Genomics Data 1 (2013) 7–10

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Genome-wide gene expression profiling to predict resistance to anthracyclines in breast cancer patients

B. Haibe-Kains [a,c,m], C. Desmedt [a], A. Di Leo [b], E. Azambuja [a], D. Larsimont [a], J. Selleslags [a], S. Delaloge [d], C. Duhem [e], J.P. Kains [f], B. Carly [g], M. Maerevoet [h], A. Vindevoghel [i], G. Rouas [a], F. Lallemand [a], V. Durbecq [a], F. Cardoso [a], R. Salgado [a], R. Rovere [a], G. Bontempi [c], S. Michiels [a], M. Buyse [j], J.M. Nogaret [a], Y. Qi [a], F. Symmans [k], L. Pusztai [l], V. D'Hondt [a], M. Piccart-Gebhart [a,1], C. Sotiriou [a,*,1]

[a] Institut Jules Bordet, Brussels, Belgium
[b] Hospital of Prato, Prato, Italy
[c] Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium
[d] Institut Gustave Roussy, Villejuif, France
[e] Centre Hospitalier du Luxembourg, Luxembourg
[f] HIS—Site Etterbeek-Ixelles, Brussels, Belgium
[g] Hopital Saint-Pierre, Brussels, Belgium
[h] Clinique Saint-Pierre, Ottignies, Belgium
[i] Clinique Ste Elisabeth, Namur, Belgium
[j] IDDI, Louvain-La-Neuve, Belgium
[k] The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
[l] Yale Cancer Center, Yale University, New Haven, CT, USA
[m] Ontario Cancer Institute, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada

ARTICLE INFO

Article history:
Received 12 September 2013
Accepted 12 September 2013
Available online 1 October 2013

Keywords:
Breast cancer
Gene expression
Microarray
Anthracycline
Bioinformatics

ABSTRACT

Validated biomarkers predictive of response/resistance to anthracyclines in breast cancer are currently lacking. The neoadjuvant Trial of Principle (TOP) study, in which patients with estrogen receptor (ER)–negative tumors were treated with anthracycline (epirubicin) monotherapy, was specifically designed to evaluate the predictive value of topoisomerase II-alpha (TOP2A) and develop a gene expression signature to identify those patients who do not benefit from anthracyclines. Here we describe in details the contents and quality controls for the gene expression and clinical data associated with the study published by Desmedt and colleagues in the Journal of Clinical Oncology in 2011 (Desmedt et al., 2011). We also provide R code to easily access the data and perform the quality controls and basic analyses relevant to this dataset.

Specifications	
Organism/cell line/tissue	*Homo sapiens*
Strain(s)	Patients' breast tumors
Sequencer or array type	Affymetrix GeneChip HG-U133PLUS2
Data format	Raw data: CEL files, normalized data: SOFT, MINIML, TXT and RData
Experimental factors	Pathological complete response; age; tumor size; histological grade; axillary lymph node status; HER2, TOP2A status by FISH, distant metastasis free and overall survival
Consent	All patients gave their written informed consent before study entry.

Direct link to deposited data

Deposited data can be found here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16446.

* Corresponding author at: Institut Jules Bordet, 121 Boulevard de Waterloo, 1000 Bruxelles, Belgium. Tel.: +32 2 541 34 28; fax: +32 2 538 08 58.
E-mail address: christos.sotiriou@bordet.be (C. Sotiriou).
[1] These authors contributed equally to the present manuscript.

2213-5960
http://dx.doi.org/10.1016/j.gdata.2013.09.001

Experimental design, materials and methods

Study population and clinical data

The neoadjuvant prospective TOP trial (NCT00162812) was conducted at different European hospitals and coordinated by the Institut Jules Bordet in Brussels, Belgium. One hundred and forty-nine patients were included in this trial.

The inclusion criteria were the following: 1/histologically-confirmed breast cancer (either operable, locally advanced or inflammatory), 2/age ≤ 70 yrs, 3/female patient, 4/tumor size ≥ 2 cm at the ultrasound examination, 5/estrogen receptor (ER)-negative tumors, 6/multifocal and multicentric breast tumors are allowed if all foci are ER-negative (Table 1.). It is reasonable to limit multifoci tumors to bifocal ones since a fixed and frozen samples should be obtained from each focus, 7/fixed and frozen samples from the primary tumor, obtained before treatment with epirubicin, must be available for evaluation of biological markers, 8/written informed consent before study registration, 9/performance status ≤ 1 (ECOG scale), 9/ANC ≥ 1500/mm^3, platelets ≥ 100,000/mm^3, Hb ≥ 10 g/dl, total bilirubin and serum creatinine ≤ 1 N, GOT/GPT ≤ 1.5 N, alkaline phosphatase ≤ 2.5 N, 10/normal left ventricular ejection fraction by echocardiography or muga scan, and 11/negative pregnancy test for all women of childbearing potential.

The exclusion criteria were the following: 1/metastatic breast cancer, 2/serious medical conditions like: a) congestive heart failure or unstable angina pectoris, previous history of myocardial infarction within 1 year from study entry, uncontrolled arrhythmias.; b) history of significant neurologic or psychiatric disorders, c) active uncontrolled infection, d) active peptic ulcer, unstable diabetes mellitus; 3/concomitant contra-lateral invasive breast cancer, 4/concurrent treatment with hormonal replacement therapy, 5/concurrent treatment with any other anti-cancer therapy, and 6/previous treatment with anthracyclines for breast cancer. One patient was not treated according to the protocol due to ineligibility (concomitant contra-lateral invasive breast cancer).

Table 1
Patient and tumor baseline characteristics for evaluable.

Patients (n = 139)		
Characteristic	No. of patients	%. of patients
Age, years		
≤50	86	61.9
>50	53	38.1
Tumor size (at baseline)		
T1	20	14.4
T2	99	71.2
T3	5	3.6
T4	15	10.8
Nodal status (at baseline)		
N0	69	49.6
N1	64	46.0
N2	3	2.2
N3	3	2.2
Histologic type		
Ductal	130	93.5
Lobular	1	0.7
Other	8	5.8
Histologic grade		
G1	2	1.4
G2	26	18.7
G3	104	74.8
Gx	7	5.0
HER2 status by FISH		
Not amplified	73	52.5
Amplified	33	23.7
Missing	33	23.7
Ki67, %		
≤25	23	16.5
>25	92	66.2
Missing	24	17.3
pCR[a]		
No	120	86.3
Yes	19	13.7

Abbreviations: HER2, human epidermal growth factor receptor 2; FISH, fluorescent in situ hybridization; pCR, pathologic complete response.

[a] The 14 patients who discontinued treatment because of lack of response (n = 11) or progression (n = 3) were considered to have residual disease for response prediction analysis.

At completion of chemotherapy, every patient underwent surgery with axillary node sampling. After surgery, adjuvant docetaxel, trastuzumab (in case of human epidermal growth factor receptor 2 (HER2)-positive tumors) and loco-regional irradiation were administered using standard criteria.

All patients underwent pretreatment core biopsies of the primary breast tumor before starting neoadjuvant chemotherapy using a 14G needle. Two biopsies were embedded in OCT (Sakura), frozen in liquid nitrogen within 5 min and transferred to a −80 °C freezer. Two biopsies were fixed in formalin and embedded in paraffin. Both fixed and frozen samples were retrieved by a specialized company and stored at the Institut Jules Bordet in Brussels (Belgium), where the HER2, topoisomerase 2A (TOP2A) and gene expression evaluations were carried out.

The pathological response assessment and the different TOP2A evaluations were carried out in a blinded fashion: pathological response determination, TOP2A gene, mRNA and protein analyses were conducted independently.

All pathology reports were centrally reviewed and 17/19 pCRs were centrally reviewed at the Jules Bordet Institute. Two pCR cases could not be centrally reviewed since the participating center did not send the slides to the Jules Bordet Institute.

The clinical data was collected, monitored and validated by the BrEAST Data Centre, Institut Jules Bordet. The anonymized clinical data were deposited in the Gene Expression Omnibus database (GEO; [2]) under accession number GSE16446.

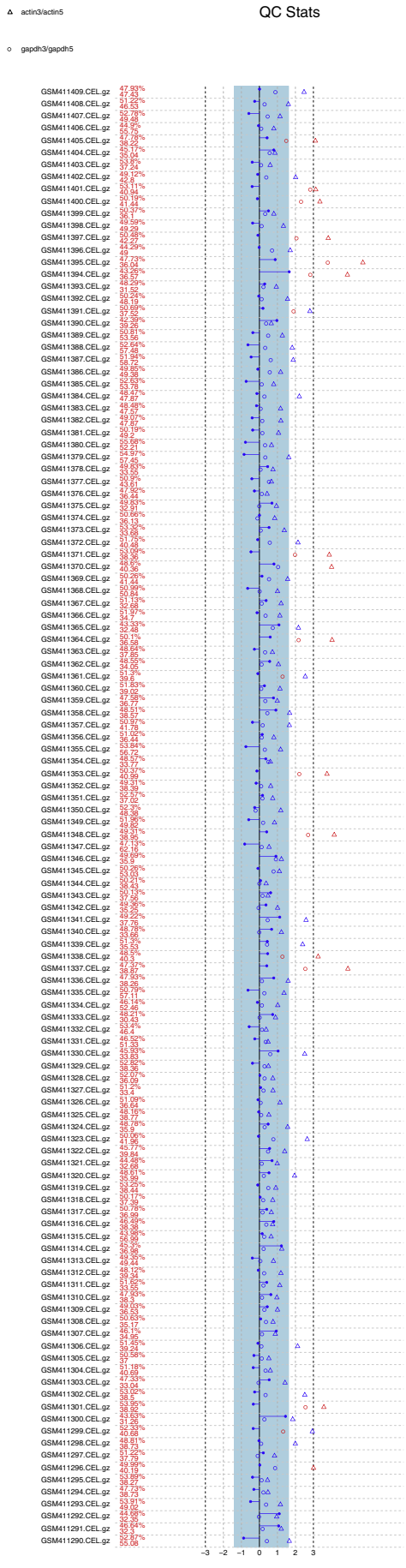
Gene expression data

One 5-μm tissue section (usually after ten 30-μm sections) of each biopsy was hematoxylin and eosin stained to monitor the tumor cell percentage of the tissue. Only specimens with more than 30% of tumor cells were included for further analysis. Isolation of RNA was performed using the Trizol method (Invitrogen) according to the manufacturer's instructions and purified using RNeasy mini-columns (Qiagen, Valencia, CA). The quality of the RNA obtained from each tumor sample was assessed based on the RNA profile generated by the bioanalyzer (Agilent Inc.). RNA amplification, hybridization and image scanning were done according to standard Affymetrix protocols.

We used the Affymetrix Human Genome U133-2.0 plus GeneChip according to standard Affymetrix protocols. Microarray and sample annotation data were deposited in GEO under accession number GSE16446.

Quality control

We used the simpleaffy Bioconductor package [10] to check the quality of each individual CEL file. As can be seen in Fig. 1, all the CEL files contained a sufficiently large percentage of present calls (>40%) and all the scale factors lie within a 3-fold range, which indicate good quality according to Affymatrix guidelines [1]. Eighteen CEL files yielded larger beta-actin and GAPDH 3′/5′ ratios than the threshold published in the Affymetrix guidelines (Fig. 1), which indicates that RNA degradation might be an issue for these samples. However there is no consensus regarding this quality metrics as to what is acceptable or should be rejected for further analysis, we therefore used all the samples in our study [3].



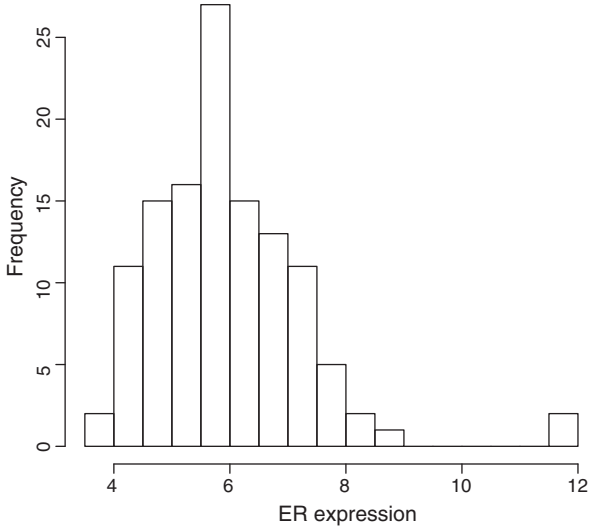


Fig. 2. Histogram of ER mRNA expression in the TOP dataset, which is composed of breast tumors identified as ER- by IHC. Two patients had high expression of ER mRNA and were therefore excluded from further analyses.

Normalization

CEL files were normalized using RMA [6] and probesets were annotated using the chip description file hgu133plus2cdf and biomaRt package [4] available from Bioconductor [5]. The sample annotation and the corresponding documentation are available as supplementary files of the series GSE16446 in GEO.

Basic analysis

As we collected patients with breast tumors determined as ER- by immunohistochemistry (IHC) we double-checked ER negativity with ER mRNA level. The most variant probeset for ER was considered (205225_at). We observed that two patients (GSM411367 and GSM411399) expressed high level of ER mRNA ($\geq$9, Fig. 2). As our study focused on ER- breast tumors these two patients were excluded from further analyses.

We then compared HER2 expression (most variant was considered: 210930_s_at) with its fluorescence in situ hybridization (FISH) status. As expected HER2 was highly expressed in HER2 FISH + tumors (two-sided Wilcoxon rank sum test p-value = 7E-12, Fig. 3).

We also paid particular attention to TOP2A as it is key in our study of resistance to anthracycline. TOP2A was represented by three different Affymetrix probesets: 201291_s_at, 201292_at and 237469_at. The first two probes have been tagged as high quality in the CleanEX database [9] while the probeset 237469_at exhibited low sequence specificity. Supporting this observation, the first two probesets were highly correlated (Pearson correlation coefficient > 0.83) while the correlation with the third one is low (0.56 and 0.52 respectively). Since the probeset 201291_s_at showed the greatest variance [variance of the 3 probesets:

Fig. 1. Quality controls for the Affymetrix Raw data generated in [3]. Cel file name for each experiment is provided on the left side, followed by the percentage of present and absent calls (in red) folowing the Affymetrix guidelines. The blue region in the middle of the plot represents the 3-fold region for scale factor as this region is considered as acceptable according to Affymetrix guidelines; any scale factor outside this region is drawn in red as it is considered an indicator of poor quality. Beta-actin and GAPDH 3′–5′ ratios are also represented on the right side by triangles and circles, respectively; ratio higher than 1.25 are drawn in red as they are considered indicators of poor quality.

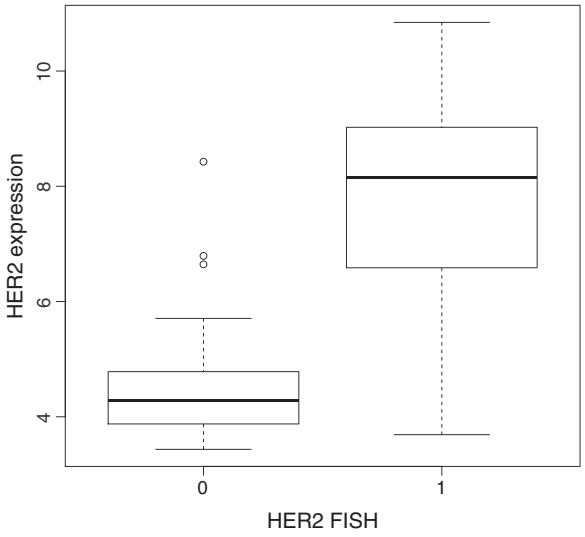


Fig. 3. Concordance between HER2 mRNA expression and HER2 FISH status. As expected HER2 mRNA expression was significantly higher in HER2 FISH positive tumors versus HER2 FISH negative tumors (P = 7E-12).

201291_s_at = 1.32, 201292_at = 0.84 and 237469_at = 0.22], we used it for further analyses.

Discussion

We described here a unique dataset of patients with ER- breast tumor treated with anthracycline. This dataset is composed of clinical data, including pathological response to anthracycline and genome-wide gene expression measured using Affymetrix GeneChip platform. We showed that the gene expression data are of high quality and are concordant with important clinical parameters. This dataset has been recently used in studies published in high impact journals [7,8], which demonstrate the importance and the benefit of data sharing for biomedical research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2013.09.001.

References

[1] Affymetrix, Quality Control Assessment in Genotyping Console. Affymetrix White Paper, 2008.
[2] T. Barrett, T.O. Suzek, D.B. Troup, S.E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A.E. Lash, W. Fujibuchi, R. Edgar, NCBI GEO: mining millions of expression profiles–database and tools. Nucleic Acids Res. 33 (2005) D562–D566.
[3] C. Desmedt, A. Di Leo, E. de Azambuja, D. Larsimont, B. Haibe-Kains, J. Selleslags, S. Delaloge, C. Duhem, J.-P. Kains, B. Carly, et al., Multifactorial approach to predicting resistance to anthracyclines. J. Clin. Oncol. 29 (2011) 1578–1586.
[4] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, W. Huber, BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21 (2005) 3439–3440.
[5] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al., Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5 (2004) R80.
[6] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4 (2003) 249–264.
[7] N. Juul, Z. Szallasi, A.C. Eklund, Q. Li, R.A. Burrell, M. Gerlinger, et al., Assessment of an RNA interference screen-derived mitotic and ceramide pathway metagene as a predictor of response to neoadjuvant paclitaxel for primary triple-negative breast cancer: a retrospective analysis of five clinical trials. Lancet Oncol. 11 (4) (2010) 358–365, http://dx.doi.org/10.1016/S1470-2045(10)70018-8.
[8] Y. Li, L. Zou, Q. Li, B. Haibe-Kains, R. Tian, Y. Li, C. Desmedt, C. Sotiriou, Z. Szallasi, J.D. Iglehart, et al., Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. Nat. Med. 16 (2010) 214–218.
[9] V. Praz, V. Jagannathan, P. Bucher, CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature. Nucleic Acids Res. 32 (2003) 542–547.
[10] C.L. Wilson, C.J. Miller, Simpleaffy: a BioConductor package for Affymetrix Quality Simpleaffy: a BioConductor package for Affymetrix Quality Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinform. Appl. Note 21 (2005) 3683–3685.