GTEx

GEO

Gene Expression Omnibus

Metabolomics Workbench

TCGA

ProteomeXchange

15/01/2023

TP 1: OMICS databases

Table of contents

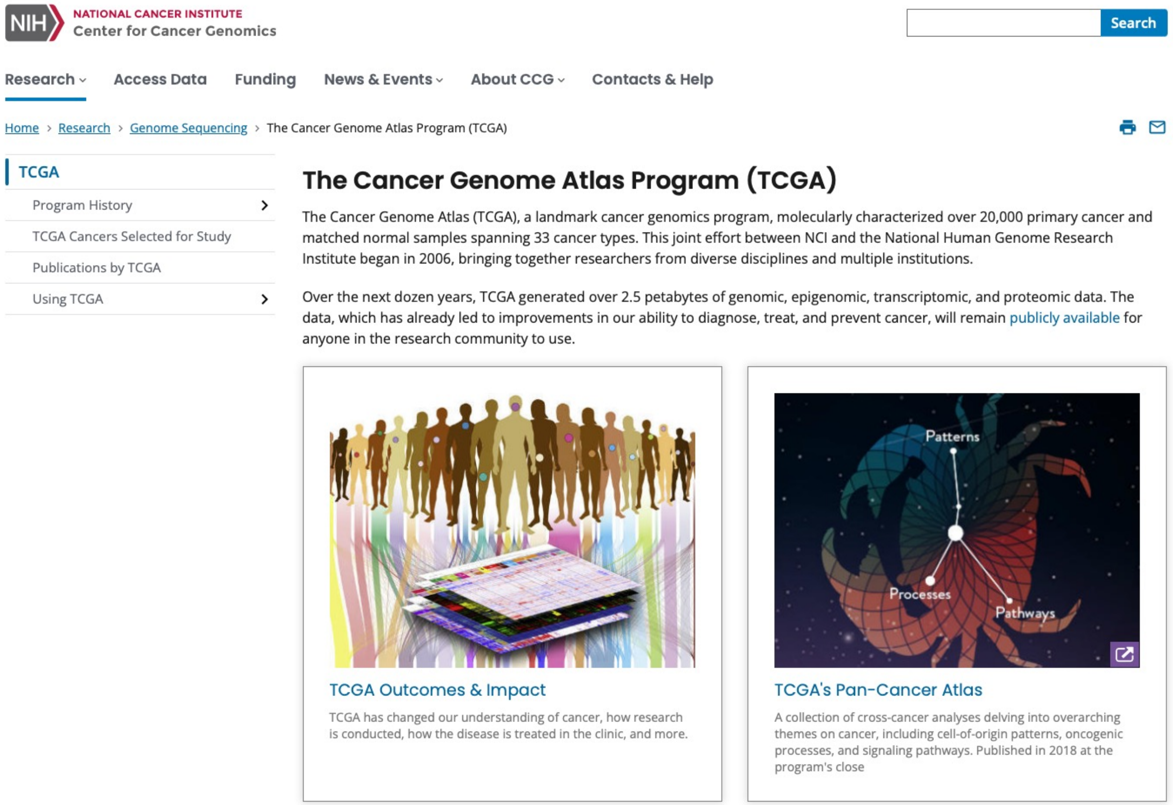
1. TCGA database
2. GTEX database
3. Finding data in GEO
4. Protein data in Proteome Exchange
5. Metabolomic data in Metabolomics workbench

1. TCGA database

TCGA database

🌐 https://www.cancer.gov/ccg/research/genome-sequencing/tcga

The Cancer Genome Atlas (TCGA) molecularly characterized over **20,000** primary **cancer** and matched **normal samples** spanning **33 cancer types**. Genomic, epigenomic, transcriptomic, and proteomic data are available.



NIH NATIONAL CANCER INSTITUTE Center for Cancer Genomics

The Cancer Genome Atlas Program (TCGA)

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already led to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use.

TCGA Outcomes & Impact

TCGA has changed our understanding of cancer, how research is conducted, how the disease is treated in the clinic, and more.

TCGA's Pan-Cancer Atlas

A collection of cross-cancer analyses delving into overarching themes on cancer, including cell-of-origin patterns, oncogenic processes, and signaling pathways. Published in 2018 at the program's close

TCGA database

- **Step 1**: Select projects in TCGA.
- **Step 2:** Select projects about prostate gland.
- **Step 3:** Select the TCGA project on prostate gland which studies 3 types of disease (hint : click on the project ID).
- **Step 4**: You should have the project summary. How many patients are involved in this project and what types of data are available? Genomic, transcriptomic, etc ... ?
- **Step 5**: Return to the table of all prostate projects and click on the number of cases in the TCGA-PRAD project.

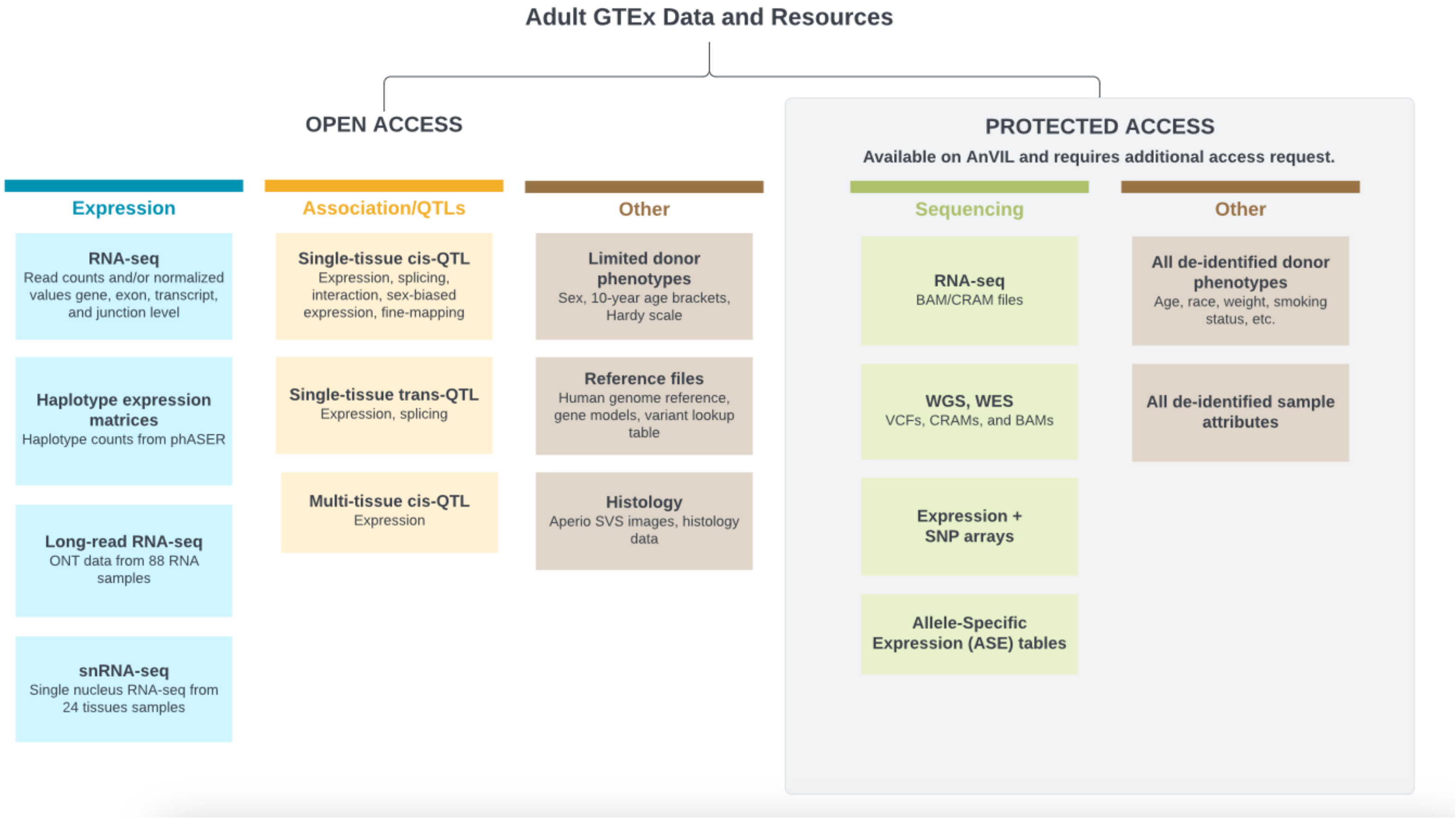
You should obtain the list of all patients included in the project.

- **Step 6:** Select only open-access data.
- **Step 7**: Now, we will select the omics data for this cohort. First select CNV (genomic data) but select only gene level copy number and ASCAT2 workflow and add the data to the Cart.
- **Step 8**: Select gene expression data (transcriptomic data) and add them to the Cart.
- **Step 9**: Select DNA methylation values (epigenetic data) and add them to the Cart.
- **Step 10**: Select proteome profiling data (proteomic data) and add them to the Cart.
- **Step 11**: Go the the Cart. How many files do you have to download? What is the total size of all files?

2. GTEX database

GTEX database

The Adult Genotype-Tissue Expression (GTEx) project is a comprehensive **public resource** for the study of **tissue-specific gene expression** and regulation. Samples were collected from **54 non-diseased tissue sites across nearly 1000 individuals**, primarily for molecular assays including WGS, WES, and RNA-Seq.



Adult GTEx Data and Resources

OPEN ACCESS

Expression

- **RNA-seq** Read counts and/or normalized values gene, exon, transcript, and junction level
- **Haplotype expression matrices** Haplotype counts from phASER
- **Long-read RNA-seq** ONT data from 88 RNA samples
- **snRNA-seq** Single nucleus RNA-seq from 24 tissues samples

Association/QTLs

- **Single-tissue cis-QTL** Expression, splicing, interaction, sex-biased expression, fine-mapping
- **Single-tissue trans-QTL** Expression, splicing
- **Multi-tissue cis-QTL** Expression

Other

- **Limited donor phenotypes** Sex, 10-year age brackets, Hardy scale
- **Reference files** Human genome reference, gene models, variant lookup table
- **Histology** Aperio SVS images, histology data

PROTECTED ACCESS

Available on AnVIL and requires additional access request.

Sequencing

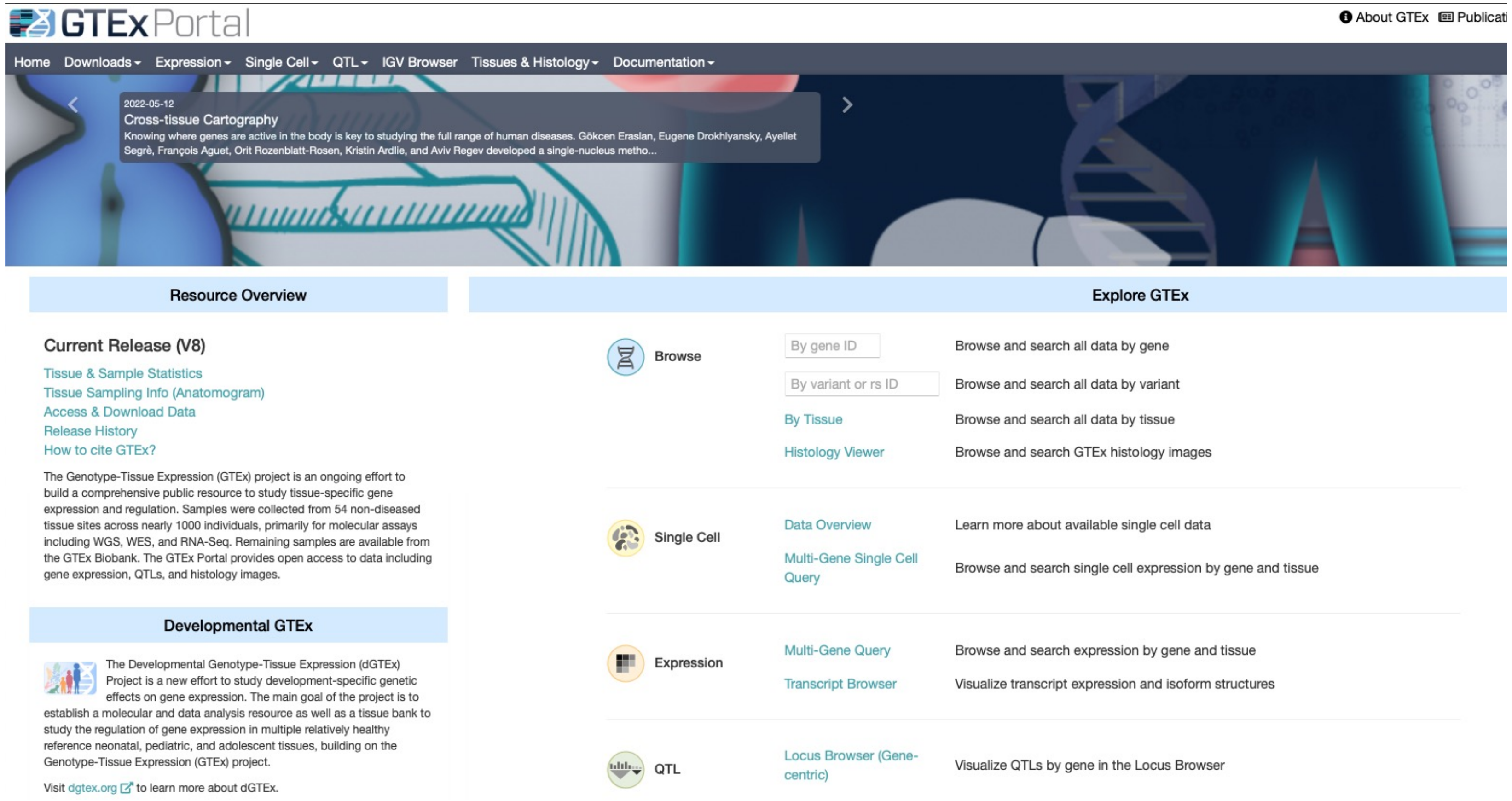
- **RNA-seq** BAM/CRAM files
- **WGS, WES** VCFs, CRAMs, and BAMs
- **Expression + SNP arrays**
- **Allele-Specific Expression (ASE) tables**

Other

- **All de-identified donor phenotypes** Age, race, weight, smoking status, etc.
- **All de-identified sample attributes**

GTEX database

https://www.gtexportal.org/home/



2022-05-12

Cross-tissue Cartography

Knowing where genes are active in the body is key to studying the full range of human diseases. Gökcen Eraslan, Eugene Drokhlyansky, Ayellet Segrè, François Aguet, Orit Rozenblatt-Rosen, Kristin Ardlie, and Aviv Regev developed a single-nucleus metho...

Resource Overview

Current Release (V8)

- Tissue & Sample Statistics
- Tissue Sampling Info (Anatomogram)
- Access & Download Data
- Release History
- How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

Developmental GTEx

The Developmental Genotype-Tissue Expression (dGTEx) Project is a new effort to study development-specific genetic effects on gene expression. The main goal of the project is to establish a molecular and data analysis resource as well as a tissue bank to study the regulation of gene expression in multiple relatively healthy reference neonatal, pediatric, and adolescent tissues, building on the Genotype-Tissue Expression (GTEx) project.

Visit dgtex.org to learn more about dGTEx.

Explore GTEx

Browse	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
	By Tissue	Browse and search all data by tissue
	Histology Viewer	Browse and search GTEx histology images
Single Cell	Data Overview	Learn more about available single cell data
	Multi-Gene Single Cell Query	Browse and search single cell expression by gene and tissue
Expression	Multi-Gene Query	Browse and search expression by gene and tissue
	Transcript Browser	Visualize transcript expression and isoform structures
QTL	Locus Browser (Gene-centric)	Visualize QTLs by gene in the Locus Browser

GTEX database

- **Step 1**: Answer the following questions

  1. What is the current version of GTEX database ? -> V8
  2. What is the total number of tissues available in GTEX

  and the number of donors? -> 54 / 948

  3. Which tissue contains the largest number of samples ? -> Muscle - Skeletal
  4. What information do we have about donors?

- **Step 2**: Download expression data from lung tissue.

3. Finding data in GEO

GEO database

🌐 https://www.ncbi.nlm.nih.gov/geo/

- The Gene Expression Omnibus (GEO) is a public repository which contains **gene expression data** from various platforms (microarray, next-generation sequencing, etc…) and experimental conditions.

- You can find data directly in GEO. But in most cases the route to getting to the data is finding a paper which describes an interesting piece of biology you want to pursue. We can search for interesting papers in the pubmed database (https://pubmed.ncbi.nlm.nih.gov/)

- We are going to look for information relating to the Prox1 gene, specifically we'd like to find out what effect knocking this gene out in embryonic tissue has.

GEO database

- **Step 1**: In pubmed, search for “prox1 embryonic knockout transcriptome” and find a paper which is obviously based around RNA-Seq data and which includes all of these terms.
- **Step 2**: Follow the link to the paper and see if you can get the full text. See if they give a GEO accession (GSEXXXX) for the data they use. If they do is it data they created, or public data they re-analysed? ->GSE69940
- **Step 3**: Search for the accession code you find in GEO (https://www.ncbi.nlm.nih.gov/geo/) and find the details for the dataset. Is the paper you found the original paper for this dataset? If not which paper first published it? ->No.

GEO database

- **Step 4**: Answer the questions

  - How many samples are included in this dataset?   -> 6
  - What experimental conditions do they represent? -> Prox1, C57
  - How many replicates of each condition are there? ->3

- **Step 5:** From the GEO entry find the SRA database accession for this data. Take this and search for it in sra explorer:

https://sra-explorer.info. -> SRP059586

Can you see all of the runs you saw in the SRA run selector? 12

Add the runs to your basket and then generate a list of download URLs where you could get the data if you wanted to download it all.

- **Step 6:** Put the GSE accession number into the text search (not the accession search) of https://www.ebi.ac.uk/ena/.

Find the relevant study page and check that you can see the samples. See that you could click on the links to the individual fastq files to download them (but don't actually download them).

->

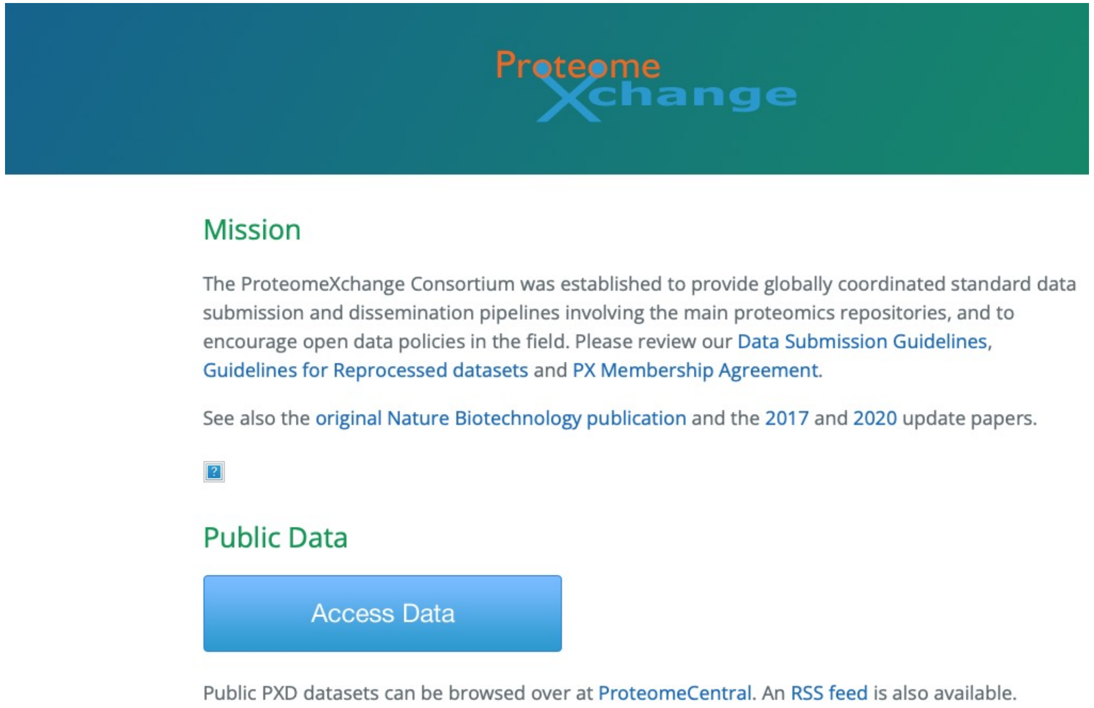
< Genome-wide analysis of embryonic gene epression in the absence of Prox1 compared to wild type>

4. Protein Data in ProteomeExchange

ProteomeExchange database

🌐 http://www.proteomexchange.org/

ProteomeXchange is a consortium for sharing mass spectrometry-based **proteomics data**. It provides a centralized platform to submit, access, and disseminate proteomics datasets generated by different laboratories and research groups.



Mission

The ProteomeXchange Consortium was established to provide globally coordinated standard data submission and dissemination pipelines involving the main proteomics repositories, and to encourage open data policies in the field. Please review our Data Submission Guidelines, Guidelines for Reprocessed datasets and PX Membership Agreement.

See also the original Nature Biotechnology publication and the 2017 and 2020 update papers.

Public Data

Public PXD datasets can be browsed over at ProteomeCentral. An RSS feed is also available.

ProteomeExchange database

- **Step 1**: Go to http://www.proteomexchange.org/ and select the option to Access public data.

You should see an interactive plot which shows you how many datasets have been deposited from different species, and using different types of spectrometer. How many studies are there from Rat? -> 766

- **Step 2**: Use the search to find datasets coming from pituitary and find a dataset from human which profiled this in 2019.
  - -> Proteomic analysis of human anterior pituitary gland
- **Step 3**: Which of the underlying hosting databases contains the full dataset for this study?
  - -> PRIDE project
- **Step 4**: Find the entry for this data in the underlying repository.
  - -> https://www.ebi.ac.uk/pride/archive/projects/PXD005819
- **Step 5**: Answer the following questions:

  1. What type of Mass Spectrometer generated this data? -> LTQ Orbitrap Velos
  2. Which publication is associated with the data? -> DOI: 10.1089/omi.2018.0160, PubMed: 30571610,

- **Step 6**: Have a look at the samples which were submitted as part of the study and see what information was recorded about each of them. -> **organism, organism part, disease,…**

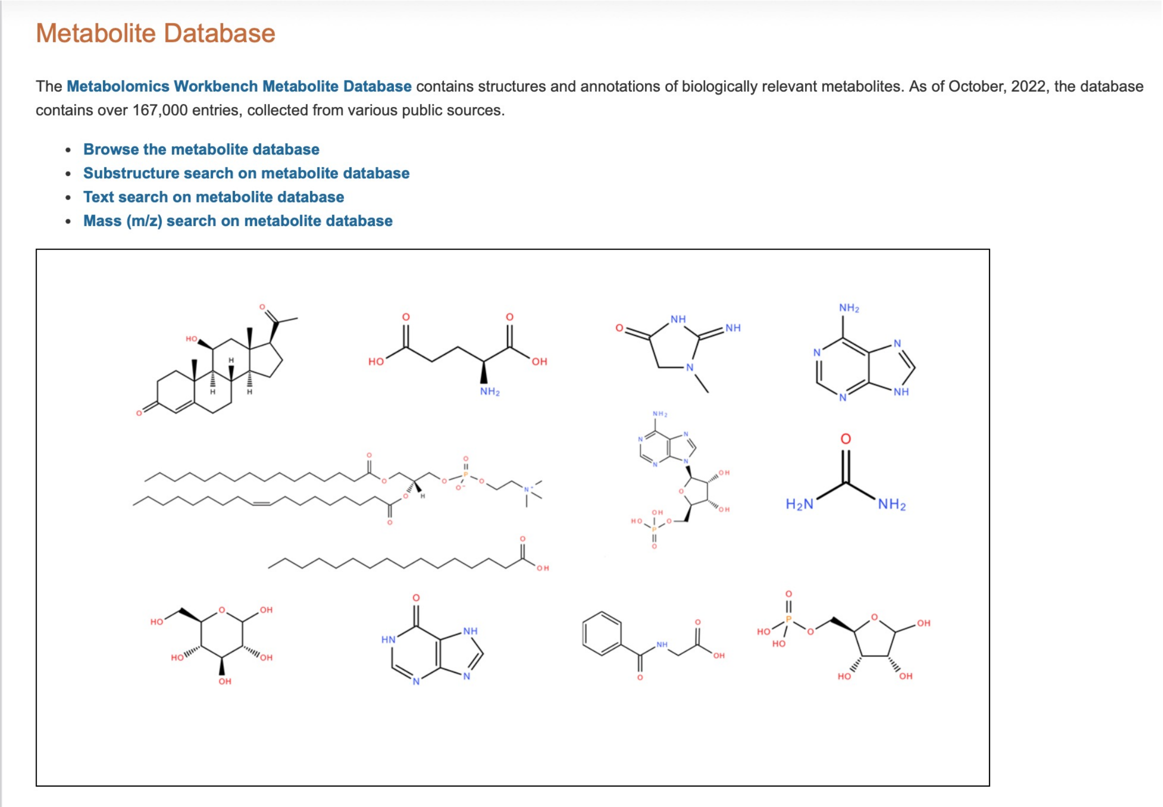
- **Step 7**: Click through to the FTP site associated with this data and check how many files you can see and what format they are in.

5. Metabolomic data in Metabolomics Workbench

Metabolomics Workbench

🌐 https://www.metabolomicsworkbench.org

The Metabolomics Workbench Metabolite Database contains **structures** and **annotations** of biologically relevant **metabolites**. The database contains over **167,000 entries**, collected from various public sources.



Metabolite Database

The **Metabolomics Workbench Metabolite Database** contains structures and annotations of biologically relevant metabolites. As of October, 2022, the database contains over 167,000 entries, collected from various public sources.

- **Browse the metabolite database**
- **Substructure search on metabolite database**
- **Text search on metabolite database**
- **Mass (m/z) search on metabolite database**

Metabolomics Workbench

- **Step 1:** Go to the main metabolomics workbench page at https://www.metabolomicsworkbench.org.

In the quick search at the top, search for the accession ST000899.

You should find one match – click through so you can see it.

- **Step 2**: Answer the following questions

- What was the purpose of this study? -> The aim of this study was to characterize serum metabolomic profiles in patients with IBD, and to assess for differences between patients with ulcerative colitis (UC), Crohn disease (CD), and non-IBD subjects.

- What biological material was collected for it? -> Serum samples from 20 UC, 20 CD, and 20 non-IBD control subjects

- What experimental conditions does it contain and how many samples from each condition are there? 20 UC, 20 CD, and 20 non-IBD control subjects

- **Step 3**: Select the option to Show Named Metabolites to see what compounds were detected in the study, remembering that these will be metabolites rather than proteins.

Note that the list of molecules is divided into sections based on the type of Ionisation which was used to detect them. Different molecules are efficiently detected using different types of ionisation (positive and negative) in different runs of the spectrometer.

Metabolomics Workbench

- **Step 4:** From the list of metabolites find adipoylcarnitine.

Select it then press the button at the top to show the values for this metabolite and then draw a bar graph.

Remembering that the samples are in groups of 20, does it look like something interesting might be happening with this metabolite?

- **Step 5:** Go back and draw a boxplot by factor. Which of the conditions does this metabolite seem to be downregulated in?

Metabolomics Workbench

- **Step 6:** We can do a larger scale analysis of the differential abundance of all metabolites between different conditions.

Go back to the main study page and select Performance statistical analysis.

We will construct a volcano plot of the results of a pairwise comparison of two conditions.
This plots the p-value (y-axis) against the magnitude of change (x-axis).
Select the volcano tool, and choose the Control as Group1 and Crohn disease as group 2 and run the analysis with the default options.

Have a look at the results, there is a table of results at the bottom of the page and you can click through to see the volcano plot and other summaries of the results. You should be able to see the adipoylcarnitine as an outlier along with several other metabolites.

- **Step 7:** Repeat the analysis on the same groups using the **negative ion mode** data instead of positive ion mode.

Note how you get a different set of hits. In negative ion mode you should see that sucrose is a strong positive hit.

- **Step 8:** Go back to the original full list of metabolites and find sucrose on the list and draw the barplot for it and check that you can see a strong increase in the crohn disease samples (21-40).