Transcriptomic analysis of BRCA1 and BRCA2 carriers

Marina Vilardell Lladó, Setareh Kompanian, Ester Aguado Flor, Sara Gutiérrez-Enríquez, Lara Nonell Mazelon

[1]Hereditary Cancer Genetics Group, Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain. [2] Bioinformatics Unit, Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain. [3] *Universitat de Vic-Universitat Central de Catalunya, Catalonia, Barcelona, Spain*

Abstract

Most frequently mutations that lead to the development of breast and ovarian cancer are *BRCA1* and *BRCA2.* Blood samples from *BRCA1, BRCA2* and non-mutation carriers were extracted for both healthy and cancer-affected individuals. These samples underwent lymphocyte stimulation and they were analyzed by using novel bioinformatic tools to gain insight into the transcriptome and to revel different traits between each of the phenotypes. A complex transcriptomic analysis was performed, including and assessing the strategies undertaken in all the necessary steps, quality control, annotation, deconvolution, filtering, variable selection, differential expression and gene set enrichment analysis. Final results suggest that cell cycle and DNA replication pathways were positively overexpressed in *BRCA1* and *BRCA2* healthy samples, as well as and in *BRCA1* affected. In contrast, negative pathways shared a similar pattern in all the studied groups.

Contact: marina@marinavilardell.cat

Supplementary information: Supplementary data are available at https://github.com/marinavilardellado/TFM

1 Introduction

According to the Global Cancer Observatory (GCO), breast (BC) and ovarian cancer (OC) rank among the most prevalent malignant tumors in women. In 2020, the worldwide diagnosis was over 2.2 million cases of BC and 313.000 cases of OC [1]. These statistics position BC as the first cause of cancer-related deaths, while OC was the eighth leading cause of deaths. *BRCA1* and *BRCA2 (BRCA1/2)* are two well-known genes associated with these two types of cancers, as individuals who inherit or acquire sporadic pathogenic germline variants in these genes present an increased risk of developing them in the future. Approximately 55%-72% of women with a harmful mutation in *BRCA1* and 45-69% with a *BRCA2* variant will develop BC. In contrast, these percentages decrease for OC, since its risk stands at 39-44% for women with a *BRCA1* variant and 11-17% for those with a *BRCA2* pathogenic variant [2].

BRCA1/2 are tumor suppressor genes that play a significant role in maintaining genome stability and the integrity of DNA. Both produce essential proteins that facilitate the repair of damaged DNA, specifically double-strand DNA breaks by homologous recombination. This repair mechanism prevents the introduction of new mutations, and therefore, cancer initiation and development.

When these genes are mutated or absent in cells, the stringency and precision of DNA damage repair mechanisms weakens. Consequently, cells are unable to effectively repair DNA breaks, leading to an increased probability of accumulating variants across cell division. While some of these cells may die, the survival and stabilization of these unrepaired DNA cells confer the capacity to independently undergo cell division, resulting in cancer development [3][4]. For this reason, cells with biallelic*BRCA1/2* deleterious mutation become highly sensitive to a wide range of DNA-damaging agents, coming both from environmental factors, such as ionizing radiation (UV, X and Gamma rays), and endogenous sources (reactive oxygen species) [5] .

A multitude of research studies have been dedicated to investigating the impact of *BRCA1/2* genes on human health [6][7]. Having a deep understanding of how these genes function and how mutations affect their roles can help to identify individuals with a heightened genetic risk, that would benefit from the implementation of more aggressive preventive measures. As researchers have a better insight into the cancer mechanisms linked to *BRCA1/2*, they can devise treatments that are both more targeted and precise, aiming to decrease cancer mortality.

Over the recent years, gene expression regulating mechanisms have been extensively explored, particularly within the context of oncology. The transcriptome contains all the information encoded in RNA that has been transcribed from DNA. Gene expression levels resulting from transcription are studied through transcriptomics, but the causes and consequences of changes in RNA expression can be examined in other omics perspectives, such genomics, epigenomics and proteomics, where each level presents its specifics advantages and limitations [8]. In contrast to the genome, which is stable, the transcriptome

depends on the dynamics of an organism at a given moment, responding to the physiological and/or pathological conditions.
Thanks to the rapid advancement and development of cutting-edge technologies, the identification of cancer biomarkers and gene signatures has become possible, opening new horizons for understanding, managing, and treating cancer.

Since the discovery of RNA molecules, quite a few technologies have been developed with the primary goal of analyzing and quantifying the transcriptome in the context of human diseases. This endeavor has advanced the understanding and exploration of the intricacies of gene expression regulation and their contributions to cellular processes and functions, driving significant progress in the comprehension of cancer. These technologies range from traditional methods like northern blotting to novel approaches such as microarrays and next generation sequencing (NGS), represented by RNA sequencing (RNA-seq). These latter two methods are key in modern transcriptomics analysis [8]. In the scope of this project, our focus will center on microarray technology.

A microarray is a collection of biomolecules containing DNA probes corresponding to known sequences. In the Affymetrix arrays, probes are attached to a solid surface, creating a platform to which DNA fragments from a sample can hybridize. The probes consist of single-stranded DNA oligonucleotides, typically of 25 bases in length, rigorously designed to perfectly match specific target sequences. RNA fragments extracted from the samples of interest undergo reverse transcription into cDNA and then are fragmented and labelled to be finally introduced to the array for hybridization. The amount of hybridization observed for each probe is quantitatively measured using fluorescence, given that the fluorescence signal is directly proportional to the number of RNA fragments present in the sample [8][9].

Microarray experiments provide information regarding which genes are differently expressed when comparing different biological samples, often RNA extracted from blood. Blood samples are composed of heterogeneous mixtures of cell types with different proportions, and this may confound the analysis of differential gene expression. Hence, it is important to understand the cellular composition of the samples. In order to address this issue, deconvolution have been developed to infer cell type proportions from transcriptomics data [10]. As a result, deconvolution helps to perform more precise differential expression analyses, which is crucial for identifying biomarkers for disease diagnosis.

To understand the biological significance of gene expression changes, there are bioinformatic functional analyses that can be performed, such as Gene Set Enrichment Analysis (GSEA). GSEA is a computational method utilized to determine whether a pre-defined list of genes present consistent statistically significant differences between biological conditions [11].To perform this enrichment analysis, several collections of gene sets can be used. These, share a common biological function or are part of the same pathway and can be obtained through different databases such as Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome or the Molecular Signature Database. The method computes an enrichment score that reflects the degree to which genes are overrepresented at the top or bottom of the ranked list. A significant score will indicate that the gene set is associated with the biological condition. This information provides insights into the biological processes that play a key role in each condition [12].

In summary, transcriptomics has applications in various areas of biological and medical research, that can be used to identify biomarkers, understand processes, and study the effects of treatments.

This project is based on a prior study whose primary purpose was to determine the impact of irradiation on blood gene expression levels in both healthy and cancer-affected individuals, considering their specific *BRCA1/2* mutation status. Carrier individuals possessed either the *BRCA1* or the *BRCA2* mutation, but not both simultaneously. A standard bioinformatics analysis was performed to examine the differential expression levels in blood samples of *BRCA1/2* carriers and non-carriers for both non-irradiated and irradiated samples. Additionally, the study also explored whether lymphocytes carrying these mutations presented an increased sensitivity to radiation when assessed in a controlled in-vitro environment [13]. The data from this study was used to perform the analysis for my project. Nevertheless, in contrast to the original study's extensive range of comparisons between samples, this project is focused exclusively on a single radiation status: non-irradiated samples. Within this context, only four main comparisons are analyzed (*BRCA1* healthy versus non-carriers healthy; *BRCA1* cancer-affected versus non-carriers cancer-affected; *BRCA2* healthy versus non-carriers healthy; *BRCA2* cancer-affected versus non-carriers cancer-affected).
Therefore, the main objective of the current study is to reassess the initial analysis by using an updated methodology and to characterize both *BRCA1* and *BRCA2* breast cancer carriers beyond the genomic mutation for non-irradiated samples.
Updated bioinformatics workflows were strategically designed, as the methodologies and techniques have evolved and differ from those used in the past years. Thereby, the results of the project should be improved by updating the bioinformatics analysis and providing more comprehensive steps.
This main objective is segmented into a series of smaller ones, covering deconvolution, differential expression analysis, and concluding with functional analysis. These steps will contribute to achieve the primary purpose of this study.

2 Methods

Subjects and data collection

The data employed in this project is derived from a study in which blood samples from 53 individuals were collected. These individuals were meticulously categorized based on the mutational status of *BRCA1* and *BRCA2* genes, alongside their disease status. The resulting categories were as follows: 18 *BRCA1* mutation carriers (10 with BC and 8 healthy), 20 *BRCA2* mutation carriers (11 with BC and 9 healthy) and 15 individuals without detected mutation (5 with BC and 10 healthy).

To establish blood cultures, researchers used 4.5ml of RPMI medium supplemented with 1% phytohaemagglutinin (PHA), a substance known for its immune system stimulation properties, facilitating the division of T-lymphocytes cells. Peripheral blood mononuclear cells (PBMC), which include lymphocytes, were isolated using a density gradient centrifugation method, resulting in 2 samples per individual. One sample was treated with 2 Gy of gamma irradiation, while the other remained untreated (baseline type). A total of 106 samples were obtained for subsequent analysis. After a 24-hour incubation period (in an incubator with 5% CO2 and 37 °C), RNA was extracted from the collected samples following a standard isolation protocol using Trizol (Invitrogen), and gene expression profiling data was obtained using the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array (Thermo Fisher). This 3'IVT microarray comprises more than 54.000 probe sets, enabling the analysis of expression levels for over 47.000 transcripts [14].

The microarray chips were scanned using a device designed to quantify the fluorescence intensity of each probe. Subsequently, the scanner process generated individual CEL files for each single microarray experiment, transforming the intensity measurements into a numerical matrix, that contains the raw intensity data for each probe.

Researchers supplied metadata, which contained information of the disease condition of each patient, as well as several covariates that need to be considered for the analysis, including age, smoking history, years of smoking, chemotherapy and radiotherapy history, and the number of years since the patient underwent those specific cancer treatments.

Bioinformatic analysis

R Studio software (version 4.3.0), along with the R programming language, was used as the primary tool to conduct data processing, analysis, visualization, and computations for this project.

The CEL files were the starting point for the analysis. To process these files, the *Affy* package [15] (version 1.78.0) from Bioconductor was opted for, since it is specially designed for 3'IVT Affymetrix GeneChip arrays [16]. The quality control and the normalization steps were performed with some functions included in this package. For the sample aggregation procedure, the packages employed were *Stats* [17] (version 4.3.0) and *dendextend* [18] (version 1.17.1). The annotation process was performed using two distinct annotation methods. The first approach involved the hgu133Plus2.db (version 3.13.0) Bioconductor's package [19], and the second involved using an annotation file provided by the microarray manufacturer for this specific array, Thermo Fisher, which was downloaded from its website [20]. Two tools, Immunedeconv [21] (version 2.1.0) and MIXTURE [22] (version 0.01) R packages were employed to conduct the deconvolution process. Variable selection was performed with *Glmnet* [23] (version 4.1-7). Linear Models for Microarray Data, abbreviated as Limma [24], was the Bioconductor package employed to perform the differential expression analysis (version 3.56.1), and finally, the gene set enrichment analysis was performed with the clusterProfiler [25] (version 4.8.1) Bioconductor's package. Graphical representation plots were performed by using ggplot2 (version 3.4.2), and ComplexHeatmap (2.15.4).

Quality control

To ensure data quality and determine whether each array was suitable for the analysis or needed to be discarded, several quality control approaches were used [26][27]. The first step involved examining the scan-generated images to identify any potential issues, such as spots or irregularities on the array.

Then, five types of graphics were performed to examine the data distribution and to identify any outlier samples. These included:

Boxplots of intensity data. This plot provides a concise summary of continuous data for each sample, showing the median and the quartiles.

Histogram of intensity data, which details the insights into the shape of the data's distribution.

MA plot. This graphical representation helps to identify which arrays are behaving differently from the rest on the probe set basis. It displays (M) the difference between the log intensity value of one array minus the intensity of a reference (pseudo median of the rest of the arrays), against (A) the average of log intensity for a given probe set.

RLE (relative Log Expression). This plot illustrates how much the expression level of a probe set in a particular array differs from the average level across all arrays. The RLE value is assessed as the log ratio of the expression level of a probe relative to the median expression level of that probe in all arrays using a linear model.

NUSE (Normalized Unscaled Standard Errors). This quality control plot assesses the standard errors of the linear models generated for the RLE analysis.

Normalization

Microarray experiments involve complex laboratory procedures that can introduce systematic biases due to differences in sample preparation, labeling, hybridization, or scanning, leading to signal intensities across the 106 arrays used in the study. In order to compare between the arrays, a normalization procedure was performed on the data to correct for these biases and remove technical variations, ensuring that the observed differences on the intensity values of the samples primarily reflect biological variations in gene expression rather than technical effects. Hence, raw expression values obtained from CEL files were preprocessed and normalized using the Robust Multichip Average (RMA) method [27], a three-step process that involves background correction, quantile normalization, and summarization of probes, resulting in a single intensity value for each of the probe sets.

Sample aggregation

Hierarchical clustering was applied to the data with the objective to examine whether samples could be grouped based on their disease status, distinguishing between cancer and healthy cases. Different distance metrics and linkage methods were used to define the clusters. Specifically, the Euclidean and correlation-based distances, as well as the Ward.D2 and average linkage methods. The purpose of using multiple approaches was to assess the consistency and similarities of the resulting cluster formations across different methodologies.

On the other hand, to address the high dimensionality of the data, given the large number of probe sets, we employed the principal component analysis (PCA) statistical method to investigate if distinct gene expression profiles existed between disease status. The proportion of data variability explained by the PCA model was also determined.

Annotation

Probe set identifiers were assigned to know annotations, which are, in this case, gene name identifiers by using the HUGO Gene Nomenclature Committee (HGNC) symbols. The annotated results from the Bioconductor's package and the annotated file from Thermo Fisher were compared to assess its effectiveness, by calculating the number of unannotated probes and the percentage of identical annotated gene symbols. The main objective was to determine which method performed better in annotating probe sets, aiming to retrieve as much information as possible.

In microarrays, a probe set may be associated with more than one annotated gene. This is because each probe set comprises multiple individual probes with different sequences, some of which might target regions that are shared by several genes or transcript isoforms. To handle this situation, the first gene symbol identification provided by Thermo Fisher was selected.

Unannotated (NA) probe sets were removed from the data set. As it may happen that different probe sets correspond to the same gene, the mean expression values for each duplicated gene and for each sample was computed, and a new matrix was created with the expression values of unique genes.

Deconvolution

As PHA was aggregated to the samples, we were interested in investigating the prevalence of T-lymphocytes within them. To achieve this, a deconvolution process was performed to determine the specific cellular composition of each sample. This step was crucial for evaluating the quality of the samples, to detect possible contaminations, and obtaining the cell fraction of different cell types present in the samples. T CD4 lymphocytes cellular fraction information was further used as a covariate for the differential expression gene (DEG) analysis, enabling us to account for any variations in this cell composition among the samples during the DEG.

Both packages employed to perform deconvolution are equipped with advanced computational algorithms designed to estimate the fractions of immune cells present in the microarray data. Within the "Immunedeconv" package, that offers multiple methods with distinct concepts, we selected two methods: CIBERSORT absolute mode and EPIC. The reason behind selecting those methods was because of the type of scores they provide, as they reflect each cell type's absolute fractions, that allow for between-samples and between-cell type comparisons. Both methods are classified into deconvolution-based approaches, which use v-support vector regression methods (v-SVR) to estimate immune cell type's proportion form a gene expression profile [28]. LM22.txt and CIBERSORT.R files were required to run Cibersort within "immunedeconv" package. These two files were downloaded from the cibersort website [29] and contain the signature gene matrix for 22 different immune cell types and a source code, respectively.

On the other hand, MIXTURE, which is a method based in cibersort, was also used to estimate the cell type fractions.

The input data for those methods was a gene expression matrix with gene symbols in rows and samples in columns.

Data had to be normalized, which was already performed in previous steps, and non-logarithmically transformed [28]. Once the cell fractions were obtained, statistical tests were applied on the data to evaluate its variability between phenotypes status. A Shapiro test was performed to examine the distribution of each cell type's data and to determine whether they followed a normal distribution or not. The test provided p-values for all cell types, and this information influenced the decision on which paired test was the most suitable for comparing the gene expression mean of the different phenotype of individuals. A Spearman's rank correlation test was employed to perform T CD4 cell type correlation analysis between CIBERSORT abs.mode, EPIC and MIXTURE.

Filtering

Only non-irradiated samples were selected to continue with the analysis of characterizing *BRCA1/2* carriers. Therefore, the findings and results obtained will specifically pertain to this subset of samples. Prior to the variable selection step, samples with missing values for the smoking variable were excluded, since this variable is important for the study and imputation methods might introduce bias into the data. Moreover, with a small sample size dataset, the data of each patient can have a significant impact on the analysis, so imputing these categorical values could lead to inaccurate results.

Standard deviation measurement was used to filter data and to avoid unexpressed or unchanged genes across different samples. The second quantile of the standard deviation was used as a threshold to select the genes to remain in the analysis. In other words, we kept the 50% of genes with higher variability.

Variable selection

Feature selection was performed with LASSO, a regression analysis method that aims to identify the most important predictor variables that are strongly associated with the response variable. LASSO introduces a penalty term to the linear regression function, forcing the coefficients of less significant variables to be shrunk towards zero [30]. In this case, we investigated the potential relationship between covariates – age, smoking history, and T CD4 cell proportions for healthy patients, along with chemotherapy, radiotherapy, and years elapsed between the end of these treatments and the blood extraction for the study, in cancer affected patients – and each patient's phenotype of the studied comparisons. Four binomial models for variable selection were constructed, one for each of the following comparisons, always comparing carriers versus non-carriers:

1.BRCA1 Affected vs NOMUT Affected
2.BRCA1 Healthy vs NOMUT Healthy
3.BRCA2 Affected vs NOMUT Affected
4.BRCA2 Healthy vs NOMUT Healthy

Since it is reasonable to expect that certain predictor variables may not have an effect in all the comparisons, employing a binomial approach allows the identification of these unique association for each comparison, provides more accurate and meaningful results.

Differential expression gene analysis

Differential expression gene (DEG) analysis was conducted on each of the four comparisons previously described, with two main objectives. The first objective aimed to detect genes that displayed different significant expression patterns between the two conditions. Secondly, it aimed to generate a ranked list of genes, which will then serve as input for the subsequent step, GSEA.

Limma, the package used for DEG, is based on a linear model analysis with empirical Bayes approach, that combines information across genes to stabilize the estimation of variance, and returns standard tests statistics, such as the fold change, B, t and p-value for each for the tested gene [31]. Since many independent tests (one per gene) are performed simultaneously, p-values were adjusted to multiple correction, in order to control the false discovery rate (FDR) using the Benjamini and Hochberg method, ensuring that the reported significant genes are unlikely to be false positives [32].

The covariates used in each of the comparisons depended on the results of variable selection, and adjusted P-value < 0.05 and the absolute value of log fold change (logFC) > 1 were the threshold to select differential expressed genes.

Gene Set Enrichment Analysis

The final step of this project involved a gene set enrichment analysis. The primary goal was to get biological pathway information from the list of previously ranked genes to identify a biomarker that can distinguish between patient's phenotypes. The Molecular Signature Database (MSigDB) was used to define two gene set collections for the analysis. The first one, was the Hallmark collection, which consist of 50 gene sets that represent specific and well-defined biological states. The second, was the Reactome subset of Canonical pathways collection, which contains 1654 gene sets derived from the Reactome pathway database [33].

Pathways were separated into positive (enriched at the top of the list) or negative (enriched at the bottom) according to its normalized enrichment score (NES), and those with an adjusted p-value < 0.05 were selected.

Finally, we assessed whether negative and positive pathways were shared between *BRCA1* healthy versus *BRCA1* affected, and *BRCA2* healthy versus *BRCA2* affected samples.

The metrics used to rank the genes was a combination of the sign of the logFC and the logarithm of the p-value.

3 Results

3.1 Preprocessing steps

Images generated by the scan were valid despite some traces of spots, marks, or signs of damage that were detected in the images 107 and 121. However, they were not significant enough to exclude them [Supplementary Figs. S1-S2]. Moreover, data quality plots outcomes demonstrated generally good quality of the arrays.

54,675 summarized probe sets were obtained after the normalization procedure, whose intensities were similarly distributed and homogenous in all the samples.

Samples did not aggregate by disease, as there was not a clear separation between healthy and cancer-affected patients in any of the hierarchical clustering dendrograms performed with different metrics and linkage methods [Fig 1.A and Supplementary Fig. S3-S4]. The PCA plot revealed overlapping ellipses in all the different phenotype conditions, suggesting similar gene expression profiles between all the individuals, independently of the *BRCA* mutation they may present, and their disease status, as visualized in Figure 1.B.

After evaluating both annotating methods, the Thermo Fisher file, was selected as the preferred method. Not only did it result in a higher number of annotated probe sets, but it is also considered the official source, providing a more reliable, curated, and informative dataset for further analysis. Once the aggregation of duplicates, the matrix contained the expression levels of 21,923 genes.

3.2 Deconvolution

CIBERSORT abs. mode and MIXTURE provided a higher level of detail regarding cell types. Therefore, the proportions of T CD4 naïve and memory activated cell types were summed and combined into a single cell type group (T CD4 aggregated) to perform correlation analysis with the T CD4 cell type proportion of EPIC, which can be found in supplementary Figs. S5-S7. CIBERSORT abs. mode and MIXTURE displayed the strongest correlation among the methods, as the Spearman's correlation value (rho) was 0.86. On the other hand, the rho value for EPIC and MIXTURE correlation analysis was 0.51, and the correlation between EPIC and CIBERSORT exhibited an even further decrease (rho= 0.4). One of the main differences in the correlation plots is that both EPIC - MIXTURE and EPIC-CIBERSORT present most of the EPIC data points concentrated between a 0.7-0.9 range of proportions, while MIXTURE and CIBERSORT concentrate their T CD4 data points between a 0.25-0.95 range. Based on these results, the T CD4 aggregated proportions obtained by CIBERSORT abs. mode, were the fractions used as a covariate in the next steps, variable selection, and differential expression analysis.

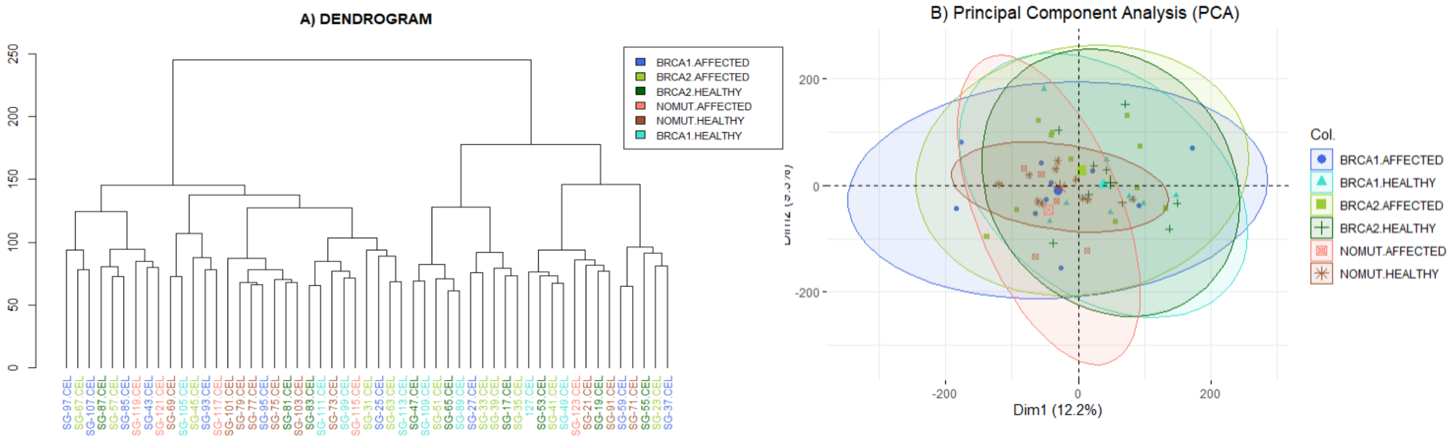


Fig 1. Sample aggregation plots. A) Hierarchical clustering dendrogram of each sample phenotype, using the Euclidean distance and the Ward.D2 linkage method. B) Principal Component Analysis (PCA) plot by sample's phenotype.

Once this method was selected, the distribution of T CD4 cell type fractions, which includes naïve and memory activated CD4 cells, were analyzed to compare the mean expression value and to detect significative differences between the phenotypes. To achieve this, a Kruskal Wallis test was performed. No statistically differences (p-value < 0.05) between the means of the phenotypes in neither T CD4 memory activated fraction and T CD4 naïve cell types, as it is showed in Fig 2, although the median level of T CD4 memory activated cells in *BRCA2* healthy samples is a little bit higher than the rest. On the other hand, a Wilcoxon test was also performed to ensure no statistically significant differences were detected between the two means of all possible phenotype comparisons.

Cell proportion results obtained through CIBERSORT method presented a notable increasement in fractions of T CD4 memory activated cells, as it is seen in Fig.3 which graphically illustrates the cell type proportions that were obtained by this method.

3.3 Filtering

Seven samples were removed from the analysis, since they contained missing values (NA) for the smoking variable, as detailed in Supplementary Table S1. Of note, *BRCA1, BRCA2* and 4 genes related to the *MYC* family (*GJA9-MYCBP, MYC, MYCBP2,* and *MYCBPAP*) passed the filtering when using a sd = 0.05.

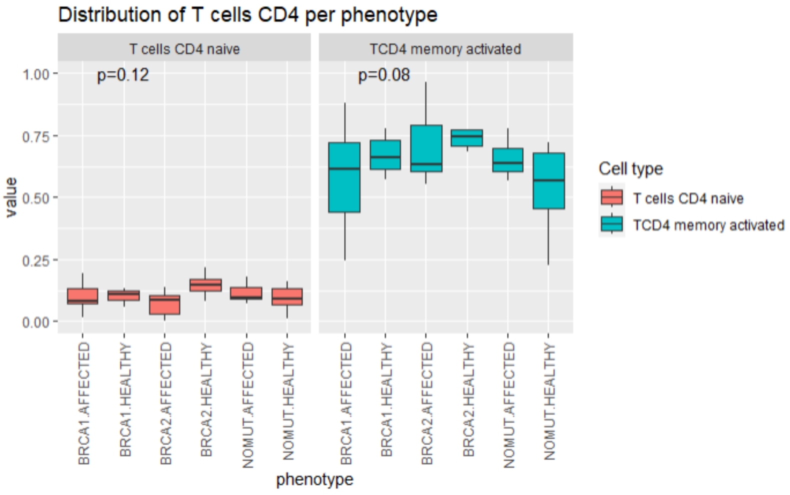


Fig 2. Distribution of T CD4 cell type per each of the phenotypes, using the CIBERSORT absolute mode deconvolution method.

3.4 Variable Selection

For each of the four comparisons studied, LASSO penalized and selected specific variables. Different variables, summarized in Table 1, were added to the model, depending on the disease status of group of samples comparing, as previously described in the methods section. The variables LASSO selected for each comparison were as follows, as its coefficient was not shrunk to 0, indicating its relation-

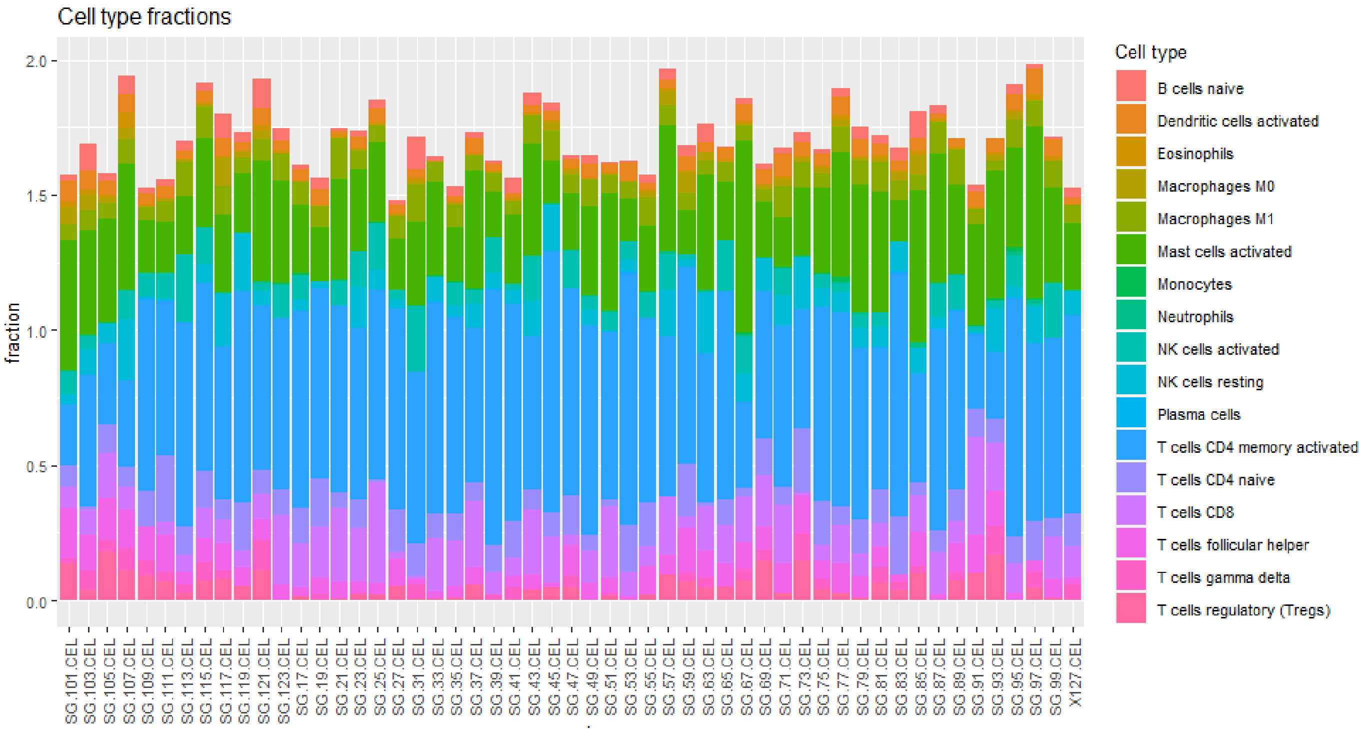


Fig 3. Cell type proportions obtained through CIBERSORT absolute mode.

ship with the phenotype of each group:

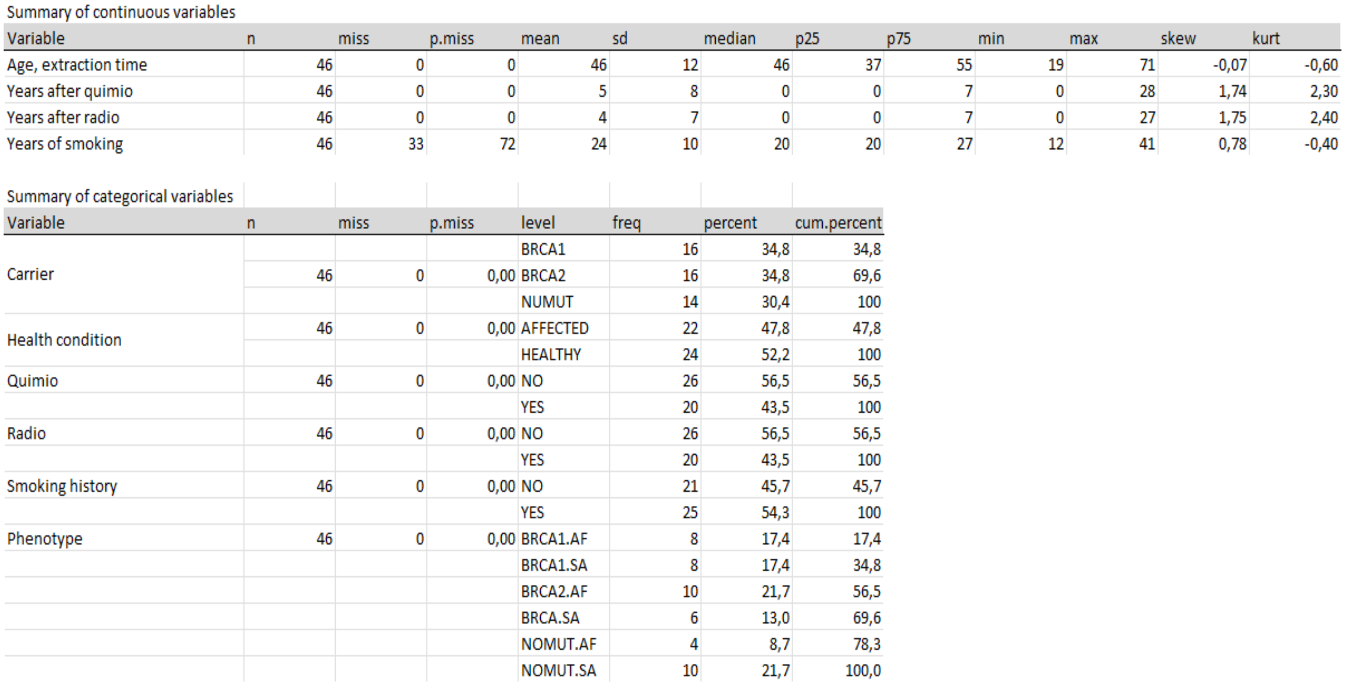
A) *BRCA1* Affected: Age and smoking.
B) *BRCA1* Healthy: no variables selected.
C) *BRCA2* Affected: Age, smoking, chemotherapy, and radiotherapy.
D) *BRCA2* Healthy: age, smoking, and T CD4 cell proportions.

The aim of the study was to finally compare results, on one hand between *BRCA1* healthy and cancer-affected samples, and in the other hand, between *BRCA2* healthy and cancer-affected, without comparing between *BRCA1* and *BRCA2.* For this reason, it is necessary to design consistent DEG models. Thus, the combination of variables selected by LASSO was chosen for both affected and healthy comparisons in each *BRCA1/2* group*.* However, the variables related with cancer treatment were retained only for cancer-affected patients, as they do not have relevance for healthy individuals. This approach led to the inclusion of age and smoking for *BRCA1* DEG models*.* For *BRCA2* models, the variables encompassed age, smoking, and T CD4 for healthy individuals, while also incorporating chemotherapy for cancer-affected individuals. Since chemotherapy and radiotherapy were highly correlated, only the variable of chemotherapy was included in the model because it had a higher LASSO coefficient.

3.5 Differential expression analysis

According to the previous described models, no significant differential expressed genes were found in *BRCA1* healthy, *BRCA1* cancer-affected, and *BRCA2* cancer-affected comparison, as detailed in Table 2. This lack of significance was due to the absence of any gene with an absolute logFC value over 1 and an adjusted p-value below 0.05 and 0.01. Fig.4 shows the expression levels of those genes with a p-value < 0.05 and logFC >1 for the *BRCA1* affected comparison, whilst the rest of the comparisons can be found in Supplementary Figs. S12-S14.
In contrast, when comparing *BRCA2* healthy and NOMUT healthy individuals, only one gene was significantly differentially expressed when using a p-value < 0.05 (Table 2). That gene was identified as *LOC100653061*, a non-specific gene name that lacks universal assignment and has an uncertain function. Given its limited significance and relevance, no specific conclusions were drawn about the gene's potential implications. The distribution of p-values across all *BRCA* comparisons deviated significantly form the expected ideal, which should have shown a right-skewed pattern. In *BRCA2* and *BRCA1* cancer-affected comparison, the distributions appeared to be uniform and left-skewed, respectively, and a little bit right-skewed in *BRCA1* healthy, although it was not considered a good distribution [Supplementary Figs S8-S11].

Table 1. Univariate descriptive analysis.



Summary of continuous variables

Variable	n	miss	p.miss	mean	sd	median	p25	p75	min	max	skew	kurt
Age, extraction time	46	0	0	46	12	46	37	55	19	71	-0,07	-0,60
Years after quimio	46	0	0	5	8	0	0	7	0	28	1,74	2,30
Years after radio	46	0	0	4	7	0	0	7	0	27	1,75	2,40
Years of smoking	46	33	72	24	10	20	20	27	12	41	0,78	-0,40

Summary of categorical variables

Variable	n	miss	p.miss	level	freq	percent	cum.percent
Carrier	46	0	0,00	BRCA1	16	34,8	34,8
				BRCA2	16	34,8	69,6
				NUMUT	14	30,4	100
Health condition	46	0	0,00	AFFECTED	22	47,8	47,8
				HEALTHY	24	52,2	100
Quimio	46	0	0,00	NO	26	56,5	56,5
				YES	20	43,5	100
Radio	46	0	0,00	NO	26	56,5	56,5
				YES	20	43,5	100
Smoking history	46	0	0,00	NO	21	45,7	45,7
				YES	25	54,3	100
Phenotype	46	0	0,00	BRCA1.AF	8	17,4	17,4
				BRCA1.SA	8	17,4	34,8
				BRCA2.AF	10	21,7	56,5
				BRCA.SA	6	13,0	69,6
				NOMUT.AF	4	8,7	78,3
				NOMUT.SA	10	21,7	100,0

Table 2. Number of significant genes for each comparison.

Comparison	Total genes	P-value < 0.05	P-value < 0.01	Adj.P-value <0.05 & logFC>1	Adj.P-value <0.01 & logFC>1
BRCA1 Healthy vs NOMUT Healthy	10961	840	4	0	0
BRCA1 Affected vs NOMUT Affected	10961	157	5	0	0
BRCA2 Healthy vs NOMUT Healthy	10961	527	10	1	0
BRCA2 Affected vs NOMUT Affected	10961	491	40	0	0

3.6 Gene Set Enrichment Analysis

Once the collection of positive and negative pathways was acquired for each phenotype, a comparative analysis within each mutation status became intriguing. Fig 5, illustrates the pathways resulting from the Hallmark gene set collection. Next, the results pertaining to Hallmark are detailed. The positive pathways jointly present in both *BRCA1* healthy and cancer-affected samples (Fig. 5A-B) encompassed *myc targets V1 and 2,* and *E2F targets.* Conversely, the negative pathways that were shared included i*nflammatory response, hypoxia, angiogenesis, epithelial mesenchymal transition, kras signaling up, and coagulation*. There was no overlap of positive pathways between *BRCA1* healthy and negative pathways of *BRCA1* affected samples, and vice versa. Concerning *BRCA2* (Fig. 5C-D), there were no positive pathways commonly shared between healthy and affected samples, neither positive pathways in one condition that were negative in other, and the other way around. Only six negative pathways were shared between *BRCA2* healthy and affected, *inflammatory response, hypoxia, kras signaling up, TNFA signaling via NFKB, interferon gamma response, and complement.*

This analysis was also performed using the Reactome collection, and similar results were obtained.

4. DISCUSSION

In this study, we have examined the transcriptional profile present in human blood samples associated with *BRCA1/2* mutation carriers. A comprehensive bioinformatic analysis was performed, which provide insights into the gene expression profile and biological processes and pathways that are overrepresented in each of the phenotype condition.

As expected, deconvolution results presented an enrichment of T cell CD4 lymphocytes in all the samples, without statistically significant differences between phenotypes. This fact indicates that the proportions of T lymphocytes were similar, independently of the disease status and whether the individual was a *BRCA* carrier or not. T lymphocytes enrichment can be attributed to the administration of PHA to the samples. This substance particularly effectively stimulated the division of T CD4 lymphocytes, leading to a heightened presence of this cell type in the samples. Therefore, the PHA treatment was successful and had a similar impact on each of the samples.

Although MIXTURE is a method that offered the potential for an improved and accurate estimation of cell type proportions, and it is based in CIBERSORT, the existing literature does not establish solid evidence on whether the kind of score it provides allows the comparison between cell type and/or sample type [22]. Then, EPIC tended to overestimate the proportion of T CD4 cell type since it calculates or predicts a higher fraction of T CD4 compared to what might truly be present, demonstrated also in internal studies. For these reasons, CIBERSORT abs. mode was the deconvolution method selected for the analysis.

Results from variable selection with LASSO indicated that the proportion of T CD4 lymphocytes in blood samples did not adjust for the phenotype in three of the four comparisons. Since LASSO selected variables based on their predictive strength, this may suggest that the T CD4 cell fraction is weakly associated with the phenotype. It does not definitively mean that cell type proportions are not related with the phenotype, but it is possible that its relationship is not strong enough to be considered for the model, as LASSO shrinks its coefficient to zero. Therefore, there was no improvement in the DEG given that our analysis unveiled an absence of significant differentially expressed genes. This minimal significance in the genes was already anticipated, in agreement with the aggregation analysis. It has been demonstrated that there isn't a specific transcriptomic profile that uniquely characterizes the phenotype of the individuals.

As differentially expressed genes were obtained taking standard thresholds, that doesn't necessarily mean that there are no regulatory changes occurring within the cell, a GSEA was performed. By focusing on gene sets rather than individual genes, GSEA allows to explore functional pathways or networks that might reveal biological processes

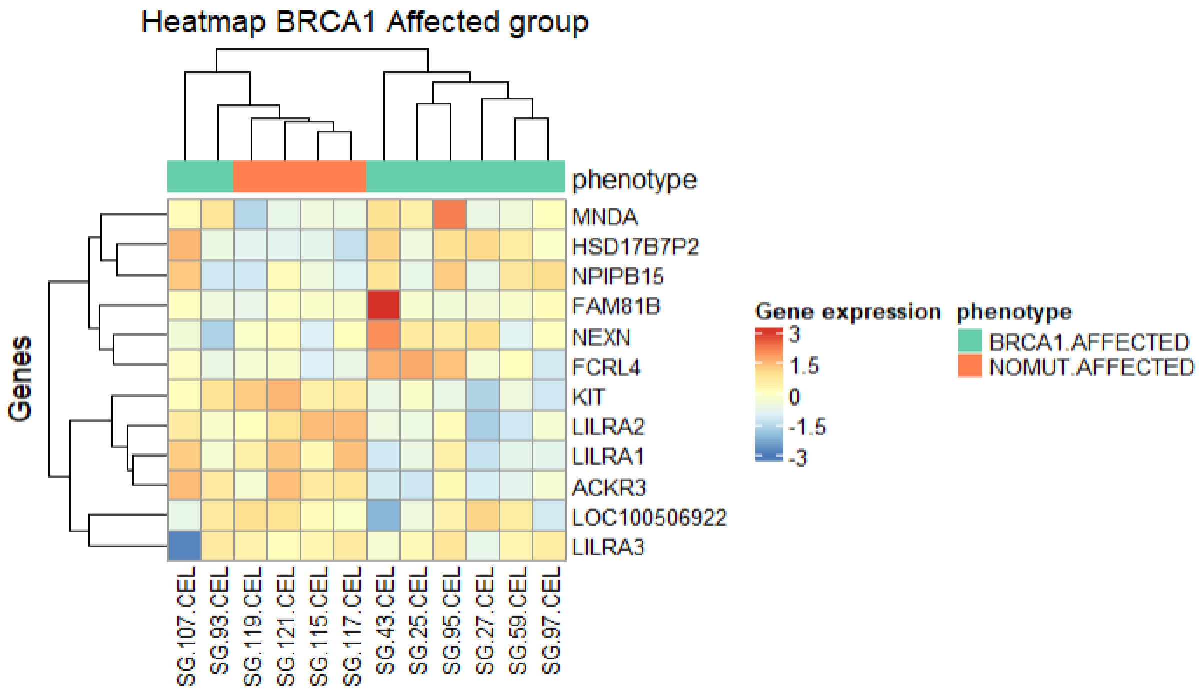


Fig 4. Heatmap of the genes with a p-value < 0.05, and absolute value LogFC>1 for the *BRCA1* affected samples.

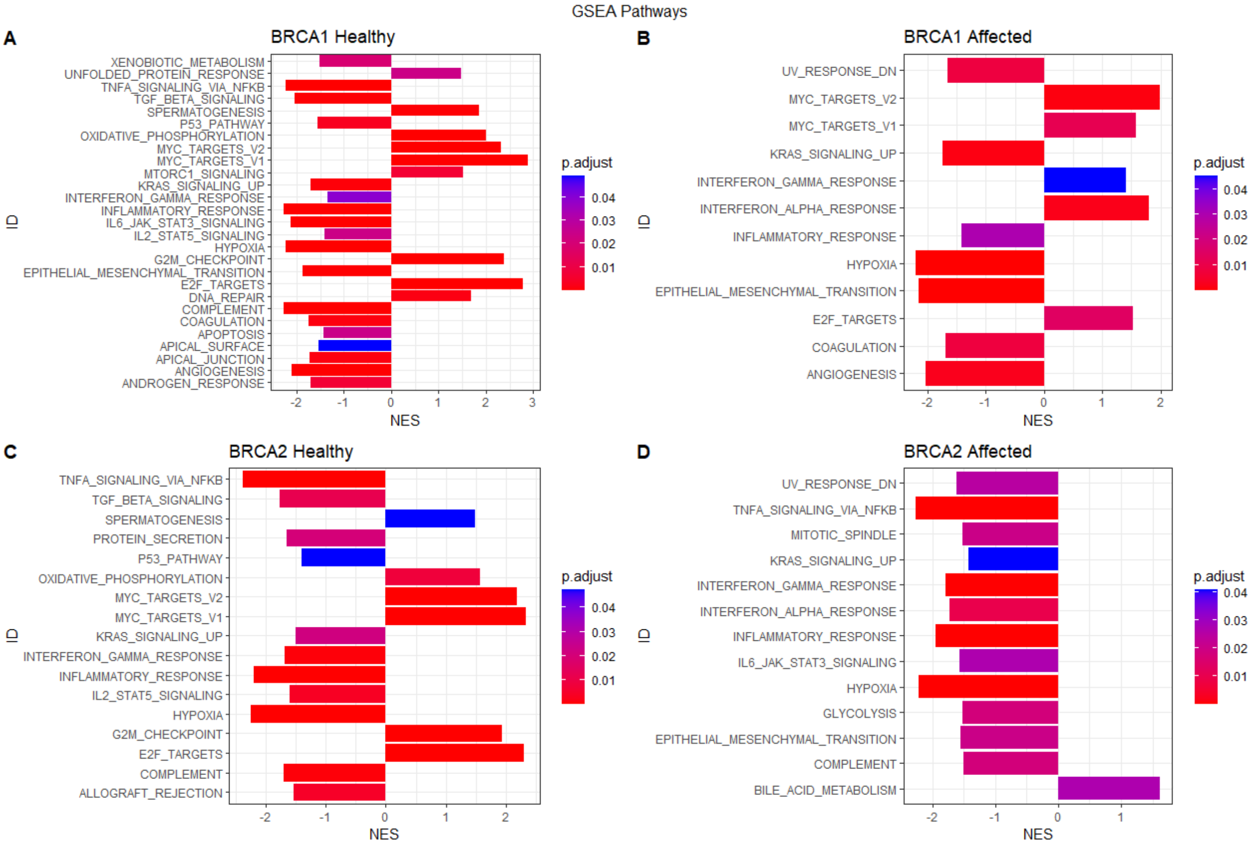


Fig 5. Barplots of the GSEA negative and positive pathways.

that have some relevance to the phenotype condition, even if individual genes don't exhibit statistical significance.

The results of the GSEA showed that cell cycle, DNA replication and proliferation pathways were by far the most positively overrepresented, specifically *MYC* and *E2F* target pathways. Since in the experimental part of the study phytohemagglutinin was used to stimulated lymphocytes, and consequently, facilitate the cell proliferation of the cells, a rapid cell division will be expected, and therefore, observing pathways related to the cell cycle and DNA replication is a key point. Furthermore, the implication of *BRCA1* and *BRCA2* genes in the DNA replication process, demonstrate the importance of observing pathways related to this event.

For *BRCA1* healthy, *BRCA1* affected, and *BRCA2* healthy individuals, both MYC_Targets_V1 and V2, and E2F_Targets were among the highest positive overexpressed pathways, which means that the individuals that have the mutation are showing more expression of *MYC* compared to the control group of them, the non-mutated samples.

That way, most of the overexpressed pathways were related with cell cycle and DNA replication, such as MYC_Target, G2M_Checkpoint, DNA_repair, and E2F_Targets in the *BRCA1* healthy group. In the *BRCA2* affected group, there are also *MYC* and *E2F* pathways, as well as in the *BRCA2* healthy group. However, the results indicated that *BRCA2* affected group, was the most different from the other three, as there was not identified any positively overexpressed pathway related with the previously mentioned.

Regarding pathways, a similar pattern was present in all the different phenotypes. Hypoxia, inflammatory response, interferon gamma (cytokines) and various signaling pathways, were the ones shared in all the groups.

GSEA results obtained were supported by the findings of other studies, which reported an enrichment of *MYC, E2F and NFKB* in lymphocytes treated with PHA [34]. Therefore, there are indeed similarities between the studies, as an enrichment of gene sets specifically associated with key biological pathways such as cell cycle, DNA replication and cell proliferation were obtained. However, in our study, the signaling pathway related to *NFKB* was found to be negatively enriched instead of positively.

E2F involves transcription factors that regulate genes involved in cell cycle progression and DNA replication, and *MYC*, also a transcription factor that is a master regulator of ell growth, proliferation, and metabolism. There are many studies that have been focusing on the study of *MYC* and *E2F* association with breast cancer, even have determined the predictive potential of *E2F* as a biomarker for the disease [35][36]. Even though these pathways are dysregulated and tend to be overexpressed in cancer [37], they are not exclusively to cancer conditions. These genes and their associated pathways are also crucial for normal cells, as they carry out essential functions related to the maintenance of cellular processes. The overexpression of *MYC* and *E2F* target pathways in both healthy and cancer samples raises interesting questions about their roles in different biological contexts and their potential contributions to cancer development. Our results did not show pathways related to cell cycle and proliferation in *BRCA2* cancer affected samples. Therefore, further investigation would be needed to study in detail and understand why these pathways are overexpressed in al the conditions except for cancer-affected *BRCA2* mutation carriers.

Although microarray technology has been a valuable tool in molecular biology for the past years, like any other technique, it has its limitations. Microarrays have a limited dynamic range, which means that the gene expression measurement may not be accurately for low and high expressed genes. Another disadvantage is the ability to detect novel transcripts, as the design of arrays is based on known sequences at the time of the design, so they may not capture unknown sequences [38]. In addition to this, the cost of microarrays can be relatively high compared to other technologies, and the high background noise can reduce the sensitivity and make it challenging to detect changes in gene expression [39].

However, since the field of genomics and transcriptomics continually evolves, novel techniques such as RNA-sequencing have largely replaced microarrays in recent years for many applications due to their improved abilities.

This study has focused on the transcriptome analysis of non-irradiated samples. Hence, results of this study pertain to this subset of samples. In a future, a similar bioinformatics approach could be employed to also analyze the irradiated samples, thus, to identify possible differential expressed genes and biological pathways in this condition.

In summary, a comprehensive and in-depth bioinformatic analysis has been conducted, delving into the complexities of performing a transcriptomics analysis. Even though no results were obtained in the differential expression analysis, they were found in the functional analysis. Moreover, GSEA results were solid and consistent in all the comparisons, as well as coherent within the biological context. Cell cycle and DNA replication related pathways were the key ones identified. Moreover, in further studies, integrating additional omics data, such as genomics and epigenomics, could unveil underlying patterns that are not apparent in the transcriptomics analysis alone.

To perform all the necessary steps to conduct the bioinformatics analysis, a variety of strategies were meticulously employed and evaluated, considering their suitability and relevance within each step of the analysis process. During this progress, we have not only discussed the power of cutting-edge bioinformatic tools, but also demonstrated a critical discernment in their application. The outcome is a robust exploration of the analysis of the transcriptome and gene expression. In this way, a data set has been re-analysed by using innovative techniques, that could potentially lead to novel discoveries and insights of the biological processes under investigation.

SUPPLEMENTARY INFORMATION

Fig. S1-S2. Array images generated by the scan.
Fig. S3. Dendrogram of the disease status, using the Euclidean distance and the linkage ward.D2 method.
Fig. S4. Dendrogram of the disease status, using the correlation-based distance and the average linkage method.
Fig. S5-S7. Correlation plots.
Fig. S8-S11. Plots representing the distribution of the p-values generated in the differential expression analysis.
Fig. S12-14. Heatmap of the most significative genes for each of the comparisons.

REFERENCES

[1] "Cancer Today." https://gco.iarc.fr/today/online-analysis-multi-bars?v=2020&mode=cancer&mode_population=countries&population=900&populations=900&key=total&sex=2&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=10&group_cancer=1&include_nmsc=0&include_nmsc_other=1&type_multiple=%257B%2522inc%2522%253Atrue%252C%2522mort%2522%253Afalse%252C%2522prev%2522%253Afalse%257D&orientation=horizontal&type_sort=0&type_nb_items=%257B%2522top%2522%253Atrue%252C%2522bottom%2522%253Afalse%257D (accessed Aug. 12, 2023).

[2] "BRCA Gene Mutations: Cancer Risk and Genetic Testing Fact Sheet - NCI." https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet (accessed Aug. 11, 2023).

[3] S. A. Narod and L. Salmena, "BRCA1 and BRCA2 Mutations and Breast Cancer," *Discov. Med.*, vol. 12, no. 66, pp. 445–453, Nov. 2011.

[4] E. R. Jang and J.-S. Lee, "DNA Damage Response Mediated through BRCA1," *Cancer Res. Treat.*, vol. 36, no. 4, p. 214, 2004, doi: 10.4143/CRT.2004.36.4.214.

[5] G. Borrego-Soto, R. Ortiz-López, and A. Rojas-Martínez, "Ionizing radiation-induced DNA injury and damage detection in patientswith breast cancer," *Genet. Mol. Biol.*, vol. 38, no. 4, p. 420, Oct. 2015, doi: 10.1590/S1415-475738420150019.

[6] M. Saleem *et al.*, "The BRCA1 and BRCA2 Genes in Early-Onset Breast Cancer Patients," *Adv. Exp. Med. Biol.*, vol. 1292, pp. 1–12, 2020, doi: 10.1007/5584_2018_147.

[7] K. Yoshida and Y. Miki, "Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage," *Cancer Sci.*, vol. 95, no. 11, pp. 866–871, Nov. 2004, doi: 10.1111/J.1349-7006.2004.TB02195.X.

[8] S. Supplitt, P. Karpinski, M. Sasiadek, and I. Laczmanska, "Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine," *Int. J. Mol. Sci.*, vol. 22, no. 3, pp. 1–22, Feb. 2021, doi: 10.3390/IJMS22031422.

[9] C. Virtanen and J. Woodgett, "Clinical Uses of Microarrays in Cancer Research," Accessed: Aug. 14, 2023. [Online]. Available: www.affymetrix.com.

[10] F. Avila Cobos, J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh, and K. De Preter, "Benchmarking of cell type deconvolution pipelines for transcriptomics data," *Nat. Commun.*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/S41467-020-19015-1.

[11] "GSEA." https://www.gsea-msigdb.org/gsea/index.jsp (accessed Aug. 14, 2023).

[12] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG.

[13] S. Gutiérrez-Enríquez *et al.*, "Ionizing radiation or mitomycin-induced micronuclei in lymphocytes of BRCA1 or BRCA2 mutation carriers," *Breast Cancer Res. Treat.*, vol. 127, no. 3, pp. 611–622, Jun. 2011, doi: 10.1007/S10549-010-1017-6/METRICS.

[14] "GeneChip ® Human Genome U133A 2.0 Array," Accessed: Aug. 12, 2023. [Online]. Available: www.affymetrix.com.

[15] "Bioconductor - affy." https://bioconductor.org/packages/release/bioc/html/affy.html (accessed Aug. 14, 2023).

[16] L. Gautier, L. Cope, B. M. Bolstad, and R. A.

Irizarry, “Affy - Analysis of Affymetrix GeneChip data at the probe level,” *Bioinformatics*, vol. 20, no. 3, pp. 307–315, Feb. 2004, doi: 10.1093/BIOINFORMATICS/BTG405.

[17] “R: The R Stats Package.” https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html (accessed Aug. 30, 2023).

[18] “dendextend package - RDocumentation.” https://www.rdocumentation.org/packages/dendextend/versions/1.17.1 (accessed Aug. 30, 2023).

[19] “Bioconductor - hgu133plus2.db.” https://bioconductor.org/packages/release/data/annotation/html/hgu133plus2.db.html (accessed Aug. 14, 2023).

[20] “GeneChip^{TM} Human Genome U133 Plus 2.0 Array.” https://www.thermofisher.com/order/catalog/product/900466 (accessed Aug. 14, 2023).

[21] “GitHub - omnideconv/immunedeconv: A unified interface to immune deconvolution methods (CIBERSORT, EPIC, quanTIseq, TIMER, xCell, MCPcounter) and mouse deconvolution methods.” https://github.com/omnideconv/immunedeconv (accessed Aug. 14, 2023).

[22] E. A. Fernández *et al.*, “MIXTURE: an improved algorithm for immune tumor microenvironment estimation based on gene expression data,” *bioRxiv*, p. 726562, Aug. 2019, doi: 10.1101/726562.

[23] “Package ‘glmnet’ Type Package Title Lasso and Elastic-Net Regularized Generalized Linear Models,” 2023, doi: 10.18637/jss.v033.i01.

[24] “Bioconductor - limma.” https://bioconductor.org/packages/release/bioc/html/limma.html (accessed Aug. 14, 2023).

[25] T. Wu *et al.*, “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data,” *Innovation*, vol. 2, no. 3, Aug. 2021, doi: 10.1016/J.XINN.2021.100141.

[26] M. Dozmorov, “Quality assessment, single channel (Affymetrix) arrays Probe level QC,” 2017, Accessed: Aug. 12, 2023. [Online]. Available: http://plmimagegallery.bmbolstad.com/.

[27] N. Lara, “Transcriptomics. Microarray Data Analysis.”

[28] “Getting started with immunedeconv • immunedeconv.” https://omnideconv.org/immunedeconv/articles/immunedeconv.html (accessed Aug. 14, 2023).

[29] “CIBERSORTx.” https://cibersortx.stanford.edu/ (accessed Aug. 23, 2023).

[30] “Least Absolute Shrinkage and Selection Operator (LASSO) | Columbia University Mailman School of Public Health.” https://www.publichealth.columbia.edu/research/population-health-methods/least-absolute-shrinkage-and-selection-operator-lasso (accessed Aug. 14, 2023).

[31] M. E. Ritchie *et al.*, “Limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, Jan. 2015, doi: 10.1093/NAR/GKV007.

[32] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/J.2517-6161.1995.TB02031.X.

[33] “GSEA | MSigDB.” https://www.gsea-msigdb.org/gsea/msigdb (accessed Aug. 14, 2023).

[34] C. Beinke, M. Port, R. Ullmann, K. Gilbertz, M. Majewski, and M. Abend, “Analysis of Gene Expression Changes in PHA-M Stimulated Lymphocytes - Unraveling PHA Activity as Prerequisite for Dicentric Chromosome Analysis,” *Radiat. Res.*, vol. 189, no. 6, pp. 579–596, Jun. 2018, doi: 10.1667/RR14974.1.

[35] Y. Chen and O. I. Olopade, “MYC in breast tumor progression,” *Expert Rev. Anticancer Ther.*, vol. 8, no. 10, p. 1689, 2008, doi: 10.1586/14737140.8.10.1689.

[36] M. Oshi *et al.*, “The E2F Pathway Score as a Predictive Biomarker of Response to Neoadjuvant Therapy in ER+/HER2− Breast Cancer,” *Cells*, vol. 9, no. 7, Jul. 2020, doi: 10.3390/CELLS9071643.

[37] Y. Li *et al.*, “Expression patterns of E2F transcription factors and their potential prognostic roles in breast cancer,” *Oncol. Lett.*, vol. 15, no. 6, pp. 9216–9230, Jun. 2018, doi: 10.3892/OL.2018.8514/HTML.

[38] “RNA-Seq vs Microarrays | Compare technologies.” https://www.illumina.com/science/technology/next-generation-sequencing/microarray-rna-seq-comparison.html (accessed Sep. 01, 2023).

[39] R. Jaksik, M. Iwanaszko, J. Rzeszowska-Wolny, and M. Kimmel, “Microarray experiments and factors which affect their reliability,” *Biol. Direct*, vol. 10, no. 1, Sep. 2015, doi: 10.1186/S13062-015-0077-2.