# SCIENTIFIC DATA

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

nature publishing group npg

SCIENTIFIC DATA

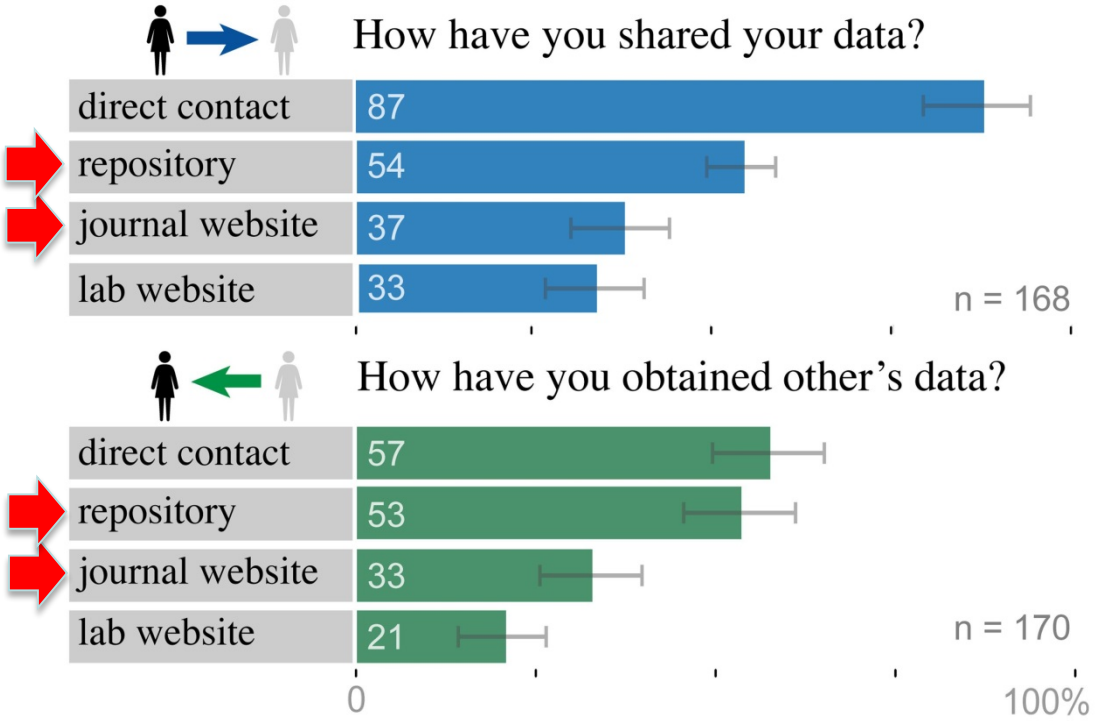
**It‘s about sharing,**

**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)




Kratz JE, Strasser C (2015) Researcher Perspectives on Publication and Peer Review of Data. *PLoS ONE* 10(2): e0117619.

ARTICLE

doi:10.1038/nature11412

# 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 HER2/EGFR/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.

4 OCTOBER 2012 | VOL 490 | NATURE | 61

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

**PDF files**

1. Supplementary Information (14.1M)

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

**Zip files**

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.*

**information is often "hidden" in supplements**

# 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | Supplemental Table 1. The list of samples used on each platform with associated subtype calls from each technolo | | | | | | | |
| 2 | Complete TCGA ID | Gender | Age at Initial Pathologi | ER Status | PR Status | HER2 Final Status | Tumor | Tumor--T1 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 | Sig Unsu ed |
| 3 | TCGA-A2-A0T2 | FEMALE | 66 | Negative | Negative | Negative | T3 | T_Other | N3 | Positive | M1 | Positive | Stage IV | No_Conversion | followup | DECEASED | 240 | 240 | 1 | 240 | Basal-like | |
| 4 | TCGA-A2-A04P | FEMALE | 36 | Negative | Negative | Negative | T2 | T_Other | N3 | Positive | M0 | Negative | Stage IIIC | No_Conversion | followup | DECEASED | 547 | 547 | 1 | 547 | Basal-like | |
| 5 | TCGA-A1-A0SK | FEMALE | 54 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | DECEASED | 594 | 967 | 1 | 967 | Basal-like | |
| 6 | TCGA-A2-A0CM | FEMALE | 40 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | DECEASED | 754 | 754 | 1 | 754 | Basal-like | |
| 7 | TCGA-AR-A1AR | FEMALE | 50 | Negative | Negative | Negative | T1 | T1 | N2 | Positive | M0 | Negative | Stage III | Stage IIIA | enrollment | DECEASED | [Not Available] | 523 | 1 | 523 | Basal-like | |
| 8 | TCGA-B6-A0WX | FEMALE | 40 | Negative | Negative | Negative | T3 | T_Other | N1 | Positive | M0 | Negative | Stage IIIA | No_Conversion | followup | DECEASED | 653 | 653 | 1 | 653 | Basal-like | |
| 9 | TCGA-BH-A1F0 | FEMALE | 80 | Negative | Indeterminate | Negative | T1 | T1 | N1 | Positive | M0 | Negative | Stage IIA | Stage IIA | enrollment | DECEASED | [Not Available] | 785 | 1 | 785 | Basal-like | |
| 10 | TCGA-B6-A0I6 | FEMALE | 49 | Negative | Negative | Not Available | T1 | T1 | N1 | Positive | M0 | Negative | Stage IIA | No_Conversion | followup | DECEASED | 991 | 997 | 1 | 997 | Basal-like | |
| 11 | TCGA-BH-A18V | FEMALE | 48 | Negative | Negative | Negative | T2 | T_Other | N1 | Positive | M0 | Negative | Stage IIB | No_Conversion | enrollment | DECEASED | 1555 | 1555 | 1 | 1555 | Basal-like | |
| 12 | TCGA-BH-A18Q | FEMALE | 56 | Negative | Negative | Negative | T2 | T_Other | N1 | Positive | M0 | Negative | Stage IIB | No_Conversion | enrollment | DECEASED | 1692 | 1692 | 1 | 1692 | Basal-like | |
| 13 | TCGA-BH-A18K | FEMALE | 46 | Positive | Positive | Negative | T1 | T1 | N0 | Negative | M0 | Negative | Stage I | Stage I | enrollment | DECEASED | 2547 | 2762 | 1 | 2762 | Basal-like | |
| 14 | TCGA-BH-A0HL | FEMALE | 56 | Positive | Positive | Negative | T2 | T_Other | N1 | Positive | M0 | Negative | Stage IIB | No_Conversion | followup | LIVING | 72 | NA | 0 | 72 | Basal-like | |
| 15 | TCGA-BH-A0E0 | FEMALE | 38 | Negative | Negative | Negative | T3 | T_Other | N3 | Positive | M0 | Negative | Stage IIIC | No_Conversion | followup | LIVING | 133 | NA | 0 | 133 | Basal-like | |
| 16 | TCGA-BH-A0RX | FEMALE | 59 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 170 | NA | 0 | 170 | Basal-like | |
| 17 | TCGA-A7-A13D | FEMALE | 46 | Negative | Positive | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 267 | NA | 0 | 267 | Basal-like | |
| 18 | TCGA-BH-A0E6 | FEMALE | 69 | Negative | Negative | Negative | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | Stage I | followup | LIVING | 292 | NA | 0 | 292 | Basal-like | |
| 19 | TCGA-AO-A0J4 | FEMALE | 41 | Negative | Negative | Negative | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | Stage I | followup | LIVING | 294 | NA | 0 | 294 | Basal-like | |
| 20 | TCGA-A7-A0CE | FEMALE | 57 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 309 | NA | 0 | 309 | Basal-like | |
| 21 | TCGA-A7-A13E | FEMALE | 62 | Positive | Negative | Negative | T2 | T_Other | N1 | Positive | M0 | Negative | Stage IIB | No_Conversion | followup | LIVING | 326 | NA | 0 | 326 | Basal-like | |
| 22 | TCGA-A7-A0DA | FEMALE | 62 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 373 | NA | 0 | 373 | Basal-like | |
| 23 | TCGA-D8-A142 | FEMALE | 74 | Negative | Negative | Negative | T3 | T_Other | N0 | Negative | M0 | Negative | Stage IIB | Stage IIB | followup | LIVING | 425 | NA | 0 | 425 | Basal-like | |
| 24 | TCGA-D8-A143 | FEMALE | 51 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 431 | NA | 0 | 431 | Basal-like | |
| 25 | TCGA-AQ-A04J | FEMALE | 45 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 499 | NA | 0 | 499 | Basal-like | |
| 26 | TCGA-BH-A0HN | FEMALE | 67 | Positive | Positive | Negative | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | Stage I | followup | LIVING | 516 | NA | 0 | 516 | Basal-like | |
| 27 | TCGA-A2-A0T0 | FEMALE | 59 | Negative | Negative | Negative | T2 | T_Other | N1 | Positive | M0 | Negative | Stage IIB | No_Conversion | followup | LIVING | 533 | NA | 0 | 533 | Basal-like | |
| 28 | TCGA-A2-A0YE | FEMALE | 48 | Negative | Negative | Negative | T2 | T_Other | N1 | Positive | M0 | Negative | Stage IIB | No_Conversion | followup | LIVING | 553 | NA | 0 | 553 | Basal-like | |
| 29 | TCGA-A2-A0YJ | FEMALE | 39 | Positive | Negative | Negative | T3 | T_Other | N2 | Positive | M0 | Negative | Stage IIIA | No_Conversion | followup | LIVING | 565 | NA | 0 | 565 | Basal-like | |
| 30 | TCGA-A2-A0D0 | FEMALE | 60 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 643 | NA | 0 | 643 | Basal-like | |
| 31 | TCGA-A2-A04U | FEMALE | 47 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 670 | NA | 0 | 670 | Basal-like | |
| 32 | TCGA-AO-A0J6 | FEMALE | 61 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 775 | NA | 0 | 775 | Basal-like | |
| 33 | TCGA-A2-A0YM | FEMALE | 67 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIA | Stage IIA | followup | LIVING | 964 | NA | 0 | 964 | Basal-like | |
| 34 | TCGA-A2-A0D2 | FEMALE | 45 | Negative | Negative | Negative | T2 | T_Other | N0 | Negative | M0 | Negative | Stage IIB | Stage IIA | followup | LIVING | 1027 | NA | 0 | 1027 | Basal-like | |
| 35 | TCGA-BH-A0B3 | FEMALE | 53 | Negative | Negative | Negative | T2 | T_Other | N1 | Positive | M0 | Negative | Stage IIB | No_Conversion | followup | LIVING | 1203 | NA | 0 | 1203 | Basal-like | |
| 36 | TCGA-A2-A04Q | FEMALE | 48 | Negative | Negative | Negative | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | Stage I | followup | LIVING | 1275 | NA | 0 | 1275 | Basal-like | |
| 37 | TCGA-A2-A0SX | FEMALE | 48 | Negative | Negative | Negative | T1 | T1 | N0 | Negative | M0 | Negative | Stage IA | Stage I | followup | LIVING | 1288 | NA | 0 | 1288 | Basal-like | |

A need of proper description/annotation of that data to facilitate re-use!

HELMHOLTZ GEMEINSCHAFT

Open Science

# 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

SCIENTIFIC DATA

Search Go

Advanced search

Home | Archive | About ▾ | For Authors ▾ | For Referees | Advisory & Editorial Board ▾ | Data Policies ▾

## 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**

Hanke *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



## Announcements

Scientific Data Updates

SCIENTIFIC DATA

## 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**

**Data Records**

All the samples used in this study are summarized in Table 1. Consistent identifiers are used in Tables 2 and 3 to allow mapping between the proteomic and transcriptomic data outputs.

**Data Record 1**

The raw data, peaklists (.mgf), ProteomeDiscoverer result files (.msf) and ProteomeDiscoverer workflow files (.xml) have been uploaded to ProteomeXchange (http://www.proteomexchange.org/) with the following accession number PXD000134 (ref. 67; Table 2).

**Data Record 2**

Microarray data are available at the NCBI Gene Expression Omnibus (GEO) database under the accession numbers GSE26451 (ref. 68) and GSE26453 (ref. 69; Table 3).

**Data Record 3**

The peptide and protein identification data sets have been annotated by The Global Proteome Machine at http://gpmdb.thegpm.org/

**Data Record 4**

The peptide and protein identification data sets have been annotated by the StemCellOmicsRepository (SCOR) at http://scor.chem.wisc.edu/

\+ All articles supported by machine-readable metadata in the ISA-tab format

# Sample case



*SCIENTIFIC DATA* | DATA DESCRIPTOR OPEN

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

**Lauren B. Becnel**, **Stacey Pereira**, **Jennifer A. Drummond**, **Marie-Claude Gingras**, **Kyle R. Covington**, **Christie L. Kovar**, **Harsha Vardhan Doddapaneni**, **Jianhong Hu**, **Donna Muzny**, **Amy L. McGuire**, **David A. Wheeler** & **Richard A. Gibbs**

Affiliations | Contributions | Corresponding author

*Scientific Data* **3**, Article number: 160010 (2016) | doi:10.1038/sdata.2016.10

# Data descriptors to increase utility of data

# Access to the data

http://txcrb.org/open.html

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.

| Case # | Sex/ Age/ Race/ Ethnicity | Prior treatment | Tumor % cellularity/ TNM | Disease Morph./ Anatomic Site | Tumor Grade |
|---|---|---|---|---|---|
| 1 | M/ 51-60/ White/ Not Hispanic or Latino | No | 10%/ T3 N1 M0 | 8500/3: infiltrating duct adenocarcinoma/ Head of pancreas | II |
| | F/ | | | | |

# There is room for detailed Methods

## Methods

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

### 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.



### Selecting the OA cohort

*Scientific Data* **3**, Article number: 160010 (2016) doi:10.1038/sdata.2016.10

# 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



*Scientific Data* **3**, Article number: 160010 (2016) doi:10.1038/sdata.2016.10

# 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



*Scientific Data* **3**, Article number: 160010 (2016) doi:10.1038/sdata.2016.10

# Usage Notes – here: ethical constraints…!

## Usage Notes

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

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.

*Scientific Data* **3**, Article number: 160010 (2016) doi:10.1038/sdata.2016.10

# 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

  (Vines et al. 2014, *Current Biology* 24(1), 94–97,2014. doi:10.1016/j.cub.2013.11.014)
- 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)

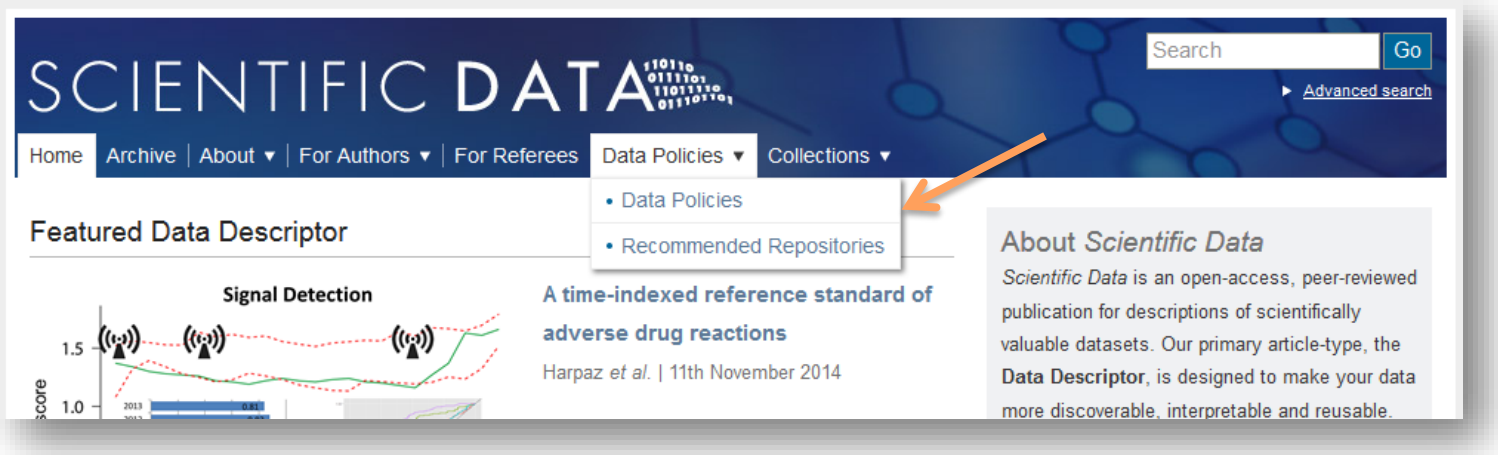
↓

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.

| | |
|---|---|
| ArrayExpress | view BioSharing entry |
| Gene Expression Omnibus (GEO) | view BioSharing entry |
| GenomeRNAi | view BioSharing entry |
| dbGAP | view BioSharing entry |
| The European Genome-phenome Archive (EGA) | view BioSharing entry |

# 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

SCIENTIFIC DATA

Search Go

Advanced search

Home | Archive | About | For Authors | For Referees | Data Policies | Collections

Home ▸ Data Descriptors ▸ Data Descriptor

SCIENTIFIC DATA | DATA DESCRIPTOR OPEN

Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies

Glenn S Cowley, Barbara A Weir […] William C Hahn

Affiliations | Contributions | Corresponding authors

Scientific Data 1, Article number: 140035 (2014) | doi:10.1038/sdata.2014.35
Received 20 May 2014 | Accepted 22 August 2014 | Published online 30 September 2014

Changes have been made to this article:
Corrigendum (11 November 2014)

PDF | ISA tab | Citation | Reprints | Rights & permissions | Article metrics

Abstract

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

Using a genome-scale, lentivirally delivered shRNA library, we performed massively parallel pooled

About Scientific Data

Scientific Data is an open-access, peer-reviewed journal 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

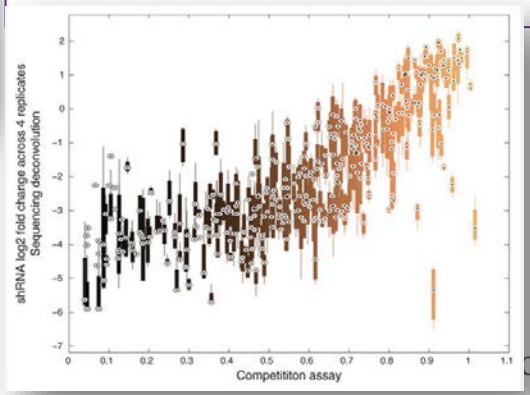
Associated Links

Cancer Discovery | Article
Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells
by A. Buzina *et al*

Proceedings of the National Academy of Sciences | Article
Highly parallel identification of essential genes in cancer cells
by A. Subramanian *et al*

Proceedings of the National Academy of Sciences | Article
Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer
by A. East *et al*

- **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

# Publish alongside: major consortium

See the Focus on RNA sequencing quality control (SEQC)
In the September 2014 issue of *Nature Biotechnology*


**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

SCIENTIFIC DATA

**Cross-platform ultradeep 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



*SCIENTIFIC DATA* | DATA DESCRIPTOR OPEN

## 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 & M… Fedak**

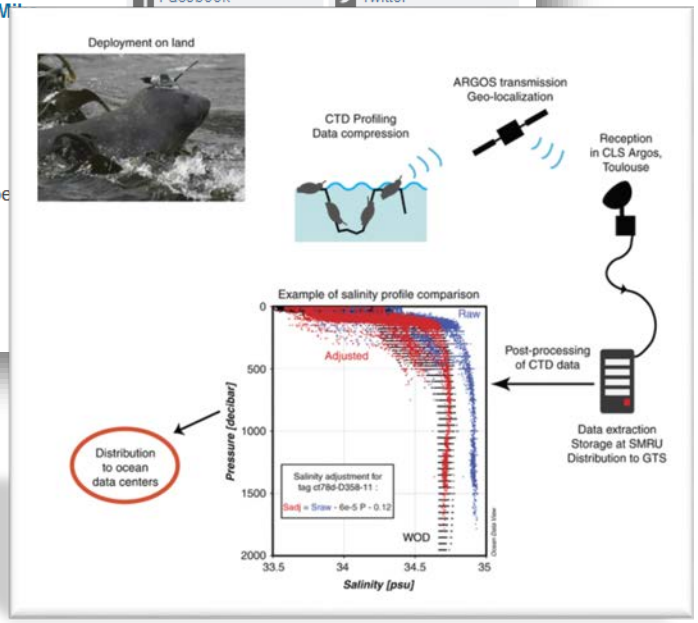
Affiliations | Contributions | Corresponding author

*Scientific Data* **1**, Article number: 140028 | doi:10.1038/sdata.2014.28

Received 23 May 2014 | Accepted 04 August 2014 | Published online 02 Septembe

PDF | ISA tab | Citation | Reprints | Rights & permissions | Article metrics

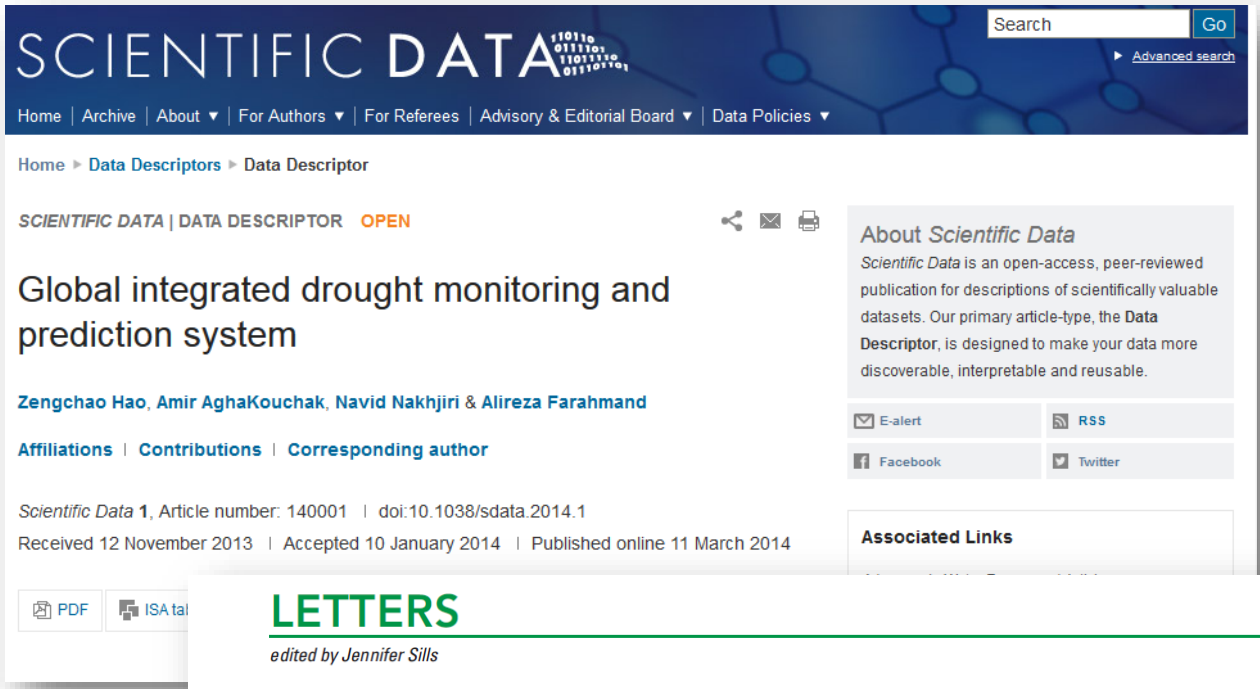


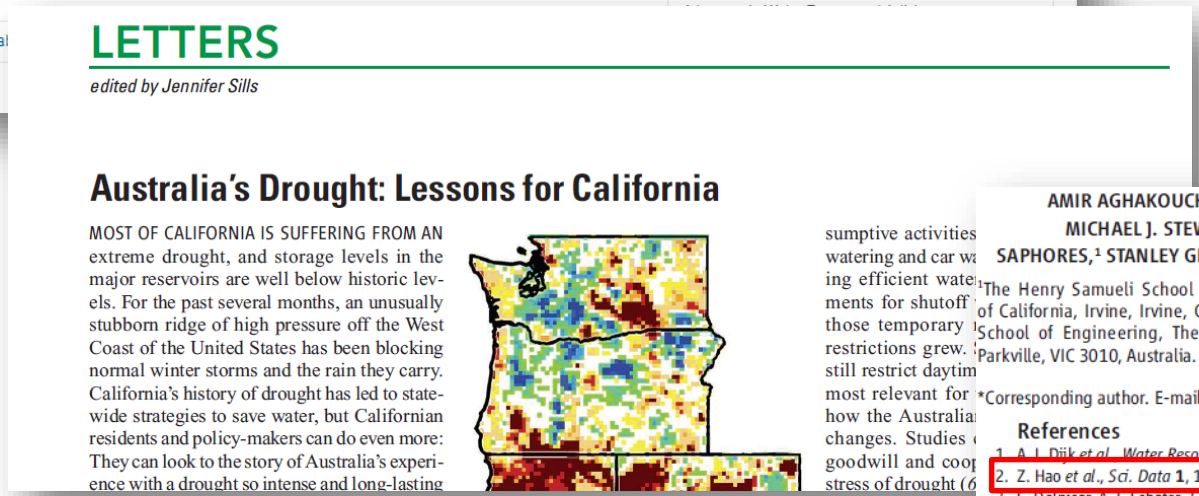


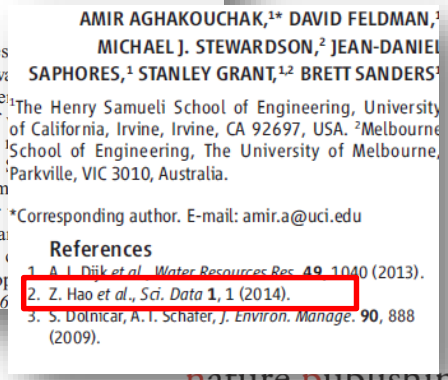

- Data in **at BODC/NERC**
- Builds on previous article at *Nature Geoscience*

SCIENTIFIC DATA

# Environmental




- **New Dataset**
- Data in **fig**share
- Code in **fig**share
- Integrated **fig**share data viewer

*Cited 47 times, according to Google Scholar!*

nature publishing group npg

## 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

# SCIENTIFIC DATA

## acknowledgments

**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 DRYAD biosharing figshare credit for all your research isatools COLLECT | CURATE | ANALYSE | SHARE | PUBLISH

nature publishing group npg