Genomics Data 2 (2014) 354–356

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Control of dataset bias in combined Affymetrix cohorts of triple negative breast cancer

Thomas Karn [a,*], Achim Rody [b], Volkmar Müller [c], Marcus Schmidt [d], Sven Becker [a], Uwe Holtrich [a], Lajos Pusztai [e]

[a] Department of Gynecology, Goethe-University Frankfurt, Frankfurt am Main, Germany
[b] Department of Obstetrics and Gynecology, University Hospital Lübeck, Germany
[c] Department of Gynecology, University Hospital Hamburg-Eppendorf, Hamburg, Germany
[d] Department of Obstetrics and Gynecology, Johannes Gutenberg-University Mainz, Mainz, Germany
[e] Yale Cancer Center, New Haven, CT, USA

ARTICLE INFO

Article history:
Received 18 September 2014
Accepted 29 September 2014
Available online 23 October 2014

Keywords:
Dataset bias
Breast cancer
Gene expression
Microarray
Pooling

ABSTRACT

Heterogenous subtypes of breast cancer need to be analyzed separately. Pooling of datasets can provide reasonable sample sizes but dataset bias is an important concern. We assembled a combined dataset of 579 Affymetrix microarrays from triple negative breast cancer (TNBC) in Gene Expression Omnibus (GEO) series GSE31519. We developed a method for selecting comparable datasets and to control for the amount of dataset bias of individual probesets.

© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Specifications	
Organism/cell line/tissue	*Homo sapiens*/breast tumor tissue
Sex	Female
Sequencer or array type	Affymetrix GeneChip HG-U133A and HG-U133PLUS2
Data format	Raw data: CEL files, normalized data: MAS5 Log2 magnitude-normalized
Experimental factors	Primary dataset origin of samples
Experimental features	Selection of comparable datasets and control for dataset bias of each probeset
Consent	Publicly available data from Gene Expression Omnibus (GEO) database
Sample source location	NA

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31519 (link to GEO Series)
http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE31519&format=file&file=GSE31519%5Fcomplete%5Fdataset%2Etxt%2Egz (direct link to normalized complete dataset in GEO supplement)
http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE31519&format=file&file=GSE31519%5FTNBC%5FSampleInfo%5FBCR%2Etxt%2Egz (direct link to sample information in GEO supplement).

Experimental design, materials and methods

Background

Breast cancer is a heterogeneous disease of different subtypes and separate analyses by subtype are mandatory. Triple negative breast cancer (TNBC) represents an aggressive disease and the use of currently available molecular prognostic signatures is limited. Reasonable sample sizes of TNBC for molecular analyses may be obtained by pooling several microarray datasets. However, because of significant inter-laboratory variation such studies require precise control of dataset bias.

Dataset

The set of 579 TNBCs in GSE31519 includes: (i) 67 CEL files in GSE31519 (GSM782523–GSM782589), (ii) 489 re-analyzed GEO samples linked in GSE31519, and (iii) 23 re-analyzed ArrayExpress samples.

* Corresponding author at: Department of Obstetrics and Gynecology, Goethe University Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt, Germany.
E-mail address: t.karn@em.uni-frankfurt.de (T. Karn).

http://dx.doi.org/10.1016/j.gdata.2014.09.014
2213-5960/© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

MAS5 values were taken from GEO if available. For samples with no MAS5 values, CEL files were downloaded from GEO and the *affy* package [1] from Bioconductor [2] was used to generate MAS5 values. Next, MAS5 values corresponding only to the 22,283 probesets from the U133A array were compiled. Subsequently, normalization of MAS5 data was performed using the command line version of the program CLUSTER 3.0 (Michael Eisen; updated by Michiel de Hoon; http://bonsai.hgc.jp/~mdehoon/software/cluster/command.txt).

The following three steps were performed in the following order:

1. log2 transformation of MAS5 values
2. median centering of arrays
3. magnitude normalization of arrays.

These three steps correspond to the following commands:

cluster.com filename -l
cluster.com filename -ca m
cluster.com filename -na

In step 3 of these procedures (magnitude normalization) the expression values of all (22,283) probesets from the U133A array are multiplied by a scale factor S so that the magnitude (sum of the squares of the values) equals one. The resulting dataset was used for the subsequent analyses. The normalized data are available under the following link:

http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE31519&format=file&file=GSE31519%5Fcomplete%5Fdataset%2Etxt%2Egz

All 579 samples in the dataset are triple negative according to the following predefined cutoffs [3] for ESR1 (205225_at) < 0.0075, PGR (208305_at) < −0.0078, and HER2 (216836_s_at) < 0.0135.

An R script of the subsequent analysis is available in the Supplementary data.

Analyses

A major concern of the pooling procedure are systematic technical differences between individual datasets ("batch effects"). Many adaptation methods as e.g. Z-normalization often do not eliminate but rather blur such effects. Thus we applied two further strategies to cope with this problem. First, we selected only highly comparable datasets for our finding cohort. Second, we controlled for biased genes which still show associations with the dataset vector. These two strategies are described below.

Comparability of datasets

The 579 arrays came from 28 different datasets. We calculated a comparability metric C for each of the datasets to identify the most comparable samples. This metric C is derived from the sum of the squared differences of the mean (μ) within a specific dataset and among all datasets, respectively, normalized by the standard deviation (σ) calculated for all genes (g) on the array:

$$C_{dataset_i} = \sum_{g=1}^{n} \left(\frac{\mu_{g,dataset_i} - \mu_{g,total}}{\sigma_{g,total}} \right)^2.$$

The metric is based on the assumption that overall the mean of a gene expression within a dataset should be similar between different datasets and gives an estimation to what extent the arrays in a specific dataset differ from the combined overall cohort. Larger datasets will dominate because of their higher impact on the global mean. All datasets were sorted according to this metric and the top 15 datasets with the lowest values (normalized $C \le 0.03$), corresponding to 394 samples in total, were used as the discovery cohort (Fig. 1).

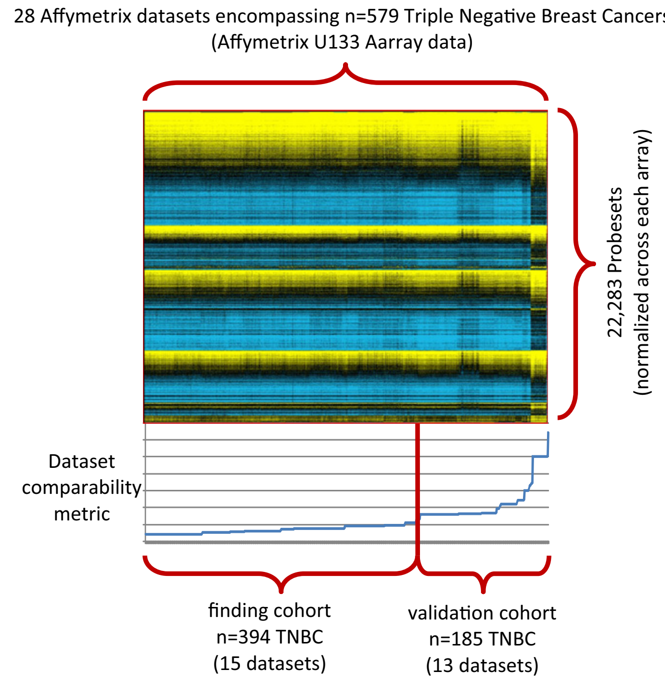

Fig. 1. Selection of the TNBC finding cohort from multiple datasets based on dataset comparability. Triple negative breast cancers (TNBCs, n = 579) from 28 datasets were sorted by dataset according to a dataset comparability metric (horizontally). Shown are the full array data of normalized Affymetrix U133A microarrays. The 15 most comparable datasets encompassing n = 394 TNBC samples were subsequently used as a finding cohort and the remaining 13 datasets (n = 185 TNBC samples) were withheld as a validation cohort.

Control for biased probesets

All probesets were checked for dataset bias (i.e. differential expression by dataset of origin that would indicate laboratory-bias or sampling

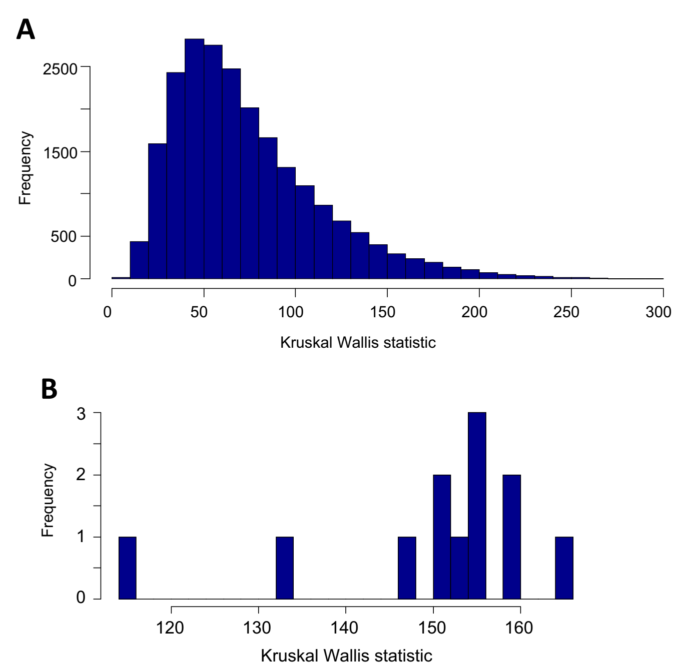

Fig. 2. Analysis of dataset bias among probesets. A) The standard Kruskal–Wallis rank test was used to analyze the dependence of each individual probeset's expression on the vector of the 15 different datasets in the finding cohort of n = 394 samples. The distribution of the rank sum statistics for all 22,283 probesets from the U133A array is shown. B) Distribution of the Kruskal–Wallis rank sum statistics among the 12 biased probesets of the hemoglobin metagene.

differences compared to the rest). To assess dataset bias, we used Kruskal–Wallis statistic comparing the expression of each probeset with the primary dataset vector across the 394 TNBCs. Each probeset was then tagged with that Kruskal–Wallis value throughout all analyses. Thus an enrichment of biased probesets can be monitored in any downstream application e.g. cluster analyses [4–6]. Cutoffs for exclusion of probesets due to strong dataset bias may be derived from the distribution of the Kruskal–Wallis statistic over all probesets. Fig. 2 demonstrates the enrichment of biased probesets in the hemoglobin metagene reported in [4]. This effect originated from the inclusion of two datasets which were obtained from fine needle aspiration (FNA) samples. Such samples generally contain relatively higher amounts of blood and lower amounts of stromal tissue as compared to surgical biopsy samples.

Acknowledgements

This work was supported by grants from the H.W. & J. Hector-Stiftung, Mannheim (grant number: M67).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.09.014.

References

[1] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry, affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 12 (20(3)) (2004) 307–315 (Feb).
[2] 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.
[3] T. Karn, D. Metzler, E. Ruckhäberle, L. Hanker, R. Gätje, C. Solbach, A. Ahr, M. Schmidt, U. Holtrich, M. Kaufmann, A. Rody, Data-driven derivation of cutoffs from a pool of 3,030 Affymetrix arrays to stratify distinct clinical types of breast cancer. Breast Cancer Res. Treat. 120 (3) (2010 Apr) 567–579.
[4] A. Rody, T. Karn, C. Liedtke, L. Pusztai, E. Ruckhaeberle, L. Hanker, R. Gaetje, C. Solbach, A. Ahr, D. Metzler, M. Schmidt, V. Müller, U. Holtrich, M. Kaufmann, A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res. 13 (5) (2011 Oct 6) R97.
[5] T. Karn, L. Pusztai, U. Holtrich, T. Iwamoto, et al., Homogeneous datasets of triple negative breast cancers enable the identification of novel prognostic and predictive signatures. PLoS One 6 (12) (2011) e28403.
[6] T. Karn, L. Pusztai, E. Ruckhäberle, C. Liedtke, et al., Melanoma antigen family A identified by the bimodality index defines a subset of triple negative breast cancers as candidates for immune response augmentation. Eur. J. Cancer 48 (1) (2012 Jan) 12–23.