Curated compendium of human transcriptional biomarker data

Nathan P. Golightly¹, Anna I. Bischoff¹, Avery Bell¹, Parker D. Hollingsworth¹,², Stephen R. Piccolo¹,³,*

¹ - Department of Biology, Brigham Young University, Provo, Utah, 84602, USA
² - Northeast Ohio Medical University, Rootstown, Ohio, 44272, USA
³ - Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, 84602, USA

* - Please address correspondence to S.R.P. at stephen_piccolo@byu.edu.
Abstract

Genome-wide transcriptional profiles provide broad insights into cellular activity. One important use of such data is to identify relationships between transcription levels and patient outcomes. These translational insights can guide the development of biomarkers for predicting outcomes in clinical settings. Over the past decades, data from many translational-biomarker studies have been deposited in public repositories, enabling other scientists to reuse the data in follow-up studies. However, data-reuse efforts require considerable time and expertise because transcriptional data are generated using heterogeneous profiling technologies, preprocessed using diverse normalization procedures, and annotated in non-standard ways. To address this problem, we curated a compendium of 45 translational-biomarker datasets from the public domain. To increase the data's utility, we reprocessed the raw expression data using a standard computational pipeline and standardized the clinical annotations in a fully reproducible manner (see https://osf.io/ssk3t). We believe these data will be particularly useful to researchers seeking to validate gene-level findings or to perform benchmarking studies—for example, to compare and optimize machine-learning algorithms' ability to predict biomedical outcomes.

Background & summary

DNA encodes a cell's instruction manual in the form of genes and regulatory sequences\(^1\). Cells behave differently, in part, because genes are transcribed into RNA at different levels within those cells\(^2\). Researchers examine gene-expression levels to understand cellular dynamics and the mechanisms behind cellular aberrations, including those that lead to disease development. Modern technologies now make it possible to profile expression levels for thousands of genes at a time for a modest expense\(^3\). Using these high-throughput technologies,
scientists have performed thousands of studies to characterize biological processes and to evaluate the potential for precision-medicine applications. One such application is to derive transcriptional biomarkers—patterns of expression that indicate disease states or that predict medical outcomes, such as relapse, survival, or treatment response\textsuperscript{4–10}. Indeed, already to date, more than 100 transcriptional biomarkers have been proposed for predicting breast-cancer survival alone\textsuperscript{11}.

Many funding agencies and academic journals have imposed policies that require scientists to deposit transcriptional data in publicly accessible databases. These policies seek to ensure that other scientists can verify the original study's findings and can reuse the data in secondary analyses. For example, Gene Expression Omnibus (GEO) currently contains data for more than 2 million biological samples\textsuperscript{12}. Upon considering infrastructure and personnel costs, we estimate that these data represent hundreds of millions—if not billions—of dollars (USD) of collective research investment. Reusing these vast resources offers an opportunity to reap a greater return on investment—perhaps most importantly via informing and validating new studies. Unfortunately, although anyone can access GEO data, researchers vastly underutilize this treasure trove because preparing data for new analyses requires considerable background knowledge and informatics expertise.

In GEO, data are typically available in two forms: 1) raw data, as produced originally by the data-generating technology, and 2) processed data, which were used in the data generators' analyses. In most cases, researchers process raw data in a series of steps that might include quality-control filtering, noise reduction, standardization, and summarization (e.g., summarizing to gene-level values and excluding outliers). Data from different profiling technologies must be handled in ways that are specific to each technology. However, even for datasets generated using the same profiling technology, the methods employed for data preprocessing vary widely
across studies. This heterogeneity makes it difficult for researchers to perform secondary analyses and to trust that analytical findings are driven primarily by biological mechanisms rather than differences in data preprocessing. In addition, when data have not been mapped to biologically meaningful identifiers, it may be difficult for researchers to draw biological conclusions from the data.

Sample-level annotations accompany each GEO dataset. For biomarker studies, such metadata might include medical diagnoses or treatment outcomes, as well as covariates such as age, sex, or ethnicity. Although GEO publishes metadata in a semi-standardized format and bioinformatics tools exist for downloading and parsing GEO data\textsuperscript{13,14}, it is difficult for many researchers to extract these data into a form that is suitable for secondary analyses. Within annotation files, values are often stored in key/value pairs with nondescript column names. Many columns are not useful for analytical purposes (e.g., when all samples have the same value). When values are missing, the columns often become shifted; accordingly, data for a given variable may be spread across multiple columns. Moreover, a variety of descriptors (e.g., “?”, “N/A”, or “Unknown”) are used to indicate missing values, thus requiring the analyst to account for these differences. In addition, seemingly minor errors, such as spelling mistakes or inconsistent capitalization, can hamper secondary-analysis efforts.

In response to these challenges, we compiled the \textit{Biomarker Benchmark}, a curated compendium of 45 transcriptional-biomarker datasets from GEO. These datasets represent a variety of human-disease states and outcomes, many related to cancer. We obtained raw gene-expression files, renormalized them using a common algorithm, and summarized the data using gene-level annotations (Figure 1). We used two techniques to check for quality-control issues in the gene-expression data. For datasets where gene-expression data were processed in multiple batches—and where batch information was available—we corrected for batch
effects. Finally, we prepared a version of the data that is suitable for direct application in machine-learning analyses. We standardized continuous values, one-hot encoded discrete values, and imputed missing values.

**Methods**

**Selecting data**

To select datasets, we executed a custom search in Gene Expression Omnibus (GEO). First, we limited our search to data series that were associated with the Medical Subject Heading (MeSH) term "biomarker" and that came from *Homo sapiens* subjects. Next we limited the search to data generated using Affymetrix gene-expression microarrays and for which raw expression data were available (so we could renormalize the data). For each dataset, we examined the metadata to ensure that each series had at least one biomarker-relevant clinical variable. These included variables such as prognosis, disease stage, histology, and treatment success or relapse. Lastly, we selected series that included data for at least 70 samples (before additional filtering, see below).

Based on these criteria, we identified 36 GEO series. Two series (GSE6532 and GSE26682) contained data for two types of Affymetrix microarray. To avoid platform-related biases, we separated each of these series into two datasets; we used a suffix for each that indicates the microarray platform (e.g., GSE6532_U133A and GSE6532_U133Plus2). For both of these datasets, the biological samples profiled using either microarray platform were distinct. The GSE2109 series—known as the Expression Project for Oncology (expO)—had been produced by the International Genomics Consortium and contains data for 129 different cancer types. To avoid confounding effects due to tissue-specific expression and because the
metadata differed considerably across the cancer types, we split GSE2109 into multiple datasets based on cancer type (Table 1). We excluded tissue types for which fewer than 70 samples were available; we also excluded the "omentum" cancer type because it was relatively heterogeneous and had relatively few samples.

We used publicly available data for this study and played no role in contacting the research subjects. We received approval to work with these data from Brigham Young University's Institutional Review Board (E 14522).

**Preparing clinical annotations**

For each dataset, we wrote custom R scripts\(^{15}\) that download, parse, and reformat the clinical annotations. Initially, these scripts download data using the `GEOquery` package\(^{13}\). Next they generate a tab-delimited text file for each dataset that contains all available clinical annotations, except those with identical values for all samples (for example, platform name, species name, submission date) or that were unique to each biological sample (for example, sample title). In addition, these scripts generate Markdown files ([https://daringfireball.net/projects/markdown/syntax](https://daringfireball.net/projects/markdown/syntax)) that summarize each dataset and indicate sources.

In some cases, multiple data values are included in the same cell in GEO annotation files. For example, in GSE5462, one patient's clinical demographics and treatment responses are listed as "female; breast tumor; Letrozole, 2.5mg/day, oral, 10-14 days; responder." We parsed these values and split them into separate columns for each sample. After these cleaning steps, the datasets contained 7.8 variables of metadata, on average (Table 1). Next we searched each dataset for missing values. Across the datasets, 11 distinct expressions had been used by the original data generators to represent missingness; these included "N/A", "NA", "unknown", "N\(\text{UM}\)\(\text{A}\)".
"MISSING", "NOT AVAILABLE", "?", and others. To support consistency, we standardized these values across the datasets, using a value of "NA". On average, 17.0% of the metadata values were missing per dataset; this proportion differed considerably across the datasets (Figure 2).

We anticipate that many researchers will use these data to develop and benchmark machine-learning algorithms. Accordingly, we have prepared secondary versions of the clinical annotations that are ready to use in machine-learning analyses. First, we identified class variables that have potential relevance for biomarker applications. In many cases, these variables were identical to those used in the original studies; we also included class variables that had not been used in the original studies. On average, the datasets contain 2.9 class variables. Second, we identified clinical variables that could be used as predictor variables (or covariates). Using these data, we generated one output file per class variable or predictor variable and named the output files using descriptive prefixes (e.g., "Prognosis", "Diagnosis", or "Stage"). The same variable can be used as a class variable in one context and a predictor in a different context. When a given sample was missing data for a given class variable, we excluded that sample from the respective output file for that class variable. After this filtering step, we identified class variables with fewer than 40 samples and excluded these class variables. When predictor variables were missing more than 20% data (Figure 2), we did not generate an output file for these variables. When predictor variables were missing less than 20% data, we imputed missing values using median-based imputation for continuous variables and mode-based imputation for discrete variables\(^6\). We scaled continuous predictor variables to have zero mean and unit variance. We transformed discrete predictor variables using one-hot encoding; each unique value, except the first, was treated as a binary variable. In cases where discrete values were rare, we merged values. For example, in GSE2109_Breast, we merged Pathological\_Stage values 3A, 3B, 3C, and 4 into a category called "3-4" because relatively few
patients fell into the individual categories (38, 8, 22, and 5 samples, respectively). In addition, some class variables were ordinal in nature (e.g., cancer stage or tumor grade); we transformed these to binary variables. Finally, some clinical outcomes were survival or relapse times; we transformed these data to (discrete) class variables, using a threshold to distinguish between "long-term" and "short-term" survivors and excluding patients who were censored after the survival threshold had been reached. Our computer scripts (see Code availability) encode these decisions for each dataset.

**Preprocessing gene-expression data**

We created a computational pipeline (using R and shell scripts) that downloads, normalizes, and standardizes the raw-expression data. We used the GEOquery package\(^{13}\) to download the CEL files and then normalized them using the SCAN.UPC package\(^ {17}\). Some heterogeneity exists even among platforms from the same manufacturer (Affymetrix). The number of probes and the probe sequences used in designing the microarray architectures vary. To help mitigate this heterogeneity and to aid in biological interpretation, we summarized the data using gene-level annotations from Brainarray\(^ {18}\).

**Code availability**

Our computer scripts are stored in the open-access Biomarker Benchmark repository (https://osf.io/ssk3t). Accordingly, other researchers can reproduce our curation process and produce alternative versions of the data.
Data records

After filtering (see Methods), we collected data for 7,037 biological samples across 45 datasets (Table 1). On average, the datasets contain values for 18,043 genes (Table 1). In total, our repository contains 129 class variables (2.8 per dataset) and 2.1 unique values per class variable.

All output data are stored in tab-delimited text files and are structured using the "tidy data" methodology\(^{19}\). Accordingly, data users can import the files directly into analytical tools such as Microsoft Excel, R, or Python. All data files are publicly and freely available in the open-access Biomarker Benchmark repository (https://osf.io/ssk3t). The original data files are available via Gene Expression Omnibus using the accession numbers listed in Table 1.

Technical validation

We evaluated each sample using the IQRray\(^{20}\) software, which produces a quality score for individual samples. Using these metrics, we applied Grubb’s statistical test (outliers package\(^{21}\)) to each dataset, identified poor-quality outliers (Figure 3), and excluded these samples (Table 2). Next we used the DoppelgangR package\(^{22}\) to identify samples that may have been duplicated inadvertently. We manually reviewed sample pairs that DoppelgangR flagged as potential duplicates. We excluded most sample pairs that were flagged (Table 2), even if the clinical annotations for both samples were distinct, under the assumption that these samples had somehow been mislabeled. In GSE46449, many samples were biological replicates, and we retained one of each replicate set. GSE5462, GSE19804, and GSE20181 contained
samples that had been profiled in a paired manner (e.g., pre- and post-treatment); we retained these samples.

When transcriptomic data are processed in multiple batches, batch assignments can lead to confounding effects\textsuperscript{23}. In the clinical annotations, we identified batch-processing information for datasets GSE25507, GSE37199, GSE39582, and GSE40292. We corrected for batch effects using the ComBat software\textsuperscript{24}. The Biomarker Benchmark repository contains pre- and post-batch-corrected data. For dataset GSE37199, we identified two variables that could have been used for batch correction ("Centre" and "Plate"). Our repository contains batch-corrected data for both of these batch variables (the default is "Plate").

**Acknowledgements**

SRP thanks Brigham Young University for research funds used in this study. AIB and PDH thank the BYU Office of Research and Creative Activities for research funds that supported this work. NPG thanks the Simmons Center for Cancer Research at Brigham Young University for a summer fellowship that supported this work. We thank researchers from many institutions who generated these data and released them to the public. We also thank the many research participants who participated in these studies.

**Author contributions**

NPG: Collected data, wrote computer scripts, evaluated data quality, prepared figures and tables, wrote the manuscript.

AIB: Collected data, wrote computer scripts, evaluated data quality, edited the manuscript.

AB: Wrote computer scripts, wrote the manuscript.
PDH: Collected data, wrote computer scripts, edited the manuscript.

SRP: Collected data, wrote computer scripts, prepared figures and tables, wrote the manuscript.

Competing interests

The author(s) declare no competing financial interests.

Figures
Figure 1: Flow diagram that illustrates the process we used to collect and curate the data. We wrote computer scripts that downloaded the data, checked for quality, normalized and standardized data values, and stored the data in analysis-ready file formats. The specific steps differed for clinical and expression data (see Methods).
Figure 2: Histogram showing the proportion of missing clinical-annotation values per dataset.

Some datasets contained no missing values, while others were missing as many as as 72.3% of data values.
Figure 3: Distribution of IQRay quality scores for each dataset. Sample qualities are plotted for each dataset. Low-quality samples were identified using Grubb’s test. Samples that fall on or below the red threshold were excluded from the data repository.
### Tables

#### Table 1: Overview of data sources used in this study.

| Series ID* | Tissue type(s) | # Samples | Clinical variable(s)                                                                 | # Genes | Affymetrix Platform(s) |
|------------|----------------|-----------|-------------------------------------------------------------------------------------|---------|------------------------|
| GSE1456⁶⁸  | Breast cancer  | 157       | Elston grade; overall survival status; overall survival time; relapse status; relapse time | 11832   | Genome U133A           |
| GSE2109⁶⁸  | Breast         | 263       | Age; alcohol consumption; days from diagnosis to excision; ER status; ethnic background; ethnic background; family history of cancer; fibrocystic disease; Her2 status; histology; hormonal therapy duration; mammogram status and findings; metastasis; metastatic sites; multiple tumors; node involvement; oophorectomy status; oral contraceptive use; PR status; prior therapy status; quality metric; relapse time; retreatment states; sex; stage; tobacco use; tumor grade; tumor size | 20024   | Genome U133 Plus 2.0   |
| GSE2109⁶⁸  | Colon          | 255       | Age; alcohol consumption; days from diagnosis to excision; diagnosis method; Dukes’ stage; ethnic background; family history of cancer; histology; metastasis; metastatic sites; multiple tumors; node involvement; primary site; prior screening status; prior therapy status; quality metric; relapse time; retreatment states; sex; stage; symptoms; tobacco use; tumor grade; tumor size | 20024   | Genome U133 Plus 2.0   |
| GSE2109⁶⁸  | Endometrium    | 51        | Age; alcohol consumption; days from diagnosis to excision; ethnic background; family history of cancer; histology; metastasis; metastatic sites; multiple tumors; node involvement; primary site; quality metric; stage; symptoms; tobacco use; tumor grade; tumor size | 20024   | Genome U133 Plus 2.0   |
| GSE2109⁶⁸  | Kidney         | 209       | Age; alcohol consumption; days from diagnosis to excision; ethnic background; family history of cancer; histology; metastasis; metastatic sites; multiple tumors; primary site; prior therapy status; quality metric; relapse time; retreatment states; sex; stage; symptoms; tobacco use; tumor grade; tumor size | 20024   | Genome U133 Plus 2.0   |
| ID       | Tissue        | Samples | Description                                                                                                                                                                                                 | Library       | Platform       |
|----------|---------------|---------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|----------------|
| GSE2109  | Lung          | 103     | Age; alcohol consumption; days from diagnosis to excision; diagnostic method; ethnic background; family history of cancer; Gleason score; histology; metastasis; multiple tumors; node involvement; primary site; prior therapy status; quality metric; relapse time; retreatment states; sex; stage; symptoms; tobacco use; tumor grade; tumor size | 20024         | Genome U133 Plus 2.0 |
| GSE2109  | Ovary         | 158     | Age; alcohol consumption; days from diagnosis to excision; esophagitis reflux history; ethnic background; family history of cancer; fibrocystic disease; histology; mammogram history; metastasis; metastatic sites; multiple tumors; node involvement; oophorectomy status; primary site; prior therapy status; quality metric; relapse time; retreatment states; screening history; stage; symptoms; tobacco use; tumor grade; tumor size | 20024         | Genome U133 Plus 2.0 |
| GSE2109  | Prostate      | 79      | Age; alcohol consumption; days from diagnosis to excision; diagnosis method; ethnic background; family history of cancer; Gleason score; histology; metastasis; multiple tumors; node involvement; prior therapy status; prostate-specific antigen (PSA) testing history; PSA finding; quality metric; stage; symptoms; tobacco use; tobacco use; tumor grade; tumor size | 20024         | Genome U133 Plus 2.0 |
| GSE2109  | Uterine       | 112     | Age; alcohol consumption; days from diagnosis to excision; ethnic background; family history of cancer; histology; human papilloma virus diagnosis history; metastasis; metastatic sites; multiple tumors; node involvement; primary site; prior therapy status; quality metric; relapse; stage; symptoms; tobacco use; tumor grade; tumor size | 20024         | Genome U133 Plus 2.0 |
| GSE4271  | Glial         | 100     | Age; recurrence status; sex; survival status; survival time; WHO grade                                                                                                                                     | 11832         | Genome U133A    |
| Study ID | Tissue Type                                      | Sample Size | Variables                                                                 | Additional Information   | Platform          |
|---------|-------------------------------------------------|-------------|---------------------------------------------------------------------------|--------------------------|------------------|
| GSE5460 | Breast cancer                                   | 127         | ER status; Her2 status; histological type; lymphovascular invasion; node status; tumor grade; tumor size | 20024                    | Genome U133 Plus 2.0 |
| GSE5462 | Breast cancer                                   | 52          | Treatment history; treatment response                                      | 11832                    | Genome U133A     |
| GSE6532 | Breast carcinoma                                | 317         | Age; distant metastasis-free survival time/status; ER status; genomic grade index; node involvement; PR status; recurrence-free survival time/status; tumor grade; tumor size | 11832                    | Genome U133A     |
| GSE6532 | Breast carcinoma                                | 87          | Age; distant metastasis-free survival time/status; ER status; genomic grade index; node involvement; PR status; recurrence-free survival time/status; tumor grade; tumor size | 20024                    | Genome U133 Plus 2.0 |
| GSE10320| Wilms Tumor                                     | 144         | Relapse                                                                   | 11832                    | Genome U133A     |
| GSE15296| Peripheral Blood                                | 75          | Kidney transplant rejection; subtype                                       | 20024                    | Genome U133 Plus 2.0 |
| GSE19804| Paired tumor and normal tissues                 | 60          | Age; tissue type; tumor stage                                             | 20024                    | Genome U133 Plus 2.0 |
| GSE20181| Breast cancer                                   | 50          | Treatment history; treatment response                                      | 11832                    | Genome U133A     |
| GSE20189| Lung adenocarcinoma                             | 162         | Case/control status; morphology; smoking status; stage                    | 11832                    | Genome U133A     |
| GSE21510| Laser capture microdissection and homogenized tissues (surgically resected material) | 104         | Metastasis; stage; tissue type                                            | 20024                    | Genome U133 Plus 2.0 |
| GSE25507| Peripheral blood lymphocyte                     | 146         | Case/control status (autism); paternal age, maternal age, subject age     | 20024                    | Genome U133 Plus 2.0 |
| GSE26682| Colorectal tumor                                | 140         | Age; microsatellite instability status; sex                              | 11832                    | Genome U133A     |
| GSE26682| Colorectal tumor                                | 160         | Age; microsatellite instability status; sex                              | 20024                    | Genome U133 Plus 2.0 |
| GSE27279| Posterior Fossa Ependymoma                      | 100         | Age; sex; tumor location                                                  | 16632                    | Exon 1.0 ST     |
| Experiment | Tissue Type | Sample Size | Characteristics | Description |
|------------|-------------|-------------|----------------|-------------|
| GSE27342   | Paired gastric tumor and normal tissue | 72 | Age; sex; stage; tissue type; tumor grade | 16632 Exon 1.0 ST |
| GSE27854   | Colorectal tumor | 115 | Metastasis; stage | 20024 Genome U133 Plus 2.0 |
| GSE30219   | Lung | 293 | Age; follow-up time; histology; metastasis; node involvement; relapse status; sex; survival; survival time; tumor size | 20024 Genome U133 Plus 2.0 |
| GSE30784   | Oral squamous cell carcinoma | 229 | Age; case/control status; sex | 20024 Genome U133 Plus 2.0 |
| GSE32646   | Breast | 115 | Age; ER status (IHC); Her2 status (FISH); histological grade; lymph node involvement; pathologic complete response; PR status (IHC); stage; tumor size | 20024 Genome U133 Plus 2.0 |
| GSE37147   | Bronchial sample | 238 | Age; case/control status (COPD); FEV1/FVC score/percentage; inhaled medication status; sex; smoking status; tobacco use | 21614 Gene 1.0 ST |
| GSE37199   | Blood sample | 94 | Disease stage (advanced castration resistant prostate cancer) | 20024 Genome U133 Plus 2.0 |
| GSE37745   | Non-small cell lung cancer | 196 | Adjuvant treatment status; age; histology; recurrence time/status; sex; stage; survival time/status; WHO performance status | 20024 Genome U133 Plus 2.0 |
| GSE37892   | Stage-II colon carcinoma | 130 | Age; diagnosis history; localisation; stage; time until metastasis | 20024 Genome U133 Plus 2.0 |
| GSE38958   | Peripheral blood mononuclear cell | 115 | Age; diagnosis (Idiopathic pulmonary fibrosis); ethnicity; predicted FVC percent; sex | 16632 Exon 1.0 ST |
| GSE39491   | Esophageal and gastric samples | 120 | Tumor cell type | 11832 Genome U133 Plus 2.0 |
| GSE39582   | Colon cancer | 566 | Adjuvant chemotherapy; age; BRAF mutation status; chromosome instability status; CIMP status; KRAS mutation status; mismatch repair status; overall survival time/status; recurrence-free survival time/status; sex; stage; TP53 mutation status; tumor location | 20024 Genome U133 Plus 2.0 |
| Series            | Sample          | Reason            |
|-------------------|-----------------|-------------------|
| GSE15296          | GSM382283       | Poor Quality      |
| GSE19804          | GSM494596       | Poor Quality      |
| GSE19804          | GSM494654       | Poor Quality      |

These identifiers represent data series in Gene Expression Omnibus. Some identifiers are listed multiple times; in these cases, we used a subset of the series data (for a specific tissue type or microarray platform).

Table 2: Summary of excluded samples. We excluded samples that did not pass our quality-control criteria or that appeared to be duplicated. The Gene Expression Omnibus series and sample identifiers are listed, along with the reason we excluded each sample.
| Sample Type       | Accession 1  | Accession 2       | Quality   |
|-------------------|-------------|-------------------|-----------|
| GSE19804          | GSM494657   | Poor Quality      |           |
| GSE20181          | GSM506289   | Duplicate         |           |
| GSE20181          | GSM506294   | Duplicate         |           |
| GSE20181          | GSM506304   | Duplicate         |           |
| GSE20181          | GSM125198   | Duplicate         |           |
| GSE20181          | GSM125210   | Duplicate         |           |
| GSE20181          | GSM125230   | Duplicate         |           |
| GSE2109_Breast    | GSM53059    | Duplicate         |           |
| GSE2109_Breast    | GSM53027    | Duplicate         |           |
| GSE2109_Colon     | GSM89040    | Duplicate         |           |
| GSE2109_Colon     | GSM152664   | Duplicate         |           |
| GSE2109_Colon     | GSM152632   | Duplicate         |           |
| GSE2109_Colon     | GSM179922   | Duplicate         |           |
| GSE2109_Colon     | GSM89044    | Duplicate         |           |
| GSE2109_Colon     | GSM152666   | Duplicate         |           |
| GSE2109_Colon     | GSM179820   | Duplicate         |           |
| GSE2109_Colon     | GSM179924   | Duplicate         |           |
| GSE2109_Colon     | GSM179903   | Duplicate         |           |
| GSE2109_Lung      | GSM203652   | Poor Quality      |           |
| GSE2109_Ovary     | GSM76554    | Duplicate         |           |
| GSE2109_Ovary     | GSM203725   | Duplicate         |           |
| GSE2109_Ovary     | GSM76567    | Duplicate         |           |
| GSE2109_Ovary     | GSM231913   | Duplicate         |           |
| GSE2109_Ovary     | GSM46839    | Poor Quality      |           |
| GSE2109_Prostate  | GSM179790   | Duplicate         |           |
| GSE2109_Prostate  | GSM179843   | Duplicate         |           |
| GSE2109_Prostate  | GSM179903   | Duplicate         |           |
| GSE25507 | GSM627091 | Duplicate |
| GSE25507 | GSM627087 | Duplicate |
| GSE25507 | GSM627096 | Duplicate |
| GSE25507 | GSM627078 | Duplicate |
| GSE25507 | GSM627085 | Duplicate |
| GSE25507 | GSM627196 | Duplicate |
| GSE25507 | GSM627153 | Duplicate |
| GSE25507 | GSM627180 | Duplicate |
| GSE25507 | GSM627099 | Duplicate |
| GSE25507 | GSM627115 | Duplicate |
| GSE25507 | GSM627118 | Duplicate |
| GSE25507 | GSM627124 | Duplicate |
| GSE25507 | GSM627154 | Duplicate |
| GSE25507 | GSM627204 | Duplicate |
| GSE25507 | GSM627209 | Duplicate |
| GSE25507 | GSM627215 | Duplicate |
| GSE26682 | GSM656833 | Duplicate |
| GSE26682 | GSM656770 | Duplicate |
| GSE26682_U133PLUS2 | GSM656860 | Poor Quality |
| GSE26682_U133PLUS2 | GSM656613 | Poor Quality |
| GSE26682_U133PLUS2 | GSM656839 | Poor Quality |
| GSE26682_U133PLUS2 | GSM656721 | Poor Quality |
| GSE27342 | GSM675945 | Duplicate |
| GSE27342 | GSM675947 | Duplicate |
| GSE27342 | GSM675933 | Duplicate |
| GSE27342 | GSM675935 | Duplicate |
| GSE    | GSM    | Quality   |
|--------|--------|-----------|
| GSE27342 | GSM676040 | Poor Quality |
| GSE27342 | GSM687519 | Poor Quality |
| GSE27854 | GSM687525 | Poor Quality |
| GSE30219 | GSM748210 | Duplicate |
| GSE30219 | GSM748212 | Duplicate |
| GSE30219 | GSM748218 | Duplicate |
| GSE30219 | GSM748219 | Duplicate |
| GSE30219 | GSM748255 | Poor Quality |
| GSE30219 | GSM748247 | Poor Quality |
| GSE30219 | GSM748057 | Poor Quality |
| GSE30219 | GSM748266 | Poor Quality |
| GSE30784 | GSM764928 | Duplicate |
| GSE30784 | GSM764930 | Duplicate |
| GSE30784 | GSM764904 | Poor Quality |
| GSE30784 | GSM764970 | Poor Quality |
| GSE32646 | GSM809214 | Duplicate |
| GSE32646 | GSM809248 | Duplicate |
| GSE32646 | GSM809251 | Duplicate |
| GSE32646 | GSM809254 | Duplicate |
| GSE37147 | GSM912230 | Duplicate |
| GSE37147 | GSM912296 | Duplicate |
| GSE37147 | GSM912296 | Duplicate |
| GSE37147 | GSM912305 | Duplicate |
| GSE37147 | GSM912291 | Duplicate |
| GSE37147 | GSM912296 | Duplicate |
| GSE37147 | GSM912305 | Duplicate |
| GSE37147 | GSM912342 | Duplicate |
|----------|-----------|------------|
| GSE37147 | GSM912273 | Duplicate |
| GSE37147 | GSM912305 | Duplicate |
| GSE37147 | GSM912342 | Duplicate |
| GSE37147 | GSM912342 | Duplicate |
| GSE37147 | GSM912342 | Duplicate |
| GSE37147 | GSM912376 | Duplicate |
| GSE37147 | GSM912376 | Duplicate |
| GSE37147 | GSM912463 | Poor Quality |
| GSE37147 | GSM912197 | Poor Quality |
| GSE37147 | GSM912300 | Poor Quality |
| GSE37199  | GSM913439 | Poor Quality |
| GSE37745  | GSM1019319 | Duplicate |
| GSE37745  | GSM1019246 | Duplicate |
| GSE37745  | GSM1019325 | Duplicate |
| GSE37745  | GSM1019247 | Duplicate |
| GSE37745  | GSM1019194 | Poor Quality |
| GSE37745  | GSM1019195 | Poor Quality |
| GSE37745  | GSM1019176 | Poor Quality |
| GSE37745  | GSM1019192 | Poor Quality |
| GSE37745  | GSM1019232 | Poor Quality |
| GSE37892  | GSM929512 | Poor Quality |
| GSE39491  | GSM970152 | Poor Quality |
| GSE39582  | GSM972249 | Duplicate |
| GSE39582  | GSM972472 | Duplicate |
| GSE   | GSM    | Status  |
|-------|--------|---------|
| GSE39582 | GSM972243 | Duplicate |
| GSE39582 | GSM972044 | Duplicate |
| GSE39582 | GSM972091 | Duplicate |
| GSE39582 | GSM972090 | Duplicate |
| GSE39582 | GSM972245 | Duplicate |
| GSE39582 | GSM972473 | Duplicate |
| GSE39582 | GSM972515 | Duplicate |
| GSE39582 | GSM972248 | Duplicate |
| GSE43176 | GSM1057835 | Poor Quality |
| GSE46449 | GSM1130404 | Duplicate |
| GSE46449 | GSM1130406 | Duplicate |
| GSE46449 | GSM1130413 | Duplicate |
| GSE46449 | GSM1130417 | Duplicate |
| GSE46449 | GSM1130426 | Duplicate |
| GSE46449 | GSM1130428 | Duplicate |
| GSE46449 | GSM1130430 | Duplicate |
| GSE46449 | GSM1130434 | Duplicate |
| GSE46449 | GSM1130436 | Duplicate |
| GSE46449 | GSM1130468 | Duplicate |
| GSE46449 | GSM1130471 | Duplicate |
| GSE46449 | GSM1130483 | Duplicate |
| GSE48390 | GSM1176924 | Poor Quality |
| GSE48390 | GSM125120 | Poor Quality |
| GSE5462 | GSM125123 | Duplicate |
| GSE5462 | GSM125125 | Duplicate |
| GSE58697 | GSM1417097 | Poor Quality |
| GSE63885 | GSM1559328 | Duplicate |
|----------|------------|-----------|
| GSE63885 | GSM1559360 | Duplicate |
| GSE63885 | GSM1559385 | Duplicate |
| GSE63885 | GSM1559370 | Duplicate |
| GSE63885 | GSM1559375 | Duplicate |
| GSE63885 | GSM1559386 | Duplicate |
| GSE63885 | GSM1559361 | Poor Quality |
| GSE6532_U133PLUS2 | GSM151294 | Poor Quality |
| GSE6532_U133PLUS2 | GSM151280 | Poor Quality |

**References**

1. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17**, 669–681 (2007).

2. Alberts, B. *Molecular Biology of the Cell: Reference edition.* (Garland Science, 2008).

3. Butte, A. The use and analysis of microarray data. *Nat. Rev. Drug Discov.* **1**, 951–960 (2002).

4. Piccolo, S. R. & Frey, L. J. Clinical and molecular models of glioblastoma multiforme survival. *Int. J. Data Min. Bioinform.* **7**, 245–265 (2013).

5. Piccolo, S. R. *et al.* Gene-expression patterns in peripheral blood classify familial breast cancer susceptibility. *BMC Med. Genomics* **8**, 72 (2015).

6. Beane, J. *et al.* Characterizing the Impact of Smoking and Lung Cancer on the Airway Transcriptome Using RNA-Seq. *Cancer Prev. Res.* **4**, 803–817 (2011).

7. Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* **3**, 111ra121 (2011).
8. Byers, L. A. et al. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin. Cancer Res. 19, 279–290 (2013).

9. Adib, T. R. et al. Predicting biomarkers for ovarian cancer using gene-expression microarrays. Br. J. Cancer 90, 686–692

10. Sirotta, M. et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci. Transl. Med. 3, 96ra77 (2011).

11. Tofigh, A. et al. The prognostic ease and difficulty of invasive breast carcinoma. Cell Rep. 9, 129–142 (2014).

12. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. 39, D1005–10 (2011).

13. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23, 1846–1847 (2007).

14. Dumas, J., Gargano, M. A. & Dancik, G. M. shinyGEO: a web-based application for analyzing gene expression omnibus datasets. Bioinformatics 32, 3679–3681 (2016).

15. Gentleman, R., Ihaka, R., Bates, D. & Others. The R project for statistical computing. R home web site: http://www.r-project.org (1997).

16. Bischl, B. et al. mlr: Machine Learning in R. J. Mach. Learn. Res. 17, 1–5 (2016).

17. Piccolo, S. R. et al. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics 100, 337–344 (2012).

18. Dai, M. et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 33, e175 (2005).

19. Wickham, H. Tidy Data. J. Stat. Softw. 59, (2014).

20. 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* **30**, 1392–1399 (2014).

21. CRAN - Package outliers. Available at: https://CRAN.R-project.org/package=outliers. (Accessed: 14th September 2017)

22. 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).

23. 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).

24. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

25. 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).

26. International Genomics Consortium. Expression Project for Oncology. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2109. (Accessed: 28th July 2017)

27. Phillips, H. S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157–173 (2006).

28. Costa, B. M. *et al.* Reversing HOXA9 Oncogene Activation by PI3K Inhibition: Epigenetic Mechanism and Prognostic Significance in Human Glioblastoma. *Cancer Res.* **70**, 453–462 (2010).

29. Lu, X. *et al.* Predicting features of breast cancer with gene expression patterns. *Breast Cancer Res. Treat.* **108**, 191–201 (2008).
30. Miller, W. R. et al. Changes in breast cancer transcriptional profiles after treatment with the aromatase inhibitor, letrozole. *Pharmacogenet. Genomics* **17**, 813–826 (2007).

31. 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 Res.* **12**, R52 (2010).

32. 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).

33. Huang, C.-C. et al. Predicting relapse in favorable histology Wilms tumor using gene expression analysis: a report from the Renal Tumor Committee of the Children’s Oncology Group. *Clin. Cancer Res.* **15**, 1770–1778 (2009).

34. Kurian, S. M. et al. Molecular classifiers for acute kidney transplant rejection in peripheral blood by whole genome gene expression profiling. *Am. J. Transplant* **14**, 1164–1172 (2014).

35. Lu, T.-P. et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol. Biomarkers Prev.* **19**, 2590–2597 (2010).

36. 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).

37. Rotunno, M. et al. A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. *Cancer Prev. Res.* **4**, 1599–1608 (2011).

38. Tsukamoto, S. et al. Clinical significance of osteoprotegerin expression in human colorectal cancer. *Clin. Cancer Res.* **17**, 2444–2450 (2011).

39. Alter, M. D. et al. Autism and increased paternal age related changes in global levels of
gene expression regulation. *PLoS One* **6**, e16715 (2011).

40. Vilar, E. *et al.* MRE11 deficiency increases sensitivity to poly(ADP-ribose) polymerase inhibition in microsatellite unstable colorectal cancers. *Cancer Res.* **71**, 2632–2642 (2011).

41. Sanz-Pamplona, R. *et al.* Gene expression differences between colon and rectum tumors. *Clin. Cancer Res.* **17**, 7303–7312 (2011).

42. Schmit, S. L. *et al.* MicroRNA polymorphisms and risk of colorectal cancer. *Cancer Epidemiol. Biomarkers Prev.* **24**, 65–72 (2015).

43. Witt, H. *et al.* Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. *Cancer Cell* **20**, 143–157 (2011).

44. Cui, J. *et al.* An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. *Nucleic Acids Res.* **39**, 1197–1207 (2011).

45. Cui, J. *et al.* Gene-expression signatures can distinguish gastric cancer grades and stages. *PLoS One* **6**, e17819 (2011).

46. Kikuchi, A. *et al.* Identification of NUCKS1 as a colorectal cancer prognostic marker through integrated expression and copy number analysis. *Int. J. Cancer* **132**, 2295–2302 (2013).

47. Rousseaux, S. *et al.* Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.* **5**, 186ra66 (2013).

48. Chen, C. *et al.* Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiol. Biomarkers Prev.* **17**, 2152–2162 (2008).

49. Miyake, T. *et al.* GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. *Cancer Sci.* **103**, 913–920 (2012).

50. Steiling, K. *et al.* A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. *Am. J. Respir. Crit. Care Med.*
187, 933–942 (2013).

51. Olmos, D. et al. Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. *Lancet Oncol.* **13**, 1114–1124 (2012).

52. Botling, J. et al. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin. Cancer Res.* **19**, 194–204 (2013).

53. Laihe, S. et al. A seven-gene signature aggregates a subgroup of stage II colon cancers with stage III. *OMICS* **16**, 560–565 (2012).

54. Huang, L. S. et al. Sphingosine-1-phosphate lyase is an endogenous suppressor of pulmonary fibrosis: role of S1P signalling and autophagy. *Thorax* **70**, 1138–1148 (2015).

55. Hyland, P. L. et al. Global changes in gene expression of Barrett’s esophagus compared to normal squamous esophagus and gastric cardia tissues. *PLoS One* **9**, e93219 (2014).

56. Marisa, L. et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).

57. Kabakchiev, B. & Silverberg, M. S. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology* **144**, 1488–96, 1496.e1–3 (2013).

58. Xu, J. et al. Dominant role of oncogene dosage and absence of tumor suppressor activity in Nras-driven hematopoietic transformation. *Cancer Discov.* **3**, 993–1001 (2013).

59. Clelland, C. L. et al. Utilization of never-medicated bipolar disorder patients towards development and validation of a peripheral biomarker profile. *PLoS One* **8**, e69082 (2013).

60. Zhao, S. G. et al. The Landscape of Prognostic Outlier Genes in High-Risk Prostate Cancer. *Clin. Cancer Res.* **22**, 1777–1786 (2016).
61. Bessho, K. *et al.* Gene expression signature for biliary atresia and a role for interleukin-8 in pathogenesis of experimental disease. *Hepatology* **60**, 211–223 (2014).

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

63. Salas, S. *et al.* Gene Expression Profiling of Desmoid Tumors by cDNA Microarrays and Correlation with Progression-Free Survival. *Clin. Cancer Res.* **21**, 4194–4200 (2015).

64. Lisowska, K. M. *et al.* Gene expression analysis in ovarian cancer - faults and hints from DNA microarray study. *Front. Oncol.* **4**, 6 (2014).

65. Kurian, S. M. *et al.* Peripheral Blood Cell Gene Expression Diagnostic for Identifying Symptomatic Transthyretin Amyloidosis Patients: Male and Female Specific Signatures. *Theranostics* **6**, 1792–1809 (2016).