Máster en Bioinformática aplicada a la Medicina Personalizada y la Salud

Escuela Nacional de Sanidad del Instituto de Salud Carlos III

Microarrays Data Analysis

Analysis and Interpretation of 'omics' data

Mar Batlle Perez

March 29, 2021

Table of contents

1. INTRODUCTION

This study aims to carry out a **microarray-based transcriptomic analysis** of the data described in the article "Direct Inhibition of the NOTCH Transcription Factor Complex" (Moellering et al., 2009) and recovered from the public functional genomic repository from NCBI; *Gene Expression Omnibus* (GEO), through the series database identifier GSE18198.

In molecular biology, **transcription factors** make up a fundamental piece for the understanding of cellular events and diseases with a high genetic component. These elements are proteins that control the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence. This translates into the control of the signaling pathways that make up the cell and, ultimately, the control of fundamental cellular processes for life, thus, proper functioning of these mechanisms is essential (Lambert et al., 2018).

Among the multitude of signaling pathways presented by the cells of complex organisms such as mammals, one of the most important is the **NOTCH signaling cascade**. This highly conserved system is associated with cell proliferation, differentiation and death; playing a major role in regulating embryonic. It is controlled by the proteins that give it its name, the NOTCH proteins, their binding results in the ICN-CSL-MAML complex, which recruits the transcription machinery of NOTCH-dependent genes (NOTCH target genes). Over-activation of the NOTCH signal has been observed in many cancers and has been studied extensively in **T-lymphocyte acute lymphoblastic leukemia** (T-ALL) where over 50% of patients have the mutated NOTCH gene. Small modulator molecules of these proteins could be important in understanding the role of NOTCH proteins in normal and malignant biological processes.

In this article, the author target to study the effect on NOTCH proteins caused by the synthetic peptide SAMH1, derived from the MAML1 protein. To show the role of SAHM1 as a repressor of the gene expression of the target genes of the signaling pathway, they carried out different approaches on cell lines with mutated NOTCH1, including a **transcriptomic analysis of gene expression profiles based on microarrays**, by which they determined SAHM1 produces the global repression of NOTCH1 signaling; this method will be replicated in the present study.

2. OBJECTIVES

The primary goal of this study is to verify that, as concluded in the article, **SAHM1 produced repression in the signaling pathway of NOTCH**. With this goal in mind, different points are also analyzed, including:

- Verifying that the normalization and correction of backgrounds are useful to scale the microarrays given different approaches, conditions and treatments.
- Through the use of clustering, validate that there are enough differences between both cell lines; KOPT-K1 and HBP-ALL to be classified as distinct groups.
- Computing whether there are significant differences in the differential expression results by following two different methodologies; study each cell line separately or, on the contrary, use both cell lines together.
- Reproducing the reference article's GSEA method and compare the results, including whether the same main enriched genes are under-expressed in the SAMH1 condition when compared with the control situation.

3. METHODS AND MATERIALS

As shown in the article, it will be intended to measuring the global changes in gene expression upon treatment of two human T-ALL cell lines; HPB-ALL and KOPT-K1. **Three replicas of each cell line are used under two different protocols**, only the control, in which the samples are treated with the compound dimethylsulfoxide (DMSO), and the control and the sample of interest, in which the synthetic peptide SAMH1 was applied.

To perform the analysis, a pipeline is built with R (v4.0.4) using the available packages from the open-source and open-development software project for the analysis and comprehension of genomic data; Bioconductor (v3.12), which enables the comparison of the results with those described by the authors. All the code developed during this study can be found in the following repository: https://gitlab.com/marbatlle/microarray-data-analysis. Below is a description of each of the tasks performed during the analysis.

3.1. Data Loading

The starting material of the article and this project is the **12 microarray samples** obtained from the GEO Database, Table 1, analyzed with the microarray technology Affymetrix U133 Plus 2.0.

Table 1. *Files information found at targets.txt*

File	Cell Line	Class
GSM455115	KOPT-K1	DMSO
GSM455116	KOPT-K1	DMSO
GSM455117	KOPT-K1	DMSO
GSM455118	HPB-ALL	DMSO
GSM455119	HPB-ALL	DMSO
GSM455120	HPB-ALL	DMSO
GSM455121	KOPT-K1	SAHM1
GSM455122	KOPT-K1	SAHM1
GSM455123	KOPT-K1	SAHM1
GSM455124	HPB-ALL	SAHM1
GSM455125	HPB-ALL	SAHM1
GSM455126	HPB-ALL	SAHM1

Given that this microarray data was produced using Affymetrix chips, this data comprises *CEL* files, which contain raw intensities for each probe on the array. Therefore, the first steps of the analysis are conducted with the *affy* package (v. 1.68.0), which consists of reading the *targets.txt* file, described in *Table 1,* with the *readTargets()* function of the *limma* package (3.46.0) and then load the *CEL* files using the *ReadAffy* function.

3.2. Preprocessing

If the HPB-ALL and KOPT-K1 samples are studied together, all the differences that exist between these cell lines may affect the differential expression analysis, but since the guidelines of this project don't delimit which way to proceed, both samples will be analyzed together and afterward, each one

separately; to see if there are significant differences between both results and thus, **determine which approach is more beneficial for the study**.

One thing to consider when analyzing microarray data is that is it known that when using this technology we can expect a certain **technical variability**, which is inherent in all microarray experiments because of the number of elements being measured and the number of steps in the process that culminates in the hybridization of RNA to the microarray slide (Subramanian et al., 2005). During microarrays, probe signals measure the abundance of specific labeled RNA sequences but they are also affected by unspecific sources such as auto-fluorescence of the chip surface, resulting in a presence of spots and different intensities between microarrays. These variations make it difficult to compare them. Therefore, to solve this problem, it is necessary to remove the systematic biases caused by the background noise and scale the microarrays to allow meaningful comparison and inference between different microarrays. For this reason, before diving deeper into the microarray data analysis, we need to proceed with another step; preprocessing, which will include background correction and intensities normalization. This process is an essential step of any microarray analysis and will allow us to transform the raw signal to expression value.

There are different methods described in the literature to **perform normalization**, this study uses the recommended method described for Affymetrix data; Robust Microarray Average (RMA), a method also implemented by the authors of the article. This method performs background correction, normalization, probe level intensity calculation, and summary of information.

Affymetrix chips are designed in such a way that each gene is matched to probes evenly distributed throughout the chip. Each of these proves are called Perfect Match probe (PM), and these are matched to a second probe, called Mismatch probe (MM). The presence of MM allows us to identify and quantify the amount of nonspecific hybridization signal that occurs in the microarray. Here, it's where the RMA background correction method comes to play and helps us differentiate the noise of genes that are expressed at low levels taking in account the presence of PM concerning the amount of MM. To solve the normalization issue, the RMA method uses the quantiles normalization, which ensures that the intensities have the same empirical distribution across arrays.

When it comes to apply the normalization to our data, we use the *expresso()* function of the *affy* library. As said previously, following the consensus, we apply the background correction (*bg.correct*) method RMA, the quantiles method for normalization. Finally, regarding the change of the PM values we set the method as *pmonly*, which translates to no method being applied and, regarding the summary method used, RMA uses *medianpolish*.

Once normalization steps are finished, we need to check how was our data modified, thus it is interesting to perform a standard statistical approach to check if there are any differences in the distribution of intensities with the corresponding histograms and boxplots. Another useful tool to analyze the differences between the different microarrays is the use of heatmaps to perform a clustering analysis through Euclidean distances and the employment of the principal component analysis (PCA) to compute the variabilities between groups.

3.3. Differential Expression Analysis

To analyze which genes are differentially expressed between the samples treated with DMSO or SAMH1, a linear model is fitted to our expression data with the *limma* package. This package is designed to analyze complex experiments involving comparisons between many experimental groups simultane-

ously while remaining reasonably easy to use for simple experiments. This package is meant to investigate complicated experiments involving comparisons between several experimental groups at the same time whereas remaining moderately simple to use for straightforward experiments. Its proficient shown results when working with a reduced number of replicates are the main reason that makes this packages the optimal option for this study.

Prior to the differential expression analysis, we have to have in mind that many of the genes on the chip won't be expressed in this experiment, or might only have small variability across the samples. It is important to eliminate these, and to do so, we **filter the data** using the interquartile range (IQR) with the *varFilter* function. This will allow us to only work by using the probes that present a greater variability between samples and thus facilitating the subsequent analysis of differential expression. Here a threshold of 0.5 for the IQR has been set, which reduces the size of the data set almost by 50%, so that, to carry the differential expression analysis, only those genes that are sufficiently variables and therefore, are more informative, will be used.

The first step to fit the model is the creation of the **design matrix**, to show which sample each file belongs to and, immediately after, we want to create the contrast matrices, which assigns which samples belong to each contrast. Thereafter, to determine the differentially expressed genes, a **linear model** (*lmFit*) is fitted and an empirical Bayesian method (*eBayes*) is used to establish the statistical differential expression, to establish the standard errors and estimated log-fold change (*logFC*). Regarding the significance thresholds imposed in the present analysis, an adjusted p-value threshold of 0.05 has been established. This correction allows us to focus on significant genes in the analysis.

Last, if we want to be able to see which genes are the most significant ones, we need to convert the labels obtained from the microarray experiment to labels with the symbols of the genes. For this, given the platform used, the library *hgu133plus2.db* (v3.2.3) applied to perform the **annotation of the probes** with different identifiers. To compare the different situations between both cell lines, a volcano plot will be used first and then a Venn diagram to determine what number of differentially expressed genes will share both cell lines.

3.4. Gene Set Analysis (GSEA)

In the past sections and analysis, all our genes were treated one by one, however genes don't act as isolated elements. The section aims to check sets of connected genes using the GSEA (*Gene Set Enrichment Analysis*, v4.1.0) tool, from the Broad Institute. This approach is powerful wherever the quantity of genes analyzed is high since it allows the large range of genes to be summarized in **sets of biologically relevant genes** (Subramanian et al., 2005). GSEA can be used through the stand-alone desktop application, a *GenePattern* module that works on the cloud or different R packages, in this case, the desktop application is to be used.

This tool works by applying a non-parametric test based on rankings; the *Kolmogorov-Smirnof* test, to find asymmetrical distributions for defined blocks of genes in datasets whole distribution and, therefore, the method will determine differences. The method works as follows as:

1. Rank the genes: based on a certain statistic obtained by comparing each of the genes in the two conditions using differential expression analysis, an analysis at the level of individual genes.
2. Calculate the enrichment score (ES): represents the amount to which the genes in the set are over-represented at either the top or bottom of the list.

3. Estimate the statistical significance of the ES: this calculation is done by a phenotypic-based permutation test to produce a null distribution for the ES.
4. Adjust for multiple hypothesis testing for when many gene sets are being analyzed at one time: the enrichment scores for each set are normalized and a false discovery rate is calculated.

To reproduce the articles approach, a set of cured genes from the *MsigDB* database is employed, specifically, the C3 collection. As the authors also did, a gene set called GSI-NOTCH, associated with associated with the effect of the GSI drug, enclosed within the same *.gmt* file. Additionally, the *.gct* file is generated with the intensity values of the different microarrays normalized by RMA and the mean of the DMSO controls were used, as normalization by the mean of DMSO controls allows inter-array variability regulation. Here, the GSEA analysis is carried out jointly in both cell lines, since the GSEA analysis is not an individual comparative analysis and the important thing is that the values used to order the genes are homogeneous, that the variability in all of them is equivalent so that the ranking is not biased.

4. RESULTS

4.1. General Approach

Preprocessing

First, an exploratory analysis of both cell lines together is carried out. In *Figure 1* the intensity distributions and the boxplots of the original data of the different microarrays are shown. We can observe how there are differences between each replicate, specially seeing the most variability between the HPB-ALL control replicas, which we can assume it is caused by technical variations. A difference in the intensities between cell lines can also be appreciated; KOPT-K1 shows lower expression levels. It can be seen that the corresponding medians represented tend to be lower and both the peak and the rest of the intensity distribution occur at lower values than the HPB-ALL cell line.

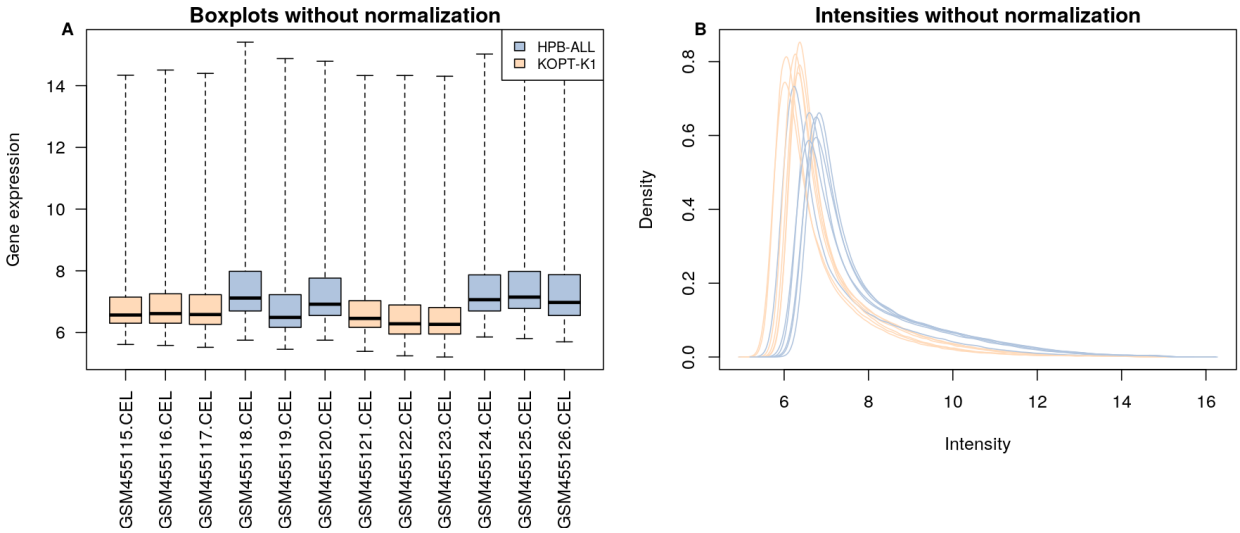


Figure 1. *Boxplots (A) and distribution of the intensities (B) of the samples of both cell lines before normalizing.*

As discussed in the Methods section, before analyzing the differential gene expression of our samples under different conditions, the normalization of the data is necessary to make the expression matrices from different microarrays comparable. Given that cell lines are being mixed in this approach,

normalization will not only eliminate technical noise, but also the biological variability presented by these lines.

Figure 2A and 2B show that, after normalization, the distribution is much more homogeneous among the samples and this exploration also shows that even though the shape of the distribution is the same, both cell lines show lower values after normalization.

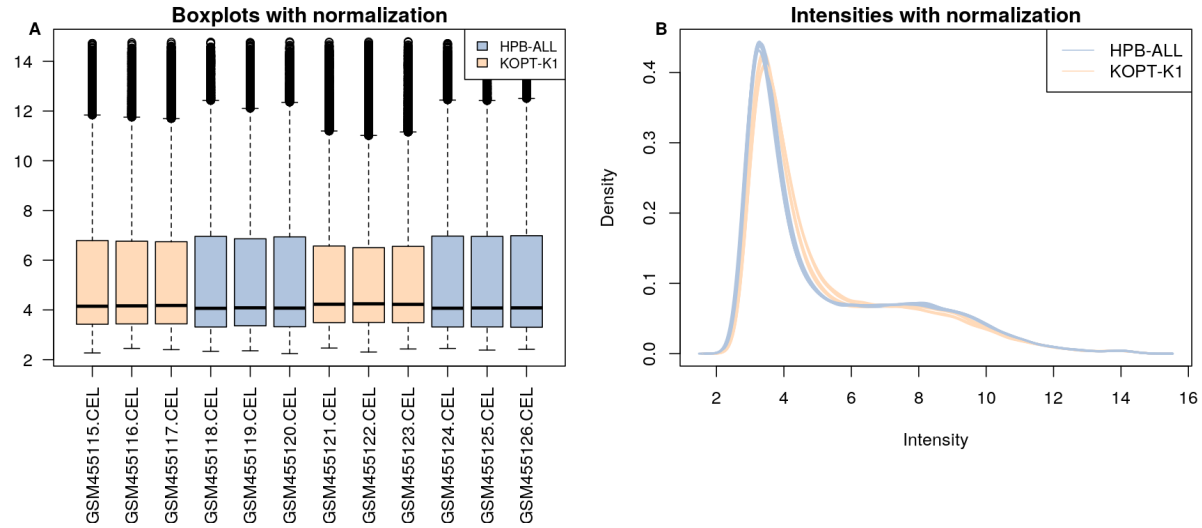


Figure 2. *Boxplots (A) and distribution of the intensities (B) of the samples of both cell lines after normalizing.*

To explore more clearly the differences between the samples, the distance matrix corresponding to the normalized samples has been calculated using the Euclidean distances as metric, Figure 3. From this graph it can be seen that the differences between the cell lines are highlighted, therefore, supporting the idea that show that this approach is incorrect for the present dataset, since the key differences are explained by cell lines, not treatments.

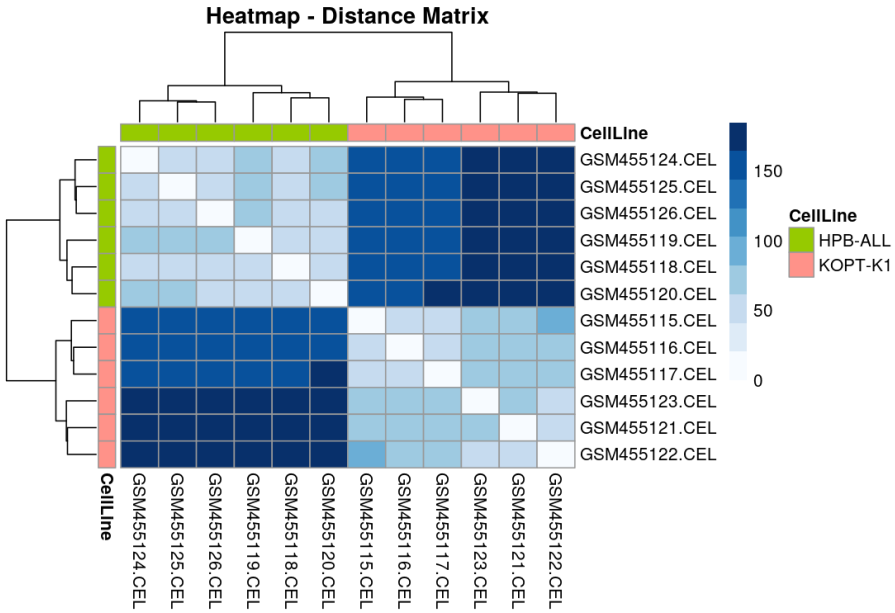


Figure 3. *Heatmap of Euclidean distances corresponding to the comparison of all normalized samples with each other. The light blue tones represent smaller distances and, therefore, greater similarity.*

Differential Expression analysis

Finally, even though the previous results already show that this approach is incorrect for the present dataset, an attempt was made to carry out a differential expression analysis between the two treatments (DMSO vs HPB-ALL) joining the samples independently of the cell line. However, differentially expressed genes were not obtained with an adjusted p-value threshold lower than 0.05, so the differences determined were not statistically significant.

4.2. One Cell Type Approach

In this section, the same steps carried out previously for the samples corresponding to each cell line will be shown separately. An approach that, theoretically, is more correct and that should allow getting significant results in the differential expression analysis.

Preprocessing

KOPT-K1 cell line

The process carried out to compare the KOPT-K1 cell line samples is the same that was used in the previous approach. Not normalized intensity data is shown in Figures 4A and 4B while normalized data is shown in Figures 4C and 4D. Unlike the first approach, we can observe that the different unnormalized replicates medians are quite similar, despite receiving different treatments.

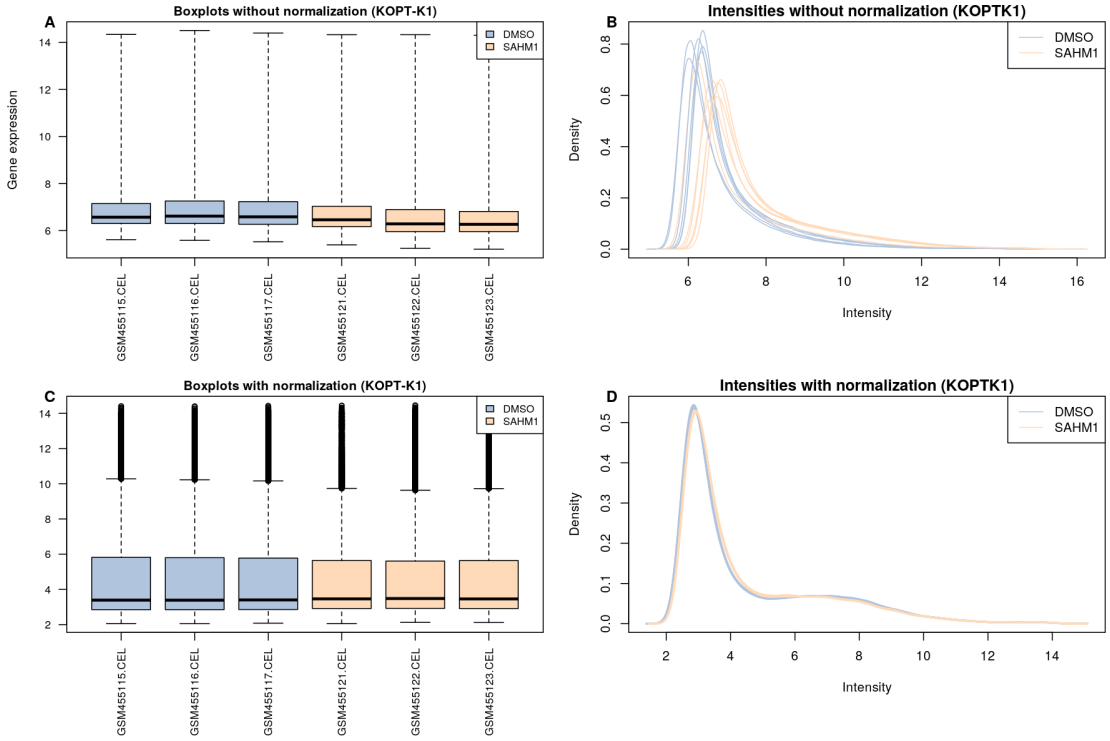


Figure 4. *Boxplots (A) and distribution of the intensities (B) of the samples of the KOPT-K1 cell lines before normalizing. Boxplots (C) and distribution of the intensities (D) of the samples of the KOPT-K1 cell lines after normalizing.*

Once the data has been normalized, these slight differences disappear; the distribution of the intensities shows a overly defined and unique peak for all the samples.

Regarding the corresponding PCA, in this case, it facilitates the analysis of the differences between the different genes analyzed from this cell line. In Figure 5 we can see how the two key components

retain 69.2% of the variance of the data, this shows that this is the total percentage of variance explained by this approach.

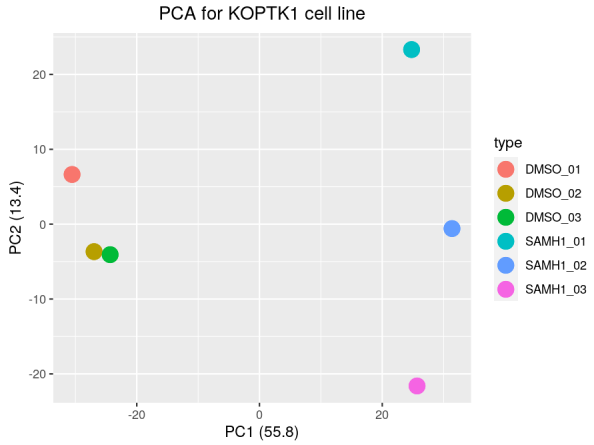


Figure 5. *PCA corresponding to the normalized samples of the KOPT-K1 cell line showing the treatment and the replicate number.*

The first component of the PCA (55.8% variance) separates the microarrays based on the treatment. Showing that the major differences observed now are attributed to treatment, and that differentially expressed genes can probably be found. Likewise, the second component of the PCA (13.4%) shows how the microarrays of the KOPT-K1 cells treated with SAHM1 present more differences among themselves than the microarrays belonging to the control replicates, as they remain grouped.

HPB-ALL cell type

Regarding the normalization of the HPB-ALL samples, when comparing the graphs in Figure 6, we verified that the normalization and correction of backgrounds are useful to scale the microarrays of the HPB-ALL cell line in the DMSO control condition and with the SAMH1 treatment.

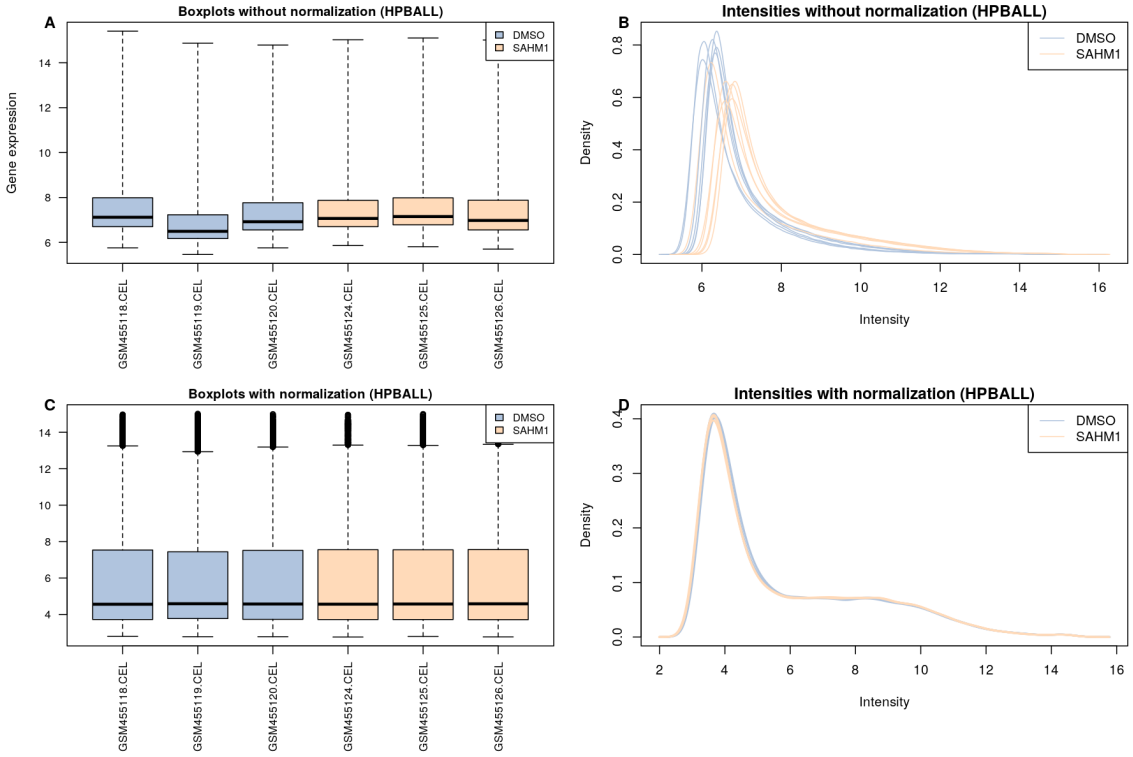


Figure 6. *Boxplots (A) and distribution of the intensities (B) of the samples of the HPB-ALL cell lines before normalizing. Boxplots (C) and distribution of the intensities (D) of the samples of the HPB-ALL cell lines after normalizing.*

Before assessing the differential expression of the genes, we use a PCA plot to visualize the differences between all the HPB-ALL samples and assess their variability. In Figure 7, we can see how the two main components retain 67.7% of the variance of the data.

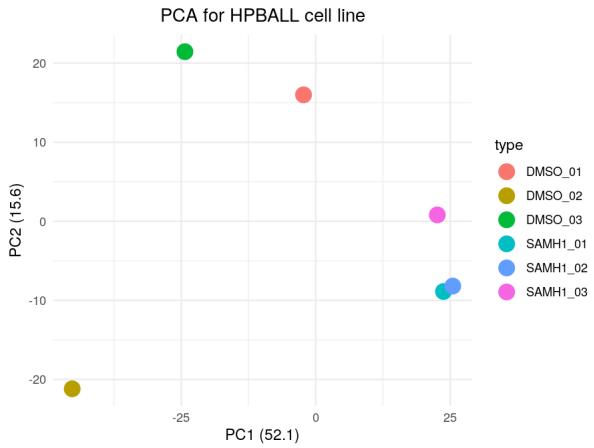


Figure 7. *PCA corresponding to the normalized samples of the HPB-ALL cell line showing the treatment and the replicate number.*

As with the analog cell type, the samples are separated according to the treatment by the first component which, in this case, explains 52.1% of the variance. Showing that the main differences observed now are attributed to the treatment and that differentially expressed genes can probably be found. Likewise, the second component of the PCA, 15.6%, mostly shows how the replicates corresponding to the treatment with SAHM1 present greater variability among themselves than the controls.

Differential Expression analysis

KOPT-K1 cell line

This section shows the results of the differential expression analysis of the KOPT-K1 cell lines. A list is obtained with the differentially expressed genes (DEGs) in the samples subjected to treatment with SAHM1 and the control replicates. To do this, first, as described in the Methods section, the probes with an IQR lower than 0.5 were filtered out, thus using only genes that show variability and that are capable of presenting different levels of expression. Afterward, a linear model was fitted with the *limma* package and by comparing the samples treated with both treatments, 6272 differentially expressed probes were obtained with a corrected p-value less than or equal to 0.05, with a total number of 4579 unique DEGs.

Given that the article aims to detect down-regulation of the genes of the NOTCH1 signaling pathway, Table 2 shows a list of the ten unique genes with the lowest logFC found.

Table 2. *Differentially expressed genes in KOPT-K1, sorted by ascending logFC*

Gene	logFC	AveExpr	t	Adj. P.Val	
SH3GL3	-1.57	6.71	-13.21	0	211565_at
LOC441666	-1.49	5.19	-12.4	0	1562527_at
FRG1BP	-1.37	5.24	-10.58	0	243689_s_at
TRIM4	-1.25	9.19	-10.36	0	224159_x_at
ZNF595	-1.24	6.83	-10.76	0	227952_at
TMEFF2	-1.23	7	-7.44	0	224321_at
RASEF	-1.23	11.1	-10.46	0	1553186_x_at
PDE4DIP	-1.19	3.74	-10.31	0	209700_x_at

FKSG49	-1.18	8.9	-10.64	0	211454_x_at
SLC6A16	-1.18	4.23	-10.29	0	1569940_at

KOPT-K1 cell line

As we did with the KOPTK1 samples, this section shows the results of the differential expression analysis of the HPD-ALL samples. Here, 1758 differentially expressed probes were obtained with a corrected p-value less than or equal to 0.05. A final number of 1409 unique DEGs is detected. In Table 3, a list of the ten unique genes with the lowest logFC found is presented.

Table 3. *Differentially expressed genes in HPB-ALL sorted by ascending logFC*

Gene	logFC	AveExpr	t	Adj. P.Val	
H1-5	-2.7	7.17	-5.54	0.01	214534_at
H1-3	-2.4	8.77	-5.08	0.01	214537_at
H1-4	-2.37	7.86	-5.52	0.01	208553_at
MALAT1	-2.32	7.46	-6.7	0	228582_x_at
JUN	-2.27	6.75	-13.11	0	201465_s_at
DDX17	-1.92	9.38	-6.42	0	230180_at
BCOR	-1.79	8.14	-6.3	0	223916_s_at
RASSF6	-1.78	4.98	-10.54	0	235638_at
ASPM	-1.71	9.62	-5.37	0.01	232238_at
CCDC18	-1.68	6.94	-5.52	0.01	236665_at

Comparison between both cell lines

Many of DEGs have been found for both cell types, seeming to confirm differences between both types of treatment; the number of differentially expressed genes with HPB-ALL is lower than in the KOPT-K1 samples.

Volcano plots are vastly useful to identify quickly changes in data sets composed of replicate data, and in this case, are helpful to identify the differences between both analysis. Figure 8 shows that the HPB-ALL samples have a much higher number of down-regulated genes than up-regulated ones, while in KOPT-K1 the opposite situation is found.

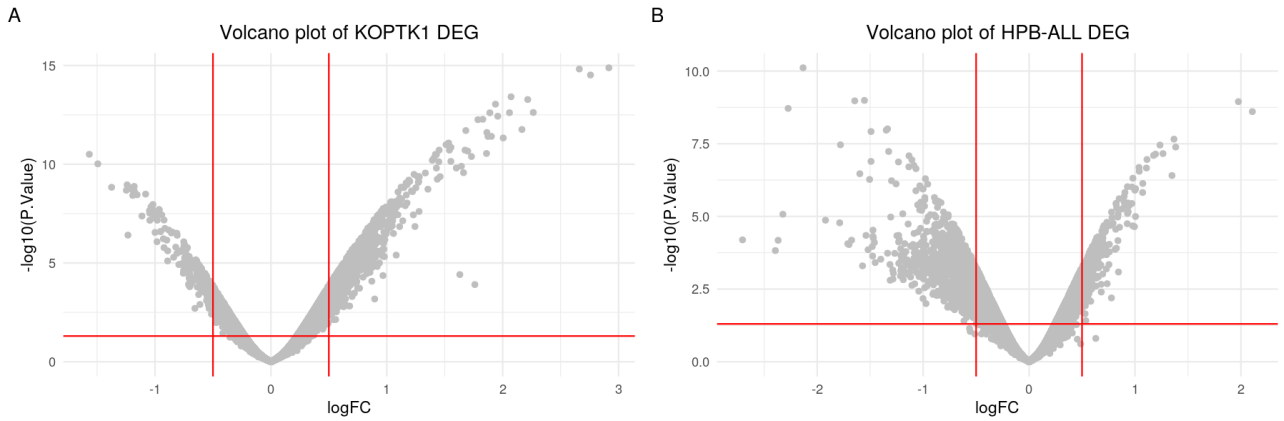


Figure 8. *Volcano plots corresponding to the analysis of differential expression DMSO vs SAHM1 in KOPT-K1 (A) and HPB-ALL (B). Thresholds are represented in a red line; p-value threshold of 0.05 and log fold between -0.5 and 0,5.*

One last tool used to compare the number of DEGs between both cell lines, and their overlapping genes, is a Venn diagram (Figure 9). In this context, it is used to detect relations between both lists of DEGs and explore the intersection of genes between them.

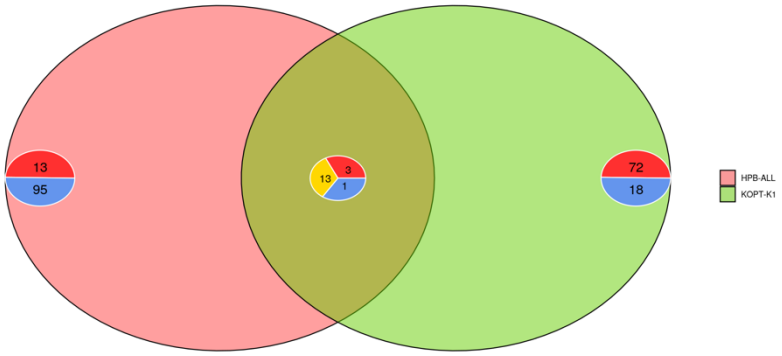


Figure 9. *Venn diagram for both cell lines, for probes with a logFC of at least 2. Red represents up-regulated and blue down-regulated genes.*

This diagram was carried out for a logFC of 2, and in this case, it shows how there are 17 genes present in both lists. This diagram also confirms that the cell line HPD-ALL shows down-regulated genes and the KOPT-K1 line the opposite. It's also interesting to note that the difference in the total number of DEGs shown previously has been reduced once we selected the genes with a logFC of 2, as now we can see the HPB-ALL cell line 125 and the KOPT-K1 only 18 genes less. There is still differences between both cell lines and this may be due to variability between them, although it is also possible that SAHM1 functions differently in each cell type; other types of approximations are necessary to produce a conclusive answer.

4.3. Gene Set Analysis (GSEA)

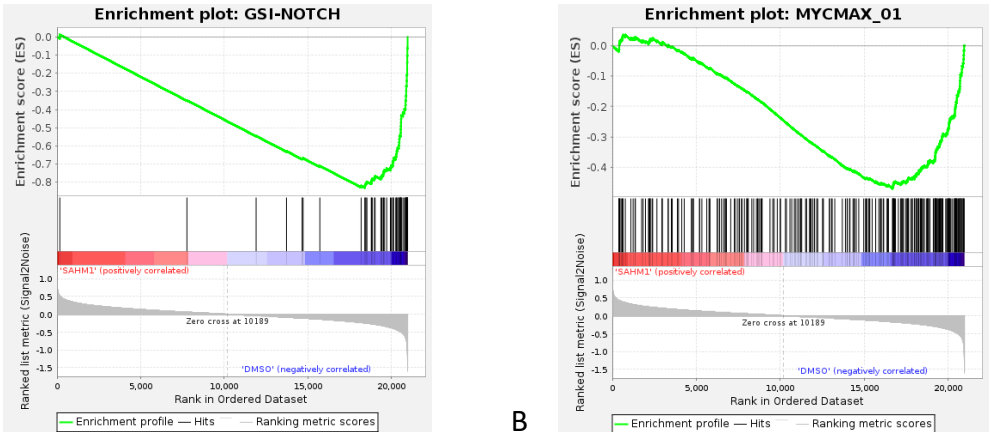
In this section, the study of enrichment in gene seats is carried out with the GSEA tool. Here, the aim is to facilitate the interpretation of the results and the capture of interactions between genes that are not detected by differential expression analysis. Regarding the results, GSI-NOTCH was obtained in the first position, while MYCMAX_01, despite not being in the second as in the original article, appeared in the third.

Table 4. *Most significative gene seats (larger NES score) obtained from the GSEA analysis.*

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
GSI-NOTCH	56	-0.8299398	-3.092909	0	0	0
EPC1_TARGET_GENES	354	-0.4924373	-2.4062092	0	0	0
MYCMAX_01	255	-0.4702582	-2.224768	0	3.99E-04	0.001
RUVBL2_TARGET_GENES	35	-0.6241652	-2.1485353	0	0.00210329	0.007
HIF1_Q3	231	-0.4509059	-2.100936	0	0.00288792	0.012
NMYC_01	272	-0.4366982	-2.0563648	0	0.00540522	0.027
ZNF165_TARGET_GENES	75	-0.5083646	-1.9977839	0	0.01066222	0.06
MIR7109_5P	75	-0.4870973	-1.9233825	0	0.02840278	0.17
MIR6085	39	-0.5560577	-1.9195241	0	0.02644728	0.177
MIR6813_5P	40	-0.5461481	-1.9030523	0	0.02861149	0.205

CSHL1_TARGET_GENES	210	-0.413183	-1.900238	0	0.02666242	0.21
MIR6887_5P	133	-0.4328631	-1.8969285	0	0.02584212	0.22
ZNF239_TARGET_GENES	37	-0.5416899	-1.8948712	0	0.0242212	0.224
KTGGYRSGAA_UNKNOWN	74	-0.4794442	-1.8943338	0	0.02257552	0.225

It can be seen that the enrichment of GSI-NOTCH, Figure 10A, is especially clear, with an adjusted p-value = 0.0 an a normalized enrichment value (NES) = -3.09, corroborating that the effect of SAMH1 is produced by under expressing target genes of NOTCH. Nonetheless in the case of MYCMAX_01, Figure 10B, a remarkable enrichment is also observed, but much lower than that of GSI-NOTCH, with an adjusted p-value = 3.99E-04 and NES = -2.23.



A B

Figure 10. *Enrichment plot for the NOTCH1 gene set (A) and for. The MYCMAX_01 (B)*

5. DISCUSSION

The first idea that needs to be considered, is that as noted earlier, when using microarray data, it is known that a **technical variability** can be expected, therefore, background correction and intensities normalization of the data is a key task before being able to analyze the data. In this case, the article's approach has been replicated, and as seen on Figures 2, 4 and 6, the histograms and boxplots used show that the differences in the distribution of intensities has diminished most undesirable systematic variations that were observed. Therefore, it can be stated that this methodology was successful, and it was verified that the normalization and correction of backgrounds are useful to scale the microarrays given different approaches, conditions and treatments.

To obtain results with a representative significance, the variability between the replicas can't be large, since this leads to high inconsistency and models with a large error. When considering the global approach, mixing both cell lines, after calculating the distance matrix corresponding to the normalized samples using the **Euclidean distances** as a metric, Figure 3, we can determine that the differences between both cell lines are clear, and therefore mixing both cell lines, with different genetic expression backgrounds, prevents us from obtaining significant results. If we wanted to carry out a comparison of the differentially expressed genes between the two lines, this approach would have been the correct one, although it is not the aim of neither, the analysis or the article.

When approaching this problem by analyzing the cell types separately, as seen on the analogous PCA plots (Figures 5 and 7), most of the **variability between the samples can be explained by the treatment** applied to these, instead of the cell types, as occurred in the last approach. Some variability can be

observed between the different replicates from the same group, this might be due to the inherent biological noise of the cell population. In this analysis, all the samples have been considered, but it may be advisable to exclude de most dissimilar ones in case that it's necessary to repeat the study since the variability between replicates can mask differential expression analysis.

As seen in the results, the **number of differentially expressed genes in the case of HPB-ALL is clearly lower than that of the KOPT-K1 samples**, which could be due to the variability shown between the replicates treated with DMSO. In the volcano plots, Figure 8, it can be seen that in the case of HPB-ALL, the number of down-regulated genes is much higher than that of up-regulated ones, while in KOPT-K1 the reverse situation is found; there is a greater number of overexpressed genes. In any case, a large number of DEGs have been found for both cell types, seeming to confirm differences between both types of treatment. However, as the restrictions imposed are increased, as seen in the Venn diagram (Figure 9), the numbers are increasingly similar until reaching the point that, with respect to the number of unique genes, the number of DEGs in HPB- ALL is greater than in KOPT-K1.

Alternatively, threshold-free gene enrichment methods such as GSEA can be especially useful, since, in addition to easy interpretation, variability between replicas at the gene level does not mask the results at the gene set level. We have reproduced the GSEA results from the reference article, finding the **GSI-NOTCH gene set as the main enriched gene set** and the MYC / MAX gene set in the third position. Both under-expressed in the SAMH1 condition when compared with the control situation. It can be said say that the genes that are under-expressed in the presence of the GSI drug also appear under-expressed in the presence of the synthetic peptide SAMH1.

6. CONCLUSIONS

Microarrays are a powerful tool to explore differential gene expression for thousands of genes at once. The raw data needs to be transformed due to systematic variations expected when using this tool. In this study, we verified that the normalization process was useful to scale the microarrays given different approaches, conditions and treatments.

Transcription factors play an important role in regulating the cell state, but their high chemical complexity complicates their targeting to discover treatments. We attempted to test whether treatment with SAHM1 could target the transcription factor NOTCH, by measuring the changes in gene expression upon treatment in two T-ALL cell lines. The results showed how the effect of variability can lead to statistically non-significant results, and thus joining the samples without taking into account the cell line to which they belong, fails to carry out a precise differential expression analysis.

Furthermore, we validated the effect that synthetic peptide of interest has in both cell lines; KOPT-K1 and HPB-ALL. Thus, we can attest that the effect of SAMH1 is independent of the cell line.

After reproducing the GSEA of the reference article, we deduced that SAHM1 exerts a specific antagonistic effect on NOTCH-directed gene expression as it provides a surprising correlation between the expression effects of SAHM1 and GSI. We can state that the main target of SAHM1 is the NOTCH signaling path.

This allows us to conclude that SAHM1 could be an attainable targeted treatment for NOTCH-controlled cancers, such as T-cell acute lymphoblastic leukemia.

7. REFERENCES

Bilban, M., Buehler, L. K., Head, S., Desoye, G., & Quaranta, V. (2002). Normalizing DNA Microarray Data. *Current Issues in Molecular Biology,4*(2), 57-64. doi:10.21775/cimb.004.057

Lambert, S. A., Jolma, A. F., Campitelli, L. K., & Das, P. U. (2018). Transcription Factors in Human Diseases. *Cell,172*, 650-665. doi:10.1016/j.cell.2018.01.029

Moellering, R. E., Cornejo, M., Davis, T. N., Bianco, C. D., Aster, J. C., Blacklow, S. C., . . . Bradner, J. E. (2009). Direct inhibition of the NOTCH transcription factor complex. *Nature,462*(7270), 182-188. doi:10.1038/nature08543

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences,102*(43), 15545-15550. doi:10.1073/pnas.0506580102