SCIENTIFIC DATA

OPEN

Data Descriptor: Curated compendium of human transcriptional biomarker data

Nathan P. Golightly[1], Avery Bell[1], Anna I. Bischoff[1], Parker D. Hollingsworth[1,2] & Stephen R. Piccolo[1,3]

Received: 20 September 2017
Accepted: 22 February 2018
Published: 17 April 2018

One important use of genome-wide transcriptional profiles is to identify relationships between transcription levels and patient outcomes. These translational insights can guide the development of biomarkers for clinical application. Data from thousands of translational-biomarker studies have been deposited in public repositories, enabling reuse. 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 45 publicly available, translational-biomarker datasets from a variety of human diseases. To increase the data's utility, we reprocessed the raw expression data using a uniform computational pipeline, addressed quality-control problems, mapped the clinical annotations to a controlled vocabulary, and prepared consistently structured, analysis-ready data files. These data, along with scripts we used to prepare the data, are available in a public repository. We believe these data will be particularly useful to researchers seeking to perform benchmarking studies—for example, to compare and optimize machine-learning algorithms' ability to predict biomedical outcomes.

Design Type(s)	data integration objective
Measurement Type(s)	Biomarker
Technology Type(s)	digital curation
Factor Type(s)	tissue • diagnosis • microarray
Sample Characteristic(s)	Homo sapiens • breast • colon • uterine endometrium • kidney • lung • ovary • prostate gland • uterus • glia • peripheral blood • peripheral blood lymphocyte • colorectum • ependymal epithelium • squamous epithelium • peripheral blood mononuclear cell • alimentary canal • blast cell • leukocyte • connective tissue

[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. Correspondence and requests for materials should be addressed to S.R.P. (email: stephen_piccolo@byu.edu).

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 in different quantities 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[4–10]. Indeed, already to date, more than 100 transcriptional biomarkers have been proposed for predicting breast-cancer survival alone[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[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[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 *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. For this version of the data, we one-hot encoded any discrete values and imputed any missing values.

Methods

Selecting data

To select datasets to be included in our compendium, we performed 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, Data Citation 1) 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

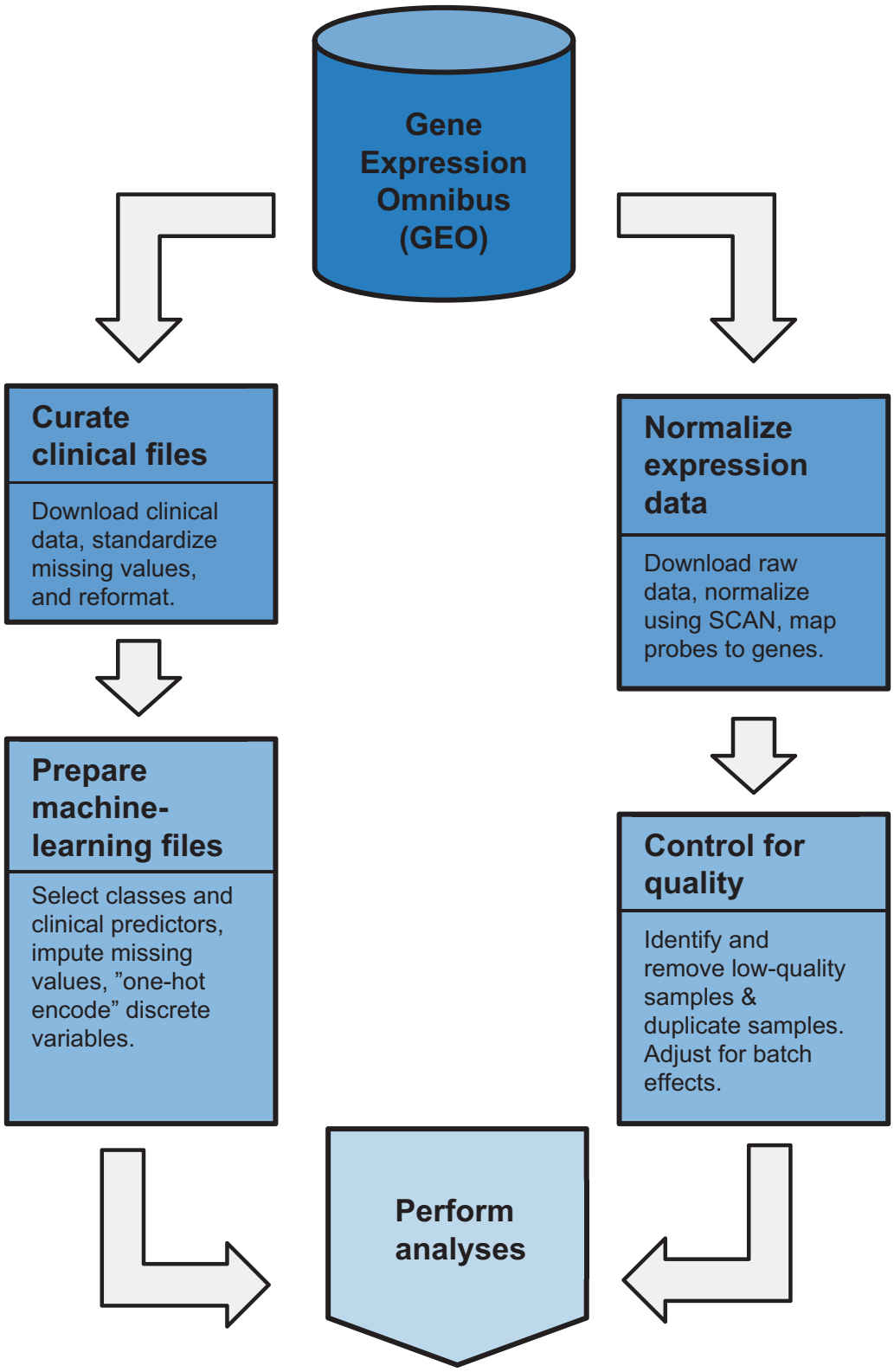

Figure 1. Flow diagram that illustrates the process we used to collect and curate the data. We wrote computer scripts that downloaded the data, checked for quality, normalized and standardized data values, and stored the data in analysis-ready file formats. The specific steps differed for clinical and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

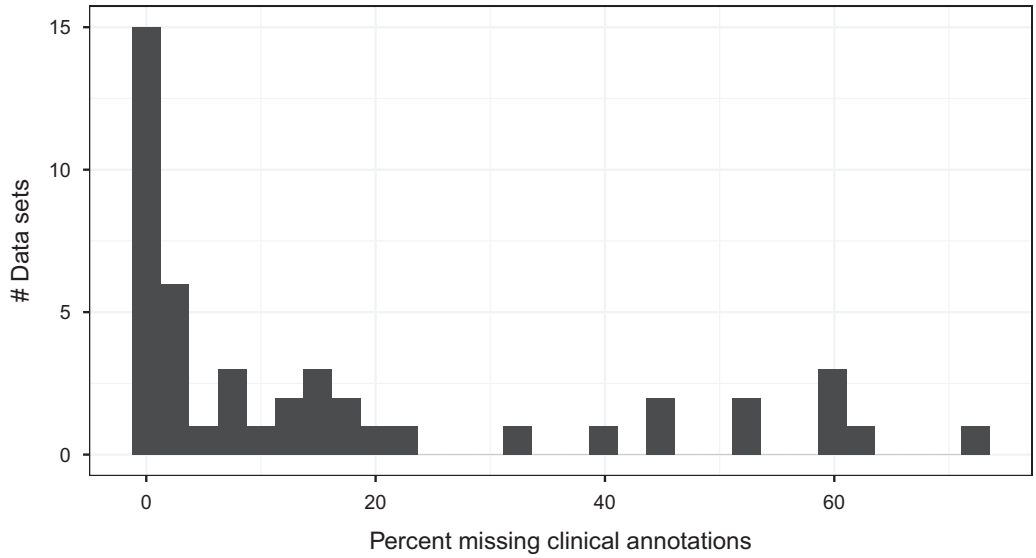

Figure 2. Histogram showing the proportion of missing clinical-annotation values per dataset. Some datasets contained no missing values, while others were missing as many as as 72.3% of data values.

platform (e.g., GSE6532_U133A and GSE6532_U133Plus2). For both of these series, 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[15]. To avoid confounding effects due to tissue-specific expression and because the metadata differed considerably across the cancer types, we split this dataset into multiple datasets based on cancer type (Table 1 (available online only)). 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[16] that download, parse, and reformat the clinical annotations. Initially, these scripts retrieve data from GEO 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 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 (Data Citation 1), one patient's clinical demographics and treatment responses are listed as "female; breast tumor; Letrozole, 2.5 mg/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 an average of 7.8 variables of metadata (Table 1 (available online only)). 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", "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 (although they can be used in many other types of analysis). Accordingly, we prepared a secondary version 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; but we also included class variables that had not been used in the original studies. On average, the datasets contain 2.8 class variables. Second, we identified clinical variables that could be used as predictor variables (covariates). Using these data, we generated one "Analysis" file per class variable that contains the class values for each sample as well as covariates that we suggest are relevant to the class variable. (A given variable may be used as a class variable in one context and a predictor variable in a different context.) We named these analysis files using descriptive prefixes (e.g., "Prognosis", "Diagnosis", or "Stage"). In addition, we identified concepts in

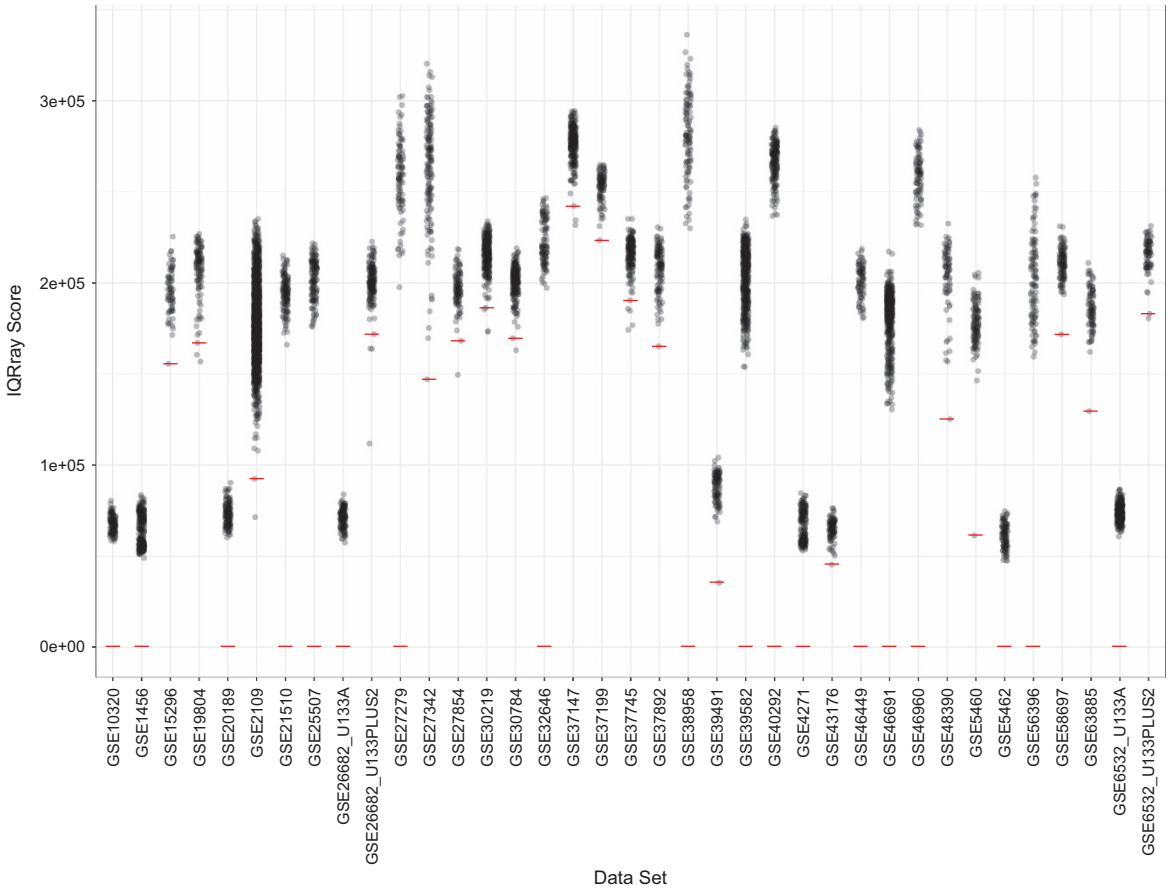

Figure 3. Distribution of IQRay quality scores for each dataset. Sample qualities are plotted for each dataset. Low-quality samples were identified using Grubb's test. Samples that fall on or below the red threshold were excluded from the data repository.

the National Cancer Institute Thesaurus[17] that map to each class and covariate variable. The name of each analysis file indicates the thesaurus term (preferred name) that corresponds to the class variable for that file. Within these files, the column names indicate the thesaurus terms that correspond to each covariate. We hope the use of this controlled vocabulary will make it easier for others to better understand the semantic meaning of these variables and identify commonalities across datasets. A tab-separated file that indicates mappings between the original annotation terms and the thesaurus terms can be found in our data repository (see https://osf.io/szwx6/).

When a given sample was missing data for a given class variable, we excluded that sample from the respective analysis file for that class variable. After this filtering step, we identified class variables with fewer than 40 samples and excluded these class variables. When covariates were missing more than 20% data (Figure 2), we excluded these variables from the analysis files. When covariates were missing less than 20% data, we imputed missing values using median-based imputation for continuous variables and mode-based imputation for discrete variables[18]. 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 (Data Citation 1), 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 into binary variables. Finally, some clinical outcomes were survival or relapse times; we transformed these data to (discrete) class variables, using dataset-specific thresholds 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[19]. 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 Ensembl-based gene-level annotations from *Brainarray*[20,21]. The SCAN algorithm log2-transforms the data and scales the data to center around zero. Relatively high values indicate relatively high gene-expression levels, and vice versa.

Code availability

Our computer scripts are stored in the open-access *Biomarker Benchmark* repository (Data Citation 1). Using these scripts, other researchers can reproduce our curation process and/or produce alternative versions of the data.

Data Records

After we filtered the original data (see Methods), our compendium includes data for 7,037 biological samples across 45 datasets (Table 1 (available online only)). On average, the datasets contain values for 18,043 genes (Table 1 (available online only)). In total, our repository contains 128 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[22]. 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 (Data Citation 1). The original data files are available via Gene Expression Omnibus using the accession numbers listed in Table 1 (available online only).

Technical Validation

We evaluated each sample using the *IQRray*[23] software, which produces a quality score for individual samples. Using these metrics, we applied Grubb's statistical test (*outliers* package[24]) to each dataset, identified poor-quality outliers (Figure 3), and excluded these samples (Table 2 (available online only)). Next we used the *DoppelgangR* package[25] 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 (available online only)), even if the clinical annotations for both samples were distinct, under the assumption that these samples had somehow been mislabeled. In GSE46449 (Data Citation 1), many samples were biological replicates; we retained one of each replicate set. GSE5462, GSE19804, and GSE20181 (Data Citation 1) contained samples that had been profiled in a paired manner (e.g., pre- and post-treatment); we retained these pairs of samples.

When transcriptomic data are processed in multiple batches, batch assignments can lead to confounding effects[26]. In the clinical annotations, we identified batch-processing information for datasets GSE25507, GSE37199, GSE39582, and GSE40292 (Data Citation 1). We corrected for batch effects using the ComBat software[27]. 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").

Machine-learning analysis

We created a document that illustrates how to programmatically download the data files and perform a simple classification analysis using our data (see https://osf.io/4n62k/). This document is coded for the R statistical package, but similar analyses could be performed using other programming languages.

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,** 111ra–121r (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 (2004).
10. Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med* **3,** 96ra–77 (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–D1010 (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. International Genomics Consortium. Expression Project for Oncology. *Gene Expression Omnibus* https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2109 (2017).

16. Gentleman, R., Ihaka, R. & Bates, D & Others. The R project for statistical computing. *R home web site* http://www.r-project.org (1997).
17. Sioutos, N. *et al.* NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* **40,** 30–43 (2007).
18. Bischl, B. *et al.* mlr: Machine Learning in R. *J. Mach. Learn. Res.* **17,** 1–5 (2016).
19. Piccolo, S. R. *et al.* A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* **100,** 337–344 (2012).
20. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** e175 (2005).
21. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
22. Wickham, H Tidy Data. *J. Stat. Softw.* **59** (2014).
23. 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).
24. Komsta, L. Package outliers. CRAN https://CRAN.R-project.org/package = outliers (2017).
25. 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).
26. 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).
27. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).
28. 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–R964 (2005).
29. 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).
30. 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).
31. Lu, X. *et al.* Predicting features of breast cancer with gene expression patterns. *Breast Cancer Res. Treat.* **108,** 191–201 (2008).
32. Miller, W. R. *et al.* Changes in breast cancer transcriptional profiles after treatment with the aromatase inhibitor, letrozole. *Pharmacogenet. Genomics* **17,** 813–826 (2007).
33. 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).
34. 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).
35. 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).
36. 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).
37. 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).
38. 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).
39. Rotunno, M. *et al.* A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. *Cancer Prev. Res* **4,** 1599–1608 (2011).
40. Tsukamoto, S. *et al.* Clinical significance of osteoprotegerin expression in human colorectal cancer. *Clin. Cancer Res.* **17,** 2444–2450 (2011).
41. Alter, M. D. *et al.* Autism and increased paternal age related changes in global levels of gene expression regulation. *PLoS ONE* **6,** e16715 (2011).
42. 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).
43. Sanz-Pamplona, R. *et al.* Gene expression differences between colon and rectum tumors. *Clin. Cancer Res.* **17,** 7303–7312 (2011).
44. Schmit, S. L. *et al.* MicroRNA polymorphisms and risk of colorectal cancer. *Cancer Epidemiol. Biomarkers Prev* **24,** 65–72 (2015).
45. Witt, H. *et al.* Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. *Cancer Cell* **20,** 143–157 (2011).
46. Cui, J. *et al.* An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. *Nucleic Acids Res.* **39,** 1197–1207 (2011).
47. Cui, J. *et al.* Gene-expression signatures can distinguish gastric cancer grades and stages. *PLoS One* **6,** e17819 (2011).
48. 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).
49. Rousseaux, S. *et al.* Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med* **5,** 186ra–66 (2013).
50. Chen, C. *et al.* Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiol. Biomarkers Prev* **17,** 2152–2162 (2008).
51. 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).
52. 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).
53. 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).
54. 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).
55. Laibe, S. *et al.* A seven-gene signature aggregates a subgroup of stage II colon cancers with stage III. *OMICS* **16,** 560–565 (2012).
56. 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).
57. 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).
58. Marisa, L *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* **10,** e1001453 (2013).
59. Kabakchiev, B. & Silverberg, M. S. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology* **144,** 1488–14961496, e1–3 (2013).

60. 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).
61. 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).
62. Zhao, S. G. *et al.* The Landscape of Prognostic Outlier Genes in High-Risk Prostate Cancer. *Clin. Cancer Res.* **22,** 1777–1786 (2016).
63. 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).
64. Huang, C.-C. *et al.* Concurrent gene signatures for han chinese breast cancers. *PLoS ONE* **8,** e76421 (2013).
65. 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).
66. Lisowska, K. M. *et al.* Gene expression analysis in ovarian cancer - faults and hints from DNA microarray study. *Front. Oncol* **4,** 6 (2014).
67. 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).

Data Citation

1. Piccolo, S, Golightly, N, Bischoff, A & Bell, A. *Open Science Framework* http://doi.org/10.17605/OSF.IO/SSK3T (2018).

Acknowledgements

S.R.P. thanks Brigham Young University for research funds used in this study. A.I.B. and P.D.H. thank the BYU Office of Research and Creative Activities for research funds that supported this work. N.P.G. 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 made these studies possible.

Author Contributions

N.P.G.: Collected data, wrote computer scripts, evaluated data quality, prepared figures and tables, wrote the manuscript. A.B.: Wrote computer scripts, wrote the manuscript, prepared figures and scripts, edited the manuscript. A.I.B.: Collected data, wrote computer scripts, evaluated data quality, edited the manuscript. P.D.H.: Collected data, wrote computer scripts, edited the manuscript. S.R.P.: Collected data, wrote computer scripts, prepared figures and tables, wrote the manuscript.

Additional information

Tables 1 and 2 are only available in the online version of this paper.

Competing interests: The authors declare no competing interests.

How to cite this article: Golightly N. P. *et al.* Curated compendium of human transcriptional biomarker data. *Sci. Data* 5:180066 doi: 10.1038/sdata.2018.66 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.