Antonio Federico[1,2], Laura Aliisa Saarimäki[1,2], Angela Serra[1,2], Giusy del Giudice[1,2], Pia Anneli Sofia Kinaret[3], Giovanni Scala[4], Dario Greco[1,2,3,5,*]

[1] Faculty of Medicine and Health Technology, Tampere University, Finland
[2] BioMediTech Institute, Tampere University, Finland
[3] Institute of Biotechnology, University of Helsinki, Helsinki, Finland
[4] Department of Biology, University of Naples Federico II, Naples, Italy
[5] Finnish Center for Alternative Methods (FICAM), Finland.

* Corresponding Author: dario.greco@tuni.fi

Running Head: Preprocessing of DNA microarray data

Microarray data preprocessing: from experimental design to differential analysis

Summary/Abstract

DNA microarray data preprocessing is of utmost importance in the analytical path starting from the experimental design and leading to a reliable biological interpretation. In fact, when all relevant aspects regarding the experimental plan have been considered, the following steps from data quality check to differential analysis will lead to robust, trustworthy results. In this chapter, all the relevant aspects and considerations about microarray preprocessing will be discussed. Preprocessing steps are organized in an orderly manner, from experimental design to quality check and batch effect removal, including the most common visualization methods. Furthermore, we will discuss data representation and differential testing methods with a focus on the most common microarray technologies, such as gene expression and DNA methylation.

Keywords: Microarray, preprocessing, experimental design, normalization, batch effect, differential analysis, omics data analysis, gene expression, DNA methylation.

1. Introduction

DNA microarrays have led the way from studies of individual genes or, at best, pathways, towards a global investigation of cellular activity. The vast amounts of data obtained through microarrays represent a valuable source of information for solving biological questions and hypotheses.

However, this cannot be achieved without proper data processing and consideration of the biases caused by the microarray experimental steps. Fortunately, these biases are recognized, and can be dealt with through different data processing steps which include careful quality assessment, sample and probe filtering, data normalization, batch effect estimation and correction, as well as probe annotation (Figure 1). These steps represent the golden standard of data preprocessing for gene expression and DNA methylation microarrays, both of which are covered in this chapter ***(1)***.

[Figure 1 near here]

The technology of DNA microarrays is based on nucleic acid hybridization, allowing the complementary sequences in the array (probes) and the sample to bind to each other by base pairing, while the signal is measured through fluorescent dyes. The several steps from experimental design to sample hybridization can introduce various sources of systematic errors that need to be carefully considered. Mindful experimental design avoids generation of unmanageable biases, paving the way for reliable, reproducible and good-quality data through the subsequent preprocessing steps. Data quality starts from proper technical execution of sample collection, RNA/DNA quality and purity assessment, and the microarray experiment itself. The experimental activities are followed by further quality assessment of the data, which helps to identify outlier samples, systematic biases or other errors. Furthermore, microarrays are intrinsically prone to a level of background noise as a result of cross-hybridizing probes and the saturation of the signal, setting the need for probe filtering. Normalization, on the other hand, is not only crucial for scaling the signals of individual measurements and arrays to allow relevant comparisons between molecular targets and samples, but also the key to meaningful cross-study

comparisons ***(2, 3)***. While normalization has the power to adjust some of the bias introduced by different batches of samples, proper mitigation of batch effects can be achieved through batch effect evaluation and correction. Finally, the probes are annotated to known targets, such as genes, genomic regions or CpGs, to support the interpretation of the results in a meaningful way. Typically, the output is further analyzed by differential expression or methylation to understand the changes introduced by the varying conditions between groups of samples.

When properly designed, performed, and processed, microarray experiments provide an effective and reliable way to study the key players of different developmental stages, pathological processes and molecular responses to perturbations, including exposure to drugs and toxic compounds, as well as the effects of gene knockout and knockdown ***(4)***. Altogether, this chapter provides an overview of the general workflow that can be applied to achieve robust and reliable results from gene expression and DNA methylation microarray data to unravel the mechanisms underlying a range of biological processes.

2. Methods

2.1. Experimental design

Microarrays provide an effective way to study and understand the state of, and changes in biological systems. However, successful microarray experimentation requires a deep understanding of the technology and all the possible pitfalls. A careful and intelligent experimental design is the first step towards effective and reliable microarray analysis. Experimental artifacts and technical effects are a common issue in scientific experimentation and results in an unwanted variation in the data. Such artifacts can heavily affect the experiment, leading to biased results or,

in the worst case, to discard the entire experiment. Thus, a rigorous study/experimental design is needed in order to mitigate, if not avoid, artifacts due to technical nuisances ***(5)***. An experimental design begins from proper sample selection. To achieve a statistical significance for meaningful downstream analysis, a proper number of replicates, with corresponding control samples are needed. The number of replicates depends on the origin of the sample. With a more homogenous sample set, derived from a cell line for example, less replicates are most likely needed to cover the biological variance than in the case of heterogeneous patient samples. As microarrays are relatively expensive assays, usually a tradeoff between the number of different treatments, sample groups and the number of replicates will be needed. Nonetheless, as a priority, a proper set of relevant control samples is required to make meaningful comparisons between the treatment/disease and the steady/healthy state. In addition to the negative (untreated or base-line) control, a positive control and/or midpoint controls might provide meaningful, comparable, replicable and usable results. After proper sample selection and extraction (RNA, DNA, protein), the second most important step is to ensure the quality of the obtained samples. For this, the sample quantity as well as its integrity need to be ensured. Data quality measures will be discussed in paragraph 2.2. Given the different microarray technologies, vendor specific platforms and solutions, array-specific procedures need to be considered case-by-case as part of the experimental design. However, common features include the use of fluorophore dyes for labeling and intensity detection. In a two-color assay, a simple mistake is to dye the same sample group with the same dye causing a dye-effect that cannot be corrected from the obtained data. This will result in consistently higher or lower intensity values in the particular sample group. Thus, when dual-colors microarrays are used, a careful randomization between samples and the dyes needs to be performed. Even with a one-color assay, a controlled sample randomization is required to

minimize the bias from sample groups or sample position on the arrays during hybridization. When the experiment is well designed with proper randomization, most of the technical biases can be corrected during the following data pre-processing steps. For this, a thorough recording of each sample feature is required to recognize the additional artefacts caused by uncontrolled technical aspects such as sample processing dates, variation in sample processing or collection, dyes and slide/chip area. This information provides an important starting point for proper data quality check, normalization and batch effect correction discussed in detail in the following paragraphs.

When appropriately designed and implemented, microarrays provide a valuable set of molecular descriptors that can be used to study the state of the biological system and draw comprehensive conclusions about the mechanism of action.

2.2. Quality check

2.2.1. DNA/RNA quality check

The output of microarray data critically depends on the quality of the sample hybridized onto the microarray ***(6–8)***. Although several different quality criteria for RNA integrity and microarray data quality have been formulated, a clear consensus has not been achieved. Despite this, using a pure and intact sample is crucial for obtaining reliable data.

The assessment of nucleic acid quality can be achieved through several methods. For instance, spectrophotometers and fluorometers can help to assess the quantity and purity of the nucleic acid while electrophoresis-based methods are typically used for more detailed quality evaluation. This type of quality control is routinely performed by running a traditional agarose gel electrophoresis or through an automated electrophoresis system that is often combined with data processing software providing numeric values that represent sample quality. The bands on the denaturing

agarose gel are evaluated to assess the presence of typical patterns of intact RNA. Similarly, the automated systems provide corresponding graphs that can be used for evaluation alongside the provided quality value, such as the RIN or RQN values for RNA in the case of Agilent Bioanalyzer and Fragment Analyzer, respectively.

2.2.2. Data quality check

2.2.2.1. Chip image analysis

The quality assessment of the data starts by evaluating the quality of the hybridization. This can be carried out through carefully inspecting the chip image post scanning to address the presence of smears, dark spots, or other irregularities in the image. Feature extraction process transforms the scanned image into raw data files comprising computable values and is generally performed with software coupled with the microarray scanner. This step can further provide a quality report summarizing aspects of the hybridization to evaluate its success and consistent quality across arrays.

Given satisfactory hybridization results, the quality assessment continues with a more thorough inspection of the data.

2.2.2.2. Data Quality check

As discussed, gene expression and methylation estimates measured through microarray experiments can be significantly affected by different sources of systematic and random technical noise that may occur at different stages of the experimental procedure ***(9)***. For this reason, Quality Check (QC) methods aim to homogenize the shape of gene expression distributions, to identify RNA/DNA degradation signals, and to increase the robustness of probe intensity measures across

different samples. Several QC methods have been proposed which are often based on visual inspection of the data ***(10)***. We will outline the main QC methods to consider in the next paragraphs.

2.2.2.3. Expression specific data quality check

In this paragraph, we outline a short list of visual methods for quality check of gene expression microarray data.

MA plot:

MA plot is a common visualization method used for general QC purposes. It is widely applied to represent probe intensity distributions of two samples or groups of samples. In the MA plot, the x axis shows the average probe intensities of the compared samples, denoted as A, while on the y axis the absolute difference (in log2 scale) of the probe intensities, denoted as M is shown. The simplest application of the MA plot is the comparison of the intensities coming from the Cy3 and Cy5 dyes. Overall, the two intensities are supposed to be constant. Significant variations or aberrations between the trend of the intensities arising from the two dyes may suggest a bias occurred during the preparation of the experiment. The MA plot is a precious instrument that can be used before and after the normalization, in order to check the correctness of the procedure. Since the majority of the genes are supposed to have constant intensity values across the samples, it is expected that all the points would lie on the zero of the y axis. Variations from this trend suggest that a normalization is necessary, and one can follow the data transformation through the inspection of the MA plot.

[Figure 2 near here]

Distance-based methods:

The quality of microarray experiments can be further evaluated using so-called distance-based methods. These methods allow the comparison among several arrays by computing pairwise distances among them and representing such distances through specific exploratory plots. For instance, one of the most utilized graphs is the heatmap of the mean absolute difference, where the pairwise distances among samples are expressed by a color scale and a dendrogram shows how the samples cluster together on the base of such distances. This method is very useful if the aim is to identify outliers ***(11)***. Furthermore, biological replicates are expected to cluster together giving indication of potential batch effects in the data if, for instance, samples are observed to cluster based on the processing date instead. Another useful visualization falling into this category is the horizontal barplot which summarizes the pairwise distances among samples. In this graph, a threshold is determined based on the distribution of values across all the arrays and represented by a vertical line. Arrays with the summarized distance larger than the threshold are considered outliers ***(1)***.

2.2.2.4. Methylation specific data quality check

Methylation arrays have their own set of quality control plots to assess the performances of different reactions using different sets of dedicated probes. Illumina control probes can be broadly divided into two types: i) Sample-Independent Controls and ii) Sample-Dependent Controls. Sample-Independent Controls are used to evaluate the quality of specific steps in the process flow while Sample-Dependent Controls are used to evaluate reaction performance across samples.

Sample-Independent Controls include:

- Staining controls, used to measure the efficiency and sensitivity of the staining step. These controls can be used to compare background and signal for Staining control in the red and green channel.
- Extension controls, used to measure the extension efficiency of A, T, C, and G nucleotides from a hairpin probe, in the red and green channel.
- Target removal controls, test the efficiency of the stripping step after the extension reaction.
- Hybridization control, test the overall performance of the Infinium Assay using synthetic targets, perfectly complementing the sequence on the array instead of amplified DNA. Synthetic targets are present in the hybridization buffer at three concentrations: low, mid and high. This control should be monitored only in the green channel.

Useful representations for this set of controls are scatter plots or bar plots for the signal of each sample or scatter plot of the median intensity of the M channel against the median intensity of the U channel as implemented in the Minfi package ***(12)***.

Sample-Independent Controls include:

- Bisulfite conversion control measures the efficiency of bisulfite conversion of the genomic DNA.
- Specificity controls check for non-specific detection of methylation signals over an unmethylated background.
- Negative control probes define the system background and provide a comprehensive measurement of background, including signal resulting from cross-hybridization, as well as non-specific extension and imaging system background.

- Non-polymorphic controls test the overall performance of the assay, from amplification to detection, by querying a particular base in a non-polymorphic region of the bisulfite genome. They let us compare assay performance across different samples.

Useful representations for this set of controls are scatter plots or bar plots for the signal of each sample or scatter plot of the median intensity of the M channel against the median intensity of the U channel as implemented in Minfi ***(12)***.

2.2.2.5. Platform independent data quality check

The first set of QC plots are utilized for general purpose exploratory data analysis. These plots are employed for the representation of general properties of the microarray data and for outlier detection and management. In detail, they are computed on raw data and aim to give insight in the sample quality, the quality of the hybridization and the overall signals. Furthermore, they evaluate the comparability of signal strength and distribution within- and between-arrays, detecting deviating arrays (bias diagnostic), and assesses the correlation and grouping of samples based on the numeric array data. Alternatively, the next set of QC plots is computed for pre-processed (annotated and normalized) data and allows evaluating the performance of the normalization.

Histograms: Histograms are a common representation of probe intensity distributions in a microarray experiment. The assumption in a microarray experiment is that the measured intensities of the probes is directly proportional to the gene expression or methylation levels. For the purpose of representing the distribution of the gene expression/methylation estimates, the values are transformed in a logarithmic scale, in order to fit the distribution

to a normal one. In this way, the x axis reports the $\log_2$ transformed values of the gene expression/methylation estimates, while on the y axis the frequencies for that estimates in the experiment are reported.

[Figure 3 near here]

Density plots: Density plots represent an alternative visualization method to the histogram. It is broadly employed to obtain a smoothened view of the gene expression/methylation estimates in a particular experiment.

Sample quality is assessed by comparing the distribution of expression/methylation values among samples and between different probe-sets (e.g., green/red channels in methylation arrays). This can be done by means of density plots (over beta- (b) of M-values) and by grouping samples based on different technical and biological features. Samples (or groups of samples) deviating from expected distribution should be marked for further investigation and eventually removed.

[Figure 4 near here]

Scatterplots: Scatterplots are a very common graphical visualization method in order to highlight variation between-arrays. The values on the x and y axes of the scatter plot are log transformed intensity values of the compared samples. Scatterplot is a precious instrument in order to identify problematic arrays and to study, for instance, the consistency of technical or biological replicates.

[Figure 5 near here]

Dimensionality reduction: Dimensionality reduction techniques help identifying problematic samples/features by representing them in a low-dimensional space and annotating them using known biological technical grouping. Samples/features deviating from the expected grouping can be candidates for further investigation and eventually filtered out. When using dimensionality reduction techniques particular attention should be paid to the amount of variance of the original dataset associated to each used dimension and to the association between each dimension and known variables. Commonly used methods to perform dimensionality reduction on microarray data are MultiDimensional Scaling (MDS), Uniform Manifold Approximation and Projection (UMAP, Figure 6) and Principal Component Analysis (PCA).

[Figure 6 near here]

2.3. Filtering

2.3.1. Filtering

The hypothesis underlying microarray analysis is that the measured intensities for each arrayed gene represent its relative expression level. Biologically relevant patterns of expression are typically identified by comparing measured expression levels between different states on a gene-by-gene basis. However, before the levels can be compared appropriately, a number of transformations must be carried out on the data to eliminate questionable or low-quality measurements, to adjust the measured intensities to facilitate comparisons, and to select genes that are significantly differentially expressed between classes of samples ***(3)***. Probe filtering is the first step of such transformations and is intended to identify and remove probes on the array with

spurious values that do not represent the underlying expression/methylation state and can thus lead to incorrect results. Common strategies to address these values are described below.

2.3.2. Expression specific probe filtering

Since microarray technology relies on complementary nucleic acid hybridization, it is often prone to a level of noise arising from non-specific binding as well as optical noise from the background. Failing to address this noise during preprocessing can lead to biased results, hence low-intensity probes should be filtered out prior to normalization. Often, the non-specific binding shows little variation across samples, thus combining low intensity with low variability ***(13)***. In many gene expression microarrays, this can be addressed by comparing the signal intensity to the negative control probes present in the platform and investigating the signal variance between samples. Probes displaying intensities in a similar range as the negative control probes across the arrays are removed from the subsequent analysis. Another approach is based on the detection p-value which is commonly applied for Illumina gene expression and methylation microarrays (14). This approach is described in more detail under the context of methylation specific probe filtering, although a comparable strategy is applied for Illumina expression arrays.

2.3.3. Methylation specific probe filtering

Probe filtering approaches over methylation can be divided in two groups: technical probe filtering and biological probe filtering. Technical probe filtering is based on detection p-values, namely the probability that a methylation value reported for a probe is distinct from the background noise. The most commonly used p-value cut-offs in the literature range from 0.05 and 0.01. Biological probe filtering, on the other hand, is based on the biological property of the particular probe sequences used to capture methylated/unmethylated oligos that can make the associated values

unreliable in certain situations. Among the latter are i) the presence of a SNP on the interrogation or the extension site and ii) the localization of the probe on a heterosome.

The presence of a mutation on the interrogation or the extension site can alter the reliability of the measured value, common polymorphisms thus increase the probability of this event. The population frequency of the SNP, along with the knowledge of genetic background of the sample, are determinant factors in deciding if a probe hosting a SNP can be kept or should be discarded.

Probes that are localized on heterosomes show different signals between male and female subjects due to the different number of copies that are found on the two chromosomes.

The choice of whether probes localized on heterosomes are to be discarded is dependent on the sample sex composition and the planned comparisons. In fact, comparison of probes on the heterosomes between male and female subjects could lead to spurious results if not adequately taken into account in the statistical model.

2.3.4. Platform independent filtering

While some filtering approaches are platform specific, some can be seamlessly applied in many expression and methylation array platforms. The first being the removal of cross hybridizing probes. This refers to probes that have a high probability of hybridization with different oligonucleotides thus lacking specificity for the desired target. Literature provides tools and lists for identifying cross-hybridizing probes in methylation arrays ***(15)*** as well as in expression array ***(16)*** experiments.

The second filtering approach that can be applied to any microarray platform is based on outlier detection, namely the inspection of the probe value distribution among samples and the identification of probes characterized by extreme values not accountable to any biological feature.

Boxplots and scatterplots are useful graphical tools to perform this kind of inspection. Automated outlier detection can be implemented by summarizing the distribution of values measured for each probe and marking outlier probes based on thresholds that can be based on percentiles (e.g., consider outliers all values outside the interval formed by the 2.5 and 97.5 percentiles) or the values outside the interval formed by the median, plus or minus 3 median absolute deviations or using dedicated statistical tests like the Grubbs's ***(17)*** or the Dixon's test ***(18)***.

2.4. Imputation

Missing measurements, due to missing signal or user filtering, can cause problems to successive downstream analyses when they require the input set of observations to be complete. Thus, depending on the analysis to be performed, a second version of the data matrix where missing data are imputed (i.e., replaced with substituted values) can be produced.

Several methods exist in literature to perform data imputation, spanning from general purpose methods (e.g., KNN, regression models, central tendency values, etc.) to methods specifically designed for data obtained from microarray experiments ***(19, 20)***.

2.5. Normalization

Normalization is a crucial step in the preprocessing of microarray data allowing robust and meaningful comparisons between molecular targets, samples, and experiments. The goal of data normalization is to adjust the signal intensities and to remove sources of variation that might affect the results, such as dye effects and hybridization artifacts ***(21)***. A plethora of different strategies for the normalization of microarray data have been proposed, yet the preferred method is highly dependent on the data. Some of the methods are also platform-specific, meaning that they are distinctly oriented to gene expression or methylation studies. Others are rather nonspecific for the

platform, and they can be employed in any kind of microarray analysis. More importantly, the normalization methods can be classified into “within-array normalization” and “between-array normalization” based on the data taken into account for the data transformation. In particular, the within-array normalization methods only take into account the statistical information of each single array ***(22)***. While this provides effective normalization between molecular targets within the array, separate arrays will likely not be comparable making reliable differential analysis virtually impossible. Conversely, in the between-array normalization, information about the statistical properties across all the arrays included in the study are retrieved and exploited in order to make them comparable for differential analysis and other multi-sample analyses (Figure 7). This is achieved through a number of statistical approaches outlined in the following paragraphs.

[Figure 7 near here]

2.5.1. Expression specific data normalization

As widely reported by Bilban et al. 2002 ***(23)***, the simplest expression microarrays specific normalization strategies are aimed at the identification of genes whose expression is stable across different arrays as well as not differentially expressed. A number of strategies have been formulated in order to identify such invariant genes. Among others, the global normalization was one of the most employed, before more sophisticated methods were developed. This approach is based on the assumption that the amount of mRNA labeled with Cy3 or Cy5 is comparable. Therefore, in DNA microarrays encompassing thousands of probes (i.e., Agilent 450K), the signals from both of the labeled probes should be averaged. In this way, the ratio of the means of the intensities across all the probes should be equal to one ***(24–27)***. However, these kinds of strategies are not preferred nowadays, since the number of invariant genes may not be wide enough to cover

the range of signal intensities, and therefore, not achieve a proper fitting for non-linear normalization ***(23, 28)***. On the other hand, more robust normalization methods have been developed in the recent years. One of these is the locally weighted scatterplot smoothing (LOWESS) method, which was used in the past for smoothing scatterplots in a weighted, least-squares fashion ***(29, 30)***. LOWESS takes into account the effects due to the probe intensities and partially corrects for background effects. Other variants of the LOWESS take also into account local effects (e.g., print-tip LOWESS). Such normalization method is commonly used for two-color arrays ***(31)***.

2.5.2. Methylation specific data normalization

Data normalization for methylation microarrays is used to normalize the values between-arrays in order to account for assay related variation but also within an array to account for the difference in the distribution of values derived from distinct types of probes used in the same array design (as in the case of type 1 and type 2 probes in some Illumina platforms). Commonly used methods for within-array data normalization are Subset-quantile Within-Array Normalization (*SWAN*) ***(32)***, that allows within-array normalization of Infinium type I and II probes by reducing the differences in their signal distribution, Beta-Mixture Quantile Method (*BMIQ*) ***(33)***, that works by adjusting the distribution of Infinium II probes using as reference the distribution of type I probes, Normal-exponential using Out-Of-Band probes (*noob*) ***(34)***, that performs background correction and dye-bias adjustments on raw intensity values and Regression on Correlated Probes (*RCP*) ***(35)*** that normalizes probe intensities by taking into account the probe spatial correlation methylation values.

Common between-array normalization methods, on the other hand, include Functional Normalization (*FunNorm*) ***(36)***, that allows to remove unwanted between-array variation by using a quantile normalization approach based on the set of control probes, *dasen* ***(37)***, that performs between-array background correction and normalization separately for each probe type and stratified quantile normalization (*pQuantile*) ***(12)***, that performs between sample quantile normalization by stratifying probes based on their genomic region.

As in many other cases, the choice of the normalization method to be applied depends on the characteristics of the dataset and on the kind of downstream analyses to be performed. Within-array techniques are useful to remove intrinsic differences linked to different chemistries implemented in 450k and Epic array, while between-array techniques help to reduce technical noise arising from the distribution of samples over different chips. Usually within-array approaches are preferred, especially when followed by the explicit application of methods to remove batch effects that include array differences among the others.

2.5.3. Platform independent data normalization

Beside technology-specific normalization methods, a number of statistical methods can be applied for DNA microarray platform-independent normalization tasks. For instance, a simple, but widely used method is the standardization of the data. The term “standardization” is generally meant to describe the transformation of the expression or methylation estimates into Z scores. In detail, this approach scales microarray data within each array, and the values for individual genes are expressed as a unit of SD from the normalized mean of zero ***(38)***. In this way, different arrays underlying distinct biological conditions become comparable for differential analysis. The

differences detected between-arrays are, therefore, differences between the Z scores (Z ratios). Another approach that is worth mentioning, is the quantile normalization. This method is one of the most widespread normalization methods ***(39)***, not only employed in microarray data preprocessing but also in high-throughput sequencing technologies. The aim of the quantile method is to make the probe intensity distributions across a number of arrays identical. This kind of normalization originated from the quantile–quantile plot representation, where the points of two data vectors having the same distribution form a straight diagonal line. This implies that we can give each array the same distribution by taking the mean quantile and substituting it as the value of the data item in the original dataset ***(31)***.

2.6. Batch effect estimation and correction

As previously discussed, microarray data is prone to undesirable variation arising from technical aspects, sample handling, and even environmental conditions. This type of systematic variation observed between groups of samples is commonly referred to as a "batch effect". While some batch effects cannot be completely avoided, their effect can be minimized through careful planning and technical execution, and further mitigated during data preprocessing, given the correct precautions have been taken during experimental design. Failing to consider and address batch effects can result in misleading conclusions as the technical bias can be strong enough to mask the true biological signal ***(40)***.

Each of the potentially interfering variables (processing date, sample handler, separate batches of reagents etc.) should be carefully reported in the metadata alongside the biological variables and exposure details (exposure time, dose, treatment, donor etc.) in order to observe meaningful patterns in the data. Additionally, batch effects can arise from hidden sources not explained by the reported variables. These types of unknown sources of variation can be estimated by running a

surrogate variable analysis (SVA) from the R library sva ***(41)***. SVA identifies variables that work as surrogates for the technical variation observed in the data. These variables can then be included in the metadata, after which they are handled similarly to the known variables.

Although powerful approaches to evaluating and correcting batch effects have been established, they can only be applied on variables that are not confounded with the biological signal or other variables of interest, as the goal is to only attenuate the variation associated with technical variables while retaining the biological variation. The relationship between variables can be assessed by evaluating their correlation, typically by using a confounding plot ***(40)***. The presence of batch effects, on the other hand, can be assessed by clustering the samples to reveal grouping related to sources of technical bias, as well as through a principal component analysis (PCA). PCA helps to quantify the effect of the technical variables by breaking down the data into principal components that work as estimates for the patterns observed in the data. If batch effects independent from biological variables are observed, they can be corrected to alleviate the technical noise in the data. Methods that tackle batch effects in high-throughput data include various statistical methods such as the use of linear mixed models (LMM) ***(42)*** and empirical Bayesian approach ***(43)***. The empirical Bayesian approach has been implemented as a method called ComBat in the R package sva ***(41)***. ComBat has been found to outperform most of the other methods proposed for batch effect correction ***(42, 44)***, which has led to its widespread use and popularity.

2.7. Probe annotation

The probes on the microarray chip represent known biological entities, such as genes or CpGs associated to them, and need to be mapped to commonly used annotations to allow meaningful interpretations. While the sequences of the oligonucleotides on the array are rarely subject to

changes, the genome assemblies are revised as the knowledge expands, requiring frequently updated sources. Probe annotations are platform specific, which can sometimes lead to discrepancies between microarray platforms. However, a general approach to probe annotation is the use of specific annotation files that can be provided and updated by the manufacturer (Agilent) or available through databases such as Brainarray (http://brainarray.mbni.med.umich.edu/, last accessed in March 2021), which provides up-to-date annotations for Affymetrix microarrays. Additionally, platform specific annotation packages are available for R, such as the Bioconductor AnnotationDbi ***(45)*** and Illumina Human Methylation EPICanno ***(46)***.

During the annotation phase, probes are mapped to known molecular entities represented in the chosen format (e.g., Ensembl IDs, Entrez gene IDs, symbols). Probes mapped to the same entity are typically aggregated, meaning that the values obtained for each of the replicated probes are combined, for instance, based on their median values.

2.8. Data representation for expression and methylation microarrays

The preprocessing steps outlined in this chapter are aimed at moderating technical biases and making microarray data coming from different biological samples comparable. An important aspect to take into consideration when evaluating gene expression or methylation estimates is the way in which we represent their differences among samples. In expression microarray experiments, the expression differences can be represented by the Expression Ratio (ER=geneA/geneB). However, this representation is not very intuitive in case the genes are up- or down-regulated, since in case of up-regulation, the scale spans from 1 (no change in gene expression) to infinite, while in case of down-regulation the value spans from 0 to 1 ***(47)***. For this reason, the logarithmic transformation of the expression ratio ($\log_2$(ER)) is almost universally used as a gene expression representation for microarray data. In fact, by reporting the expression ratio

in a logarithmic scale, the expression ratios are much more intuitively interpretable, being distributed symmetrically around zero (up-regulated genes will have positive values whereas down-regulated ones will have negative values) ***(47)***. Moreover, by applying a log transformation, the data will be projected on a normal distribution, making them exploitable for differential expression analysis based on linear regression.

Regarding methylation estimates, they are in general represented in two alternative ways: beta values and M-values. Both of the methods are widely employed having, however, their strengths and weaknesses ***(48)***. Beta values reflect the formula beta=*meth*/*meth*+*unmeth*+α, being *meth* the value of the methylated probe intensities, *unmeth* the value of the unmethylated probe intensities and α an offset generally posed equal to 100 in order to stabilize *meth* and *unmeth* in case they are very small ***(49)***. Beta values range between 0 and 1 and can be interpreted as the percentage of methylation. For these characteristics, the beta values do not comply with the normal distribution assumption on which many statistical methods for differential analysis are based.

On the other hand, M-values reflect the formula $M=\log_2(meth/unmeth)$, so, in its formulation is very similar to the abovementioned log ratio used in gene expression studies, since it represents the log scaled ratio between methylated versus unmethylated probes. This makes the M-values suitable to be utilized in differential methylation analyses. A relationship exists between beta and M-values and it is expressed by the following formula: $M=\log_2(beta/1\text{-}beta)$, so it is quite straightforward to transform beta values into M-values and vice versa. As a general comment, beta values provide a more intuitive biological significance to the methylation estimates, since it expresses the percentage of methylation of a certain CpG site. M-values, instead, suffer from a lower biological interpretability, but they are more robust than beta values in statistical terms. Therefore, the use of M-values is highly recommended for differential methylation analysis.

3. Differential testing

Differential analysis aims at identifying genes (in case of gene expression) or CpGs (in case of DNA methylation) whose values show a statistically significant difference between two experimental conditions (e.g., treated vs. untreated samples). The classical analysis for microarray experiments uses linear models to estimate the variability between the experimental conditions (also defined as contrasts) and its covariate dependencies, which can be distinguished from the random variation. Classical covariates can be experimental variables (e.g., slide, array, date of sample handling) or biological variables (tissue/cell type, dose, time). The log-fold-change is a quantitative estimation of the differences between the gene expression values in two experimental conditions. For every gene or CpG, the log-fold-change is computed as the ratio between the mean value of the samples belonging to the first condition *vs.* the mean value of the samples in the second condition. A positive log-fold-change indicates a positive increase in the gene expression or DNA methylation in the first condition *vs.* the second condition, while a negative log-fold-change indicates a decrease in the gene expression or DNA methylation, respectively. Moreover, the log-fold-change is usually associated with a p-value that estimates its statistical significance. Classical thresholds for identifying differentially expressed genes are absolute log-fold change > 0.58 and p-value < 0.05. A typical method used to graphically represent the log-fold changes and p-values after a differential analysis is the volcano plot (Figure 8).

Due to the high dimensionality of microarray data, the p-values are usually corrected to decrease the risk of errors. The strongest correction technique is the Bonferroni methodology, which divides the p-values by the number of tests (in this case the number of genes/molecules). A more sophisticated and less conservative correction is the false discovery rate (FDR) which addresses

the proportion of false positive in the dataset. A classical differential expression analysis can be conducted by using the R library Limma from the Bioconductor repository ***(50)***.

[Figure 8 near here]

4. Conclusions

DNA microarrays represent a milestone in the history of biomedical science. This technology opened the doors to the so-called omics data, which have revolutionized the way of thinking to profile a molecular compartment of the cell, such as the transcriptome or the methylome. However, the nature of data coming from a DNA microarray experiment posed several challenges, starting from the assessment of standard quality check procedures to data transformation. Over time, countless algorithms and methods were developed in order to carry out a rigorous data preprocessing and assure reliable downstream analysis. These procedures are mostly devoted to the mitigation of biases that might occur during the different steps of the preparation of the experiment. In this chapter, we outlined the main principles and techniques employed in microarray preprocessing, starting from the experimental design of a microarray experiment to differential analysis, focusing our attention mainly on the transcriptome and DNA methylation analysis.

References

1. Marwah VS, Scala G, Kinaret PAS, et al. (2019) eUTOPIA: solUTion for Omics data PreprocessIng and Analysis. Source Code Biol Med 14:1. doi: 10.1186/s13029-019-0071-7
2. Rudy J, Valafar F (2011) Empirical comparison of cross-platform normalization methods for gene expression data. BMC Bioinformatics 12:467. doi: 10.1186/1471-2105-12-467
3. Quackenbush J (2002) Microarray data normalization and transformation. Nat Genet 32 Suppl:496–501. doi: 10.1038/ng1032
4. Stoughton RB (2005) Applications of DNA microarrays in biology. Annu Rev Biochem 74:53–82. doi:

10.1146/annurev.biochem.74.082803.133212
5. Risso D, Ngai J, Speed TP, Dudoit S (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol 32:896–902. doi: 10.1038/nbt.2931
6. Tumor Analysis Best Practices Working Group (2004) Expression profiling--best practices for data generation and interpretation in clinical trials. Nat Rev Genet 5:229–237. doi: 10.1038/nrg1297
7. Wilkes T, Laux H, Foy CA (2007) Microarray data quality - review of current developments. OMICS 11:1–13. doi: 10.1089/omi.2006.0001
8. Raman T, O'Connor TP, Hackett NR, et al. (2009) Quality control in microarray assessment of gene expression in human airway epithelium. BMC Genomics 10:493. doi: 10.1186/1471-2164-10-493
9. Lee E-K, Park T (2007) Exploratory methods for checking quality of microarray data. Bioinformation 1:423–428. doi: 10.6026/97320630001423
10. Eijssen LMT, Jaillard M, Adriaens ME, et al. (2013) User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. Nucleic Acids Res 41:W71-6. doi: 10.1093/nar/gkt293
11. Kauffmann A, Gentleman R, Huber W (2009) arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics 25:415–416. doi: 10.1093/bioinformatics/btn647
12. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30:1363–1369. doi: 10.1093/bioinformatics/btu049
13. Federico A, Serra A, Ha MK, et al. (2020) Transcriptomics in toxicogenomics, part II: preprocessing and differential expression analysis for high quality data. Nanomaterials (Basel). doi: 10.3390/nano10050903
14. Du P, Kibbe WA, Lin SM (2008) lumi: a pipeline for processing Illumina microarray. Bioinformatics 24:1547–1548. doi: 10.1093/bioinformatics/btn224
15. Chen Y, Lemire M, Choufani S, et al. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 8:203–209. doi: 10.4161/epi.23470
16. Uva P, de Rinaldis E (2008) CrossHybDetector: detection of cross-hybridization events in DNA microarray experiments. BMC Bioinformatics 9:485. doi: 10.1186/1471-2105-9-485
17. Grubbs FE (1969) Procedures for Detecting Outlying Observations in Samples. Technometrics 11:1–21. doi: 10.1080/00401706.1969.10490657
18. Dean RB, Dixon WJ (1951) Simplified statistics for small numbers of observations. Anal Chem 23:636–638. doi: 10.1021/ac60052a025
19. Faisal S, Tutz G (2017) Missing value imputation for gene expression data by tailored nearest neighbors. Stat Appl Genet Mol Biol 16:95–106. doi: 10.1515/sagmb-2015-0098
20. Lena PD, Sala C, Prodi A, Nardini C (2020) Methylation data imputation performances under different representations and missingness patterns. BMC Bioinformatics 21:268. doi: 10.1186/s12859-020-03592-5
21. Park T, Yi S-G, Kang S-H, et al. (2003) Evaluation of normalization methods for microarray data. BMC Bioinformatics 4:33. doi: 10.1186/1471-2105-4-33
22. Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31:265–273. doi: 10.1016/s1046-2023(03)00155-5
23. Bilban M, Buehler LK, Head S, et al. (2002) Normalizing DNA microarray data. Curr Issues Mol Biol

4:57–64.
24. Marton MJ, DeRisi JL, Bennett HA, et al. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301. doi: 10.1038/3282
25. Alizadeh AA, Eisen MB, Davis RE, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511. doi: 10.1038/35000501
26. Ross DT, Scherf U, Eisen MB, et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24:227–235. doi: 10.1038/73432
27. Yue H, Eastman PS, Wang BB, et al. (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Res 29:E41-1. doi: 10.1093/nar/29.8.e41
28. Tseng GC, Oh MK, Rohlin L, et al. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 29:2549–2557. doi: 10.1093/nar/29.12.2549
29. Berger JA, Hautaniemi S, Järvinen A-K, et al. (2004) Optimized LOWESS normalization parameter selection for DNA microarray data. BMC Bioinformatics 5:194. doi: 10.1186/1471-2105-5-194
30. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74:829–836. doi: 10.1080/01621459.1979.10481038
31. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193. doi: 10.1093/bioinformatics/19.2.185
32. Maksimovic J, Gordon L, Oshlack A (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13:R44. doi: 10.1186/gb-2012-13-6-r44
33. Teschendorff AE, Marabita F, Lechner M, et al. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29:189–196. doi: 10.1093/bioinformatics/bts680
34. Triche TJ, Weisenberger DJ, Van Den Berg D, et al. (2013) Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 41:e90. doi: 10.1093/nar/gkt090
35. Niu L, Xu Z, Taylor JA (2016) RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics 32:2659–2663. doi: 10.1093/bioinformatics/btw285
36. Fortin J-P, Labbe A, Lemire M, et al. (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol 15:503. doi: 10.1186/s13059-014-0503-2
37. Pidsley R, Y Wong CC, Volta M, et al. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 14:293. doi: 10.1186/1471-2164-14-293
38. Cheadle C, Vawter MP, Freed WJ, Becker KG (2003) Analysis of Microarray Data Using Z Score Transformation. J Mol Diagn 5:73–81. doi: 10.1016/S1525-1578(10)60455-2
39. Qiu X, Wu H, Hu R (2013) The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC Bioinformatics 14:124. doi: 10.1186/1471-2105-14-124
40. Leek JT, Scharpf RB, Bravo HC, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11:733–739. doi: 10.1038/nrg2825
41. Leek JT, Johnson WE, Parker HS, et al. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28:882–883. doi:

10.1093/bioinformatics/bts034
42. Espín-Pérez A, Portier C, Chadeau-Hyam M, et al. (2018) Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data. PLoS One 13:e0202947. doi: 10.1371/journal.pone.0202947
43. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8:118–127. doi: 10.1093/biostatistics/kxj037
44. Chen C, Grennan K, Badner J, et al. (2011) Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One 6:e17238. doi: 10.1371/journal.pone.0017238
45. Pagès H, Carlson M, Falcon S, Li N (2020) *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. R package version 1.52.0, https://bioconductor.org/packages/AnnotationDbi
46. Hansen KD (2016) *IlluminaHumanMethylationEPICanno.ilm10b2.hg19: Annotation for Illumina's EPIC methylation arrays*. R package version 0.6.0, https://bitbucket.com/kasperdanielhansen/Illumina_EPIC.
47. Babu MM (2004) Introduction to microarray data analysis. In Grant RP (ed) Computational Genomics: Theory and Application. Taylor & Francis
48. Du P, Zhang X, Huang C-C, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11:587. doi: 10.1186/1471-2105-11-587
49. Weinhold L, Wahl S, Pechlivanis S, et al. (2016) A statistical model for the analysis of beta values in DNA methylation studies. BMC Bioinformatics 17:480. doi: 10.1186/s12859-016-1347-4
50. Ritchie ME, Phipson B, Wu D, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47. doi: 10.1093/nar/gkv007

Figure Captions

Figure 1 - Microarray preprocessing workflow covered in this chapter. Steps 1–7 represent crucial points to consider for achieving high-quality data from microarray experiments regardless of the

selected platform and application. Pink box marks the input data for preprocessing, while the blue color indicates the output data.

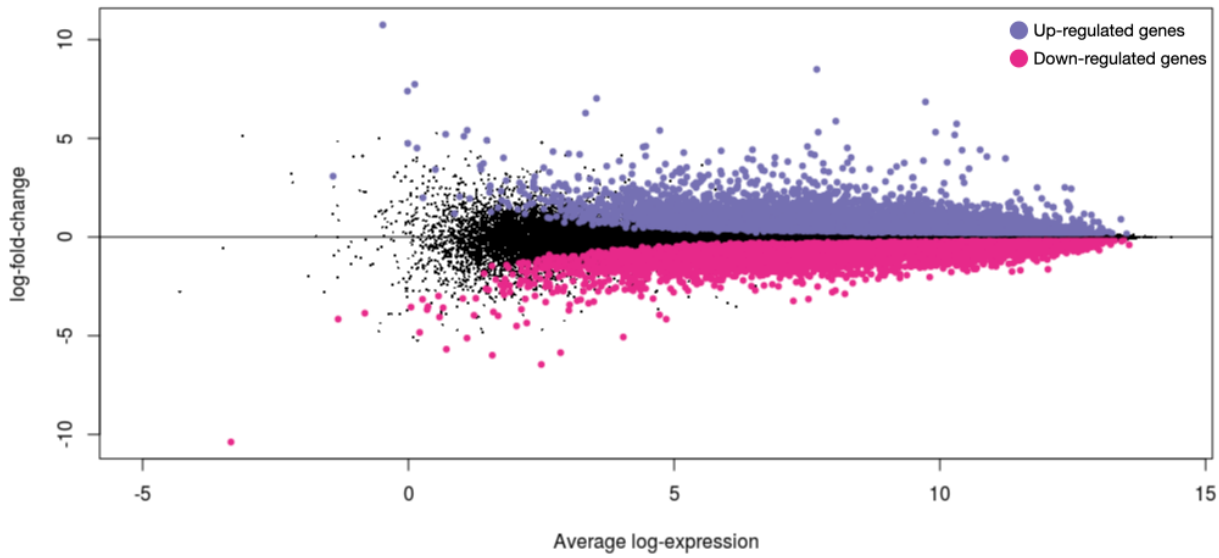

Figure 2 - MA plot representation of gene expression distribution between lesional and unaffected skin of patients affected by atopic dermatitis (AD). Microarray data for this and the following figures were retrieved from GEO (GEO Dataset ID: GSE34248).

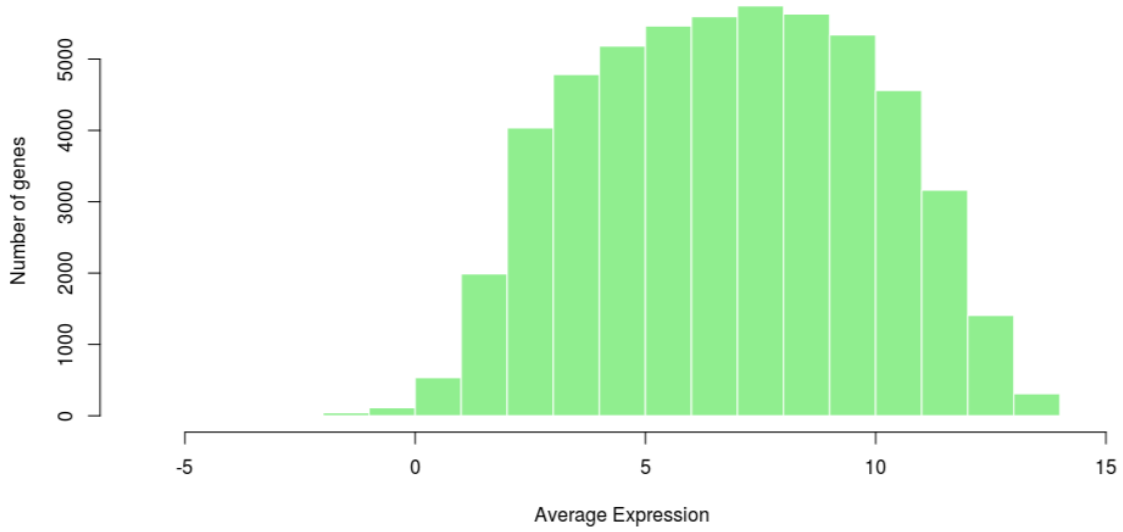

Figure 3 - Histogram showing the distribution of gene expression estimates in a microarray experiment.

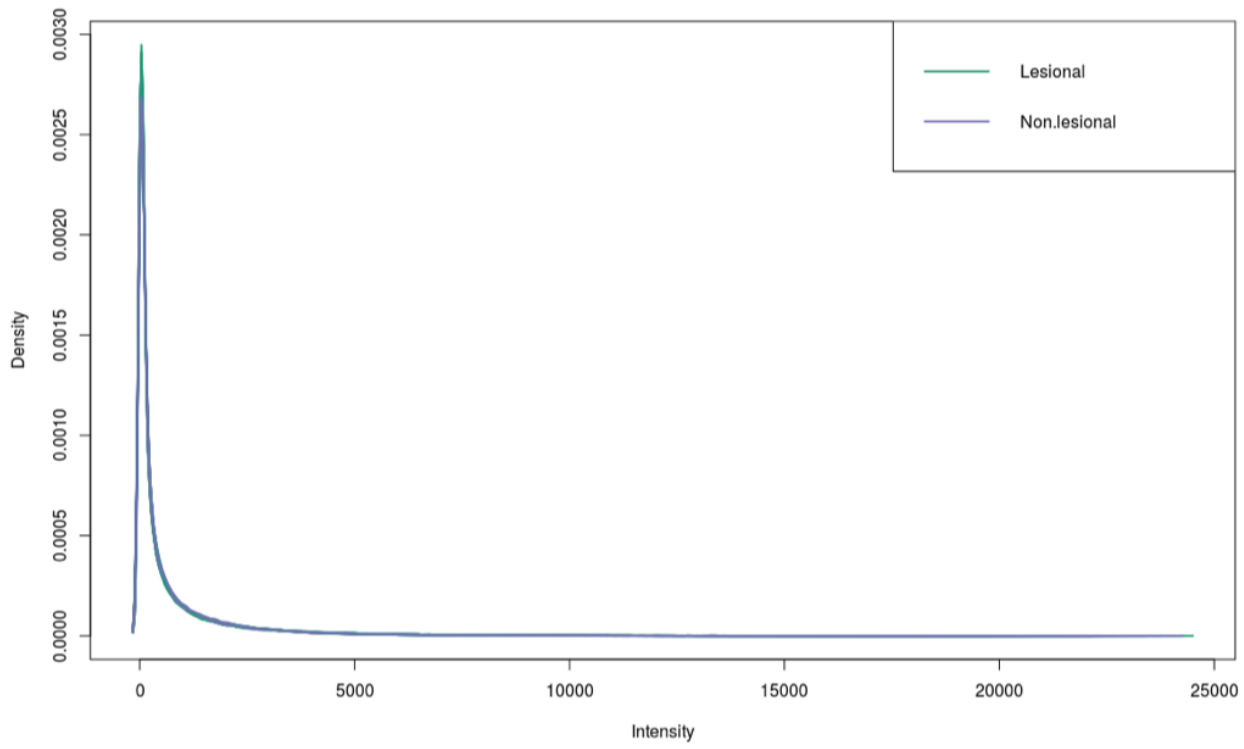

Figure 4 - Density plot of gene expression distribution between lesional and non-lesional skin of atopic dermatitis (AD) patients. The curves (mostly overlapping in this figure) indicate the amount of probes showing a certain intensity. In this experiment, most of the probes show a low intensity range.

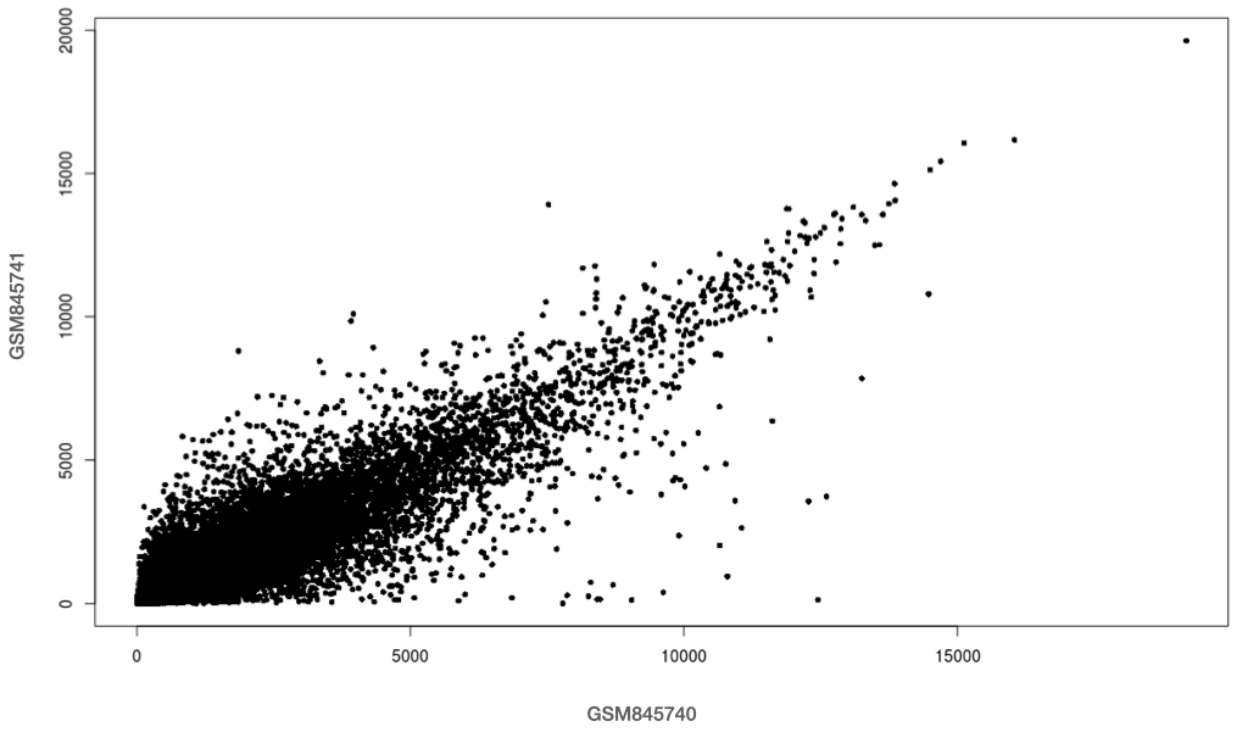

Figure 5 - Scatter plot showing the comparison between gene expression estimates of two samples of the dataset GSE34248.

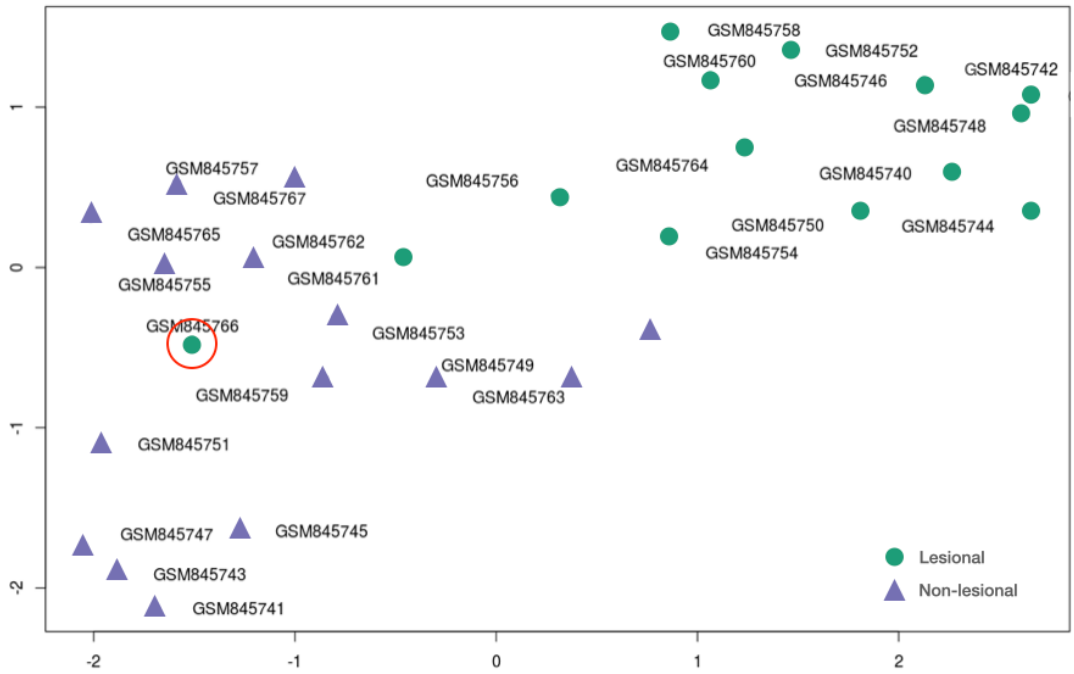

Figure 6 - Uniform Manifold Approximation and Projection (UMAP) plot showing the projection of samples in a low-dimensional space. Given its position in the plot, the sample rounded by the red circle could be a candidate outlier.

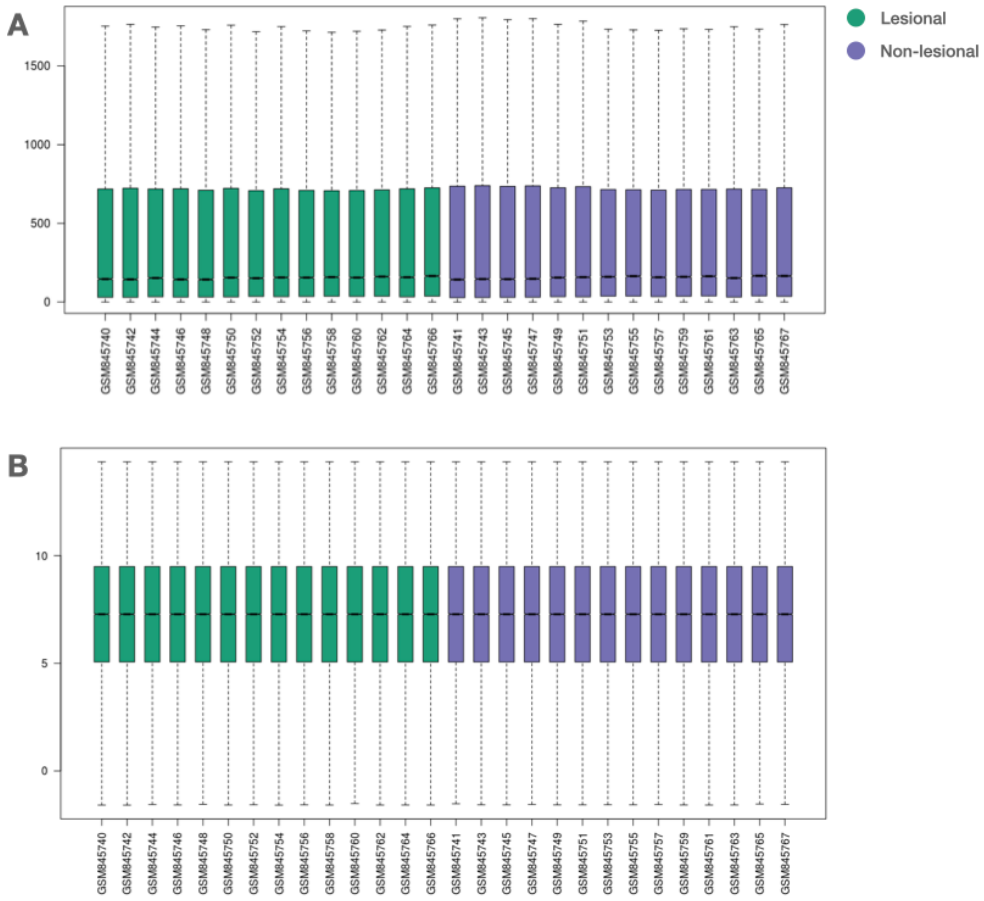

Figure 7 - Boxplot showing gene expression distribution of lesional and non-lesional skin of atopic dermatitis (AD) patients before (A) and after between-arrays normalization (B).

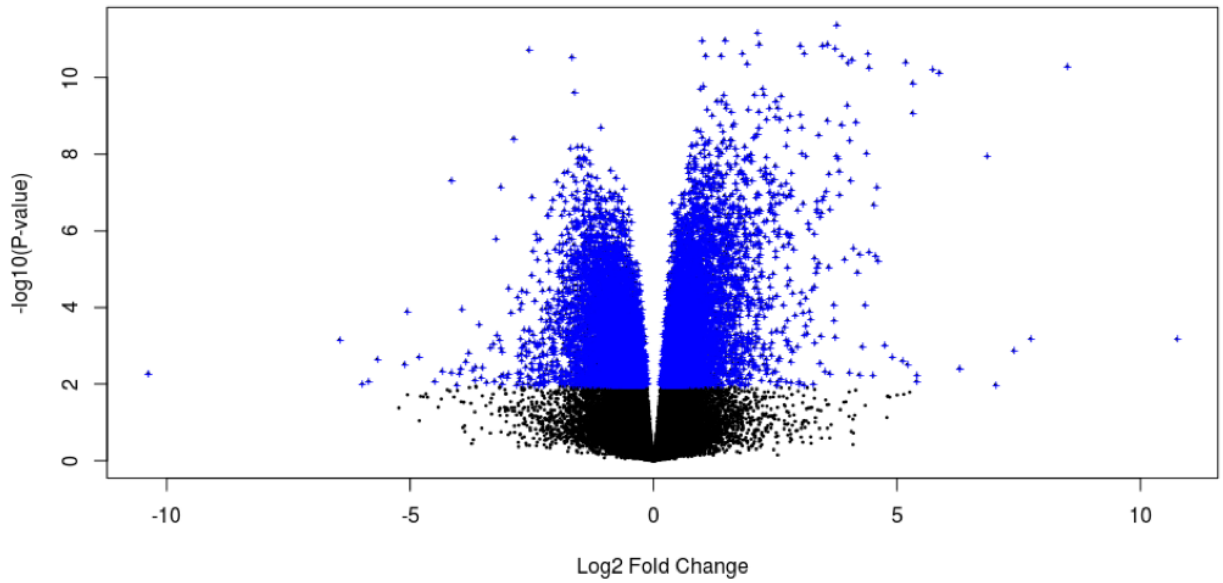

Figure 8 - Volcano plot showing the magnitude of deregulation in terms of log-fold changes and the p-values of the genes after differential analysis. The blue dots represent differential expressed genes.