Data descriptors to enhance utility and utilization of data sets

Stefan Wiemann  
Division Molecular Genome Analysis  
German Cancer Research Center  
Member Editorial Board, Scientific Data

Andrew L. Hufton  
Managing Editor, Scientific Data  
Nature Publishing Group  
andrew.hufton@nature.com

Helmholtz Open Science Webinars on Research Data  
Webinar 35 – 3 / 9 May 2016
It‘s about sharing, utility & utilization of research data (in this webinar)
The current situation

- Most researchers are sharing data, and using the data of others
- Direct contact between researchers (on request) is a common way of sharing data
- Repositories are second most common method of sharing, followed by papers (supplements)

Comprehensive molecular portraits of human breast tumours

The Cancer Genome Atlas Network

We analysed primary breast cancers by genomic DNA copy number arrays, DNA methylation, exome sequencing, messenger RNA arrays, microRNA sequencing and reverse-phase protein arrays. Our ability to integrate information across platforms provided key insights into previously defined gene expression subtypes and demonstrated the existence of four main breast cancer classes when combining data from five platforms, each of which shows significant molecular heterogeneity. Somatic mutations in only three genes (TP53, PIK3CA and GATA3) occurred at >10% incidence across all breast cancers; however, there were numerous subtype-associated and novel gene mutations including the enrichment of specific mutations in GATA3, PIK3CA and MAP3K1 with the luminal A subtype. We identified two novel protein-expression-defined subgroups, possibly produced by stromal/microenvironmental elements, and integrated analyses identified specific signalling pathways dominant in each molecular subtype including a HER2/phosphorylated EGFR signature within the HER2-enriched expression subtype. Comparison of basal-like breast tumours with high-grade serous ovarian tumours showed many molecular commonalities, indicating a related aetiology and similar therapeutic opportunities. The biological finding of the four main breast cancer subtypes caused by different subsets of genetic and epigenetic abnormalities raises the hypothesis that much of the clinically observable plasticity and heterogeneity occurs within, and not across, these major biological subtypes of breast cancer.

Supplementary Information is available in the online version of the paper.

1. Supplementary Information (14.1M)

This file contains Supplementary Figures 1-20, Supplementary Methods 1-15 (with additional figures and tables) and Supplementary References.

1. Supplementary Tables (1M)

This zipped file contains Supplementary Tables 1-8. *This file was replaced on 15 November 2012 to correct an error in Supplementary Table 5.*
### Part of supplementary Table 1 – patient data

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
| Complete TCGA ID | Gender | Age at Initial Pathology | ER Status | PR Status | HER2 Final Status | Tumor | Tumor-IT Coded | Node | Node-Coded | Metastasis | Metastasis-Coded | AJCC Stage | Converted Stage | Survival Data Form | Vital Status | Days to Date of Last Contact | Days to Date of Death | OS Event | OS Time | PAM50 mRNA | Sign Unfolded |
| TCGA-A2-AM? | FEMALE | 56 | Negative | Negative | Negative | T3 | T_Other | N3 | Positive | M1 | Positive | Stage IV | No Conversion | follow-up | DECEASED | 790 | 1 | 790 | Basa-Like |
| TCGA-A2-AN? | FEMALE | 54 | Negative | Negative | Negative | T2 | T_Other | N3 | Positive | M0 | Negative | Stage IIC | No Conversion | follow-up | DECEASED | 547 | 1 | 547 | Basa-Like |
| TCGA-A2-AM? | FEMALE | 56 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | No Conversion | follow-up | DECEASED | 534 | 1 | 534 | Basa-Like |
| TCGA-A2-AH? | FEMALE | 50 | Negative | Negative | Negative | T1 | T1 | N2 | Positive | M0 | Negative | Stage IIIA | No Conversion | follow-up | DECEASED | 764 | 1 | 764 | Basa-Like |
| TCGA-A2-AV? | FEMALE | 50 | Negative | Negative | Negative | T3 | T1 | M1 | Positive | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 563 | 1 | 563 | Basa-Like |
| TCGA-BH-AF? | FEMALE | 50 | Negative | Intermediate | Intermediate | T3 | T1 | M1 | Positive | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 591 | 1 | 591 | Basa-Like |
| TCGA-BH-AE? | FEMALE | 40 | Negative | Intermediate | Intermediate | T3 | T1 | M1 | Positive | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 591 | 1 | 591 | Basa-Like |
| TCGA-BH-BF? | FEMALE | 40 | Negative | Intermediate | Intermediate | T2 | T1 | M1 | Positive | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 555 | 1 | 555 | Basa-Like |
| TCGA-BH-AF? | FEMALE | 56 | Positive | Positive | Positive | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 153 | 0 | 153 | Basa-Like |
| TCGA-BH-AE? | FEMALE | 56 | Positive | Positive | Positive | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 170 | 0 | 170 | Basa-Like |
| TCGA-BH-BF? | FEMALE | 40 | Negative | Negative | Negative | T3 | T1 | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 257 | 0 | 257 | Basa-Like |
| TCGA-BH-AE? | FEMALE | 56 | Positive | Positive | Positive | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 257 | 0 | 257 | Basa-Like |
| TCGA-BH-BE? | FEMALE | 56 | Positive | Positive | Positive | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 257 | 0 | 257 | Basa-Like |
| TCGA-BH-CF? | FEMALE | 38 | Negative | Intermediate | Intermediate | T3 | T1 | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 257 | 0 | 257 | Basa-Like |
| TCGA-BH-AD? | FEMALE | 59 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | No Conversion | follow-up | DECEASED | 303 | 0 | 303 | Basa-Like |
| TCGA-BH-CF? | FEMALE | 52 | Positive | Positive | Positive | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 303 | 0 | 303 | Basa-Like |
| TCGA-BH-BF? | FEMALE | 74 | Negative | Negative | Negative | T4 | T1 | N0 | Negative | M0 | Negative | Stage IB | No Conversion | follow-up | DECEASED | 313 | 0 | 313 | Basa-Like |
| TCGA-BH-CB? | FEMALE | 51 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | No Conversion | follow-up | DECEASED | 425 | 0 | 425 | Basa-Like |
| TCGA-BH-AM? | FEMALE | 45 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | No Conversion | follow-up | DECEASED | 431 | 0 | 431 | Basa-Like |
| TCGA-A2-AM? | FEMALE | 59 | Negative | Negative | Negative | T2 | T1 | N0 | Negative | M0 | Negative | Stage IB | No Conversion | follow-up | DECEASED | 513 | 0 | 513 | Basa-Like |
| TCGA-A2-AF? | FEMALE | 50 | Negative | Negative | Negative | T3 | T_Other | N1 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 513 | 0 | 513 | Basa-Like |
| TCGA-A2-AE? | FEMALE | 48 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 513 | 0 | 513 | Basa-Like |
| TCGA-A2-AD? | FEMALE | 58 | Negative | Negative | Negative | T3 | T_Other | N2 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 513 | 0 | 513 | Basa-Like |
| TCGA-A2-AM? | FEMALE | 47 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 513 | 0 | 513 | Basa-Like |
| TCGA-A2-AF? | FEMALE | 61 | Negative | Negative | Negative | T2 | T1 | N0 | Negative | M0 | Negative | Stage IIA | No Conversion | follow-up | DECEASED | 553 | 0 | 553 | Basa-Like |
| TCGA-A2-AD? | FEMALE | 67 | Negative | Negative | Negative | T3 | T_Other | N1 | Negative | M0 | Negative | Stage IA | No Conversion | follow-up | DECEASED | 553 | 0 | 553 | Basa-Like |
| TCGA-A2-AE? | FEMALE | 53 | Negative | Negative | Negative | T3 | T1 | N0 | Negative | M0 | Negative | Stage IIA | No Conversion | follow-up | DECEASED | 553 | 0 | 553 | Basa-Like |
| TCGA-A2-AM? | FEMALE | 52 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | No Conversion | follow-up | DECEASED | 553 | 0 | 553 | Basa-Like |

---

A need of proper description/annotation of that data to facilitate re-use!
How do you make your data useful?

Open data is about more than disclosure – it must be “FAIR”

- Findable
- Accessible
- Interoperable
- Re-usable

Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship
Scientific Data 3, Article number: 160018 (2016) http://dx.doi.org/10.1038/sdata.2016.18
the data paper

– a link between original paper and the data

i.e., annotation of data set(s)
Launched in May 2014

Featured Data Descriptor

Systematic global assessment of reef fish communities by the Reef Life Survey program
Graham J. Edgar and Rick D. Stuart-Smith
27 May 2014 | doi: 10.1038/sdata.2014.7

Founded in 2007, the Reef Life Survey uses volunteer divers to assess biodiversity on ocean reefs around the world. Here, the authors release and describe the data collected by this project in detail, opening up this important citizen-science dataset to the wider scientific community.

Latest content

Data Descriptor | 27 May 2014
microclim: Global estimates of hourly microclimate based on long-term monthly climate averages
Kearney et al.

Data Descriptor | 27 May 2014
A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie
Hanko et al.

Data Descriptor | 27 May 2014
miRNA expression atlas in male rat
Minami et al.

Data Descriptor | 27 May 2014
Time-resolved gene expression profiling during reprogramming of C/EBPα-pulsed B

About Scientific Data

*Scientific Data* is an open-access, peer-reviewed publication for descriptions of scientifically valuable datasets. Our primary article-type, the Data Descriptor, is designed to make your data more discoverable, interpretable and reusable.

E-alert RSS
Facebook Twitter

Submit manuscript

nature OUTLOOK

Produced with support from Otsuka Pharmaceutical Development and Commercialization, Inc.

Announcements

Scientific Data Updates
Get Credit for Sharing Your Data
Publications will be indexed and citeable.

Open-access
Articles are published by default under a Creative Commons Attribution licence (CC BY). Each publication supported by CC0 metadata.

Focused on Data Reuse
All the information others need to reuse the data; no interpretative analysis, or hypothesis testing

Peer-reviewed
Rigorous peer-review focused on technical data quality and reuse value

Promoting Community Data Repositories
Not a new data repository; data stored in community data repositories
The “Data Descriptor” article-type

Does not contain tests of new scientific hypotheses
(no Results, no Discussion)

Sections:
• Title
• Abstract
• Background & Summary
• Methods
• Data Records
• Technical Validation
• Usage Notes
• Figures & Tables
• References
• Data Citations

All articles supported by machine-readable metadata in the ISA-tab format
An open access pilot freely sharing cancer genomic data from participants in Texas


Affiliations | Contributions | Corresponding author

*Scientific Data* 3, Article number: 160010 (2016)  doi:10.1038/sdata.2016.10
Data descriptors to increase utility of data

An open access pilot freely sharing cancer genomic data from participants in Texas

Abstract

Genomic data sharing in cancer has been restricted to aggregate or controlled-access initiatives to protect the privacy of research participants. By limiting access to these data, it has been argued that the autonomy of individuals who decide to participate in data sharing efforts has been superseded and the utility of the data as research and educational tools reduced. In a pilot Open Access (OA) project from the CPRIT-funded Texas Cancer Research Biobank, many Texas cancer patients were willing to openly share genomic data from tumor and normal matched pair specimens. For the first time, genetic data from 7 human cancer cases with matched normal are freely available without requirement for data use agreements nor any major restrictions except that end users cannot attempt to re-identify the participants (http://txcrb.org/open.html).
Access to the data

The Texas Cancer Research Biobank (TCRB) was created to bridge the gap between doctors and scientific researchers to improve the prevention, diagnosis and treatment of cancer. This work occurred with funding from the Cancer Prevention & Research Institute of Texas (CPRIT) from 2010-2014.

Access the Clinical Data Annotations by Specimen Label

The table below contains a list of the data available through the BCM-HGSC SFTP server.

To access this data, you must first register for an account and verify that you have read the Conditions for Data Use.

Register for SFTP account

Once you have registered, you can download the data through the web interface or SFTP. Please refer to the download instructions for more information.

<table>
<thead>
<tr>
<th>Case #</th>
<th>Sex/ Age/ Race/ Ethnicity</th>
<th>Prior treatment</th>
<th>Tumor % cellularity/ TNM</th>
<th>Disease Morph./ Anatomic Site</th>
<th>Tumor Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M/ 51-60/ White/ Not Hispanic or Latino</td>
<td>No</td>
<td>10%/ T3 N1 M0</td>
<td>8500/3: infiltrating duct adenocarcinoma/ Head of pancreas</td>
<td>II</td>
</tr>
</tbody>
</table>
There is room for detailed Methods

Methods

Obtaining informed consent
All work was carried out as part of an IRB-approved protocol (BCM IRB H-32711), which utilized a main consent document for general participation in TCRB with an opt-in consent addendum for OA data release. Of 194 TCRB participants offered the option of signing the opt-in addendum participating in OA sharing out of >2,500 total participants, more than half agreed to open access data sharing at time of consent. Annotated TCRB specimen and data collection consent and the OA opt in consent documents are available at http://txcrb.org/resources.html. To address concerns about whether patients can provide truly informed consent regarding the potential risks of genomic data sharing, a subset of the OA participants (n=37) were educated on risks and societal benefits of data sharing. The educational materials are available at http://txcrb.org/privacy.html. Participants were surveyed to assess their comprehension, risk tolerance, and subjective comfort with OA data release. Each participant was again queried, post-survey, to reconfirm their choice to take part in the OA data sharing option. The majority demonstrated adequate understanding of the possible privacy and discrimination risks, yet still elected to allow their data to be openly shared. The work described in Pereira et al. ⁹ is one clear example that many, though not all, cancer patients indeed desire to participate in activities that could have broad-reaching, positive impacts to public health for reducing cancer mortality and morbidity, and have the capability to make an informed choice.
Data sets are comprehensively described

Data Records

Abstract • Background & Summary • Methods • Data Records • Technical Validation • Usage Notes • Additional Information • References • Data Citations • Acknowledgements • Author information

FASTQ reads and BAM data records for tumor (T) and normal (N) specimens from each case are freely available along with their conditions of use are freely available on the Texas Cancer Research Biobank website, http://txcrb.org/open.html (Data Citation 1: TCRB Open Access Repository TCRBOA1). Clinical annotations available for these cases are defined in Table 1. Other than a click-through agreement to acknowledge the conditions of use, requirement to create an access account for auditing purposes, and include these conditions within any re-sharing of the data, there are no additional barriers to data access on this portal. User accounts are valid for 30 days and can be renewed. All or some of these data may be downloaded, shared and redistributed for research and educational purposes in accordance with their conditions of use.

To ensure sustainable availability of the data, they are also deposited within SRA. We created the Texas Cancer Research Biobank Open Access Data Sharing Umbrella Project (Accession: PRJNA285925) under which two platform-specific projects were created—the subproject entitled the Texas Cancer Research Biobank Open Access Data Sharing: Exome Project (Data Citation 2: NCBI Sequence Read Archive PRJNA284596) that includes all seven cases and the subproject entitled
Means of validation having been applied

Technical Validation

Abstract · Background & Summary · Methods · Data Records · Technical Validation · Usage Notes · Additional Information · References · Data Citations · Acknowledgements · Author information

The TCRB utilized a secure, database-backed web application called Acquire$^{30}$ (code available at https://github.com/BCM-DLDCC/Acquire) for tracking specimens and their annotations. Through its modules, it supports the full lifecycle of biobanking operations, from collections to quality control testing. Public researchers can use the specimen request module to electronically search for and request available specimens. Acquire greatly facilitated non-OA TCRB donations to the TCGA and ICGC.

As TCRB tissue advocates at sites across the state of Texas consented patients, collected specimens, and entered data into Acquire. The system automatically assigns a barcode and UUID (universally unique identifier) to each specimen, aliquot and derivative. These identifiers are completely masked and contain no PHI or other data that can be mapped back to participants, such that the system’s administrators held the mapping for the UUIDs to participant identifiers acted as the TCRB honest broker. All specimens underwent initial review by expert pathologists for disease diagnosis at the Texas hospital or clinic at which the patients were consented. The TCRB’s own
Usage Notes – here: ethical constraints...!

By downloading or utilizing any part of this dataset, end users must agree to the following conditions of use:

- No attempt to identify any specific individual represented by these data or any derivatives of these data will be made.

- No attempt will be made to compare and/or link this public data set or derivatives in part or in whole to private health information.

- These data in part or in whole may be freely downloaded, used in analyses and repackaged in databases.

- Redistribution of any part of these data or any material derived from the data will include a copy of this notice.

- The data are intended for use as learning and/or research tools only.

- This data set is not intended for direct profit of anyone who receives it and may not be resold.

- Users are free to use the data in scientific publications if the providers of the data (Texas Cancer Research Biobank and Baylor College of Medicine Human Genome Sequencing Center) are properly acknowledged.
Some problems with sharing upon request

• Relies heavily on trust (have you tried “cloning by phoning”?)

• Data associated with published works disappears at a rate of ~17% per year


• Datasets not referenced in a manuscript are essentially invisible (a.k.a “Dark data”)

• Data producers do not get appropriate credit for their work
Stability of databases is another problem!

New Project -> collection of data)

Database developed -> paper written)

End of project

Death of database
-> loss of data!
Where to deposit data?

Browse our recommended data repository online.

- We currently list more than 60 repositories, across the biological, physical and social sciences
- We advise authors on the best place to store their data
Some recommended Data Repositories

**Omics**

**Functional genomics**

Functional genomics is a broad experimental category, and *Scientific Data*’s recommendations in this discipline likewise bridge disparate research disciplines. Data should be deposited following the relevant community requirements where possible.

Please refer to the MIAME standard for microarray data. Molecular interaction data should be deposited with a member of the International Molecular Exchange Consortium (IMEx), following the MIMIx recommendations.

For data linking genotyping and phenotyping information in human subjects, we strongly recommend submission to dbGAP or EGA, which have mechanisms in place to handle sensitive data.

<table>
<thead>
<tr>
<th>Repository</th>
<th>View BioSharing entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArrayExpress</td>
<td>view BioSharing entry</td>
</tr>
<tr>
<td>Gene Expression Omnibus (GEO)</td>
<td>view BioSharing entry</td>
</tr>
<tr>
<td>GenomeRNAi</td>
<td>view BioSharing entry</td>
</tr>
<tr>
<td>dbGAP</td>
<td>view BioSharing entry</td>
</tr>
<tr>
<td>The European Genome-phenome Archive (EGA)</td>
<td>view BioSharing entry</td>
</tr>
</tbody>
</table>
Broad scope of Scientific Data

View data repositories

- Biological sciences:
  - nucleic acid sequence; protein sequence; molecular & supramolecular structure; neuroscience; omics; taxonomy & species diversity; mathematical & modelling resources; cytometry; organism-focused resources
- Health sciences
- Chemistry & chemical biology
- Earth and environmental sciences
- Physics, astrophysics & astronomy
- Social sciences
- Generalist repositories
- Institutional or project-specific repositories
Other use cases: Screening data

- Screen results and in-depth analysis published in 2011 at *PNAS*
- Full screen data published at *Scientific Data* in 2014
- Data at *figshare*
- Data Descriptor cited 26 times according to Google Scholar!

**doi:**10.1038/sdata.2014.35
A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium
SEQC/MAQC-III Consortium | doi:10.1038/nbt.2957

The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance
Wang et al. | doi:10.1038/nbt.3001

Cross-platform ultradecp transcriptomic profiling of human reference RNA samples by RNA-Seq
Xu et al. | doi:10.1038/sdata.2014.20

Transcriptomic profiling of rat liver samples in a comprehensive study design by RNA-Seq
Gong et al. | doi:10.1038/sdata.2014.21
Earth sciences

A Southern Indian Ocean database of hydrographic profiles obtained with instrumented elephant seals

Fabien Roquet, Guy Williams, Mark A. Hindell, Rob Harcourt, Clive McMahon, Christophe Guinet, Jean-Benoit Charrassin, Gilles Reverdin, Lars Boehme, Phil Lovell & Mike Fedak

About Scientific Data
Scientific Data is an open-access, peer-reviewed publication for descriptions of scientifically valuable datasets. Our primary article-type, the Data Descriptor, is designed to make your data more discoverable, interpretable and reusable.

Data in at BODC/NERC
Builds on previous article at Nature Geoscience
Environmental

- New Dataset
- Data in figshare
- Code in figshare
- Integrated figshare data viewer

Cited 47 times, according to Google Scholar!
Who benefits from enhanced re-use of data:

- Individual researcher -> citations
- Scientific community -> access to valuable data
- Society -> progress in research
- Funding agencies -> justification of funding
- Tax payer -> output per € / $ / Yen...
  -> funding for new research
Managing Editor, Scientific Data
Andrew L. Hufton
andrew.hufton@nature.com

Honorary Academic Editor
Susanna-Assunta Sansone

Advisory Panel and Editorial Board including senior researchers, funders, librarians and curators

Visit    nature.com/scientificdata
Email    scientificdata@nature.com
Tweet    @ScientificData

Supported by

acknowledgments