生醫研究中的資料再利用

Chia-Lang Hsu (許家郎), Ph.D.
Department of Medical Research
National Taiwan University Hospital (NTUH)

Open data - Why make data available?

Accelerated Discovery

Openly available datasets enable researchers to build upon existing work, validate each other's findings, and accelerate scientific discovery without repeating costly and time-consuming data collection.

Helps ensure we don't miss breakthroughs

There are countless ways to analyze any given dataset. What appears as noise to one researcher might represent a key discovery to another, depending on their perspective or analytical approach.

Improved Integrity and Reproducibility

When the underlying data are accessible, researchers can verify one another's work and ensure that conclusions rest on a solid foundation.

Accountability and Transparency

Making data publicly available promotes transparency and accountability, and enables informed civic participation in policy decisions.

FAIR your data

https://www.biomeris.it/en/introduction-fair-principles/

DATA SHARING TO ADVANCE GENETIC AND GENOMIC KNOWLEDGE

Data sharing has been common practice in genetics and genomics research for decades, accelerating biomedical discoveries and improving human health. Sustained funding to the National Institutes of Health (NIH) will ensure that genetic and genomic data remain accessible to researchers in ways that advance science and also protect the privacy of participants and minimize the stigmatization of groups of people.

Large-Scale Data is Transformative

Researchers use biobanks – large collections of biological samples and health information – and other research resources with shared genetic and genomic data from large numbers of diverse research participants to investigate new scientific questions. Open data empowers investigators with the ability to pool data, effectively increasing the sample size for robust studies and promoting reproducible science. Reusing existing data this way helps scientists learn more about the health of individuals and populations and drives medical innovations.

What is Data Sharing?

Data sharing is the practice of making de-identified sociodemographic, genomic, and medical record data used for research available to other investigators. Sharing research findings benefits the scientific community, fosters collaboration, and increases transparency with the public. Since science progresses by building upon the work of others and innovates from prior discoveries, ethical data sharing accelerates knowledge.[1]

https://www.ashg.org/wp-content/uploads/2023/10/ASHG_Success_Stories-Data-Sharing-Final.pdf

‘Data Availability Statements’ in a research paper

- An article’s data availability statement lets a reader know where and how to access data that support the results and analysis.
- Data availability statements are important because they support **validation**, **reuse** and **citation** of research data.

Major repositories for public high-throughput data

Repository type	Repository names	Typical data type	Note
Public repositories for non-sensitive data	GEO, ENA, SRA, ArrayExpress	Bulk & single-cell RNA-seq, microarray, ChIP-seq, ATAC-seq	Processed or raw expression data, no identifiable human info
Public repositories for sensitive data	dbGAP, EGA, JGA	**Human** WGS, WES, genotype, clinical / phenotype, RNA-seq	Require DAC approval, human identifiable data
Project-specific repositories	GDC (TCGA, TARGET), ICGC	Cancer multi-omics (WES, RNA-seq, methylation, clinical)	NIH- or consortium-based data portals
Single-cell and spatial data portals	Human Cell Atlas, Single Cell Portal	scRNA-seq	
Proteomics repositories	PRIDE, PeptideAtlas	Mass spectrometry proteomics	Processed protein identification and quantification
Metabolomics repositories	MetaboLights, Metabolomics Workbench	LC-MS, GC-MS metabolomics	

GEO & SRA

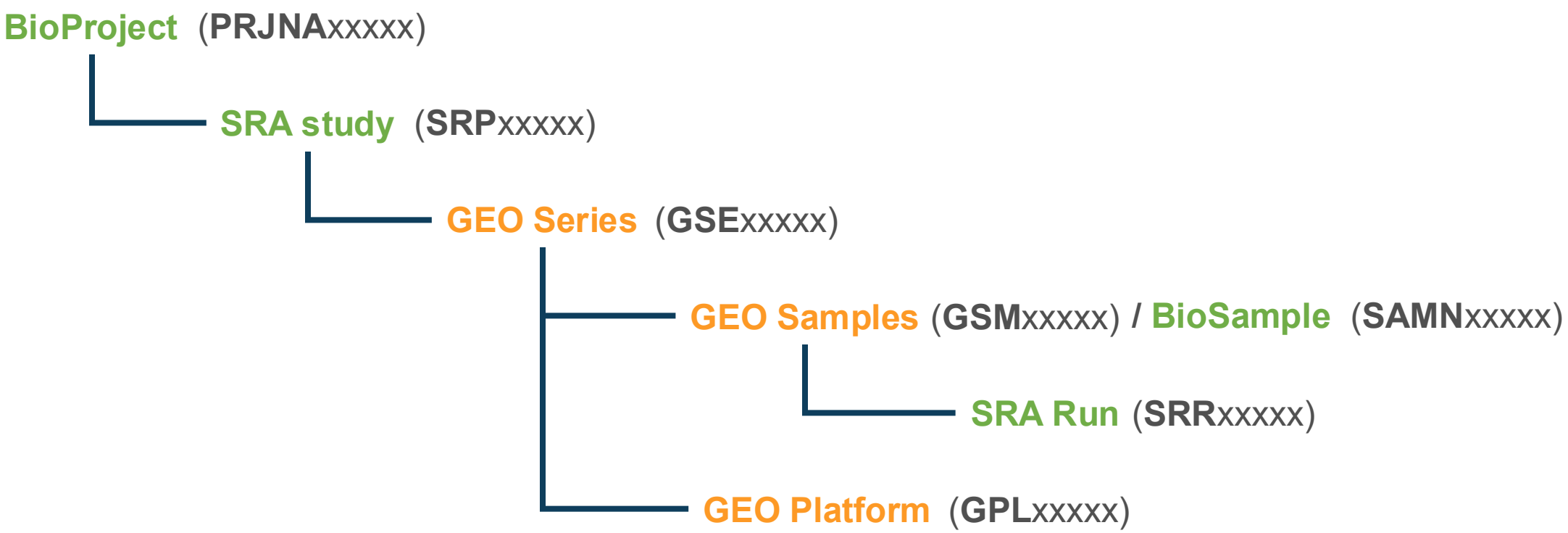
SRA
Sequence Read Archive

- A global repository jointly maintained by NCBI, EBI, and DDBJ.
- Stores **RAW SEQUENCING** data (FASTQ) generated by high-throughput technologies.
- Enables users to re-analyze data using their own pipelines or updated reference genomes.

GEO
Gene Expression Omnibus

- A public repository maintained by NCBI for sharing gene expression and functional genomics data.
- Stores **PROCESSED or NORMALIZED** data such as expression matrices and sample annotations.
- Common data types: microarray, RNA-seq, ChIP-seq, ATAC-seq, and single-cell RNA-seq.

Relationship between SRA and GEO



A single **GSE** may correspond to multiple **GSM**s, and each **GSM** can be linked to one or more **SRR**s.

Sequencing-based data

Series page

Series GSE255949 UPDATE — Query DataSets for GSE255949

Status	Public on Jul 05, 2024
Title	Gene expression of alveolar macrophage in mice with AT2 cell-specific Cpt1a deletion
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	Carnitine palmitoyltransferase 1a (CPT1a) is the key regulator of mitochondrial long-chain fatty acid beta-oxidation (LCFAO). However, the functional significance of AT2 cell-specific LCFAO at baseline and during acute lung injury (ALI) is not fully understood. In this study, Murine models of AT2 cell-specific Cpt1a deletion were generated for investigating the role of CPT1a in regulating AT2 cell function and the severity of lipopolysaccharide-induced murine ALI models.
Overall design	To investigate whether impaired LCFAO in AT2 cells regulates the responses of alveolar macrophages, we also isolated alveolar macrophages from Cpt1a-KO and control mice for trancriptomic analysis.
Contributor(s)	Chung K
Citation(s)	Chung KP, Cheng CN, Chen YJ, Hsu CL et al. Alveolar epithelial cells mitigate neutrophilic inflammation in lung injury through regulating mitochondrial fatty acid oxidation. *Nat Commun* 2024 Aug 22;15(1):7241. PMID: 39174557
NIH grant(s)	Add grant
Submission date	Feb 16, 2024
Last update date	Sep 27, 2024
Contact name	Chia-Lang Hsu
E-mail(s)	chialanghsu@ntuh.gov.tw
Organization name	National Taiwan University Hospital
Department	Department of Medical Research
Street address	No. 7, Zhongshan S. Rd.
City	Taipei
ZIP/Postal code	100
Country	Taiwan
Platforms (1)	GPL24247 Illumina NovaSeq 6000 (Mus musculus)
Samples (11)	GSM8083254 Alveolar macrophages from control mouse (sample #1) GSM8083255 Alveolar macrophages from control mouse (sample #2) GSM8083256 Alveolar macrophages from control mouse (sample #3)
	⊿ More...

Relations

BioProject	PRJNA1077286

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE255949_read_count_macrophage_baseline.txt.gz	702.7 Kb	(ftp)(http)	TXT

SRA Run Selector

Raw data are availa... SRA

Next page

Sample GSM8083254 — Query DataSets for GSM8083254

Status	Public on Jul 05, 2024
Title	Alveolar macrophages from control mouse (sample #1)
Sample type	SRA
Source name	Lung
Organism	Mus musculus
Characteristics	tissue: Lung cell type: Primary alveolar macrophages genotype: Sftpc-CreERt2(+/-)
Extracted molecule	total RNA
Extraction protocol	Cell as input and lysis accroding to manufacturer's instructions of SMART-Seq Stranded Kit (Takara) Libraries were prepared using the SMART-Seq Stranded Kit (Takara)
Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina NovaSeq 6000
Description	N11301_01
Data processing	Trimming adaptor sequences and removing reads of low quality were performed by cutadapt (v2.4). Quantified reads were aligned to the mouse genome (GRCm38) by STAR (v2.7.2a) and then gene-level read counts were generated based on the annotations of Gencode (vM25) Assembly: GRCm38 Supplementary files format and content: tab-delimited text files include read count values for each sample
Submission date	Feb 16, 2024
Last update date	Jul 05, 2024
Contact name	Chia-Lang Hsu
E-mail(s)	chialanghsu@ntuh.gov.tw
Organization name	National Taiwan University Hospital
Department	Department of Medical Research
Street address	No. 7, Zhongshan S. Rd.
City	Taipei
ZIP/Postal code	100
Country	Taiwan
Platform ID	GPL24247
Series (1)	GSE255949 Gene expression of alveolar macrophage in mice with AT2 cell-specific Cpt1a deletion

Relations

BioSample	SAMN39964386
SRA	SRX23642287

Supplementary data files not provided

SRA Run Selector

Raw data are available in SRA

Sample page

SRA project page

Accession PRJNA1077286 Search

Common Fields

BioProject	PRJNA1077286
Consent	PUBLIC
Assay Type	RNA-Seq
AvgSpotLen	302
cell_type	Primary alveolar macrophages
Center Name	DEPARTMENT OF MEDICAL RESEARCH, NATIONAL TAIWAN UNIVERSITY HOSPITAL
Collection_Date	missing
DATASTORE filetype	FASTQ, RUN.ZQ, SRA
DATASTORE provider	GS, NCBI, S3

Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	11	56.38 Gb	179.26 G	Metadata or Accession List		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Found 11 Items

Project → Sample → Experiment → Run

	Run	BioSample	Bases	Bytes	Experiment	genotype	Library Name	create_date	Sample Name
1	SRR27989218	SAMN39964376	22.04 G	6.93 Gb	SRX23642297	Cpt1a-KO (AT2 cell-specific)	GSM8083264	2024-02-16 12:57:00Z	GSM8083264
2	SRR27989219	SAMN39964377	16.01 G	5.04 Gb	SRX23642296	Cpt1a-KO (AT2 cell-specific)	GSM8083263	2024-02-16 13:15:00Z	GSM8083263
3	SRR27989220	SAMN39964378	16.06 G	5.03 Gb	SRX23642295	Cpt1a-KO (AT2 cell-specific)	GSM8083262	2024-02-16 12:46:00Z	GSM8083262

Non-sequencing-based data

Series page

Series GSE124226 Query DataSets for GSE124226

Status	Public on Jan 24, 2019
Title	Gene Expression Data of Adipose Stem Cells from Normal-Weight Polycystic Ovary Syndrome Women vs. Controls
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	In vitro studies of subcutaneous (SC) abdominal adipose stem cells (ASC) from women with polycystic ovary syndrome (PCOS) show altered ASC commitment to preadipocytes and differentiation to mature adipocytes related to hyperandrogenism. The goal of the study is to use microarrays to examine whether SC abdominal ASC gene expression are altered in normal-weight PCOS women and correlated with hyperandrogenemia and/or insulin resistance, which are prevalent clinical pathologies of PCOS.
Overall design	ASCs were isolated from SC adipose tissue following SC abdominal biopsy in 4 PCOS women and 4 age and BMI matched controls for gene expression comparison. First generation of stem cells were cultured until cells reached confluency and isolated for RNA extraction and hybridization on Affymetrix microarrays. Grant#: P50 HD071836 Title: Androgen excess as a mechanism for adipogenic dysfunction in PCOS women Period: 2018-2022
Contributor(s)	Phan JD, Leung KL, Ding X, Li X, Chazenbalk GD
Citation(s)	Dumesic DA, Phan JD, Leung KL, Grogan TR et al. Adipose Insulin Resistance in Normal-Weight Women With Polycystic Ovary Syndrome. *J Clin Endocrinol Metab* 2019 Jun 1;104(6):2171-2183. PMID: 30649347

Analyze with GEO2R

Submission date	Dec 20, 2018
Last update date	Mar 25, 2019
Contact name	Greogorio Chazenbalk
E-mail(s)	gchazenbalk@mednet.ucla.edu
Phone	1-818-599-4175
Organization name	University of California, Los Angeles
Department	Obstetrics and Gynecology
Street address	10833 Le Conte Ave
City	Los Angeles
State/province	California
ZIP/Postal code	90095
Country	USA

Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (8)	GSM3526020 ASC-3004-control GSM3526021 ASC-3006-PCOS GSM3526022 ASC-3010-PCOS
≪ More...	

Relations

BioProject PRJNA516631

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE124226_RAW.tar	38.6 Mb	(http)(custom)	TAR (of CEL)

Processed data included within Sample table

Sample GSM3526020 Query DataSets for GSM3526020

Status	Public on Jan 24, 2019
Title	ASC-3004-control
Sample type	RNA
Source name	subcutaneous adipose tissue, control
Organism	Homo sapiens
Characteristics	female type: control
Growth protocol	Adipose stem cells were plated in 10 cm dish containing DMEM/10% FCS, 0.05 U/ml penicillin, 0.05 mg/ml streptomycin, 1.25 mg/ml fungizone and cultured at 37°C until cells reached confluency.
Extracted molecule	total RNA
Extraction protocol	Total cellular RNA was isolated using RNeasy mini kit (Qiagen, Carlsbad, CA) according to the manufacturer's protocol.
Label	biotin
Label protocol	Biotinylated cRNA were prepared according to the standard Affymetrix protocol from 500 ng total RNA (Ambion MessageAmp™ Premier RNA Amplification Protocol)
Hybridization protocol	Following fragmentation, 20 ug of cRNA were hybridized for 16 hr at 45C on GeneChip Human U133 plus 2.0 Array. GeneChips were washed and stained in the Affymetrix Fluidics Station 450.
Scan protocol	GeneChips were scanned using the Affymetrix GeneChip Scanner.
Description	Gene expression data from control ASCs of first pair-match
Data processing	The data was analyzed with Partek Genomics suite using the RMA as normalization method.

Submission date	Dec 20, 2018
Last update date	Jan 24, 2019
Contact name	Greogorio Chazenbalk
E-mail(s)	gchazenbalk@mednet.ucla.edu
Phone	1-818-599-4175
Organization name	University of California, Los Angeles
Department	Obstetrics and Gynecology
Street address	10833 Le Conte Ave
City	Los Angeles
State/province	California
ZIP/Postal code	90095
Country	USA

Platform ID	GPL570
Series (1)	GSE124226 Gene Expression Data of Adipose Stem Cells from Normal-Weight Polycystic Ovary Syndrome Women vs. Controls

Data table header descriptions

ID_REF	
VALUE	RMA normalized log2 transformed signal intensity

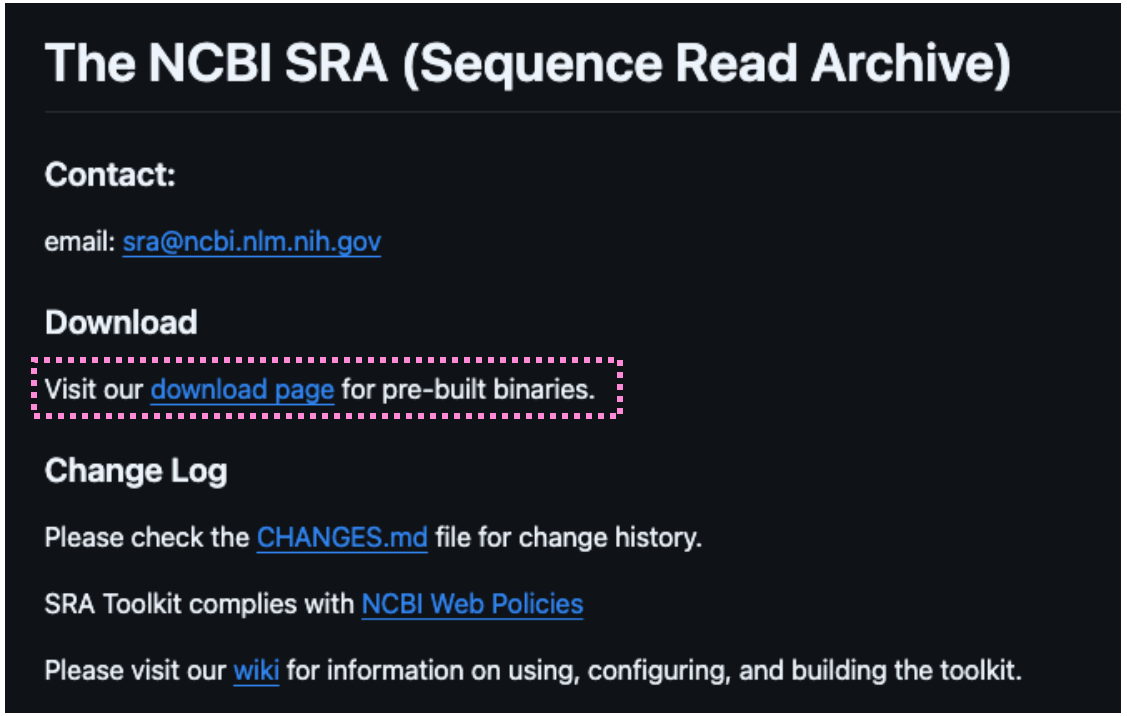
Data table

ID_REF	VALUE
1007_s_at	9.54885
1053_at	8.61334
117_at	6.20375
121_at	9.09007
1255_g_at	3.8519
1294_at	7.95796

Sample page

Extract FASTQ-files from SRA-accessions using SRA Toolkit

https://github.com/ncbi/sra-tools



The NCBI SRA (Sequence Read Archive)

Contact:

email: sra@ncbi.nlm.nih.gov

Download

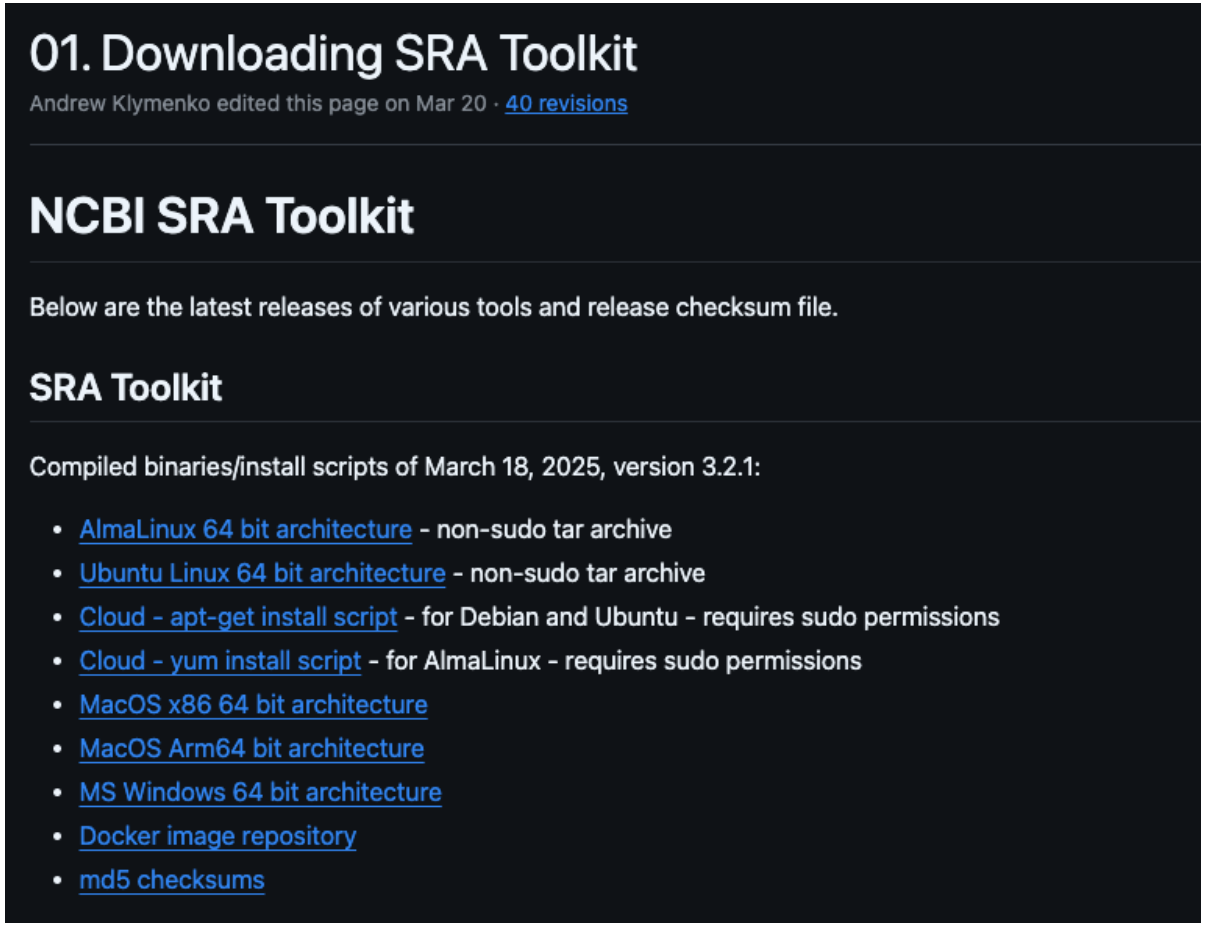
Visit our download page for pre-built binaries.

Change Log

Please check the CHANGES.md file for change history.

SRA Toolkit complies with NCBI Web Policies

Please visit our wiki for information on using, configuring, and building the toolkit.



01. Downloading SRA Toolkit

Andrew Klymenko edited this page on Mar 20 · 40 revisions

NCBI SRA Toolkit

Below are the latest releases of various tools and release checksum file.

SRA Toolkit

Compiled binaries/install scripts of March 18, 2025, version 3.2.1:

- AlmaLinux 64 bit architecture - non-sudo tar archive
- Ubuntu Linux 64 bit architecture - non-sudo tar archive
- Cloud - apt-get install script - for Debian and Ubuntu - requires sudo permissions
- Cloud - yum install script - for AlmaLinux - requires sudo permissions
- MacOS x86 64 bit architecture
- MacOS Arm64 bit architecture
- MS Windows 64 bit architecture
- Docker image repository
- md5 checksums

SRA Toolkit – Step 1: download datasets

prefetch retrieves SRA files (.sra) from a given accession ID (SRR, SRX, SRP, etc.)

```
# Example: download a single run
prefetch SRR1234567

# Example: download multiple runs listed in a text file
prefetch --option-file SRR_AccList.txt

# Example: specify download directory
prefetch --output-directory /data/SRA SRR1234567
```

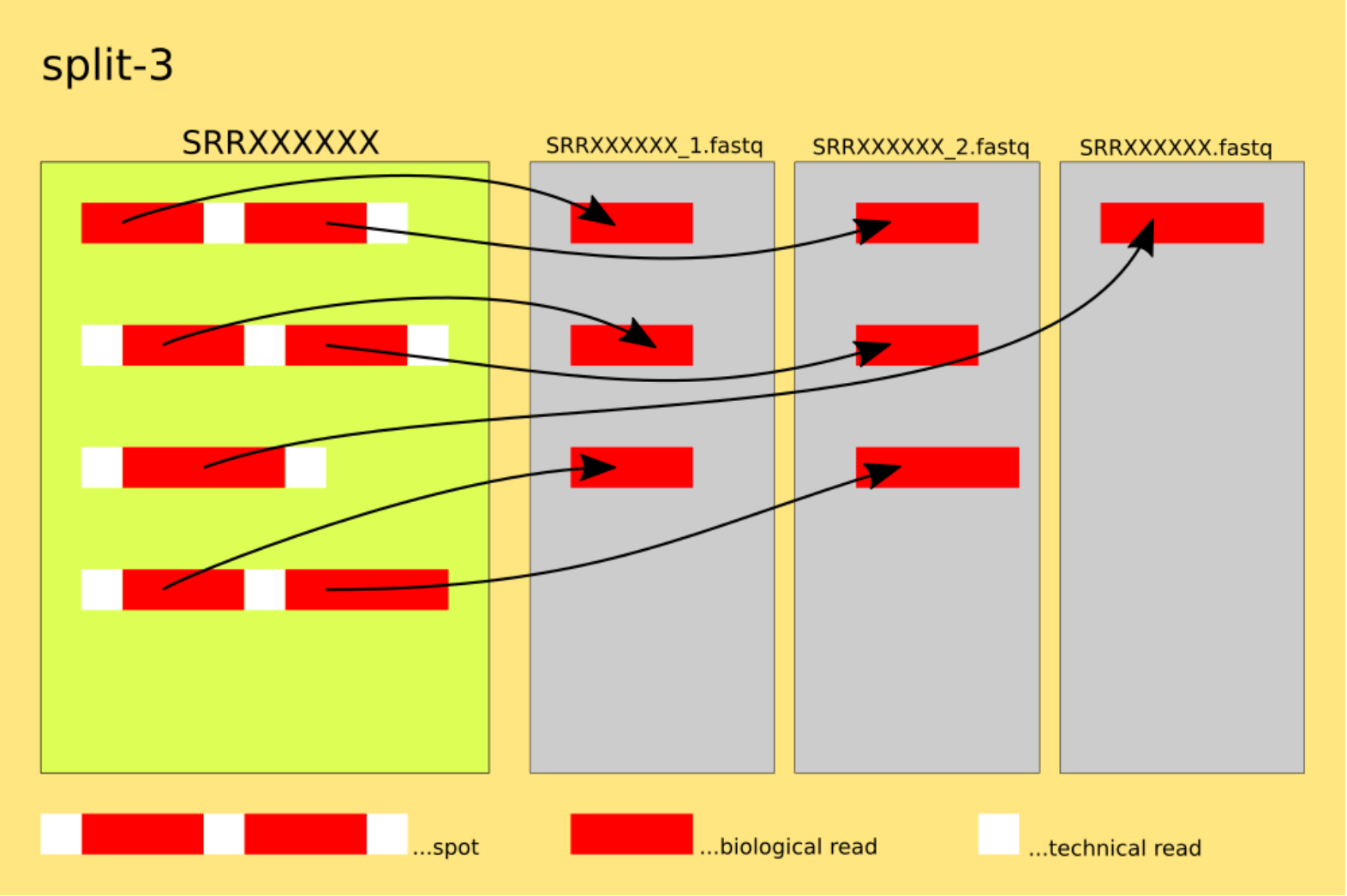
SRA Toolkit – Step 2: convert SRA to FASTQ files

After downloading .sra, convert to FASTQ format using fasterq-dump

```
# Basic conversion
fasterq-dump SRR1234567

# Specify output directory
fasterq-dump SRR1234567 -O /data/fastq/

# Use multi-threading for large files
fasterq-dump SRR1234567 -e 8 -O /data/fastq/
```

For spots having 2 reads (paired-end reads), the reads are written into the ***_1.fastq** and ***_2.fastq** files. Unmated reads are placed in ***.fastq**.

How the fasterq-dump work



Analyzed processed GEO data using GEO2R

Series GSE124226 Query DataSets for GSE124226

Status	Public on Jan 24, 2019
Title	Gene Expression Data of Adipose Stem Cells from Normal-Weight Polycystic Ovary Syndrome Women vs. Controls
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	In vitro studies of subcutaneous (SC) abdominal adipose stem cells (ASC) from women with polycystic ovary syndrome (PCOS) show altered ASC commitment to preadipocytes and differentiation to mature adipocytes related to hyperandrogenism. The goal of the study is to use microarrays to examine whether SC abdominal ASC gene expression are altered in normal-weight PCOS women and correlated with hyperandrogenemia and/or insulin resistance, which are prevalent clinical pathologies of PCOS.
Overall design	ASCs were isolated from SC adipose tissue following SC abdominal biopsy in 4 PCOS women and 4 age and BMI matched controls for gene expression comparison. First generation of stem cells were cultured until cells reached confluency and isolated for RNA extraction and hybridization on Affymetrix microarrays. Grant#: P50 HD071836 Title: Androgen excess as a mechanism for adipogenic dysfunction in PCOS women Period: 2018-2022
Contributor(s)	Phan JD, Leung KL, Ding X, Li X, Chazenbalk GD
Citation(s)	Dumesic DA, Phan JD, Leung KL, Grogan TR et al. Adipose Insulin Resistance in Normal-Weight Women With Polycystic Ovary Syndrome. *J Clin Endocrinol Metab* 2019 Jun 1;104(6):2171-2183. PMID: 30649347

Analyze with GEO2R

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. Full instructions

GEO accession GSE124226 Set

Gene Expression Data of Adipose Stem Cells from Normal-Weight Polycystic Ovary Syndrome Women vs. Controls

Samples Define groups Selected **0** out of **8** samples

Columns Set

Group	Accession	Title	Source name	Female type
-	GSM3526020	ASC-3004-control	subcutaneous adipose tissue, control	control
-	GSM3526021	ASC-3006-PCOS	subcutaneous adipose tissue, PCOS	PCOS
-	GSM3526022	ASC-3010-PCOS	subcutaneous adipose tissue, PCOS	PCOS
-	GSM3526023	ASC-3019-PCOS	subcutaneous adipose tissue, PCOS	PCOS
-	GSM3526024	ASC-3027-control	subcutaneous adipose tissue, control	control
-	GSM3526025	ASC-3034-control	subcutaneous adipose tissue, control	control
-	GSM3526026	ASC-3036-control	subcutaneous adipose tissue, control	control
-	GSM3526027	ASC-3042-PCOS	subcutaneous adipose tissue, PCOS	PCOS

GEO2R | Options | Profile graph | R script

Quick start

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, source and characteristics) columns to help determine which Samples belong to which group.
- Click 'Analyze' to perform the calculation with default settings.
- You may change settings in the Options tab.

How to use

Analyze

Functions of GEO2R

- Web-based analysis tool integrated into the NCBI GEO database.
- Allows users to compare two or more groups of samples within a GEO Series (GSE).
- Provides an interactive interface without requiring programming skills.
- Performs differential expression analysis using the limma package.
- Automatically normalizes and log-transforms data when appropriate.
- Enables sample grouping and visualization of group differences.
- Generates volcano plots and boxplots for quick exploratory analysis.
- Outputs a table of differentially expressed genes with logFC, p-values, and adjusted p-values (FDR).
- Allows downloading of results in text or Excel format.
- Provides links to annotation and gene information directly from the result table.
- Works primarily with **processed microarray or RNA-seq expression matrices available** in GEO.

GEO2R: To learn more, please read this page

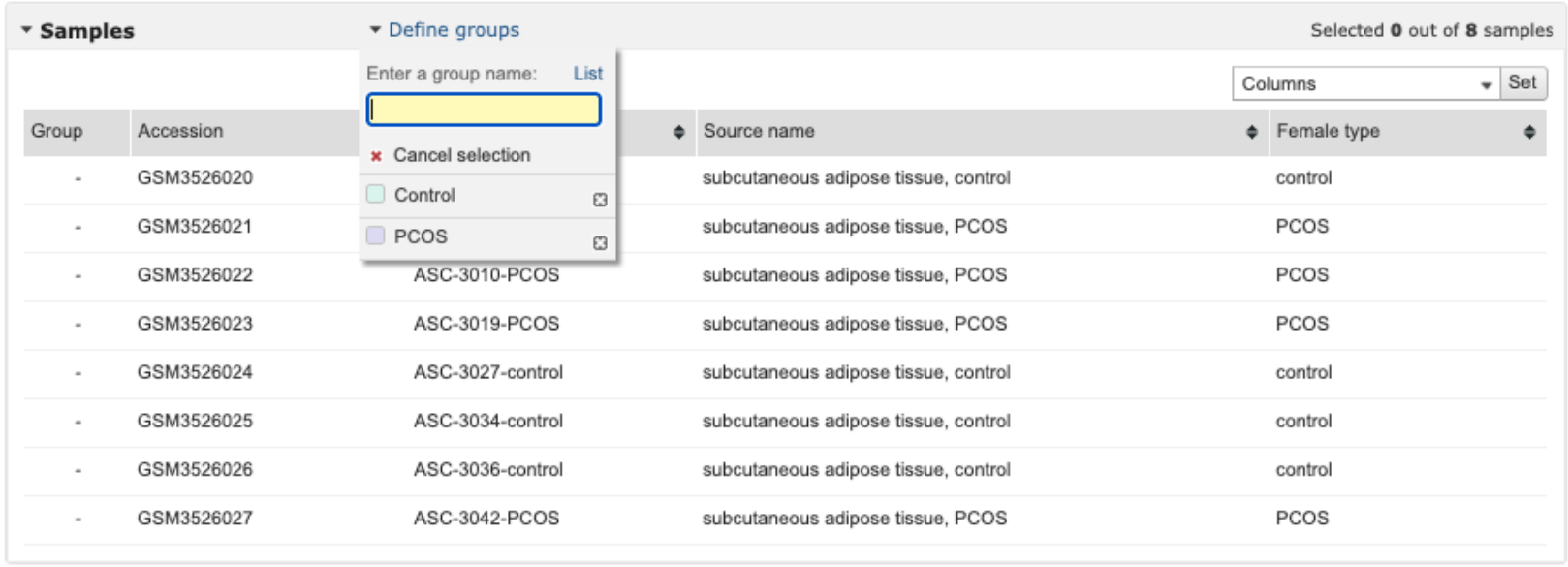
https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html

About GEO2R

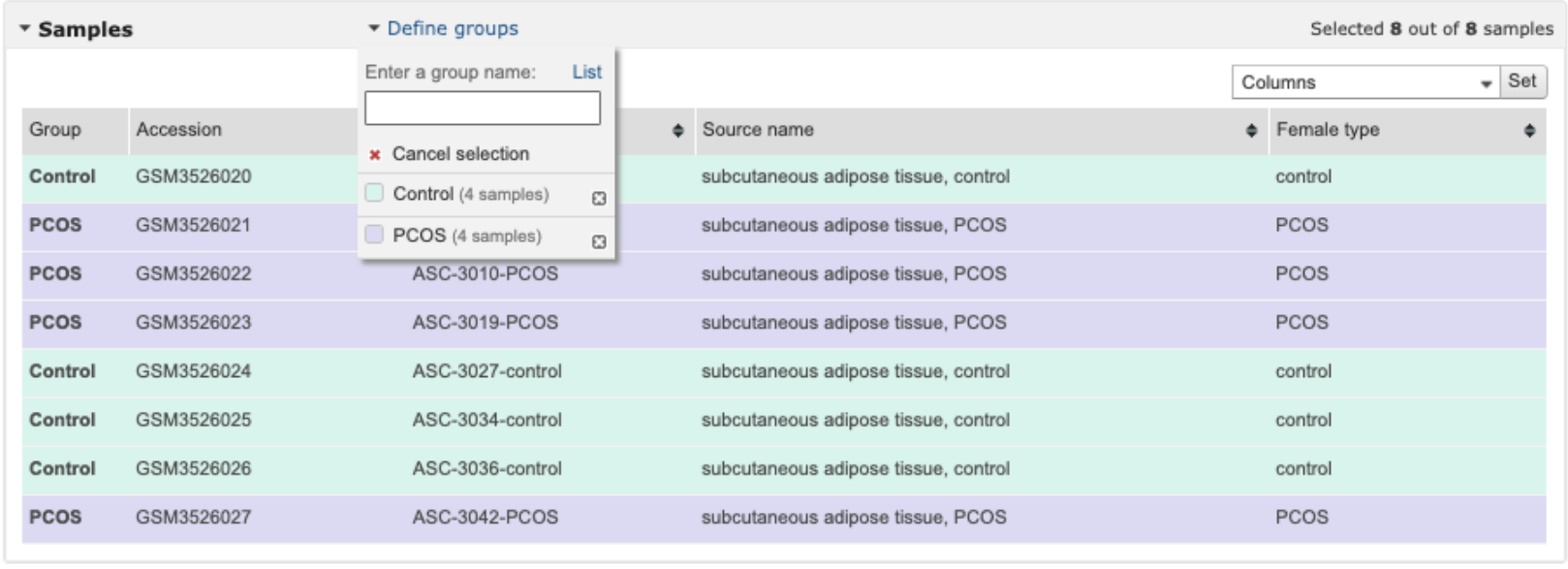
- Background
  - RNA-seq data
  - Microarray data
- How to use
  - Enter a Series accession number
  - Define Sample groups
  - Assign Samples to each group
  - Perform the analysis
  - Top differentially expressed genes
  - Visualization
  - Tutorial video
- Edit options and features
  - Options
  - Profile graph
  - R script
- Limitations and caveats
- More information and references
  - Summary statistics
  - General references
  - Adjustment test references

GEO2R: define groups

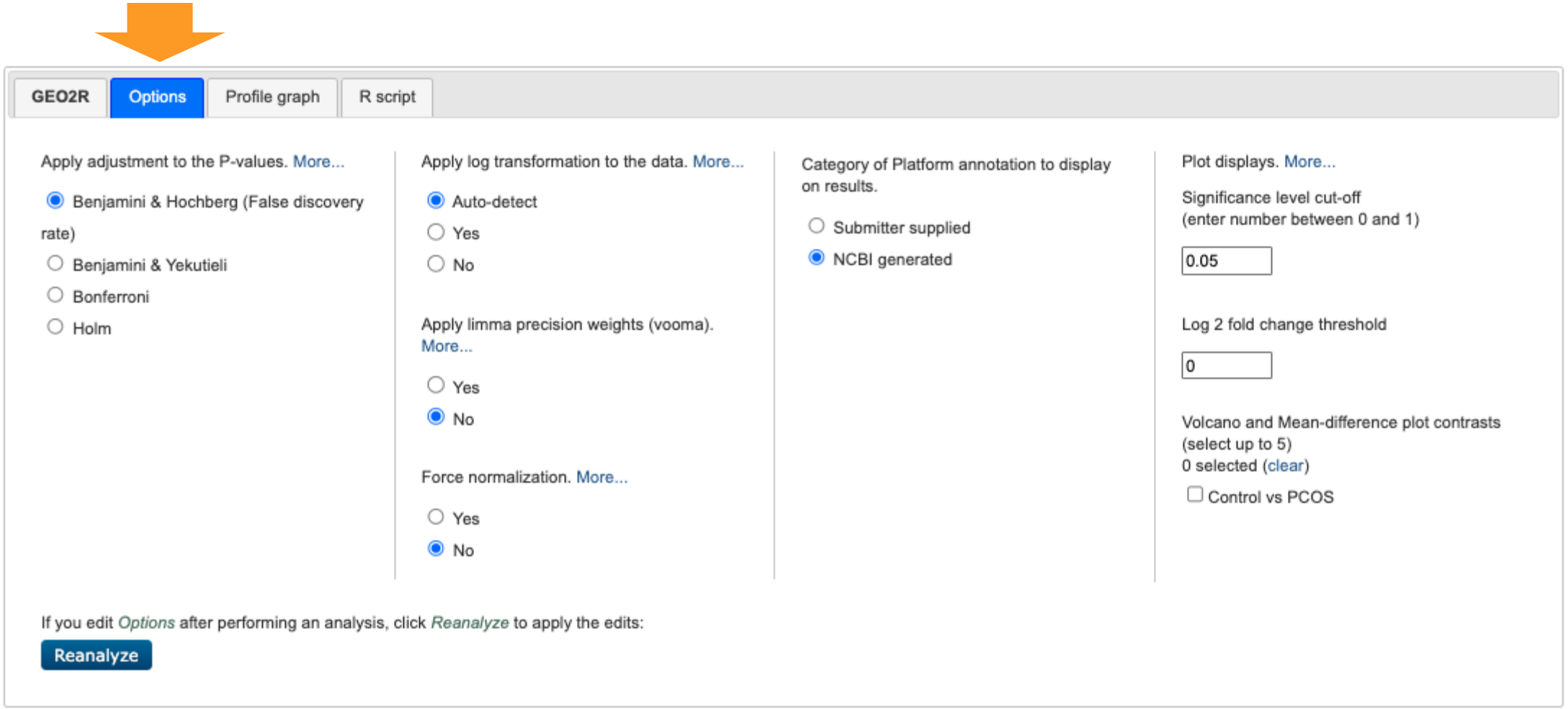
Step 1.



Step 2.



GEO2R: set options

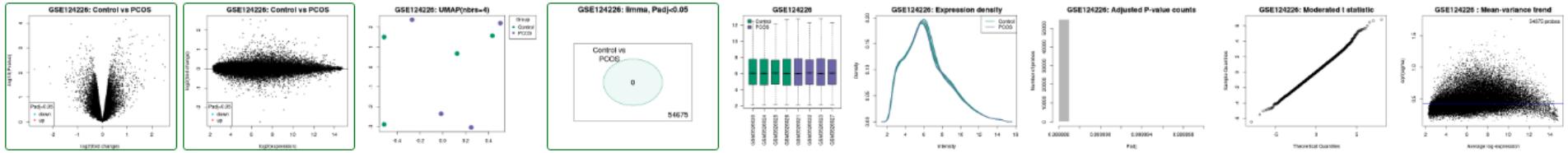


GEO2R: analysis results

GEO2R | Options | Profile graph | R script

Reanalyze if you changed any options.

Visualization



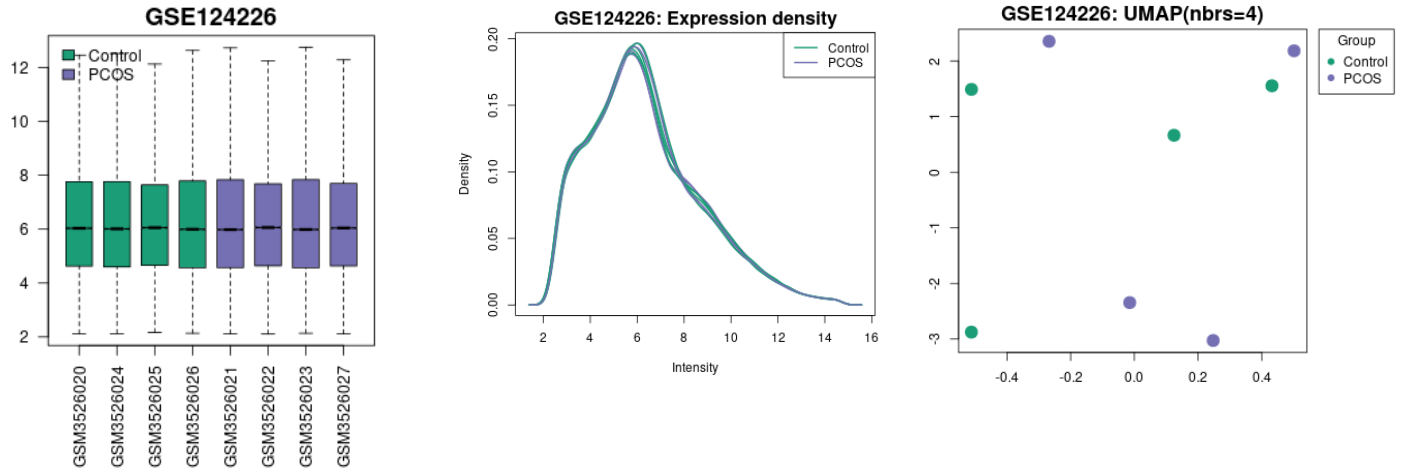
Top differentially expressed genes

Download full table Select columns

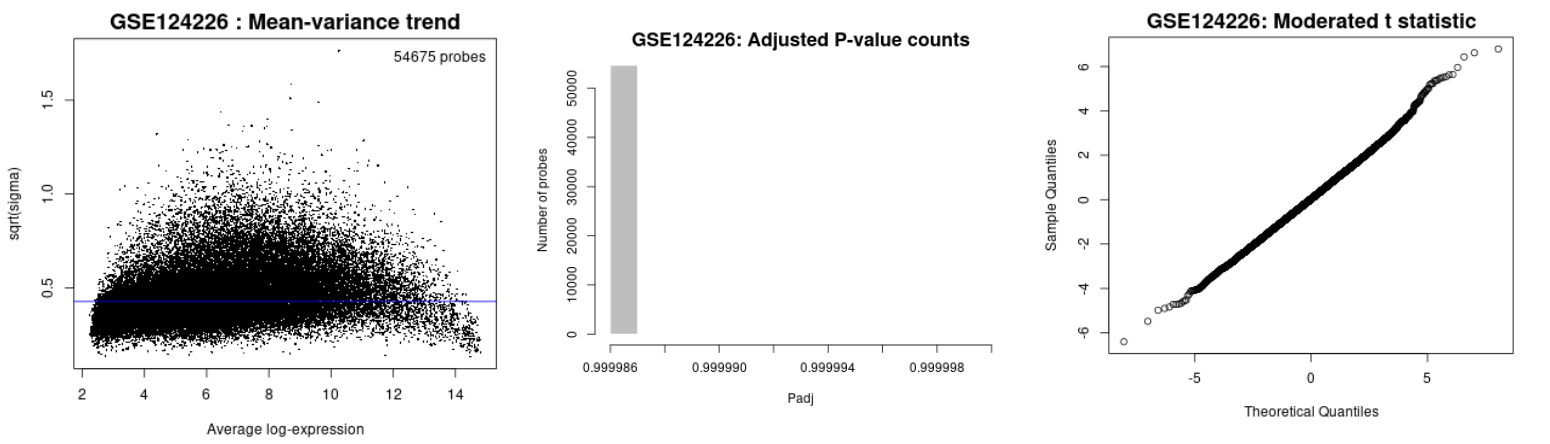
ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
219073_s_at	1	0.0000705	6.79	-2.9	0.941	OSBPL10	oxysterol binding protein like 10
213802_at	1	0.0000861	6.62	-2.93	0.861	PRSS12	protease, serine 12
203200_s_at	1	0.0001075	6.43	-2.95	0.541	MTRR	5-methyltetrahydrofolate-homoc…
213975_s_at	1	0.000111	-6.4	-2.95	-1.269	LYZ	lysozyme
204320_at	1	0.0001923	5.96	-3.02	1.146	COL11A1	collagen type XI alpha 1 chain
37892_at	1	0.0002895	5.64	-3.07	1.396	COL11A1	collagen type XI alpha 1 chain
241404_at	1	0.0002957	5.62	-3.07	2.301		
225275_at	1	0.0003315	5.53	-3.08	1.295	EDIL3	EGF like repeats and discoidin d…
204337_at	1	0.0003439	5.51	-3.09	1.207	RGS4	regulator of G-protein signaling 4
204529_s_at	1	0.0003509	5.49	-3.09	1.805	TOX	thymocyte selection associated …
235587_at	1	0.0003553	-5.48	-3.09	-0.613	PAXIP1-AS1	PAXIP1 antisense RNA 1 (head …
230660_at	1	0.0003737	5.44	-3.1	1.083	SERTAD4	SERTA domain containing 4
206704_at	1	0.0004093	5.37	-3.11	0.654	CLCN5	chloride voltage-gated channel 5
201976_s_at	1	0.0004149	5.36	-3.11	0.702	MYO10	myosin X

GEO2R: visualization

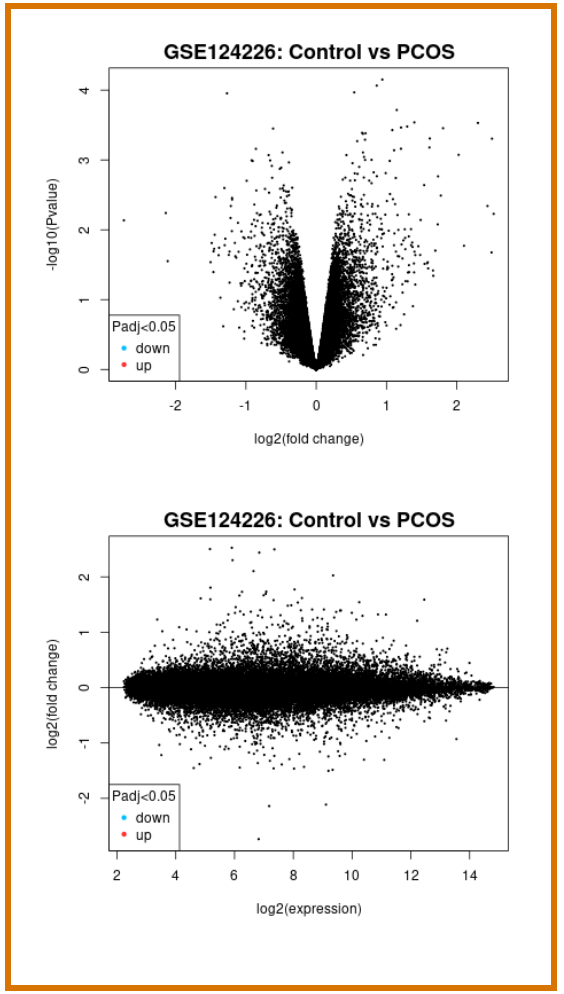
To evaluate the quality of data



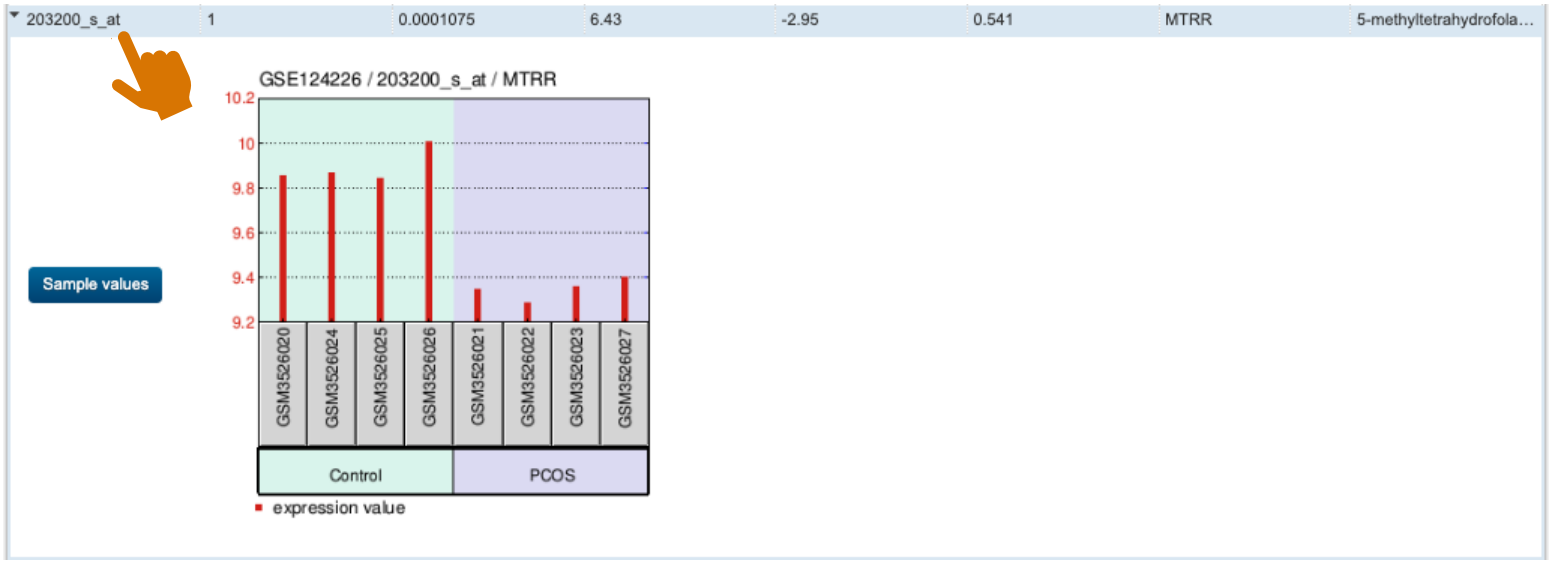
To evaluate the quality of statistic test

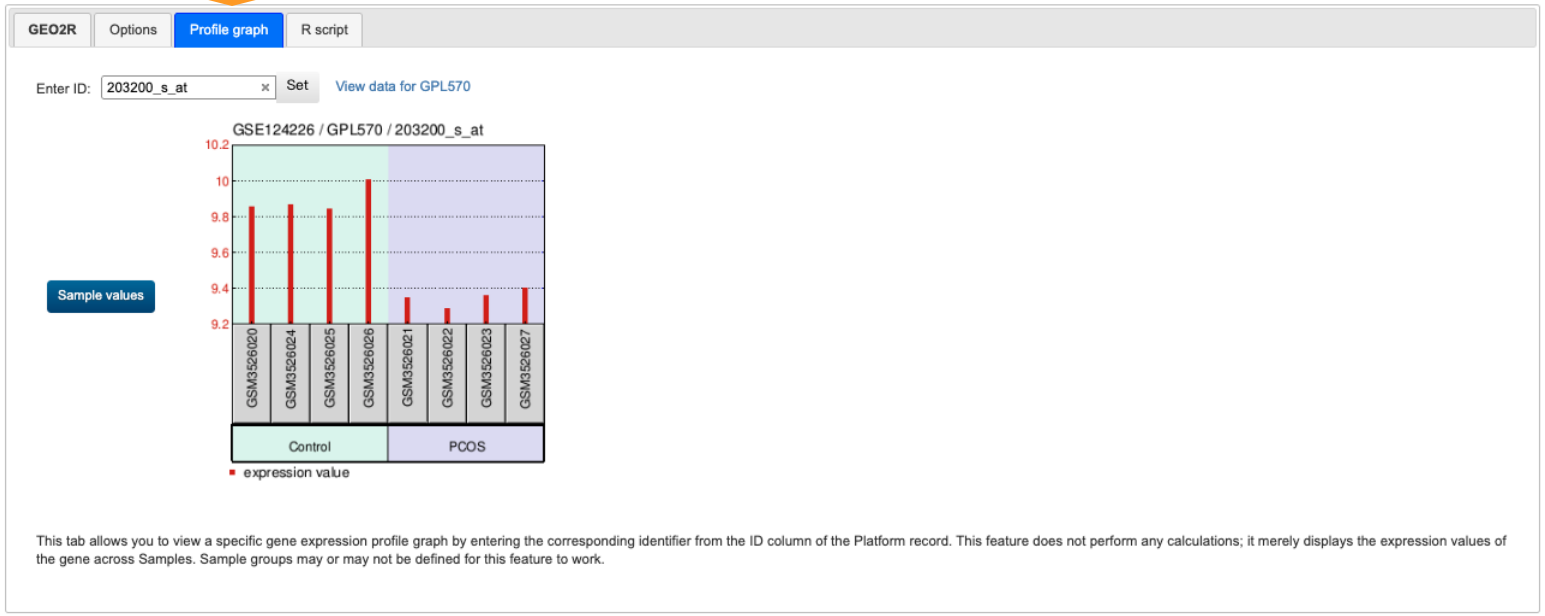


To display the differentially expressed genes

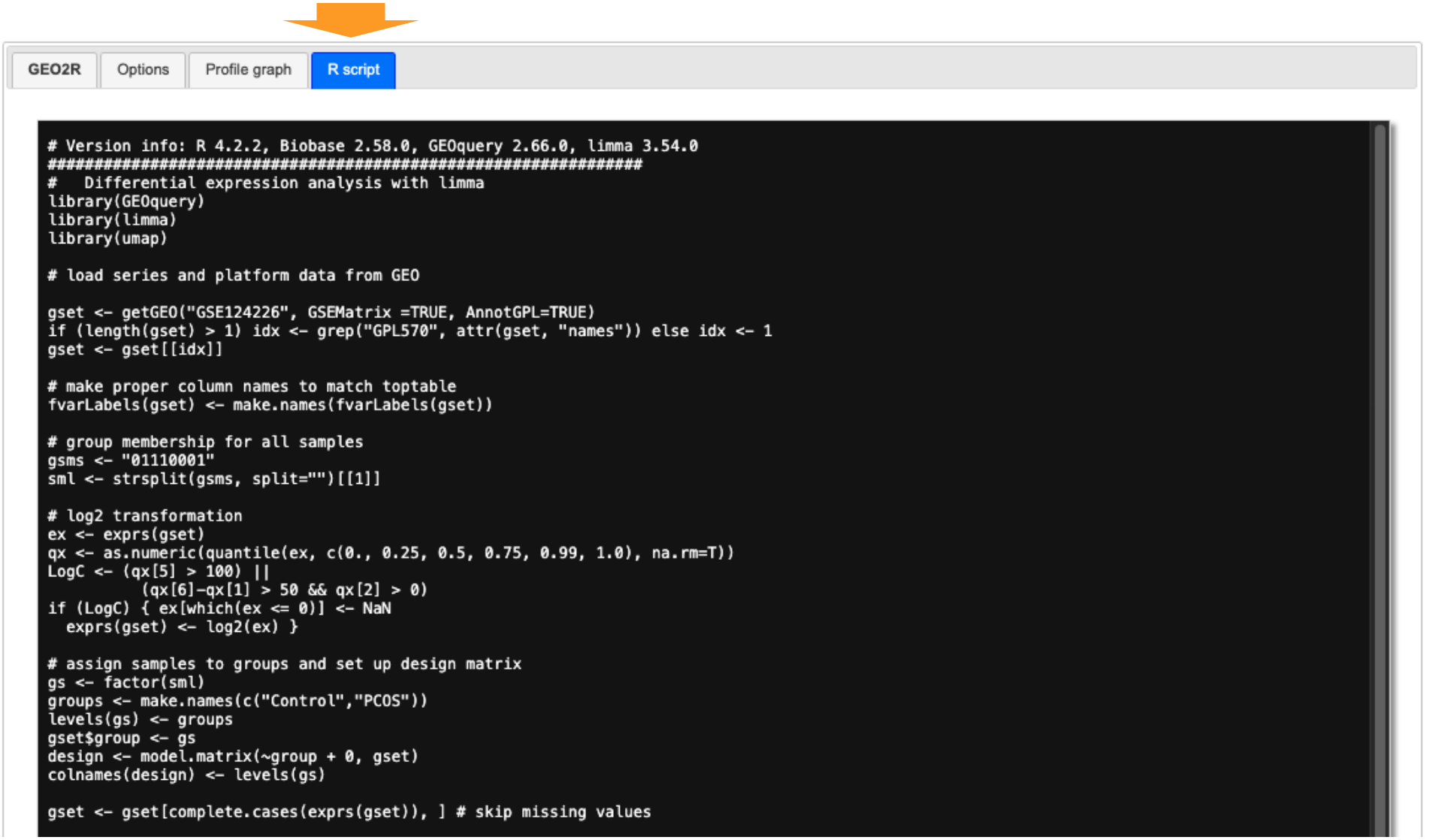


GEO2R: gene expression exploration





GEO2R: automatically generated analysis script



GEQquery (R package)

GEOquery

This is the **released** version of GEOquery; for the devel version, see GEOquery.

Get data from NCBI Gene Expression Omnibus (GEO)

platforms all | rank 61 / 2361 | support 0 / 0 | in Bioc 19.5 years | build error | updated before release | dependencies 76

DOI: 10.18129/B9.bioc.GEOquery

Bioconductor version: Release (3.22)

The NCBI Gene Expression Omnibus (GEO) is a public repository of microarray data. Given the rich and varied nature of this resource, it is only natural to want to apply BioConductor tools to these data. GEOquery is the bridge between GEO and BioConductor.

Author: Sean Davis [aut, cre]

Maintainer: Sean Davis <seandavi at gmail.com>

Citation (from within R, enter citation("GEOquery")):

Davis S, Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics*, **14**, 1846–1847. doi:10.1093/bioinformatics/btm254.

Installation

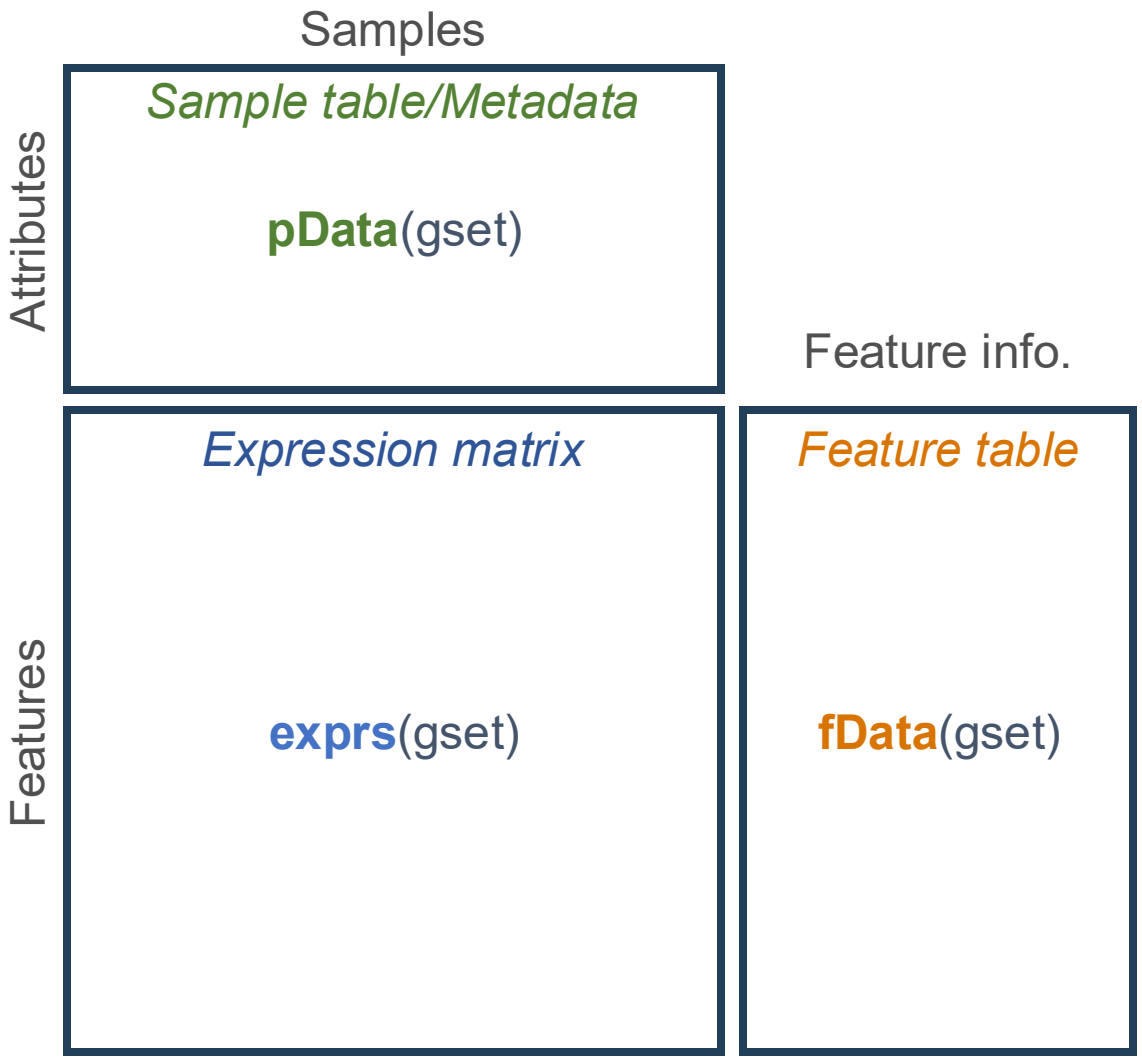
To install this package, start R (version "4.5") and enter:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GEOquery")
```

Fetch GEO processing data via GEOquery

```
library("GEOquery")

gseid <- "GSE124226"
gset <- getGEO(gseid, GSEMatrix =TRUE, AnnotGPL=TRUE)
gset1 <- getGEO(gseid, GSEMatrix =TRUE, AnnotGPL=FALSE)

class(gset)
#> 'list’
names(gset)
#> 'GSE124226_series_matrix.txt.gz'

gset <- gset[[1]]
gset1 <- gset1[[1]]

class(gset)
#> 'ExpressionSet'
```

Data structure of ExpressionSet object



Expression matrix: exprs()

```
exprs(gset)
```

A matrix: 54675 × 8 of type dbl

	GSM3526020	GSM3526021	GSM3526022	GSM3526023	GSM3526024	GSM3526025	GSM3526026	GSM3526027
1007_s_at	9.54885	9.76681	9.67118	9.67082	9.70476	9.66295	9.67862	9.21119
1053_at	8.61334	7.69760	8.53742	8.06942	8.17666	7.85192	7.75730	7.99570
117_at	6.20375	6.34466	6.46556	6.30172	6.45548	6.62242	6.21767	6.53739
121_at	9.09007	8.87642	9.23535	8.73644	8.78167	9.16652	8.67744	9.43601
1255_g_at	3.85190	3.68445	3.89538	3.59723	3.91981	3.71919	3.68602	3.75776
1294_at	7.95796	7.97800	7.81674	8.06442	7.66563	6.96192	8.01771	7.41316
1316_at	7.43911	7.65307	8.00262	7.18998	7.71863	7.73945	7.14290	7.64515
1320_at	6.55236	6.48452	6.36801	6.44498	6.65787	6.30494	6.25951	6.44187

Phenotype/sample data: pData()

```
pData(gset)
```

	title	geo_accession	status	submission_date	last_update_date	type	channel_count	source_name_ch1	organism_ch1	characteristics_ch1	...	con
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	...	
GSM3526020	ASC-3004-control	GSM3526020	Public on Jan 24 2019	Dec 20 2018	Jan 24 2019	RNA	1	subcutaneous adipose tissue, control	Homo sapiens	female type: control	...	
GSM3526021	ASC-3006-PCOS	GSM3526021	Public on Jan 24 2019	Dec 20 2018	Jan 24 2019	RNA	1	subcutaneous adipose tissue, PCOS	Homo sapiens	female type: PCOS	...	
GSM3526022	ASC-3010-PCOS	GSM3526022	Public on Jan 24 2019	Dec 20 2018	Jan 24 2019	RNA	1	subcutaneous adipose tissue, PCOS	Homo sapiens	female type: PCOS	...	
GSM3526023	ASC-3019-PCOS	GSM3526023	Public on Jan 24 2019	Dec 20 2018	Jan 24 2019	RNA	1	subcutaneous adipose tissue, PCOS	Homo sapiens	female type: PCOS	...	

AnnotGPL=TRUE / FALSE

AnnotGPL: A boolean defaulting to FALSE as to whether or not to use the Annotation GPL information. These files are nice to use because they contain up-to-date information remapped from Entrez Gene on a regular basis.

AnnotGPL=TRUE

```
names(fData(gset))
```

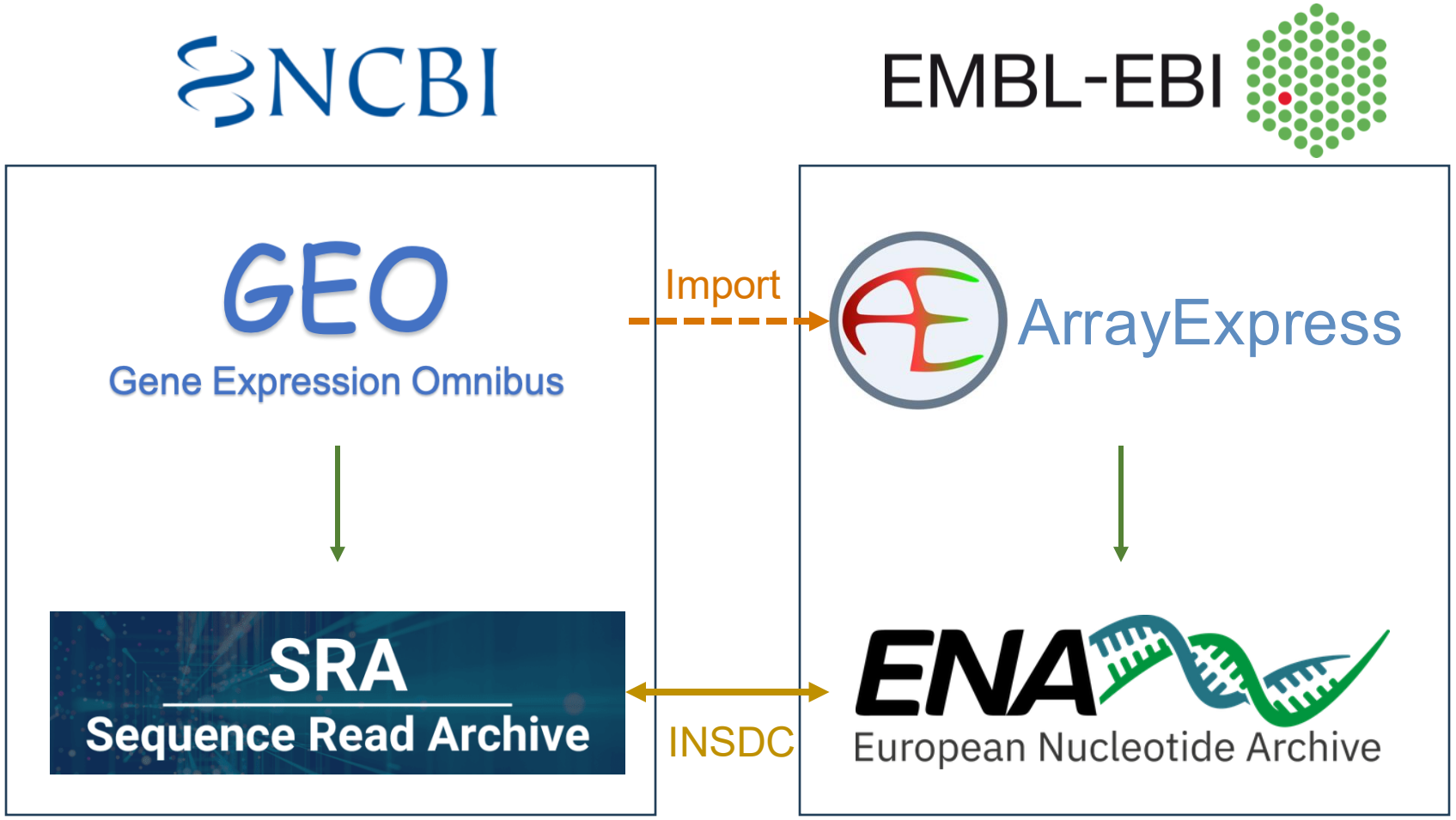
'ID' · 'Gene title' · 'Gene symbol' · 'Gene ID' · 'UniGene title' · 'UniGene symbol' · 'UniGene ID' · 'Nucleotide Title' · 'GI' · 'GenBank Accession' · 'Platform_CLONEID' · 'Platform_ORF' · 'Platform_SPOTID' · 'Chromosome location' · 'Chromosome annotation' · 'GO:Function' · 'GO:Process' · 'GO:Component' · 'GO:Function ID' · 'GO:Process ID' · 'GO:Component ID'

AnnotGPL=FALSE

```
names(fData(gset1))
```

'ID' · 'GB_ACC' · 'SPOT_ID' · 'Species Scientific Name' · 'Annotation Date' · 'Sequence Type' · 'Sequence Source' · 'Target Description' · 'Representative Public ID' · 'Gene Title' · 'Gene Symbol' · 'ENTREZ_GENE_ID' · 'RefSeq Transcript ID' · 'Gene Ontology Biological Process' · 'Gene Ontology Cellular Component' · 'Gene Ontology Molecular Function'

Collaboration and data exchange between NCBI and EMBL-EBI repositories



INSDC (International Nucleotide Sequence Database Collaboration) ensures data synchronization among NCBI (SRA), EBI (ENA), and DDBJ (DRA, Japan).

Databases for archiving and distributing human-sensitive data

Human genomics data, Controlled data

NCBI (U.S.)
https://dbgap.ncbi.nlm.nih.gov/
NIH-funded human genomic studies

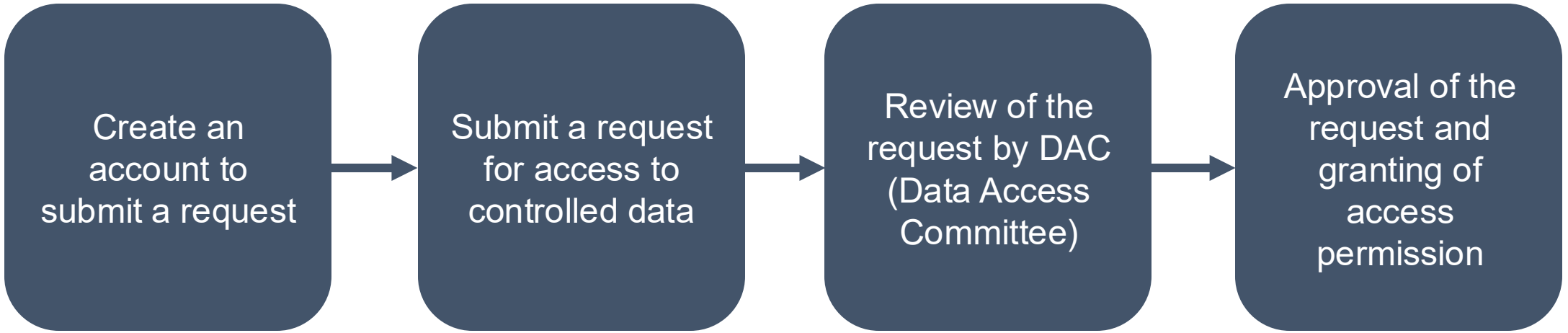
EMBL-EBI (Europe)
https://ega-archive.org/

Common Features of dbGaP and EGA

- **Purpose**
  Both are repositories for human genomic and phenotypic data that require controlled access due to privacy and ethical concerns.
- **Data type**
  Store individual-level, potentially identifiable human data — such as genotype, sequencing, clinical, and phenotype data.
- **Controlled access model**
  Researchers must apply for access via a Data Access Committee (DAC), specifying study purpose and data-use agreement.
- **Ethical and legal compliance**
  Governed by informed consent, IRB/ethics approval, and data-use restrictions defined by the original study.
- **Metadata and study registration**
  Each study includes metadata describing cohort design, sample type, and consent group.
- **Data security**
  Both provide secure infrastructure for storage, transfer, and auditing of sensitive datasets.

dbGaP / EGA Data Access Procedure

https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/accessing-data/dbgap
https://ega-archive.org/access/request-data/how-to-request-data/



- dbGAP: only **PIs** can create the account (eRA)

- Research proposal
- Data Access Agreement (Institute to Institute)

Who can apply for access to controlled data in dbGaP?

Those eligible to apply for data access must be a permanent employee of their institution and either:

- At a level equivalent to a tenure-track professor
- Senior scientist with responsibilities that may include laboratory or research program administration and oversight.

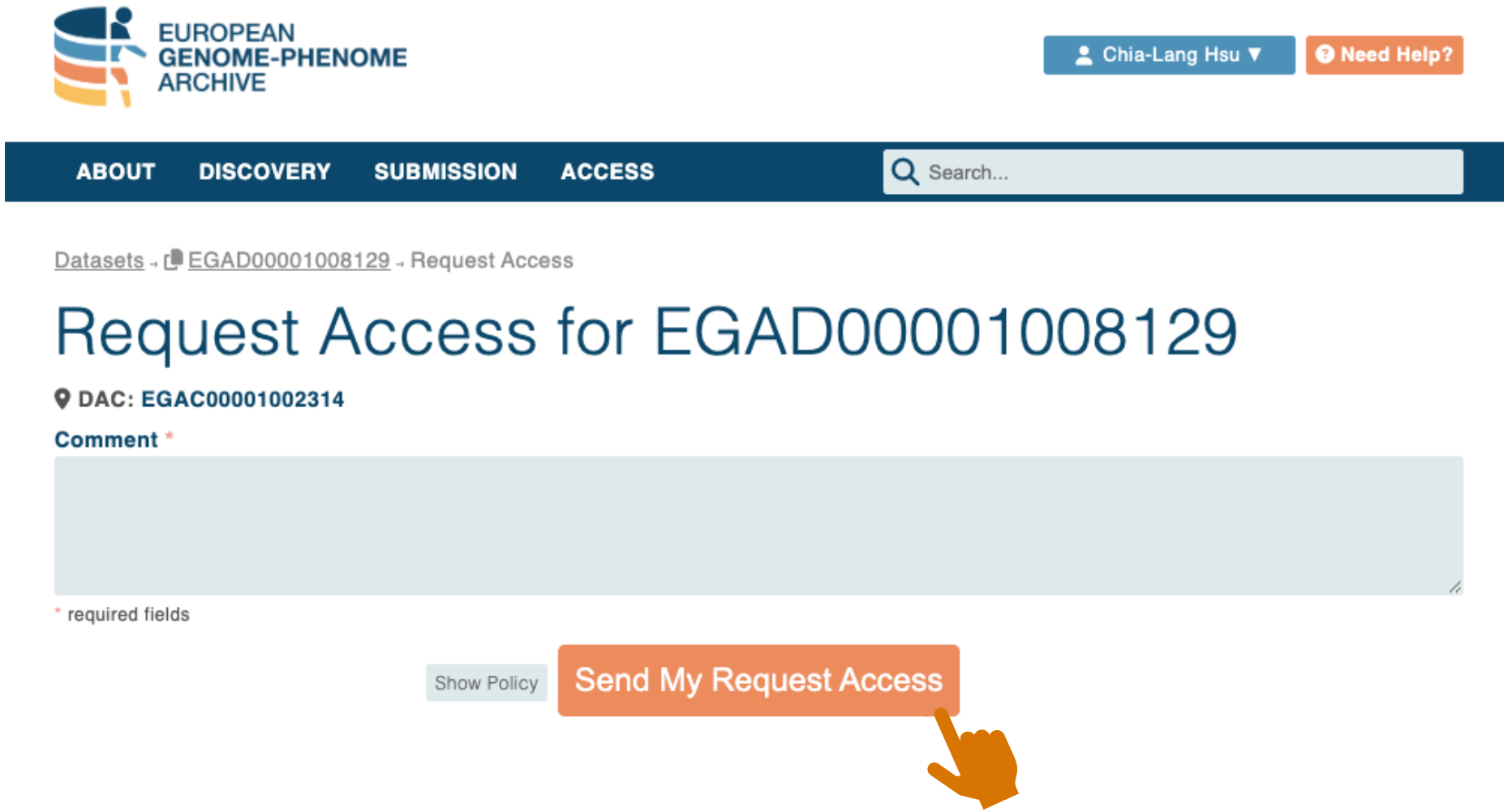
Laboratory staff and trainees such as graduate students and postdoctoral fellows are **NOT** permitted to submit data access requests in dbGaP. However, they may be part of projects using such data when the projects are overseen by an eligible user.

Investigators not affiliated with NIH must have an **eRA** account to access dbGaP.

https://www.youtube.com/watch?v=m0xp_cCO7kA

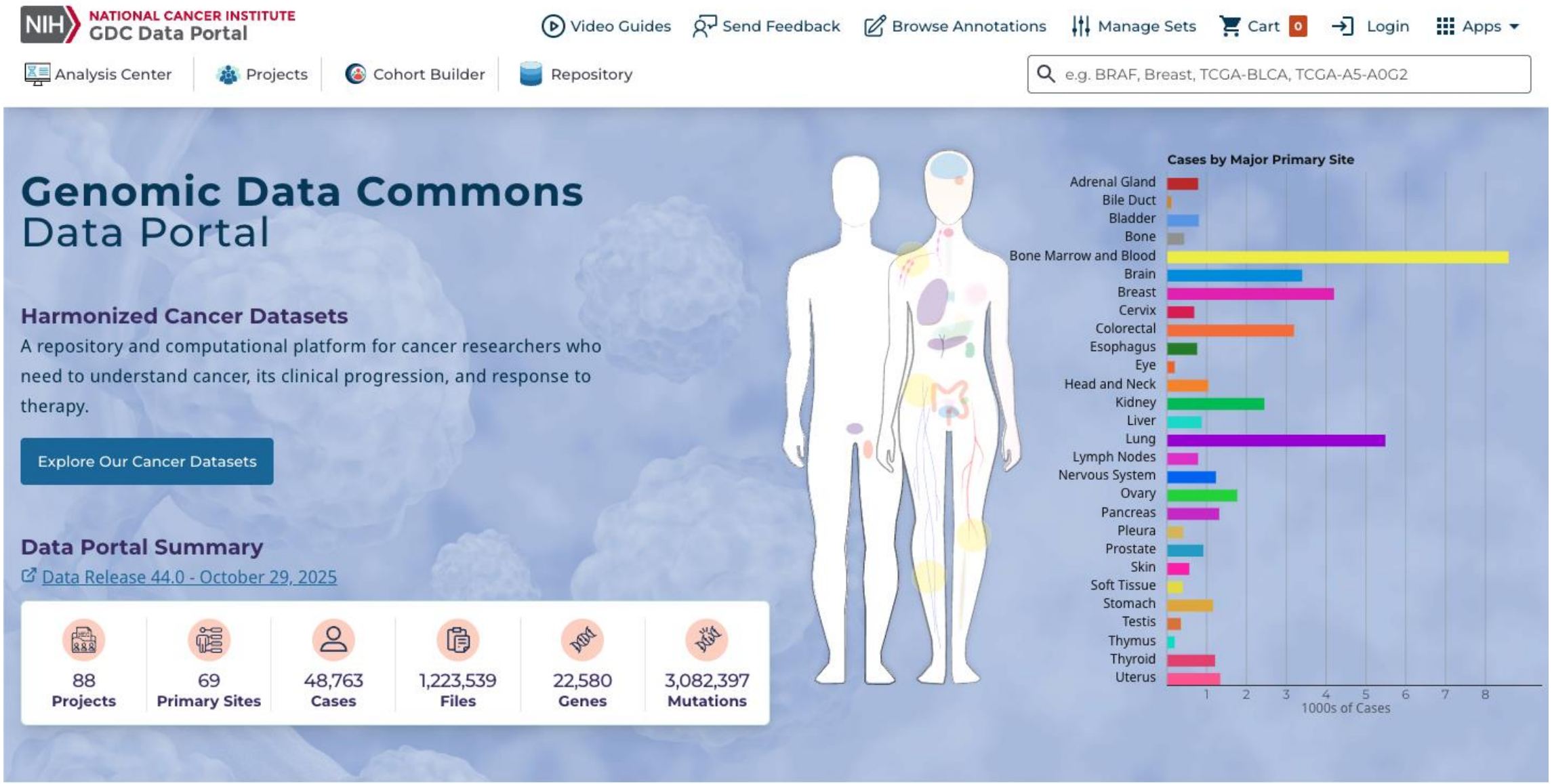
Submit a request to access controlled data in EGA

Submit a request to access controlled data in EGA



Accessing TCGA data: GDC data portal

https://portal.gdc.cancer.gov/



Accessing TCGA data: GDC data portal

Open-access data can be downloaded directly,
whereas **controlled-access data** require approval through dbGaP before download.

Cart	Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
	Open	6157fcee-36c0-4e9f-a086-01d64b8db3de_noid_Grn.idat	1	TCGA-BRCA	DNA Methylation	IDAT	719.56 kB	0
	Controlled	907fec58-2e0f-4501-b0a8-150b98e7ca7b.wxs.muse.raw_somatic_mutation.vcf.gz	1	TCGA-BRCA	Simple Nucleotide Variation	VCF	34.33 kB	2
	Open	TCGA-E2-A15C-01A-31D-A895-36.WholeGenome.RP-1657.cr.igv.reheader.seg.txt	1	TCGA-BRCA	Copy Number Variation	TXT	31.08 kB	0
	Controlled	TCGA_BRCA.637493d7-3d92-4df3-8442-e1c2b624264a.wxs.Pindel.somatic_annotation.vcf.gz	1	TCGA-BRCA	Simple Nucleotide Variation	VCF	186.18 kB	0
	Controlled	TCGA-E2-A15C-01A-31R-A12C-13_mirna_gdc_realn.bam	1	TCGA-BRCA	Sequencing Reads	BAM	165.84 MB	0
	Controlled	df5d5895-ac41-43f5-a9af-ec3f0f8a940e.wxs.Pindel.aliquot.maf.gz	1	TCGA-BRCA	Simple Nucleotide Variation	MAF	85.84 kB	0

Accessing processed TCGA data: R package TCGABiolink

TCGAbiolinks

This is the **released** version of TCGAbiolinks; for the devel version, see TCGAbiolinks.

TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

DOI: 10.18129/B9.bioc.TCGAbiolinks

Bioconductor version: Release (3.22)

The aim of TCGAbiolinks is : i) facilitate the GDC open-access data retrieval, ii) prepare the data using the appropriate pre-processing strategies, iii) provide the means to carry out different standard analyses and iv) to easily reproduce earlier research results. In more detail, the package provides multiple methods for analysis (e.g., differential expression analysis, identifying differentially methylated regions) and methods for visualization (e.g., survival plots, volcano plots, starburst plots) in order to easily develop complete analysis pipelines.

Author: Antonio Colaprico, Tiago Chedraoui Silva, Catharina Olsen, Luciano Garofano, Davide Garolini, Claudia Cava, Thais Sabedot, Tathiane Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, Houtan Noushmehr

Maintainer: Tiago Chedraoui Silva <tiagochst at gmail.com>, Antonio Colaprico <axc1833 at med.miami.edu>

Citation (from within R, enter citation("TCGAbiolinks")):

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H (2015). "TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." *Nucleic Acids Research*. doi:10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507.

Silva, C T, Colaprico, Antonio, Olsen, Catharina, D'Angelo, Fulvio, Bontempi, Gianluca, Ceccarelli, Michele, Noushmehr, Houtan (2016). "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages." *F1000Research*, **5**.

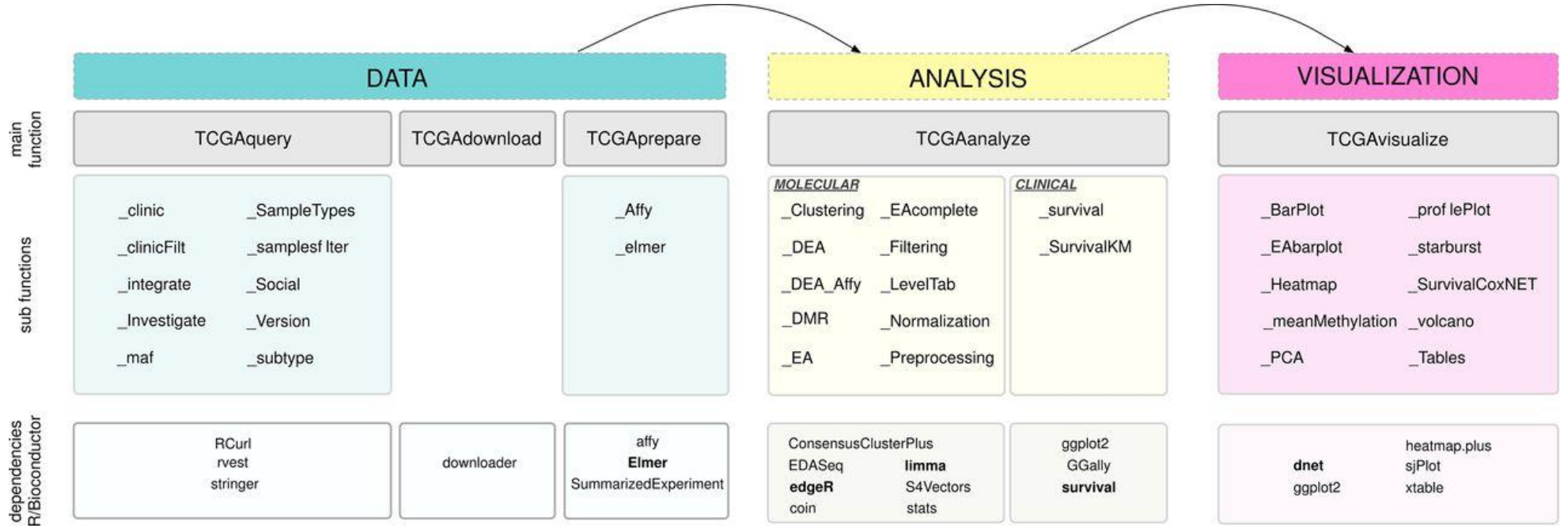
Mounir, Mohamed, Lucchetta, Marta, Silva, C T, Olsen, Catharina, Bontempi, Gianluca, Chen, Xi, Noushmehr, Houtan, Colaprico, Antonio, Papaleo, Elena (2019). "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx." *PLoS computational biology*, **15**(3), e1006701.

Installation

To install this package, start R (version "4.5") and enter:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("TCGAbiolinks")
```

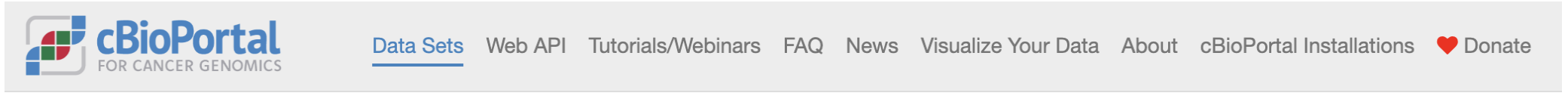
Overview of TCGAbiolinks functions.



Colaprico A. et al. 2016. NAR

Accessing processed TCGA data: cBioPortal

https://www.cbioportal.org/



Name ▲		Reference	All	Mutations	CNA	RNA-Seq
Acinar Cell Carcinoma of the Pancreas (JHU, J Pathol 2014)		Jial et al. J Pathol 2014	23	23	0	0
Acral Melanoma (TGEN, Genome Res 2017)		Liang et al. Genome Res 2017	38	38	38	36
Acute Leukemias of Ambiguous Lineage (TARGET GDC, 2025)			251	123	0	0
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2015)		Andersson et al. Nat Genet 2015	93	93	0	0
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2016)		Zhang et al. Nat Genet 2016	73	73	0	0
Acute Myeloid Leukemia (OHSU, Cancer Cell 2022)		Bottomly et al. Cancer Cell 2022	942	903	0	698
Acute Myeloid Leukemia (OHSU, Nature 2018)		Tyner et al. Nature 2018	672	622	0	451
Acute Myeloid Leukemia (TARGET GDC, 2025)			2766	584	92	0
Acute Myeloid Leukemia (TCGA GDC, 2025)			200	72	190	0
Acute Myeloid Leukemia (TCGA, Firehose Legacy)			200	197	191	173
Acute Myeloid Leukemia (TCGA, NEJM 2013)		TCGA, NEJM 2013	200	200	191	173
Acute Myeloid Leukemia (TCGA, PanCancer Atlas)		TCGA, Cell 2018	200	200	191	173

Accessing processed TCGA data: Xena

https://xenabrowser.net/

DATA SETS VISUALIZATION TRANSCRIPTS SINGLECELL DATA HUBS VIEW MY DATA HELP MORE TOOLS

237 Cohorts, 2256 Datasets

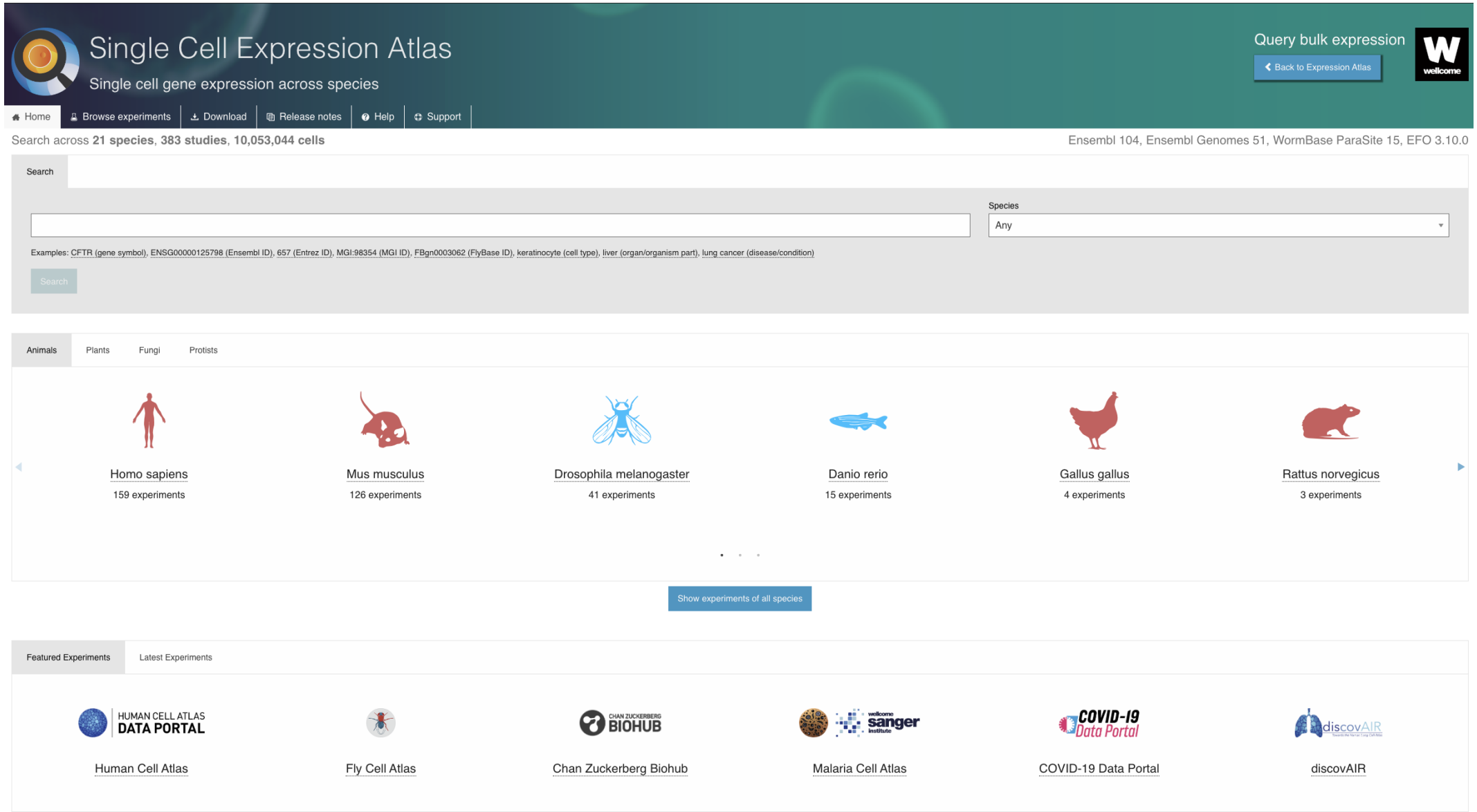
10x visium Human Ovarian Cancer (7 datasets)
10x visium Human Ovarian Cancer enhanced resolution (3 datasets)
10x visium Mouse Brain Coronal (6 datasets)
10x visium Mouse Sagittal Anterior1 (6 datasets)
10x visium V1_Breast_Cancer_Block_A_Section_1 (4 datasets)
Acute lymphoblastic leukemia (Mullighan 2008) (3 datasets)
Breast Cancer (Caldas 2007) (3 datasets)
Breast Cancer (Chin 2006) (3 datasets)
Breast Cancer (Haverty 2008) (2 datasets)
Breast Cancer (Hess 2006) (2 datasets)
Breast Cancer (Miller 2005) (2 datasets)
Breast Cancer (vantVeer 2002) (2 datasets)
Breast Cancer (Vijver 2002) (2 datasets)
Breast Cancer (Yau 2010) (2 datasets)
Breast Cancer Cell Lines (Heiser 2012) (4 datasets)
Breast Cancer Cell Lines (Neve 2006) (2 datasets)
Cancer Cell Line Encyclopedia (Breast) (4 datasets)
Cancer Cell Line Encyclopedia (CCLE) (9 datasets)

Active Data Hubs

- [x] My computer hub
- [x] UCSC Public Hub
- [x] TCGA Hub
- [x] Pan–Cancer Atlas Hub
- [x] ICGC Hub
- [x] PCAWG Hub
- [x] UCSC Toil RNA–seq Recompute
- [ ] Treehouse Hub
- [x] GDC Hub
- [x] ATAC–seq Hub
- [x] Kids First Hub
- [x] https://previewsinglecell.xenahubs.net
- [x] jupyter notebook

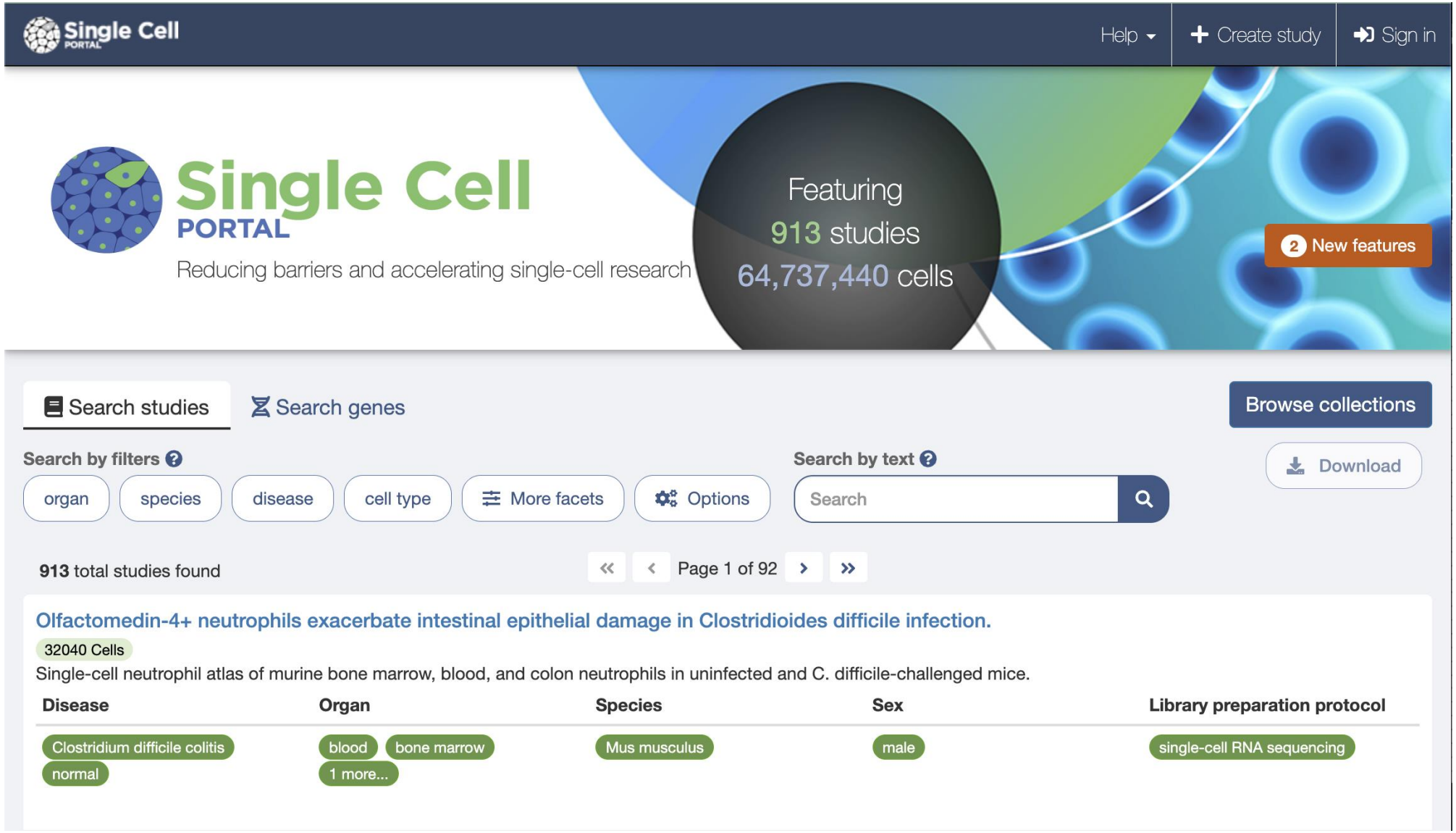
Single Cell Expression Atlas

https://www.ebi.ac.uk/gxa/sc/home



Single Cell Portal

https://singlecell.broadinstitute.org/



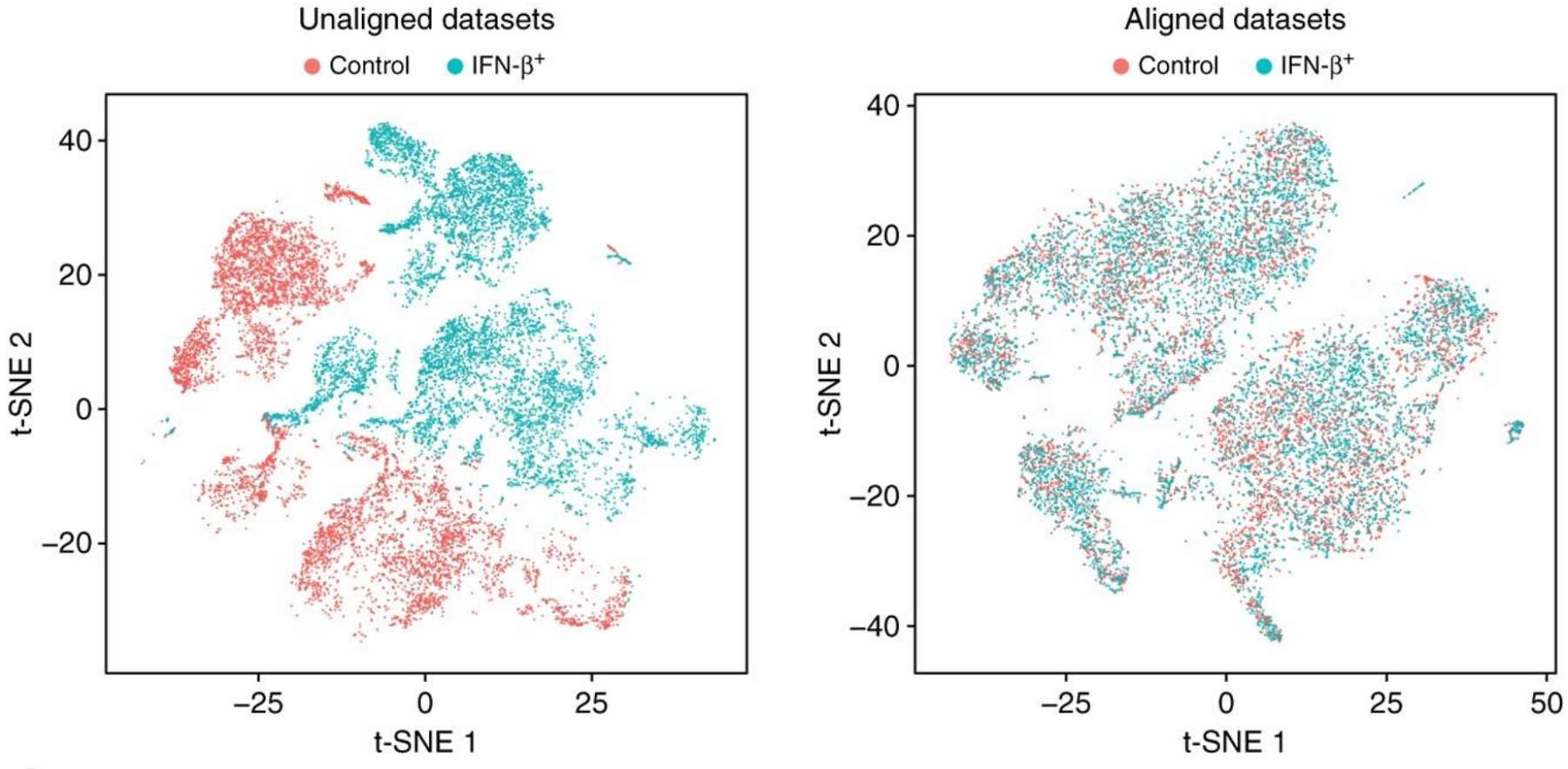
Comparison: Single Cell Expression Atlas vs. Single Cell Portal

	Single Cell Expression Atlas	Single Cell Portal
Host Organization	EMBL-EBI	Broad Institute
Purpose	Centralized resource for exploring standardized, reprocessed single-cell datasets across multiple species	Platform for uploading, sharing, and interactively exploring single-cell studies
Data Source	Public datasets collected and re-analyzed using uniform pipelines	Researcher-submitted studies, often directly from publications
Standardization Level	High — datasets are normalized and annotated for cross-study comparison	Variable — depends on individual study submissions
User Uploads	Not user-uploadable (curated by EMBL-EBI team)	Users can upload and publish their own datasets
Data Download Options	Processed (normalized) data, metadata, and cell type annotations	Raw and/or processed files; downloadable visualizations

What is a batch effect?

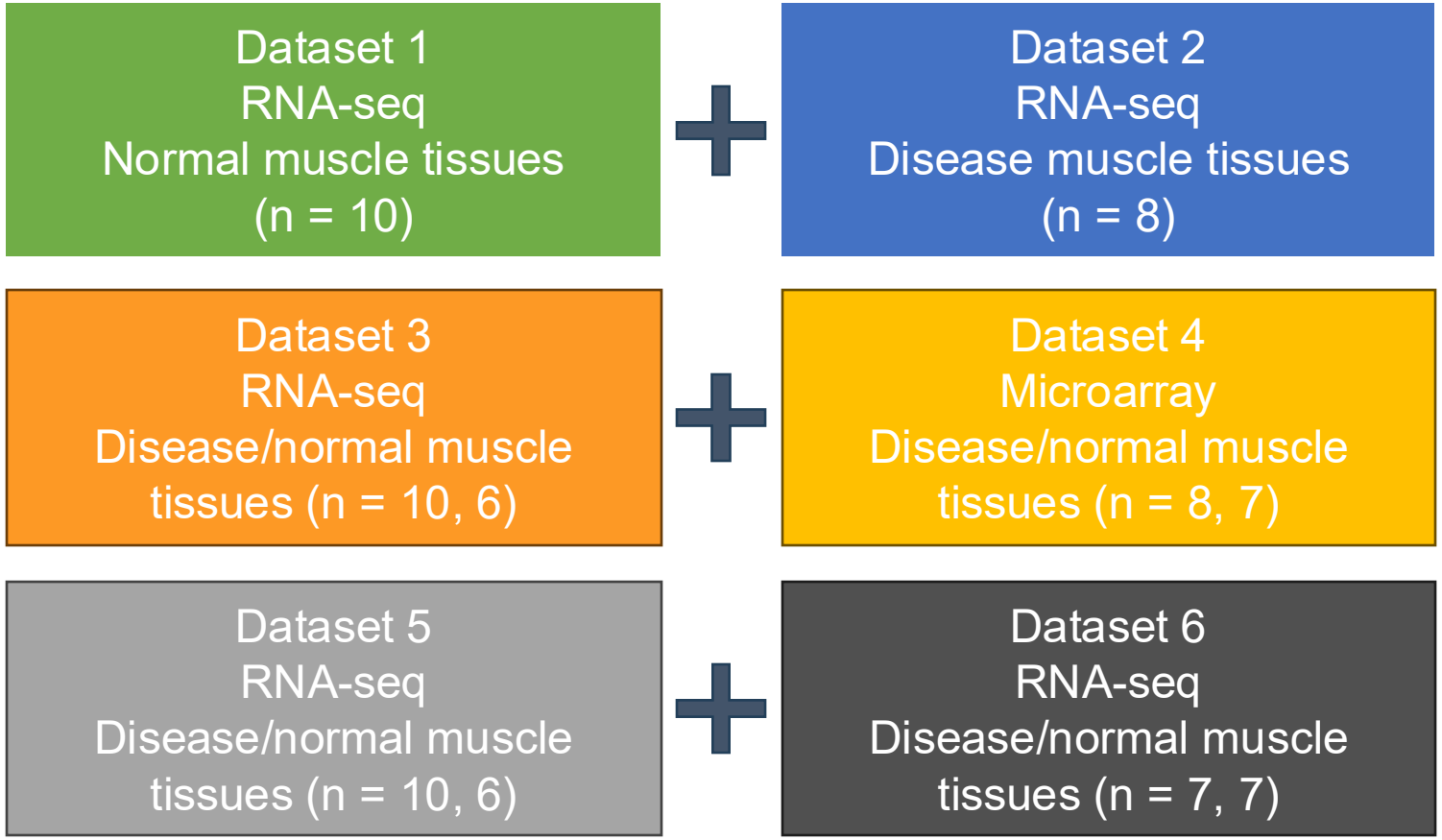
- Non-biological variations introduced by experimental or technical factors.
- Can mask true biological signals or create false-positive differences.
- Common in public datasets integrated from different studies or platforms.

Batch correction is very important for data integration

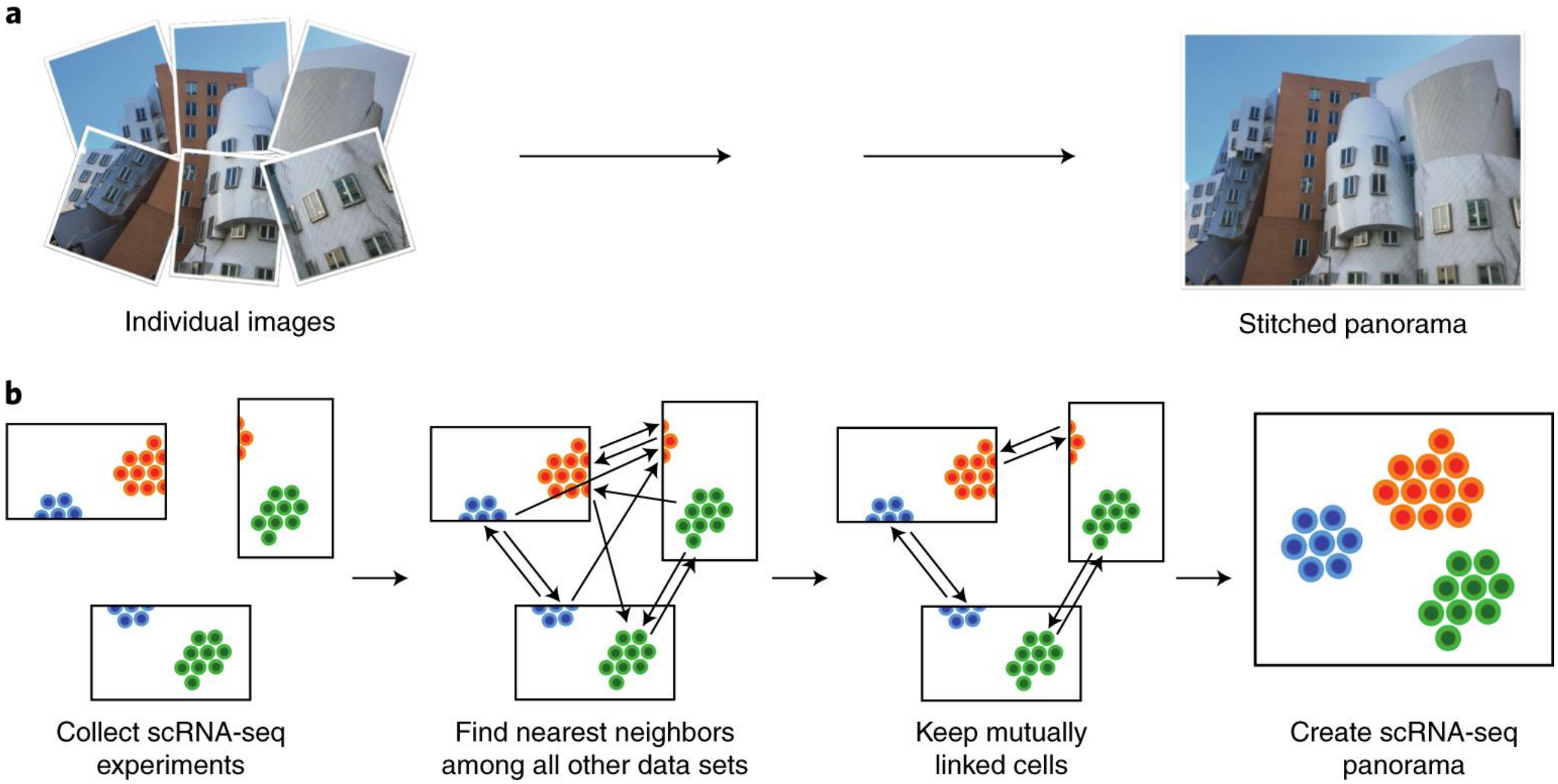


doi: 10.1038/nbt.4096

Scenarios: Is it possible to remove the batch effect?

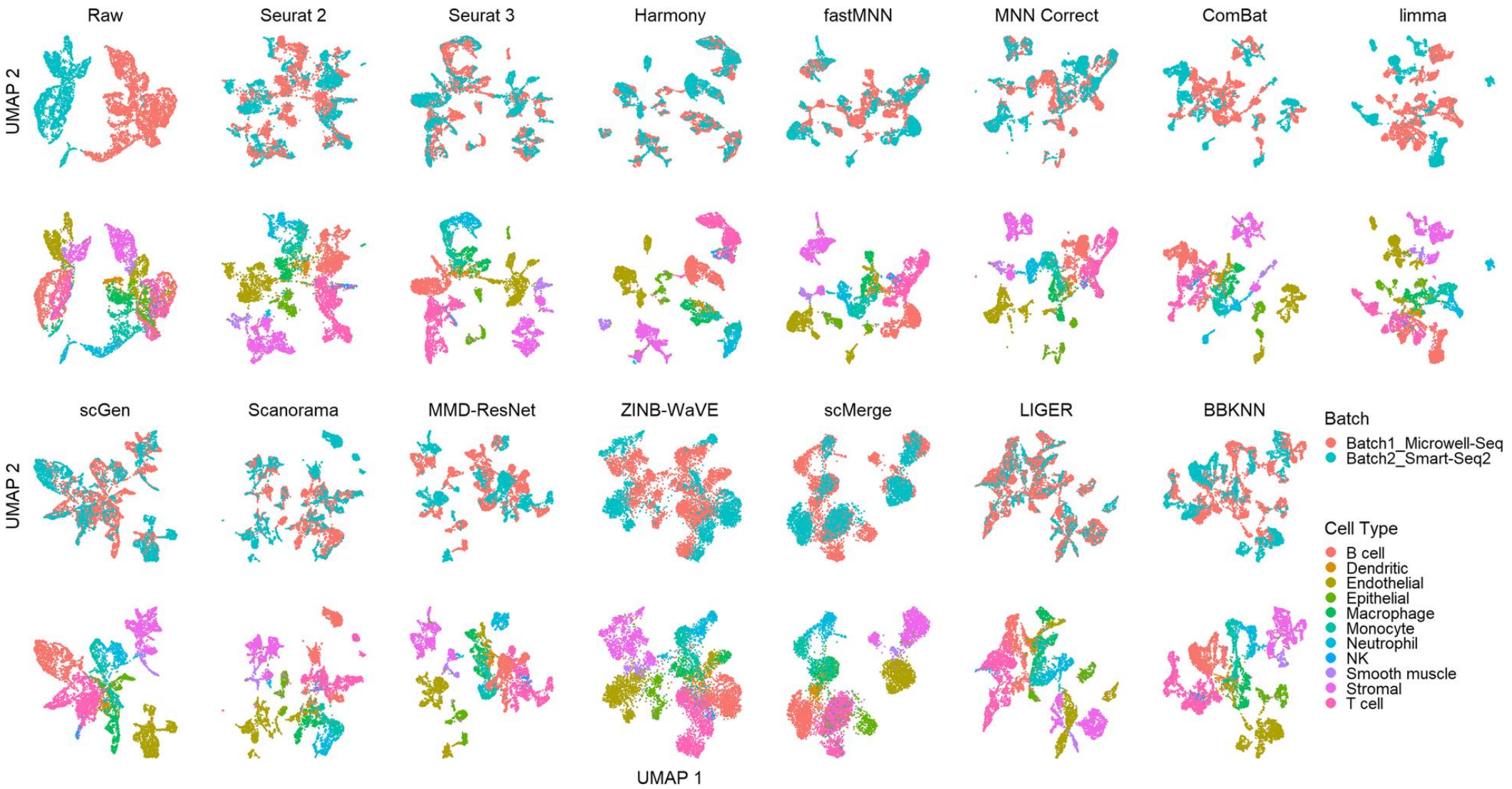


Single-cell data integration and batch effect correction



10.1038/s41587-019-0113-3

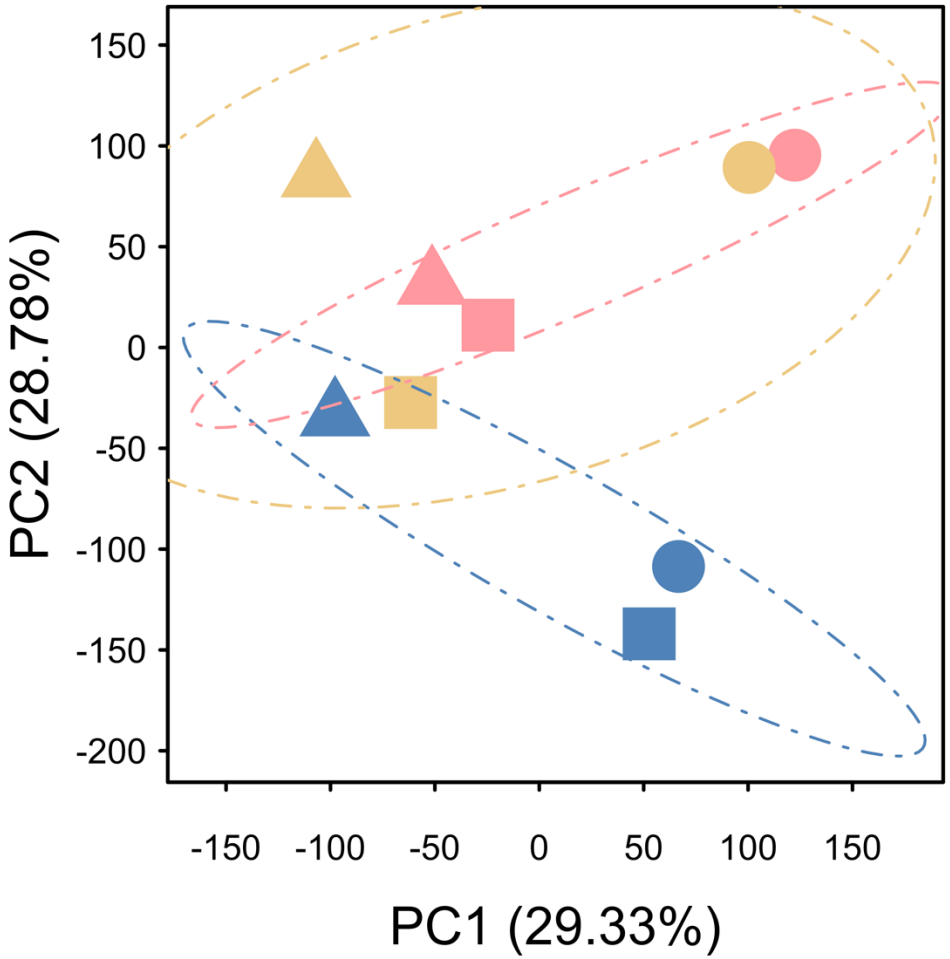
Batch-effect correction methods: which is best?

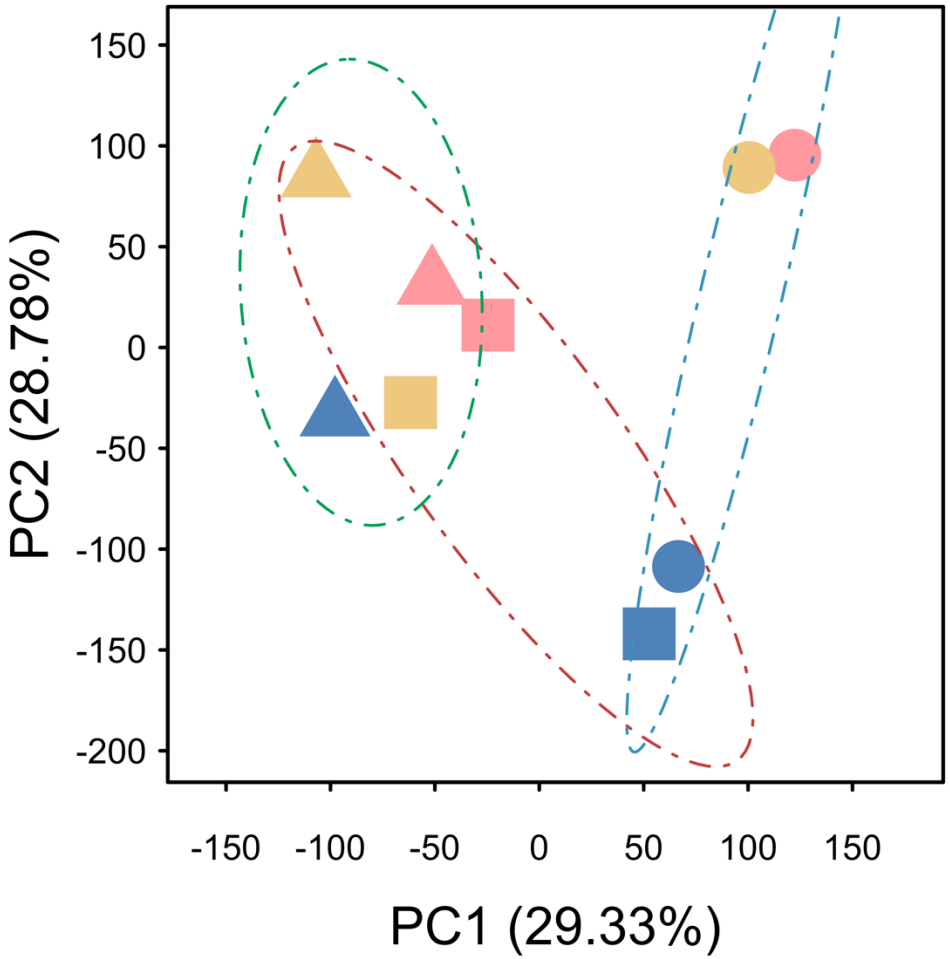


10.1186/s13059-019-1850-9

Before batch-effect removal

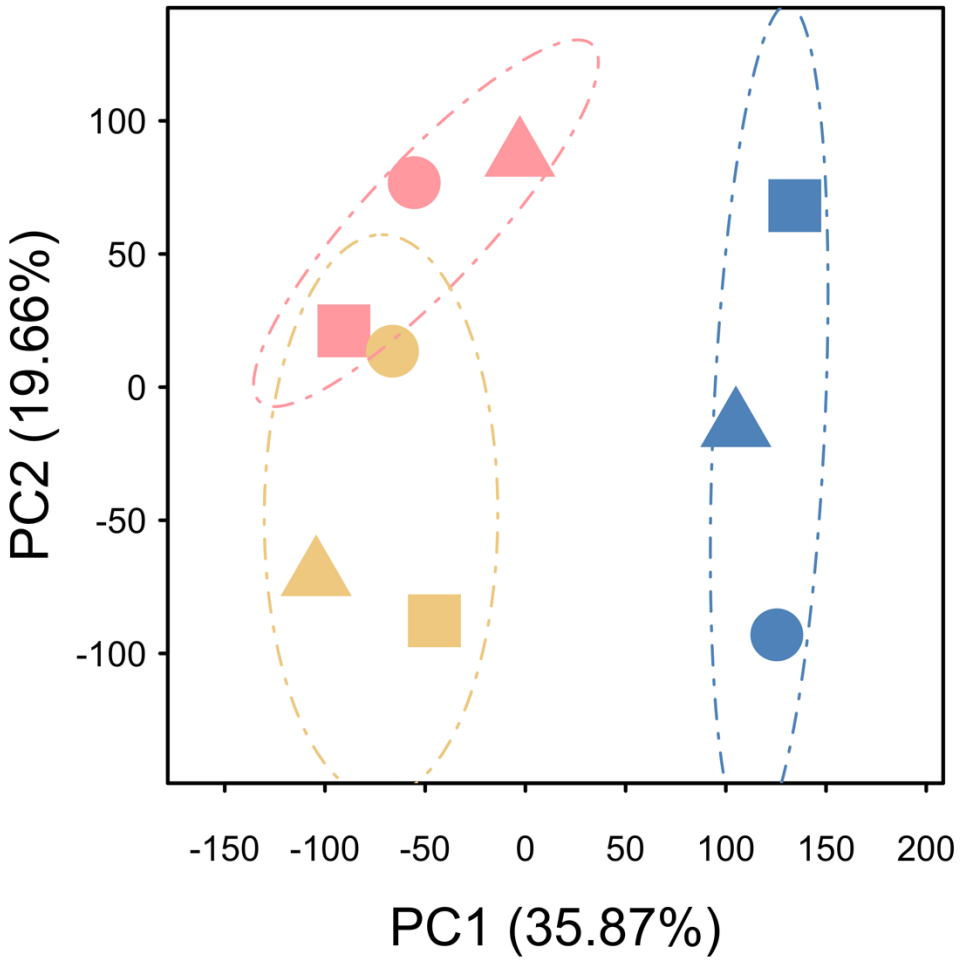
Color indicates biological groups; shape indicates batches

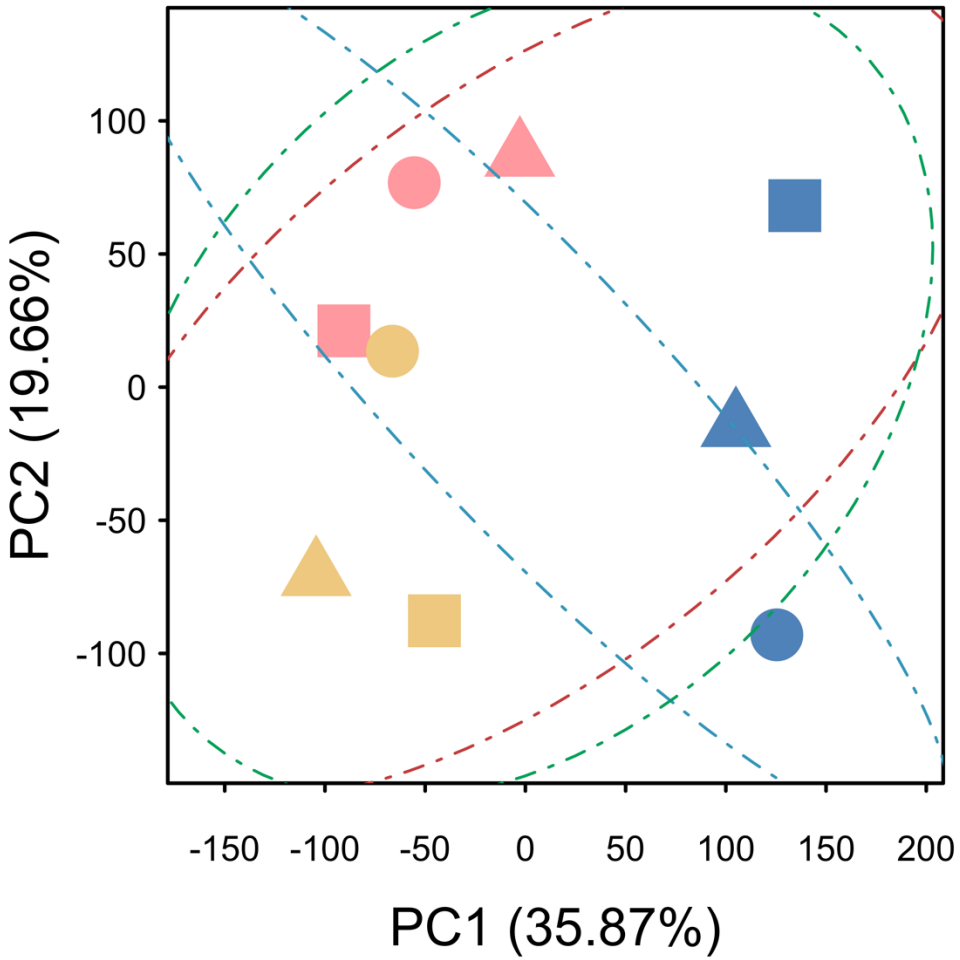




After batch-effect removal

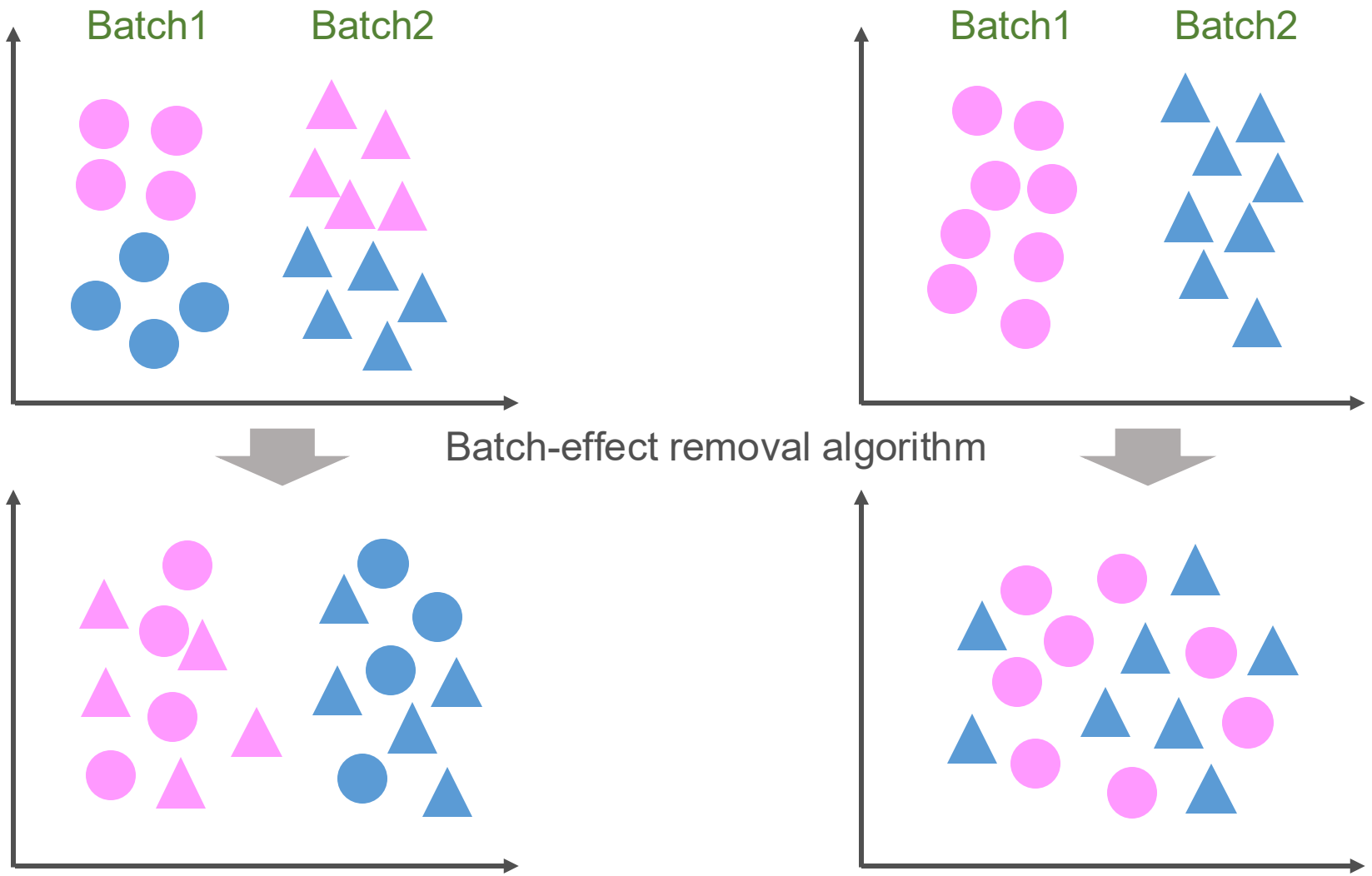
Color indicates biological groups; shape indicates batches





How dose ‘batch effect removal’ work?

Assumption: The distribution of biological groups (e.g., disease vs control) is similar across batches.



Batch effect adjustment for RNA-seq count data: ComBat-seq

https://github.com/zhangyuqing/ComBat-seq

```
# Install from GitHub
install.packages("devtools")
devtools::install_github("zhangyuqing/sva-devel")

library(sva)
adjusted <- ComBat_seq(count_matrix, batch=batch)
```

Batch effect adjustment for normalized expression matrix (RNA-seq/microarray)

```
## Using limma
library(limma)
adjusted <- limma::removeBatchEffect(exprs_matrix, batch=batch)
```

```
## Using ComBat
library(sva)
adjusted <- sva::ComBat(exprs_matrix, batch=batch)
```

An alternative to integration: Meta-Analysis

Meta-analysis strategies:

- Identify common or consensus DE genes
- Combine p-values across studies
- Combine effect sizes across datasets
- Combine gene rankings

R packages:
GeneMeta, *metaRNASeq*, *metafor*, *RobustRankAggreg*