Final Project

Submission due dates

- Workflow – 28/02/2022 or earlier
- Final project – 20/03/2022

<u>We recommend submitting your workflow ASAP to give yourself more time for the main project</u>

General instructions

- The project should be submitted in in pairs.
- The project should be submitted in English (we will allow submissions in Hebrew, though it is not recommended).
- Use Calibri font in size 12 (except for headers).

Project steps

1. Workflow submission (due 28/02/2022)
2. Literature review
3. Bioinformatic analysis
4. Results
5. Discussion

Project goal

The main objective in many bioinformatics projects is to extract insights from raw data by reducing noise to uncover meaningful biological information. In this project, you will analyze gene expression data generated from bulk or single-cell RNA sequencing. In class and tutorials, we practiced analyzing gene expression data using two popular pipelines:

- Differential expression analysis (DESeq2)
- Single cell RNA-seq clustering (Seurat)

The **goal** of the project is to investigate **how genomic data can be useful to better understand diseases**. To do so you will choose a <u>disease, or some aspect of the disease</u>, such as a group of patients that receive some treatment, a complication of the disease, etc. You will then identify all the gene expression datasets that have been deposited in the public domain. You will choose one (or more) datasets and perform various analyses to extract insights from the data.

Details instructions

1. **Workflow submission:** find datasets and phrase a biological question

We encourage you to work on a biological question related to a disease you are personally interested in. Although there is an endless amount of publicly available datasets, it is sometimes challenging to find those that meet your needs. Therefore, we recommend first making a list of all datasets that are related to a specific disease, and only then coming up with a biological question that can be answered by analyzing one or more of those datasets.

For instructions on finding the "right" datasets, click here

Workflow structure

We ask you to submit a workflow so we can ensure your efforts are focused on the right direction. The workflow should include the following points:

- Students' names and IDs.

- Your disease of choice.

Example: Asthma

- What biological question do you want to answer?

Example: We would like to explore how different medications effect the airway smooth muscle cells in asthma.

- How RNA-seq data can help?

Example: We will analyze bulk RNA-seq expression data to run a differential expression analysis to compare samples of asthma patients that consumed different medications. In addition, we use the differentially expressed genes to identify enriched gene sets and pathways.

- A table of the dataset(s) you collected with information on:
    - Dataset accession ID
    - Link for the website where the data can be found
    - Is it a count matrix data?
    - Different groups that can be found in the data
    - A citation for the study that generated this data

Example:

Accession ID	***Location***	***Count matrix***	***Groups***	***Citation***
GSE58434	*https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58434*	*yes*	*12 samples in total, 6 of them are healthy control, and 6 are asthma patients in which 3 have taken Dexamethasone and 3 Albuterol*	*Himes BE, Koziol-White C, Johnson M, Nikolos C et al. Vitamin D Modulates Expression of the Airway Smooth Muscle Transcriptome in Fatal Asthma. PLoS One 2015;10(7):e0134057. PMID: 26207385*

- If you are not sure that this is a count matrix data, download the data, load it into R and make sure that you have a genes x samples/cells matrix.

Please send the workflow to Almog (almog.angel@campus.technion.ac.il) as a PDF file with the students' IDs as the file name (ID1_ID2.pdf). The name of the e-mail should be "Workflow submission ID1 ID2".

2. Literature review (20 points)

You should begin with this step after we approved your workflow

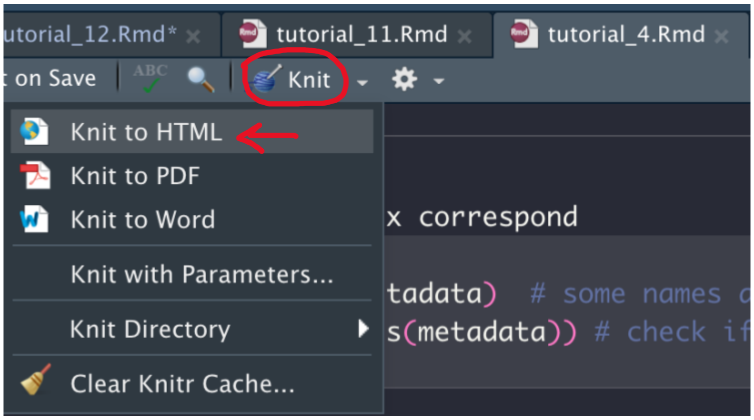
- Write a literature review (up to 3 pages)
- Identify **four** relevant studies related to the disease and biological question you chose (with a major focus on studies that are based on analysis of gene expression data)
- The review should be written as an "upside-down pyramid":
  a. Begin with writing a general overview of the disease (~1 page)
  b. Write about what is still unknown, what main challenges do we face in the context of this disease? (~1 page)
  c. How analyzing genomic data can help to overcome those challenges? (Give examples based on results of the studies you collected) (~0.5 pages)
  d. Finish by writing about the biological question you are about to address. What are you looking for? How is it going to help address those challenges? (~0.5 pages)
- **Use Mendeley (or any other reference manager software) to cite (Cell format) every statement or result that relies on prior knowledge.**

Here is a useful guide of to use Mendeley inside Word.

3. Bioinformatic analysis (30 points)

- Use the dataset(s) you chose in step 1 and perform all relevant analyses.
- The analysis should follow the entire pipeline of RNA-seq and/or single-cell RNA-seq analysis as we practiced in the tutorials.
- The analysis should be comprehensive, ideally covering as many methods we learned as possible:

  - Differential expression analysis
  - Unsupervised clustering
  - Dimensionality reduction (PCA/tSNE/UMAP),
  - Classification and prediction (inear/logistic regression, Random Forest, etc.)
  - Gene set enrichment analyses (and pathway analysis in general)
  - Cell type composition (xCell)
  - Cell type annotation (SingleR)
  - Survival analysis (if relevant)
  - And more
- You may use any publicly available additional datasets to complement your analysis, such as GWAS catalog, DepMap and TCGA.
- Make sure that you create beautiful plots during your analysis, save them, you will use them in the next step.
- Keep your code clean as possible, use comments with useful information about each step in your analysis.
- Document your analysis in R markdown and when you finish save it as an HTML file. To do so you need to click on "Knit to HTML".



```
x correspond
tadata)  # some names c
s(metadata)) # check if
```

We will grade this step based on:

a. How comprehensive your analysis (use as many methods as possible).
b. The quality of your analysis (make sure you do not forget steps and use the right functions properly).
c. How well you documented the analysis.
d. The plots (figures) you generate.

We will **not** grade this step based on your findings. Namely, we do not expect you to successfully generate novel biological findings.

4. Results (35 points)

- Summarize your results in 5-8 pages (depends on figures size).

- This section should be written in a way that we can completely understand why and how you ended up with some meaningful results from the data you downloaded.

- The results section tells us the story of how you answered the question raised in the introduction. You do not need to present here all the analyses and figures you made in step 3, only those that lead to the conclusions.

- The results section should first start with a short description of the goal or hypothesis leading the analysis.

For example:
Dexamethasone or Albuterol are currently the two most used medication to treat asthma. However, patients do not always respond similarly to those medications, namely, some asthma patients show weak respond to Albuterol compared to Dexamethasone and vice versa (here should be a citation). We hypothesis that genetic variation in certain genes (i.e., SNPs) among different patients may be the reason to why patients respond differently. However, it is still unknown which genes are mostly affected by those medication…

- Then, describe the data you chose for the analysis.

For example:
We used RNA-seq gene expression data from GSE58434 (here should be a citation) to compare samples of airway smooth muscle cells from patients that consumed Dexamethasone with 4 samples of…

- Next, for **each** analysis you wish to report use the following structure:

- Explain **why** you did this analysis?

For example:
We aim to identify genes that are differentially expressed in patients under Dexamethasone to…

- **What** did you do in the analysis?

For example:
We preformed differential expression analysis using DESeq2 (here should be a citation) by using the raw count data and corresponding metadata regarding the medication and the sex of each sample…

- **What** did you get out of the analysis? **What** does it mean?

For example:
We found 10 genes that are significantly (FDR adjusted p-value < 0.05) highly expressed in samples treated with Dexamethasone (Figure 1)…

… Among them the genes ABC DEF and GHI are all related to the BlaBla pathway which play key role in cell inflammation (citation) …

- Explain briefly what can be learned from these results. This sentence should lead to the next result.
-

For example:
The BlaBla pathway might be activated in response to Dexamethasone. Therefore, we were interested in preforming gene-set enrichment analysis to identify more pathways…

- Each result should be accompanied by a figure.
- When you report a figure give it a number and write the title below.
- The title should be a short (1-2 lines) description of what we see in the figure.

For example:

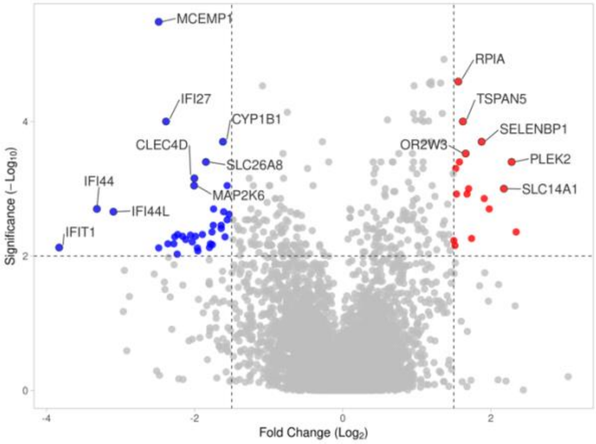


Figure 1: Volcano plot displaying genes that were most differentially expressed in Dexamethasone (red) relative to Albuterol (blue) in of airway smooth muscle cells of asthma patients.

5. Discussion (15 points)

Here you should discuss your results from step 4 in 1-2 pages.

Try to answer the following questions:

- What conclusions have you drawn from the analysis? Do they provide any insight into the biological question?
- Mention the limitations of your analysis.
- What would you do next? Are there any ways to overcome those limitations? What future experiment can you suggest answering your biological question that will address what is still unknown?

The structure of the project

- **Title**

Provide a title that gives a glimpse of what you did and what you found. Also include your names, affiliations, and IDs.

- **Abstract**

One paragraph of up to 250 words.
Should include a sentence or two for each of the following:

a. Brief introduction on the disease.
b. The knowledge gaps.
c. The main goal of the analysis.
d. Brief overview of your analysis.
e. Your key result(s).
f. Main conclusion(s).

- **Introduction**

This is the literature review you performed in step 2.
Remember to provide references to the relevant manuscripts!

- **Results**

Description of your analyses and figures from step 4.
Pay attention to put a number and a title under each figure.
Refer to the figure numbers when you describe the result.

- **Discussion**

Your discussion from step 5.

- **References (Cell format)**

Add references to your document using a citation manager. You can use Mendeley (Technion has a subscription) or Zotero, but you are welcome to use any citation manager.

Instructions for submission

Submit your project as a ZIP file that contain:

- Workflow signed and approved by Almog/Dvir
- PDF file of your work
- HTML R markdown file of your analysis

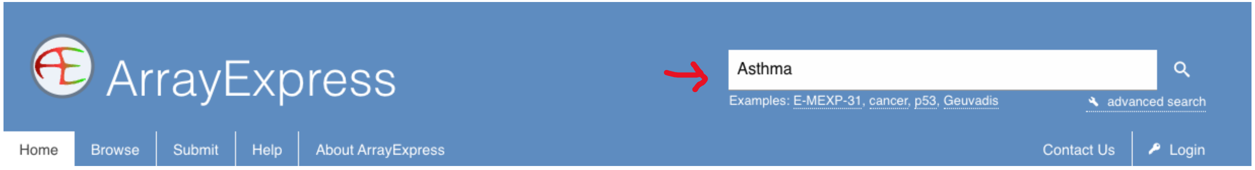
Good luck!

How to find the “right” dataset

Once you have a disease in mind (i.e., asthma) you should start by looking for data in the ArrayExpress repository.

* Notice: ArrayExpress is a good place to start looking for count matrix of **bulk** RNA-seq. If you are interested in **single cell RNA-seq** skip to the next repository (GEO).

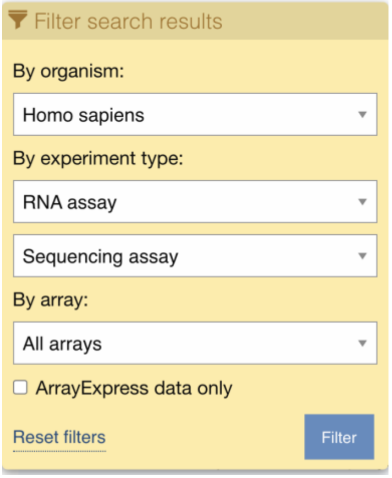
- Search for the disease on the top right:



Asthma

Examples: E-MEXP-31, cancer, p53, Geuvadis

- Filter the results by choosing an organism (we recommend sticking with humans or mouse), experiment type should be RNA assay and Sequencing assay.



Filter search results

By organism: Homo sapiens

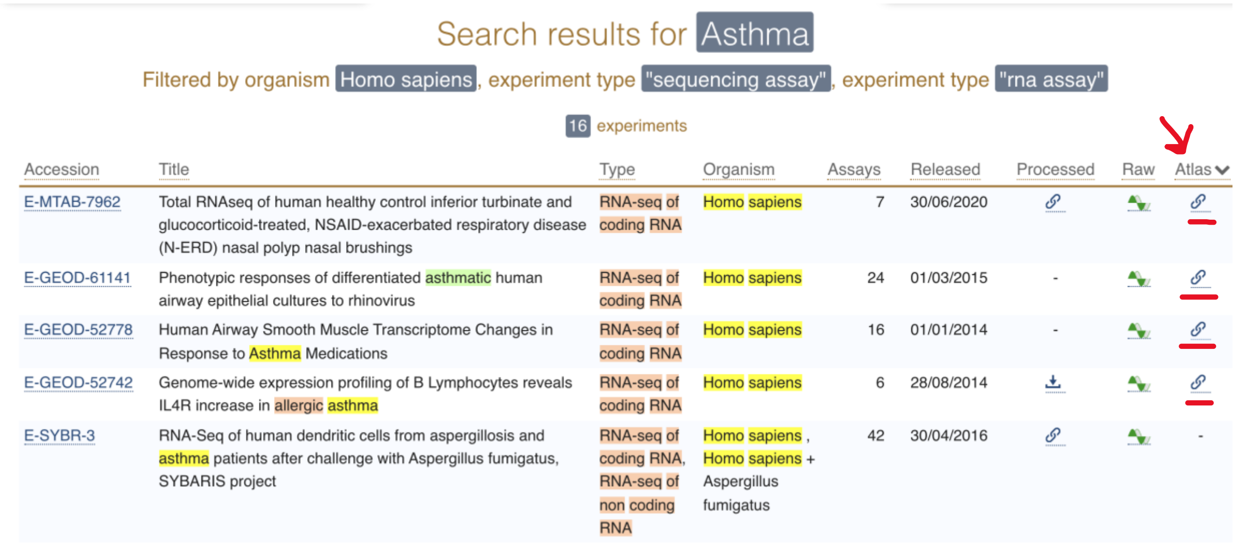
By experiment type: RNA assay

Sequencing assay

By array: All arrays

□ ArrayExpress data only

- Sort datasets by “Atlas” (those are studies with ready to download count matrix):



Search results for Asthma

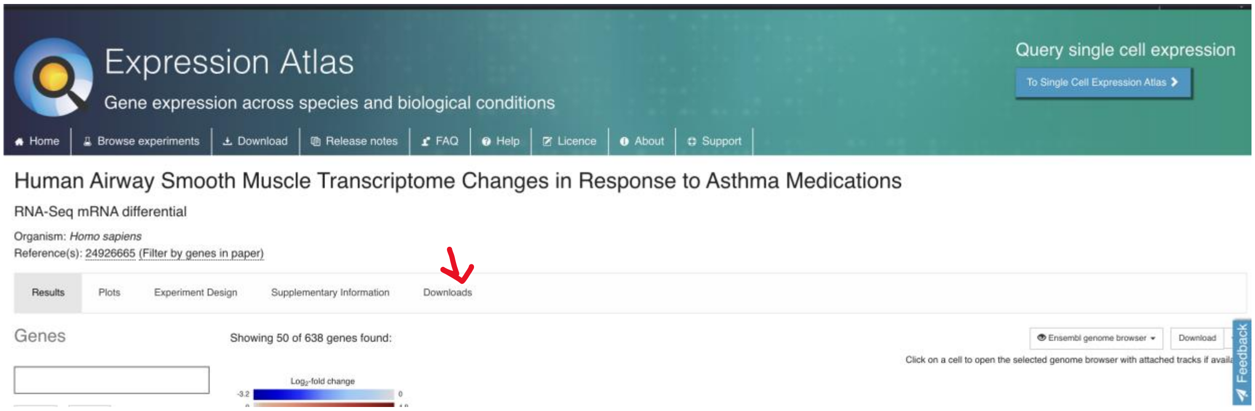
Filtered by organism Homo sapiens, experiment type "sequencing assay", experiment type "rna assay"

16 experiments

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-MTAB-7962	Total RNAseq of human healthy control inferior turbinate and glucocorticoid-treated, NSAID-exacerbated respiratory disease (N-ERD) nasal polyp nasal brushings	RNA-seq of coding RNA	Homo sapiens	7	30/06/2020	🔗	✓	🔗
E-GEOD-61141	Phenotypic responses of differentiated asthmatic human airway epithelial cultures to rhinovirus	RNA-seq of coding RNA	Homo sapiens	24	01/03/2015	-	✓	🔗
E-GEOD-52778	Human Airway Smooth Muscle Transcriptome Changes in Response to Asthma Medications	RNA-seq of coding RNA	Homo sapiens	16	01/01/2014	-	✓	🔗
E-GEOD-52742	Genome-wide expression profiling of B Lymphocytes reveals IL4R increase in allergic asthma	RNA-seq of coding RNA	Homo sapiens	6	28/08/2014	⤓	✓	🔗
E-SYBR-3	RNA-Seq of human dendritic cells from aspergillosis and asthma patients after challenge with Aspergillus fumigatus, SYBARIS project	RNA-seq of coding RNA, RNA-seq of non coding RNA	Homo sapiens , Homo sapiens + Aspergillus fumigatus	42	30/04/2016	🔗	✓	-

- Make a list for the datasets (in this example we have 4), click on the accession ID under the “Accession” column for more information about this data (write notes somewhere).

- If you wish to download the data, click on the icon below the “Atlas” column. Then, go to “Downloads” on the new window.



Human Airway Smooth Muscle Transcriptome Changes in Response to Asthma Medications

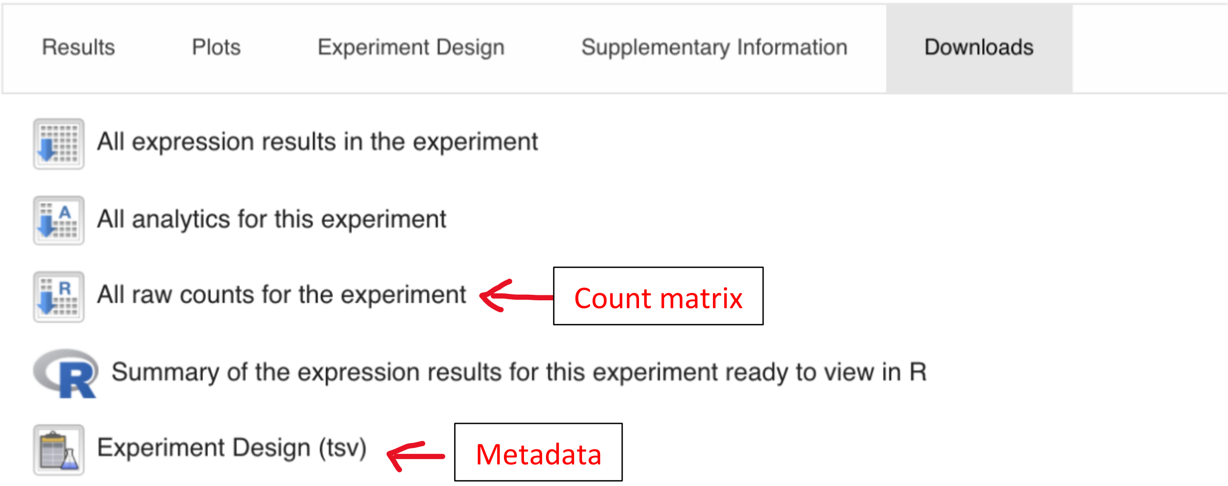
RNA-Seq mRNA differential

Organism: *Homo sapiens*
Reference(s): 24926665 (Filter by genes in paper)

Showing 50 of 638 genes found:

Click on a cell to open the selected genome browser with attached tracks if availa

- Finally, you can download the count matrix and metadata


Results
Plots
Experiment Design
Supplementary Information
Downloads
All expression results in the experiment
All analytics for this experiment
All raw counts for the experiment
Summary of the expression results for this experiment ready to view in R
Experiment Design (tsv)

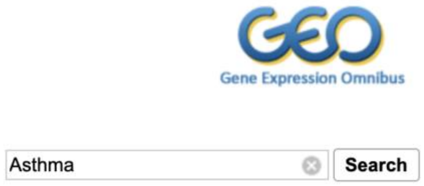
Count matrix

Metadata

Another repository that you should try is GEO (here you might also find single cell RNA-seq data).

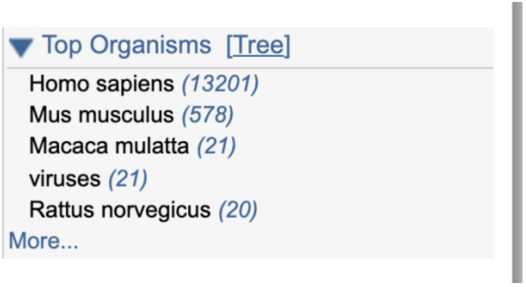
* Notice: in GEO you need to make sure that you are downloading the count matrix and not the raw data:

1. Type the disease name in the search bar and click search **(if you are looking for single cell data type "<your disease name> + single cell", i.e., "asthma single cell")**.



Asthma

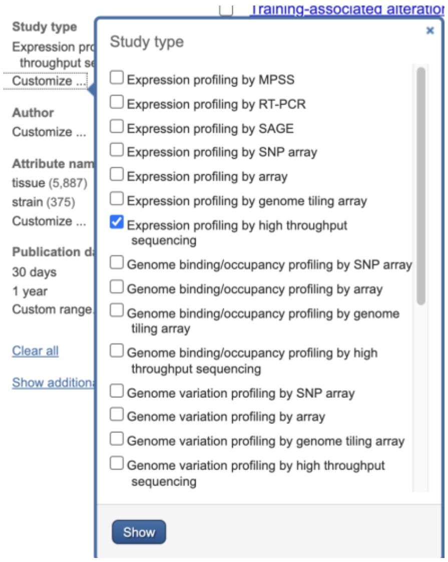
2. Filter by an organism from the top right



Top Organisms [Tree]

Homo sapiens (13201)
Mus musculus (578)
Macaca mulatta (21)
viruses (21)
Rattus norvegicus (20)
More...

3. In the right click "Customize" under "Study type" and remove everything expect "Expression profiling by high throughput sequencing" then click show.



Study type
Expression pro
throughput se
Customize ...

Author
Customize ...

Attribute nam
tissue (5,887)
strain (375)
Customize ...

Publication da
30 days
1 year
Custom range

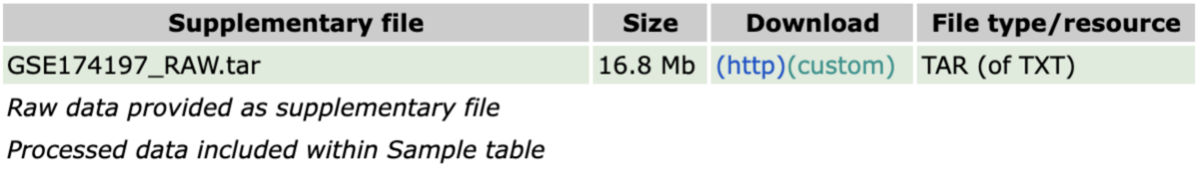
Clear all

Show addition

Study type

- □ Expression profiling by MPSS
- □ Expression profiling by RT-PCR
- □ Expression profiling by SAGE
- □ Expression profiling by SNP array
- □ Expression profiling by array
- □ Expression profiling by genome tiling array
- ☑ Expression profiling by high throughput sequencing
- □ Genome binding/occupancy profiling by SNP array
- □ Genome binding/occupancy profiling by array
- □ Genome binding/occupancy profiling by genome tiling array
- □ Genome binding/occupancy profiling by high throughput sequencing
- □ Genome variation profiling by SNP array
- □ Genome variation profiling by array
- □ Genome variation profiling by genome tiling array
- □ Genome variation profiling by high throughput sequencing

4. Go over each study by clicking on the study title to learn more about it (make notes).
5. Avoid studies that only have raw data available – for example:

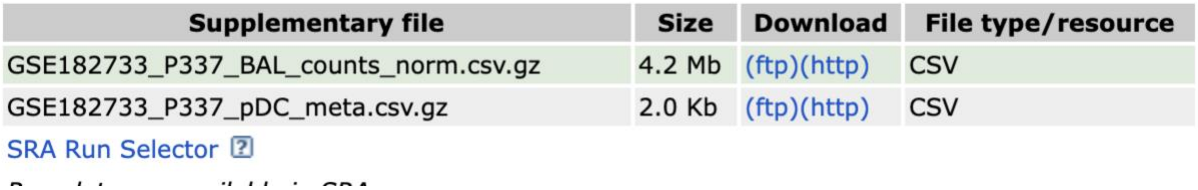


Supplementary file	Size	Download	File type/resource
GSE174197_RAW.tar	16.8 Mb	(http)(custom)	TAR (of TXT)

Raw data provided as supplementary file

Processed data included within Sample table

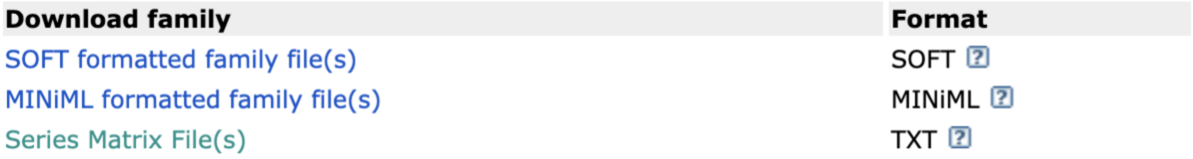
6. You should look for the count matrix (sometimes it might be already normalized and sometimes not) – for example:



Supplementary file	Size	Download	File type/resource
GSE182733_P337_BAL_counts_norm.csv.gz	4.2 Mb	(ftp)(http)	CSV
GSE182733_P337_pDC_meta.csv.gz	2.0 Kb	(ftp)(http)	CSV

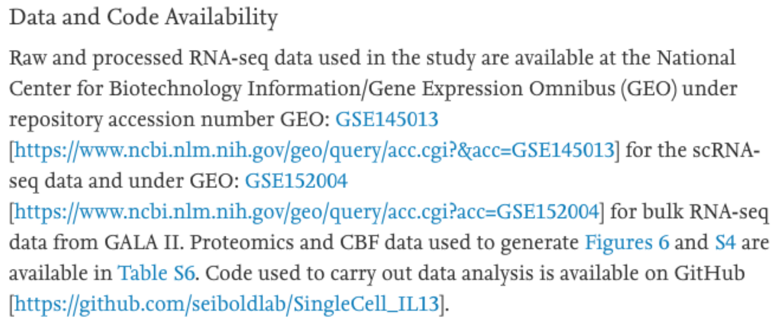
SRA Run Selector ?

7. You can find the metadata under “Series Matrix File(s)”:



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

ArrayExpress and GEO are the two main repositories, however, you can also find data in multiple other resources such as: recount2 (bulk + single cell), 10X genomics datasets (single cell), this curated database of single-cell studies, Hemberg lab’s collection (single cell), scRNASeqDB (single cell) and of course you can always Google: “<your disease name> single cell/bulk”. Even better, you can search in Google Scholar and look in different studies for their data (most studies write an accession ID of GEO or ArrayExpress), for example:



Data and Code Availability

Raw and processed RNA-seq data used in the study are available at the National Center for Biotechnology Information/Gene Expression Omnibus (GEO) under repository accession number GEO: GSE145013 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE145013] for the scRNA-seq data and under GEO: GSE152004 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152004] for bulk RNA-seq data from GALA II. Proteomics and CBF data used to generate Figures 6 and S4 are available in Table S6. Code used to carry out data analysis is available on GitHub [https://github.com/seiboldlab/SingleCell_IL13].