Genomics Data 3 (2015) 75–79

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

miRNA expression profiling of formalin-fixed paraffin-embedded (FFPE) hereditary breast tumors

Miljana Tanić [a,1], Kira Yanowski [a], Eduardo Andrés [b], Gonzalo Gómez-López [b], María Rodríguez-Pinilla Socorro [c], David G. Pisano [b], Beatriz Martinez-Delgado [a,2], Javier Benítez [a,d,*]

[a] Human Genetics Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[b] Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[c] Pathology Department, Fundación Jimenez Díaz, Madrid, Spain
[d] Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Madrid, Spain

ARTICLE INFO

Article history:
Received 30 October 2014
Received in revised form 13 November 2014
Accepted 17 November 2014
Available online 22 November 2014

Keywords:
Hereditary breast cancer
Microarray
miRNA

ABSTRACT

Hereditary breast cancer constitutes only 5–10% of all breast cancer cases and is characterized by strong family history of breast and/or other associated cancer types. Only ~25% of hereditary breast cancer cases carry a mutation in BRCA1 or BRCA2 gene, while mutations in other rare high and moderate-risk genes and common low penetrance variants may account for additional 20% of the cases. Thus the majority of cases are still unaccounted for and designated as BRCAX tumors. MicroRNAs are small non-coding RNAs that play important roles as regulators of gene expression and are deregulated in cancer. To characterize hereditary breast tumors based on their miRNA expression profiles we performed global microarray miRNA expression profiling on a retrospective cohort of 80 FFPE breast tissues, including 66 hereditary breast tumors (13 BRCA1, 10 BRCA2 and 43 BRCAX), 10 sporadic breast carcinomas and 4 normal breast tissues, using Exiqon miRCURY LNA™ microRNA Array v.11.0. Here we describe in detail the miRNA microarray expression data and tumor samples used for the study of BRCAX tumor heterogeneity (Tanic et al., 2013) and biomarkers associated with positive BRCA1/2 mutation status (Tanic et al., 2014). Additionally, we provide the R code for data preprocessing and quality control.

Specifications	
Organism/tissue	*Homo sapiens*/breast tissue
Array type	Exiqon miRCURY LNA™ microRNA Array v.11.0
Data format	Raw data: .txt files; normalized data: SOFT, MINIML, TXT; Code: .RData file
Experimental factors	Hereditary breast tumors, sporadic breast tumors, normal breast tissue
Experimental features	Global miRNA expression profiling of formalin-fixed paraffin – embedded tissue (FFPE) breast tissues
Consent	All patients gave their written informed consent for use of exceeding pathological material in research
Sample source location	Samples were collected from Spanish hospitals

* Corresponding author at: Human Genetics Group, Spanish National Cancer Research Center (CNIO), C\ Melchor Fernández Almagro, 3, E-28029 Madrid, Spain. Tel.: +34 912 246 900; fax: +34 912 246 911.
E-mail address: jbenitez@cnio.es (J. Benítez).
[1] Present address: Laboratory for Molecular Genetics, Experimental Oncology Department, Institute for Oncology and Radiology of Serbia (IORS), Belgrade, Serbia.
[2] Present address: Molecular Genetics Unit. Human Genetics Section. Instituto de Investigación en Enfermedades Raras IIER, Instituto de Salud Carlos III (ISCIII), Madrid, Spain.

Direct link to deposited data

The data is deposited in the GEO database under the accession number GSE44899: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44899.

Experimental design, materials and methods

Study population

Breast tumor samples were ascertained from patients diagnosed with breast cancer between 1988 and 2011 through Spanish hospitals: Gregorio Marañon, Hospital San City, Pablo, Fundacion Jimenez Dıaz, Hospital La Paz, H Santa Caterina (Girona), P. de Hierro, H. Severo Ochoa and H. Ramon y Cajal. Sporadic breast tumor samples were collected from breast cancer patients without suggestive family history of breast or ovarian cancer. Normal breast tissue samples were acquired from patients undergoing cosmetic breast reduction surgery with no personal or family history of breast cancer. All patients signed informed consent for genetic testing and use of exceeding material in research,

http://dx.doi.org/10.1016/j.gdata.2014.11.008

and the research project has the approval of the ethics committee of the Spanish National Cancer Research Centre (CNIO), named Comité de ética de la investigación y de bienestar animal del Instituto de Salud Carlos III.

Patients belonging to high-risk breast cancer families were selected according to the Spanish Medical Oncology Society (http://www.seom.org/) inclusion criteria for hereditary breast cancer: breast cancer diagnosed in <40-year-old female; both breast and ovarian cancer diagnosed in same patient; families with at least three female first-degree relatives affected with breast cancer; at least two females affected with breast cancer (at least one of them diagnosed before 50) or at least one case of female breast cancer and at least one case of ovarian, female bilateral breast or male breast cancer. All patients belonging to high-risk breast cancer families have undergone full BRCA1/2 gene testing for mutations (dHPLC and Sanger sequencing) and large rearrangements (MLPA). Patients fulfilling the criteria for hereditary breast cancer with no identifiable mutations in either BRCA1 or BRCA2 genes were designated as BRCAX cases. Hereditary breast tumor characteristics are shown in Table 1.

Sample processing

For each sample tumoral area was marked by pathologist, FFPE blocks were cut in 3–5 10 μm-sections, mounted on slides and tumor tissue was scraped into 1.5 ml tubes by needle macrodissection for subsequent RNA extraction. Total RNA was extracted using miRNeasy FFPE kit (QIAGEN) according to the manufacturer's instructions, with modified tissue digestion step using 20 mg/μl Proteinase K (Roche, Basel, Switzerland) for overnight incubation at 55 °C. RNA quantity was assessed using NanoDrop Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

Table 1
Clinico-pathological data for hereditary breast tumors.

Total no.	BRCA1 n = 13		BRCA2 n = 10		BRCAX n = 43	
	n	(%)	n	(%)	n	(%)
Age at diagnosis	**9**		**6**		**40**	
Mean	40.33		42.5		47.75	
Range	28–55		35–56		25–95	
Grade	**12**		**10**		**43**	
1	0	(0%)	2	(20%)	4	(9.3%)
2	1	(8.3%)	3	(30%)	20	(48.8%)
3	11	(91.7%)	5	(50%)	17	(41.5%)
Estrogen receptor	**13**		**10**		**37**	
Positive	2	(15.4%)	7	(70%)	18	(48.6%)
Negative	11	(84.6%)	3	(30%)	19	(51.4%)
Progesteron receptor	**12**		**10**		**38**	
Positive	2	(15.4%)	7	(70%)	14	(36.8%)
Negative	11	(84.6%)	3	(30%)	24	(63.2%)
HER2	**13**		**10**		**38**	
Positive	0	(0%)	2	(20%)	9	(23.7%)
Negative	13	(100%)	8	(80%)	29	(76.3%)
Ki-67	**13**		**7**		**31**	
1 (0–5%)	4	(30.8%)	3	(42.9%)	15	(48.3%)
2 (6–25%)	5	(23.1%)	2	(28.6%)	11	(35.5%)
3 (>25%)	4	(30.8%)	2	(28.6%)	5	(21.6%)
Subtype	**13**		**10**		**35**	
Luminal A	2	(15.4%)	6	(60%)	13	(37.1%)
Luminal B	0	(0%)	2	(20%)	6	(17.1%)
HER2	0	(0%)	0	(0%)	3	(8.6%)
Triple negative	11	(84.6%)	2	(20%)	13	(37.1%)
Lymph node	**13**		**10**		**37**	
Positive	5	(50%)	4	(57.1%)	18	(48.6%)
Negative	5	(50%)	3	(42.9%)	17	(51.4%)

Breast cancer cases were classified into four subtypes based on IHC-model [15]. In bold is the number of samples per category (BRCA1, BRCA2 and BRCAX) for which there was available information on clinico-pathological feature in question.

Experimental design and array description

Microarray expression profiling of microRNAs was performed on a retrospective cohort of 80 FFPE breast tissues, including 66 hereditary breast tumors (13 BRCA1, 10 BRCA2 and 43 BRCAX), 10 sporadic breast carcinomas and 4 normal breast tissues, using miRCURY LNA™ microRNA Array v.11.0 — hsa, mmu & rno (Exiqon A/S, Vedbaek, Denmark), in a single-color experiment. Experimental design included pairwise comparisons between BRCA1 tumors, BRCA2 tumors, BRCAX tumors, sporadic breast tumors and normal breast tissue; pairwise comparisons of BRCA-mutation carriers (BRCA1 and BRCA2 tumors), non-carriers (BRCAX and sporadic tumors), and normal breast tissue; and finally, pairwise comparisons of hereditary breast tumors (BRCA1, BRCA2 and BRCAX tumors), sporadic breast tumors and normal breast tissue.

The miRCURY LNA™ microRNA Array v.11.0 — hsa, mmu & rno contains capture probes for over 1700 microRNAs in human, mouse, rat and their related viruses as annotated in miRBase Release v.11.0 including 1940 capture probes, in 4 replicates, representing 829 human miRNAs annotated in miRBasev.11 database and 434 hsa-miRPlus™ probes (Exiqon proprietary). Forty three control capture probes were included in the probe set including 10 synthetic microRNAs spike-in control probes in 48 replicates to evaluate labeling and hybridization, and seven negative control capture probes and twenty six capture probes complementary to small nuclear RNAs in 4 replicates.

Microarray hybridization

Labeling and hybridization procedure was performed as recommended by the manufacturer, using miRCURY LNA™ microRNA Power Labeling Kit (Exiqon, Denmark). First, 300 ng of total RNA was treated with Calf Intestinal Alkaline Phosphatase (CIP) to remove the 5′-phosphates from the microRNA termini prior to labeling with Hy3 green fluorescent dye. A set of 10 synthetic spike-in RNAs (Spike-in miRNA kit, Exiqon) was added to the RNA sample before the labeling reaction and later used for quality control for RNA labeling reaction and inter-array reproducibility. Labeling reaction was performed using 2 μl of CIP treated total RNA, 1.5 μl of Hy3 fluorescent dye, 2 μl DMSO and 2 μl of labeling enzyme, reaction was incubated at 16 °C for 1 h and heat inactivated by incubation at 65 °C for 15 min and left at 4 °C until hybridization step. Labeled samples were subsequently loaded onto a miRNA microarray slide and hybridized over 16 h at 56 °C. Washing of the slides was performed according to the manufacturer's instruction. Washed slides were dried by centrifuging at 1200 rpm for 5 min. Processed slides were scanned with Agilent G2565AA Microarray Scanner System (Agilent Technologies, Santa Clara, CA, USA), with the laser set to 635 nm, at Power 80 and PMT 70 setting, and a scan resolution of 10 μm. To avoid ozone bleaching, microarrays were scanned in an ozone-free environment (less than 2 ppb ozone). Fluorescence intensities on scanned images were quantified using Agilent Feature Extraction software version 9.5.3 (Agilent Technologies) using the modified Exiqon protocol and corresponding GAL files.

Quality control

Reliability of each microarray was assessed using FE Quality Control report data. No values were flagged indicating that salt/hair/dust and other anomalies did not affect the results. Additionally, expression values and distribution of spike-in RNAs were inspected for each of the arrays to perform quality control of labeling and hybridization (Supporting Fig. 1), to estimate the variance of replicated measurements within arrays and to assess the technical variability between different parts of the array. Spike-in CV values calculated between the different slides in the experiment did not exceed 12%, with median CV

of 1.9% between arrays, and the Pearson correlation of >0.974 indicating very high inter-array reproducibility. Further between array quality control was performed using *ArrayDataMetrics* R/Bioconductor package [6]. Fig. 1 shows boxplots of the array intensities before and after quantile normalization. Outlier detection was performed by computing the Kolmogorov–Smirnov statistic Ka between each array's distribution and the distribution of the pooled data. Two arrays were marked as outliers by an asterisk (*) for having the distribution of an array different from the others.

Patterns in the plot Fig. 2 indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). There was no indication of batch effect given that there was no correlation of the date of hybridization (array order) and clustering pattern. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b\, d_{ab}$ was exceptionally large. Three such arrays were detected, and they are marked by an asterisk, *. (See Fig. 2)

Finally, we inspected the MA plots of each array compared to median-intensity "pseudo-array" (See Fig. 3). Outlier detection was performed by computing Hoeffding's statistic D_a on the joint distribution of A and M for each array. Shown are first the 4 arrays with the highest values of D_a, then the 4 arrays with the lowest values. There were no outlier arrays with $D_a > 0.15$. Given that no single array was highlighted as an outlier by more than one method, and that there was no evidence of batch effect we used all arrays for downstream data analysis.

Data preprocessing and probe annotation

Raw data values were background subtracted using Normexp background correction method and normalized applying quantile method in *limma* R/Bioconductor package [4,5]. Normalized intensity data were log2 transformed and subjected to further analysis. Data preprocessing was performed using GEPAS 4.0 [13], however since January 2013 GEPAS has migrated to Babelomics 4.0 (http://babelomics.bioinfo.cipf.es/) [8]. Gene patterns containing more than 70% missing values were discarded, while other missing values were imputed using 3-k nearest neighbors. Prior to hierarchical clustering miRNA data was preprocessed to eliminate miRNAs with uniformly low expression and/or with low expression variation (var < 0.1) across the experiments, retaining 444 features (276 hsa-miR, 168 hsa-miRPlus). In the original publications probes were annotated according to miRBase v.11. Due to subsequent changes in miRNA nomenclature here we provide the up-dated probe annotation file for miRBase v.21 (Supporting file 1).

Data analysis

Differential expression analysis was performed with linear models (limma) moderated t-test implemented in the POMELO II tool, available in Asterias package (http://asterias.bioinfo.cnio.es) [9]. The estimated significance level (unadjusted p-values) was corrected for multiple hypotheses testing using Benjamini and Hochberg

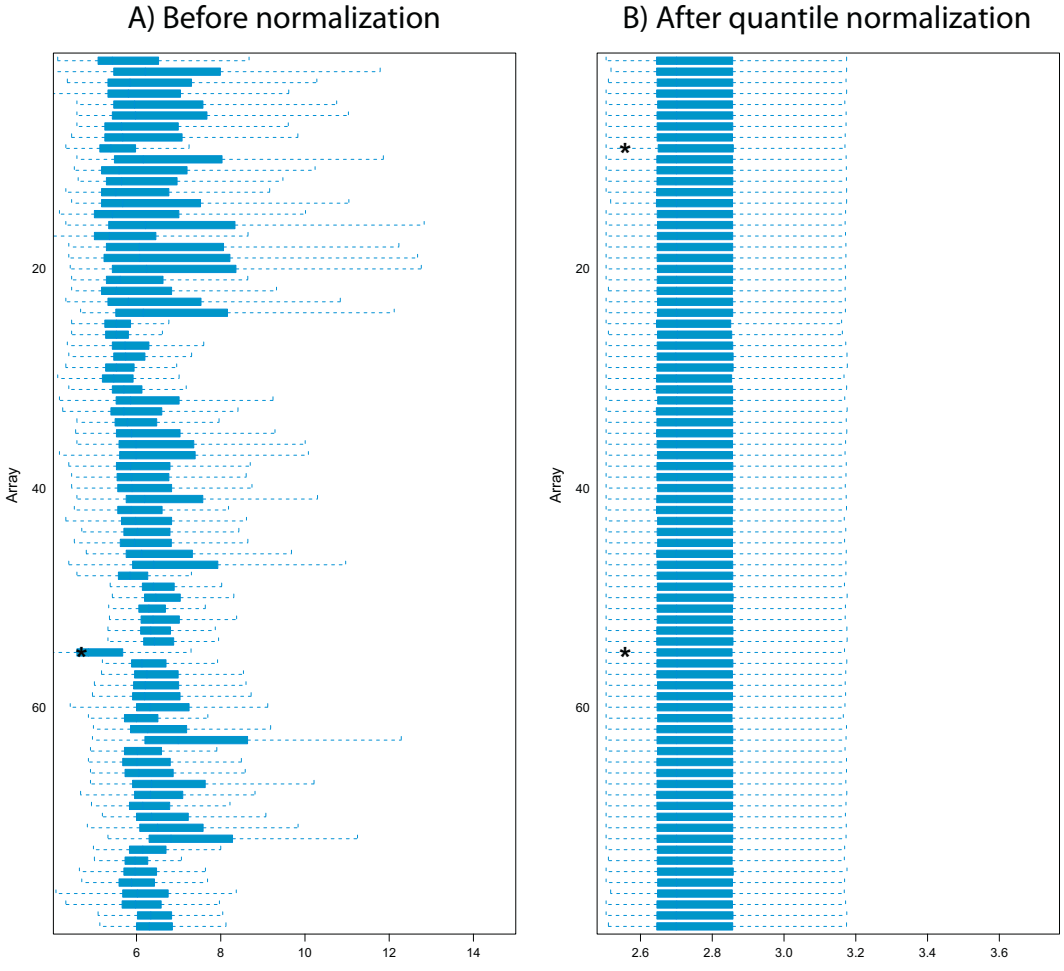

Fig. 1. Shows boxplots representing summaries of the signal intensity distributions of the arrays A) before normalization, B) after quantile normalization. Each box corresponds to one array.

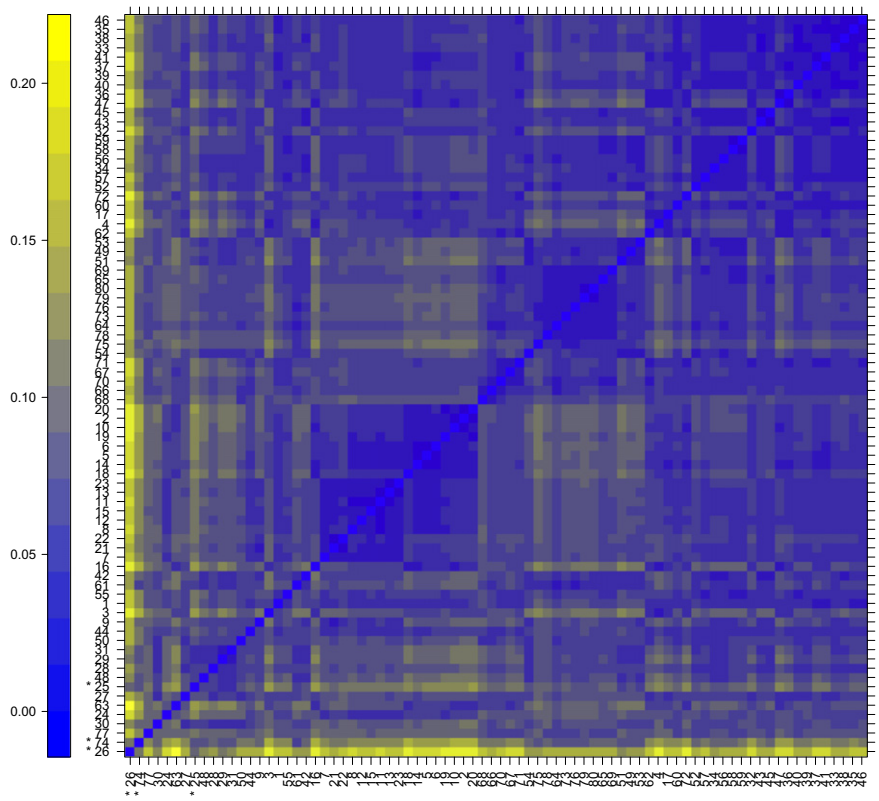

Fig. 2. Shows a pseudocolor heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. The distance d_{ab} between two arrays *a* and *b* is computed as the mean absolute difference (L_1-distance) between the data of the arrays (using the data from all probes without filtering). Identified outliers are marked with an asterisk (*).

method to control the False Discovery Rate (FDR) [1]. Those miRNAs with q-value <0.05 were selected as significantly differentially expressed.

Average linkage hierarchical cluster analysis using Pearson correlation with uncentered metrics was performed using Gene Cluster and data were visualized by Treeview v.1.6 [3]. Consensus Clustering module available in the Gene Pattern suite [10] was used for the discovery of biologically meaningful clusters among the BRCAX tumors [11] by assessing the stability of the discovered clusters by applying a KNN means resampling algorithm, with 2, 3, 4, and 5 centroids using 500 re-sampling iterations. The consensus among the multiple runs was assessed and summarized in a consensus matrix and ΔG plot to estimate the composition and number of clusters, and the change in free energy with every additional group added.

For the building of predictive miRNA classifier described in Tanic et al. [12], we used the Prophet [7] tool implemented in GEPAS 4.0 and a split sample approach. Genes discriminating between BRCA1/2 mutated tumors in the training set were ranked by their F-ratio based on the between to within sum of squares, and predictors were built using the best 2, 5, 10, 20, 35, 50 and 100 genes and several methods for classification [2,14]: support vector machines (SVM), k-nearest neighbor (KNN), diagonal linear discriminant analysis (DLDA), self-organizing maps (SOM) and shrunken centroids (PAMR). Classifier performance was evaluated by the leave-one-out cross-validation procedure. The final miRNA- classifier yielding the minimum misclassification error after cross-validation in the training set and optimal sensitivity and specificity was selected. The performance of the selected classifier was validated by applying the specified model to the corresponding test set which has never been used for the training of the classifier or for feature selection, and estimating sensitivity, specificity, positive and negative predictive values based on the confusion matrix.

Conclusion

In summary, here we have described, to our knowledge, the largest dataset on miRNA expression profiling in hereditary breast tumors used recently in studies published in specialty journals. We have shown that the data are of high quality and described in detail the microRNA data set including an updated miRNA nomenclature to match miRBase v21 to enable future studies.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.11.008.

Disclosures

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Asociación Española Contra el Cancer (AECC) and the Spanish Fondo de Investigaciones Sanitarias (grant numbers FIS-PI081298 and FIS-PI081120), by grants from INNPRONTA from the Ministry of Science and Innovation PI11/01059, from the Fundación Mutua Madrileña 2011 (PI BMD), and from the Fundación Sandra Ibarra 2011 (PI BMD). The CIBER de Enfermedades Raras is an initiative of the Instituto de Salud Carlos III (ISCIII). MT had financial support from the Fundación La Caixa call 2008.

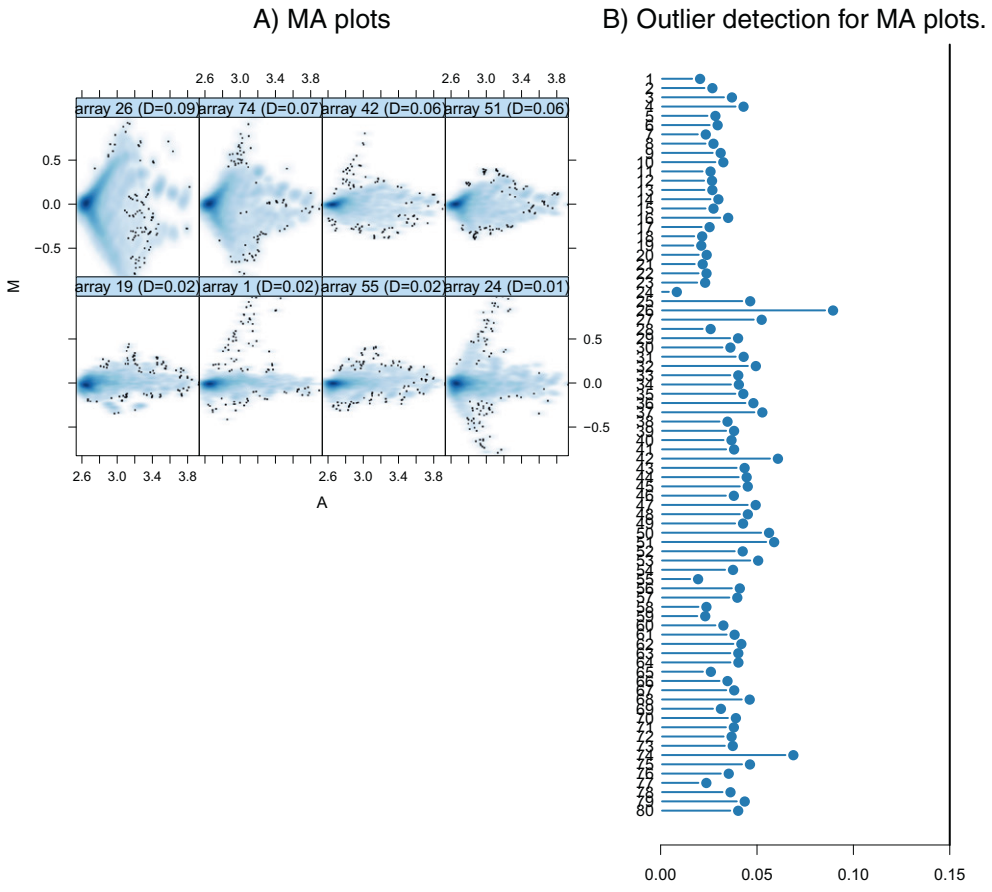

Fig. 3. A) shows representative MA plots for arrays with lowest (top 4) and highest (bottom 4) Hoeffing's D-statistic. M and A are defined as: $M = \log_2(I_1) - \log_2(I_2)$, $A = 1/2\ (\log_2(I_1) + \log_2(I_2))$, where I_1 is the intensity of the array studied, and I_2 is the intensity of a "pseudo"-array that consists of the median across arrays. The value of Da is shown in the panel headings. B) Shows a bar chart of the Da, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. A threshold of 0.15 was used, which is indicated by the vertical line. None of the arrays exceeded the threshold and was considered an outlier.

References

[1] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, I. Golani, Controlling the false discovery rate in behavior genetics research. Behav. Brain Res. 125 (1–2) (2001) 279–284.

[2] S. Dudoit, J. Fridlyand, A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. 3 (7) (2002) (RESEARCH0036).

[3] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. 95 (25) (1998) 14863–14868.

[4] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, J. Zhang, Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5 (10) (2004) R80.

[5] G.K. Smyth, Limma: linear models for microarray data. in: R. Gentleman, CV, S. Dudoit, R. Irizarry, W. Huber (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, New York, 2005, pp. 397–420.

[6] A. Kauffmann, R. Gentleman, W. Huber, arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics 25 (3) (2009) 415–416.

[7] I. Medina, D. Montaner, J. Tarraga, J. Dopazo, Prophet, a web-based tool for class prediction using microarray data. Bioinformatics 23 (3) (2007) 390–391.

[8] I. Medina, J. Carbonell, L. Pulido, S.C. Madeira, S. Goetz, A. Conesa, J. Tarraga, A. Pascual-Montano, R. Nogales-Cadenas, J. Santoyo, F. Garcia, M. Marba, D. Montaner, J. Dopazo, Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 38 (Web Server issue) (2010) W210–W213.

[9] E.R. Morrissey, R. Diaz-Uriarte, Pomelo II: finding differentially expressed genes. Nucleic Acids Res. 37 (Web Server issue) (2009) W581–W586.

[10] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, J.P. Mesirov, GenePattern 2.0. Nat. Genet. 38 (5) (2006) 500–501.

[11] M. Tanic, E. Andres, S.M. Rodriguez-Pinilla, I. Marquez-Rodas, M. Cebollero-Presmanes, V. Fernandez, A. Osorio, J. Benitez, B. Martinez-Delgado, MicroRNA-based molecular classification of non-BRCA1/2 hereditary breast tumours. Br. J. Cancer 109 (10) (2013) 2724–2734.

[12] M. Tanic, K. Yanowski, G. Gomez-Lopez, M. Socorro Rodriguez-Pinilla, I. Marquez-Rodas, A. Osorio, D.G. Pisano, B. Martinez-Delgado, J. Benitez, MicroRNA expression signatures for the prediction of BRCA1/2 mutation-associated hereditary breast cancer in paraffin-embedded formalin-fixed breast tumors. Int. J. Cancer 136 (3) (2015) 593–602.

[13] J.M. Vaquerizas, L. Conde, P. Yankilevich, A. Cabezon, P. Minguez, R. Diaz-Uriarte, F. Al-Shahrour, J. Herrero, J. Dopazo, GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Res. 33 (Web Server issue) (2005) W616–W620.

[14] L.F. Wessels, M.J. Reinders, A.A. Hart, C.J. Veenman, H. Dai, Y.D. He, L.J. van't Veer, A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 21 (19) (2005) 3755–3762.

[15] P. Tang, K.A. Skinner, D.G. Hicks, Molecular classification of breast carcinomas by immunohistochemical analysis: are we ready? Diagn. Mol. Pathol. 18 (3) (2009) 125–132.