Instructions: Data Retrieval Assignment

Deadline: Sunday, 14 September, 2025 (midnight)

1. Objective

From now onward, you will work with your own selected dataset throughout the internship. Your first task is to retrieve a suitable microarray dataset from public repositories.

2. Dataset Requirements

- Must include two biological states
  - Example: *disease vs. healthy control (No disease subtypes)*
- Must have at least 30 samples in total
- You may choose any sample type (e.g., tissue, blood etc)
- This dataset will be used for all further modules: preprocessing, differential expression, and machine learning analysis.

3. Where to Retrieve Data

- ArrayExpress (EMBL-EBI)
  https://www.ebi.ac.uk/biostudies/arrayexpress
- NCBI GEO (Gene Expression Omnibus)
  https://www.ncbi.nlm.nih.gov/geo/

4. Files to Download

From ArrayExpress

- Raw data (e.g., CEL, Txt or depending on platform)
- Sample and Data Relationship Format-SDRF (.sdrf file)

From NCBI GEO

- Raw data (CEL or equivalent raw files)
- Series Matrix File (e.g GSEXXXX_series_matrix.txt.gz)

Notes

- These files can be directly imported into R.
- Raw data can be very large. If your internet connection is unstable, downloads inside R may fail. In such cases, download raw files directly from the repository website.
- CEL files (common for Affymetrix platforms) are provided as separate files for each sample. Make sure you download all sample files.
- Save all files in a folder named with the dataset accession ID.

5. Submission via Google Form

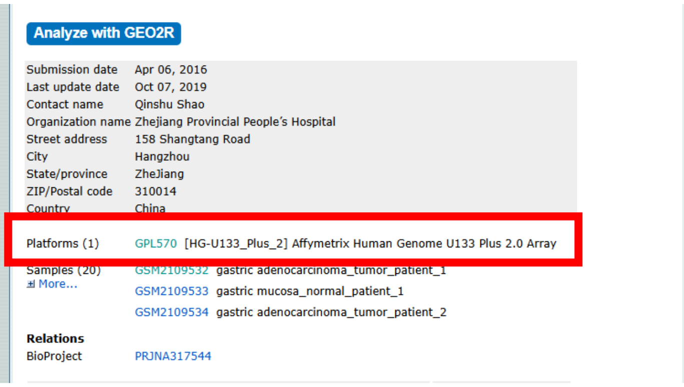
Submit your dataset information using the provided Google Form link. https://forms.gle/hHiQZNCidbiAxwQo8

You will need to enter the following details:

1. Dataset ID (ArrayExpress or GEO accession number)
2. Dataset title
3. Study Design. (Example: normal vs disease patients, normal vs adjacent tissues from same individual, treated vs untreated etc.)
4. Disease Type (Example: breast cancer, type 2 diabetes, Alzheimer’s disease etc.)
5. Total number of samples
6. Number of disease samples

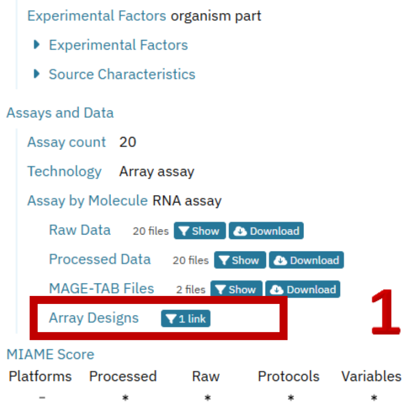
7. Number of control samples
8. Database used (ArrayExpress or NCBI GEO)
9. Platform: Affymetrix, illumine, Agilent etc

- For NCBI datasets, (Mention platform number as well GPL570 etc)



Submission date	Apr 06, 2016
Last update date	Oct 07, 2019
Contact name	Qinshu Shao
Organization name	Zhejiang Provincial People's Hospital
Street address	158 Shangtang Road
City	Hangzhou
State/province	ZheJiang
ZIP/Postal code	310014
Country	China
Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (20) ⊞ More...	GSM2109532 gastric adenocarcinoma_tumor_patient_1 GSM2109533 gastric mucosa_normal_patient_1 GSM2109534 gastric adenocarcinoma_tumor_patient_2
Relations	
BioProject	PRJNA317544

- For ArrayExpress (mention Array Designs: Affymetrix GeneChip etc)


Experimental Factors organism part

- Experimental Factors
- Source Characteristics

Assays and Data

Assay count 20

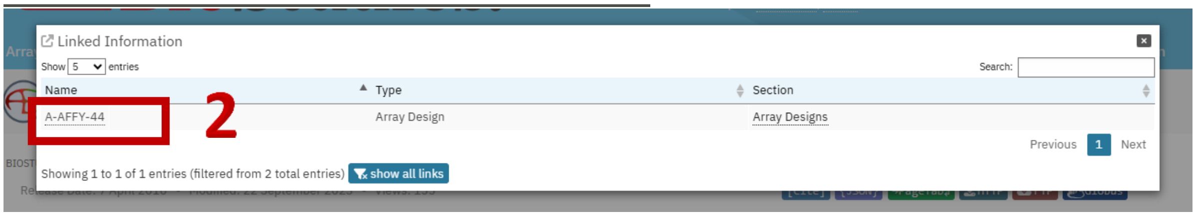
Technology Array assay

Assay by Molecule RNA assay

Raw Data 20 files
Processed Data 20 files
MAGE-TAB Files 2 files
Array Designs

MIAME Score

Platforms	Processed	Raw	Protocols	Variables
-	*	*	*	*

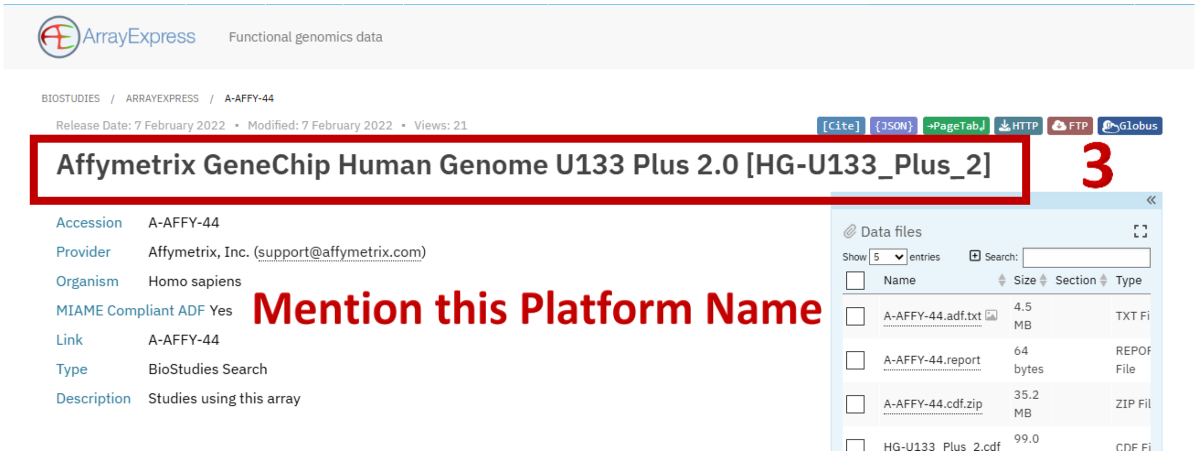

Linked Information

Show 5 entries

Search:

Name	Type	Section
A-AFFY-44	Array Design	Array Designs

Showing 1 to 1 of 1 entries (filtered from 2 total entries)



ArrayExpress Functional genomics data

BIOSTUDIES / ARRAYEXPRESS / A-AFFY-44

Release Date: 7 February 2022 • Modified: 7 February 2022 • Views: 21

[Cite] {JSON} →PageTab HTTP FTP Globus

Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133_Plus_2]

3

Accession	A-AFFY-44
Provider	Affymetrix, Inc. (support@affymetrix.com)
Organism	Homo sapiens
MIAME Compliant ADF	Yes
Link	A-AFFY-44
Type	BioStudies Search
Description	Studies using this array

Mention this Platform Name

Data files

Show 5 entries Search:

□	Name	Size	Section	Type
□	A-AFFY-44.adf.txt	4.5 MB		TXT Fi
□	A-AFFY-44.report	64 bytes		REPOR File
□	A-AFFY-44.cdf.zip	35.2 MB		ZIP Fil
□	HG-U133_Plus_2.cdf	99.0		CDF Fi

10. Additional information: Does the dataset include metadata such as patient age, sex, clinical features, treatment response, etc.?