Annotated Compendium of 102 Breast Cancer Gene-Expression Datasets

Ifeanyichukwu O. Nwosu
Department of Biology
Brigham Young University
Provo, UT, USA
ifeanyichukwu_nwosu@yahoo.com

Daniel D. Tabler
Department of Biology
Brigham Young University
Provo, UT, USA
daniel.d.tabler@gmail.com

Greg Chipman
Medical Oncology and Hematology
Utah Cancer Specialists
Salt Lake City, UT, USA
gchipman@utahcancer.com

Stephen R. Piccolo*
Department of Biology
Brigham Young University
Provo, UT, USA
stephen_piccolo@byu.edu

*Please address correspondence to S.R.P.

Abstract

Transcriptomic data from breast-cancer patients are widely available in public repositories. However, before a researcher can perform statistical inferences or make biological interpretations from such data, they must find relevant datasets, download the data, and perform quality checks. In many cases, it is also useful to normalize and standardize the data for consistency and to use updated genome annotations. Additionally, researchers need to parse and interpret metadata: clinical and demographic characteristics of patients. Each of these steps requires computational and/or biomedical expertise, thus imposing a barrier to reuse for many researchers. We have identified and curated 102 publicly available, breast-cancer datasets representing 17,151 patients. We created a reproducible, computational pipeline to download the data, perform quality checks, renormalize the raw gene-expression measurements (when available), assign gene identifiers from multiple databases, and annotate the metadata against the National Cancer Institute Thesaurus, thus making it easier to infer semantic meaning and compare insights across datasets. We have made the curated data and pipeline freely available for other researchers to use. Having these resources in one place promises to accelerate breast-cancer research, enabling researchers to address diverse types of questions, using data from a variety of patient populations and study contexts.

Background

Among women worldwide, breast cancer is the most frequently diagnosed cancer type and the leading cause of cancer death; it accounts for 23% of total cases and 14% of deaths[1], with over 2 million new cases diagnosed and 600,000 deaths in 2020[2]. This is a huge public health burden. Breast cancers exhibit considerable diversity and complexity[3] and are driven by a combination of factors, including genomic material inherited from parents and an accumulation of somatic events within individual cells; these events include mutations, chromosomal rearrangements, gene amplifications, and deletions that eventually lead to evasion of apoptosis and unregulated cellular proliferation[4]. While earlier diagnoses and treatment advances have improved survival and lowered mortality rates[5,6], researchers continue to seek to understand the molecular underpinnings of this disease, identify subtypes of the disease, and explore ways to tailor treatments for individual patients[3,7].

Microarrays and RNA sequencing have been used in hundreds of thousands of gene-expression studies (spanning many biomedical contexts)[8], helping to identify differentially expressed genes, define disease subtypes, inform treatment and patient-management strategies, predict patient outcomes, etc.[9] Much of the data from these studies has been deposited in public databases such as Gene Expression Omnibus (GEO)[8,10] and ArrayExpress[11]. These databases offer opportunities for reuse, including studies that combine insights from multiple datasets. However, making data accessible is not enough. To be most useful for addressing biomedical questions, the data must also be findable, interoperable, and reusable[12]. First, researchers must locate the data. Gene-expression datasets are available in diverse repositories. These repositories support keyword searches; however, our experience is that these searches miss many relevant datasets and require considerable manual effort to filter through search results. Although in some cases researchers can reuse data exactly in the form they are found in public repositories, data depositors use a heterogeneous mix of methods for verifying data quality, normalizing expression measurements, and then filtering, summarizing, and annotating these measurements. Sometimes steps are skipped or may be completed using invalid methods. Furthermore, genome annotations are regularly updated, so datasets that were annotated years ago may have stale annotations. Fortunately, gene-expression data are frequently provided in raw form, so researchers can perform these steps on their own, thus resulting in datasets that are more consistent and of higher quality.

A key challenge to reusing publicly available datasets is working with metadata. Nearly all gene-expression data from human patients have clinical and/or demographic data associated with them. However, researchers use different terms, abbreviations, and spellings (including misspellings) to describe the same clinical and demographic phenomena. For example, we have seen variable names as diverse as "drfs_even_time_years", "distal_recurrence_time", "distant_recurrence_free_survival_mo", "dmfs_time", "t_tdm", and "time" to describe the length of time a patient survived without a metastatic event after diagnosis. Some variable names are short and/or vague (e.g., "cell_type", "chemo", "nhg", "pt"). Sometimes measurement units are included in names of numeric variables (e.g., "age_years", "size_mm", "mfsdel_month"), but other times they are not. Some variable names incorrectly or inadequately reflect the semantics of those variables. For example, "ethnicity" sometimes refers to population ancestral groups (often referred to as "race"); "gender" sometimes refers to a person's biological sex assigned at

birth. Clarifying these semantic inconsistencies is especially important when researchers wish to identify datasets that have particular metadata variables and/or to combine multiple datasets.

Formatting differences also cause challenges. In GEO, data are provided in the "series matrix" and SOFT formats[13], whereas ArrayExpress uses standardized formats like MAGE-TAB[14]. Software tools are available to download and convert the data into formats and data structures that are not specific to gene-expression data[15,16]; however, other challenges remain. Researchers sometimes store multiple values in a single cell (e.g., "female|tamoxifen|58"), use key/value pairs within cells (e.g., "sex: female", "treatment=tamoxifen", "age=58 years"), or include missing values (using descriptors such as "NA", "n/a", or "?")[17]. Addressing these problems requires time and expertise. To ensure semantic consistency, biomedical domain experts must manually review, interpret, and standardize metadata. To deal with formatting problems, researchers often must write custom computer code. These barriers have caused these valuable data resources to be underutilized and have slowed the research enterprise.

We have addressed these challenges for breast cancer by finding and curating 102 gene-expression datasets from the public domain. These datasets represent 17,151 patient samples. For each dataset, we performed quality-control checks, uniformly processed the expression data, mapped the expression data to multiple gene-reference databases, and standardized the metadata variables against the National Cancer Institute Thesaurus (NCIT)[18], a popular reference standard with unique codes and synonyms for biomedical terms. Additionally, a practicing medical oncologist reviewed ambiguous term mappings. We have deposited these datasets, along with the code we used to process them, in an open-access repository, thus making it easy for other researchers to find, access, integrate, and reuse the data, as well as to reproduce our process. This repository is not tied to any programming language—the data can be imported using a variety of tools, including Microsoft Excel and software libraries available for commonly used programming languages.

Methods

Data finding

Initially, we searched for datasets within GEO and ArrayExpress. Our initial research interest was the interaction between race and HER2 amplification status among breast-cancer patients, and this guided our choice of search parameters. For GEO, our search parameters were ("Breast Cancer"[All Fields] AND "Her2"[All Fields] AND "race"[All Fields]). We selected the study types as “Expression profiling by array” or “Expression profiling by high-throughput sequencing”. We individually screened the available datasets for relevance and duplicates. In ArrayExpress, our search parameters were "Breast Cancer" AND "Her2" AND "race". Next, we searched within PubMed for articles that matched the following criteria: "Breast Cancer" AND "Her2" AND "race". These criteria resulted in a limited number of datasets, so we expanded our searches in PubMed to breast cancer more broadly. We also searched for datasets using the Google search engine. As part of this process, we identified two existing collections that included breast-tumor expression data: MetaGxData[19] and curatedBreastData[20]. We used these collections as resources for identifying relevant breast-cancer datasets. Finally, we examined cancer-genomics repositories for additional datasets.

The International Cancer Genome Consortium (ICGC) data portal[21,22] contained five breast-cancer datasets, two of which provided gene-expression data; raw data were available for one of them. cBioPortal for Cancer Genomics[23] provided a copy of gene-expression data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project[24]. The Cancer Genome Atlas (TCGA) provided gene-expression data for many cancer types, including breast cancer[25]. We used a reprocessed version of the data stored in GEO[26].

We found a mix of different data-generation platforms, including Affymetrix, Agilent and Illumina microarrays, custom microarrays, as well as Illumina-based RNA sequencing. Because the Affymetrix platforms were the most common—and well-recognized normalization and annotation packages could be applied to these platforms—we focused on these. We also included datasets that had been processed using RNA sequencing technology because of its advantages over microarray technologies[27]. With one exception, we excluded datasets that used Agilent microarrays, Illumina microarrays, or uncommon microarray platforms to reduce heterogeneity. The METABRIC project used Illumina microarrays; we included this dataset in our compendium because it is large (1980 samples) and provides extensive metadata. We excluded Affymetrix datasets for which raw data were unavailable or that were from cell lines or single cells. Raw data were not publicly available for the METABRIC dataset or the RNA-sequencing datasets; we used the processed versions of these datasets. This resulted in 102 datasets with a total of 17,151 samples. Most of the samples are from breast tumors; however, a small number are from normal breast tissue as indicated in the metadata.

Data curation

We automated all steps of the curation process using custom computer scripts. Some scripts used the bash interpreter (Linux operating system); other scripts used the R statistical software[28]. We configured the scripts to be executed within Docker containers so that our curation process is reproducible and portable across computer systems[29]. We obtained the GEO data using the GEOquery[15] package. For the other sources, our scripts downloaded the data directly from the repositories. The data comprised both gene-expression data and metadata, which we processed to achieve consistency across all datasets.

Our workflow was divided into six major steps described below.

In step one, we downloaded the metadata for Affymetrix datasets, cleaned up column names by removing non-informative suffixes (e.g "_ch1") and removing columns that we determined would not add value in downstream analyses. For example, some columns had only one unique value (e.g., "extraction protocol", "last_update_date"), while others had all unique values such (e.g., "sample_name"). Because this process was subjective, and other researchers might find these columns useful, we retained the unfiltered metadata as a separate part of the outputs. Upon reviewing the metadata, we made additional customizations on a case-by-case basis. When a single dataset was processed using more than one platform, we separated them into multiple data files, according to their respective platforms. Finally, we summarized the metadata variables from each dataset. These summaries indicated the unique values for each variable, the number of unique values, and the minimum, maximum and mean (for numeric variables).

In step two, we downloaded raw CEL files for the Affymetrix datasets and normalized them using the SCAN.UPC[30] package, yielding zero-centered, log2-transformed intensities. This normalization step made use of custom chip definition files (CDF) for Affymetrix GeneChips using updated probe-set definitions from the Brainarray Microarray consortium[31]. Custom CDFs are necessary because the manufacturer's probe-set definitions have become outdated and are sometimes incorrectly assigned to genes or to multiple genes. Some GEO datasets had a mixture of Affymetrix and Agilent data. Our scripts selected the Affymetrix samples only.

In step three, we performed quality-control measures on the expression data. We used the DoppelgangR[32] package to identify Affymetrix samples that may have been duplicated. Duplicates can occur when samples are inadvertently mislabeled or when studies include samples from previously published datasets. After manually reviewing the sample pairs that had been flagged as duplicates by *DoppelgangR*, we determined that an expression-similarity threshold of 0.99 may be appropriate for excluding samples; this score errs on the side of only excluding samples that were most likely to be duplicates. The sample pairs that exceed this threshold are indicated in our output files. However, we retained expression data for all samples and leave it to data users to choose if and when to exclude these samples.

Next, we evaluated each dataset using the IQRray[33] method, which identifies poor quality arrays by ranking all probe intensities in an array, computing the average rank for each probe set and returning a quality score for each sample in the dataset. Based on platform-specific quality thresholds from the authors' prior analysis, we excluded samples that did not meet these thresholds. The IQRray method is specific to samples processed on Affymetrix platforms.

In steps four and five, we processed the metadata and expression data, respectively, for non-Affymetrix datasets. We created a separate script for each dataset because they were heterogeneous. For instance, the files were downloaded from different sources, and the variables we excluded differed widely in each of these datasets. We also had to separate the TCGA dataset into tumor-only and normal-only datasets. Otherwise, the tasks performed in these steps paralleled what we did in steps one and two.

In step six, we standardized the names of individual samples, ensuring consistency with capitalization and removing suffixes. We also removed the control probeset values from the Affymetrix datasets. Next we filtered out poor quality arrays identified by the IQRray method. To ensure consistency, we matched the samples in the expression data to samples in the metadata; thus the samples that were filtered by the IQRray method were also removed from the metadata. We also excluded any gene-expression samples for which metadata were missing. Finally, because some datasets used Entrez Gene identifiers while others used HUGO Gene Nomenclature Committee (HGNC)[34] identifiers, we used the biomaRt[35,36] package from Bioconductor[37] to map gene identifiers across Entrez Gene, HGNC, and Ensembl[38]. When a single gene symbol was associated with multiple Ensembl identifiers, we retained genes from the main chromosomes and excluded gene identifiers from haplotypic regions. While this removed most duplicated genes, some remained. Through manual inspection, we found that the first identifier (sorted alphanumerically) was the most appropriate identifier, so we retained those. We determined that these identifiers were most

appropriate by looking up each gene in The Human Gene Database[39]. This database has a section for each gene named "External Ids" where different identifiers are listed. We found that the first identifiers matched those listed in this section. We also excluded data for which any gene identifier was missing.

Thesaurus mapping

We summarized the metadata variables into a single spreadsheet. For each variable, we used the variable name and data-value summaries to guide our interpretations of the semantic meaning. We manually mapped the metadata variables to NCIT standardized terms, using version:22.11d (release date: 2022-11-28). This was accessed via the NCI Term browser[40]. NCIT terms typically have synonyms and parent-child relationships, which proved useful in identifying the most appropriate term(s) for each variable. For this process, there were two curators (authors ION and SRP). The curators worked independently to match each variable name and value to respective NCIT terms. At the end of this process, the curators discussed the mappings jointly and addressed differences in interpretation. When we were uncertain about the semantic meaning, we searched the journal article associated with the dataset (if there was one) to see if any relevant explanations were available to decipher the meaning. These searches usually focused on the Methods sections or supplementary materials. For some datasets, we found relevant information in the supplementary files section on GEO. For variables that remained unclear, a practicing medical oncologist (author GC) provided suggestions. We excluded any variables that still remained unclear.

The spreadsheet has two columns: "NCIT_Term_field_name" and "NCIT_Term_values," which contain the NCIT preferred name(s) and unique identifier(s) that we selected for each metadata variable name and its associated values, respectively. The former typically reflects the biomedical context of the variable; the latter typically indicates the manner in which the values were represented (e.g., units of measurement). In many instances, a single NCIT term was insufficient to convey the semantic meaning, so we used multiple NCIT terms. For example, for the variable *pr_status_by_ihc*, we used NCIT terms "Progesterone Receptor Status" and "Immunohistochemical Test Result". In most cases, we retained the original metadata variable names; however, in a few cases, the names were particularly generic (e.g., *title* or *description*), so we renamed them. In other cases, a single column had multiple values per cell, so we split the column and named the new columns appropriately.

Availability of data and materials

All gene expression data, metadata, and computer scripts generated during the current study are in a publicly accessible Open Science Framework (OSF)[41] repository (accessible from https://osf.io/eky3p). The list of datasets is summarized in Table 1.

Discussion

Huge sums are budgeted each year by the National Institutes of Health and other grant-awarding agencies to generate gene-expression data and accompanying metadata. Additionally, researchers put much time and effort into generating the original data. Placing the data in the public domain is invaluable for research, but that alone is not enough. To move the field forward, we need compendia of datasets that follow common themes. By reprocessing, curating, and sharing these breast-cancer datasets, we are able to save time that other researchers would have used in completing the same steps, thus reducing duplicated efforts. These efforts make the data more accessible to researchers who lack the expertise to complete these reprocessing and curation steps. Accordingly, we have sought to increase the benefit to the public of prior investments.

A major takeaway from this project is the number of samples that we have compiled. This resource includes 102 datasets, representing 17,151 patient samples. While groups like TCGA[25] have generated large data collections (11,000 tumor samples across 33 cancer types), we have collated data from multiple sources and have focused solely on breast cancer. Projects like *recount3*[42] focus on RNA sequencing data for a wide range of biomedical conditions. Out of necessity, that project provides relatively little metadata, which is uncurated. We have taken considerable time to clean and manually curate the metadata variables in the datasets we reprocessed because it is difficult for many researchers to decipher the semantic meaning of a variable solely by its name or contents. Human review is essential (and time consuming). We envision that researchers will search our compendium for datasets associated with given NCIT term(s) and perform diverse types of meta-analyses using those data. Supplementary File 1 contains these NCIT term mappings.

To ensure reproducibility, we tested our scripts using an iterative process. The scripts have been configured to be executed within Docker containers so that they can be re-executed more easily by others on different operating systems and obtain the same outputs. Finally, our output files are simple text files and thus are programming-language agnostic. Other research groups have created valuable packages like the MetaGxBreast[19] and curatedBreastData[20] packages. These packages are accessible using the R programming language and are implemented within the Bioconductor framework, which is used heavily in the biomedical community[43], but the outputs from our scripts can be accessed on any platform.

Although our compendium consists of many datasets relevant to a common biomedical theme, care must be taken when combining datasets. Systematic differences between data-generation platforms, laboratory equipment, and environmental conditions can bias analytical results[44]. We did not adjust for these differences but rather leave it to consumers of the data to identify scenarios and techniques for adjusting the data. Analyses that combine *inferences* across datasets—rather than combining data directly—may be less prone to such biases.

We mapped the metadata to NCIT terms but did not standardize the field names or values in the text files. Therefore, when combining datasets, researchers may wish to perform this step. For instance, one dataset might have a column that indicates estrogen-receptor status using values of "1" or "0", whereas another dataset might represent the same information

using values of "positive" and "negative" or "P" and "N". To provide insight into the frequency with which biomedical concepts are represented in the metadata, we summarized the NCIT terms mapped to the metadata fields. Summaries for the NCIT terms used to describe field names can be found in Supplementary Files 2 and 3. Some of the most commonly observed terms for field names were Estrogen Receptor Status (n = 73), Event-Free Survival (n = 63), Disease Stage Qualifier (n = 52), Progesterone Receptor Status (n = 46) and HER2/Neu Status (n = 43). Thus, even though some variables are quite common, these results illustrate that breast-cancer studies differ considerably in the metadata collected. Some terms like sex (n = 15) and race (n = 11) were less common than we expected. Even though breast cancer largely affects females, approximately 1% of cases affect males[45]; recording research participants' sex would make it possible to learn more about differences between male and female cases. Collecting race information may make it possible to shed more light on mechanisms that drive differential responses to treatments across racial groups[46]. Summaries for the NCIT terms used to describe data values can be found in Supplementary Files 4 and 5. Across all metadata fields, the most commonly observed value was “binary” (n = 138). Researchers used either 0 or 1 to represent many variables, including estrogen-receptor status, tumor size and lymph node status. In other cases, terms like Estrogen Receptor Positive and Estrogen Receptor Negative were used. In these cases the values can be more easily interpreted, whereas binary values must be interpreted in combination with field names. In future studies it may be beneficial to use more explicit terms to describe these values.

The large number of unique field names and values indicates the huge diversity in data that researchers collect. However, only 23% of terms (25% for values) occurred in 5 or more datasets. With this low concordance across datasets, it may be difficult to draw broad conclusions across the datasets. In future studies, it would be beneficial to collect a common set of variables across studies (although researchers might collect additional variables that are specific to each study).

In addition to aiding breast-cancer researchers in identifying datasets and samples associated with research topics of interest, biomedical-semantics researchers can use the ontology mappings for developing and testing algorithms that automate the ontology-mapping process. Although we were careful to ensure consistency when annotating the metadata variables, this process is subjective in nature. In many cases, several different NCIT terms could have been used to represent a given biomedical concept. In nearly every case, more generic (parent) or specific (child) terms could have been used as alternatives. The NCI Term Browser provides mappings between terms and links to other ontologies, thus making it possible for researchers to use our annotations as a starting point to find alternative terms.

Conclusions

This standardized resource has the potential to reduce time that breast-cancer researchers spend on finding and preparing publicly available gene-expression data and allow them to focus more on answering biomedical questions. We have previously[9] surveyed the literature on types of analyses that have been commonly performed using gene-expression data from breast-cancer patients. Examples include predicting patient outcomes, defining breast-

cancer subtypes, and identifying differentially expressed genes. We believe this resource will help in making these types of studies easier and to have more statistical power.

Table 1: Gene-expression data sources. This table summarizes all gene-expression datasets that we found in the public domain related to breast cancer.

Author Name	Pubmed ID	Data Source	Number of samples*	Platform
Zhou et al.[47,48]	17407600, 18631401	GSE7378	52 (54)	Affymetrix GeneChip HT-HG_U133A Early Access Array
Yau et al.[48,49]	17850661, 18631401	GSE8193	45 (47)	Affymetrix GeneChip HT-HG_U133A Early Access Array
Chin et al.[50]	17157792	E-TABM-158	129 (130)	Affymetrix GeneChip HT-HG_U133A Early Access Array
Knudsen et al.[51]	22134623	GSE33692	45	Affymetrix Human Exon 1.0 ST Array [transcript (gene) version]
Tofigh et al.[52]	25284793	GSE58644	321	Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]
Winslow et al.	No associated publication found	GSE59772	9	Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]
Lehmann et al.[53]	27310713	GSE81838	20	Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]
Rebollar-Vega et al.	No associated publication found	GSE86374	159	Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]
Gregory et al.[54]	31248446	GSE118432	41 (42)	Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]
Chang et al.[55]	15718313	GSE6434	19 (24)	Affymetrix Human Genome U95 Version 2 Array
Lee et al.[56]	22751464	GSE41194	49 (50)	Affymetrix Human Genome U95 Version 2 Array
Lee et al.[56]	22751464	GSE41196	11 (14)	Affymetrix Human Genome U95 Version 2 Array

Lee et al.[56]	22751464	GSE41197	16 (18)	Affymetrix Human Genome U95 Version 2 Array
Farmer et al.[57]	15897907	GSE1561	49	Affymetrix Human Genome U133A Array
Wang et al.[58]	15721472	GSE2034	286	Affymetrix Human Genome U133A Array
Minn et al.[59,60]	16049480, 23982961	GSE2603	121	Affymetrix Human Genome U133A Array
Sotiriou et al.[61]	16478745	GSE2990	189	Affymetrix Human Genome U133A Array
Karn et al.[62–67]	17317819, 18058224, 19272155, 19455418, 21203899, 25274406	GSE4611	216 (218)	Affymetrix Human Genome U133A Array
Minn et al.[68]	17420468	GSE5327	28 (58)	Affymetrix Human Genome U133A Array
Boersma et al.[69–71]	17999412, 19225562, 20978357	GSE5847	95	Affymetrix Human Genome U133A Array
Desmedt et al.[72,73]	17545524, 25788628	GSE7390	198	Affymetrix Human Genome U133A Array
Tripathi et al.[74]	18058819	GSE9574	29	Affymetrix Human Genome U133A Array
Schmidt et al.[75–78]	18593943, 20500820, 25485508, 33003293	GSE11121	200	Affymetrix Human Genome U133A Array
Zhang et al.[79]	18821012	GSE12093	134 (136)	Affymetrix Human Genome U133A Array
Zhang et al.[80,81]	19573813, 25670079	GSE14018	36	Affymetrix Human Genome U133A Array
Emery et al.[82]	19700746	GSE16873	40	Affymetrix Human Genome U133A Array

Symanns et al.[83]	20697068	GSE17705	296 (298)	Affymetrix Human Genome U133A Array
Miller et al.[84,85]	20646288, 20697427	GSE20181	147 (176)	Affymetrix Human Genome U133A Array
Shi et al.[86,87]	20064235, 20676074	GSE20194	251 (278)	Affymetrix Human Genome U133A Array
Tabchy et al.[88,89]	20829329, 23185353	GSE20271	156 (178)	Affymetrix Human Genome U133A Array
Graham et al.[90]	20197764	GSE20437	42	Affymetrix Human Genome U133A Array
Graham et al.[91]	21059815	GSE21947	30	Affymetrix Human Genome U133A Array
Iwamoto et al.[89,92]	21191116, 23185353	GSE22093	97 (103)	Affymetrix Human Genome U133A Array
Iwamoto et al.[92]	21191116	GSE23988	56 (61)	Affymetrix Human Genome U133A Array
Creighton et al.[93]	21750966	GSE24185	91 (103)	Affymetrix Human Genome U133A Array
Hatzis et al.[94,95]	21558518, 36293478	GSE25055	295 (310)	Affymetrix Human Genome U133A Array
Hatzis et al.[94–96]	21558518, 24337596, 36293478	GSE25065	184 (198)	Affymetrix Human Genome U133A Array
Karn et al.[97–100]	21978456, 22220191, 21741824, 26484129,	GSE31519	67	Affymetrix Human Genome U133A Array
Bianchini et al.[101,102]	20805453, 22235097	GSE32518	73 (74)	Affymetrix Human Genome U133A Array
Nagalla et al.[103]	23618380	GSE45255	135 (139)	Affymetrix Human Genome U133A Array
Rody et al.[67,104–107]	18340528, 19054665, 23135572, 24737166	GSE46184	74	Affymetrix Human Genome U133A Array

	25274406,			
Casey et al[108]	18373191	GSE10797	25 (66)	Affymetrix Human Genome U133A 2.0 Array
Reyngold et al.[109,110]	25083786, 28813674	GSE57968	71 (72)	Affymetrix Human Genome U133A 2.0 Array
Lu et al.[111]	18297396	GSE5460	127 (129)	Affymetrix Human Genome U133 Plus 2.0 Array
Turashvili et al.[112]	17389037	GSE5764	30	Affymetrix Human Genome U133 Plus 2.0 Array
Richardson et al[113].	16473279	GSE7904 (an extension of GSE3744)	62	Affymetrix Human Genome U133 Plus 2.0 Array
Karnoub er al.[114]	17914389	GSE8977	22	Affymetrix Human Genome U133 Plus 2.0 Array
Loi et al.[115,116]	18498629, 20479250	GSE9195	77	Affymetrix Human Genome U133 Plus 2.0 Array
Budczies et al.[117]	21339180	GSE11001	14 (30)	Affymetrix Human Genome U133 Plus 2.0 Array
Creighton et al.[118]	19666588	GSE10281	36	Affymetrix Human Genome U133 Plus 2.0 Array
Chen et al.[119]	19266279	GSE10780	179 (185)	Affymetrix Human Genome U133 Plus 2.0 Array
Pedraza et al.[120]	20029976	GSE10810	58	Affymetrix Human Genome U133 Plus 2.0 Array
Bos et al.[121]	19421193	GSE12276	204	Affymetrix Human Genome U133 Plus 2.0 Array
Hoeflich et al.[122,123]	19567590, 21673316	GSE12763	30	Affymetrix Human Genome U133 Plus 2.0 Array
Marty et al.[124 125]	19055754, 21047405	GSE13787	23	Affymetrix Human Genome U133 Plus 2.0 Array
Zhang et al.[80,81]	19573813, 25670079	GSE14017	29	Affymetrix Human Genome U133 Plus 2.0 Array
Desmedt et al.[126]	19573224	GSE16391	55	Affymetrix Human Genome U133 Plus 2.0 Array

Li et al.[127–130]	20098429, 20189874, 21422418, 26484051	GSE16446	120	Affymetrix Human Genome U133 Plus 2.0 Array
Sircoulomb et al.[131]	20932292	GSE17907	55 (109)	Affymetrix Human Genome U133 Plus 2.0 Array
Korde et al.[132]	20012355	GSE18728	61	Affymetrix Human Genome U133 Plus 2.0 Array
Li et al.[127,128,133]	20098429, 20100965, 20189874	GSE18864	71	Affymetrix Human Genome U133 Plus 2.0 Array
Li et al.[127]	20098429	GSE19615	115	Affymetrix Human Genome U133 Plus 2.0 Array
Lin et al.[134]	19967557	GSE19697	24	Affymetrix Human Genome U133 Plus 2.0 Array
Bauer et al.[135]	20062080	GSE20086	12	Affymetrix Human Genome U133 Plus 2.0 Array
Kao et al.[136]	21501481	GSE20685	327	Affymetrix Human Genome U133 Plus 2.0 Array
Dedeurwaerder et al. [137]	21910250	GSE20711	90	Affymetrix Human Genome U133 Plus 2.0 Array
Kretschmer et al. [138,139]	21314937, 21468687	GSE21422	19	Affymetrix Human Genome U133 Plus 2.0 Array
Sabatier et al.[140,141]	20490655, 22110708	GSE21653	266	Affymetrix Human Genome U133 Plus 2.0 Array
Bauer et al.[142–144]	20068102, 20878462, 21633166	GSE22513	28	Affymetrix Human Genome U133 Plus 2.0 Array
Hawthorn et al.[145]	20799942	GSE22544	19 (20)	Affymetrix Human Genome U133 Plus 2.0 Array
Bekhouche et al.[141,146,147]	21339811, 22110708, 22666420	GSE23720	197 (370)	Affymetrix Human Genome U133 Plus 2.0 Array
Latimer et al.[148]	21118987	GSE25407	10	Affymetrix Human Genome U133 Plus 2.0 Array

Planche et al.[149]	21611158	GSE26910	12 (24)	Affymetrix Human Genome U133 Plus 2.0 Array
LaBreche et al.[150]	21781289	GSE27562	161 (162)	Affymetrix Human Genome U133 Plus 2.0 Array
Lehmann et al.[144]	21633166	GSE28821	89	Affymetrix Human Genome U133 Plus 2.0 Array
Cuadros et al.[151,152]	22622817, 22832278	GSE29431	60 (66)	Affymetrix Human Genome U133 Plus 2.0 Array
Jones et al.	No associated publication found	GSE31138	6	Affymetrix Human Genome U133 Plus 2.0 Array
Harvel et al.[153]	23479404	GSE31192	33	Affymetrix Human Genome U133 Plus 2.0 Array
Sabatier et al.[141]	22110708	GSE31448	357	Affymetrix Human Genome U133 Plus 2.0 Array
Miyake et al.[154]	22320227	GSE32646	115	Affymetrix Human Genome U133 Plus 2.0 Array
Clarke et al.[155]	23740839	GSE42568	115 (121)	Affymetrix Human Genome U133 Plus 2.0 Array
Metzger-Filho et al.[156]	23990869	GSE43365	111	Affymetrix Human Genome U133 Plus 2.0 Array
D'Alfonso et al.[157]	23774991	GSE47109	229 (246)	Affymetrix Human Genome U133 Plus 2.0 Array
Huang et al.[158]	24098497	GSE48390	80 (81)	Affymetrix Human Genome U133 Plus 2.0 Array
Prat et al.[159]	24443618	GSE50948	78 (156)	Affymetrix Human Genome U133 Plus 2.0 Array
Fumagalli et al.	No associated publication found	GSE58984	94	Affymetrix Human Genome U133 Plus 2.0 Array
Aswad et al.[160,161]	26474389, 26517092	GSE61304	62	Affymetrix Human Genome U133 Plus 2.0 Array

Burstein MD et al.[162,163]	25208879, 26921331	GSE76275	265	Affymetrix Human Genome U133 Plus 2.0 Array
Brouwers et al.[164]	28673354	GSE90521	17	Affymetrix Human Genome U133 Plus 2.0 Array
Santucci-Pereira et al.[165]	30922380	GSE111662	27	Affymetrix Human Genome U133 Plus 2.0 Array
Dhage et al.[166]	30325954	GSE120129	107 (108)	Affymetrix Human Genome U133 Plus 2.0 Array
Hartung et al.[167]	35082572	GSE167213	87 (124)	Affymetrix Human Genome U133 Plus 2.0 Array
Lehmann et al.[144]	21633166	GSE28796	(24)	Affymetrix Human Genome U133 Plus 2.0 Array
Jose et al.[168]	30169535	GSE93332	(8)	Affymetrix Human Genome U133 Plus 2.0 Array
Pawitan et al.[169,170]	16280042, 16813654	GSE1456	317 (318)	Affymetrix Human Genome U133A / U133B Array
Miller et al.[171–173]	16141321, 29136509, 28972043	GSE3494	501 (502)	Affymetrix Human Genome U133A / U133B Array
Ivshina et al.[174]	17079448	GSE4922	570 (578)	Affymetrix Human Genome U133A / U133B Array
Loi et al.[115,116,175,176]	17401012, 18498629, 20479250, 19552798	GSE6532	741	Affymetrix Human Genome U133A / U133B / U133 Plus 2.0 Array
Rahman et al.[25,26]	26209429, 23000897	GSE62944 (TCGA)	1171	Illumina Genome Analyzer
Brueffer et al.[177–179]	32913985, 32926574, 33937624	GSE81538	405	Illumina HiSeq 2000
Brueffer et al.[177–180]	32913985, 32926574, 33937624, 35304506	GSE96058	3273 (3409)	Illumina HiSeq 2000, Illumina NextSeq 500

	No associated publication	https://dcc.icgc.org/releases/current/Projects/BRCA-KR (ICGC)	50	Illumina HiSeq
Curtis et al.[24]	22522925	https://www.cbioportal.org/study/summary?id=brca_metabric (METABRIC)		Illumina Human HT-12 v3 Expression Beadchips

* When the sample size was different after filtering, we indicate the original sample size in parenthesis.

Supplementary Data

Supplementary File 1 is a summary of the metadata in which the variables have been mapped to the NCIT terms.

Supplementary FIle 2 is a frequency table of the individual NCIT terms mapped to field names.

Supplementary FIles 3 is a frequency table of combinations of NCIT terms mapped to field names.

Supplementary FIles 4 is a frequency table of the individual NCIT terms mapped to data values.

Supplementary FIles 5 is a frequency table of combinations of NCIT terms mapped to data values.

References

1. Jemal, A. *et al.* Global cancer statistics. *CA: A Cancer Journal for Clinicians* vol. 61 69–90 Preprint at https://doi.org/10.3322/caac.20107 (2011).
2. The International Agency for Research on Cancer (IARC). Global Cancer Observatory. https://gco.iarc.fr/.
3. Lal, S., McCart Reed, A. E., de Luca, X. M. & Simpson, P. T. Molecular signatures in breast cancer. *Methods* vol. 131 135–146 Preprint at https://doi.org/10.1016/j.ymeth.2017.06.032 (2017).
4. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* vol. 144 646–674 Preprint at https://doi.org/10.1016/j.cell.2011.02.013 (2011).
5. Arranz, E. E., Vara, J. Á. F., Gámez-Pozo, A. & Zamora, P. Gene signatures in breast cancer: current and future uses. *Transl. Oncol.* **5**, 398–403 (2012).
6. Keenan, T. *et al.* Comparison of the Genomic Landscape Between Primary Breast Cancer in African American Versus White Women and the Association of Racial Differences With Tumor Recurrence. *Journal of Clinical Oncology* vol. 33 3621–3627 Preprint at https://doi.org/10.1200/jco.2015.62.2126 (2015).
7. Ringnér, M., Fredlund, E., Häkkinen, J., Borg, Å. & Staaf, J. GOBO: gene expression-based outcome for breast cancer online. *PLoS One* **6**, e17911 (2011).
8. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991–5 (2013).
9. Nwosu, I. O. & Piccolo, S. R. A systematic review of datasets that can help elucidate relationships among gene expression, race, and immunohistochemistry-defined subtypes in breast cancer. *Cancer Biol. Ther.* **22**, 417–429 (2021).
10. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
11. Parkinson, H. *et al.* ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35**, D747–50 (2007).
12. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
13. Barrett, T. & Edgar, R. Gene expression omnibus: microarray data storage, submission, retrieval,

and analysis. *Methods Enzymol.* **411**, 352–369 (2006).

14. Rayner, T. F. *et al.* A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* **7**, 489 (2006).
15. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).
16. Kauffmann, A. *et al.* Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics* **25**, 2092–2094 (2009).
17. Mecham, A., Stephenson, A., Quinteros, B. I., Salmons, G. & Piccolo, S. R. TidyGEO: Preparing analysis-ready datasets from Gene Expression Omnibus. *bioRxiv* 2023.02.09.527930 (2023) doi:10.1101/2023.02.09.527930.
18. Sioutos, N. *et al.* NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* **40**, 30–43 (2007).
19. Gendoo, D. M. A. *et al.* MetaGxData: Clinically Annotated Breast, Ovarian and Pancreatic Cancer Datasets and their Use in Generating a Multi-Cancer Gene Signature. *Sci. Rep.* **9**, 8770 (2019).
20. Planey, C. R. & Gevaert, O. CoINcIDE: A framework for discovery of patient subtypes across multiple datasets. *Genome Med.* **8**, 27 (2016).
21. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database* **2011**, bar026 (2011).
22. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
23. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
24. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
25. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
26. Rahman, M. *et al.* Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* **31**, 3666–3672 (2015).
27. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

28. The R Project for Statistical Computing. https://www.R-project.org/.
29. Piccolo, S. R. & Frampton, M. B. Tools and techniques for computational reproducibility. *Gigascience* **5**, 30 (2016).
30. Piccolo, S. R. *et al.* A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* **100**, 337–344 (2012).
31. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).
32. Waldron, L., Riester, M., Ramos, M., Parmigiani, G. & Birrer, M. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. *J. Natl. Cancer Inst.* **108**, (2016).
33. Rosikiewicz, M. & Robinson-Rechavi, M. IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinformatics* vol. 30 1392–1399 Preprint at https://doi.org/10.1093/bioinformatics/btu027 (2014).
34. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).
35. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
36. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
37. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
38. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
39. GeneCards Human Gene Database. GeneCards - Human Genes. https://www.genecards.org/.
40. NCI Term Browser. http://nciterms.nci.nih.gov/.
41. Foster, E. D. & Deardorff, A. Open Science Framework (OSF). *J. Med. Libr. Assoc.* **105**, 203 (2017).
42. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
43. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).

44. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

45. Fentiman, I. S., Fourquet, A. & Hortobagyi, G. N. Male breast cancer. *Lancet* **367**, 595–604 (2006).

46. Morris, G. J. *et al.* Differences in breast carcinoma characteristics in newly diagnosed African–American and Caucasian patients. *Cancer* **110**, 876–884 (2007).

47. Zhou, Y. *et al.* Enhanced NFκB and AP-1 transcriptional activity associated with antiestrogen resistant breast cancer. *BMC Cancer* **7**, 59 (2007).

48. Yau, C. & Benz, C. C. Genes responsive to both oxidant stress and loss of estrogen receptor function identify a poor prognosis group of estrogen receptor positive primary breast cancers. *Breast Cancer Res.* **10**, R61 (2008).

49. Yau, C. *et al.* Aging impacts transcriptomes but not genomes of hormone-dependent breast cancers. *Breast Cancer Res.* **9**, R59 (2007).

50. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541 (2006).

51. Knudsen, E. S. *et al.* Progression of ductal carcinoma in situ to invasive breast cancer is associated with gene expression programs of EMT and myoepithelia. *Breast Cancer Res. Treat.* **133**, 1009–1024 (2012).

52. Tofigh, A. *et al.* The Prognostic Ease and Difficulty of Invasive Breast Carcinoma. *Cell Reports* vol. 9 129–142 Preprint at https://doi.org/10.1016/j.celrep.2014.08.073 (2014).

53. Lehmann, B. D. *et al.* Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. *PLoS One* **11**, e0157368 (2016).

54. Gregory, K. J. *et al.* Gene expression signature of atypical breast hyperplasia and regulation by SFRP1. *Breast Cancer Res.* **21**, 76 (2019).

55. Chang, J. C. *et al.* Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. *J. Clin. Oncol.* **23**, 1169–1177 (2005).

56. Lee, S. *et al.* Differentially expressed genes regulating the progression of ductal carcinoma in situ to invasive breast cancer. *Cancer Res.* **72**, 4574–4586 (2012).

57. Farmer, P. *et al.* Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* **24**, 4660–4671 (2005).

58. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679 (2005).

59. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518–524 (2005).

60. Azzam, D. J. *et al.* Triple negative breast cancer initiating cell subsets differ in functional and molecular characteristics and in γ-secretase inhibitor drug responses. *EMBO Mol. Med.* **5**, 1502–1522 (2013).

61. Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262–272 (2006).

62. Rody, A. *et al.* Poor Outcome in Estrogen Receptor–Positive Breast Cancers Predicted by Loss of Plexin B1. *Clin. Cancer Res.* **13**, 1115–1122 (2007).

63. Ruckhäberle, E. *et al.* Microarray analysis of altered sphingolipid metabolism reveals prognostic significance of sphingosine kinase 1 in breast cancer. *Breast Cancer Res. Treat.* **112**, 41–52 (2008).

64. Rody, A. *et al.* T-cell metagene predicts a favorable prognosis in estrogen receptor-negative and HER2-positive breast cancers. *Breast Cancer Res.* **11**, R15 (2009).

65. Karn, T. *et al.* 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**, 567–579 (2010).

66. Karn, T. *et al.* Gene expression profiling of luminal B breast cancers reveals NHERF1 as a new marker of endocrine resistance. *Breast Cancer Res. Treat.* **130**, 409–420 (2011).

67. Győrffy, B. *et al.* Dynamic classification using case-specific training cohorts outperforms static gene expression signatures in breast cancer. *Int. J. Cancer* **136**, 2091–2098 (2015).

68. Minn, A. J. *et al.* Lung metastasis genes couple breast tumor size and metastatic spread. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 6740–6745 (2007).

69. Boersma, B. J. *et al.* A stromal gene signature associated with inflammatory breast cancer. *Int. J. Cancer* **122**, 1324–1332 (2008).

70. Martin, D. N. *et al.* Differences in the tumor microenvironment between African-American and European-American breast cancer patients. *PLoS One* **4**, e4531 (2009).

71. Glynn, S. A. *et al.* Increased NOS2 predicts poor survival in estrogen receptor–negative breast cancer patients. *J. Clin. Invest.* **120**, 3843–3854 (11 2010).

72. Desmedt, C. *et al.* Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series. *Clinical Cancer Research* vol. 13 3207–3214 Preprint at https://doi.org/10.1158/1078-0432.ccr-06-2765 (2007).
73. Patil, P., Bachant-Winner, P.-O., Haibe-Kains, B. & Leek, J. T. Test set bias affects reproducibility of gene signatures. *Bioinformatics* **31**, 2318–2323 (2015).
74. Tripathi, A. *et al.* Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *Int. J. Cancer* **122**, 1557–1566 (2008).
75. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68**, 5405–5413 (2008).
76. Hellwig, B. *et al.* Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics* **11**, 276 (2010).
77. Cadenas, C. *et al.* Loss of circadian clock gene expression is associated with tumor progression in breast cancer. *Cell Cycle* **13**, 3282–3291 (2014).
78. Heimes, A.-S. *et al.* Prognostic Significance of Interferon-γ and Its Signaling Pathway in Early Breast Cancer Depends on the Molecular Subtypes. *Int. J. Mol. Sci.* **21**, (2020).
79. Zhang, Y. *et al.* The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. *Breast Cancer Res. Treat.* **116**, 303–309 (2009).
80. Zhang, X. H.-F. *et al.* Latent bone metastasis in breast cancer tied to Src-dependent survival signals. *Cancer Cell* **16**, 67–78 (2009).
81. Xu, J. *et al.* 14-3-3ζ turns TGF-β's function from tumor suppressor to metastasis promoter in breast cancer by contextual changes of Smad partners from p53 to Gli2. *Cancer Cell* **27**, 177–192 (2015).
82. Emery, L. A. *et al.* Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. *Am. J. Pathol.* **175**, 1292–1302 (2009).
83. Symmans, W. F. *et al.* Genomic index of sensitivity to endocrine therapy for breast cancer. *J. Clin. Oncol.* **28**, 4111–4119 (2010).
84. Miller, W. R. & Larionov, A. Changes in expression of oestrogen regulated and proliferation genes with neoadjuvant treatment highlight heterogeneity of clinical resistance to the aromatase

inhibitor, letrozole. *Breast Cancer Research* vol. 12 Preprint at https://doi.org/10.1186/bcr2611 (2010).

85. Miller, W. R., Larionov, A., Anderson, T. J., Evans, D. B. & Dixon, J. M. Sequential changes in gene expression profiles in breast cancers during treatment with the aromatase inhibitor, letrozole. *Pharmacogenomics J.* **12**, 10–21 (2012).
86. Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838 (2010).
87. Popovici, V. *et al.* Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res.* **12**, R5 (2010).
88. Tabchy, A. *et al.* Evaluation of a 30-Gene Paclitaxel, Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy Response Predictor in a Multicenter Randomized Trial in Breast Cancer. *Clinical Cancer Research* vol. 16 5351–5361 Preprint at https://doi.org/10.1158/1078-0432.ccr-10-1265 (2010).
89. Shen, K. *et al.* A systematic evaluation of multi-gene predictors for the pathological response of breast cancer patients to chemotherapy. *PLoS One* **7**, e49529 (2012).
90. Graham, K. *et al.* Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *British Journal of Cancer* vol. 102 1284–1293 Preprint at https://doi.org/10.1038/sj.bjc.6605576 (2010).
91. Graham, K., Ge, X., de las Morenas, A., Tripathi, A. & Rosenberg, C. L. Gene Expression Profiles of Estrogen Receptor–Positive and Estrogen Receptor–Negative Breast Cancers Are Detectable in Histologically Normal Breast Epithelium. *Clinical Cancer Research* vol. 17 236–246 Preprint at https://doi.org/10.1158/1078-0432.ccr-10-1369 (2011).
92. Iwamoto, T. *et al.* Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. *J. Natl. Cancer Inst.* **103**, 264–272 (2011).
93. Creighton, C. J. *et al.* A gene transcription signature of obesity in breast cancer. *Breast Cancer Res. Treat.* **132**, 993–1000 (2012).
94. Hatzis, C. A Genomic Predictor of Response and Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer. *JAMA* vol. 305 1873 Preprint at https://doi.org/10.1001/jama.2011.593 (2011).

95. Baldasici, O. *et al.* Circulating Small EVs miRNAs as Predictors of Pathological Response to Neo-Adjuvant Therapy in Breast Cancer Patients. *Int. J. Mol. Sci.* **23**, (2022).

96. Itoh, M. *et al.* Estrogen receptor (ER) mRNA expression and molecular subtype distribution in ER-negative/progesterone receptor-positive breast cancers. *Breast Cancer Res. Treat.* **143**, 403–409 (2014).

97. Rody, A. *et al.* A clinically relevant gene signature in triple negative and basal-like breast cancer. *Breast Cancer Res.* **13**, R97 (2011).

98. Karn, T. *et al.* Homogeneous datasets of triple negative breast cancers enable the identification of novel prognostic and predictive signatures. *PLoS One* **6**, e28403 (2011).

99. Karn, T. *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**, 12–23 (2012).

100. Karn, T. *et al.* Control of dataset bias in combined Affymetrix cohorts of triple negative breast cancer. *Genom Data* **2**, 354–356 (2014).

101. Bianchini, G. *et al.* Molecular anatomy of breast cancer stroma and its prognostic value in estrogen receptor-positive and -negative cancers. *J. Clin. Oncol.* **28**, 4316–4323 (2010).

102. Gonzalez-Angulo, A. M. *et al.* Gene expression, molecular class changes, and pathway analysis after neoadjuvant systemic therapy for breast cancer. *Clin. Cancer Res.* **18**, 1109–1119 (2012).

103. Nagalla, S. *et al.* Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis. *Genome Biol.* **14**, R34 (2013).

104. Rody, A. *et al.* Gene expression of topoisomerase II alpha (TOP2A) by microarray analysis is highly prognostic in estrogen receptor (ER) positive breast cancer. *Breast Cancer Res. Treat.* **113**, 457–466 (2009).

105. Rody, A. *et al.* Loss of Plexin B1 is highly prognostic in low proliferating ER positive breast cancers--results of a large scale microarray analysis. *Eur. J. Cancer* **45**, 405–413 (2009).

106. Milde-Langosch, K. *et al.* Validity of the proliferation markers Ki67, TOP2A, and RacGAP1 in molecular subgroups of breast cancer. *Breast Cancer Res. Treat.* **137**, 57–67 (2013).

107. Milde-Langosch, K. *et al.* Prognostic relevance of glycosylation-associated genes in breast cancer. *Breast Cancer Res. Treat.* **145**, 295–305 (2014).

108. Casey, T. *et al.* Molecular signatures suggest a major role for stromal cells in development of

invasive breast cancer. *Breast Cancer Res. Treat.* **114**, 47–62 (2009).

109. Reyngold, M. *et al.* Remodeling of the methylation landscape in breast cancer metastasis. *PLoS One* **9**, e103896 (2014).

110. Walsh, L. A. *et al.* An Integrated Systems Biology Approach Identifies TRIM25 as a Key Determinant of Breast Cancer Metastasis. *Cell Rep.* **20**, 1623–1640 (2017).

111. Lu, X. *et al.* Predicting features of breast cancer with gene expression patterns. *Breast Cancer Res. Treat.* **108**, 191–201 (2008).

112. Turashvili, G. *et al.* Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* **7**, 55 (2007).

113. Richardson, A. L. *et al.* X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* **9**, 121–132 (2006).

114. Karnoub, A. E. *et al.* Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. *Nature* **449**, 557–563 (2007).

115. Loi, S. *et al.* Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* **9**, 239 (2008).

116. Loi, S. *et al.* PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor–positive breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10208–10213 (2010).

117. Budczies, J. *et al.* Genome-wide gene expression profiling of formalin-fixed paraffin-embedded breast cancer core biopsies using microarrays. *J. Histochem. Cytochem.* **59**, 146–157 (2011).

118. Creighton, C. J. *et al.* Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 13820–13825 (2009).

119. Chen, D.-T. *et al.* Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. *Breast Cancer Res. Treat.* **119**, 335–346 (2010).

120. Pedraza, V. *et al.* Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness. *Cancer* **116**, 486–496 (2010).

121. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459**, 1005–1009 (2009).

122. Hoeflich, K. P. *et al.* In vivo antitumor activity of MEK and phosphatidylinositol 3-kinase inhibitors in basal-like breast cancer models. *Clin. Cancer Res.* **15**, 4649–4664 (2009).

123. Stinson, S. *et al.* TRPS1 targeting by miR-221/222 promotes the epithelial-to-mesenchymal transition in breast cancer. *Sci. Signal.* **4**, ra41 (2011).

124. Marty, B. *et al.* Frequent PTEN genomic alterations and activated phosphatidylinositol 3-kinase pathway in basal-like breast cancer cells. *Breast Cancer Res.* **10**, R101 (2008).

125. Servant, N. *et al.* EMA - A R package for Easy Microarray data analysis. *BMC Res. Notes* **3**, 277 (2010).

126. Desmedt, C. *et al.* The Gene expression Grade Index: a potential predictor of relapse for endocrine-treated breast cancer patients in the BIG 1-98 trial. *BMC Med. Genomics* **2**, 40 (2009).

127. Li, Y. *et al.* Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nat. Med.* **16**, 214–218 (2010).

128. Juul, N. *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**, 358–365 (2010).

129. Desmedt, C. *et al.* Multifactorial approach to predicting resistance to anthracyclines. *J. Clin. Oncol.* **29**, 1578–1586 (2011).

130. Haibe-Kains, B. *et al.* Genome-wide gene expression profiling to predict resistance to anthracyclines in breast cancer patients. *Genom Data* **1**, 7–10 (2013).

131. Sircoulomb, F. *et al.* Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer* **10**, 539 (2010).

132. Korde, L. A. *et al.* Gene expression pathway analysis to predict response to neoadjuvant docetaxel and capecitabine for breast cancer. *Breast Cancer Res. Treat.* **119**, 685–699 (2010).

133. Silver, D. P. *et al.* Efficacy of neoadjuvant Cisplatin in triple-negative breast cancer. *J. Clin. Oncol.* **28**, 1145–1153 (2010).

134. Lin, Y. *et al.* A gene expression signature that predicts the therapeutic response of the basal-like breast cancer to neoadjuvant chemotherapy. *Breast Cancer Res. Treat.* **123**, 691–699 (2010).

135. Bauer, M. *et al.* Heterogeneity of gene expression in stromal fibroblasts of human breast carcinomas and normal breast. *Oncogene* **29**, 1732–1740 (2010).

136. Kao, K.-J., Chang, K.-M., Hsu, H.-C. & Huang, A. T. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer* **11**, 143 (2011).

137. Dedeurwaerder, S. *et al.* DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol. Med.* **3**, 726–741 (2011).

138. Kretschmer, C. *et al.* Identification of early molecular markers for breast cancer. *Mol. Cancer* **10**, 15 (2011).

139. Kretschmer, C., Conradi, A., Kemmner, W. & Sterner-Kock, A. Latent transforming growth factor binding protein 4 (LTBP4) is downregulated in mouse and human DCIS and mammary carcinomas. *Cell. Oncol.* **34**, 419–434 (2011).

140. Sabatier, R. *et al.* A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res. Treat.* **126**, 407–420 (2011).

141. Sabatier, R. *et al.* Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. *PLoS One* **6**, e27656 (2011).

142. Bauer, J. A. *et al.* Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiation. *Clin. Cancer Res.* **16**, 681–690 (2010).

143. Adams, S. *et al.* Preoperative concurrent paclitaxel-radiation in locally advanced breast cancer: pathologic response correlates with five-year overall survival. *Breast Cancer Res. Treat.* **124**, 723–732 (2010).

144. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**, 2750–2767 (2011).

145. Hawthorn, L., Luce, J., Stein, L. & Rothschild, J. Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast. *BMC Cancer* **10**, 460 (2010).

146. Bekhouche, I. *et al.* High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. *PLoS One* **6**, e16950 (2011).

147. Bertucci, F. *et al.* 8q24 Cancer risk allele associated with major metastatic risk in inflammatory breast cancer. *PLoS One* **7**, e37943 (2012).

148. Latimer, J. J. *et al.* Nucleotide excision repair deficiency is intrinsic in sporadic stage I breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21725–21730 (2010).

149. Planche, A. *et al.* Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. *PLoS One* **6**, e18640 (2011).

150. LaBreche, H. G., Nevins, J. R. & Huang, E. Integrating factor analysis and a transgenic mouse

model to reveal a peripheral blood predictor of breast tumors. *BMC Med. Genomics* **4**, 61 (2011).

151. Lopez, F. J., Cuadros, M., Cano, C., Concha, A. & Blanco, A. Biomedical application of fuzzy association rules for identifying breast cancer biomarkers. *Med. Biol. Eng. Comput.* **50**, 981–990 (2012).

152. Cuadros, M., Cano, C., López, F. J., López-Castro, R. & Concha, A. Expression profiling of breast tumors based on human epidermal growth factor receptor 2 status defines migration-related genes. *Pathobiology* **80**, 32–40 (2013).

153. Harvell, D. M. E. *et al.* Genomic signatures of pregnancy-associated breast cancer epithelia and stroma and their regulation by estrogens and progesterone. *Horm. Cancer* **4**, 140–153 (2013).

154. Miyake, T. *et al.* GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. *Cancer Science* vol. 103 913–920 Preprint at https://doi.org/10.1111/j.1349-7006.2012.02231.x (2012).

155. Clarke, C. *et al.* Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* **34**, 2300–2308 (2013).

156. Metzger-Filho, O. *et al.* Genomic Grade Index (GGI): feasibility in routine practice and impact on treatment decisions in early breast cancer. *PLoS One* **8**, e66848 (2013).

157. D'Alfonso, T. M. *et al.* BreastPRS is a gene expression assay that stratifies intermediate-risk Oncotype DX patients into high- or low-risk for disease recurrence. *Breast Cancer Res. Treat.* **139**, 705–715 (2013).

158. Huang, C.-C. *et al.* Concurrent gene signatures for han chinese breast cancers. *PLoS One* **8**, e76421 (2013).

159. Prat, A. *et al.* Research-Based PAM50 Subtype Predictor Identifies Higher Responses and Improved Survival Outcomes in HER2-Positive Breast Cancer in the NOAH Study. *Clinical Cancer Research* vol. 20 511–521 Preprint at https://doi.org/10.1158/1078-0432.ccr-13-0239 (2014).

160. Aswad, L. *et al.* Genome and transcriptome delineation of two major oncogenic pathways governing invasive ductal breast cancer development. *Oncotarget* **6**, 36652–36674 (2015).

161. Grinchuk, O. V. *et al.* Sense-antisense gene-pairs in breast cancer and associated pathological pathways. *Oncotarget* **6**, 42197–42221 (2015).

162. Burstein, M. D. *et al.* Comprehensive genomic analysis identifies novel subtypes and targets of

triple-negative breast cancer. *Clin. Cancer Res.* **21**, 1688–1698 (2015).

163. den Hollander, P. *et al.* Phosphatase PTP4A3 Promotes Triple-Negative Breast Cancer Growth and Predicts Poor Patient Survival. *Cancer Res.* **76**, 1942–1953 (2016).

164. Brouwers, B. *et al.* The footprint of the ageing stroma in older patients with breast cancer. *Breast Cancer Res.* **19**, 78 (2017).

165. Santucci-Pereira, J. *et al.* Genomic signature of parity in the breast of premenopausal women. *Breast Cancer Res.* **21**, 46 (2019).

166. Dhage, S. *et al.* A genomic ruler to assess oncogenic transition between breast tumor and stroma. *PLoS One* **13**, e0205602 (2018).

167. Hartung, C. *et al.* Identifying High-Risk Triple-Negative Breast Cancer Patients by Molecular Subtyping. *Breast Care* **16**, 637–647 (2021).

168. Jose, V. *et al.* Feasibility of developing reliable gene expression modules from FFPE derived RNA profiled on Affymetrix arrays. *PLoS One* **13**, e0203346 (2018).

169. Pawitan, Y. *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* **7**, R953–64 (2005).

170. Hall, P. *et al.* Hormone-replacement therapy influences gene expression profiles and is associated with breast-cancer prognosis: a cohort study. *BMC Med.* **4**, 16 (2006).

171. Miller, L. D. *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13550–13555 (2005).

172. Palazon, A. *et al.* An HIF-1α/VEGF-A Axis in Cytotoxic T Cells Regulates Tumor Progression. *Cancer Cell* **32**, 669–683.e5 (2017).

173. Lundberg, A. *et al.* Gene Expression Signatures and Immunohistochemical Subtypes Add Prognostic Value to Each Other in Breast Cancer Cohorts. *Clin. Cancer Res.* **23**, 7512–7520 (2017).

174. Ivshina, A. V. *et al.* Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* **66**, 10292–10301 (2006).

175. Loi, S. *et al.* Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.* **25**, 1239–1246 (2007).

176. Loi, S. *et al.* Gene expression profiling identifies activated growth factor signaling in poor prognosis (Luminal-B) estrogen receptor positive breast cancer. *BMC Med. Genomics* **2**, 37 (2009).

177. Brueffer, C. *et al.* Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative. *JCO Precis Oncol* **2**, (2018).

178. Brueffer, C. *et al.* The mutational landscape of the SCAN □B real□world primary breast cancer transcriptome. *EMBO Molecular Medicine* vol. 12 Preprint at https://doi.org/10.15252/emmm.202012118 (2020).

179. Dahlgren, M. *et al.* Preexisting Somatic Mutations of Estrogen Receptor Alpha (*ESR1*) in Early-Stage Primary Breast Cancer. *JNCI Cancer Spectrum* vol. 5 Preprint at https://doi.org/10.1093/jncics/pkab028 (2021).

180. Dalal, H. *et al.* Clinical associations of ESR2 (estrogen receptor beta) expression across thousands of primary breast tumors. *Scientific Reports* vol. 12 Preprint at https://doi.org/10.1038/s41598-022-08210-3 (2022).