DATA NOTE

A collection of annotated and harmonized human breast cancer transcriptome datasets, including immunologic classification [version 1; peer review: 2 approved with reservations]

Jessica Roelands [1], Julie Decock [2], Sabri Boughorbel [3], Darawan Rinchai[1], Cristina Maccalli[1], Michele Ceccarelli[4], Michael Black[5], Cris Print[6], Jeff Chou[7], Scott Presnell[8], Charlie Quinn[8], Puthen Jithesh[9], Najeeb Syed[10], Salha B.J. Al Bader[11], Shahinaz Bedri[12], Ena Wang[13], Francesco M. Marincola[14], Damien Chaussabel [3], Peter Kuppen[15], Lance D. Miller[7*], Davide Bedognetti[1*], Wouter Hendrickx [1*]

[1]Tumor Biology, Immunology and Therapy section, Sidra Medical and Research Center, Doha, Qatar
[2]Qatar Biomedical Research Institute, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar
[3]Systems Biology Department, Sidra Medical and Research Center, Doha, Qatar
[4]Qatar Computing Research Institute, Doha, Qatar
[5]Department of Biochemistry, Otago School of Medical Sciences, University of Otago, Dunedin, 9054, New Zealand
[6]Department of Molecular Medicine and Pathology and Maurice Wilkins Institute, Faculty of Medical and Health Sciences, The University of Auckland, Auckland, 1142, New Zealand
[7]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC, 27101, USA
[8]Benaroya Research Institute at Virginia Mason, Seattle, WA, 98101, USA
[9]Translational Bioinformatics, Division of Biomedical Informatics Research, Sidra Medical and Research Center, Doha, Qatar
[10]Technical Bioinformatics team, Biomedical Informatics Division, Sidra Medical and Research Center, Doha, Qatar
[11]National Center for Cancer Care and Research (NCCCR), Hamad General Hospital, Doha, Qatar
[12]Weill Cornell Medicine - Qatar, Doha, Qatar
[13]Division of Translational Medicine, Research Branch, Sidra Medical and Research Center, Doha, Qatar
[14]Office of the Chief Research Officer (CRO), Research Branch, Sidra Medical and Research Center, Doha, Qatar
[15]Department of Surgery, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

[*] Equal contributors

v1 **First published:** 20 Mar 2017, **6**:296 (https://doi.org/10.12688/f1000research.10960.1)
Latest published: 09 Feb 2018, **6**:296 (https://doi.org/10.12688/f1000research.10960.2)

Abstract

The increased application of high-throughput approaches in translational research has expanded the number of publicly available data repositories. Gathering additional valuable information contained in the datasets represents a crucial opportunity in the biomedical field. To facilitate and stimulate utilization of these datasets, we have recently developed an interactive data browsing and visualization web application, the Gene Expression Browser (GXB). In this note, we describe a curated compendium of 13 public datasets on human breast cancer, representing a total of 2142 transcriptome profiles. We classified the samples according to

Open Peer Review

Reviewer Status ✔ ✔

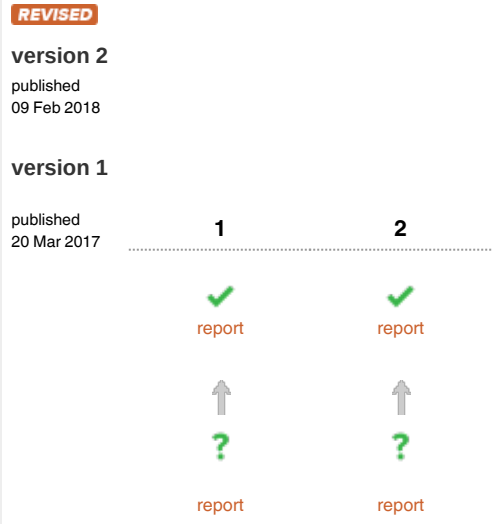
Invited Reviewers

different immune based classification systems and integrated this information into the datasets. Annotated and harmonized datasets were uploaded to GXB. Study samples were categorized in different groups based on their immunologic tumor response profiles, intrinsic molecular subtypes and multiple clinical parameters. Ranked gene lists were generated based on relevant group comparisons. In this data note, we demonstrate the utility of GXB to evaluate the expression of a gene of interest, find differential gene expression between groups and investigate potential associations between variables with a specific focus on immunologic classification in breast cancer. This interactive resource is publicly available online at: http://breastcancer.gxbsidra.org/dm3/geneBrowser/list.

Keywords
Breast Cancer, Immune Subtypes, Cancer Immune Phenotype, Gene Expression Browser, Immunologic Constant of Rejection

This article is included in the Sidra Medicine gateway.

This article is included in the Data: Use and Reuse collection.



1 **Benjamin Haibe-Kains**, University Health Network, Toronto, Canada

2 **Christos Hatzis**, Yale School of Medicine, New Haven, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Lance D. Miller (ldmiller@wakehealth.edu), Davide Bedognetti (dbedognetti@sidra.org), Wouter Hendrickx (whendrickx@sidra.org)

Competing interests: No competing interests were disclosed.

Grant information: JD, SB, DR, DC, DB, WH received support from the Qatar Foundation. JR received support from Qatar National Research Fund (grant number: JSREP07-010-3-005).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Roelands J, Decock J, Boughorbel S *et al.* **A collection of annotated and harmonized human breast cancer transcriptome datasets, including immunologic classification [version 1; peer review: 2 approved with reservations]** F1000Research 2017, **6**:296 (https://doi.org/10.12688/f1000research.10960.1)

First published: 20 Mar 2017, **6**:296 (https://doi.org/10.12688/f1000research.10960.1)

Introduction

Technological progress in the field of biomedical research has resulted in an increased utilization of platforms generating information on a system-scale, e.g. genome, transcriptome and proteome. As researchers are typically willing and often required to share their data collections, the availability of 'big data' is expanding rapidly. At this moment, the NCBI Gene Expression Omnibus (GEO), a public repository of transcriptome profiles, holds over 2 million individual transcriptome profiles from more than 76,000 studies ("Home - GEO - NCBI", 2016). This large amount of available transcriptomic data provides major opportunities as well as challenges to researchers. Identification of differential gene expression in healthy versus diseased individuals, for example, has the potential to increase our understanding of the disease process, can lead to the identification of novel disease biomarkers or to the recognition of potential therapeutic targets. However, utilization of the available system-scale information can be challenging, since data repositories often lack the analytical and visualization tools needed for data assessment and interpretation. For this reason, proper analysis relies on elevated bioinformatics skills.

To overcome the challenges faced when analyzing transcriptomic data, we previously developed a web application called gene expression browser (GXB), which makes datasets more accessible and interactive (Speake *et al.*, 2015). The application graphically visualizes gene expression data in bar chart or box plot representation and is capable of dynamically changing its interface views upon user input. GXB allows users to upload microarray data, add data annotations, which enables overlay of clinical data, explore gene rank lists based on their differential expression patterns between groups, view the data on a gene-by-gene basis and compare different datasets and diseases. These capabilities stimulate the acquisition of new knowledge from public datasets, as demonstrated by the first paper that employed GXB to identify a previously unknown role of a specific transcript during immune-mediated processes (Rinchai *et al.*, 2015).

In recent years, a large number of transcriptional studies have been conducted with the aim to characterize breast cancer on a genetic basis. GEO holds about 1297 datasets relating to breast cancer. One of the main impacts gene expression profiling has had on our understanding of breast cancer has been through the classification of breast cancer into intrinsic molecular subtypes (IMS). Three main methods have been described to achieve this, which have the same subtypes, but actually use different gene sets to stratify the patients (Hu *et al.*, 2006; Parker *et al.*, 2009; Sorlie *et al.*, 2003;). Four major IMS of breast cancer have been identified: Luminal A, Luminal B, HER2-enriched and Basal-like. A less common molecular subtype called Claudin-low has been characterized at a later time point (Prat *et al.*, 2010). Stratified IMS groups present critical differences in incidence, survival and response to treatment, and most importantly add prognostic information that is not provided by classical stratifications, like estrogen receptor status, histologic grade, tumor size, and node status (Parker *et al.*, 2009).

Recent breakthroughs in the field of cancer immunotherapy and especially the application of checkpoint blockade inhibitors has ignited a fierce drive to understand the genetic basis for the huge differences observed between patients with different immune phenotypes. Several papers have shown that expression profiles are able to distinguish between those patients that have an active immune environment and those that do not (Galon *et al.*, 2013; Herbst *et al.*, 2014; Ji *et al.*, 2012; Ribas *et al.*, 2015; Wang *et al.*, 2013). A clear correlation can be seen both regarding prognosis (survival) and prediction of therapeutic effectiveness of immune regulatory therapies. The expression of genes observed in association with tissue-specific destruction in a broader context, defined as the immunological constant of rejection (ICR), can distinguish between breast cancer patients with different prognosis. This immunological classification is based on the consensus clustering of ICR genes (Galon *et al.*, 2013), e.g. genes underlying Th1 polarization, related chemokines, adhesion molecules and cytotoxic factors, in combination with immune regulatory genes IDO1 and FOXP3, PDCD1, CTLA4 and CD274/PD-L1 (Figure 1A) (Bedognetti *et al.*, 2015). In Miller *et al.* (2016), a novel survival-based immune classification system was devised for breast cancer based on the relative expression of immune gene signatures that reflect different effector immune cell subpopulations, namely antibody-producing plasma B cells (the B/P metagene), cytotoxic T and/or NK cells (the T/NK metagene), and antigen-presenting myeloid/dendritic cells (the M/D metagene). The system defines a tumor's immune subclass based on its survival-associated immunogenic disposition status (IDS), which discriminates between *poor immunogenic disposition* (PID), *weak immunogenic disposition* (WID) and *favorable immunogenic disposition* (FID). The ability of IDS to distinguish patients with differential prognosis is dependent on the tumor's immune benefit status (IBS), which is defined by IMS and the expression of cell proliferation markers. The IBS classification segregates immune benefit-enabled (IBE) and immune benefit-disabled (IBD) tumors. In IBE tumors, but not IBD tumors, FID status confers a protective survival benefit compared to WID and PID status (Figure 1B) (Miller *et al.*, 2016; Nagalla *et al.*, 2013). In this data note, we demonstrate the use of GXB to evaluate cancer gene expression across immunologic classifications of breast cancer.

Since the amount of possible datasets to be included in GXB is enormous, we chose to start with the GEO datasets underlying the immunologically classified breast cancer datasets by (Miller *et al.*, 2016). In Hendrickx *et al.* (2017), these same datasets were classified according to ICR. This will allow us to share our immune related classifications in a comprehensible way and allow others to reuse them. A harmonization effort of the other available clinical data had been undertaken and should help the downstream analysis of the expression data. Therefore, gathering these datasets with their detailed study and sample information will facilitate the identification of clinically-relevant genetic signatures for biomarker and/or therapeutic purposes.

In this data note, using GXB, we have made available a curated compendium of 13 public datasets relevant to human breast cancer, representing a total of 2142 cases.

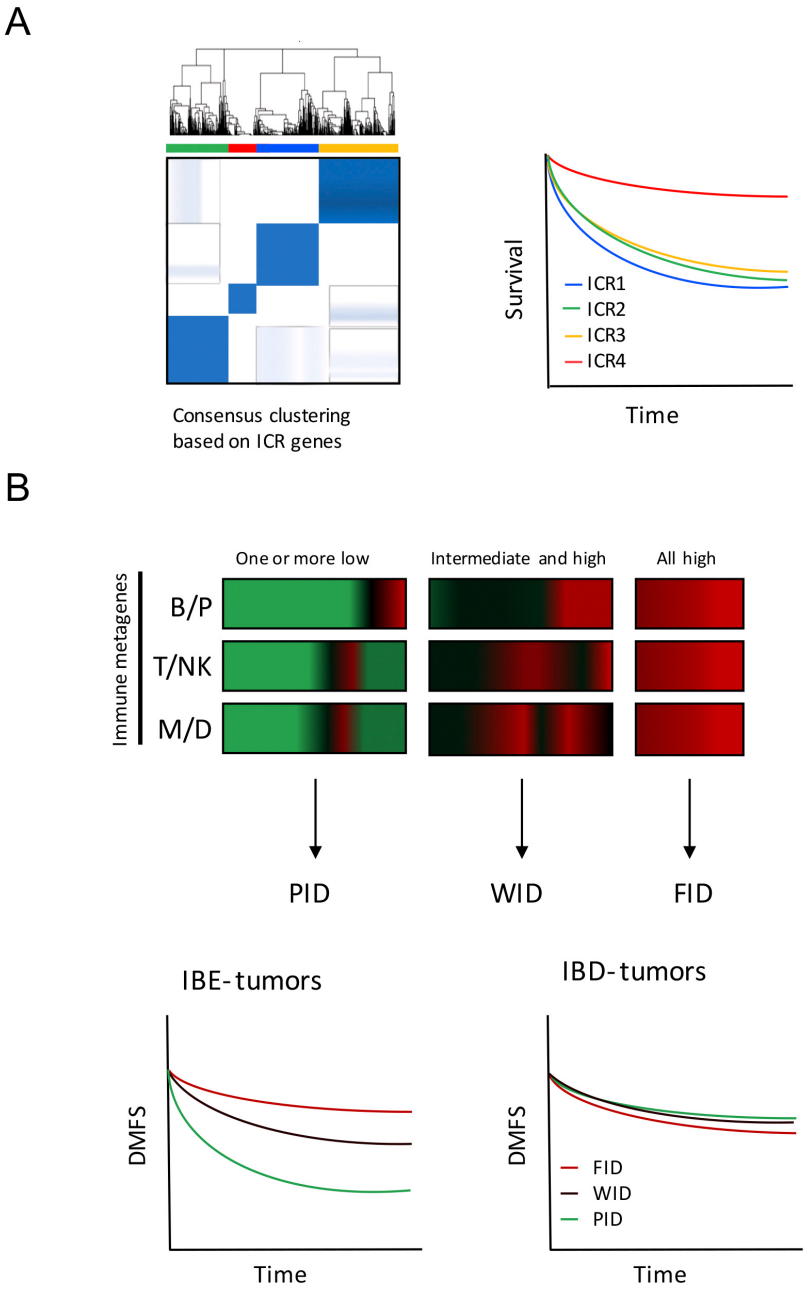


Figure 1. Basis of ICR and IDS/IBS classifications and prognostic value. (**A**) Consensus clustering based on ICR genes segregates breast cancer patient in four different groups: ICR1, 2, 3 and 4. Patients with tumors categorized as ICR4 have the highest expression of the ICR gene signature and have a better prognosis compared with other ICR groups. (**B**) Immune metagene model based on the relative expression of immune metagenes (B/P, T/NK and M/D) distinguishes PID, WID and FID tumors (horizontal axis: genes, vertical axis: individual cases). This classification has prognostic value in IBE tumors, and not in IBD tumors. Diagrams are based on Hendrickx *et al.*, 2017 (**A**) and Miller *et al.*, 2016 (**B**). ICR, Immunologic Constant of Rejection; IBE/D, Immune Benefit Enabled OR Disabled; F/P/WID, Favorable OR Poor OR Weak Immune Disposition.

Methods

Selection of breast cancer datasets

The starting point of our selection of breast cancer datasets are the patient cohorts included in the multi-study breast cancer database described by Nagalla *et al.* (2013). These 13 NCBI GEO datasets (GEO accession numbers: GSE45255, GSE2034, GSE5327, GSE12093, GSE9195, GSE11121, GSE1456, GSE2603, GSE6532, GSE7390, GSE7378 and GSE4922) resulted in 2142 cases initially uploaded in GXB. 22 of these cases reflect data from breast cancer cell lines and were therefore excluded from our data

collection. A total of 1839 cases represent primary invasive breast tumors sampled at the time of surgical resection without prior neoadjuvant treatment and were therefore annotated with survival data, IMS, IBS, IDS and ICR status (Hendrickx *et al.*, 2017; Miller *et al.*, 2016). 281 of the cases did not fulfill these criteria and were therefore not annotated. Of note, 115 cases of original meta-cohort used Nagalla's study (n=1954) were not shared within GEO, but shared within other platforms (caArray and ArrayExpress). For this reason, these samples were not included in our GXB collection (Figure 2).

The datasets that comprise our collection are listed in Table 1 and can be searched interactively in GXB. All GEO datasets consist of unique cases with the exception for 36 cases from NUH Singapore, which are both present in the Bordet Radcliff NUH (GSE45255) dataset and the Uppsala and Singapore (GSE4922) dataset.

Data of the 1839 GEO-cases annotated with survival data that were previously combined and used in the Nagalla study, have been uploaded to GXB in the dataset "Nagalla 2013 reconstituted public dataset".

Dataset upload into GXB

All datasets were downloaded from NCBI GEO in SOFT file format and were uploaded into GXB with the exception of the Guy's hospital dataset (GUYT2; GSE9195). Expression data in the SOFT file of this dataset was expressed as fold change. Therefore, we had to revert to reprocessing of the CEL files found attached to the GSE on GEO. In this case, the cell files were read into R (v3.2.2) using the 'affy' package (v1.50.0). Data was normalized using the RMA (Robust multichip averaging) and gene annotation data was added using the hgu133plus2.db package (v3.2.3).

GSE records containing data generated with different or multiple platforms have been split by platform using the import process of GXB. GSEs containing data from both clinical as *in vitro* origin (GSE2603) have been split manually using the GXB Graphical interface.

Metadata of the different studies was added to GXB both from the descriptive information found on GEO or from the method sections of the publications linked to these datasets. Short links to PMID (Pubmed) and GEO records were added.

Construction of the Nagalla's dataset

The constitution of the complete cohort has previously been described by (Nagalla *et al.*, 2013). The dataset "Nagalla 2013 reconstituted public dataset" available in GXB contains only the samples that were publicly available via GEO. Briefly, raw data (CEL files) were extracted from GEO. The array platforms employed for these 13 datasets were Affymetrix U133A, U133A2, and U133 PLUS 2.0 gene chips; the 22,268 probe sets present in each of these platforms were included in the gene expression file. Data were MAS5.0 normalized using the justMAS function in the simpleaffy library from Bioconductor (Gentleman *et al.*, 2004) using a trimmed mean target intensity of 600 without background correction. COMBAT empirical Bayes method was used to correct for batch effects (Johnson *et al.*, 2007).

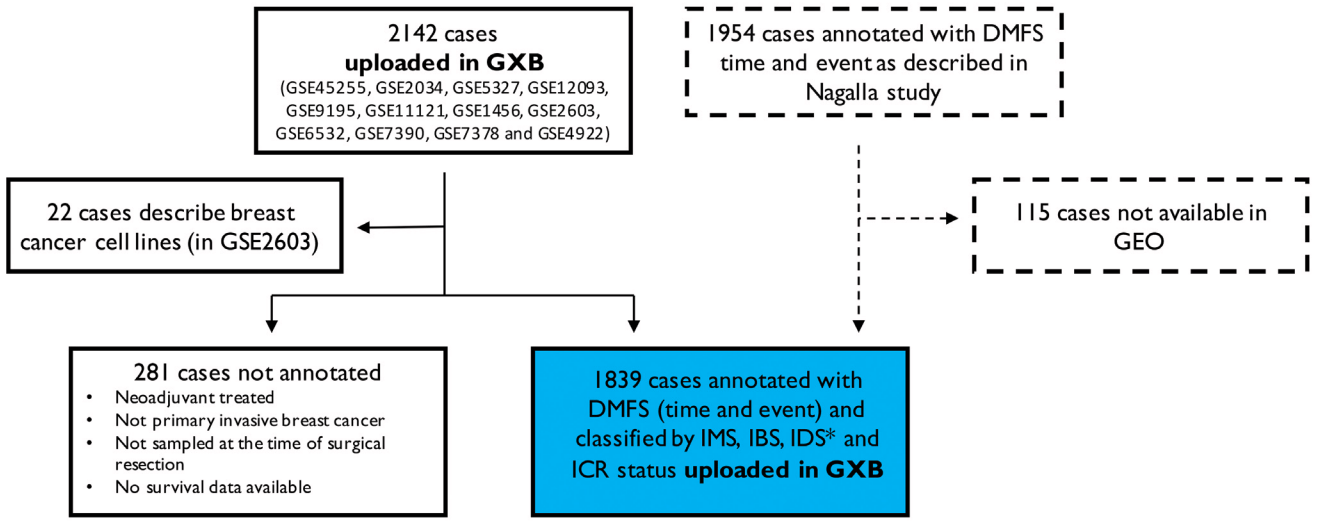


Figure 2. Schematic representation of dataset selection and annotation. Breast cancer cases included in 13 NCBI GEO datasets were uploaded in GXB (n=2142). 22 cases described data from breast cell lines and were excluded from our data collection. We annotated 1839 cases with survival data, IMS, IBS, IDS and ICR status. 281 cases were either neoadjuvant treated, did not represent a primary invasive tumor, were not sampled at the time of surgery or without available survival data and were therefore not annotated. The total collection includes 1839 cases from the original cohort described in Nagalla *et al.* (2013) (n=1954). Of note, 115 cases of this cohort are not included in our collection as these were not shared via GEO. *251/1839 cases have been classified for IMS "Normal-like". IDS is not applicable for normal-like breast cancer tissue; therefore, IDS is non-classified for these samples. DMFS, Distant Metastasis Free Survival; GXB, Gene Expression Browser; IMS, intrinsic molecular subtype; IBS, immune benefit status; IDS, immune disposition status; ICR, immunologic constant of rejection.

Table 1. List of datasets uploaded to GXB.

Dataset	Platforms	Diseases	Number of samples	GEO ID	References
Bordet Radcliffe NUH dataset - GSE45255.GPL96	Affymetrix Human Genome U133A Array	Breast Cancer	139	GSE45255	(**Nagalla *et al.*, 2013**)
Erasmus Medical Center (EMC) dataset 1 - GSE2034.GPL96	Affymetrix Human Genome U133A Array	Lymph Node Negative Breast Cancer	286	GSE2034	(**Y. Wang *et al.*, 2005**)
Erasmus Medical Center (EMC) dataset 2 - GSE5327.GPL96	Affymetrix Human Genome U133A Array	Lymph Node Negative Breast Cancer	58	GSE5327	(**Minn *et al.*, 2007**)
Europe and Cleveland (EMCT) dataset - GSE12093.GPL96	Affymetrix Human Genome U133A Array	ER + Breast Cancer	136	GSE12093	(**Zhang *et al.*, 2009**)
Guy's hospital dataset (GUYT2) - GSE9195.GPL570.fCEL	Affymetrix Human Genome U133 Plus 2.0 Array	ER+ Breast Cancer	77	GSE9195	(**Loi *et al.*, 2008**)
Johannes Gutenberg University (MAINZ) dataset - GSE11121.GPL96	Affymetrix Human Genome U133A Array	LN- Breast Cancer	200	GSE11121	(**Schmidt *et al.*, 2008**)
Karolinska (STO) dataset - GSE1456.GPL96 +GPL97	Affymetrix Human Genome U133A Array & Affymetrix Human Genome U133B Array	Breast Cancer	159	GSE1456	(**Pawitan *et al.*, 2005**)
Memorial Sloan-Kettering Cancer Center (MSKCC) dataset - GSE2603.GPL96_Clinical samples	Affymetrix Human Genome U133A Array	Breast Cancer	99	GSE2603	(**Minn *et al.*, 2005**)
Nagalla 2013 reconstituted public dataset	Affymetrix Human Genome U133A Array & Affymetrix Human Genome U133A2 Array & Affymetrix Human Genome U133 Plus 2.0 Array	Breast Cancer	1839	multiple	(**Nagalla *et al.*, 2013**)
Princess Margaret Cancer Centre dataset (GUYT) - GSE6532.GPL570	Affymetrix Human Genome U133 Plus 2.0 Array	ER+ Breast Cancer	87	GSE6532	(**Loi *et al.*, 2007**)
Princess Margaret Cancer Centre dataset - GSE6532.GPL96 +GPL97	Affymetrix Human Genome U133A & U133B Array	ER+ Breast Cancer	327	GSE6532	(**Loi *et al.*, 2007**)
TRANSBIG (TBIG) dataset - GSE7390.GPL96	Affymetrix Human Genome U133A Array	Lymph Node Negative Breast Cancer	198	GSE7390	(**Desmedt *et al.*, 2007**)
University of California San Francisco (YAU) dataset - GSE7378.GPL4685	Affymetrix GeneChip HT-HG_U133A Early Access Array	ER+ Breast Cancer	54	GSE7378	(**Zhou *et al.*, 2007**)
Uppsala and Singapore dataset - GSE4922.GPL96 +GPL97	Affymetrix Human Genome U133A & U133B Array	Breast Cancer	289	GSE4922	(**Ivshina *et al.*, 2006**)

Clinical data annotation

Gene expression data is accompanied with clinical data in CSV file format. Gene expression data and clinical data are coupled to the sample via variable "Sample ID". We annotated a total of 1839 cases with 10-year survival (time and event), IBS (IBE, IBD), IDS (PID, WID and FID) (Miller *et al.*, 2016) and ICR (ICR1, ICR2, ICR3 and ICR4) immune classifications (Hendrickx *et al.*, 2017) (Figure 2). IMS (i.e., Basal-like, HER2-enriched, Luminal A and Luminal B) were defined using the Single Sample Predictor (SSP) algorithm by Hu (Hu *et al.*, 2006) utilized by (Fan *et al.*, 2006). Claudin-low tumors were identified using the method of (Prat *et al.*, 2010). Of the 1839 samples, 251 samples were "Normal-like" in IMS classification. Therefore, these samples are not classified according to IDS. For the separate dataset containing samples of *in vitro* origin (GSE2603), survival annotations and immune classifications are not applicable. A final 281 cases were not annotated and non-classified, since for these cases either samples were not taken at the time of surgical resection, were neoadjuvant-treated or cases were not annotated with distant metastasis free survival (DMFS) time and event.

To enable comparisons between datasets and to facilitate efficient data analysis, the clinical data was harmonized to reflect a nomenclature similar to that of The Cancer Genome Atlas (TCGA).

Clinical variable names and availability in datasets are listed in Table 2. In general, variable values have been replaced by descriptive values (e.g. "1" and "0" are replaced by "ER+" and "ER-", respectively). For disease free survival, variable values have been adapted to "DiseaseFree" or "Recurred/Progressed", and for distant metastasis survival to "DistantMetastasisFree" and "DistantMetastasis". Numeric values of variable "tumor size" have been converted to units in cm for all datasets. This variable was used to generate the additional variable pathology T stage according to the 7th edition of the AJCC staging system for breast cancer (Edge & Compton, 2010). For tumors with a diameter larger than 5 cm, pathology T stage could be either T3 or T4, therefore value "T3/T4" has been assigned to these cases.

Table 2. Clinical data availability.

Clinical variable	Available in *N* datasets
IMS	13
IBS	13
IDS	13
ICR	13
DMFS 10Y EVENT	13
DMFS 10Y TIME	13
Disease free survival event	11
Disease free survival time	11
Distant metastasis free survival event	6
Distant metastasis free survival time	6
Age at initial pathologic diagnosis	8
Lymph node status	8
ER status	8
PR status	5
Histology differentiation grade	7
Tumor size	8
Pathology T Stage	8
Type treatment, bone metastasis event, bone metastasis free survival time, breast cancer cause of death, HER2 status, histologic diagnosis, lung metastasis event, lung metastasis free survival time, lung metastasis gene expression signature status, vital status, angio invasion indicator, disease specific survival time, genetic grade signature status sws classifier, GGI indicator, lymph nodes examined count, number of lymph nodes positive, lymphocyte infiltration, molecular subtype, NPI, overall survival, p53 mutation status, probability by sws classifier, RFS 5Y EVENT, risk AOL indicator, risk NPI indicator, risk SG, risk veridex indicator, tissue type, van 't Veer signature.	<2

Standardized clinical datasets can be found in the 'downloads' tab in GXB under the heading "additional files". All datasets start with the following 21 clinical variables in fixed order: "sample.ID", "array sample id", "sample title", "series", "IMS", "IBS", "IDS", "ICR", "DMFS_10Y_EVENT", "DMFS_10Y_TIME", "disease free survival event", "disease free survival years", "distant met free survival event", "distant met free survival", "age at initial pathologic diagnosis", "lymph node status", "ER status", "PR status", "histology differentiation grade", "tumor size cm", and "pathology T stage". In case one of these variables is not available in a specific dataset, values in this column are all NA.

Group sets for IBS/IDS, ICR cluster, Lymph Node (LN) Status, IMS, Histological grade, stage and Estrogen Receptor (ER) status were defined with matching differential gene expression rank lists. Rank lists are based on differential gene expression between two relevant groups for each group set: IBD-FID *vs* IBE-FID (IBS/IDS); ICR1 *vs* ICR4 (ICR1/ICR4); LN+ *vs* LN- (LN status); G1 *vs* G3 (histological grade); ER+ *vs* ER- (ER status). For IMS, no rank list was generated, as this variable is not ordered. For tumor stage, no rank list was generated because the spread of samples between categories was small.

Dataset demonstration

Utilization of GXB

The GXB software has been described in detail in a recent publication (Speake *et al.*, 2015). This custom software interface provides users with a means to easily navigate and filter the dataset collection available at http://breastcancer.gxbsidra.org/dm3/landing.gsp. A web tutorial is also available online: http://breastcancer.gxbsidra.org/dm3/tutorials.gsp#gxbtut.

Example case: Expression of HLA-G across ICR groups

In GXB, users can search interactively for a specific gene of interest. Differential expression across different group sets can be observed in the graphical interface, either in bar or box plots. For illustrative purposes, we choose to evaluate the abundance of the HLA-G transcripts across ICR groups.

HLA-G is a non-classical class I gene of human Major Histocompatility Complex that is primarily expressed on fetal derived placental cells (Ellis *et al.*, 1990). In contrast to its classical counterparts, HLA-G does not initiate immune responses, but instead has immunosuppressive effects (Naji *et al.*, 2014; Rouas-Freiss *et al.*, 1997). Expression of HLA-G has been reported in a variety of cancers, including breast cancer, and has been assigned a role in tumor immune escape (Naji *et al.*, 2014; Rouas-Freiss *et al.*, 1997; Swets *et al.*, 2016; Zeestraten *et al.*, 2014).

Concerning its role in tumor immunity, it may be of interest to investigate whether HLA-G expression is elevated in breast tumors of specific immune phenotypes. The ICR gene signature segregates breast tumors into four immune phenotype groups based on the expression of genes underlying immune-mediated tissue-specific destruction, with ICR1 having the lowest and ICR4 the highest expression of this signature (Bedognetti D *et al.*, in press).

To compare HLA-G expression across ICR groups using the breast cancer datasets uploaded to GXB, we start by selecting one of the datasets. After opening this dataset: 1) the gene of interest, HLA-G, can be identified using the search box in the upper left corner of the user interface. Upon selection of "HLA-G" in the left panel, the central panel displays the expression values of this gene for all samples as a bar chart. 2) Sample grouping is default as "All sample", it is changed by selecting "Immunologic Constant of Rejection" and 3) plot type is set to "Box Plot" in drop down menus in the central panel. The central panel now presents a graphical display of the observed abundance of HLA-G transcripts in breast cancer samples across the different ICR groups, each sample is represented by a single point in a boxplot (Figure 3A). A tendency of increased HLA-G expression in groups with the highest expression of ICR genes can be observed. 4) To verify whether this trend can also be observed in other breast cancer datasets, GXB's "Cross Project View" is used. By selecting "Cross Project View" in the "Tools" drop-down menu located in the top right corner of the user interface, a list of available datasets/projects appears in the left pane. By consecutive selection of single datasets, box plots with HLA-G transcripts across ICR groups are displayed for each individual dataset.

Figure 3. Illustrative example of abundance of HLA-G transcripts across ICR groups in multiple breast cancer datasets in GXB. (**A**) Cross Project View in GXB showing HLA-G expression across ICR groups. ICR represents the immune gene signatures observed in association with tissue-specific destruction. In this view of GXB, expression of HLA-G can be visualized across projects listed on the left. (**B**) Boxplots of HLA-G expression across ICR groups of three additional representative datasets selected from the dataset collection and the complete dataset including all annotated cases (right bottom plot). Plots indicate an increased HLA-G expression in breast tumors with a high expression of ICR genes. ICR, Immunologic Constant of Rejection.

Each of the boxplots corresponding to the 13 datasets show a similar pattern, indicating an increased HLA-G expression in breast tumors with a high expression of ICR genes (representative plots are shown in Figure 3B). In the combined dataset containing the total of 1839 annotated cases from these datasets, this trend is also observed (Figure 3B). From a biological perspective, increased expression of an immunosuppressant in an immunologically active tumor would be in line with our current view of the tumor microenvironment. Pro-inflammatory tumor environments, as observed in ICR4 tumors, also show counter regulatory mechanisms to suppress the immune system (Bedognetti *et al.*, 2015; Galon *et al.*, 2013).

This observation made by exploring transcriptome data in GXB provides an interesting starting point for further analysis. Statistical analysis of this potential association is required and, of course, the clinical relevance of the observed difference in abundance of transcripts should be determined. Most importantly, the functional relevance of HLA-G expression depends on its interaction with inhibitory receptors including ILT2, ILT4 and KIR2DL4 (LeMaoult *et al.*, 2005). Therefore, combined analysis of both HLA-G and these inhibitory receptors is suggested in future analyses.

This example illustrates the convenience of exploring gene expression data in GXB. The browser facilitates intuitive navigation and visualization of gene expression across different group sets.

Differential gene expression between IBS/IDS subgroups

The breast cancer datasets uploaded in GXB are provided with a rich context of immune classifications and clinical parameters. As opposed to start a search with a specific gene of interest, as presented in the HLA-G example case, differential gene expression between *groups* of interest can be explored in GXB by evaluation of gene rank lists. Here, we demonstrate the use of GXB to explore differential gene expression across IBS/IDS groups.

The IDS group set is based on an immune metagene model segregating breast tumors in groups of different immunogenic dispositions: PID, WID and FID (Nagalla *et al.*, 2013). The prognostic value of this classification is dependent on the molecular subtype and the proliferative capacity of the tumor, hereby segregating tumors in IBE and IBD groups, with and without prognostic value of the IDS, respectively. Since the hypothesis is that IBE-FID tumors confer metastasis-protective potential and IBD-FID tumors do not, transcriptional differences between these specific subgroups are of particular interest and have systematically been analyzed by Miller *et al.* (2016).

The Nagalla 2013 reconstituted dataset containing all annotated cases of this GXB breast cancer instance (n=1839) is used to explore differential gene expression between IBE-FID and IBD-FID tumors in GXB. Group set "Immune Benefit Status" is selected and corresponding gene rank list "IBD-FID *vs* IBE-FID" will load in the left panel by default. Filtering for specific immune gene categories, e.g. cytokine and chemokine ligands, cytokine and chemokine receptors, B and T cell signaling, and antigen presenting cell processing, is possible by selecting gene list category in the rank list menu. Exploring the expression of genes with known roles in tumor immunology reveals two important observations: 1) markers of immune cell infiltration, including CD8, CD3, CD19 and CD2, show similar expression in IBD-FID and IBE-FID subgroups (Figure 4A); while (2) markers of immune functional orientation, including CXCL10 (tissue rejection chemokine), GZMB (cytotoxic effector molecule), INFG and STAT1 (Th1 polarization), show differential expression across IBD-FID and IBE-FID groups (Figure 4B). A comprehensive statistical analysis of expression of these and other immune-related genes confirmed these observations, suggesting that while the composition of the immune infiltrate is similar in these tumors, the functional molecular orientation determines the metastasis-protective phenotype (Miller *et al.*, 2016).

This demonstration indicates that GXB allows for easy and efficient visualization of differential gene expression between subgroups. Subsequently, elaborate statistical analysis is required to confirm the differences in gene expression observed in GXB.

Overview of breast cancer immune classifications in GXB

Since this GXB data collection is provided with multiple immune classifications of breast cancer, it is interesting to visualize the relationship between these classifications in GXB. The overlay feature in GXB can be used to visualize the assignment of different classifications to individual samples simultaneously.

To illustrate this overlay option, we choose to select the Erasmus Medical Center dataset 2 (EMC2) with CXCL9 expression, as this is one of the chemokines included in the ICR gene signature. Graphical representation in GXB is set to bar plot and group set ICR is selected. As anticipated, the CXCL9 expression gradually increases from ICR1-ICR4. The drop down menu "Overlays" is used to add multiple layers of additional variables, "IBS", "IDS" and "IMS". Boxes underneath the individual bars (each bar represents a single case) display the assigned classifications (Figure 5A). When comparing IBS classifications across ICR groups, it is evident that IBE tumors are frequently assigned to the higher ICR clusters, ICR3 and ICR4, while IBD tumors tend to concentrate to the clusters with a low expression of the ICR signature (ICR1, ICR2) (Figure 5A). This result is consistent with our previous observations: pathways that distinguish IBE and IBD are associated with the immune functional orientation of the tumor, and genes in these same pathways are crucial components of the ICR signature (Bedognetti *et al.*, 2015; Miller *et al.*, 2016).

IDS relates to the ICR classification in a similar manner. FID tumors are mostly assigned to ICR4, while WID tumors are frequently classified to intermediate clusters (ICR2 and ICR3), and PID tumors prevail in the ICR1 cluster (Figure 5A). This observation is also in line with our expectations, the IDS classification is based on an immune metagene model that relies on immune gene

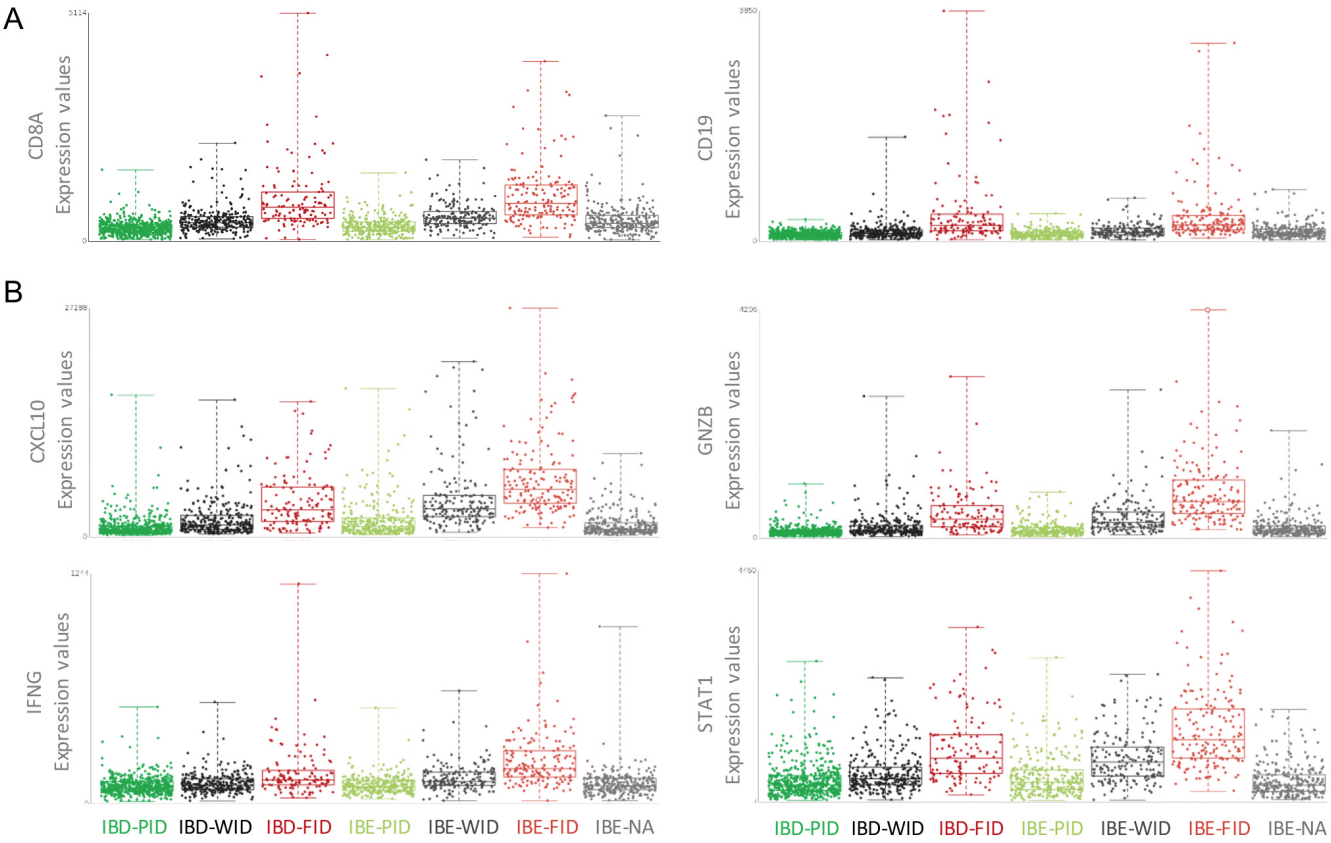


Figure 4. GXB overview of expression of genes with known roles in tumor immunology across IBS/IDS subgroups in reconstituted Nagalla's breast cancer dataset. (**A**) Expression values of CD8 and CD19, indicators of immune cell infiltration, are similar in IBD-FID and IBE-FID groups, indicating equal immune cell infiltration in these subgroups. (**B**) Expression values of CXCL10, GNZB, IFNG and STAT1, markers of immune functional orientation, are increased in the IBE-FID group compared with IBD-FID, indicating a differential functional orientation of the immune infiltrate between IBD-FID and IBE-FID tumors. IBE/D, Immune Benefit Enabled OR Disabled; F/P/WID, Favorable OR Poor OR Weak Immune Disposition.

subclusters that reflect the relative abundance of tumor-infiltrating immune cells (Nagalla *et al.*, 2013). As markers of immune cell infiltration are also included in the ICR signature, IDS is closely associated with ICR.

For a more comprehensive overview of the relationship between different immune classifications in breast cancer, the overlay of immune classifications was evaluated in the Nagalla 2013 reconstituted public dataset (n=1839). The observations made in the EMC2 dataset (n=58; Figure 5A) are also apparent in the dataset containing all annotated cases of this GXB breast cancer instance (Figure 5B). Moreover, in this dataset it is clearly visible that IBS/IDS subgroups with an improved prognosis are more prevalent in the ICR4 cluster. For example, IBE-FID tumors are relatively more frequently assigned to ICR4 compared with IBD-FID. Vice versa, IBD-PID tumors are proportionally more frequently observed in the ICR1 cluster compared with IBE-PID tumors, which are in comparison more frequently assigned to ICR2 ICR3.

The overlay of the different immune classifications demonstrates a coherency between the IBS/IDS classification and the ICR clusters. Bearing in mind that the ICR signature is associated with a broader phenomenon of immune-mediated, tissue-specific destruction, this coherency strengthens the hypothesis of a common final pathway of tissue destruction.

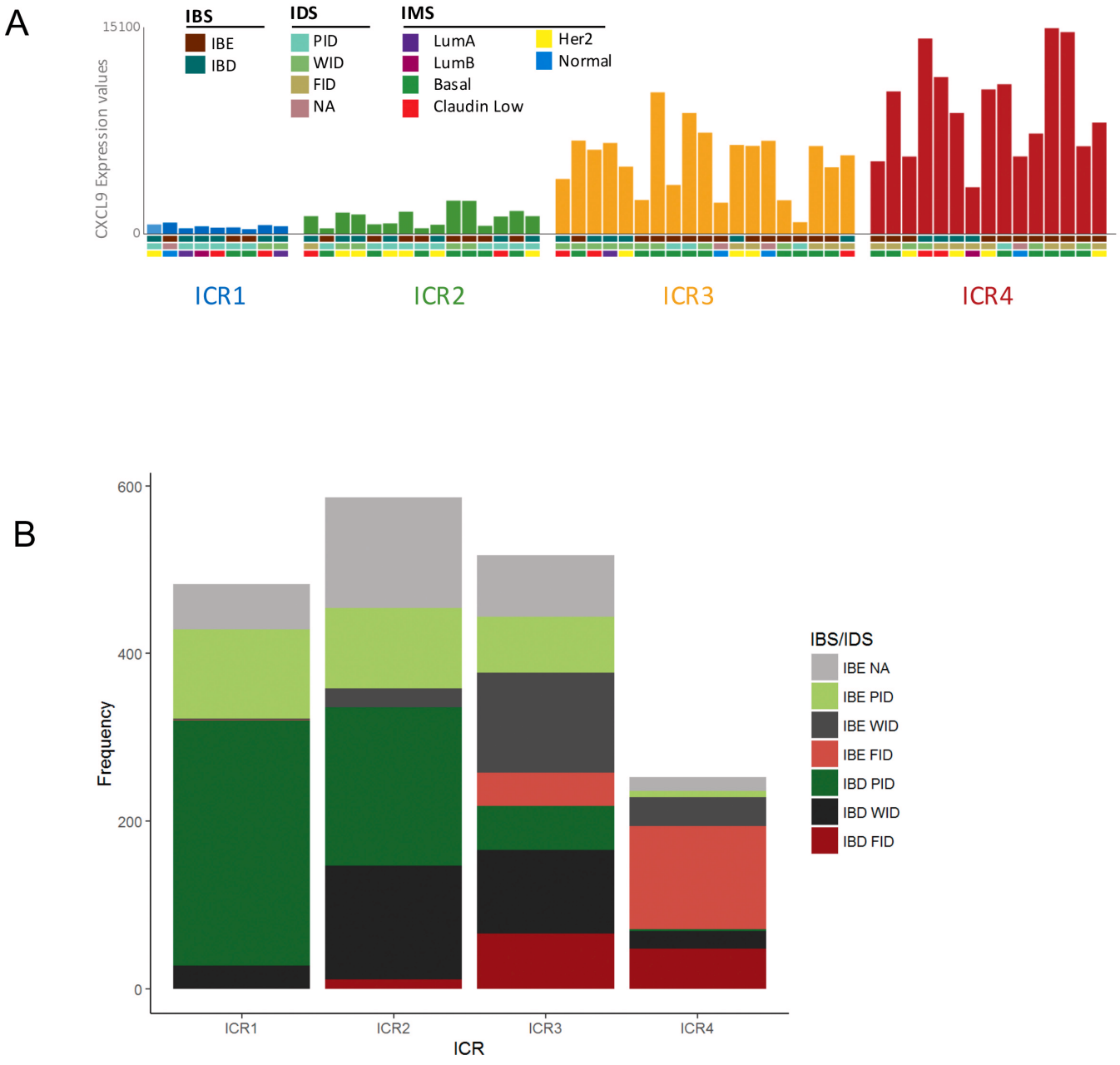


Figure 5. Overlay of immunologic classifications in breast cancer as evaluated in GXB. (**A**) Bar graph showing CXCL9 expression in individual samples from Erasmus Medical Center (EMC) dataset 2 split by ICR (single bar represents single case). Overlay of additional variables IBS, IDS and IMS is shown (http://breastcancer.gxbsidra.org/dm3/miniURL/view/Lv). (**B**) Frequency plot showing number of breast cancer cases across IBS/IDS subgroups split by ICR cluster. ICR, Immunologic Constant of Rejection; IBE/D, Immune Benefit Enabled OR Disabled; F/P/WID, Favorable OR Poor OR Weak Immune Disposition.

Conclusions

In this data note, we highlighted the opportunities provided by the availability of public datasets. We uploaded 13 public datasets on human breast cancer, including a combined dataset, with harmonized clinical data annotation and immune classification to GXB to facilitate the reuse of gene expression data. The use of GXB to explore gene expression and the different possible approaches were illustrated by the following: (1) an example case of a specific gene of interest, HLA-G; (2) comparison of gene expression between specific subgroups, IBD-FID vs IBE-FID; and (3) the evaluation of the relationship between different categorical variables, IBS/IDS and ICR immune classifications. To conclude, GXB provides a convenient environment to explore gene expression profiles in the context of breast cancer.

Data availability

All datasets included in our curated collection are available publicly via the NCBI GEO website: http://www.ncbi.nlm.nih.gov/geo/, and are referenced throughout the manuscript by their GEO accession numbers (e.g. GSE7390). Signal files and sample description files can also be downloaded from the GXB tool under the "downloads" tab.

Author contributions

JR: curated, uploaded and annotated datasets, interpreted the data and drafted the manuscript and accepted final version. **JD:** critically reviewed the datasets annotation, drafted the manuscript and accepted final version. **SBo:** installed the software, uploaded datasets, programmed portions of the web application, tested the software, and assisted in drafting the manuscript. **DR:** critically reviewed gene expression data, sample annotation, drafted the manuscript and accepted final version. **CM:** reviewed manuscript and accepted final version. **MC:** technical support annotating datasets and accepted final version of manuscript. **MB:** assembled and curated datasets and accepted final version of manuscript. **CP:** assembled and curated datasets and accepted final version of manuscript. **JC:** assembled and curated datasets and accepted final version of manuscript. **SP:** participated in the design of the software, programmed portions of the original web application, installed the software, tested the software, assisted in drafting the manuscript and accepted final version. **CQ:** participated in design and programmed portions of the original web application, tested the software, assisted in drafting the manuscript and accepted final version. **PJ:** critically reviewed sample annotation and gene expression data and accepted final version of manuscript. **NS:** critically reviewed quality of gene expression data and accepted final version of manuscript. **SBa:** reviewed manuscript and accepted final version. **SBe:** reviewed manuscript and accepted final version. **DC:** participated in software and study design, tested the software, assisted in drafting the manuscript and accepted final version. **EW, FM:** critically reviewed the manuscript. **PK:** Interpreted the data, critically reviewed manuscript and accepted final version. **LM:** assembled and curated datasets, provided immunological attributes, interpreted the data, drafted the manuscript and accepted final version. **DB:** designed the study, provided immunological attributes, supervised the project, interpreted the data, drafted the manuscript and accepted final version. **WH:** coordinated the uploading and annotation of datasets, contributed to the study design, provided immunological attributes, interpreted the data, drafted the manuscript and accepted final version.

Competing interests

No competing interests were disclosed.

Grant information

JD, SB, DR, DC, DB, WH received support from the Qatar Foundation. JR received support from Qatar National Research Fund (grant number: JSREP07-010-3-005).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to acknowledge all the investigators who decided to make their datasets publically available by sharing them in NCBI GEO.

References

Bedognetti D, Tomei S, Hendrickx W, *et al.*: **Toward the Identification of Genetic Determinants of Responsiveness to Cancer Immunotherapy.** In: *Developments in T Cell Based Cancer Immunotherapies.* Cancer Drug Discovery and Development 2196–9906. Springer, 2015. in press; 99–127.
Publisher Full Text

Bedognetti D, Tomei S, Hendrickx W, *et al.*: **Toward the Identification of Genetic Determinants of Responsiveness to Cancer Immunotherapy.** In: *Developments in T Cell Based Cancer Immunotherapies*, edited by Paolo A. Ascierto, David F. Stroncek, and Ena Wang, Cancer Drug Discovery and DevelopmentCancer Drug Discovery and Development. Springer International Publishing. 2015; 99–127.
Publisher Full Text

Desmedt C, Piette F, Loi S, *et al.*: **Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series.** *Clin Cancer Res.* 2007; **13**(11): 3207–3214.
PubMed Abstract | Publisher Full Text

Edge SB, Carolyn CC: **The American Joint Committee on Cancer: the 7th edition of the *AJCC cancer staging manual* and the future of TNM.** *Ann Surg Oncol.* 2010; **17**(6): 1471–74.
PubMed Abstract | Publisher Full Text

Ellis SA, Palmer MS, McMichael AJ: **Human trophoblast and the choriocarcinoma cell line BeWo express a truncated HLA Class I molecule.** *J Immunol.* 1990; **144**(2): 731–35.
PubMed Abstract

Fan C, Oh DS, Wessels L, *et al.*: **Concordance among gene-expression-based predictors for breast cancer.** *N Engl J Med.* 2006; **355**(6): 560–69.
PubMed Abstract | Publisher Full Text

Galon J, Angell HK, Bedognetti D, *et al.*: **The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures.** *Immunity.* 2013; **39**(1): 11–26.
PubMed Abstract | Publisher Full Text

Gentleman RC, Carey VJ, Bates DM, *et al.*: **Bioconductor: open software**

development for computational biology and bioinformatics. *Genome Biol.* 2004; **5**(10): R80.
PubMed Abstract | Publisher Full Text | Free Full Text

Hendrickx W, Simeone I, Anjum S, *et al.*: **Identification of Genetic Determinants of Breast Cancer Immune Phenotypes by Integrative Genome-Scale Analysis.** *Oncolmmunology.* 2017; **6**(2): e1253654.
Publisher Full Text

Herbst RS, Soria JC, Kowanetz M: **Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients.** *Nature.* 2014; **515**(7528): 563–67.
PubMed Abstract | Publisher Full Text | Free Full Text

Home - GEO - NCBI. Accessed November 28. 2016.
Reference Source

Hu Z, Fan C, Oh DS, *et al.*: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics.* 2006; **7**: 96.
PubMed Abstract | Publisher Full Text | Free Full Text

Ivshina AV, George J, Senko O, *et al.*: **Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer.** *Cancer Res.* 2006; **66**(21): 10292–10301.
PubMed Abstract | Publisher Full Text

Ji RR, Chasalow SD, Wang L, *et al.*: **An immune-active tumor microenvironment favors clinical response to ipilimumab.** *Cancer Immunol Immunother.* 2012; **61**(7): 1019–31.
PubMed Abstract | Publisher Full Text

Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics.* 2007; **8**(1): 118–27.
PubMed Abstract | Publisher Full Text

LeMaoult J, Zafaranloo K, Le Danff C, *et al.*: **HLA-G up-regulates ILT2, ILT3, ILT4, and KIR2DL4 in antigen presenting cells, NK cells, and T Cells.** *FASEB J.* 2005; **19**(6): 662–64.
PubMed Abstract | Publisher Full Text

Loi S, Haibe-Kains B, Desmedt C, *et al.*: **Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade.** *J Clin Oncol.* 2007; **25**(10): 1239–1246.
PubMed Abstract | Publisher Full Text

Loi S, Haibe-Kains B, Desmedt C, *et al.*: **Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen.** *BMC Genomics.* 2008; **9**: 239.
PubMed Abstract | Publisher Full Text | Free Full Text

Miller LD, Chou JA, Black MA, *et al.*: **Immunogenic Subtypes of Breast Cancer Delineated by Gene Classifiers of Immune Responsiveness.** *Cancer Immunol Res.* 2016; **4**(7): 600–10.
PubMed Abstract | Publisher Full Text | Free Full Text

Minn AJ, Gupta GP, Padua D, *et al.*: **Lung metastasis genes couple breast tumor size and metastatic spread.** *Proc Natl Acad Sci U S A.* 2007; **104**(16): 6740–6745.
PubMed Abstract | Publisher Full Text | Free Full Text

Minn AJ, Gupta GP, Siegel PM, *et al.*: **Genes that mediate breast cancer metastasis to lung.** *Nature.* 2005; **436**(7050): 518–524.
PubMed Abstract | Publisher Full Text | Free Full Text

Nagalla S, Chou JW, Willingham MC, *et al.*: **Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis.** *Genome Biol.* 2013; **14**(4): R34.
PubMed Abstract | Publisher Full Text | Free Full Text

Naji A, Menier C, Morandi F, *et al.*: **Binding of HLA-G to ITIM-Bearing Ig-like Transcript 2 Receptor Suppresses B Cell Responses.** *J Immunol.* 2014; **192**(4): 1536–46.
PubMed Abstract | Publisher Full Text

Parker JS, Mullins M, Cheang MC, *et al.*: **Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes.** *J Clin Oncol.* 2009; **27**(8): 1160–7.
PubMed Abstract | Publisher Full Text | Free Full Text

Pawitan Y, Bjöhle J, Amler L, *et al.*: **Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.** *Breast Cancer Res.* 2005; **7**(6): R953–964.
PubMed Abstract | Publisher Full Text | Free Full Text

Prat A, Parker JS, Karginova O, *et al.*: **Phenotypic and Molecular Characterization of the Claudin-Low Intrinsic Subtype of Breast Cancer.** *Breast Cancer Res.* 2010; **12**(5): R68.
PubMed Abstract | Publisher Full Text | Free Full Text

Ribas A, Robert C, Hodi FS, *et al.*: **Association of Response to Programmed Death Receptor 1 (PD-1) Blockade with Pembrolizumab (MK-3475) with an Interferon-Inflammatory Immune Gene Signature.** *J Clin Oncol Res.* suppl; abstr 3001. Oral Abstract Session, Developmental Therapeutics—Immunotherapy. 2015.
Reference Source

Rinchai D, Kewcharoenwong C, Kessler B, *et al.*: **Abundance of ADAM9 transcripts increases in the blood in response to tissue damage [version 1; referees: 3 approved with reservations].** *F1000Res.* 2015; **4**: 89.
Publisher Full Text

Rouas-Freiss N, Gonçalves RM, Menier C, *et al.*: **Direct evidence to support the role of HLA-G in protecting the fetus from maternal uterine natural killer cytolysis.** *Proc Natl Acad Sci U S A.* 1997; **94**(21): 11520–25.
PubMed Abstract | Publisher Full Text | Free Full Text

Schmidt M, Böhm D, von Törne C, *et al.*: **The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer.** *Cancer Res.* 2008; **68**(13): 5405–5413.
PubMed Abstract | Publisher Full Text

Sorlie T, Tibshirani R, Parker J, *et al.*: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci U S A.* 2003; **100**(14): 8418–23.
PubMed Abstract | Publisher Full Text | Free Full Text

Speake C, Presnell S, Domico K, *et al.*: **An interactive web application for the dissemination of human systems immunology data.** *J Transl Med.* 2015; **13**: 196.
PubMed Abstract | Publisher Full Text | Free Full Text

Swets M, König MH, Zaalberg A, *et al.*: **HLA-G and classical HLA class I expression in primary colorectal cancer and associated liver metastases.** *Hum Immunol.* 2016; **77**(9): 773–79.
PubMed Abstract | Publisher Full Text

Wang E, Bedognetti D, Marincola FM: **Prediction of response to anticancer immunotherapy using gene signatures.** *J Clin Oncol.* 2013; **31**(19): 2369–71.
PubMed Abstract | Publisher Full Text

Wang Y, Klijn JG, Zhang Y, *et al.*: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet.* 2005; **365**(9460): 671–679.
PubMed Abstract

Zeestraten EC, Reimers MS, Saadatmand S, *et al.*: **Combined analysis of HLA class I, HLA-E and HLA-G predicts prognosis in colon cancer patients.** *Br J Cancer.* 2014; **110**(2): 459–68.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhang Y, Sieuwerts AM, McGreevy M, *et al.*: **The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy.** *Breast Cancer Res Treat.* 2009; **116**(2): 303–309.
PubMed Abstract | Publisher Full Text

Zhou Y, Yau C, Gray JW, *et al.*: **Enhanced NF kappa B and AP-1 transcriptional activity associated with antiestrogen resistant breast cancer.** *BMC Cancer.* 2007; **7**: 59.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 26 September 2017

https://doi.org/10.5256/f1000research.11813.r24824

?

Christos Hatzis
Breast Medical Oncology, Yale School of Medicine, New Haven, CT, USA

Roelands *et al* present a web application that provides real-time analysis and visualization of a large compendium of microarray datasets generated from primary breast cancer biopsies. The uniqueness and distinct utility of this dataset is the focus on immune-related signatures. Before discussing potential shortcomings according to this reviewer's opinion, the authors should be commended for investing the energy and resources to develop and implement a focused public repository that would facilitate further analyses.

Although the overall objective behind this website and tools is laudable, I believe that the approach followed, as presented in the manuscript, has several potential shortcomings that would be worth addressing.

1. Regarding the protocols followed, I would echo the comments by the first reviewer and question the use of MAS5 and COMBAT, given that more robust algorithms are now available. This may be a key issue given the quantitative nature of many of the indices used. Also, it is not clear that this approach removes bias between the older Affymetrix platforms and the newer ones (PLUS2.0). On page 5 it is mentioned that "the 22,268 probe sets present in each of these platforms were included", but the PLUS 2.0 chips include twice as many. They probably meant to say "present in all platforms".

2. In terms of whether sufficient details were provided to allow replication, it would be great if the authors provided the complete set of scripts used to pre-process and normalize the data included in the analysis. This will allow transparent assessment of the methods and replication of the dataset if needed.

3. Related to the type of data included, I believe that the resource would be more valuable if additional clinically-relevant information was provided with each sample. For example, disease progression clearly depends on treatment and treatment information is missing entirely - we do not know whether patients even received any treatment at all. It would be useful to curate treatment information available in the original datasets. Also, it may be worth providing annotations for additional immune signatures (e.g. Rody *et al*, 2011[1]) and also of additional breast cancer

subtypes (e.g. Lehmann *et al*, 2016[2]). This will allow a more thorough assessment of the overlap between the various signature-defined subtypes. Finally, although there are not as many profiles from breast cancer metastasis available in GEO, it may be worth considering curating those and including them in the platform. It would be extremely interesting to know how the immune signatures differ between primary and metastatic samples, and whether the trends suggest immune evasion.

4. Finally, from a usability standpoint, I have a few suggestions outlined below:

   - Although the functionality may be available, I found it very difficult to filter samples based on a particular feature. For example, I was trying to do the HLA-G by ICR group analysis only for the ER-negative or basal cases, but could not find an easy way to do that.

   - It would be useful to provide tools for correlational analysis of two or more genes (e.g. co-expression patterns) and heat maps. The one gene at a time visualization is very restrictive.

Minor points:

1. Worth considering having another table after Table 1 that summarizes the immune subtype signatures and annotations with references. Intrinsic subtype signatures used can also be summarized in that table. That will also eliminate the need to explain the acronyms (e.g. Table 2 is marred by too many acronyms IMS, IBC, IDS, ICR that are not explained).

2. When running the HLA-G example as outlined in the manuscript, I was presented with 4 different HLA-G transcripts with no explanation as to what was the difference between them. Are these different probe sets? If so, that should probably be explained. Also, in that case, is there a recommendation as to which of these probe sets is the most representative (e.g. based on specificity, range, variance etc).

References

1. Rody A, Karn T, Liedtke C, Pusztai L, Ruckhaeberle E, Hanker L, Gaetje R, Solbach C, Ahr A, Metzler D, Schmidt M, Müller V, Holtrich U, Kaufmann M: A clinically relevant gene signature in triple negative and basal-like breast cancer.*Breast Cancer Res*. 2011; **13** (5): R97 PubMed Abstract | Publisher Full Text
2. Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, Moses HL, Sanders ME, Pietenpol JA: Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection.*PLoS One*. 2016; **11** (6): e0157368 PubMed Abstract | Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Jan 2018

Wouter Hendrickx, Sidra Medical and Research Center, Qatar

Major points:

- Regarding the protocols followed, I would echo the comments by the first reviewer and question the use of MAS5 and COMBAT, given that more robust algorithms are now available. This may be a key issue given the quantitative nature of many of the indices used. Also, it is not clear that this approach removes bias between the older Affymetrix platforms and the newer ones (PLUS2.0). On page 5 it is mentioned that "the 22,268 probe sets present in each of these platforms were included", but the PLUS 2.0 chips include twice as many. They probably meant to say "present in all platforms".

The primary goal of this data note was to share the transcriptomic and annotation data associated with this previous publication (in GXB defined as "the Nagalla 2013 reconstituted public dataset") and the breast cancer transcriptomic datasets from GEO that it constitutes of in a more interactive, comprehensible format to facilitate usage of the data and did not involve data processing. The selection of 22,268 probe sets, combination and normalization of the 13 datasets was also previously performed (Nagalla et al). For your second point, we indeed meant the probe sets that are present in all platforms, for this reasons probe sets that are exclusively present in Affymetrix U133 PLUS2.0 are not included in these 22,268 probe sets. We have revised this sentence in the manuscript accordingly.

- In terms of whether sufficient details were provided to allow replication, it would be great if the authors provided the complete set of scripts used to pre-process and normalize the data included in the analysis. This will allow transparent assessment of the methods and replication of the dataset if needed.

The uploads from individual GEO datasets into GXB did not involve any data processing. For the Nagalla et al this has been performed as part of the work described in Nagalla et al. This data note did involve clinical data harmonization performed in R, scripts have been made available on Github.

- Related to the type of data included, I believe that the resource would be more valuable if additional clinically-relevant information was provided with each sample. For example, disease progression clearly depends on treatment and treatment information is missing entirely - we do not know whether patients even received any treatment at all. It would be useful to curate treatment information available in the original datasets. Also, it may be worth providing annotations for additional immune signatures (e.g. Rody *et al*, 2011[1]) and also of additional breast cancer subtypes (e.g. Lehmann *et al*, 2016[2]). This will allow a more thorough assessment of the overlap between the various signature-defined subtypes. Finally, although there are not as many profiles from breast cancer metastasis available in GEO, it may be worth considering curating those and including them in the platform. It would be extremely interesting to know how the immune signatures differ between primary and metastatic samples, and whether the trends suggest immune evasion.

When uploading these GEO datasets into GXB, all available clinical data on GEO was also transferred to GXB. For some of the datasets (e.g. GSE4922), a treatment parameter is available which can be selected as a parameter using the overlay function in the bar plot generated in GXB. We completely agree that information on adjuvant treatment is essential for interpretation of the supplied survival data (Distant Metastasis Free Survival and Disease Free Survival). For proper survival analysis, we recommend users to download the csv file supplied which can be found under the "Downloads"-tab of each dataset and use available treatment information for stratification.
Addition of more gene signatures can indeed be valuable to compare immune based classifications as demonstrated in this data note for Immunological Constant of Rejection (ICR) classification and Immune Benefit Status (IBS)/ Immune Disposition Status (IDS). Similarly, uploading additional GEO datasets that contain matched primary and metastatic samples would be very interesting and allow for comparison of both immune microenvironments. Although not in the scope of the current data note, we will definitely take these suggestions into consideration for further development of the GXB breast cancer instance.

- Finally, from a usability standpoint, I have a few suggestions outlined below:

-Although the functionality may be available, I found it very difficult to filter samples based on a particular feature. For example, I was trying to do the HLA-G by ICR group analysis only for the ER-negative or basal cases, but could not find an easy way to do that.
It is indeed possible to visualize the expression of a gene of interest, for example HLA-G, in a specific subgroup of samples like all ER-negative or basal cases. The most convenient way to do this is to (1) select barplot, (2) set group by Molecular Subtype and (3) overlay the barplots with ICR group annotation. To make the figure easier to analyze, you can subsequently sort by ICR.

- It would be useful to provide tools for correlational analysis of two or more genes (e.g. co-expression patterns) and heat maps. The one gene at a time visualization is very restrictive.
We realize that gene per gene visualization has its limitations and analysis of groups coordinately expressed genes provides more biological significance. For this reason, the GXB development team has already created a module analysis tool (MAT) that analyses expression data to find pre-defined groupings of co-expressed genes (modules) to obtain a molecular fingerprint of gene expression for each individual sample in your dataset. MAT has already been applied to many datasets with samples of various immune related diseases and we are considering to also use this platform for cancer transcriptomic datasets. As the desired number of group comparisons in breast cancer (i.e., Molecular Subtypes, Immunologic classifications, Pathology T Stage) is higher compared to the existing group comparisons, some work is required before we can implement these dataset in this tool.

Minor points:

- Worth considering having another table after Table 1 that summarizes the immune subtype signatures and annotations with references. Intrinsic subtype signatures used can also be summarized in that table. That will also eliminate the need to explain the acronyms (e.g. Table 2 is marred by too many acronyms IMS, IBC, IDS, ICR that are not explained).

We have added the suggested table to the manuscript. This is both useful when reading the manuscript as well as for the use of this breast cancer GXB application, thank you for this suggestion.

- When running the HLA-G example as outlined in the manuscript, I was presented with 4 different HLA-G transcripts with no explanation as to what was the difference between them. Are these different probe sets? If so, that should probably be explained. Also, in that case, is there a recommendation as to which of these probe sets is the most representative (e.g. based on specificity, range, variance etc).

These different HLA-G transcripts are indeed different probe sets. If you are interested to find and select a specific probe ID, you can select "Show Probe ID" under "Tools". This GXB data portal does not give an explanation on the difference between these probes. For this information we would like to refer to the Affymetrix website. For the HLA-G example, we just took a single probe at random: 211528_X_AT. We added this information in version 2 of the manuscript.

Competing Interests: authors reply, no competing intrest

Reviewer Report 24 July 2017

https://doi.org/10.5256/f1000research.11813.r24330

?

Benjamin Haibe-Kains
Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada

In this manuscript, Roelands *et al.* describe the implementation of the gene expression browser (GXB), a database and web-application integrating 13 breast cancer datasets containing gene expression and clinical data for a total of almost 2000 patients. The authors focused their case studies on the investigation of immune-related genes and subtypes and made their curated data publicly available. I think GXB is a welcome contribution to the breast cancer research community and its public availability will help other scientists to leverage this valuable collection of datasets. However, I have several concerns that need to be addressed before indexing, as listed below.

The manuscript is well written and easy to understand.

Major comments

In Table 1 and on the front page of the web-application, Princess Margaret Cancer Centre is listed in the sample set name even though these data have not been generated in this institution, which is misleading.

When selecting "Breast Cancer", which I assumed contained the whole database, many datasets got

filtered out, which is confusing.

I tried AURKA and selected GSE9195 as dataset. Then I played a bit with the barplot to overlay different kinds of information. I added DMFS 10Y (categorical) and sorted the patients based on this but there were some patients with very low DMFS at the beginning and the end of the plot, which is confusing.

Since the authors are dealing with survival data, I was expecting survival statistics and/or survival curves but I did not see any in GXB or the manuscript. Figure 1 is somehow misleading as these plots do not seem to be part of the web-application.

The authors decided to limit their database to Affymetrix data. In that case, I suggest the authors reprocess all the CEL files consistently using (f)RMA and up-to-date chip description files (CDF), such as BrainArray CDFs. This would increase the consistency across all the datasets.

Given the authors' large collection of datasets, mining each dataset separately is cumbersome. Implementing meta-analysis pipelines would help draw an overall conclusion before digging into dataset-specific results (e.g., Figure 3). This is what the authors seem to have done with the "Nagalla" dataset that is a compendium of all the datasets reprocessed using MAS5 and further corrected with ComBat. It is not clear why the authors used MAS% to normalize the data in this case vs (f)RMA for some other datasets. Please clarify. Why not recommending users to first explore Nagallia before going after each dataset separately to detect trends that may be (in)consistent with the majority of the datasets?

There have been many molecular subtyping schemes published, including the Integrative Subtypes (IntClust; Curtis et al, Nature 2012), the Subtype Classification Models (SCMGENE, SCMOD1, SCMOD2; Haibe-Kains et al, JNCI 2012) and the Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype (AIMS; Paquet et al, JNCI 2015). The authors should discuss the rationale for their choice of relying solely on PAM50.

Minor comments

Table 2 would be better represented as a barplot or a similar figure.

GXB would benefit from the inclusion of normal samples (from healthy patients or adjacent normal samples). METABRIC, TCGA and GTEx are relevant data sources for such gene expression profiles.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 Sep 2017

Wouter Hendrickx, Sidra Medical and Research Center, Qatar

Answers are in Italic and bold

Major comments

In Table 1 and on the front page of the web-application, Princess Margaret Cancer Centre is listed in the sample set name even though these data have not been generated in this institution, which is misleading.
Names of datasets have been changed from "Princess Margaret Cancer Centre dataset - GSE6532.GPL96" and "Princess Margaret Cancer Centre dataset - GSE6532.GPL97", to "John Radcliff Hospital (OXFU, OXFT) dataset- GSE6532.GPL96" and "John Radcliff Hospital (OXFU, OXFT) dataset- GSE6532.GPL97" respectively. "Princess Margaret Cancer dataset- GSE6532.GPL570" was changed to "Guys hospital (GUYT) dataset- GSE6532.GPL570". Tables and figures throughout the manuscript have been revised accordingly.

When selecting "Breast Cancer", which I assumed contained the whole database, many datasets got filtered out, which is confusing.
Thank you for noticing this. Only a single disease type can be assigned per dataset. We can change ER+ and LN- datasets to disease type: Breast Cancer , effectively disabling the user to filter datasets based on ER+ and LN status. We preferred changing the name of Breast cancer (general) to "mixed breast cancer types" for clarity.

I tried AURKA and selected GSE9195 as dataset. Then I played a bit with the barplot to overlay different kinds of information. I added DMFS 10Y (categorical) and sorted the patients based on this but there were some patients with very low DMFS at the beginning and the end of the plot, which is confusing.
We were able to reproduce the barplot you described. By sorting by the categorical variable DMFS 10Y EVENT, you only sort based on this variable (DistantMetastasisFree or DistantMetastasis). The continuous counterpart of this variable (DMFS 10Y TIME) is not taken into account when you perform the sorting, explaining your observation. An alternative is to sort based on the continuous variable DMFS 10Y TIME, though this sorting results in a plot that mixes DistantMetastasisFree and DistantMetastasis categories. Unfortunately, subsequent sorting based on 2 variables is not possible using GXB. We have suggested this improvement to the GXB developing team.

Since the authors are dealing with survival data, I was expecting survival statistics and/or survival curves but I did not see any in GXB or the manuscript. Figure 1 is somehow misleading as these plots do not seem to be part of the web-application.
Survival statistics are not yet part of the GXB web-application. To prevent any misconception, we have added an extra sentence in the legend of figure 1: "This figure is for explanatory purposes only and does not serve as a demonstration of the GXB web

application."

The authors decided to limit their database to Affymetrix data. In that case, I suggest the authors reprocess all the CEL files consistently using (f)RMA and up-to-date chip description files (CDF), such as BrainArray CDFs. This would increase the consistency across all the datasets.
The purpose of this data note is to share existing data from GEO in a more comprehensible format and does not involve any data processing. We had to make an exception for dataset GSE9195.GPL570, since only raw format was available on GEO.

Given the authors' large collection of datasets, mining each dataset separately is cumbersome. Implementing meta-analysis pipelines would help draw an overall conclusion before digging into dataset-specific results (e.g., Figure 3). This is what the authors seem to have done with the "Nagalla" dataset that is a compendium of all the datasets reprocessed using MAS5 and further corrected with ComBat. It is not clear why the authors used MAS5 to normalize the data in this case vs (f)RMA for some other datasets. Please clarify.
Normalization of the complete Nagalla cohort was previously performed as described in Nagalla et al. 2013 and Miller et al. 2016 To make this data available as published in this article, we used the same data file to upload to GXB. Normalization is different between datasets because the data was transferred straight from GEO and different contributors use different methods depending on the times it was performed and the need of their specific work.

Why not recommending users to first explore Nagallia before going after each dataset separately to detect trends that may be (in)consistent with the majority of the datasets?
In this dataset collection, it is indeed possible to first explore the Nagalla dataset, which is a dataset that includes all samples from the collection. This advantage is specific for this breast cancer compendium of datasets, so your suggestion is a very good strategy to explore these datasets. We have added this recommendation in the text of the Dataset Demonstration section of the data note.

There have been many molecular subtyping schemes published, including the Integrative Subtypes (IntClust; Curtis et al, Nature 2012), the Subtype Classification Models (SCMGENE, SCMOD1, SCMOD2; Haibe-Kains et al, JNCI 2012) and the Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype (AIMS; Paquet et al, JNCI 2015). The authors should discuss the rationale for their choice of relying solely on PAM50.
Categorization using the PAM50 molecular subtyping was performed as part of the research performed by Nagalla et al. 2013 and Miller et al. 2016. These annotated datasets were uploaded in GXB as described in this data note. The choice to rely on this 50-gene classifier was dependent on its high prognostic and predictive value of this subtyping method in combination with the high clinical applicability of PAM50 testing.

Minor comments

Table 2 would be better represented as a barplot or a similar figure.
For the purpose of listing which clinical variables are available in how many datasets from the collection, we found a table the more appropriate option.

GXB would benefit from the inclusion of normal samples (from healthy patients or adjacent normal

samples). METABRIC, TCGA and GTEx are relevant data sources for such gene expression profiles.
Comparing healthy control tissue with cancerous tissue on gene expression level indeed represents an interesting aspect that can be explored using the GXB browser. At this moment, different types of dataset collections have already been uploaded to GXB, including case versus control comparisons. For example, a dataset with gene expression of whole blood samples from both lung cancer patients and healthy controls and a dataset with head and neck cancer and cervical cancer tissue samples, matched with site-matched normal epithelial samples are currently available. The TCGA, METABRIC dataset and datasets in the GTEx portal are valuable resources that can be used to upload new datasets into GXB. We will definitely take this suggestion into consideration for our new dataset collection to upload into GXB.

Discussed changes will be applied in version 2 of the manuscript.

Competing Interests: none