhou, Bo Xiang, Fuyan Wang, Xiaoling Li, Yong

- Li, Guiyuan Li, Zhaoyang Zeng, Can Guo, and Wei Xiong. The role of alternative splicing in human cancer progression. *Am J Cancer Res*, 11(10):4642–4667, 2021.
- [57] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017.
- [58] Q Pan, O Shai, L J Lee, B J Frey, and B J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, 12 2008.
- [59] Stanley Chun-Wei Lee and Omar Abdel-Wahab. Therapeutic targeting of splicing in cancer. *Nature medicine*, 22(9):976–986, 2016.
- [60] Julia K Nussbacher, Ricardos Tabet, Gene W Yeo, and Clotilde Lagier-Tourenne. Disruption of RNA Metabolism in Neurological Diseases and Emerging Therapeutic Interventions. *Neuron*, 102(2):294–320, 4 2019.
- [61] Ruben G de Bruin, Ton J Rabelink, Anton Jan van Zonneveld, and Eric P van der Veer. Emerging roles for RNA-binding proteins as effectors and regulators of cardiovascular disease. *European heart journal*, 38(18):1380–1388, 5 2017.
- [62] D Deamer, M Akeson, and D Branton. Three decades of nanopore sequencing. *Nature biotechnology*, 34(5):518–524, 5 2016.
- [63] Y Hu, L Fang, X Chen, J F Zhong, M Li, and K Wang. LIQA: long-read isoform quantification and analysis. *Genome biology*, 22(1):182–188, 6 2021.

- [64] Isak Holmqvist, Alan B  ckerholm, Yarong Tian, Guojiang Xie, Kaisa Thorell, and Ka-Wei Tang. FLAME: long-read bioinformatics tool for comprehensive spliceome characterization. *RNA (New York, N.Y.)*, 27(10):1127–1139, 10 2021.
- [65] Alison D. Tang, Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals down-regulation of retained introns. *Nature Communications* 2020 11:1, 11(1):1–12, 3 2020.
- [66] Yupei You, Michael B Clark, and Heejung Shim. NanoSplicer: accurate identification of splice junctions using Oxford Nanopore sequencing. *Bioinformatics*, 38(15):3741–3748, 2022.
- [67] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111, 5 2009.
- [68] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477, 2011.
- [69] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J M Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

- [70] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Ar-

naud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

- [71] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 1 2000.
- [72] Isha Sethi, Rose Anne Romano, Christian Gluck, Kirsten Smalley, Borivoj Vojtesek, Michael J. Buck, and Satrajit Sinha. A global analysis of the complex landscape of isoforms and regulatory networks of p63 in human cells and tissues. *BMC Genomics*, 16(1):1–15, 8 2015.
- [73] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D. Pruitt, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 46(Database issue):D41, 1 2018.
- [74] Fergal J. Martin, M. Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G. Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Alexandra Bignell, Sanjay Boddu, Paulo R. Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Mehrnaz Charkhchi,

Alexander Cockburn, Luca Da Rin Fiorretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Gurpreet S. Ghattaoraya, Jose Gonzalez Martinez, Cristi Guijarro, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Mike Kay, Vinay Kaykala, Tuan Le, Diana Lemos, Diego Marques-Coelho, Jose Carlos Marugán, Gabriela Alejandra Merino, Louisse Paola Mirabueno, Aleena Mush-taq, Syed Nakib Hossain, Denye N. Ogeh, Manoj Pandian Sakthivel, Anne Parker, Malcolm Perry, Ivana Piliota, Irina Prosovetkaia, Jose G. Perez-Silva, Ahamed Imran Abdul Salam, Nuno Saraiva-Agostinho, Helen Schuilenburg, Dan Shep-pard, Swati Sinha, Botond Sipos, William Stark, Emily Steed, Ranjit Sukumaran, Dulika Sumathipala, Marie Marthe Suner, Likhitha Surapaneni, Kyösti Sutinen, Michal Szpak, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veiden-berg, Thomas A. Walsh, Brandon Walts, Elizabeth Wass, Natalie Willhoft, Jamie Allen, Jorge Alvarez-Jarreta, Marc Chakiachvili, Bethany Flint, Stefano Giorgetti, Leanne Haggerty, Garth R. Ilsley, Jane E. Loveland, Benjamin Moore, Jonathan M. Mudge, John Tate, David Thybert, Stephen J. Trevanion, Andrea Winterbottom, Adam Frankish, Sarah E. Hunt, Magali Ruffier, Fiona Cunningham, Sarah Dyer, Robert D. Finn, Kevin L. Howe, Peter W. Harrison, Andrew D. Yates, and Paul Flicek. Ensembl 2023. *Nucleic Acids Research*, 51(D1):D933–D941, 1 2023.

- [75] Marcus Stoiber, Joshua Quick, Rob Egan, Ji Eun Lee, Susan Celniker, Robert K Neely, Nicholas Loman, Len A Pennacchio, and James Brown. De novo Identifica-tion of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv*, page 94672, 2017.
- [76] Andrey D. Prjibelski, Alla Mikheenko, Anoushka Joglekar, Alexander Smetanin,

- Julien Jarroux, Alla L. Lapidus, and Hagen U. Tilgner. Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology* 2023 41:7, 41(7):915–918, 1 2023.
- [77] Roger Volden, Kayla Schimke, Ashley Byrne, Danilo Dubocanin, Matthew Adams, and Christopher Vollmers. Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion. *bioRxiv*, page 2022.06.29.498139, 6 2022.
- [78] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, 21(1):30, 2020.
- [79] Ying Chen, Nadia M Davidson, Yuk Kei Wan, Harshil Patel, Fei Yao, Hwee Meng Low, Christopher Hendra, Laura Watten, Andre Sim, Chelsea Sawyer, Viktoriia Iakovleva, Puay Leng Lee, Lixia Xin, Hui En Vanessa Ng, Jia Min Loo, Xuewen Ong, Hui Qi Amanda Ng, Jiaxu Wang, Wei Qian Casslynn Koh, Suk Yeah Polly Poon, Dominik Stanojevic, Hoang-Dai Tran, Kok Hao Edwin Lim, Shen Yon Toh, Philip Andrew Ewels, Huck-Hui Ng, N Gopalakrishna Iyer, Alexandre Thiery, Wee Joo Chng, Leilei Chen, Ramanuj DasGupta, Mile Sikic, Yun-Shen Chan, Boon Ooi Patrick Tan, Yue Wan, Wai Leong Tam, Qiang Yu, Chiea Chuan Khor, Torsten Wustefeld, Ploy N Pratanwanich, Michael I Love, Wee Siong Sho Goh, Sarah B Ng, Alicia Oshlack, Jonathan Goke, and SG-NEx consortium. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. *bioRxiv*, 2021.
- [80] Marcus J Viscardi and Joshua A Arribere. Poly(a) selection introduces bias and undue noise in direct RNA-sequencing. *BMC Genomics*, 23(1):530, 2022.

- [81] Michael R Green and Joseph Sambrook. Isolation of Poly(A)+ Messenger RNA Using Magnetic Oligo(dT) Beads. *Cold Spring Harbor Protocols*, 2019(10), 1 2019.
- [82] U Gubler and B J Hoffman. A simple and very efficient method for generating cDNA libraries. *Gene*, 25(2-3):263–269, 1 1983.
- [83] Mandy Jayne Peffers, Xuan Liu, and Peter David Clegg. Transcriptomic profiling of cartilage ageing. *Genomics Data*, 2:27–28, 12 2014.
- [84] Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Fran c. coise Thibaud-Nissen, Derek Albracht, Robert S Fulton, Milinn Kremitzki, Vince Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, Will Chow, Joanna Collins, Glenn Harden, Tim Hubbard, Sarah Pelan, Jared T Simpson, Glen Threadgold, James Torrance, Jonathan Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Heng Li, Chen-Shan Chin, Adam M Phillippy, Richard Durbin, Richard K Wilson, Paul Flicek, and Deanna M Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *bioRxiv*, 2016.
- [85] Zbigniew Dominski and William F. Marzluff. Formation of the 3’ end of histone mRNA: Getting closer to the end. *Gene*, 396(2):373, 7 2007.
- [86] Shimyn Slomovic, David Laufer, Dan Geiger, and Gadi Schuster. Polyadenylation of ribosomal RNA in human cells. *Nucleic Acids Research*, 34(10):2966–2975, 5 2006.
- [87] M Davari, B Amani, F Mokarian, M Hoseini, A Akbarzadeh, and N Heidarzadeh Khoramabadi. Effectiveness of trastuzumab as adjuvant therapy in patients with

- early stage breast cancer: A systematic review and meta-analysis. *Medical Journal of the Islamic Republic of Iran*, 31:88, 12 2017.
- [88] T W Nilsen and B R Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 1 2010.
- [89] E T Wang, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, S F Kingsmore, G P Schroth, and C B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 11 2008.
- [90] C Gutierrez and R Schiff. HER2: biology, detection, and clinical implications. *Archives of Pathology & Laboratory Medicine*, 135(1):55–62, 1 2011.
- [91] D C Allred, G M Clark, R Molina, A K Tandon, S J Schnitt, K W Gilchrist, C K Osborne, D C Tormey, and W L McGuire. Overexpression of HER-2/neu and its relationship with other prognostic factors change during the progression of in situ to invasive breast cancer. *Human pathology*, 23(9):974–979, 9 1992.
- [92] A M Namboodiri and J P Pandey. Differential inhibition of trastuzumab- and cetuximab-induced cytotoxicity of cancer cells by immunoglobulin G1 expressing different GM allotypes. *Clinical and experimental immunology*, 166(3):361–365, 12 2011.
- [93] E Y Lasfargues, W G Coutinho, and E S Redfield. Isolation of two human tumor epithelial cell lines from solid breast carcinomas. *Journal of the National Cancer Institute*, 61(4):967–978, 10 1978.
- [94] X Dai, H Cheng, Z Bai, and J Li. Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. *Journal of Cancer*, 8(16):3131–3141, 9 2017.

- [95] D L Holliday and V Speirs. Choosing the right cell line for breast cancer research. *Breast cancer research : BCR*, 13(4):215, 8 2011.
- [96] Ryan R Wick, Louise M Judd, Claire L Gorrie, and Kathryn E Holt. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial genomics*, 3(10):e000132, 9 2017.
- [97] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.
- [98] Diego Garrido-Martín, Emilio Palumbo, Roderic Guigó, and Alessandra Breschi. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLOS Computational Biology*, 14(8):e1006360, 8 2018.
- [99] Kristina Subik, Jin-Feng Lee, Laurie Baxter, Tamera Strzepek, Dawn Costello, Patti Crowley, Lianping Xing, Mien-Chie Hung, Thomas Bonfiglio, David G Hicks, and Ping Tang. The expression patterns of eR, pR, HeR2, cK5/6, eGFR, Ki-67 and AR by Immunohistochemical Analysis in Breast cancer cell Lines. *Breast Cancer: Basic and Clinical Research*, 4:4, 2010.
- [100] Grace A. Hall, Terence P. Speed, and Christopher J. Woodruff. Strain-level sample characterisation using long reads and MAPQ scores. *bioRxiv*, page 2020.10.18.344739, 10 2020.
- [101] Steve S. Sommer and Joel E. Cohen. The size distributions of proteins, mRNA, and nuclear RNA. *Journal of Molecular Evolution*, 15(1):37–57, 3 1980.

- [102] Jing Luo, Kai Xie, Xiang Gao, Yu Yao, Gaoming Wang, Chenye Shao, Xiaokun Li, Yang Xu, Binhui Ren, Liwen Hu, and Yi Shen. Long Noncoding RNA Nuclear Paraspeckle Assembly Transcript 1 Promotes Progression and Angiogenesis of Esophageal Squamous Cell Carcinoma Through miR-590-3p/MDM2 Axis. *Frontiers in Oncology*, 10:618930, 2 2021.
- [103] Tarek M.A. Abdel-Fatah, Robert C. Rees, A. Graham Pockley, Paul Moseley, Graham R. Ball, Stephen Y.T. Chan, Ian O. Ellis, Amanda K. Miles, Tarek M.A. Abdel-Fatah, Robert C. Rees, A. Graham Pockley, Paul Moseley, Graham R. Ball, Stephen Y.T. Chan, Ian O. Ellis, and Amanda K. Miles. The localization of pre mRNA splicing factor PRPF38B is a novel prognostic biomarker that may predict survival benefit of trastuzumab in patients with breast cancer overexpressing HER2. *Oncotarget*, 8(68):112245–112257, 11 2017.
- [104] Cristina Casalou, Andreia Ferreira, and Duarte C. Barral. The Role of ARF Family Proteins and Their Regulators and Effectors in Cancer Progression: A Therapeutic Perspective. *Frontiers in Cell and Developmental Biology*, 8:217, 4 2020.
- [105] Jin Wang, Xueting Yu, Xiyuan Cao, Lirong Tan, Beibei Jia, Rui Chen, and Jianxiang Li. GAPDH: A common housekeeping gene with an oncogenic role in pan-cancer. *Computational and Structural Biotechnology Journal*, 21:4056, 1 2023.
- [106] Takehiro Ichikawa, Fumiaki Sato, Kazuya Terasawa, Soken Tsuchiya, Masakazu Toi, Gozoh Tsujimoto, and Kazuharu Shimizu. Trastuzumab Produces Therapeutic Actions by Upregulating miR-26a and miR-30b in Breast Cancer Cells. *PLOS ONE*, 7(2):e31422, 2 2012.

- [107] Yuan Lu, Juan Bi, Fei Li, Gang Wang, Junjie Zhu, Jiqing Jin, and Yueyun Liu. Differential Gene Analysis of Trastuzumab in Breast Cancer Based on Network Pharmacology and Medical Images. *Frontiers in physiology*, 13, 7 2022.
- [108] Henrik Hoffmann and Cordelia Schiene-Fischer. Functional aspects of extracellular cyclophilins. *Biological Chemistry*, 395(7-8):721–735, 2014.
- [109] Surya Ramachandran, Anila Venugopal, V. R. Kutty, Vinitha A, Divya G, V. Chitrasree, Ajit Mullassari, N. S. Pratapchandran, K. R. Santosh, M. R. Pillai, and C. C. Kartha. Plasma level of cyclophilin A is increased in patients with type 2 diabetes mellitus and suggests presence of vascular disease. *Cardiovascular Diabetology*, 13(1), 2 2014.
- [110] Yuan Wei, Yan Jinchuan, Liang Yi, Wu Jun, Wang Zhongqun, and Wang Cuiping. Antiapoptotic and proapoptotic signaling of cyclophilin A in endothelial cells. *Inflammation*, 36(3):567–572, 6 2013.
- [111] Teruhisa Kazui, Norio Inoue, Osamu Yamada, and Sakuzo Komatsu. Selective cerebral perfusion during operation for aneurysms of the aortic arch: A reassessment. *The Annals of Thoracic Surgery*, 53(1):109–114, 1992.
- [112] Puja R. Mehta, Anna Leigh Brown, Michael E. Ward, and Pietro Fratta. The era of cryptic exons: implications for ALS-FTD. *Molecular Neurodegeneration*, 18(1):1–9, 12 2023.

APPENDICES

Appendix A

Supplementary Figures

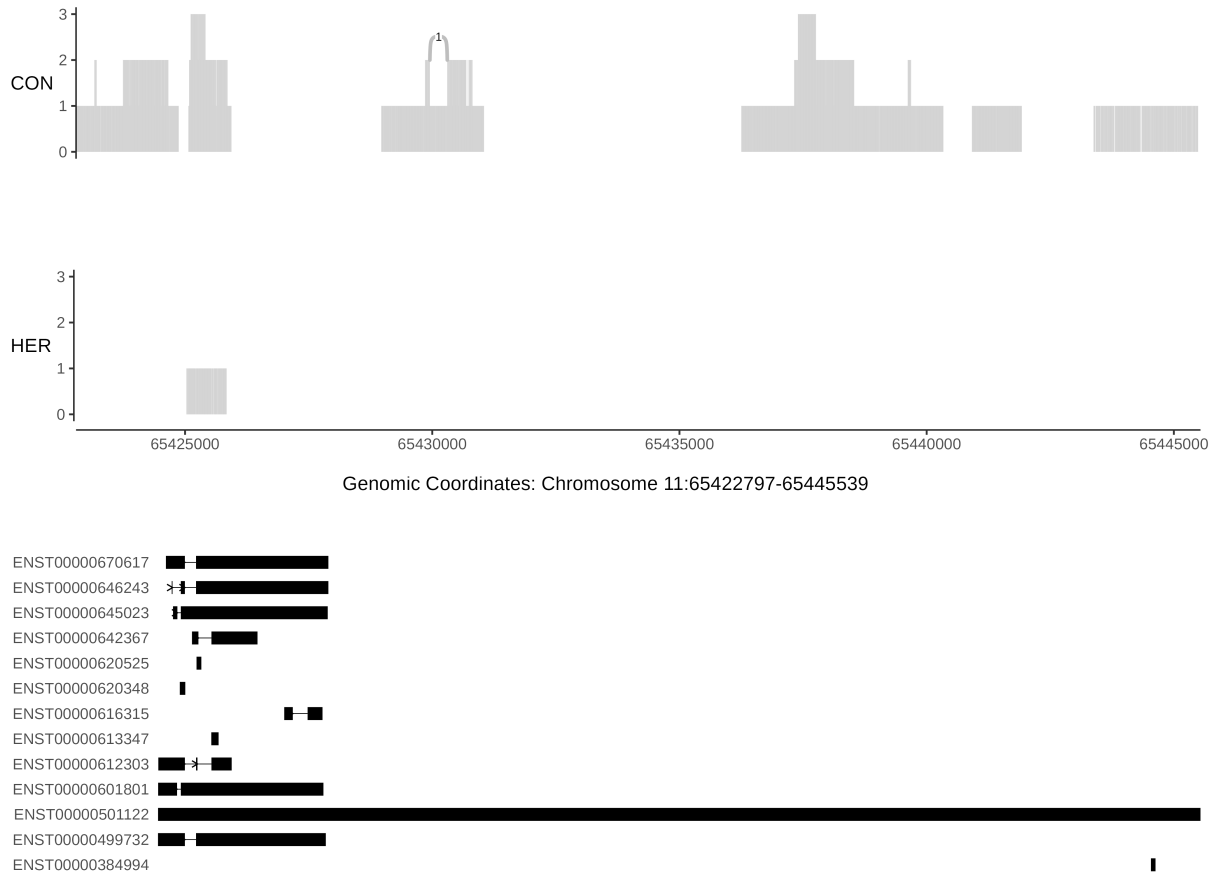


Figure A.1: **Sashimi plot of AS isoforms in the NEAT1 gene, cell line BT474.** The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

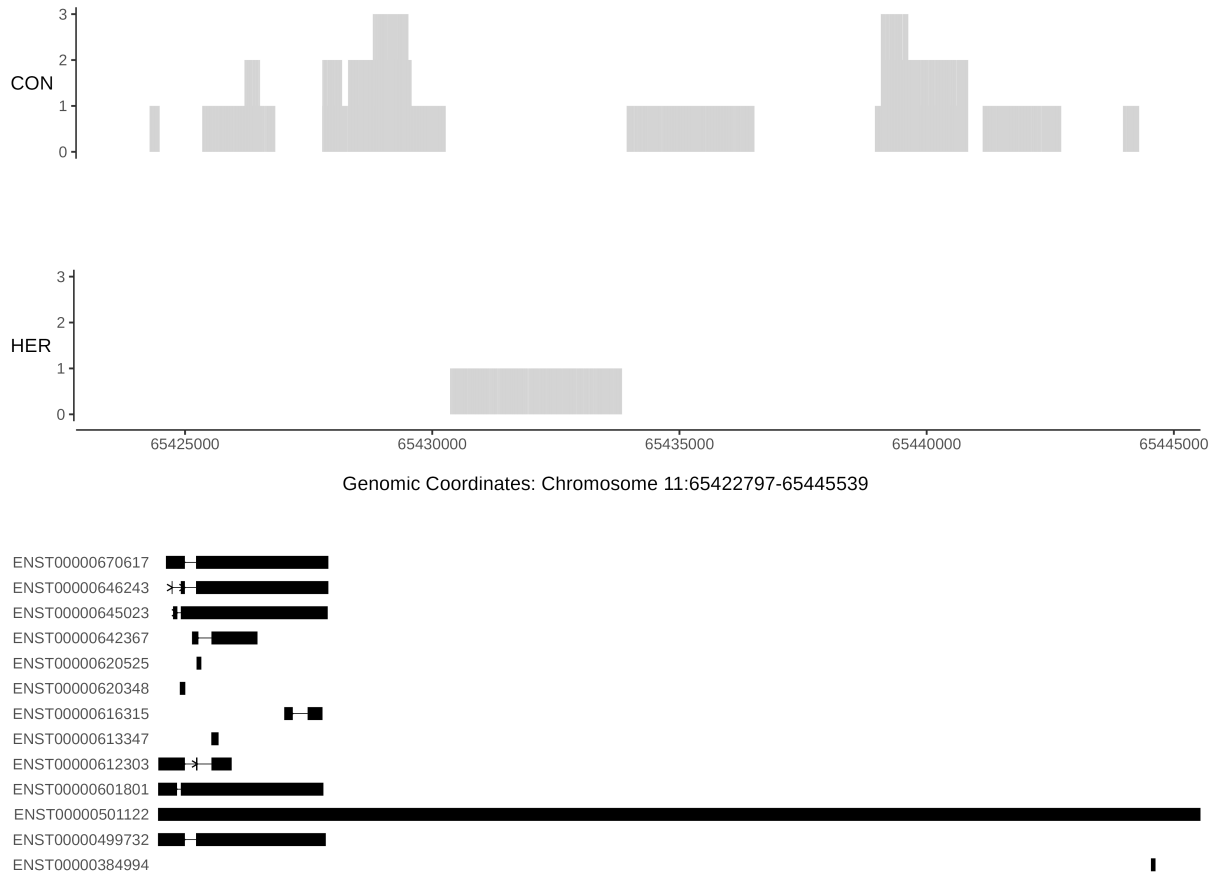


Figure A.2: Sashimi plot of AS isoforms in the NEAT1 gene, cell line MCF7. The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

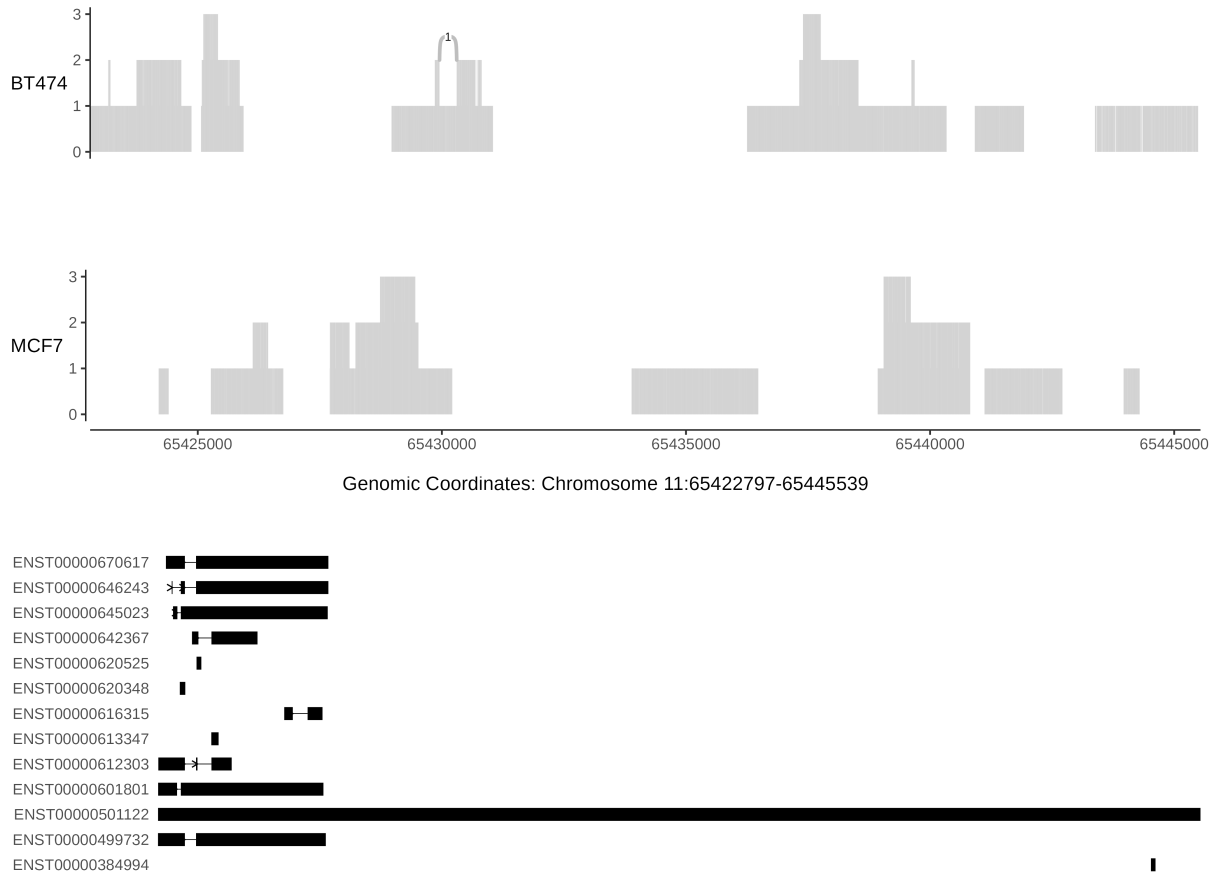


Figure A.3: Sashimi plot of AS isoforms in NEAT1 gene between cell lines BT474 and MCF7. The top plot shows the isoforms in cell line BT474 and abundances on the y axis. The middle plot shows the isoforms in cell line MCF7 and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

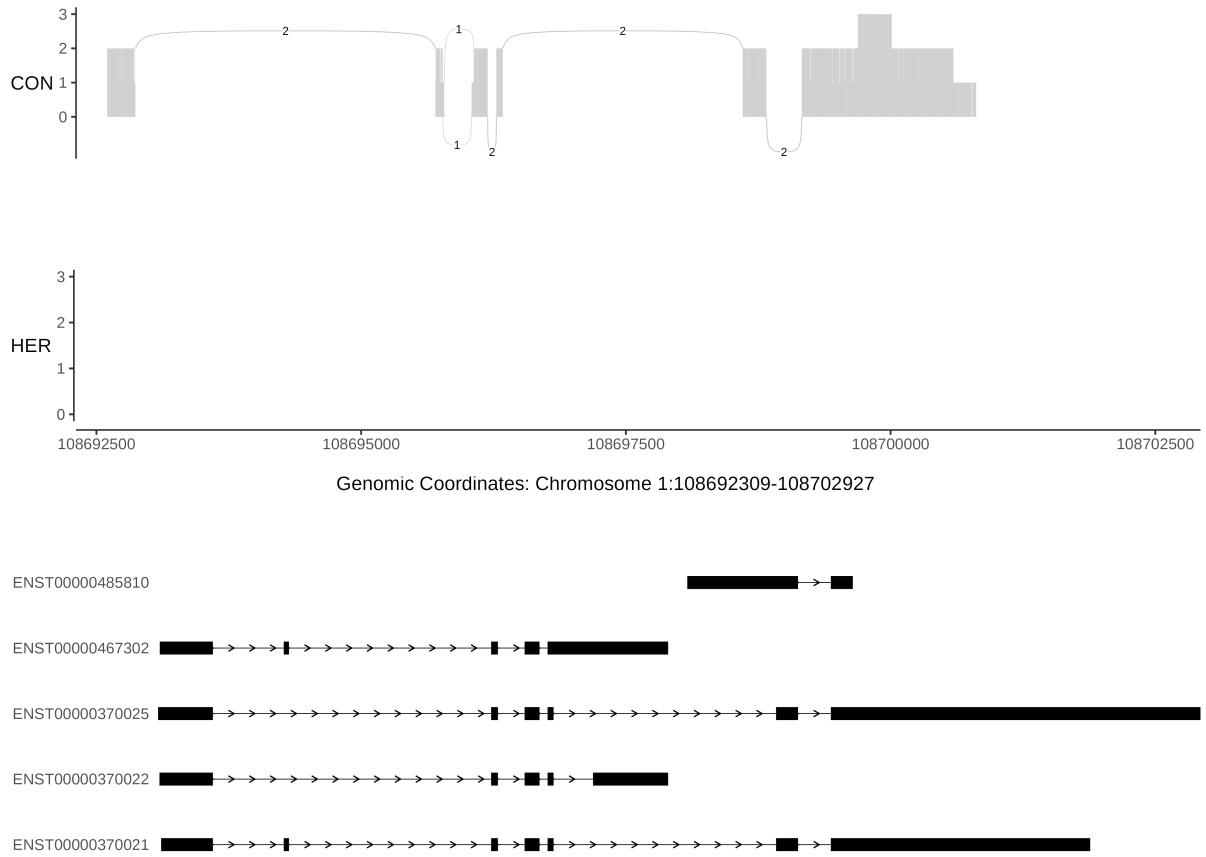


Figure A.4: **Sashimi plot of AS isoforms in the PRPF38B gene, cell line SKBR3.** The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

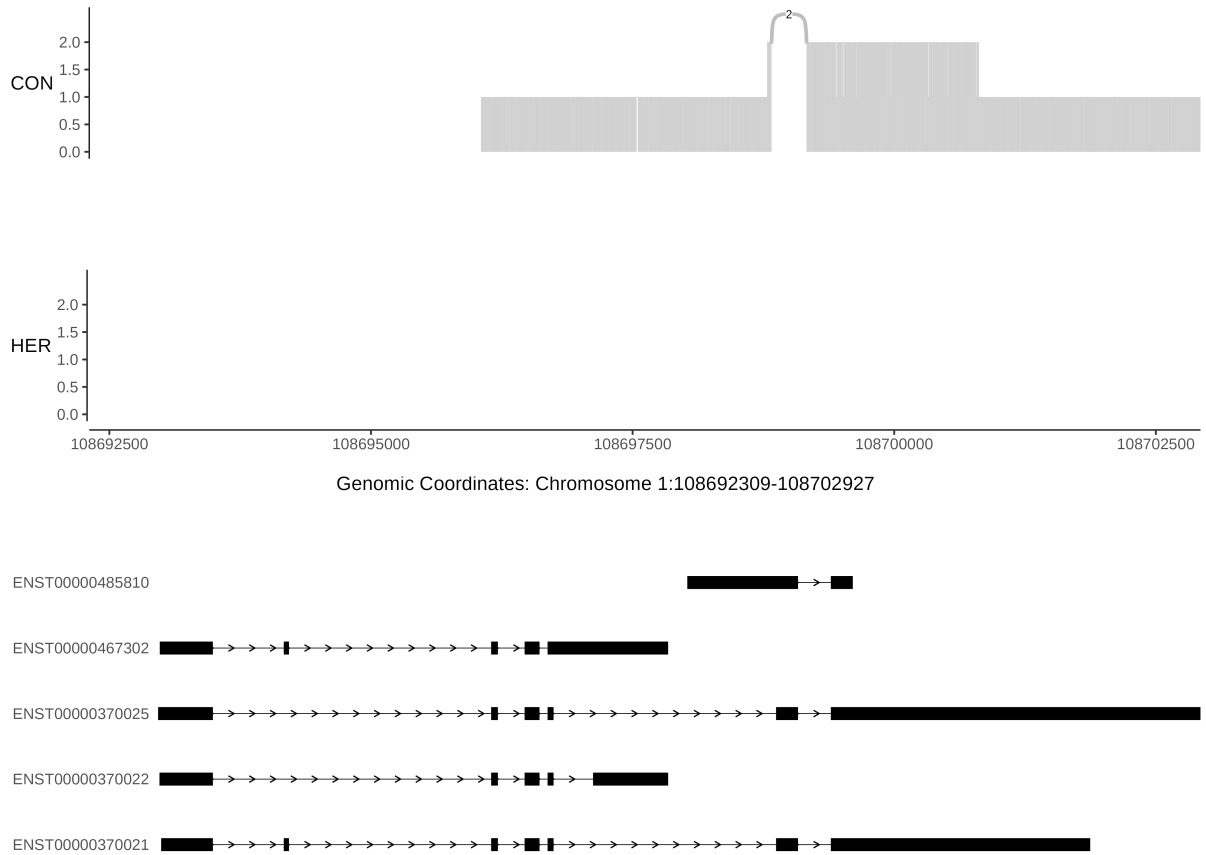


Figure A.5: **Sashimi plot of AS isoforms in the PRPF38B gene, cell line MCF7.** The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

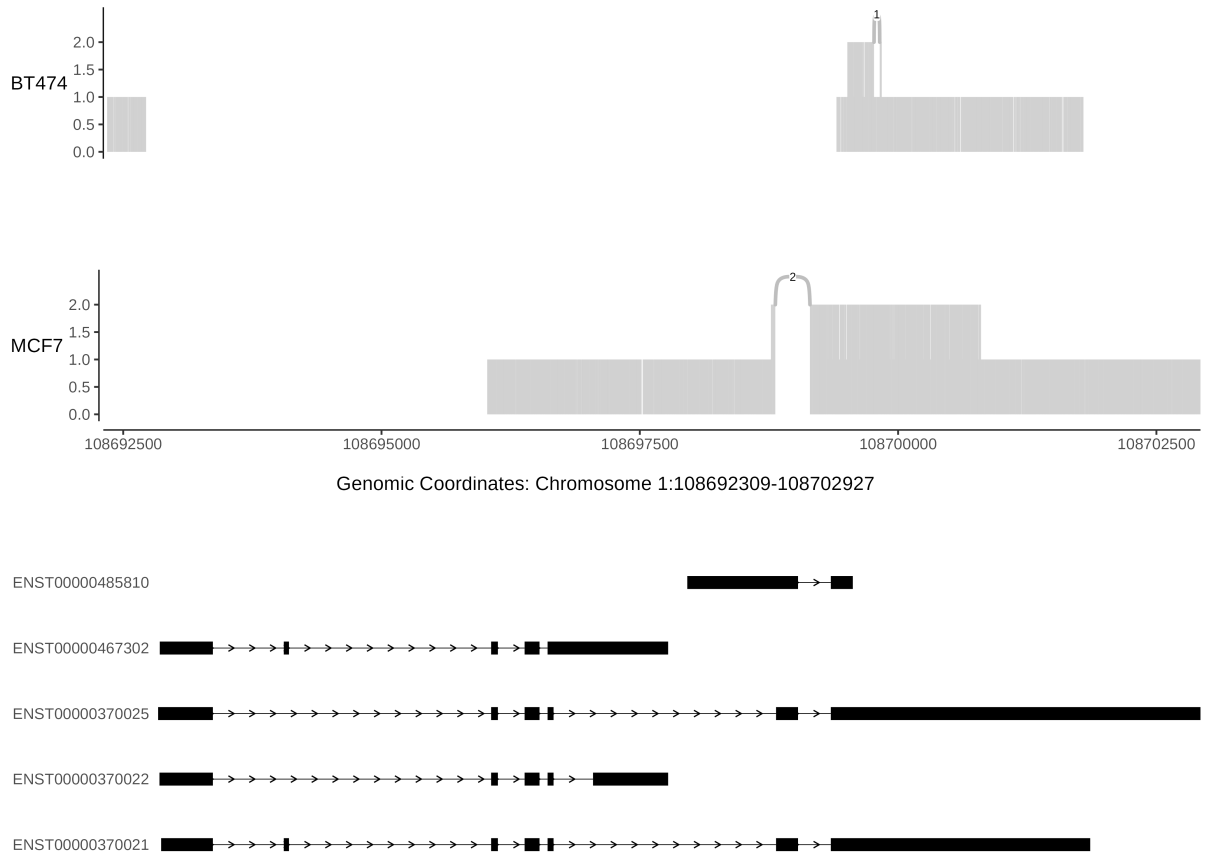


Figure A.6: **Sashimi plot of AS isoforms in PRPF38B gene between cell lines BT474 and MCF7.** The top plot shows the isoforms in cell line BT474 and abundances on the y axis. The middle plot shows the isoforms in cell line MCF7 and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

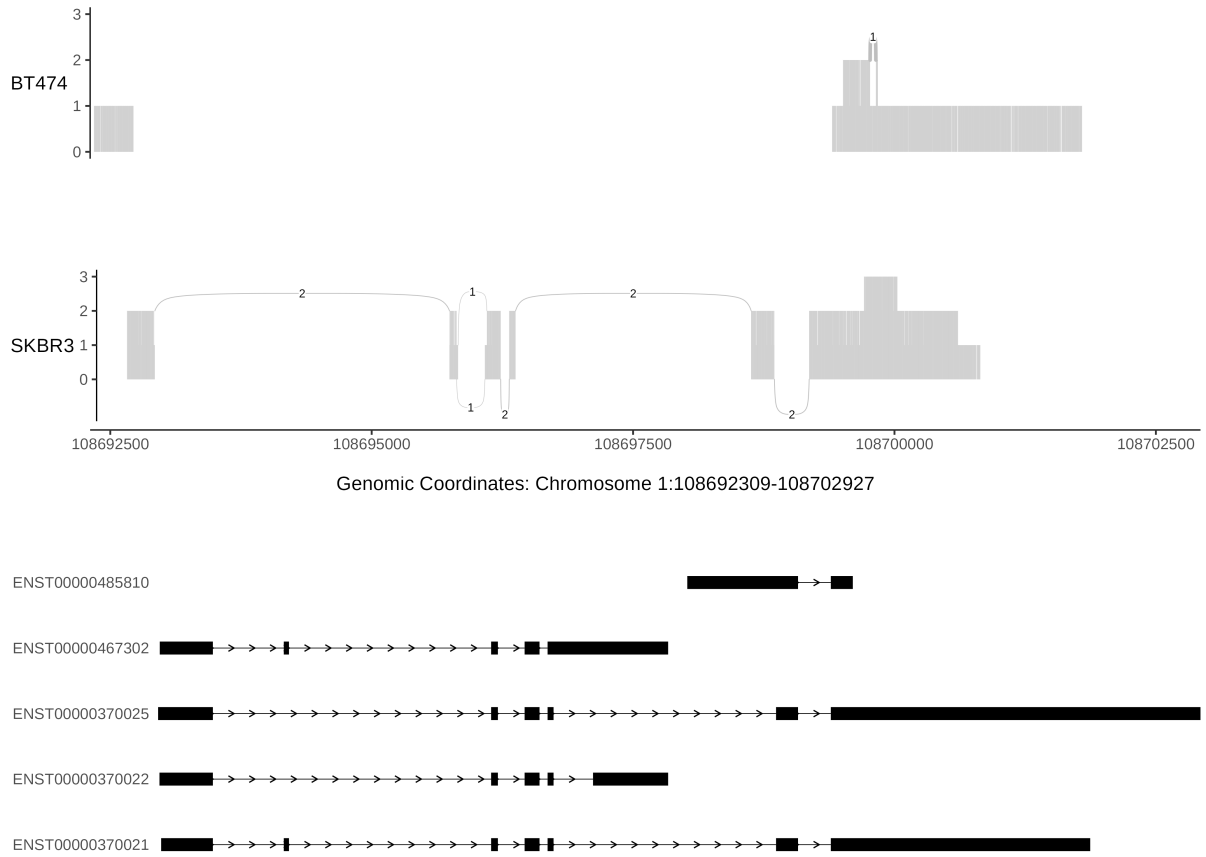


Figure A.7: Sashimi plot of AS isoforms in PRPF38B gene between cell lines BT474 and SKBR3. The top plot shows the isoforms in cell line BT474 and abundances on the y axis. The middle plot shows the isoforms in cell line SKBR3 and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

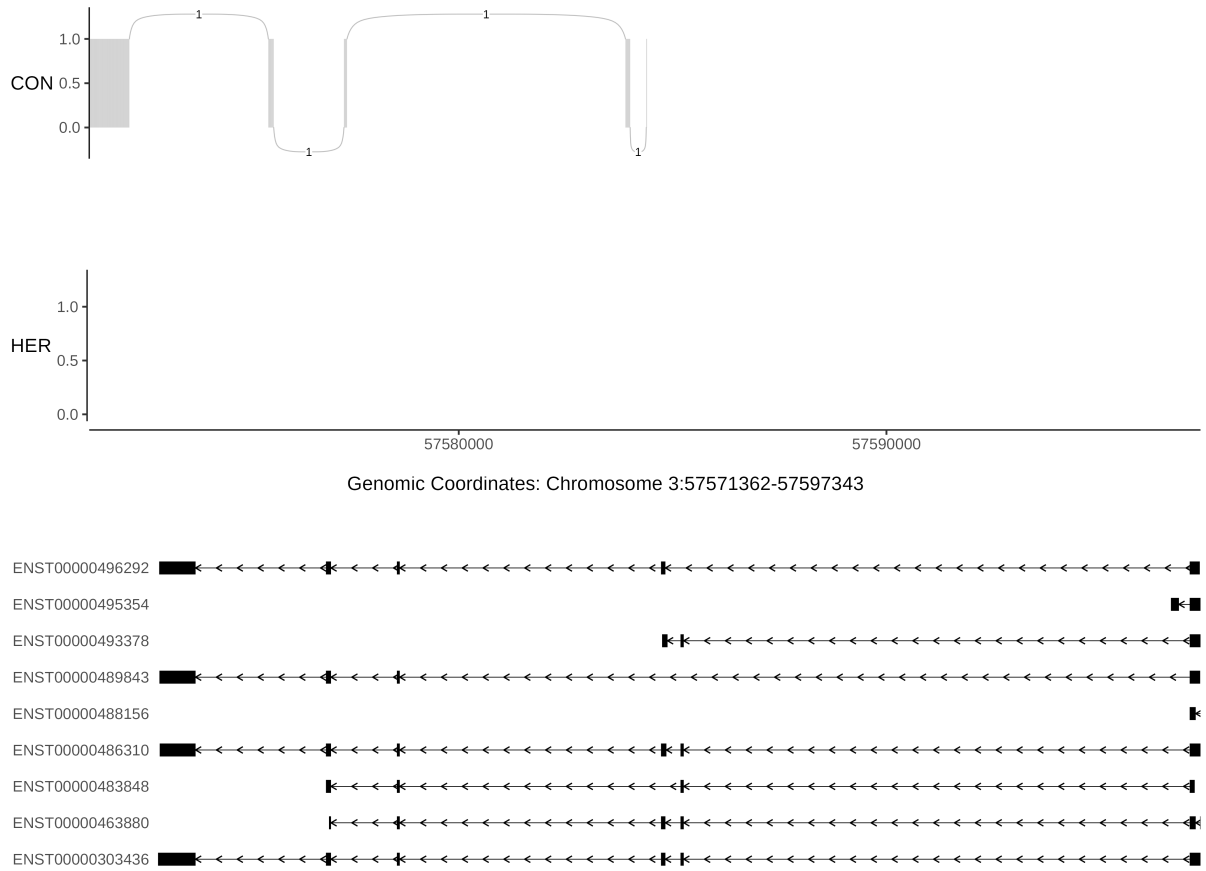


Figure A.8: **Sashimi plot of AS isoforms in the ARF4 gene, cell line BT474.** The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

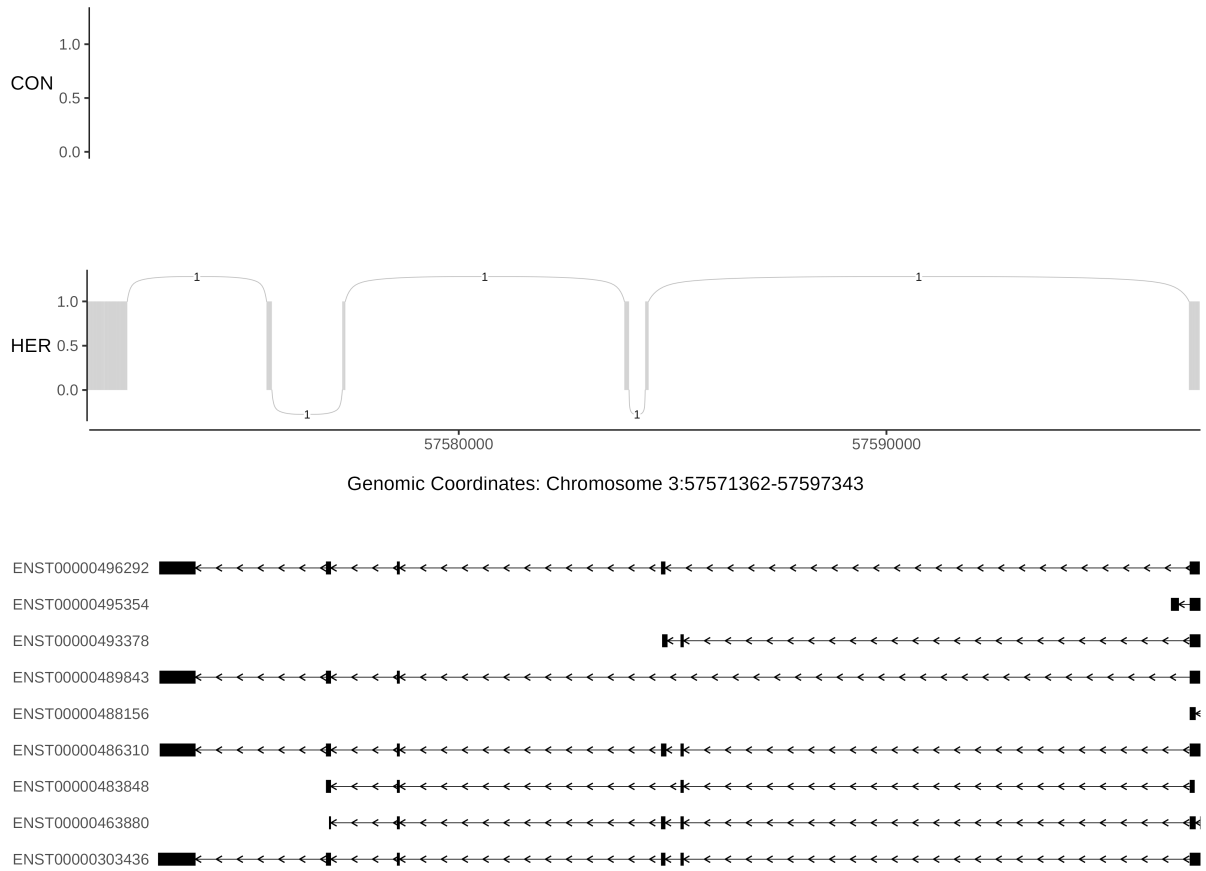


Figure A.9: **Sashimi plot of AS isoforms in the ARF4 gene, cell line SKBR3.** The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

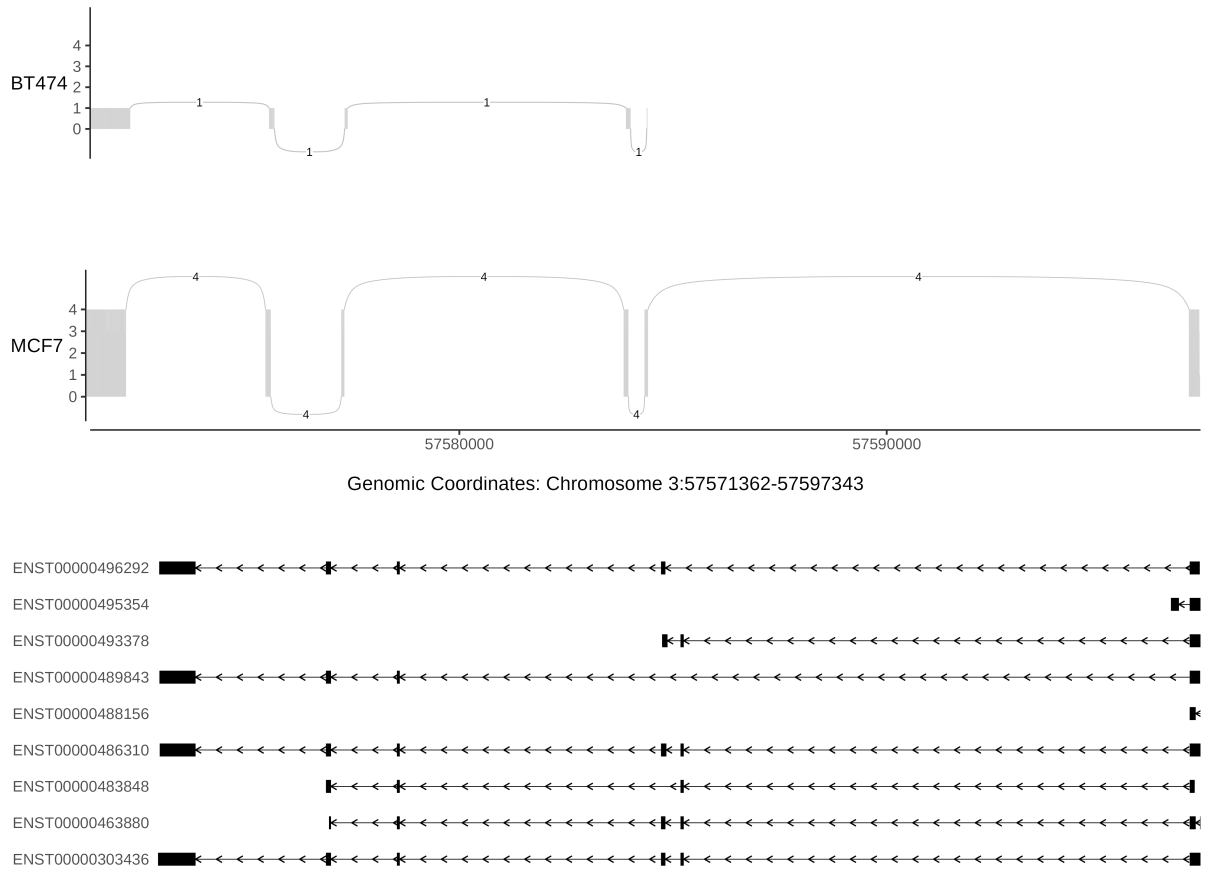


Figure A.10: Sashimi plot of AS isoforms in ARF4 gene between cell lines BT474 and MCF7. The top plot shows the isoforms in cell line BT474 and abundances on the y axis. The middle plot shows the isoforms in cell line MCF7 and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

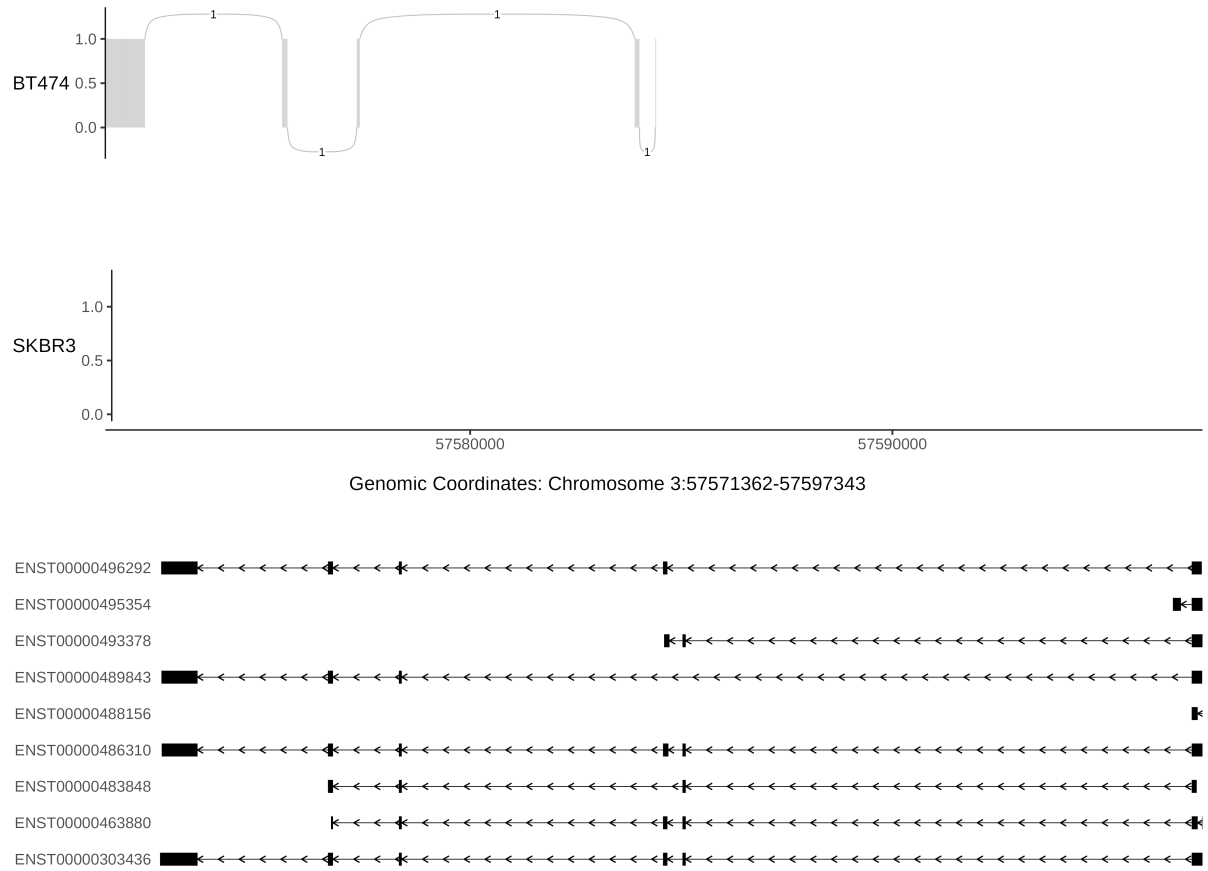


Figure A.11: **Sashimi plot of AS isoforms in ARF4 gene between cell lines BT474 and SKBR3.** The top plot shows the isoforms in cell line BT474 and abundances on the y axis. The middle plot shows the isoforms in cell line SKBR3 and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

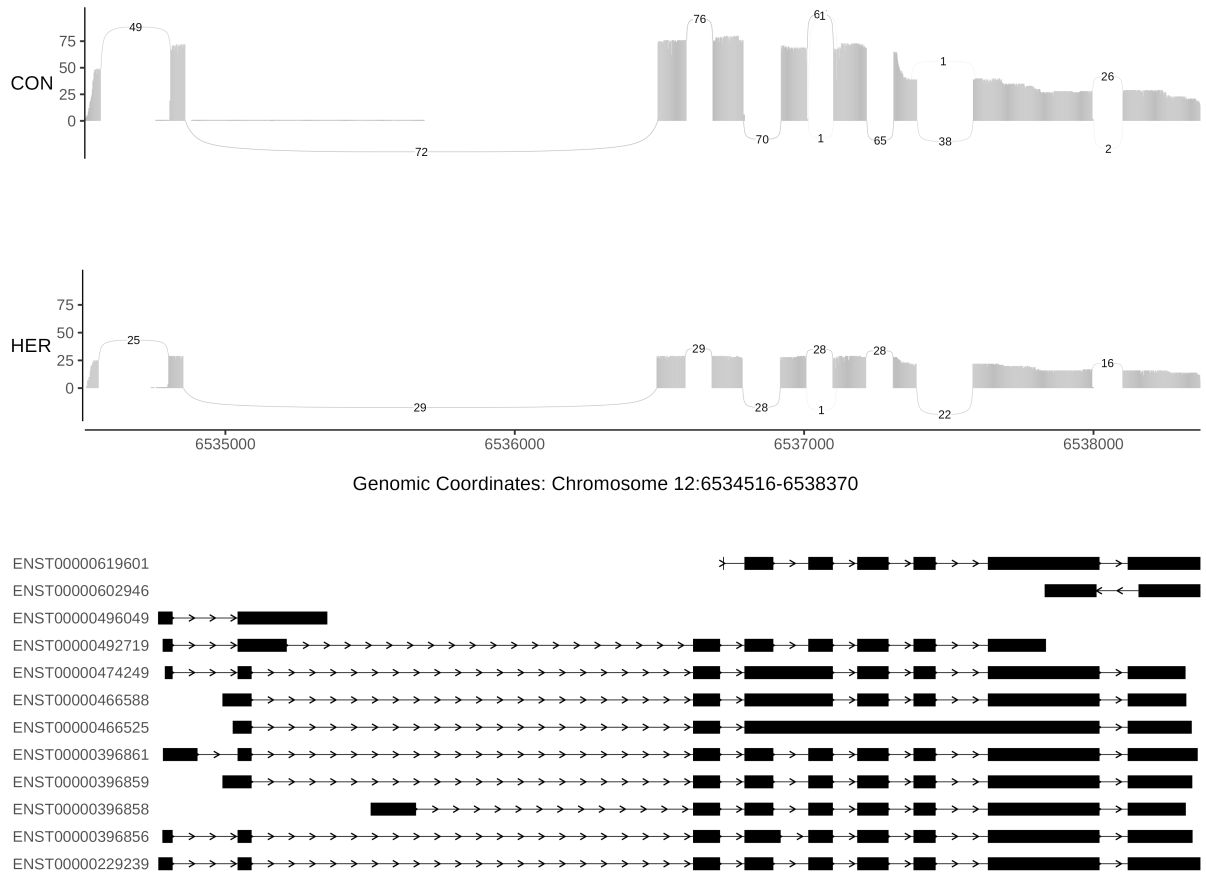


Figure A.12: Sashimi plot of AS isoforms in the GAPDH gene, cell line SKBR3. The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.

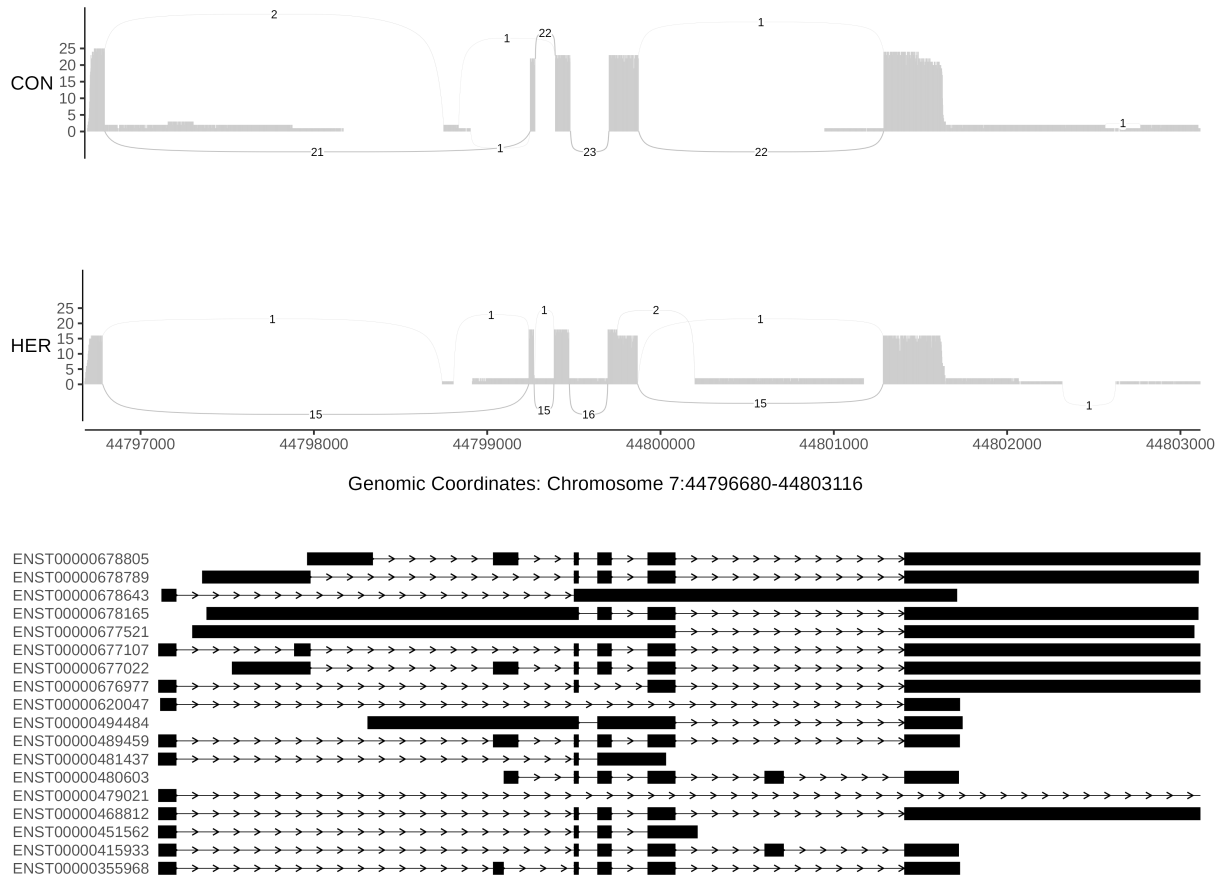


Figure A.13: Sashimi plot of AS isoforms in the PPIA gene, cell line BT474. The top plot shows the control isoforms and abundances on the y axis. The middle plot shows the isoforms from the Trastuzumab/Herceptin treated cells and abundances on the y axis. The bottom plot shows known isoform structures found in the Homo Sapiens GRCh38 gene annotation GTF file.