Breast cancer microarrays analyses

Anaëlle Urio

1 Introduction

We report on an analysis of breast tumors mRNA with positive (ER+) or negative (ER-) estrogen-receptor statuses. Forty-nine samples are hybridized to Affymetrix GeneChip HG-U133A, containing 22 283 probe sets (genes).

First, a quality assessment is carried out to detect possible outliers. Then, after a normalization of the data, we compute expression measures for each gene to identify differential expression between the two tissues. We also carry out a cluster analysis on the arrays.

2 Quality Assessment

We first do some preliminary exploratory analysis. Figure 1 shows boxplots of perfect matches intensities for the different arrays. The $\log_2$ transformation of PM intensities is considered for more readability. There is no big changes among the arrays, hence no sign of outliers here.

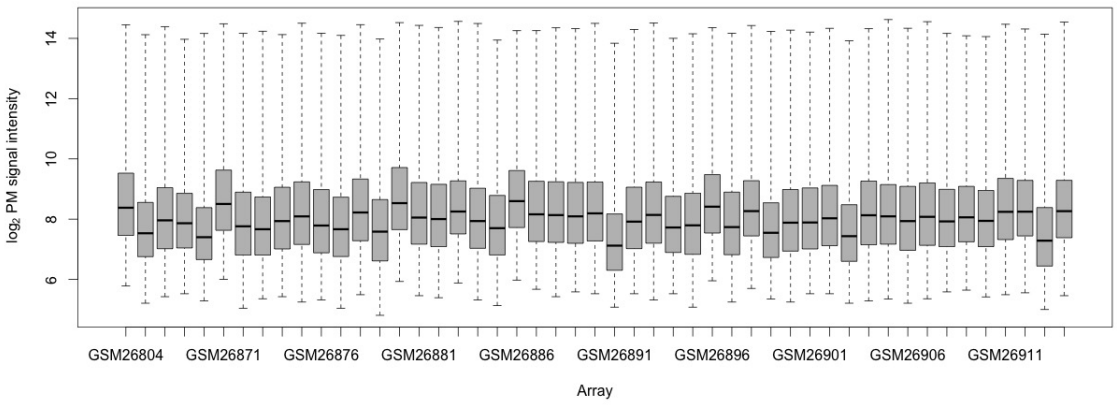


Figure 1: Boxplots of $\log_2$ PM signal intensities for each chip.

With the aim of computing a quality measures controls, we fit a probe level model (PLM) to the data, modeling the pre-processed signal intensities and fitting it using M-estimation following Bolstad 2004 [2]. This robust procedure is chosen because it lacks sensitivity to outliers.

More precisely, we assume the model

$$\mathrm{PM}^*_{kij} = C_{kj} + P_{ki} + \varepsilon_{kij}\,, \tag{1}$$

where the subscript j indicates array, k indicates probeset and i the probe in probeset k. PM^* are the pre-processed PM values, *i.e.* the $\log_2$ of the quantile normalized PM. Furthermore, C_{kj} is the $\log_2$ gene expression value on chip j for probeset k. Hence, C represents the array effect. P_{ki} is the effect of probe i in probeset k. For identifiability, we add the constraint $\sum_{i=1}^{I_k} P_{ki} = 0$. The ε_{kij} represent the errors.

The parameters of Model 1 are estimated by the iteratively reweighted least squares (IRLS) algorithm. To estimate the parameters, we solve the M-estimation

$$\min_{P_{ki},C_{kj}} \sum_{k,i,j} \rho\left(\frac{\log_2 \mathrm{PM}_{kij} - C_{kj} - P_{ki}}{\hat{\sigma}}\right), \tag{2}$$

where $\rho(x)$ is a loss function and $\hat{\sigma}$ is a robust estimate of scale. Here, we use the Huber loss function

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} & \text{for} |x| \leq \delta \\ \delta|x| - \dfrac{\delta^2}{2} & \text{for} |x| > \delta \,, \end{cases} \tag{3}$$

with δ a tuning constant that we fix at 1.345. Equation 2 does not have any explicit solution in general, hence we solve it using IRLS. We start with initial estimate of the parameters $\tilde{P}_{ki}^0$ and $\tilde{C}_{kj}^0$, obtained by ordinary least squares. Then, we compute iteratively the quantities

$$r_{kij}^{l+1} = \log_2 \mathrm{PM}_{kij} - \tilde{C}_{kj}^l - \tilde{P}_{ki}^l, \qquad \text{(residuals)}$$

$$u_{kij}^{l+1} = \frac{r_{kij}^{l+1}}{\hat{\sigma}}, \qquad \text{(rescaled residuals)}$$

$$w_{kij}^{l+1} = \frac{\rho'(u_{kij}^{l+1})}{u_{kij}^{l+1}} \,. \qquad \text{(weights)}$$

Then, we compute the next estimators of P_{kj} and C_{ki} by weighted least squares. The weight function w described above gives low weights to outliers.

Figure 2 shows pseudo-images of the weights for some of the arrays. Low weights are colored green while high weights are colored light gray. Figure 2(h) is the pseudo-images of Array GSM26906 which is representative of most of the chips. Chip GSM26870 has lower weights than the other chips but no identifiable pattern (Figure 2(a)). Hence, we can not conclude that this array is an outlier for now. Chip GSM26903 (Figure 2(f)) has low weights in the top left of its pseudo-image. This is a diffuse, rather weak, artifact of the hybridization. Other chips in Figure 2 have locally low weights which could be the effects of the carry-over from the printing process.

Two other measures of quality are the relative log expression (RLE) and normalized unscaled standard error (NUSE) values. The RLE is the expression values of each array relative to the virtual median chip

$$RLE(\tilde{C}_{kj}) = \tilde{C}_{kj} - \underset{j}{med}(\tilde{C}_{kj}) \,, \tag{4}$$

where k indexes probeset (gene) and j indexes chip. The median RLE value for each chip should be around 0 and the RLE values should not have a large variability. Figure 3 shows no chips with more variability than the other.

We also consider the NUSE values. The unscaled standard error is defined as

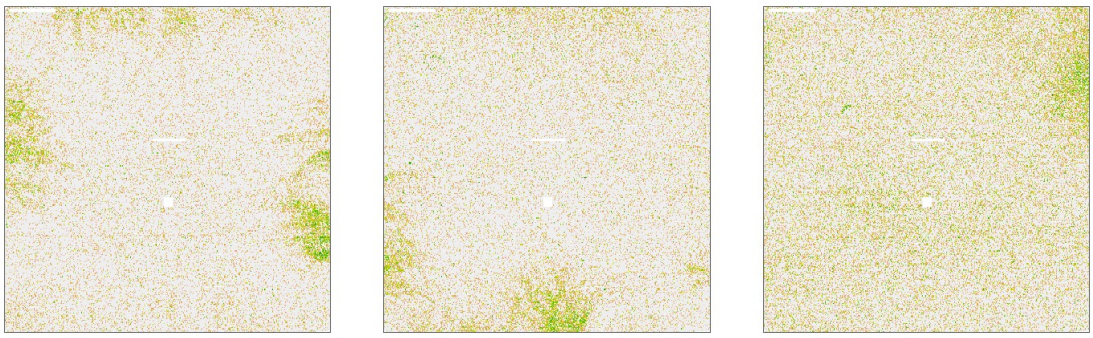
$$USE(\tilde{C}_{kj}) = \frac{1}{\sqrt{\sum_i w_{kij}}} \,, \tag{5}$$

where again k refers to probeset (gene) and j indexes chip. The NUSE is then the unscaled standard error normalized by dividing by the median (across chips) of each standard error:

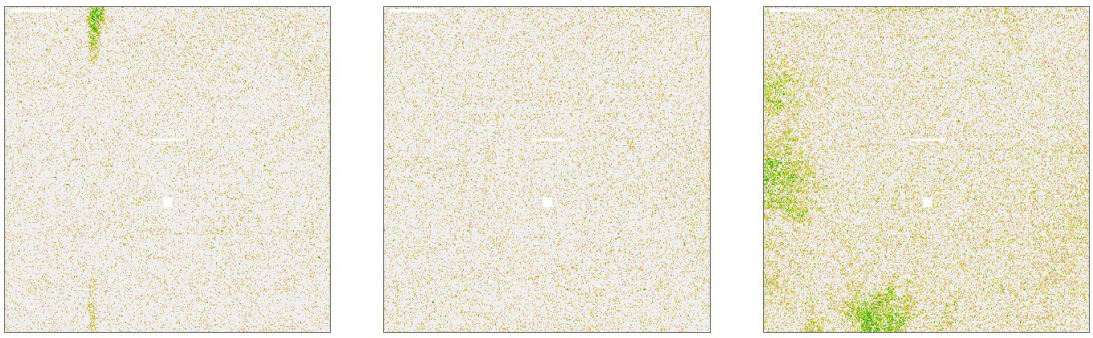
$$NUSE(\tilde{C}_{kj}) = \frac{USE(\tilde{C}_{kj})}{\underset{j}{med} \frac{1}{\sqrt{\sum_i w_{kij}}}} \,. \tag{6}$$

The NUSE values should fluctuate around 1. A chip with a median NUSE value larger than 1.05 is an outlier and should be excluded. Orange boxplots on Figure 4 are the ones with their 75% values above 1.05. They correspond to respectively Chips GSM26870, GSM26903, GSM26910 and GSM26914. The two first were already identified as having low weights in their pseudo-images. Despite that, no chip has its median NUSE value above the threshold 1.05. Hence, we can not identify any outlier and keep all the arrays for all the further analyses.

(a) GSM26870 (b) GSM26880 (c) GSM26885



(d) GSM26890 (e) GSM26901 (f) GSM26903



(g) GSM26904 (h) GSM26906 (i) GSM26908

Figure 2: Pseudo-images of the weights of nine chips. Colors: green for low weights and light gray for high weights.

3 Gene Expression Quantification

To quantify gene expression, we use the robust multichip average (RMA), (Irizarry et al. [3], [4]). RMA consists of three steps: background correction, quantile normalization and probeset summarization.

Background correction is performed to remove the effects of background noise signals caused by experimental procedures. The PM signal intensity is modeled as $\mathrm{PM}_{kij} = s_{kij} + bg_{kij}$, where s_{kij} is the true signal intensity in array j of probe i into probeset k and bg_{kij} is the experimental background effect. We assume that s_{kij} follows an exponential distribution with parameter α and the error term bg_{kij} follows a normal distribution with parameters μ and σ^2. The parameters are estimated from the data. Under these assumptions, the background correction of the PM can be write as

$$B(\mathrm{PM}_{kij}) = \mathbb{E}[s_{kij}|\mathrm{PM}_{kij}]\,. \tag{7}$$

Normalization is carried out to remove the effects of artifacts, thereby reducing the variation

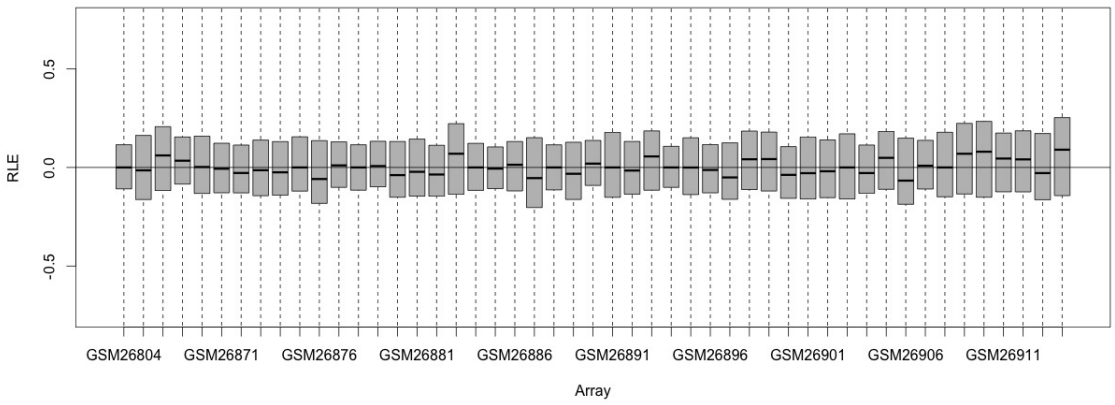


Figure 3: Boxplots of RLE values for each chip.

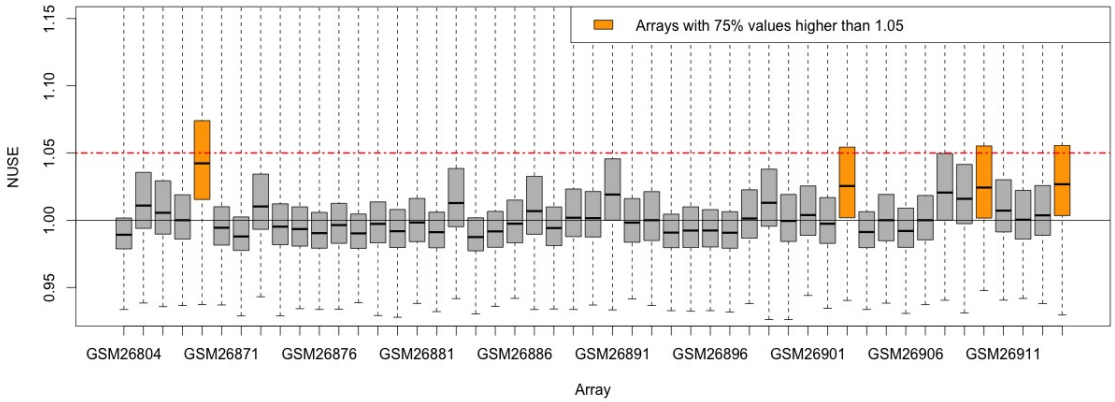


Figure 4: Boxplots of NUSE values for each chip with threshold red dashed line at 1.05.

between the arrays. RMA uses quantile normalization. Quantile normalization makes the distribution of probe intensities the same for each chip. The steps of the normalization are

1. create a matrix X of dimension $K \times J$ where each column represents an array with its K genes intensities;

2. compute X_o by ordering in a decreasing way each column of X;

3. compute the means of each row of X_o and replace each row by its row-mean to obtain X'_o;

4. compute X_{norm} by rearranging each column of X'_o to have the same ordering of the original X, *i.e.* the normalized genes intensities are again in the original ordering.

Columns of X_{norm} contains now the background-corrected, normalized PM signals. These values are then $\log_2$ transformed.

To obtain an expression measure, we fit an additive model to the transformed PM intensities. This model is the same as Model 1 above:

$$\mathrm{PM}^*_{kij} = C_{kj} + P_{ki} + \varepsilon_{kij}\,, \tag{8}$$

where PM^*_{kij} is the background-corrected, quantile normalized $\log_2$ PM value for probeset k, probe i and array j. For dataset GSE1561, we have $K = 22\,283$ genes and $J = 49$ arrays.

The estimates of the parameters are then obtained using a robust fitting method. By default, the robust method is median polish; however, we use M-estimation as in Section 2.

4 Identification of Differentially Expressed Genes

We aim to indentify genes that are differentially expressed between ER+ and ER- status. For this purpose, we fit a linear model for every probeset k:

$$\tilde{C}_k = \begin{pmatrix} \tilde{C}_{k1} \\ \vdots \\ \tilde{C}_{k49} \end{pmatrix} = X\beta_k + \varepsilon_k = X \begin{pmatrix} \beta_{1,k} \\ \beta_{2,k} \end{pmatrix} + \begin{pmatrix} \varepsilon_{k1} \\ \vdots \\ \varepsilon_{k49} \end{pmatrix}, \tag{9}$$

where $\tilde{C}_{kj}$ is the $\log_2$ value of gene expression (RMA value) from array j, X is the design matrix (drawn in Table 1 of Appendices) with the first column of 1's and the second column of 0's and 1's depending on the positive or negative ER status, $\beta_{1,k}$ is the coefficient representing ER+ status expression, $\beta_{2,k}$ is the coefficient for the difference in expression between ER- and ER+ status and ε_k is a vector of error terms. Model 9 is fit by ordinary least squares.

To assess differential expression between the two status, we test

$$H_{0,k} : \beta_{2,k} = 0 \quad \text{against} \quad H_{1,k} : \beta_{2,k} \neq 0\,,$$

for $k = 1, \ldots, K$. If the null hypothesis $H_{0,k}$ is rejected, we conclude that the level expression is different between the two statuses samples for gene k.

To test these hypotheses, we could use the ordinary t-statistic

$$t_k = \frac{\hat{\beta}_{2,k}}{SE(\hat{\beta}_{2,k})} = \frac{\hat{\beta}_{2,k}}{s_k\sqrt{v}}\,, \tag{10}$$

where s_k^2 is an estimate of the variance σ_k^2 and v is the second diagonal element of the matrix $(X^TX)^{-1}$. These t_k follow a t-distribution with d_k degrees of freedom.

In this t-test, the variance estimates are specific to each gene k and s_k^2 is of the form $\frac{1}{K-p}\hat{\sigma}^2$ with K the number of genes (22 283) and p the number of arrays (49). Hence, in a small sample size, the risk to have a lot of s_k near 0 is big. We then use a slightly modified standard error.

We apply an empirical Bayes approach as in Smyth 2004 [6], incorporating prior information about the variation across probesets of the coefficients $\beta_{2,k}$ and of the variances σ_k^2. We assume $\mathbb{P}[\beta_{2,k} \neq 0] = p_0$ is a known probability, where p_0 represents the expected proportion of truly differentially expressed genes. We also assume that σ_k^2 follows an inverse weighted χ^2 distribution

$$\frac{1}{\sigma_k^2} \sim \frac{1}{d_0 s_0^2}\chi^2_{d_0}\,, \tag{11}$$

where s_0^2 is the prior estimate with d_0 degrees of freedom. From these priors, we can compute the posterior variance

$$\tilde{s}_k = \frac{d_o s_0^2 + d_k s_k^2}{d_0 + d_k}\,. \tag{12}$$

We can now define the moderated t-statistic

$$\operatorname{mod} t_k = \frac{\hat{\beta}_{2,k}}{\tilde{s}_k\sqrt{v}}\,, \tag{13}$$

where the initial variance has been replaced by the posterior variance. This implies that the moderated t is the classical t-statistic if d_0 equals zero. The $\operatorname{mod} t_k$ statistic still follows a t-distribution, but now with $d_k + d_0$ degrees of freedom due to the extra information borrowed by the empirical Bayes procedure.

We simultaneously test the null hypotheses for all genes. Because of the large multiplicity issue, the chance of having at least one false positive result (Type I error) increases. If we

reject the individual null hypotheses at 5% level, we expect to have $22\,283 \cdot 0.05 = 1114.15$ false positive results if all nulls are true. But we are not only intersted by the question wether an error was made in the test but also by the number of wrong rejections we have done. The False Discovery Rate (FDR) is

$$\text{FDR} = \mathbb{E}[Q] \quad \text{where} \quad Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}, \tag{14}$$

with V the number of Type I errors which is unknown and R the known number of null hypotheses rejected. This gives the expected ratio of errors among the rejected nulls. We want to control the FDR and keep it at level 0.05. Hence, we introduce adjusted p-values.

Here, we consider the Benjamini-Hochberg (BH) method explained in Benjamini-Hochberg 1995 [1]. Suppose we order the K p-values and denote $\{r_k\}_{k=1}^K$ the ordering such that $|p_{r_1}| \leq |p_{r_2}| \leq \ldots \leq |p_{r_K}|$. The BH method is a step-up adjustment method, *i.e.* the p-values are adjusted from the largest to the smallest. Explicitly, we have

$$p^*_{r_j} = \min_{k=j,\ldots,K} \left\{ \min \left(\frac{K}{k} p_{r_k}, 1 \right) \right\}. \tag{15}$$

When we carry out the tests with the FDR at 5%, the BH method rejects 3647 nulls. Hence, we identify these 3647 genes as differentially expressed between the two statuses. We thus identify about 16% of the genes as differentially expressed.

Figure 5 shows a volcano plot with points corresponding to differentially expresssed genes colored in green. It plots the absolute value of the moderated t_k-statistics against $\log_2$ fold change in expression between testis and placenta, $\log_2 \hat{\beta}_{2,k}$. The horizontal red dashed line represents the 5% statistical significance threshold, *i.e.* the level where the p-values are equal to 0.05. The vertical red dashed lines are at -1 and 1, respectively when the ratio of gene expression between the statuses is of 1/2 and 2. As explained in Li 2012 [5], the volcano plot can be used to detect differentially expressed genes under two aspects. The first criterion can be the $\log_2$ fold-change level (here halfed and doubled) and the second the p-value level (here 5%). Then, either we can apply a single filtering criterion like we did with the p-value or we can consider the double filtering and take only the genes in the top corners as differentially expressed. The genes in bottom corners are often not taken into account because genes with high $\log_2$ fold-change but insignificant test results may be produced by few outliers with large values in ER+ or ER- status. The risk to take the genes with large moderated t but low $\log_2$ fold-change could be that they can be produced by a false signal due to variance near zero. But this point was already avoid by taking the moderated t-statistics and not the classical t-statistics. Here, since the area with large moderated t and low $\log_2$ fold-change contains the majority of the points with significant test, we decide to keep the single filtering criterion of significant test at level 5%. However, we keep in mind that a lot of the rejected points have low $\log_2$ fold-changes.

The top 50 differentially expressed genes are presented in Table 2 of Appendices.

5 Cluster Analysis

Finally, we carry out a cluster analysis on the 22 arrays with ER negative status. We take into account only the 100 genes with the largest variance across the arrays. If we find clusters well adjusted to the genes which are the most variable, the other genes should fit easily into these clusters.

There are several different ways to obtain clusters. First, we try a hierarchical clustering with an agglomerative method. An agglomerative method starts with as many clusters as samples

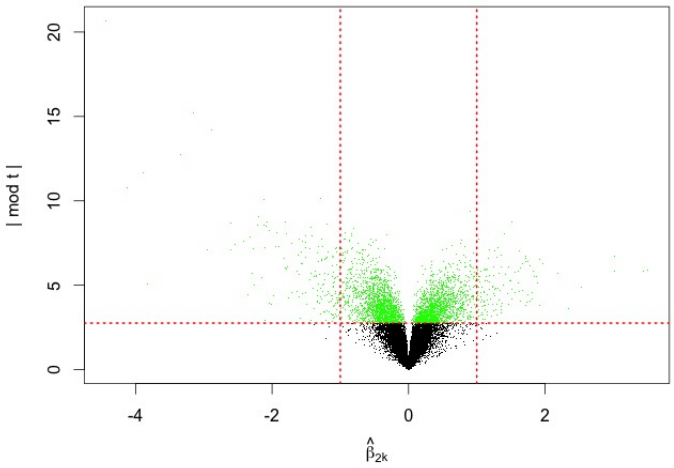


Figure 5: Volcano plot for genes of GSE1561 with BH rejected nulls in green. Horizontal red dashed line is the 5% statistical significance. Vertical red dashed lines are at -1 and 1, respectively halved or doubled gene expression change between ER status.

(here 22). At each step, the two closest clusters are merged using a between-object (or within-cluster) dissimilarity criterion. Here, we use the correlation as between-object dissimilarity

$$1 - \mathrm{Corr}(\tilde{C}_j, \tilde{C}_l)\,, \qquad \text{(correlation dissimilarity)}$$

where $\tilde{C}_j$ is the $\log_2$ gene expression value (RMA value) from array j. This dissimilarity is chosen because it does not take into account the variation of measurement between the different arrays.

There exist different methods to determine which clusters we merge together, *i.e.* on which criteria two clusters are the closest. Here, we use Ward's method based on

$$\triangle(A, B) = \frac{n_A n_B}{n_A + n_B} \left\| \vec{m}_A - \vec{m}_B \right\| , \tag{16}$$

where m_L is the center of the cluster L and n_L is the number of arrays in the cluster L. Hence, $\triangle(A, B)$ is the merging "cost" of combining clusters A and B. The clusters with the lowest Δ are merged. If there are two pairs of clusters whose centers are equally far apart, Ward's method preferentially merges the ones with the smaller number of elements.

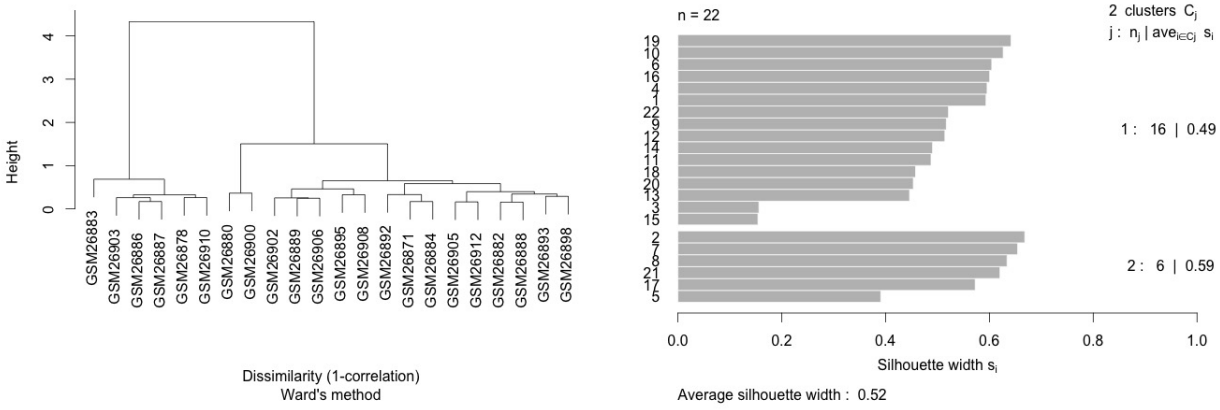
A measure to estimate if an array j is in an appropriate cluster is the silhouette

$$s_j = \frac{b(j) - a(j)}{\max\left(a(j), b(j)\right)}\,, \tag{17}$$

where $a(j)$ is the average between-cluster dissimilarity of array j and $b(j)$ is the lowest average dissimilarity of array j with the other clusters. A large value of $b(j)$ means that array j is far from the other clusters, thus, chip j would be well placed in its cluster. We note that $-1 \leq s_j \leq 1$. For s_j near 1, array j should be in the correct cluster. But if s_j is around 0, the clustering for array j is poor. Moreover, if s_j is negative, then array j is probably not in the right cluster.

Figure 6 shows the summary of the hierarchical clustering with correlation dissimilarity and Ward's method. The dendrogram on Figure 6(a) shows two distinct groups. The silhouette on Figure 6(b) has an average width of 0.52 which indicates a good structure.

Another way to do a clustering is using a partitioning method. These methods provide a grouping (partition) of the objects (samples) into a pre-determined number of clusters K. The clustering is obtained by satisfying an optimality criterion. The disadvantage of partition



(a) (b)

Figure 6: Summary of the hierarchical clustering with Ward's method and correlation dissimilarity. (a) Hierarchical dendrogram. (b) Silhouette.

methods is that an initial number K of clusters must be specified. For high-dimensional data, partitioning methods are computationally demanding.

Here, we use two partitioning methods. First, the K-means method minimizes the sum of squares from arrays to their assigned cluster

$$\min_{A_1,\ldots,A_K} \sum_{j=1}^{K} \sum_{x \in A_j} \left\| \tilde{C} - \mu_j \right\|^2, \tag{18}$$

where $\tilde{C} = (\tilde{C}_1, \ldots, \tilde{C}_{49})$ is the set of RMA values, $A_1, \ldots, A_K$ is the set of clusters that we want to determine and μ_j is the mean of objects in cluster S_j. Here, we run the K-means with K equals 2. We obtain an average silhouette width of 0.34. So this clustering does not gives better results than the hierarchical clustering.

Secondly, we use the partitioning around medoids (PAM) method. PAM selects K objects to be initial medoids. Then it associates each object to the closest medoid calculated with the between-cluster Manhattan dissimilarity. Define the total cost as follows

$$cost(\tilde{C}_j, M) = \sum_{k=1}^{100} \left| \tilde{C}_{kj} - M_k \right|, \tag{19}$$

where M is one of the initial medoid. Then, PAM exchanges the role of the medoid M with the non-medoid $\tilde{C}_j$ while the total cost decreases. Here, we compute the PAM with K equals to 2. The average silhouette width is about 0.34. Hence, the hierarchical clustering with Ward's method and correlation dissimilarity gives the best silhouette.

Finally, we display a heatmap of PLM weights for the arrays and genes (Figure 7). The dendrogram on x axis is the one from the hierarchical clustering on the arrays with correlation dissmilarity and Ward's method. The dendrogram on y axis of the 100 more variable genes is displayed on Figure 8. The list of 100 genes from the hierarchical clustering can be found in Table 3. Figure 7 shows a slight difference between the two array's clusters. The second one has lower weights than the first.

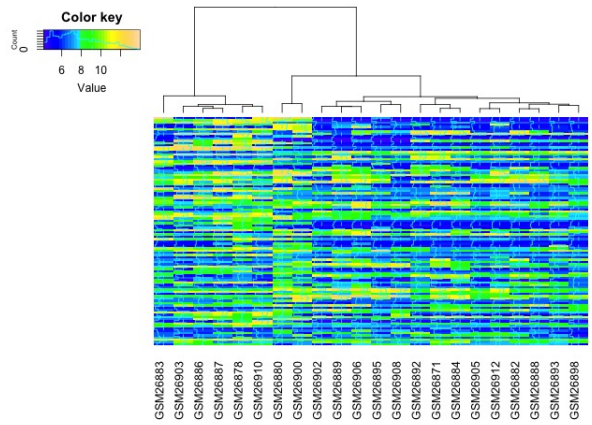


Figure 7: Heatmap of weights with hierarchical dendrograms of correlation dissimilarity and Ward's method for respectively arrays (x axis) and genes (y axis). Coloring: genes in dark blue are the ones with less weights and genes in dark yellow are the ones with large weights.

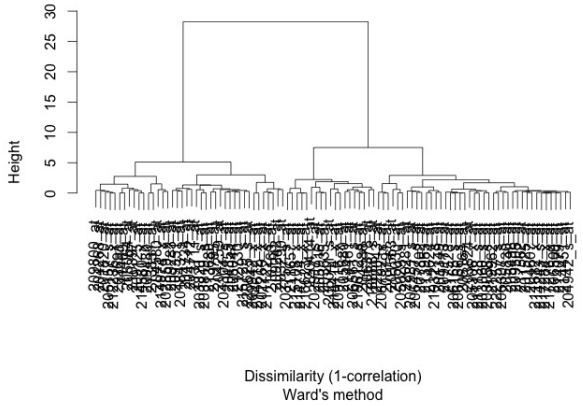


Figure 8: Dendrogram of the hierarchical clustering with Ward's method and correlation dissimilarity of the genes.

6 Conclusion

We first carried out a quality assessment on GSE1561. No outlier was detected. Only small atrifact effects were present on the weights of the probe level model but not enough to remove an array of the sample.

Then, the gene expression quantification was extract from the 49 arrays of the sample using RMA. The three steps applied to the PM signal intensities were the background correction, a quantile normalization and finally the probeset summarization.

The aim was then to identify differentially expressed genes between ER- and ER- status in the breast cancer cells. Therefore, a linear model was fit for every genes. The nulls were tested with the moderated t-statistic. To avoid a problem in multiple testing, we decided to control the FDR with Benjamini-Hochberg p-values. This method identified 3647 genes as differentially expressed, hence 16% of the total number of genes.

Finally, a cluster analysis was carried out on the samples with negative ER status and with the 100 genes with largest variance. We applied different clusterings to the data. The hierarchical clustering with Ward's method and correlation dissimilarity gave the higher silhouette width average of 0.52. This clustering gave two different clusters for the arrays with the second one with lower weights.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[2] Benjamin M. Bolstad. *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization.* PhD thesis, University of California, Berkeley, 2004.

[3] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15–e15, 2003.

[4] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[5] Wentian Li. Volcano Plots in Analyzing Differential Expressions with mRNA Microarrays. *Journal of bioinformatics and computational biology*, 10(06):1231003:1–24, 2012.

[6] Gordon K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25, 2004.

Appendices

Figure 9: Pseudo-images of the weights for chips 1-18. Colors: green for low weights and light gray for high weights.

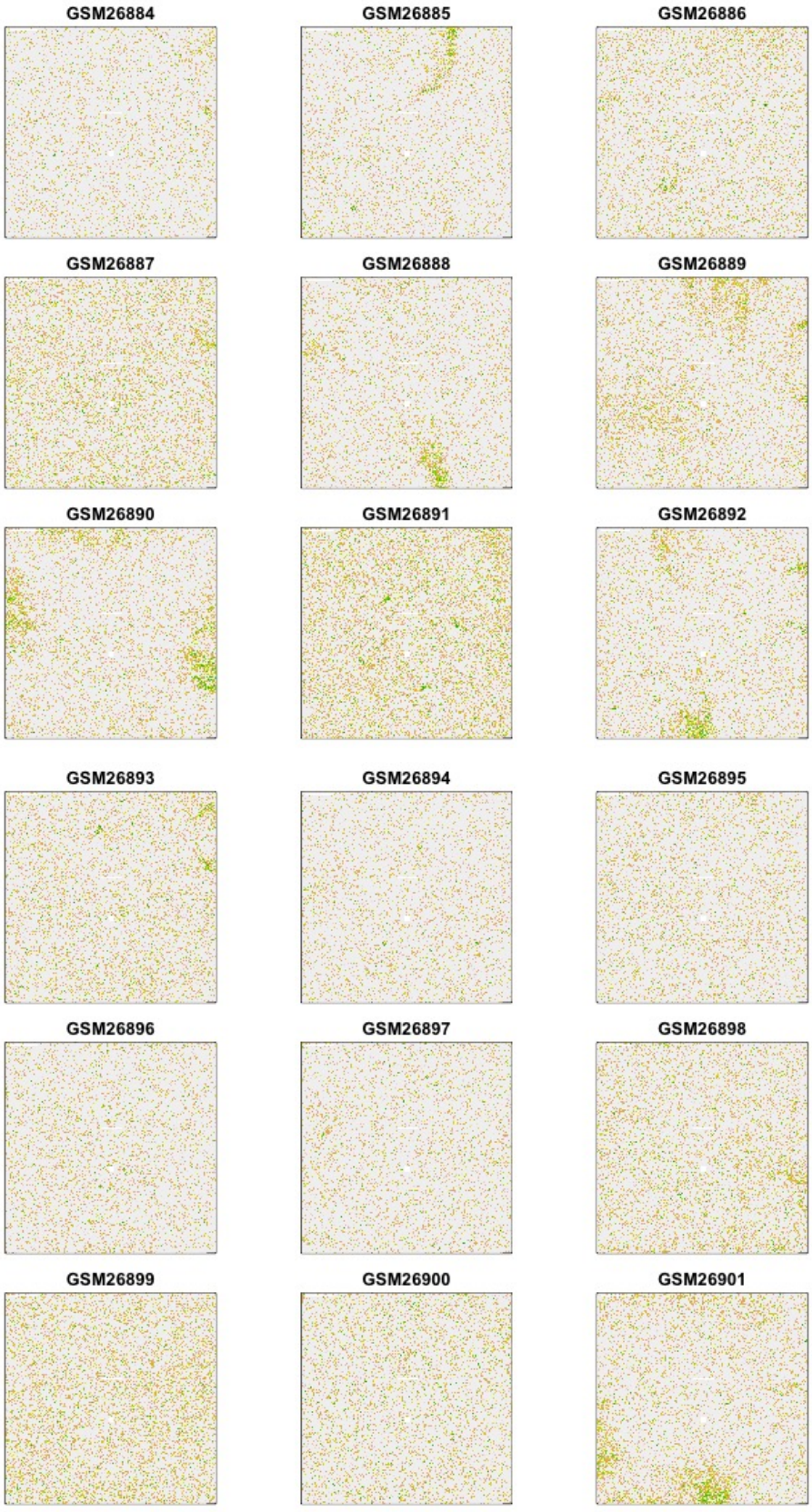


Figure 10: Pseudo-images of the weights for chips 19-36. Colors: green for low weights and light gray for high weights.

Figure 11: Pseudo-images of the weights for chips 37-49. Colors: green for low weights and light gray for high weights.

	ER+	ER- vs ER+
GSM26804	1	0
GSM26867	1	0
GSM26868	1	0
GSM26869	1	0
GSM26870	1	0
GSM26871	1	1
GSM26872	1	0
GSM26873	1	0
GSM26874	1	0
GSM26875	1	0
GSM26876	1	0
GSM26877	1	0
GSM26878	1	1
GSM26879	1	0
GSM26880	1	1
GSM26881	1	0
GSM26882	1	1
GSM26883	1	1
GSM26884	1	1
GSM26885	1	0
GSM26886	1	1
GSM26887	1	1
GSM26888	1	1
GSM26889	1	1
GSM26890	1	0
GSM26891	1	0
GSM26892	1	1
GSM26893	1	1
GSM26894	1	0
GSM26895	1	1
GSM26896	1	0
GSM26897	1	0
GSM26898	1	1
GSM26899	1	0
GSM26900	1	1
GSM26901	1	0
GSM26902	1	1
GSM26903	1	1
GSM26904	1	0
GSM26905	1	1
GSM26906	1	1
GSM26907	1	0
GSM26908	1	1
GSM26909	1	0
GSM26910	1	1
GSM26911	1	0
GSM26912	1	1
GSM26913	1	0
GSM26914	1	0

Table 1: Design matrix for ER- against ER+ status in GSE1561 dataset.

Gene's name	$\hat{\beta}_{2,k}$	moderated t-statistic	p-value	BH p-value
205225_at	-4.43	-20.68	8.07e-26	1.80e-21
209603_at	-3.15	-15.18	3.90e-20	4.34e-16
209604_s_at	-2.88	-14.17	6.15e-19	4.57e-15
209602_s_at	-3.34	-12.74	3.73e-17	2.08e-13
204623_at	-3.89	-11.66	1.00e-15	4.46e-12
205009_at	-4.13	-10.80	1.50e-14	5.56e-11
202088_at	-1.29	-10.16	1.23e-13	3.91e-10
212956_at	-2.12	-10.10	1.49e-13	4.14e-10
214431_at	0.90	9.37	1.70e-12	4.22e-09
209443_at	-2.20	-9.06	4.84e-12	1.08e-08
204508_s_at	-1.80	-8.78	1.30e-11	2.41e-08
202089_s_at	-2.07	-8.78	1.30e-11	2.41e-08
203574_at	1.51	8.72	1.59e-11	2.72e-08
205862_at	-2.61	-8.69	1.76e-11	2.80e-08
205186_at	-1.19	-8.61	2.31e-11	3.44e-08
214164_x_at	-2.12	-8.57	2.66e-11	3.70e-08
203963_at	-2.21	-8.54	2.95e-11	3.87e-08
215867_x_at	-2.08	-8.50	3.34e-11	4.13e-08
221823_at	-1.43	-8.43	4.31e-11	5.06e-08
35666_at	-0.80	-8.35	5.78e-11	6.44e-08
210652_s_at	-2.04	-8.23	8.54e-11	9.06e-08
204811_s_at	-0.92	-8.13	1.24e-10	1.22e-07
203685_at	-1.93	-8.12	1.26e-10	1.22e-07
201976_s_at	1.40	8.05	1.62e-10	1.48e-07
204862_s_at	-0.83	-8.04	1.67e-10	1.48e-07
205696_s_at	-1.80	-8.03	1.73e-10	1.48e-07
212770_at	-0.32	-8.02	1.83e-10	1.51e-07
212195_at	-1.41	-7.95	2.29e-10	1.82e-07
201413_at	-0.83	-7.92	2.55e-10	1.95e-07
209696_at	-1.50	-7.91	2.63e-10	1.95e-07
209460_at	-2.33	-7.83	3.54e-10	2.55e-07
213540_at	-0.55	-7.76	4.59e-10	3.12e-07
212441_at	-0.93	-7.74	4.82e-10	3.12e-07
203988_s_at	-1.28	-7.74	4.92e-10	3.12e-07
211712_s_at	-1.60	-7.73	5.03e-10	3.12e-07
218224_at	-0.94	-7.73	5.04e-10	3.12e-07
219806_s_at	1.02	7.71	5.35e-10	3.22e-07
210731_s_at	-0.48	-7.68	5.98e-10	3.49e-07
202390_s_at	-0.48	-7.67	6.11e-10	3.49e-07
218424_s_at	0.77	7.66	6.48e-10	3.61e-07
219833_s_at	-0.86	-7.65	6.69e-10	3.63e-07
219197_s_at	-2.41	-7.61	7.78e-10	4.05e-07
218204_s_at	-0.87	-7.61	7.81e-10	4.05e-07
204542_at	-1.56	-7.58	8.56e-10	4.34e-07
212207_at	-0.53	-7.56	9.11e-10	4.51e-07
217838_s_at	-1.50	-7.53	1.02e-09	4.93e-07
203929_s_at	-1.71	-7.50	1.15e-09	5.37e-07
212199_at	-0.74	-7.49	1.17e-09	5.37e-07
218259_at	-0.74	-7.49	1.18e-09	5.37e-07
208627_s_at	0.77	7.46	1.28e-09	5.72e-07

Table 2: Top of fifty genes differentially expressed.

Gene's name	Gene's name
206378_at	219795_at
217528_at	202269_x_at
214243_s_at	220281_at
205916_at	209173_at
209125_at	213953_at
206165_s_at	205680_at
209242_at	205041_s_at
205030_at	203058_s_at
219612_s_at	209116_x_at
217562_at	210397_at
205029_s_at	204607_at
206166_s_at	205064_at
219962_at	213711_at
206509_at	211122_s_at
214079_at	205157_s_at
205044_at	208792_s_at
201525_at	216623_x_at
209942_x_at	206561_s_at
213680_at	218888_s_at
211682_x_at	204304_s_at
214451_at	201843_s_at
214612_x_at	215108_x_at
215729_s_at	205306_x_at
204268_at	214974_x_at
203535_at	214580_x_at
202917_s_at	201820_at
209480_at	208791_at
213831_at	209800_at
204580_at	212236_x_at
205498_at	204734_at
203820_s_at	204988_at
203757_s_at	209505_at
217284_x_at	202859_x_at
209351_at	204712_at
222257_s_at	204855_at
217276_x_at	204942_s_at
214774_x_at	203824_at
203290_at	212730_at
204259_at	214461_at
203691_at	207802_at
206457_s_at	211138_s_at
220625_s_at	219327_s_at
201563_at	203819_s_at
203060_s_at	41469_at
211657_at	220414_at
205307_s_at	210163_at
206164_at	214624_at
205830_at	201349_at
207981_s_at	209283_at
214414_x_at	222043_at

Table 3: List of 100 genes with largest variance.