Màster Universitari
Anàlisi de Dades Òmiques / Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

RNA sequencing role in the genetic diagnosis of hereditary breast and ovarian cancer.

by

Sergio Manzano Sánchez

Supervisor: Lara Nonell, Bioinformatics Unit, Vall d'Hebron Institute of Oncology (VHIO).

Co-supervisor: Sara Gutiérrez- Enríquez, Hereditary Cancer Genetics Group, Vall d'Hebron Institute of Oncology (VHIO).

Academic tutor: Mireia Olivella García, Biosciences Department, University of Vic.

Biosciences Department

University of Vic – Central University of Catalonia

10/09/2024

RNA sequencing role in the genetic diagnosis of hereditary breast and ovarian cancer.

Sergio Manzano Sánchez[1,3], Irene Agustí Barea[1], Setareh Kompanian[2], Sara Gutiérrez-Enríquez[2], Lara Nonell[1]

[1]Bioinformatics Unit, Vall d´Hebron Institute of Oncology (VHIO), Barcelona, Spain. [2]Hereditary Cancer Genetics Group, Vall d´Hebron Institute of Oncology (VHIO), Barcelona, Spain. [3]Universitat de Vic – Universitat Central de Catalunya, Barcelona, Spain.

Abstract

Hereditary breast and ovarian cancer (HBOC) is a syndrome defined by an increased risk of developing breast cancer (BC) and/or ovarian cancer (OC), mainly due to germline pathogenic variants in *BRCA1* and *BRCA2*. While clinical genetic testing typically focuses on DNA coding regions, RNA sequencing (RNA-seq) offers a potential improvement in patient diagnosis the study of the expression of these regions. The goal of this study was to evaluate whether splicing and expression analysis of whole blood RNA, tested using Illumina sequencing, can provide a genetic diagnosis of HBOC in patients with high suspicion for the disorder, where conventional DNA testing has yielded inconclusive results. Consequently, the expression of 15 genes associated with HBOC was examined through RNA-seq in 45 patients who tested negative for *BRCA1* and *BRCA2* mutations. This study employed various bioinformatic approaches to assess the impact of transcriptomics on a cohort of HBOC patients, focusing on patient-specific alterations, using RNA from blood samples: Differential Expression Analysis (DEA), Outlier in RNA-seq Finder (OUTRIDER), Differential Exon Usage (DEU), SUPPA, and Find Rare Splicing Events in RNA-seq Data (FRASER). The results revealed 2 aberrantly expressed genes detected by OUTRIDER, 24 local splicing events identified by SUPPA, and 14 significant findings discovered by FRASER. After performing this study, we can conclude that bioinformatics approaches, such as transcriptomics, are powerful but require laboratory validation. This initial exploration should be followed by additional computational and experimental analyses.

Code Availability: https://github.com/SergioManzano10/RNA-sequencing_FMP

Contact: sergioprof2022@gmail.com

Supplementary information: Supplementary_Materials.pdf

1 INTRODUCTION

Hereditary breast and ovarian cancer (HBOC) is a syndrome defined by an increased risk of developing breast cancer (BC) and/or ovarian cancer (OC) due to inherited germline pathogenic variants in various genes that confer a range of moderate to high cancer risks (Marmolejo et al., 2021). Global estimates indicate that BC is the most common female cancer diagnosis, accounting for approximately 24.8% of all female cancer diagnoses worldwide. On the other hand, OC accounts for approximately 4.4% of female cancer diagnoses (McDevitt et al., 2024). Approximately a 10% of breast and ovarian cancers result from hereditary causes (Beitsch et al., 2019).

HBOC is primarily caused by germline pathogenic variants in *BRCA1* and *BRCA2*, which confer a high risk of cancer predisposition. The *BRCA1* and *BRCA2* (*BRCA1/2*) are tumour suppressor genes that encode for proteins involved in DNA double strand break repair by homologous recombination (HR), one of the critical maintenance mechanisms of DNA integrity. To complete this function, the BRCA1/2 proteins interact with a host of other molecules which together form a protein complex. Without a functional BRCA complex, the cell relies on alternative mechanisms for DNA repair, some of which are error prone and may further contribute to the development of genetic aberrations. Because of this phenomenon, HBOC patients with germline *BRCA1* and *BRCA2* pathogenic variants have an increased risk for the development of several neoplasms, particularly those arising in the breast as well as ovary (Hodgson & Turashvili, 2020).

Since the discovery of these genes in the early 1990s, *BRCA1/2* genetic testing has been offered to an increasing number of HBOC patients and families due to its proven clinical benefits (Montalban et al., 2021), such as medical management based on personalized cancer treatments, cancer surveillance and prevention programs of patients and families (González-Santiago et al., 2020). However, only a proportion of HBOC families are explained by deleterious variants in these two genes (Montalban et al., 2021). Pathogenic variants in other genes such as *PALB2, TP53, CDH1 and PTEN* are also associated with a high risk of BC (Dorling et al., 2021; Hu et al., 2021), while those in *ATM, BARD1* ,*CHEK2*, *RAD51C*, *RAD51D* are BC risk genes associated with a moderate risk (Dorling et al., 2021; Hu et al., 2021). Additionally, *RAD51C*, *RAD51D* and *BRIP1* together with those in Lynch syndrome genes (*MLH1, MSH2, MSH6*), contribute to an increased risk of OC (Bellcross, 2022).

Due to multigene causality, the massive sequencing of customized gene panels is now widely used in clinical genetic diagnosis. This approach allows the simultaneous analysis of high- and moderate-risk susceptibility genes in a single test, increasing the likelihood of detecting cancer-predisposing variants (Piccinin et al., 2019)

Current clinical practices for genetic testing are commonly focused on the study of DNA coding regions, whereas the existence and significance of genetic variants in non-coding regions including promoters, untranslated regions (UTRs) and deep intronic regions remains largely unexplored. Consistent with this, it has been demonstrated that DNA changes in both coding and non-coding regions can affect gene expression. These alterations can be measured at the transcriptomic level through RNA sequencing (RNA-seq) analyses (PCAWG Transcriptome Core Group et al., 2020). In the same line, it has been documented that the use of RNA-seq, which detects both aberrant gene expression and splicing alterations, has systematically obtained diagnostic rates increased by 8-36% over DNA sequencing alone for a variety of disorders (Yépez et al., 2021). Therefore, RNA analysis can allow the identification of alterations that could go undetected by conventional testing, which is relevant for high-risk families with negative DNA results (Montalban et al., 2021).

Focusing on RNA-seq, different types of analyses can be performed: 1) **Differential gene expression**: compares the expression of the genes between different conditions in order to obtain the differentially expressed genes (DEGs) (Koch et al., 2018). 2) **Outlier in RNA-seq finder**: identifies outliers in RNA-seq data by comparing individual values to the rest of the dataset, irrespective of specific experimental conditions (Brechtmann et al., 2018). 3) **Alternative splicing**: considering the exon boundaries, splicing events are detected and compared to identify those that are differentially spliced between a pair of conditions (Trincado et al., 2018). 4) **Rare splicing events**: finds aberrant splicing events in RNA-seq samples, from exon skipping over alternative donor usage to intron retention (Mertes et al., 2021).

Considering the different analyses that can be performed with RNA-seq data shown above, it is evident that this next-generation sequencing (NGS) technology also plays a critical role in the study of pre-mRNA splicing. Removal of introns from messenger RNA precursors (pre-mRNA splicing) is an essential step for the expression of most eukaryotic genes. However, disruption of splicing can lead to the generation of cancer cells with distinct splicing patterns. Thus, the monitoring of splicing alterations can provide effective biomarkers for use in the diagnosis, prognostication and monitoring of patients with cancer (Bonnal et al., 2020). Moreover, it has been demonstrated that approximately 15-30% of the variants responsible for inherited diseases impact alternative splicing (Mertes et al., 2021), which may serve as a hallmark in tumorigenesis.

To investigate whether RNA sequencing enhances the diagnostic rate of pathogenic variants in families at high risk for hereditary breast and ovarian cancer syndrome who tested negative on conventional targeted DNA panel testing, whole blood RNA was massively sequenced using Illumina technology from 53 individuals. The study included 45 index cases from high-risk Spanish families with breast and/or ovarian cancer in whom no pathogenic variants were detected by DNA panel testing of coding regions. The study also included 2 positive cases carrying previously characterised RNA alterations Finally, the study also included 6 healthy controls.

Bioinformatic analyses were designed to include software packages to detect impact on RNA splicing and RNA abundance in the gene panel related to HBOC risk, which is typically screened in routine DNA testing: *BRCA1, BRCA2, MLH1, MSH2, MSH6, TP53, PALB2, CHEK2, ATM, BRIP1, BARD1, RAD51C, RAD51D, PTEN* and *CDH1*.

Different analyses were conducted using two pipelines: 1) The **nf-core/rnaseq** pipeline was used to perform a differential gene expression analysis (DEA). Additionally, with the same input data, an aberrant gene expression study was conducted using the OUTRIDER package in R studio. 2) The **nf-core/rnasplice** pipeline was utilized to analyze differential exon usage (DEU) and local events (SUPPA). Moreover, the .bam files generated by this pipeline were employed to study rare splicing events with the FRASER method.

2 METHODS

Data description

Data from 47 RNA samples (isolated from whole blood), obtained and sequenced from index cases from breast and/or ovarian cancer high-risk Spanish families ascertained between years 2008-2018 through the Unit of Familial Cancer Unit at Vall d'Hebron Hospital in Barcelona, were analyzed. These index cases were the best patients to be analyzed in each family according to established selection clinical (González-Santiago et al., 2020) criteria. Of these index cases, 46 were suffering from cancer and 1 was healthy (but belonging to a family suffering from cancer). In addition, there were samples corresponding to 6 healthy individuals that were classified as controls, who were unrelated to the individuals with cancer.

Families with hereditary breast/ovarian cancer were recruited following clinical eligibility criteria for DNA testing of the related HBOC panel gene, including: *BRCA1, BRCA2, MLH1, MSH2, MSH6, TP53, PALB2, CHEK2, ATM, BRIP1, BARD1, RAD51C, RAD51D, PTEN* and *CDH1* genes. All cases were screened for single nucleotide variants in coding exons and exon-intron boundaries (by Sanger or massively sequencing) and large rearrangements (by Multiplex Ligation-dependent Probe Amplification, MLPA) and no pathogenic variant was identified.

Additionally, data from the GTEx (Genotype-Tissue Expression) database (Lonsdale et al., 2013), which provides information on gene expression across a wide variety of tissues, were used to obtain median TPM (Transcripts Per Million) values for 15 genes of interest. These median TPM values from GTEx were then used to correlate with the TPM values of the genes of interest in our original dataset.

RNA-seq analysis

Raw sequencing reads (50X paired-end sequencing) in the FASTQ files were processed through the nf-core/rnaseq pipeline (Ewels et al., 2020) version 3.12.0 with default parameters and using the GRCh38 genome from NCBI. This pipeline includes trimming, alignment, and quantification steps, which are essential for obtaining gene counts. These counts were then used to perform differential expression analysis (DEA).

Prior to performing the DEA, we conducted different preprocessing steps. To do that, the library size differences of gene raw counts were normalized using the Trimmed Mean of M-values (TMM) method (Robinson & Oshlack, 2010) implemented in the edgeR package version 3.42.4 (Robinson et al., 2010; McCarthy et al., 2012; Chen et al., 2024) using RStudio. The normalized counts were used in the unsupervised analysis: hierarchical clustering of samples (using different methods), and principal component analysis (PCA).

The DEA analysis was assessed following the voom + limma approach from limma package version 3.56.2 (Ritchie et al., 2015). First, gene raw counts were filtered to include those with at least 10 reads in at least 6 samples, matching the smallest group size (the controls one). The filtered counts were transformed using *voom() function*. Following this, a linear model was created using *lmFit() function* and finally the DEA analysis was done using *eBayes() function*. Correction for multiple comparisons was performed using false discovery rate (FDR) (Benjamini & Hochberg, 1995) to obtain the adjusted p-values. Genes were differentially expressed between the two conditions if the adjusted p-value was < 0.05 and the $|logFC| > 1$. Also, a GSEA was performed using the C6 oncogenic signature gene sets from MSigDB. This analysis was conducted using the clusterProfiler R package version 4.10.0 by applying the *GSEA() function*.

OUTRIDER (OUTlier in RNA-Seq fInDER) analysis

The output corresponding to the gene counts matrix generated through the nf-core/rnaseq pipeline was used to create the OutriderDataSet object, needed to perform the OUTRIDER analysis using the OUTRIDER R package version 1.20.0 (Brechtmann et al., 2018) by applying the *OutriderDataSet() function*. Then, the low-expressed genes were filtered out using the *filterExpression() function*. Later, the remaining genes that were not filtered out were used to run the full OUTRIDER pipeline with the *OUTRIDER() function*. Finally, the significant results (adjusted p-value < 0.05) were retrieved using *results() function.*

RNA splice analysis

Raw sequencing reads in the FASTQ files were processed through the nf-core/rnasplice pipeline (Ewels et al., 2020) version 1.0.1 with default parameters and using the GRCh38 release 110 from Ensembl. Before generating the objects used for analysis, the pipeline performs preliminary steps such as trimming and alignment.

To perform the differential exon usage (DEU) analysis we used the *.rds* objects (exon raw counts) for each patient (created after comparing each patient vs the 6 controls) obtained directly from the nf-core/rnasplice pipeline, that internally utilizes the DEXSeq R package version 1.48.0 (Anders et al., 2012; Reyes et al., 2013). Initially, this package generated a DEXSeqDataSet object containing the counts per exon. This object was normalized using the *estimateSizeFactors() function* from DEXSeq package. Subsequently, dispersion was estimated with the *estimateDispersions() function*. Finally, the test for differential exon usage was conducted using the *testForDEU() function* and the results were extracted with the *DEXSeqResults() function*. From these results, the genes of interest were selected and plotted using the *plotDEXSeq() function.*

To perform the local events analysis (SUPPA), we used the *.dpsi* files for each patient (created by comparing each patient to the 6 controls) obtained directly from the nf-core/rnasplice pipeline, that internally uses the SUPPA tool (Alamancos et al., 2015; Trincado et al., 2018). The mentioned tool works with *Python scripts* and uses a command/subcommand structure: *suppa.py subcommand options* where the subcommand can be one of these five: *generateEvents* (generates events from an annotation), *psiPerEvent* (quantifies event inclusion levels (PSIs) from multiple samples), ***psiPerIsoform*** (quantifies isoform inclusion levels (PSIs) from multiple samples), *diffSplice* (calculate differential splicing across multiple conditions with replicates) and *clusterEvents* (cluster events according to PSI values across conditions). For our purpose, we used the previously mentioned *.dpsi* files that contained: gene name, local event, chromosome, event positions, ΔPSI, and p-value. From these files, we selected the local events of the genes of interest with a p-value < 0.05.

FRASER (Find RAre Splicing Events in RNA-seq Data) analysis

The *.bam* files generated through the nf-core/rnasplice pipeline were used to create the FraserDataSet object using the *FraserDataSet() function*, necessary for performing the FRASER analysis (Mertes et al., 2021) using the FRASER R package version 1.14.1. Then, the counting reads (raw data) were extracted from this object using the *countRNAData() function.* Following that, the PSI values were computed using the *calculatePSIValues() function* and junctions with low expression were filtered out using the *filterExpressionAndVariability() function,* selecting those with a minimum of 20 counts in one sample and a minimum ΔPSI of 0.1. Later, the splicing model was fitted using the *optimHyperParams()* and *bestQ() functions.* Subsequently, the full FRASER pipeline was ran using the *FRASER() function* and introns were annotated using *annotateRanges()* and *rowRanges() function*. Finally, significant results were obtained by applying the *results() function* with an adjusted p-value cutoff of 0.05 and a ΔPSI cutoff of 0.1.

Processing details

The execution of both nf-core pipelines was performed with Nextflow v23.10.0 (Di Tommaso et al., 2017) in a high-performance cluster (HPC) using Slurm as the task manager.

For most analyses, R Studio version 4.3.2 was used. Additionally, the ggplot2 package version 3.4.3 was used to create the plots.

Moreover, the results were manually visualized using IGV v2.16.0 with the reference genome GRCh38 (Thorvaldsdottir et al., 2013).

In addition, it is important to consider that **TPM** values have been used in some of the analyses. TPM is a metric used in gene expression analysis to allow the comparison between different samples (Zhao et al., 2020). The formula for calculating TPM is:

$$TPM = \frac{\text{reads mapped to transcript/transcript length}}{\text{Sum(reads mapped to transcript/transcript length)}} \times 10^6$$

Furthermore, **coverage** is also used by IGV for data visualization. It is defined as the number of times a specific region of the genome has been read during sequencing. The formula for calculating coverage is:

$$Coverage = \frac{\text{read length} \times number\ of\ reads}{\text{genome length}}$$

3 RESULTS

In Figure 1, the workflow followed in this study is outlined. We started by processing the FASTQ files (obtained from the bulk RNA-Seq data) through the nf-core rnaseq and rnasplice pipelines, which gave different outputs. Regarding the rnaseq pipeline, the TPM values for the 15 genes of interest were used to perform gene correlations with those from GTEx (Genotype-Tissue Expression). The gene raw counts were used to perform PCA, hierarchical clustering and heatmap after TMM normalization. Additionally, low-count genes were filtered out from the gene raw counts, and a differential expression analysis was performed on the remaining ones using the voom + limma approach. Moreover, gene raw counts were used to perform the OUTRIDER analysis.

On the other hand, we applied the rnasplice pipeline. In this case we retrieved .rds objects that contained the counts per exon for each gene, which were subjected to different functions before performing the final DEU analysis. In addition, we had .dpsi objects containing all the necessary information to perform the SUPPA analysis. The results of both analyses were visualized in IGV. Furthermore, the .bam files obtained after the alignment were used to conduct the FRASER analysis.

As we have mentioned previously, there are some genes that are usually analyzed when dealing with individuals suffering from HBOC. For this reason, all the analyses performed in this study will be focused on these 15 genes of interest: *BRCA1, BRCA2, MLH1, MSH2, MSH6, TP53, PALB2, CHEK2, ATM, BRIP1, BARD1, RAD51C, RAD51D, PTEN and CDH1.*

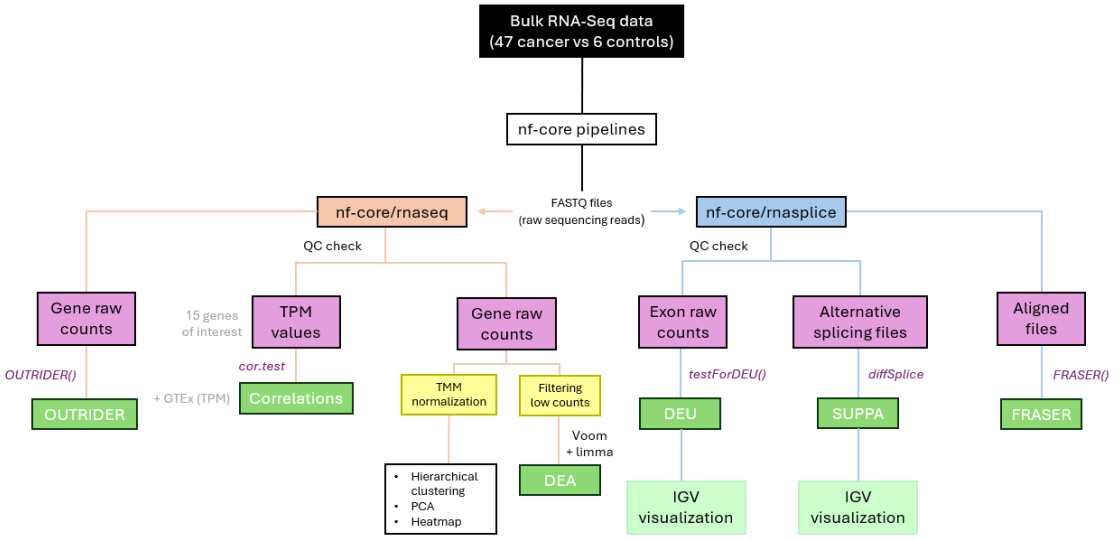

***Figure 1:* Diagram of the workflow followed in this study.** First, the FASTQ files from the bulk RNA-Seq data were processed through the nf-core pipelines, which generated a QC report and various outputs. After confirming that the QC report was satisfactory in both cases, we proceeded to the next steps. For the rnaseq pipeline: TPM values for the genes of interest were subjected to a correlation test with data from GTEx. Raw gene counts were normalized for hierarchical clustering, PCA, and heatmap analyses. They were also filtered for differential expression analysis (DEA) using the voom + limma approach. Additionally, raw gene counts were used to perform OUTRIDER analysis. Regarding the rnasplice pipeline: The exon raw counts (.rds objects) were used to perform DEU analysis, while the alternative splicing files (.dpsi objects) were used for SUPPA analysis. The results were visualized using the IGV tool. The aligned files (.bam files) were used to perform FRASER analysis. QC: Quality Control, TPM: Transcript Per Million, GTEx: Genotype-Tissue Expression, TMM: Trimmed Mean of M-values, OUTRIDER: Outlier in RNA-Seq Finder, DEA: Differential Expression Analysis, DEU: Differential Exon Usage, FRASER: Find Rare Splicing Events in RNA-seq Data, IGV: Integrative Genomics Viewer.

Quality control (QC)

The nf-core/rnaseq pipeline, which is used as a standard approach to perform the RNA-seq analyses at the first stages produces a report, using MultiQC (Ewels et al., 2020), encompassing several quality measures. We focused on the following: 1) Sequence quality histograms: We saw that the mean quality value across each base position in the reads indicated high-quality data (Supplementary Figure 1). 2) Percentage of duplications: we observed a high presence of duplications in some samples, but we did not remove them because without the use of UMIs (Unique Molecular Identifier), the pipeline is not able to differentiate technical duplications from those caused by high expression (Supplementary Table 1).

Since the QC metrics met the required standards, we decided to continue with the analysis by examining the TPM values. To have a summary across all samples, median TPM values (TPM VHIO) were calculated for the genes of interest. Some of them like *PTEN*, *ATM* or *TP53* had high TPM values compared to other genes, especially compared to *BRCA2*, *CDH1* and *BRIP1* (Figure 2A). With the objective of determining whether this pattern of expression was followed in healthy population of other studies we performed a correlation test using the median TPM expression values of the 15 selected genes obtained from the GTEx portal in whole blood (TPM GTEx), resulting in a R^2 = 0.98 (Figure 2B). This result demonstrated a high positive correlation between the two data sources, indicating a high-quality of our data compared to reference data from GTEx.

Moreover, the coverage for each sample was visually checked using the IGV programme. As an additional step in quality control, we analyzed the distribution of the data. For this, we used hierarchical clustering and PCA; we observed that samples did not separate based on condition (cancer or control) and appeared mixed, as shown in Figure 3A and Supplementary Figure 2. This could indicate that the comparison between cancer cases and control samples may not be as decisive as expected.

A

Table 1: TPM expression values

Gene ID	Median TPM (VHIO)	Median TPM (GTEx)
PTEN	194.66	40.27
ATM	46.849	3.5
TP53	22.886	7.69
MLH1	11.576	6.61
MSH6	8.697	2.32
MSH2	6.600	1.35
RAD51C	6.205	0.67
RAD51D	5.086	1.072
CHEK2	3.482	1.25
PALB2	2.555	0.84
BRCA1	2.450	0.88
BARD1	1.381	0.65
BRCA2	0.567711	0.2
CDH1	0.295017	0.21
BRIP1	0.258056	0.09

B

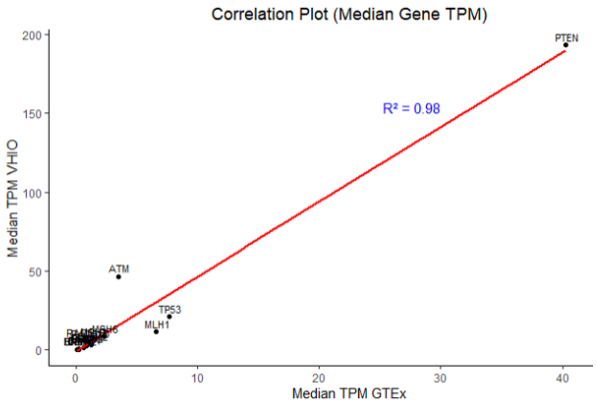

***Figure 2:* TPM expression values and correlation plot.** (A) Table of median TPM for the 15 genes of interest in which they have been ordered from highest to lowest TPM (VHIO) value. The median was calculated across all samples in the study. (B) A correlation plot was performed to compare median TPM values of VHIO and GTEx. TPM: Transcript Per Million.

Gene expression and DEA

Cancer is a heterogeneous disease involving more than one gene, and for that reason, the identification of differentially expressed genes (DEG) is a crucial step towards its understanding. It has been previously reported that there are some genes affecting HBOC risk, so this study will focus on determining their expression.

The first step was to visualize the expression levels of the 15 genes of interest previously mentioned. To do that, the raw gene counts after using the TMM normalization method were represented. The heatmap (Figure 3A) represents the z-scores of gene expression and also shows the type of cancer, gender, and the condition corresponding to each sample. Moreover, cancer and control patients are mixed, which indicates that there is not clear separation between the two conditions when focusing on these genes.

The second step was to perform a DEA, which is the type of analysis that allows us to identify the genes that are differentially expressed between different conditions. To do that, all samples were subjected to preprocessing steps (raw counts filtering and data transformation), and final analysis was performed by comparing the 47 cancer samples against the 6 control samples, using the voom + limma approach and adjusting for the first two surrogate variables (Supplementary Figure 3). Although some results were obtained, none of our genes of interest met the selection criteria (adjusted p-value < 0.05 and |logFC| > 1) . The position of these genes with respect to the significance limits can be observed in the volcano plot (Figure 3B).

Even though our 15 genes of interest were not differentially expressed, other genes were classified as DEGs: 120 were down-regulated and 2 were up-regulated (Supplementary Figure 4A). These ones were used to conduct a GSEA using the C6 oncogenic signature gene sets from MSigDB for oncogenic signature analysis. Three sets related to breast cancer research were identified: RAF_UP.V1_UP, LTE2_UP.V1_UP and CYCLIN_D1_KE_.V1_UP; all containing up-regulated genes in breast cancer cells (Supplementary Figure 4B).

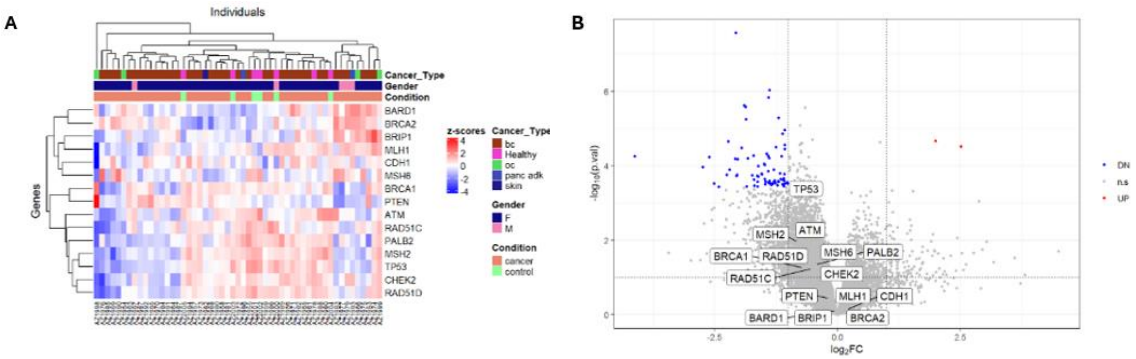

***Figure 3:* Heatmap and volcano plot for gene expression analysis.** (A) Heatmap representing the z-scores of TMM values, including information regarding the studied condition, cancer type, and gender. (B) Volcano plot of genes of interest following differential expression analysis. Down-regulated (DN) genes are coloured in blue, not significant (n.s) genes are coloured in grey and up-regulated (UP) genes are coloured in red (adjusted p-value < 0.05 and |logFC| > 1).

Patient specific analyses

As mentioned above, none of the 15 genes of interest was differentially expressed when comparing cancer cases with control cases, and therefore, none of them served as biomarker for the series of patients. In order to find potential patient-specific alterations, we performed different analyses to examine the genes of interest: 1) comparison of each patient classified as cancer against the six patients classified as controls in the case of DEU and SUPPA. 2) comparison of each patient to the rest of the samples in the case of OUTRIDER and FRASER.

OUTRIDER analysis

This analysis allows to find aberrantly expressed genes in RNA-seq samples. It does so by fitting a negative binomial model to RNA-seq read counts, correcting for variations in sequencing depth and apparent co-variations across samples and removing confounding factors. Read counts that significantly deviate from the distribution are detected as outliers. These outliers are detected within a given population by comparing each sample to all others without considering any specific conditions (Brechtmann et al., 2018).

After applying the OUTRIDER workflow, we selected all our genes of interest without considering the adjusted p-value (padjCutoff = 1) and discarding the rest of the genes. Once we had this list, the associated p-values were manually adjusted using the *p.adjust() function* in order to restrict the adjustment to only the genes of interest. As a result of this adjustment, two genes were found to be significant: *BARD1*, corresponding to sample AZ1971 and *BRIP1*, found in sample AZ1974 (Table 2).

In addition, to compare the expression of these two genes with that of the rest in the different samples, we generated a dot plot in which the normalized counts (normcounts) of the genes, provided by OUTRIDER, were represented (the significant genes are highlighted in red, with their names in blue). We can observe that the expression values of both genes were low, although they were not among the most extreme values (Figure 4). This may be due to the presence of genes with lower expression. However, when compared to the rest of the samples, other genes were not significant because they also showed low expression in those samples (data not shown).

Table 2: **OUTRIDER analysis results.** Samples in which aberrantly expressed genes are present with their respective adjusted p-value and normalized counts.

Table 2: OUTRIDER analysis results

Gene ID	Sample ID	Adjusted p-value	Normcounts
BARD1	AZ1971	0.0349	274
BRIP1	AZ1974	0.0439	89

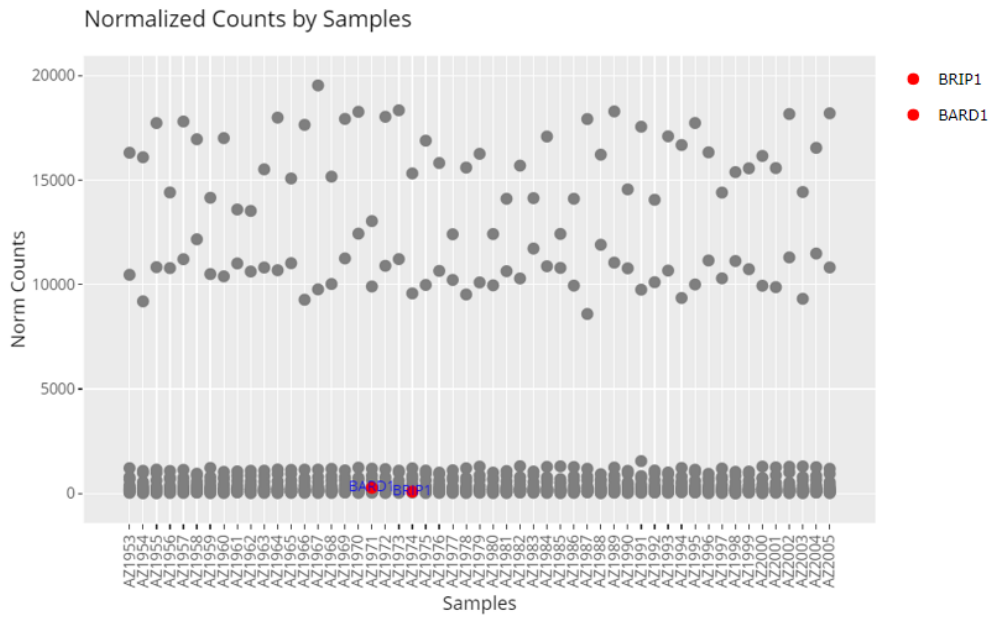

Figure 4: **Representation of the normalized counts derived from the OUTRIDER analysis.** Plot of the normalized counts for the genes of interest across all samples, with aberrantly expressed genes highlighted in red and listed in Table 2.

DEU analysis

In eukaryotic organisms, alternative splicing is a very important process for the regulation of gene expression, since alterations in it can have serious consequences. By applying the DEU analysis, which departs from exon raw counts (*.rds objects*), we studied the differential use of exons in a specific way by comparing each cancer patient against controls.

We found, as expected, that in most of the genes there were differentially expressed exons. We then manually checked in IGV and observed that there were no clear differences between cancer and control samples (data not shown). We concluded that they were false positive findings.

Local events analysis

When analyzing alternative splicing, various events can occur, each of which has different effects on the regulation of gene expression. To identify these local events, we used a tool called SUPPA, which can identify: skipping exon (SE), mutually exclusive exons (MX), alternative 5´ splice site (A5), alternative 3´ splice site (A3), retained intron (RI), alternative first exon (AF) and alternative last exon (AL).

A results table was obtained for each patient containing the following main information: gene name, type of local event, position and p-value. A final list of 24 significant local events was found (p-value < 0.05): 8 SE, 15 AF and 1 AL (Table 3). It is remarkable that some patients have more than one event. However, when these results were checked manually in IGV, there were no clear differences between cancer and control samples. For example, the reads mapped to the exons followed the same pattern, the coverage was very similar, and Sashimi Plots were almost the same in both cases (cancer and control). In addition, a small signal was observed in regions where exons were missing, which could be and indicative of artifact introduction in the results. Some examples can be seen in the Supplementary Figure 5.

Table 3: **Local events retrieved from SUPPA analysis:** Table displaying each significant local splicing event for each gene and the corresponding sample.

Table 3: Local events retrieved from SUPPA analysis

Sample ID	Gene ID	Local Event
AZ1953	*RAD51C*	AF
AZ1954	*MLH1*	AF
AZ1956	*RAD51C*	AL
AZ1963	*CDH1*	SE
AZ1967	*MLH1*	AF
AZ1976	*RAD51D*	AF
AZ1976	*TP53*	AF
AZ1976	*TP53*	SE
AZ1976	*ATM*	AF
AZ1977	*MLH1*	AF
AZ1977	*PTEN*	SE
AZ1977	*PTEN*	SE
AZ1979	*MLH1*	AF
AZ1979	*PALB2*	SE
AZ1980	*CDH1*	SE
AZ1983	*TP53*	SE
AZ1984	*TP53*	SE
AZ1985	*MLH1*	AF
AZ1985	*ATM*	AF
AZ1991	*ATM*	AF
AZ1993	*MLH1*	AF
AZ1996	*MLH1*	AF
AZ1998	*PALB2*	AF
AZ1998	*RAD51D*	AF

FRASER analysis

This analysis allows the detection of rare aberrant splicing events in transcriptome profiles (alternative splicing and intron retention) departing from aligned files (*.bam files*) without the use of explicit controls. It is done by modelling the expected read count ratios to control for confounding factors. Given these expectations, the ratios are assumed to follow a beta-binomial distribution with a junction specific dispersion. Outlier events are then identified as read-count ratios that deviate significantly from this distribution. FRASER works on the splice metrics ψ5 (used to assess the acceptor site usage in a specific context, indicating how often a particular acceptor site is used with a given donor site compared to all other acceptor sites), ψ3 (used to assess the donor site usage in a specific context , indicating how often a particular donor site is used with a given acceptor site compared to all other donor sites) and θ (that measures the overall splicing efficiency of a splice site, to detect partial or full intron retention events) to be able to detect any type of aberrant splicing event from exon skipping over alternative donor usage to intron retention (Mertes et al., 2021).

We applied the FRASER workflow to subsequently select our genes of interest. We obtained 14 significant results belonging to different samples but only for two genes: *PTEN* and *ATM*. The 8 *PTEN* results were observed in both cancer cases (6) and control cases (2); all of them were ψ5 indicating 5´ splice sites with their associated coordinates. Regarding the 6 *ATM* results, all of them were observed in cancer cases. In this case, we identified ψ5 and ψ3, indicating 5' and 3' splice sites, respectively, together with θ, which represents a partial or a full intron retention, with their associated coordinates (Table 4).

Table 4: **FRASER analysis results.** Table displaying each aberrant splicing event for each gene and the corresponding sample. It includes the genomic positions (start and end), the type of splicing event, and the adjusted p-value.

Table 4: FRASER analysis results

Sample ID	Start	End	Gene ID	Type	Adjusted p-value
AZ1976	87880439	87894024	*PTEN*	ψ5	0.00035226
AZ1976	87880439	87925512	*PTEN*	ψ5	0.00035226
AZ2000	87878353	87894063	*PTEN*	ψ5	0.0025412
AZ2000	87878353	87925512	*PTEN*	ψ5	0.0025412
AZ1979	108284473	108284474	*ATM*	θ	0.0057795
AZ1971	87864080	87864150	*PTEN*	ψ5	0.0041119
AZ1971	87864080	87864158	*PTEN*	ψ5	0.0041119
AZ1954	108247128	108248932	*ATM*	ψ5	0.0044958
AZ1954	108247128	108250700	*ATM*	ψ5	0.0044958
AZ1954	108247128	108250700	*ATM*	Ψ3	0.018524
AZ1954	108249103	108250700	*ATM*	Ψ3	0.018524
AZ1998	87958596	87960893	*PTEN*	ψ5	0.018478
AZ1998	87958596	87960936	*PTEN*	ψ5	0.018478
AZ1954	108345909	108347278	*ATM*	Ψ3	0.039571

4 DISCUSSION

Cancer is a disease caused by a large number of factors, which makes it very difficult to find a specific and common cause that allows its treatment. Among these causes, genetic components play a crucial role in cancer development and progression. The main goal of this study was to evaluate whether RNA-seq data analysis of 15 genes of interest can provide a genetic diagnosis of hereditary breast and ovarian cancer in patients with high suspicion for HBOC, where conventional DNA testing has yielded inconclusive results. Typically, in RNA-seq data analysis, patients are grouped by conditions which are compared to identify DEGs across all samples of a condition. In this study we first compared all cancer and control samples, focusing on the genes of interest, to gain a deeper understanding of the data. After confirming that none of the 15 genes were commonly differentially expressed, we performed patient individual approaches using OUTRIDER and alternative splicing analyses.

To start, we processed the FASTQ files obtained from the samples sequenced at the Vall d'Hebron laboratories with the goal of comparing the resulting expression data to the expression values from the GTEx project. The GTEx project is a well-established resource database with associated tissue bank designed to study the relationship between genetic variation and gene expression in human tissues

(Lonsdale et al., 2013). This dataset can be used to assess whether the gene expression behaviour in our samples is similar to this reference. When we performed the correlation of TPM values, we obtained a high positive number (R^2 = 0.98), indicating that our data exhibited approximately the same distribution of expression values as in the reference study.

As outlined in the formulas for TPM and coverage in the Processing Details section of the Methods, genes with high expression levels are expected to have higher TPM values compared to genes with low expression levels due to the number of reads mapped. This effect is also visible in the coverage and becomes more evident when examining raw counts in IGV, as no normalization (such as that applied in TPM) is performed in this case, leading to significantly different coverages between genes. Consequently, genes with low TPMs and, therefore, low coverage could not be well studied because they had not enough reads. Additionally, the insufficiency of reads causes magnification of differences when applying the base 2 logarithm in studies. However, these differences are not as pronounced, making them difficult to confirm in IGV. These issues are likely due to the insufficient number of reads per sample (Supplementary Table 1). A possible solution could be to increase the number of sequenced reads per sample to improve the coverage of those genes with lower coverage.

Of note, we observed that patients AZ1956 and AZ1976, who belong to the same family (Supplementary Table 2), were grouped together in the hierarchical clustering when we used the Ward.D2 method (Supplementary Figure 2A), indicating that in this case, the family relationship (genetic background) was stronger than the condition itself. Despite these issues, we performed the DEA with the available samples, obtaining the DEGs showed in Supplementary Figure 4A. Among them, we did not find genes directly related to HBOC, which may be attributed to the imbalance in the comparison. However, this is also an important finding, as a deeper analysis of these genes could reveal potential links to HBOC. In fact, the GSEA performed on the DEGs revealed three gene sets that were significantly enriched and associated with breast cancer (Supplementary Figure 4B).

Regarding the OUTRIDER and RNA-splice analyses, we identified events in the OUTRIDER, DEU, and SUPPA analyses. However, when we checked manually, none of these events were present and seemed to be artifacts. This could be due to an insufficient number of reads per sample and low coverage (explained previously), which likely magnify the expression differences and led to false positives. Furthermore, when analyzing the list of differential local events, obtained from SUPPA, to check for the positive control (pseudoexon in *BRCA1*), we found that it was absent. This is because SUPPA does not detect this type of event (Trincado et al., 2018). Given these issues, we believe that the possible solutions are: 1) Invest in sequencing with a higher number of reads per sample, and 2) Use a control event detectable by SUPPA or consider enhancing the tool to include detection of pseudoexon events, which could also benefit other researchers. Additionally, the alternative splicing pattern of the *BRCA1* gene (high-risk breast cancer predisposition gene) in control cases was manually analyzed in IGV. In sample AZ2005, a deletion of exons 8 to 10 (Δ8_10) was observed (Supplementary Figure 6), which has been reported as a minor variant (Colombo et al., 2014) and may affect the normal gene function (Rohlfs et al., 2000). This is not unusual, as physiological splicing isoforms can sometimes be present in control samples. The same pattern was observed in the cancer sample (AZ1953) (see Supplementary Figure 6). Thus, we should reconsider using this sample as a control in future analyses since it is not a typical isoform.

In an initial study using the FRASER method, we did not obtain any significant results when focusing on our genes of interest. This could be because the FRASER package was adjusting the p-value by considering all genes identified as aberrant splicing events. For this reason, we decided to manually adjust the p-value by considering only the 15 genes of interest, which yielded significant results. In this case, the two genes exhibiting alternative splicing were among those with the highest TPM values. Of note, one of the events corresponding to the *PTEN* gene was observed in the control sample AZ2000. This observation underscores the need to increase the number of control samples to ensure that they are representative of the entire healthy population. Since the main objective of this study was to explore splicing patterns, the alternative splicing event in the *ATM* gene had already been investigated through other bioinformatic approaches and later validated in the wet laboratory using PCR and gel electrophoresis (Supplementary Figure 7).

To summarize, although the DEA did not reveal any DEGs in the genes of interest, the OUTRIDER analysis identified 2 aberrantly expressed genes. SUPPA identified 24 local splicing events in 16 patients, but these were false positive due to the low coverage and expression of some of the genes of interest in whole blood. Fourteen significant results were identified in the FRASER analysis corresponding to 7 patients. One of them was confirmed experimentally. From the hierarchical clustering (ward.D2 method) and SUPPA analyses, only two patients belonged to the same family (AZ1956 and AZ1976). No other family association was observed in the rest of the analyses.

One of the main challenges in our study was the small sample size (53). We had to study 47 case samples for which only 6 control samples were available. We considered including additional samples from public databases, but the sequencing techniques were different, and this could affect the results. To solve this problem, we should increase the number of samples analyzed as much as possible, especially the number of controls.

As a final remark, we would like to remark that to carry out this study, we worked with whole blood RNA samples with the aim of finding a germline genetic diagnosis for each patient in a non-invasive way. Although the primary source of RNA for oncological studies is typically tumour tissue, which has higher gene expression levels, using tumour samples is much more

invasive. Additionally, tumour biopsies present challenges, such as being embedded in paraffin, which can also affect the results. Another issue was that genes (such as *BRCA1/2*) were expressed at low levels in whole blood, which represents a limitation. Since few RNA aberrations were found, other approaches, such as working with tumour samples, might be useful to complement our study.

5 CONCLUSION

The main goal of this study was to identify genetic abnormalities that might impact the development of HBOC. All of our analyses, except for DEA, were conducted using a patient-specific approach with the aim of providing personalized diagnostics and solutions for each affected patient. In most of our analyses, we found events that appear promising and could help explain the phenotypes observed in the patients.

However, although bioinformatics approaches, including transcriptomics, are highly powerful, the results still need to be validated in the laboratory. Furthermore, this is only an initial exploration, and all our findings should be validated with further analyses, both computationally and experimentally in the laboratory.

Acknowledgements

I would like to express my gratitude to my supervisor, Lara, for all the advice and knowledge that she has shared with me throughout this process. I also want to thank Sara for giving me the opportunity to participate in this project, and Setareh for all her help during the work.

Additionally, I want to thank all the members of the Bioinformatics Unit: Irene, Ángel, Pau, Alba, and Gilles, for their support over these months. They have created a working environment that made everything much easier.

References

Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N., & Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, *21*(9), 1521-1531. https://doi.org/10.1261/rna.051557.115

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-Seq data. *Nature Precedings*. https://doi.org/10.1038/npre.2012.6837.2

Beitsch, P. D., Whitworth, P. W., Hughes, K., Patel, R., Rosen, B., Compagnoni, G., Baron, P., Simmons, R., Smith, L. A., Grady, I., Kinney, M., Coomer, C., Barbosa, K., Holmes, D. R., Brown, E., Gold, L., Clark, P., Riley, L., Lyons, S., … Nussbaum, R. L. (2019). Underdiagnosis of Hereditary Breast Cancer: Are Genetic Testing Guidelines a Tool or an Obstacle? *Journal of Clinical Oncology*, *37*(6), 453-460. https://doi.org/10.1200/JCO.18.01631

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bonnal, S. C., López-Oreja, I., & Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer—Implications for care. *Nature Reviews Clinical Oncology*, *17*(8), 457-474. https://doi.org/10.1038/s41571-020-0350-x

Breast Cancer Association Consortium. (2021). Breast Cancer Risk Genes—Association Analysis in More than 113,000 Women. *New England Journal of Medicine*, *384*(5), 428-439. https://doi.org/10.1056/NEJMoa1913948

Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V. A., Avsec, Ž., Herzog, M., Bader, D. M., Prokisch, H., & Gagneur, J. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *The American Journal of Human Genetics*, *103*(6), 907-917. https://doi.org/10.1016/j.ajhg.2018.10.025

Chen, Y., Chen, L., Lun, A. T. L., Baldoni, P. L., & Smyth, G. K. (2024). *edgeR 4.0: Powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets*. https://doi.org/10.1101/2024.01.21.576131

Colombo, M., Blok, M. J., Whiley, P., Santamariña, M., Gutiérrez-Enríquez, S., Romero, A., Garre, P., Becker, A., Smith, L. D., De Vecchi, G., Brandão, R. D., Tserpelis, D., Brown, M., Blanco, A., Bonache, S., Menéndez, M., Houdayer, C., Foglia, C., Fackenthal, J. D., … De La Hoya, M. (2014). Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: A report from the ENIGMA consortium. *Human Molecular Genetics*, *23*(14), 3666-3680. https://doi.org/10.1093/hmg/ddu075

Hodgson, A., & Turashvili, G. (2020). Pathology of Hereditary Breast and Ovarian Cancer. *Frontiers in Oncology*, *10*, 531790. https://doi.org/10.3389/fonc.2020.531790

Hu, C., Hart, S. N., Gnanaolivu, R., Huang, H., Lee, K. Y., Na, J., Gao, C., Lilyquist, J., Yadav, S., Boddicker, N. J., Samara, R., Klebba, J., Ambrosone, C. B., Anton-Culver, H., Auer, P., Bandera, E. V., Bernstein, L., Bertrand, K. A., Burnside, E. S., … Couch, F. J. (2021). A Population-Based Study of Genes Previously Implicated in Breast Cancer. *New England Journal of Medicine*, *384*(5), 440-451. https://doi.org/10.1056/NEJMoa2005936

Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A Beginner's Guide to Analysis of RNA Sequencing Data. *American Journal of Respiratory Cell and Molecular Biology*, *59*(2), 145-157. https://doi.org/10.1165/rcmb.2017-0430TR

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580-585. https://doi.org/10.1038/ng.2653

Marmolejo, D. H., Wong, M. Y. Z., Bajalica-Lagercrantz, S., Tischkowitz, M., Balmaña, J., Patócs, A. B., Chappuis, P., Colas, C., Genuardi, M., Haanpää, M., Vetti, H. H., Hoogerbrugge, N., Irmejs, A., Kahre, T., Klink, B., Krajc, M., Milagre, T. H., De Putter, R., Steinke-Lange, V., … Wimmer, K. (2021). Overview of hereditary breast and ovarian cancer (HBOC) guidelines across Europe. *European Journal of Medical Genetics*, *64*(12), 104350. https://doi.org/10.1016/j.ejmg.2021.104350

McDevitt, T., Durkie, M., Arnold, N., Burghel, G. J., Butler, S., Claes, K. B. M., Logan, P., Robinson, R., Sheils, K., Wolstenholme, N., Hanson, H., Turnbull, C., & Hume, S. (2024). EMQN best practice guidelines for genetic testing in hereditary breast and ovarian cancer. *European Journal of Human Genetics*, *32*(5), 479-488. https://doi.org/10.1038/s41431-023-01507-5

Mertes, C., Scheller, I. F., Yépez, V. A., Çelik, M. H., Liang, Y., Kremer, L. S., Gusic, M., Prokisch, H., & Gagneur, J. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. *Nature Communications*, *12*(1), 529. https://doi.org/10.1038/s41467-020-20573-7

Montalban, G., Bonache, S., Bach, V., Gisbert-Beamud, A., Tenés, A., Moles-Fernández, A., López-Fernández,

A., Carrasco, E., Balmaña, J., Diez, O., & Gutiérrez-Enríquez, S. (2021). BRCA1 and BRCA2 whole cDNA analysis in unsolved hereditary breast/ovarian cancer patients. *Cancer Genetics*, *258-259*, 10-17. https://doi.org/10.1016/j.cancergen.2021.06.003

PCAWG Transcriptome Core Group, Calabrese, C., Davidson, N. R., Demircioğlu, D., Fonseca, N. A., He, Y., Kahles, A., Lehmann, K.-V., Liu, F., Shiraishi, Y., Soulette, C. M., Urban, L., Calabrese, C., Davidson, N. R., Demircioğlu, D., Fonseca, N. A., He, Y., Kahles, A., Lehmann, K.-V., … Von Mering, C. (2020). Genomic basis for RNA alterations in cancer. *Nature*, *578*(7793), 129-136. https://doi.org/10.1038/s41586-020-1970-0

Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., & Huber, W. (2013). Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences*, *110*(38), 15377-15382. https://doi.org/10.1073/pnas.1307202110

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47-e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Rohlfs, E. M., Puget, N., Graham, M. L., Weber, B. L., Garber, J. E., Skrzynia, C., Halperin, J. L., Lenoir, G. M., Silverman, L. M., & Mazoyer, S. (2000). AnAlu-mediated 7.1 kb deletion ofBRCA1 exons 8 and 9 in breast and ovarian cancer families that results in alternative splicing of exon 10. *Genes, Chromosomes and Cancer*, *28*(3), 300-307. https://doi.org/10.1002/1098-2264(200007)28:3<300::AID-GCC8>3.0.CO;2-1

the SEOM Hereditary Cancer Working Group, González-Santiago, S., Ramón Y Cajal, T., Aguirre, E., Alés-Martínez, J. E., Andrés, R., Balmaña, J., Graña, B., Herrero, A., Llort, G., & González-del-Alba, A. (2020). SEOM clinical guidelines in hereditary breast and ovarian cancer (2019). *Clinical and Translational Oncology*, *22*(2), 193-200. https://doi.org/10.1007/s12094-019-02262-0

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178-192. https://doi.org/10.1093/bib/bbs017

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

Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*, *26*(8), 903-909. https://doi.org/10.1261/rna.074922.120