**TRABAJO FIN DE Máster**

**Máster en Bioinformática**

**Facultad de Biología - Universidad de Murcia**

**Dpto. Oncología Clínica - Universidad Médica de Graz**

# Cross-omics regulation of gene expression in breast cancer

Autor: **Vladimir Estal Daries**

Tutor Universidad Murcia: **Manuel Franco Nicolas**

Tutor externo (Universidad Médica de Graz): **Sebastian Vosberg**

CURSO 2023-24

# INDEX


ABSTRACT ..... 1
RESUMEN ..... 1
INTRODUCTION ..... 2
MATERIALS AND METHODS ..... 3
DATA DOWNLOAD ..... 3
SAMPLES AND CLINICAL DATA ..... 3
GENES SELECTION ..... 3
DATA PROCESSING ..... 3
RESULTS ..... 6
INITIAL GENE SELECTION ..... 6
COPY NUMBER ..... 7
METHYLATION ..... 9
RESULT OF GENE INTERACTIONS ..... 9
DISCUSSION ..... 12
INITIAL GENE SELECTION ..... 12
COPY NUMBER ..... 13
METHYLATION ..... 13
GENE INTERACTIONS ..... 13
Groups by MUTg & Methylation by SDg ..... 13
Groups by MUTg & Methylation by MUTg ..... 14
Groups by CNg & Methylation by SDg ..... 14
Groups by CNg & Methylation by CNg ..... 15
CONCLUSIONS ..... 15
REFERENCES ..... 16

## ABSTRACT

Breast cancer is one of the most deadly forms of cancer and is the most common in women. This high prevalence and mortality rate highlight the urgent need for a deeper understanding of its underlying mechanisms.

Most studies typically consider gene expression in conjunction with either methylation, mutations, or copy number variations independently. This fragmented approach limits our understanding of the complex interactions and regulatory networks involved in breast cancer, so multifactorial research that integrates these factors is necessary to provide a more comprehensive and accurate picture.

To identify genes that can be regulated by DNA methylation within a specific genetic context, influenced by the effect of another gene, a dataset of breast cancer primary tumors was studied. This work used the integration of genetic mutation profiles, copy number variation, DNA methylation, and gene expression profiles through a systematic approach.

The results showed numerous gene interactions with high statistical significance in both genes that play critical roles in cancer (TP53, PIK3CA, GATA3, CDH1, MAP3K1, FOXA1, BRCA1, PGR, IDO1) and others not previously documented, uncovering interactions between genetic and epigenetic factors that have the potential to regulate gene expression.

This systematic approach could facilitate the discovery of interactions between cross-omics alterations and improve our understanding of molecular tumor biology.

**Keywords**: Breast cancer, methylation, mutations, copy number variation, epigenetics, tumor environment, gene expression, hypermethylation.

## RESUMEN

El cáncer de mama es una de las formas de cáncer más mortales y es el más común en las mujeres. Esta alta prevalencia y tasa de mortalidad resaltan la necesidad urgente de una comprensión más profunda de sus mecanismos subyacentes.

La mayoría de los estudios suelen considerar la expresión genética junto con la metilación, las mutaciones o las variaciones del número de copias de forma independiente. Este enfoque fragmentado limita nuestra comprensión de las complejas interacciones y redes regulatorias involucradas en el cáncer de mama, por lo que es necesaria una investigación multifactorial que integre estos factores para proporcionar una imagen más completa y precisa.

Para identificar genes que pueden regularse mediante la metilación del ADN dentro de un contexto genético específico, influenciado por el efecto de otro gen, se estudió un conjunto de datos de tumores primarios de cáncer de mama. Este trabajo utilizó la integración de perfiles de mutación genética, variación del número de copias, metilación del ADN y perfiles de expresión génica mediante un enfoque sistemático.

Los resultados mostraron numerosas interacciones genéticas con alta significancia estadística en ambos genes que desempeñan papeles críticos en el cáncer (TP53, PIK3CA, GATA3, CDH1, MAP3K1, FOXA1, BRCA1, PGR, IDO1) y otros no documentados previamente, descubriendo interacciones entre factores genéticos y epigenéticos que tienen el potencial de regular la expresión genética.

Este enfoque sistemático podría facilitar el descubrimiento de interacciones génicas y mejorar la comprensión de la biología molecular de los tumores.

**Palabras clave**: Cáncer de mama, metilación, mutaciones, variación en el número de copias, epigenética, entorno tumoral, expresión génica, hipermetilación.

## INTRODUCTION

Breast cancer (**BC**) is one of the most common types of cancer worldwide, being the most common in women in both developed and developing countries and represents around 25% of all cancer cases in women globally, that is, approximately 1 in 4 cases of cancer diagnosed in women is BC, being one of the most deadly cancer types especially in patients with advanced disease [1].

Taking both sexes into account globally, 1 in 8 cancer diagnoses correspond to BC, meaning a total of 2.3 million cases in 2020, with an estimated 685,000 deaths of women from this type of cancer [2]. These data represent 16% of the cases, meaning that 1 in 6 women who suffer from this disease die from BC.Due to the resistance to treatment that this type of cancer can present, especially in patients with advanced disease, its elimination is a challenge that is usually addressed with a treatment that varies depending on the classification of the tumor according to its genetic characteristics [3].

Human tumors are caused by DNA mutations and epigenetic alterations like DNA methylation[4]. Epigenetic modifications can affect the activity of a gene without altering the DNA sequence. DNA methylation, one of the most studied epigenetic modifications, has been shown to be of great importance in the regulation of gene expression[5]. This consists of the addition of methyl groups to the sequence of DNA molecules and can influence gene transcription and, ultimately, cellular function and behavior.

Epigenetic modifications can activate or deactivate gene expression partially or completely, both directly and indirectly, playing a crucial role in a multitude of biological processes, including the development of cancer [6].

In BC, hypermethylation in certain genes can lead to their silencing, which could contribute to cancer development. An example would be hypermethylation of the tumor suppressor gene BRCA1, which can deactivate this gene and promote tumor development [7].

The alteration in copy number (**CN**) can also have an effect on the level of gene expression, affecting the expression of a gene directly, which in turn could downstream regulate the expression of other genes [8].

Both genetic and epigenetic alterations can have effects on the regulation of gene expression, but the promoting role of mutations in a certain epigenetic background is still unclear [9]. Studies have identified abnormal methylation patterns in breast tumors, and some have suggested that these changes may be related to cancer development and progression [10]. Although numerous studies have been conducted to correlate DNA methylation with gene expression in BC, the conclusions obtained have often been inconsistent. Furthermore, these studies generally only take into account gene expression together with methylation, mutations or CN independently, so multifactorial research is necessary in this area.

A better understanding of this background would allow the identification of clinically relevant biomarkers that would ultimately serve to obtain a deeper insight into the molecular mechanisms, with the potential to reveal new therapeutic targets. The discovery of interactions between cross-omics alterations would improve understanding of molecular tumor biology.

To evaluate the interactions between genetic and epigenetic events, this work use the integration of genetic mutation profiles, CN variation, DNA methylation and gene expression profiles of BC patients, using a systematic approach.

The goal is to identify genes that can be regulated by DNA methylation within a specific genetic context, such as the mutation or CN variation of another gene. This means uncovering interactions between genetic and epigenetic that have the potential to regulate gene expression in BC.

A more comprehensive understanding of these processes could lead to new avenues for targeted therapies and preventive strategies in BC.

## MATERIALS AND METHODS

The entire project has been developed using Rstudio. Codes used for downloading, data processing and obtaining results can be consulted at the GitHub link https://github.com/vlaesda/Master-Tesis-Vladimir. *Figure 1* represents the processing of the different data sets.

### DATA DOWNLOAD

Expression, clinical data and mutation files were downloaded from http://firebrowse.org/ (TCGA data version 2016_01_28 for BRCA). Expression data were obtained using mRNAseq.

Methylation positions or CpG sites (**CpG**), were obtained using "IlluminaHumanMethylation450kmanifest" package (CpG site is a DNA region where a cytosine is followed by a guanine. DNA methylation often occurs in CpG sites). This package is part of the Bioconductor project and contains detailed information which allows establishing the CpG corresponding to each gene. CN values and the methylation status or CpG values of each patient, were obtained using "BiocManager, library (TCGAbiolinks)".

### SAMPLES AND CLINICAL DATA

The expression data, clinical data and mutation files contained data for a total of 977 patients. For CN values, data were obtained for 1064 patients and 893 for methylation.

### GENES SELECTION

Expression: The $\log_2$ transformation was used to standardise the variance of the gene expression data. A selection of genes was carried out based on those with 90% of patients with expression >2 and with a standard deviation >2 (**SDg**). Not all genes were included because it would have led to over-correction by multiple testing and represented a significant computational burden. Only genes that showed consistent changes between patients, i.e. provided relevant information, were selected. Genes with limited information and variations were excluded.

Mutations: A selection was carried out of those mutations affecting open reading frame, that is, those that may have significant consequences on the function of the protein encoded by the gene. For this selection, only samples from primary tumor tissue were used, discarding those from normal or metastatic tissue. To determine the selection of genes that have mutations, those that had at least 20 patients with mutations for that gene were selected (**MUTg**). This selection was made with the aim of using only the common variants. Furthermore, this ensures that the subgroups that are defined within the mutated cases in the statistical analysis still have a sufficient number of cases.

Copy Number: Only genes that coded for proteins were selected with at least 20 patients with more than 4 CN for that gene, that is, those who had 5,6 or 7 CN (**CNH**). This selection allowed to focus on NACs that can be considered to have a strong effect on gene expression. These genes were also filtered based on a result for the T-test with p-value <0.5 between patients 2CN vs CNH, and foldchanges with a value >=Log2(CN/2*0.8). This foldchanges value represents 80% of the expected value for each CN (5,6,7), assuming that the expression will increase along with the increase in the CN value. The 0.8 threshold was chosen to allow for some variability or noise in expression values.

Finally, filtering was performed selecting those genes with at least a correlation of 0.4 between the CN and the gene expression of the same gene (**CNg**).

### DATA PROCESSING

To determine which genes (CNg or MUTg) statistically significantly affect the expression of SDg, T-test and foldchange were obtained in each gen selection. *Figure 1* represents the processing of the different data sets.

For the MUTg, T-test was performed for each gene, comparing the expression of each SDg between patients with mutations for MUTg and the rest of patients considered wild type (**WT**). The foldchanges value of these groups was also obtained. Following this methodology, the same values were obtained with the CNg - SDg, forming the groups of patients to be compared according to the 2CN value (two copies of the gene) vs CNH.

In this case, before performing the multiple T-test, gene combinations (CNg – SDg comparison) whose genes were located on the same chromosome were eliminated. Through this selection of combinations, we avoid the genes that we would obtain in our analysis if both were duplicated on the same chromosome, because duplications occur in chromosomal segments that may contain multiple

genes. This co-duplication could increase the expression of SDg together with increasing the CN of the CNg. Furthermore, in this way we achieve a more precise level of significance in the Benjamini-Hochberg correction method, since it uses a correction based on the number of tests performed and the number of significant discoveries expected.

For the multiple T-tests of MUTg and CNg, Benjamini-Hochberg correction was applied to the results obtained. This procedure is applied by performing many simultaneous statistical tests, adjusting p-values to reduce false positives and balance significant discovery by error control.

After Benjamini-Hochberg correction, gene combinations with p-value >0.1 and/or foldchanges between -Log2(1.5) and Log2(1.5) were removed. Using foldchanges as a selection criterion allows a more strict selection based on quantifiable differences in the gene expression of the groups compared in each case.

The objective is to determine, with the fold change, if the significant p-value of these groups reflects a real difference. A drawback of using foldchange in this environment is that it can miss genes expressed with large differences but small ratios, leading to a high failure rate at high expression. To reduce this problem, expression data were standardised using Log2. It solves this problem in two main ways: On the one hand, it approximates the data to a normal distribution, stabilising the variance and reducing the bias. It simplifies the interpretation of the changes by converting them into multiples (fold changes), so that the regulation is expressed as positive or negative values, respectively, without the need to explicitly consider the direction of change.

CpG means of each gene/patient made it possible to establish 3 methylation levels:

low methylation : CpG averages < 0.33 / medium methylation : 0.33 < CpG averages < 0.66 / high methylation : CpG averages > 0.66.

These methylation levels were incorporated into the analysis of the T-tests and foldchanges, to form the groups of patients with CNg and MUTg.

Three levels of patients were formed for each gene according to the methylation level of each patient. In the case of combinations of MUTg, 3 groups were obtained according to the methylation level for patients with mutated genes, and another 3 groups according to methylation for WT patients.

The same was done with the groups according to CNg. In both gene selections, gene clusters were classified according to the methylation levels of the SDg as well as the methylation levels of CNg and MUTg respectively. In this way 4 sets of data were obtained:

- Groups by **MUTg** & methylation levels
- Methylation by **SDg**
- Expression by **SDg**

- Groups by **MUTg** & methylation levels
- Methylation by **MUTg**
- Expression by **SDg**

- Groups by **CNg** & methylation levels
- Methylation by **SDg**
- Expression by **SDg**

- Groups by **CNg** & methylation levels
- Methylation by **CNg**
- Expression by **SDg**

The results obtained were filtered according to the genes combination with one group foldchange variation < 75% and pval < 0.05 and other group foldchange variation < 50% + pval > 0.2.

This selection allows us to establish whether the differences between groups of the same methylation level can be caused by the effect of the CNg and MUTg respectively, or if it is due to methylation.

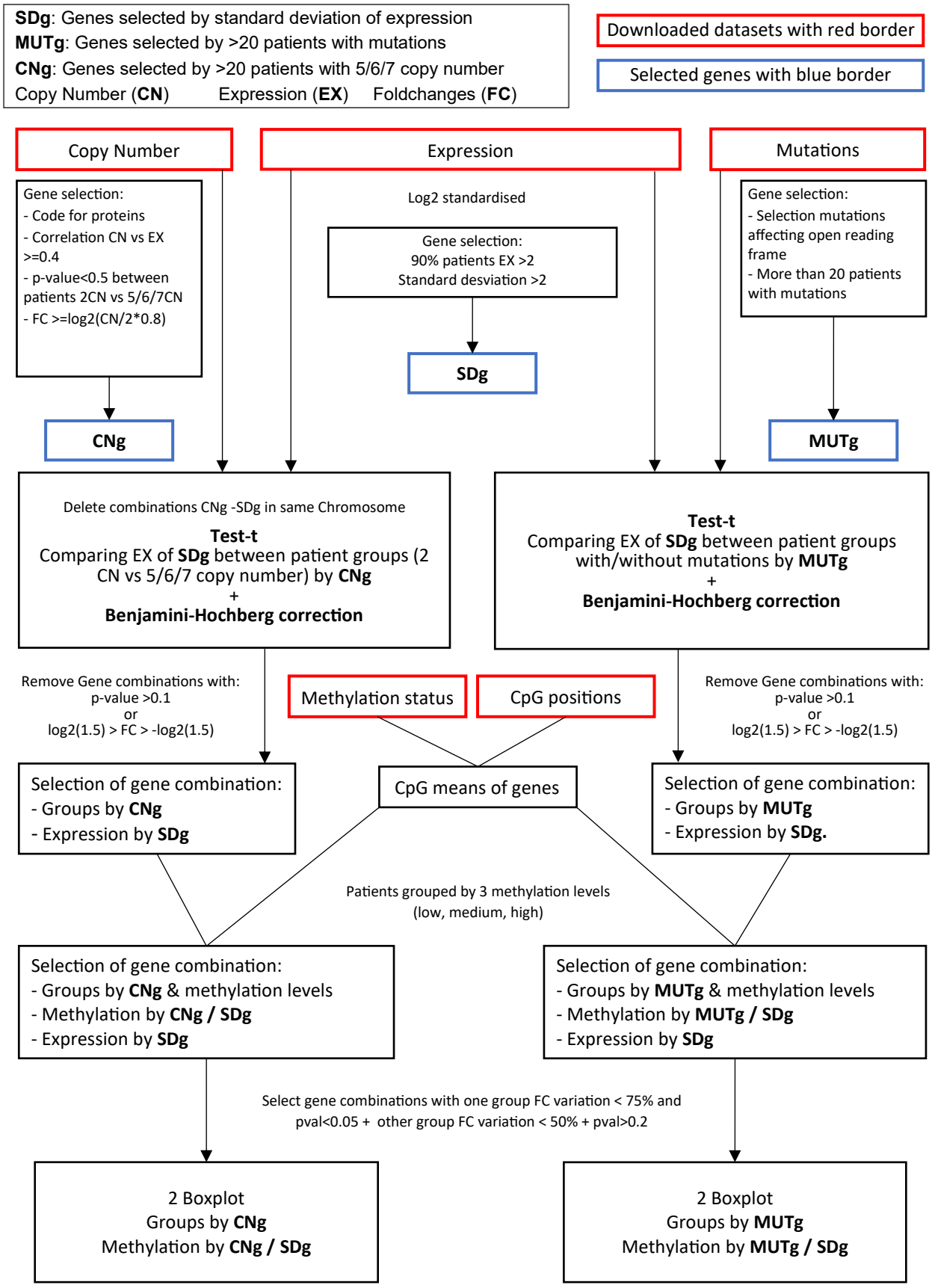


**Figure 1**. Classification of genes according to the standard deviation, the number of patients with mutations for a given gene, and the number of patients with 5/6/7 copies for a given gene. After making the relevant corrections using Benjamini-Hochberg, methylation data have been integrated to allow patients to be classified based on this. Subsequently, the filtering of the data and obtaining the results represented in several boxplots are shown.

## RESULTS

INITIAL GENE SELECTION

In the initial selection of genes according to the different criteria (see Materials and Methods), were selected n=497 for SDg, n=156 for MUTg and n=165 for CNg. T-testing of the patient groups (mutations vs WT) by MUTg with expression by SDg, and subsequent Benjamini Hochberg correction, resulted in a matrix with p-values comparing the groups. For these groups, the Log2(foldchanges) were also calculated, obtaining the matrix with which the heatmap in *Figure 2* was elaborated.

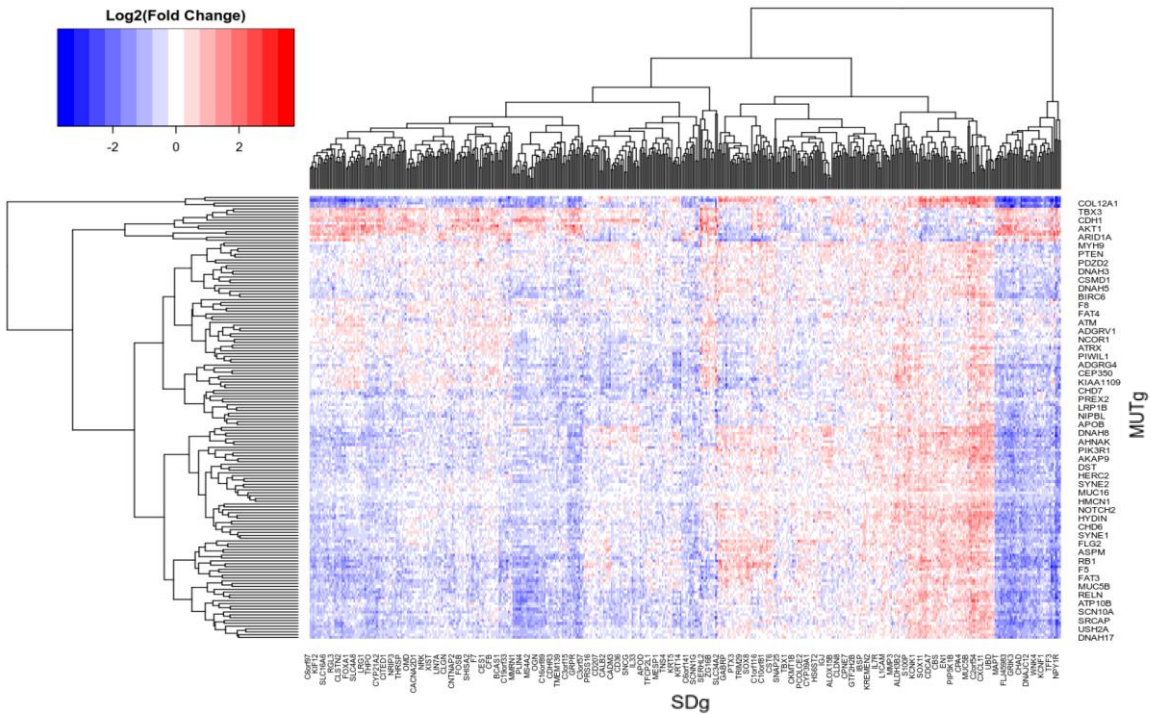


**Figure 2.** Heatmap representing the Log2(foldchanges) values for the patient groups formed by the MUTg (patients with mutations vs WT patients) and the expression of the SDg.

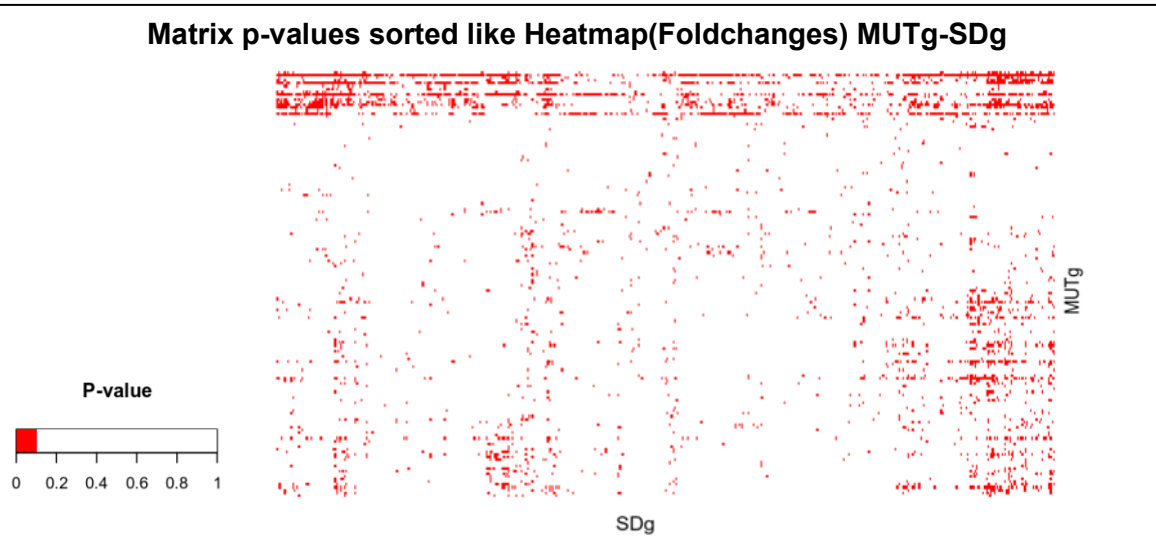


**Figure 3.** Matrix with p-values <0.1 marked in red and genes arranged in the same order as the genes in the Heatmap with Log2(foldchanges). The groups compared by T-test are formed by patients with mutations vs. WT patients for MUTg. The expression of the SDg was used. The arrangement of the significant p-values coincides with the position of the clusters established in the heatmap of the foldchanges of the *Figure 2*.

The genes of the matrix with the p-values were arranged in the same order as the genes of the Heatmap, resulting in the matrix of *Figure 3*. With both matrices, the Heatmap of *Figure 4*, was created, in which they are represented only the results for gene combinations with p-value <0.1 and foldchange out of range -Log2(1.5) and Log2(1.5). On this matrix, a selection of the genes that met both requirements was carried out, obtaining 152 mutated genes and 490 transcribed genes.

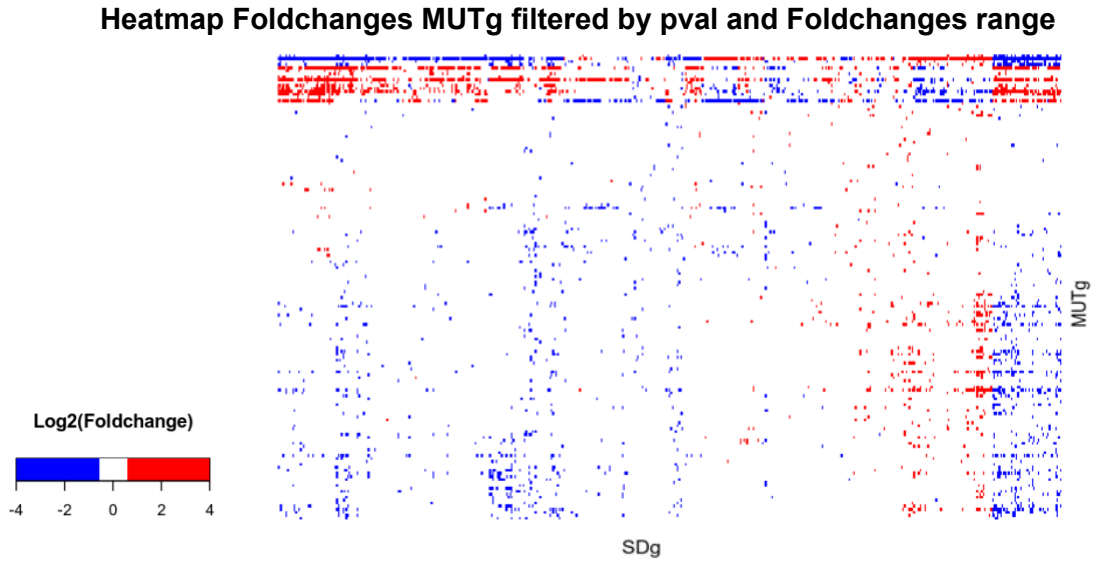


**Figure 4.** Selection of values from the heatmap Log2(foldchange) in *Figure 2*, according to the significance of the p-values <0.1 in *Figure 3*. The gene combinations were selected for presenting a foldchange out of range -Log2(1.5) and Log2( 1.5). These parameters allow us to select only those groups that have a real and quantitative difference in gene expression.

COPY NUMBER

After downloading and processing the copy number data per gene-patient, 11,795 protein-coding genes that had more than two copies in at least 20 patients were obtained.

*Figure 5* represents the genes ordered according to the number of patients for each gene. To the right of the figure, in red, you can see a set of genes forming a group that is characterized by having the majority of patients with more than two copies.

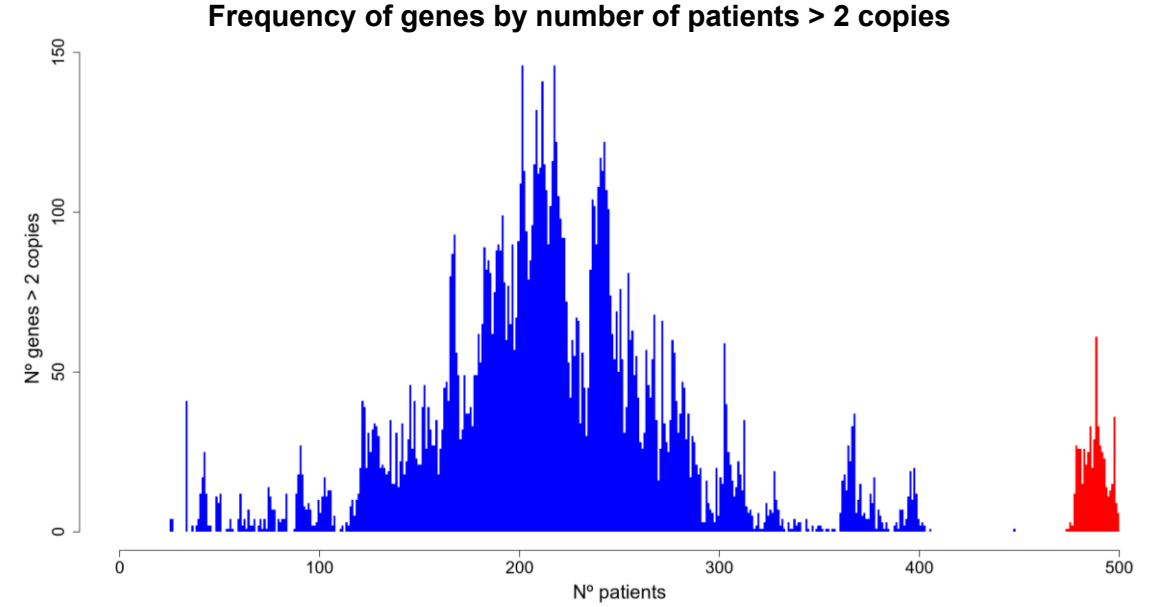


**Figure 5.** Histogram representing the 11795 protein-coding genes with more than 20 patients with more than two copies. Each bar represents a set of genes with a given number of patients. On the right, in red, a set of 830 genes with at least 474 patients. This group of genes belongs to chromosome 1.

To determine which genes formed this group, all the genes with more than 473 patients were selected (number of patients that delimits the group), giving rise to 830 genes, all of them being found on chromosome 1, specifically between positions 150149183 – 247622372 corresponding to the 1q end (long arm of chromosome 1).

In *Figure 6*, the expression of the 11.795 genes that code for proteins is represented, with the patients organized according to the number of copies. Of these genes, the 165 CNg were selected, as described in the Materials and Methods section. With these, *Figure 7* was created, representing the joint expression of the CNg according to their CN.

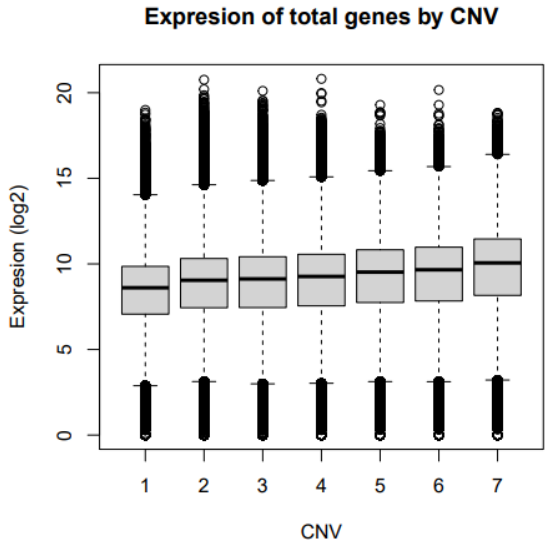


**Figure 6.** Expression of the patients according to the CN for all of the 11.795 genes that coded for proteins.

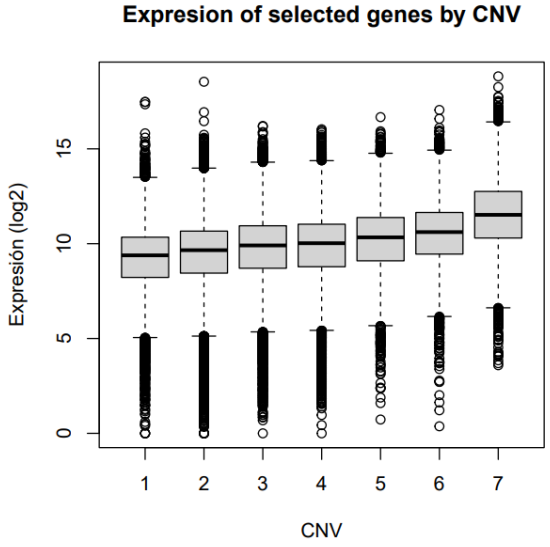


**Figure 7.** Expression of the patients according to the CN for the selection of CNg. A greater increase in gene expression is observed as the CN of the genes increases in the case of CNg.

In *Appendix 1*, for each of the CN genes, the expression according to CN is shown. A sample of the results for 3 CNg is shown in *Figure 8*.

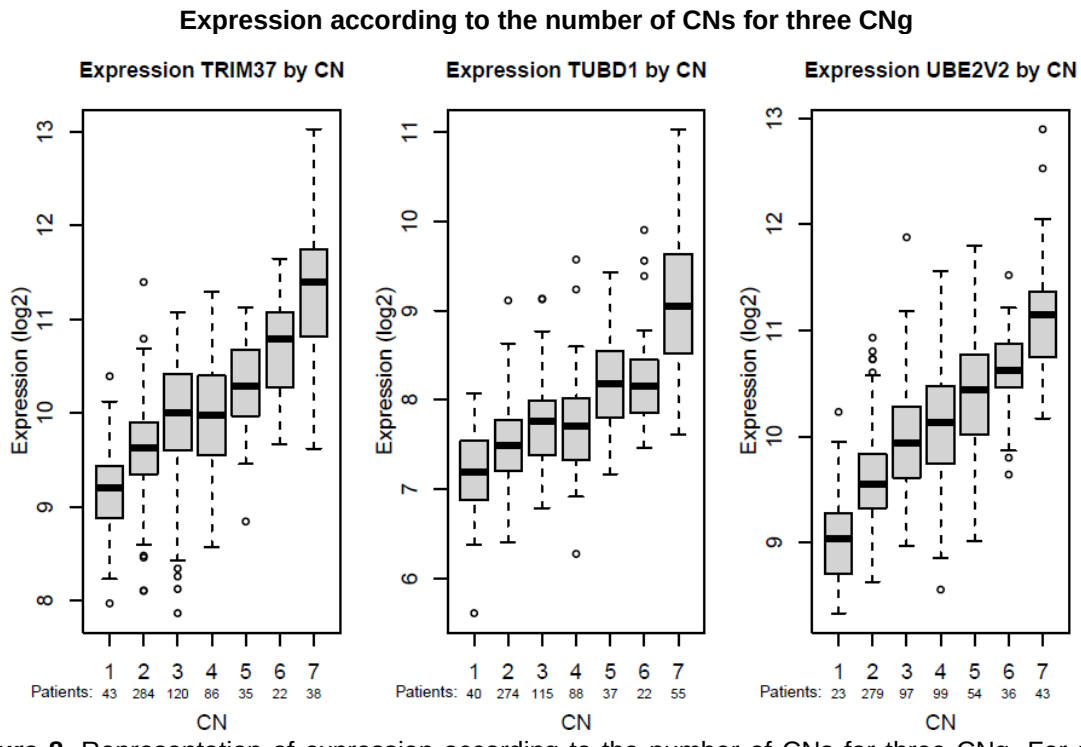


**Figure 8.** Representation of expression according to the number of CNs for three CNg. For most genes *(see Appendix 1)* an increase in expression is observed with increasing copy number. In many cases there is a higher increase in expression at 7CN.

METHYLATION

With the three selected gene groups (MUTg, SDg, CNg), the CpG means of gene-patients were calculated and are represented in the density plot in *Figure 9* on the left for SDg, and on the right for MUTg. Due to the distribution of values, patients were divided into three methylation levels.

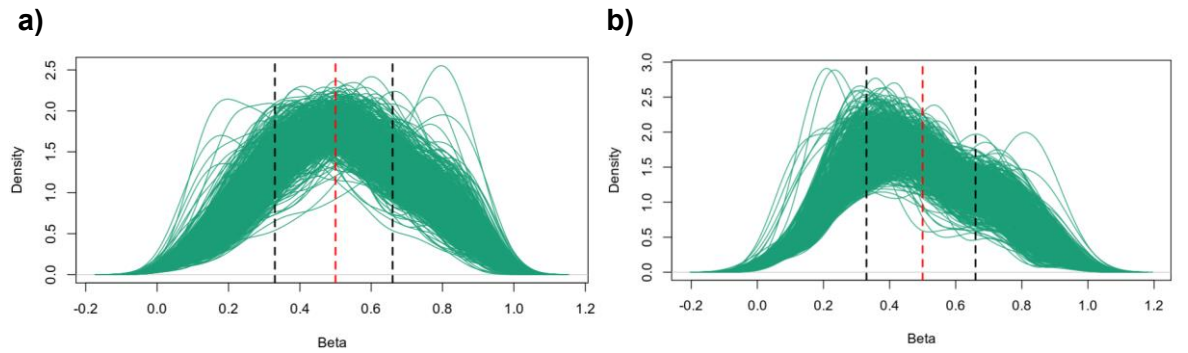


**Figure 9.** Density plot for SDg (**a**) and for MUTg (**b**). Each line represents the density of a patient's methylation averages showing which methylation values are most common. The X-axis represents gene methylation values by Beta value, a quantitative measure reflecting the level of DNA methylation, with values ranging from 0 (unmethylated) to 1 (fully methylated). Values outside this range are due to experimental noise. The Y-axis represents the density of the methylation averages. The vertical red line represents the beta methylation value of 0.5. Patients were divided into three methylation levels marked with black dashed vertical lines, being low for patients with a Beta value <0.33, high for patients with a Beta value >0.66 and medium for those with values between 0.33 and 0.66.

RESULT OF GENE INTERACTIONS

After classification of patients/genes into three methylation levels and obtaining the respective CNg, MUTg and SDg, gene combinations were processed and selected (see Materials and Methods). For each of the gene selections there is an annex with the results in the corresponding section. Some results are shared here:

Groups by MUTg & Methylation by SDg *(Appendix 2)*

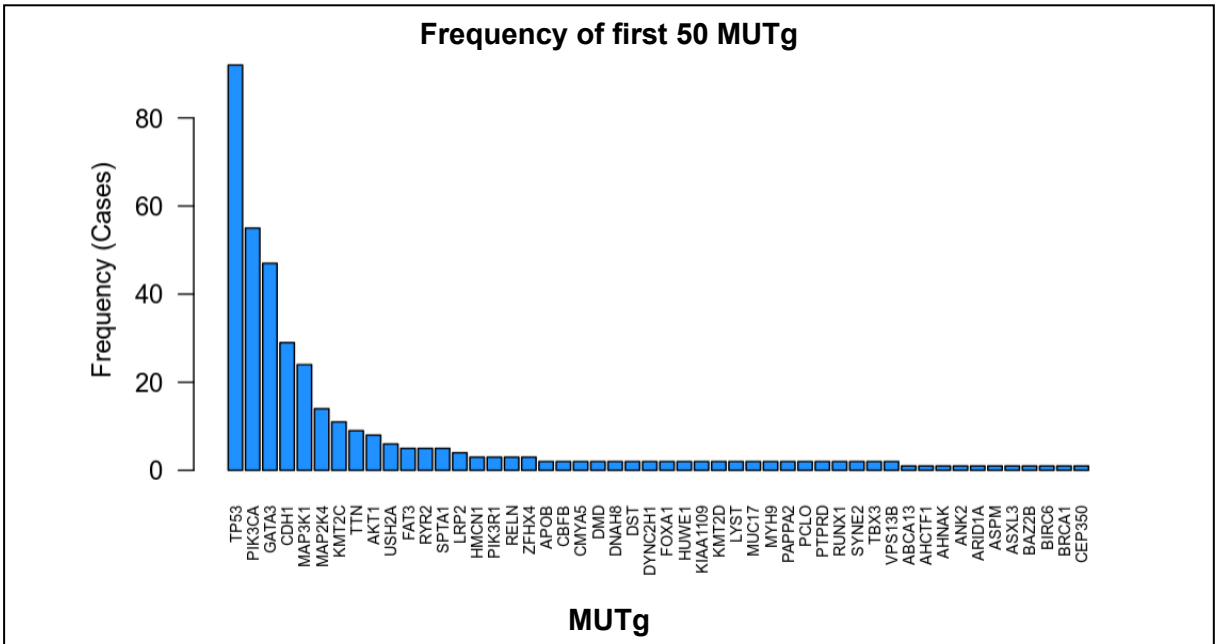


**Figure 10.** Frequency of first 50 MUTg in selected gene (Methylation by SDg).

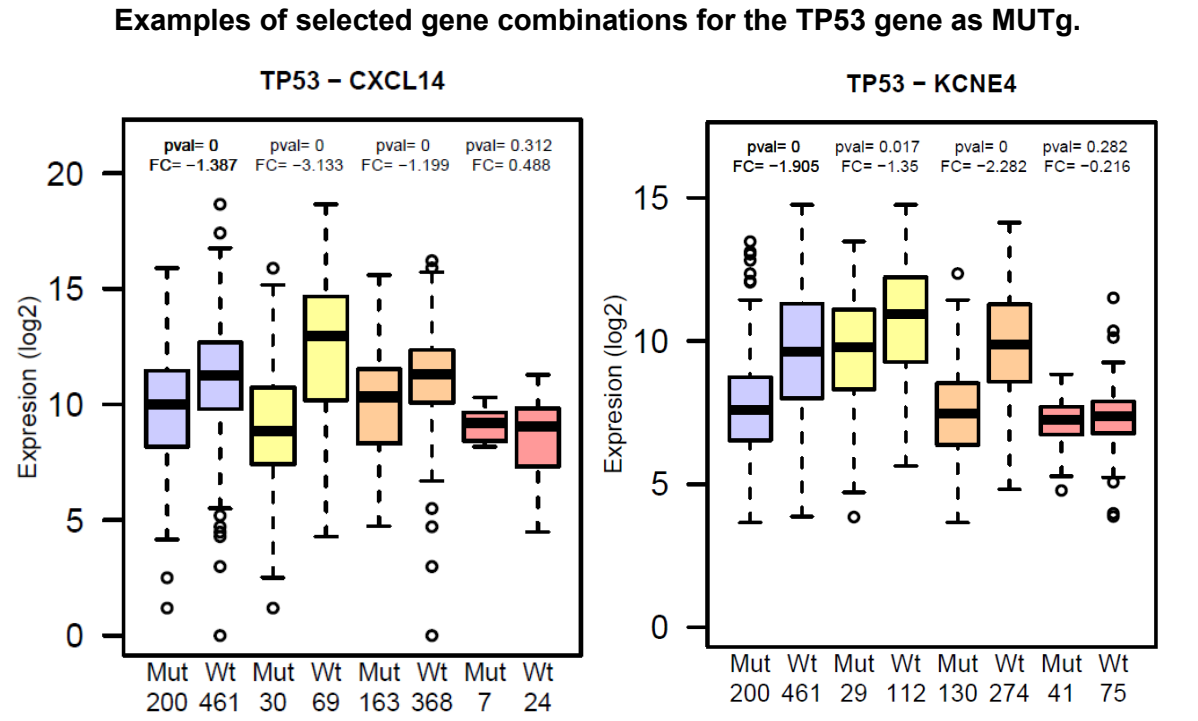


**Figure 11.** Two examples for the TP53 gene as MUTg. In both cases the results are significant since there is a group with p-value> 0.05 and another with p-value <0.05.

Groups by MUTg & Methylation by MUTg *(Appendix 3)*

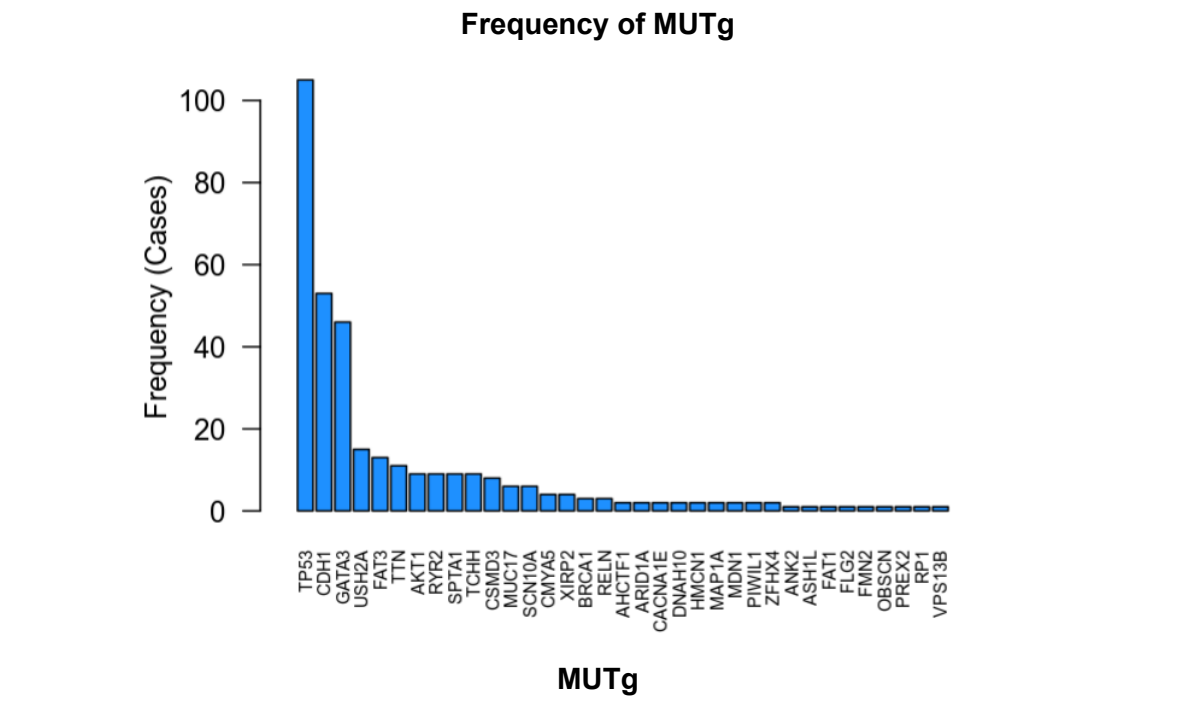


**Figure 12.** Frequency of MUTg in selected gene combinations (Methylation by MUTg).

**Example of selected gene combination for the CDH1 gene as MUTg**

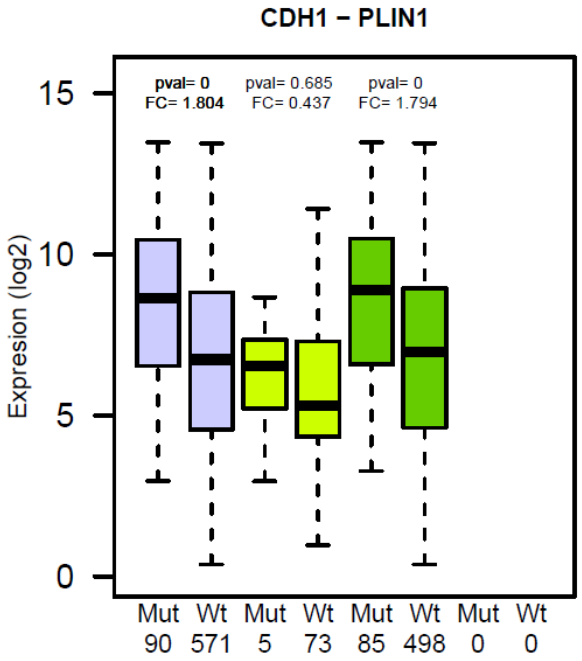


**Figure 13.** Selected gene combinations for CDH1 like MUTg & Methylation by MUTg. The expression is from PLIN1.

Groups by CNg & Methylation by SDg *(Appendix 4)*

**Example of selected gene combination for the LRRC59 gene as CNg**

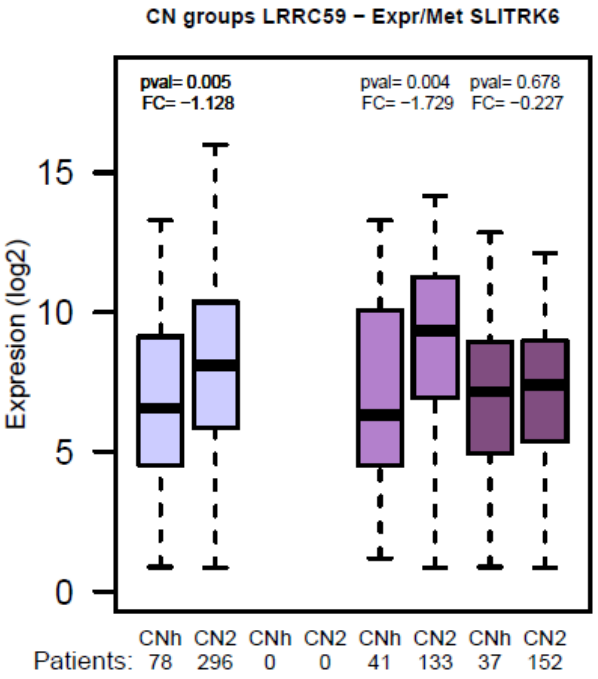


**Figure 14.** Selected gene combinations LRRC59 gene as CNg & Methylation and expression by SLITRK6.

Groups by CNg & Methylation by CNg *(Appendix 5)*

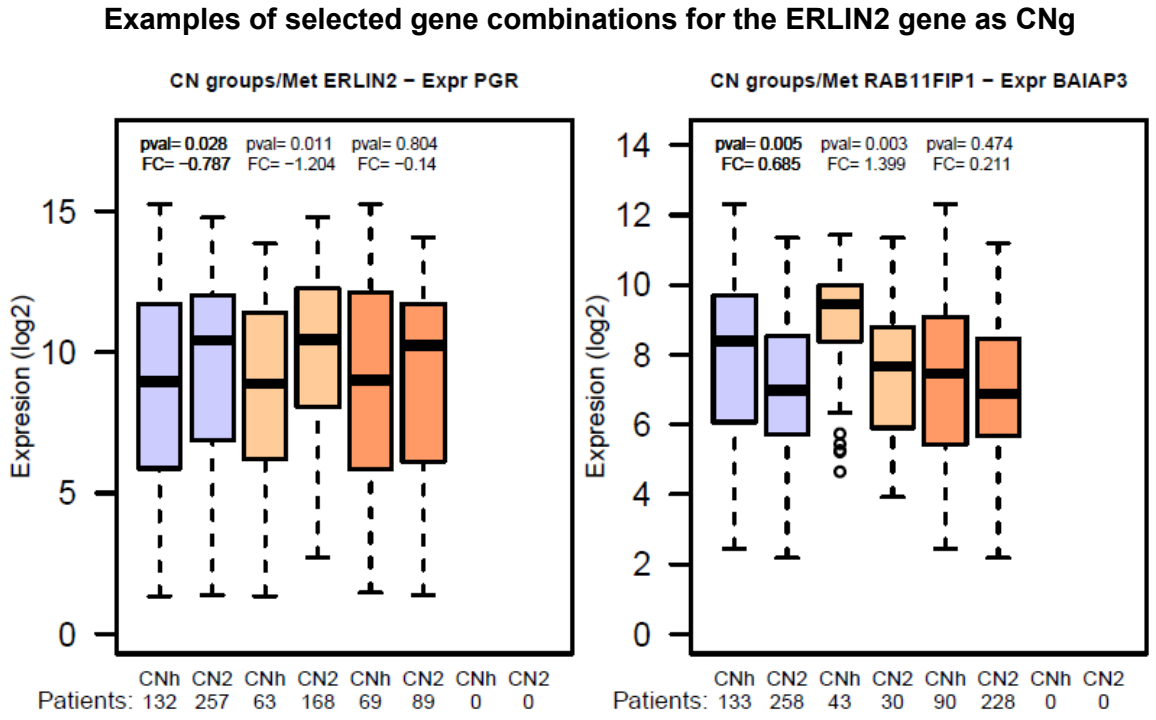


**Figure 15.** Selected gene combinations ERLIN2 gene as CNg & Methylation by ERLIN2. Expression by SDg.

## DISCUSSION

INITIAL GENE SELECTION

The results for the heatmap in *Figure 2.* representing the Log2(fold changes) values for the patient groups formed by the MUTg (patients with mutations vs WT patients) and the expression of the SDg, show clusters with groupings of genes for negative and positive foldchange values. In *Figure 3*, the rearrangement of the p-values following the same order as the genes in the heatmap foldchanges shows significant p-values mainly for those groups of genes defined in various clusters in the heatmap foldchanges graph. In *Figure 4* we observe the selection of selected foldchange values according to the significance of the p-values. The objective of this plot is to determine if the significant p-value of these groups reflects a real difference. The groups are again shown but in this case validated by the significance level p-value<0.1.

This result suggests that there are sets of genes that may have regulatory interactions. The joint study of these genes could facilitate the identification of gene hubs, which are points of integration and control in the gene and metabolic networks of an organism. The identification of these gene hubs could have a high value as a tumour marker or be a possible point of action on which to develop therapeutic strategies.

The combined matrix (p-value / foldchange) allowed us to establish a selection of MUTg with an impact on the expression of SDg. A gene that appears with a high frequency is a gene that influences the expression or function of other genes, known as a master regulator gene. This term is used to describe genes that play a central role in the regulation and coordination of multiple biological or metabolic pathways. Master genes are often transcription factors or other regulatory proteins that act in signalling cascades or gene regulatory networks. These genes may directly control the expression of other genes by binding to their promoters, or they may indirectly influence expression through complex signalling pathways. An example would be TP53 which has the highest frequency among the MUTg. This gene has a well-known role as a tumour suppressor gene regulating the expression of genes involved in apoptosis, DNA repair and other cellular processes. Mutations in this gene are common in a multitude of cancers, among them BC [11].

## COPY NUMBER

*Figure 5* represents the selection of 11795 protein-coding genes. On the right side of the graph, a group of genes that are separate from the rest and have more than two copies in the majority of patients is shown in red. Further analysis revealed that this set of genes belongs to chromosome 1q, which often has chromosomal aberrations leading to increased copy number in BC patients [12]. The distribution of the genes with respect to the x-axis (number of patients), disregarding the gene cluster on chromosome 1, resembles a normal distribution. This graph allows us to identify whether the distribution of genes, according to the number of patients, follows a pattern, as in the case of the chromosome 1 gene cluster. In

*Figure 6* left, the expression, at each copy number, of all 11795 genes is plotted. By filtering these genes, the CNg (165 genes) were obtained and used to produce *Figure 7*. This represents the combined expression of the CNg according to their copy number. Higher averages and a greater increase in gene expression are observed as the CN of the genes increases.

*Appendix 1* shows, for each of the CNg the expression according to CN. Many of the genes showed a progressive increase in expression values, with a quantitative jump from 6CN to 7CN. A sample boxplot for 3 CNg is shown in *Figure 8.*

The left graph shows a progressive increase in expression with increasing CN. The central graph shows a greater increase in expression values from 6CN to 7CN, with respect to the rest of the levels. In many genes there is a quantitative jump in gene expression in patients with 7CN. The statistical significance of these results was not assessed as the aim of the analysis was to visually verify whether there is an increase in expression as CN increases both globally and individually for each gene.

## METHYLATION

The density plots of the methylation means for MUTg and SDg (*Figure 9*) showed a similar distribution of the methylation measurements of the different patients. This distribution made it possible to establish 3 methylation levels to form the groups of patients as there was a greater probability of finding, for each gene, patients at all three methylation levels. In addition, this systematic approach would facilitate the establishment of methylation levels in other datasets where appropriate. A higher number of methylation levels would result in statistically worthless clusters due to insufficient patients. A smaller number of methylation levels would reduce the precision of the results by making it difficult to identify significant differences between groups at each level.

## GENE INTERACTIONS

Numerous significant results were obtained in all analyses, despite stringent screening criteria. This stringent screening criterion was used to obtain meaningful and consistent results, as a result with too many significant interactions can be counterproductive by making it difficult to choose for further analysis either in silico or in vitro. One option to reduce this problem could be to perform an even stricter selection, but it must be taken into account that relevant information regarding gene interactions could be lost.

Some genes stand out due to their frequency, such as: TP53, PIK3CA, GATA3, CDH1, MAP3K1, FOXA1, BRCA1, PGR and IDO1. These genes play critical roles in the biology of BC as well as other cancers and are therefore widely documented in the literature. Given their importance, it makes biological sense that they are among the most frequently selected genes.

All selected gene combinations showed significant differences (see Materials and Methods) between patient groups of the same methylation level and also a methylation level with no significant differences between groups.

### Groups by MUTg & Methylation by SDg

This analysis aims to establish whether MUTg affects SDg expression depending on the different methylation levels of SDg. For example, in *Figure 11* we have two interactions of the TP53 gene with different SDg. In both cases, the total groups (without taking into account methylation levels), which are represented in lilac, had p-values <0.05. However, the high level of methylation (in red) did not show a significant p-value (<0.05), but the low and medium level of methylation did show a p-value <0.05 and a high foldchange.

On the other hand, the methylation of SDg (which in this case is CXCL14 and KCNE4), based on which the methylation levels have been established, silences the differences produced by the

mutations of the TP53 gene, at the high level of methylation . That is, patients with a high level of methylation for these SDg do not show significant differences in the expression of these genes between the WT and the mutant genotype. Taking into account that the expression, at the other methylation levels, decreases in the groups of patients with mutations for TP53, we can affirm that there is evidence that the effect of the mutation in TP53 and high methylation produces a reduction of SDg expression. A recent article from 2022 describes how the CXCL14 gene has oncosuppressive activity in BC [13] so silencing this gene could favor tumor development. Regarding the KCNE4 gene, another very recent article (2023) correlates high expression of this gene with pathological characteristics in colonorectal cancer [14]. Silencing this oncogene through methylation would have an oncosuppressive effect.

These statistically significant differences, together with the quantitative differences in foldchange and a biological reasoning consistent with the results, make these SDg good candidates to evaluate in depth the possibility of opening a research avenue to elucidate what their relationship is, how they act in the tumor environment. and its use as a molecular marker or as a therapeutic target.

Groups by MUTg & Methylation by MUTg

This selection aims to establish whether the mutations and methylation of the MUTg affect the expression of SDg depending on the different methylation levels.

We see an example in *Figure 13* with the MUTg CDH1 and the SDg PLIN1, where the patient groups were organized into the different methylation levels according to the methylation of the MUTg. As in the previous examples, if the expression of the SDg presents a level with significant values and another with non-significant values, we can affirm that the composition of the patient groups (based on the methylation levels of the MUTg) affects the SDg expression. In other words, the methylation of MUTg on the basis of which the patient groups were formed would affect the expression of SDg.

In this case, a significant p-value and greater foldchange are seen, only at the average methylation level. This means that MUTg methylation affects SDg expression, producing an increase in it.

It is also observed that the comparison between the total groups (boxplot in lilac color) has a significant p-value, so the division of the patients into three methylation levels allows us to locate the level where the difference between groups is significant.

The genes for this selection are documented in the BC. The CDH1 (Cadherin-1) gene encodes the protein E-cadherin, which is a crucial cell adhesion protein that plays an important role in maintaining tissue structure and cohesion. E-cadherin acts as a tumor suppressor. Its loss or dysfunction may contribute to cancer progression by facilitating the spread of malignant cells and metastasis. CDH1 is associated with a specific subtype known as invasive lobular BC[15]. The PLIN1 gene is documented as a gene with prognostic significance in BC [16]. Again, the results of the analysis agree with the literature.

This approach facilitates a systematic approach that identifies gene interactions in a context where mutations and methylation are involved. From this starting point, new avenues of research could be opened that allow us to clarify the molecular mechanisms that exist both between well-documented genes and among other de novo ones.

Groups by CNg & Methylation by SDg

One of the selected CNgs was LRRC59, which encodes a protein with leucine-rich repeats, involved in protein-protein interactions, transport and cellular signaling. Multiple recent studies reveal that this gene has a potential prognostic biomarker [17,18]. In *Figure 14* we find, for the patient groups according to the CN of LRRC59, the expressions of SLITRK6.

SLITRK6 plays a crucial role in the development of the nervous system and although there is some literature that links this gene with different types of cancer, no study was found that establishes a relationship with BC. For this gene, the groups with high methylation (dark purple) did not present significant differences, however, the average methylation level did present them with a lower average expression value in the CNH subgroup. It follows that high methylation reduces the expression of SDg (SLITRK6 in this case), eliminating the differences between these groups. At the average methylation level there are significant differences between groups, with a higher expression in patients with 2CN. The increase in LRRC59 in CN reduces the expression of SLITRK6 in CNH, as well as high methylation values. Given that there is no bibliography on the matter, an investigation into the possible interactions at the molecular level could provide more information to establish lines of research,

although the information provided by this systematic approach constitutes an excellent starting point for further analysis.

Groups by CNg & Methylation by CNg

*Figure 15* shows the CNg ERLIN2 with the SDg PGR and BAIAP3, with non-significant p-values in the medium methylation level (there are no data for the high methylation level), indicating that methylation of the CNg gene prevents it from affecting the SDg expression. For the low methylation level, both cases show significant values in the comparison of groups according to their CN, meaning that the copy number variation in the CNg (when methylation is low and probably does not silence the CNg), affects the expression of the corresponding SDg.

ERLIN2 is involved in the degradation of misfolded proteins and in the regulation of lipid metabolism in the endoplasmic reticulum. It has been established as a key gene associated with metastasis in BC [19]. PGR is a well-documented gene essential for breast development[20]. BAIAP3 Regulates synaptic vesicular trafficking and neurotransmitter release in the nervous system. The literature for this gene is scarce and without direct connection with BC.

## CONCLUSIONS

The strict selective filters applied have allowed numerous results to be obtained. These results have shown gene interactions well documented by the literature, validating the gene selection system with these already known results. Results have also been obtained that agree with very recent and promising publications on genes that are currently little studied. The results also showed little-known genetic interactions, but with coherent statistical results, with a solid theoretical basis that provides a starting point to clarify the interactions between genetic and epigenetic events.

The methodology used facilitates a systematic approach that identifies gene interactions in a context where mutations, CN, methylation and expression participate. This systematic approach can be a great advantage, as it would allow easy application to other data sets. However, it would be essential to apply this system in other data sets to debug errors and improve strategies, with the aim of obtaining an improved methodology that can be applied through pipelines that facilitate the work.

These pipelines would be an important advantage in terms of application to other data sets, but the complexity that the preprocessing of this data sometimes entails is an important problem to take into account.

Another possible drawback is the large number of results obtained since a large number of genes makes their study difficult, allowing a wide range of candidates. Performing a stricter selection could be a good option to improve results.

Performing an analysis on the prognostic value of the selected genes with respect to patient survival could reduce the number of results as well as show that there is biological relevance in the selected genes.

The selection of genes carried out has given very promising results and is a starting point that could open new avenues of research that allow us to clarify the molecular mechanisms that exist both among well-documented genes and among others de novo. This methodology not only selects promising candidates, but also provides information on the molecular mechanisms involved, facilitating possible avenues of action on which to initiate possible in vitro research.

This approach facilitates the discovery of new gene interactions regulated by DNA methylation in a broader context and paves the way towards innovative therapeutic strategies and personalized treatments against BC as well as other types of cancer.

Finally, this work identifies genes regulated by DNA methylation within specific genetic contexts, offering insights into the interactions between genetic and epigenetic factors, enhancing our understanding of tumor biology in BC.

## REFERENCES


[1] Heer, Emily, et al. "Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study." The Lancet Global Health 8.8 (2020): e1027-e1037.

[2] Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 71.3 (2021): 209-249.

[3] Sørlie, T., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America, 98(19), 10869–10874. doi.org/10.1073/pnas.191367098

[4] Baylin, S. B., & Jones, P. A. (2016). Epigenetic Determinants of Cancer. Cold Spring Harbor perspectives in biology, 8(9), a019505. https://doi.org/10.1101/cshperspect.a019505

[5] Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology, 38(1), 23–38. doi.org/10.1038/npp.2012.112

[6] Kar, Swayamsiddha, et al. "Expression profiling of DNA methylation-mediated epigenetic gene-silencing factors in breast cancer." Clinical epigenetics 6 (2014): 1-13.

[7] Yoshida, K., & Miki, Y. (2004). Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. Cancer science, 95(11), 866–871. https://doi.org/10.1111/j.1349-7006.2004.tb02195.x

[8] Hakkaart, Christopher et al. “Copy number variants as modifiers of breast cancer risk for BRCA1/BRCA2 pathogenic variant carriers.” Communications biology vol. 5,1 1061. 6 Oct. 2022, doi:10.1038/s42003-022-03978-6

[9] Thakur, Chitra et al. “Epigenetics and environment in breast cancer: New paradigms for anti-cancer therapies.” Frontiers in oncology vol. 12 971288. 15 Sep. 2022, doi:10.3389/fonc.2022.971288

[10] Ennour-Idrissi, Kaoutar et al. “Epigenome-wide DNA methylation and risk of breast cancer: a systematic review.” BMC cancer vol. 20,1 1048. 31 Oct. 2020, doi:10.1186/s12885-020-07543-4

[11] Guimaraes, D. P., and P. Hainaut. "TP53: a key gene in human cancer." Biochimie 84.1 (2002): 83-93.

[12] Privitera, Anna Provvidenza, Vincenza Barresi, and Daniele Filippo Condorelli. "Aberrations of chromosomes 1 and 16 in breast cancer: a framework for cooperation of transcriptionally dysregulated genes." Cancers 13.7 (2021): 1585.

[13] Gibbs, Carla et al. “CXCL14 Attenuates Triple-Negative Breast Cancer Progression by Regulating Immune Profiles of the Tumor Microenvironment in a T Cell-Dependent Manner.” International journal of molecular sciences vol. 23,16 9314. 18 Aug. 2022, doi:10.3390/ijms23169314

[14] Tian, Kui et al. “KCNE4 expression is correlated with the pathological characteristics of colorectal cancer patients and associated with the radioresistance of cancer cells.” Pathology, research and practice vol. 241 (2023): 154234. doi:10.1016/j.prp.2022.154234

[15] Corso, Giovanni et al. “CDH1 germline mutations and hereditary lobular breast cancer.” Familial cancer vol. 15,2 (2016): 215-9. doi:10.1007/s10689-016-9869-5

[16] Zhou, Cefan et al. “Prognostic significance of PLIN1 expression in human breast cancer.” Oncotarget vol. 7,34 (2016): 54488-54502. doi:10.18632/oncotarget.10239

[17] Zeng, Meiqi, et al. "An integrated pan-cancer analysis of leucine-rich repeat containing protein 59: a potential biomarker for prognostic and immunotherapy." Genome Instability & Disease 4.6 (2023): 333-348.

[18] Zhang, Peng, et al. "LRRC59 serves as a novel biomarker for predicting the progression and prognosis of bladder cancer." Cancer Medicine 12.19 (2023): 19758-19776.

[19] Li, Wei, et al. "Transcriptome analysis reveals key genes and pathways associated with metastasis in breast cancer." OncoTargets and therapy (2020): 323-335.

[20] Fuqua, Suzanne AW, and Douglas M. Wolf. "Molecular aspects of estrogen receptor variants in breast cancer." Breast cancer research and treatment 35 (1995): 233-241.