RNA-Seq Practice

Public Datasets

Day 03

https://tttorres.github.io/transcriptomics/

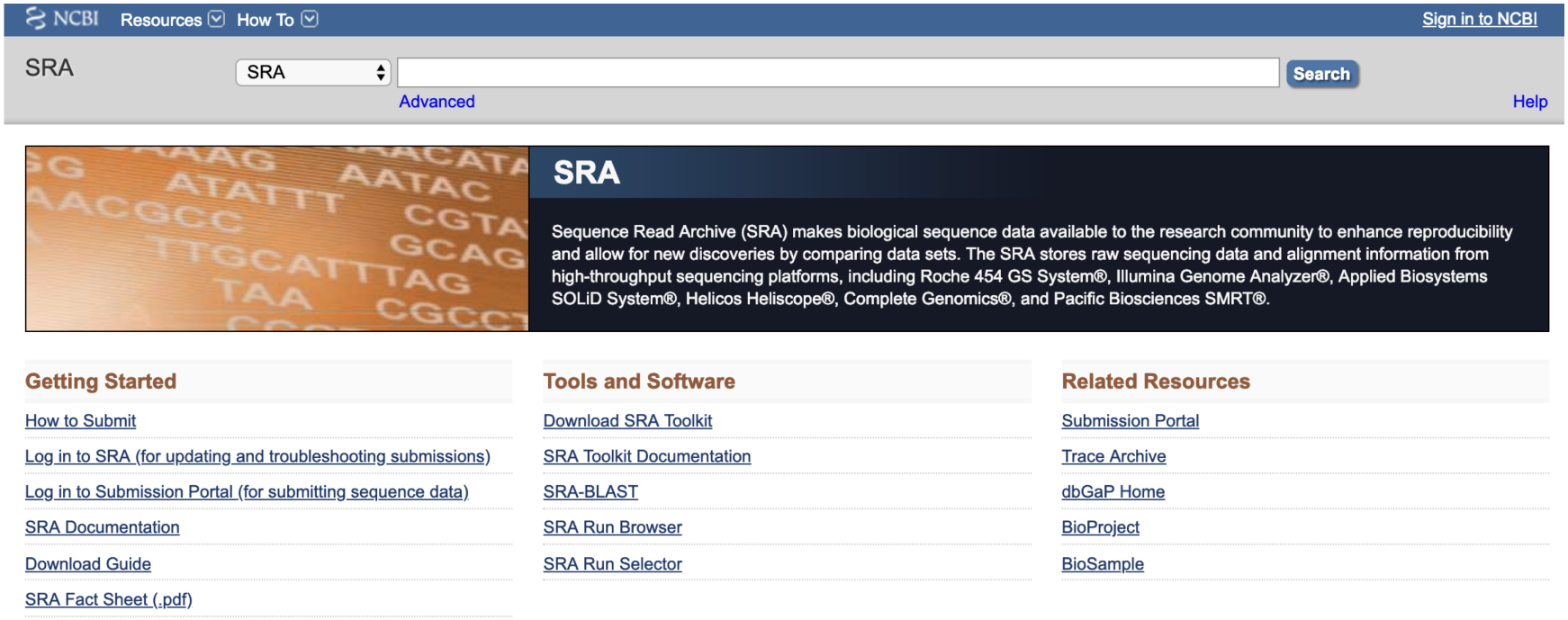
Public data

Why use public data?

1. You liked a paper and want to look at their data
2. Metanalyses with many samples or multiple studies
3. New biological questions that were not explored using the public dataset

The Sequence Read Archive (SRA)

Home to all sequence data



https://www.ncbi.nlm.nih.gov/sra

The Sequence Read Archive (SRA)

Home to all sequence data

1. **Focus**:
    - Stores raw sequencing data generated by high-throughput sequencing technologies, such as Illumina, Oxford Nanopore, and PacBio.
    - Includes raw reads from DNA or RNA sequencing experiments.
2. **Data Type**:
    - Raw sequence reads (FASTQ or equivalent formats).
    - Metadata about the sequencing experiment (e.g., sample preparation, sequencing method, study objectives).

The Sequence Read Archive (SRA)

Home to all sequence data

3. **Purpose**:
    - To preserve the unprocessed sequence data that can be re-analyzed using updated tools or pipelines.
    - Allows researchers to verify analyses or develop new approaches for processing raw data.
4. **Examples of Data**:
    - Whole-genome sequencing.
    - RNA-Seq raw reads.
    - Metagenomic data.

The Sequence Read Archive (SRA)

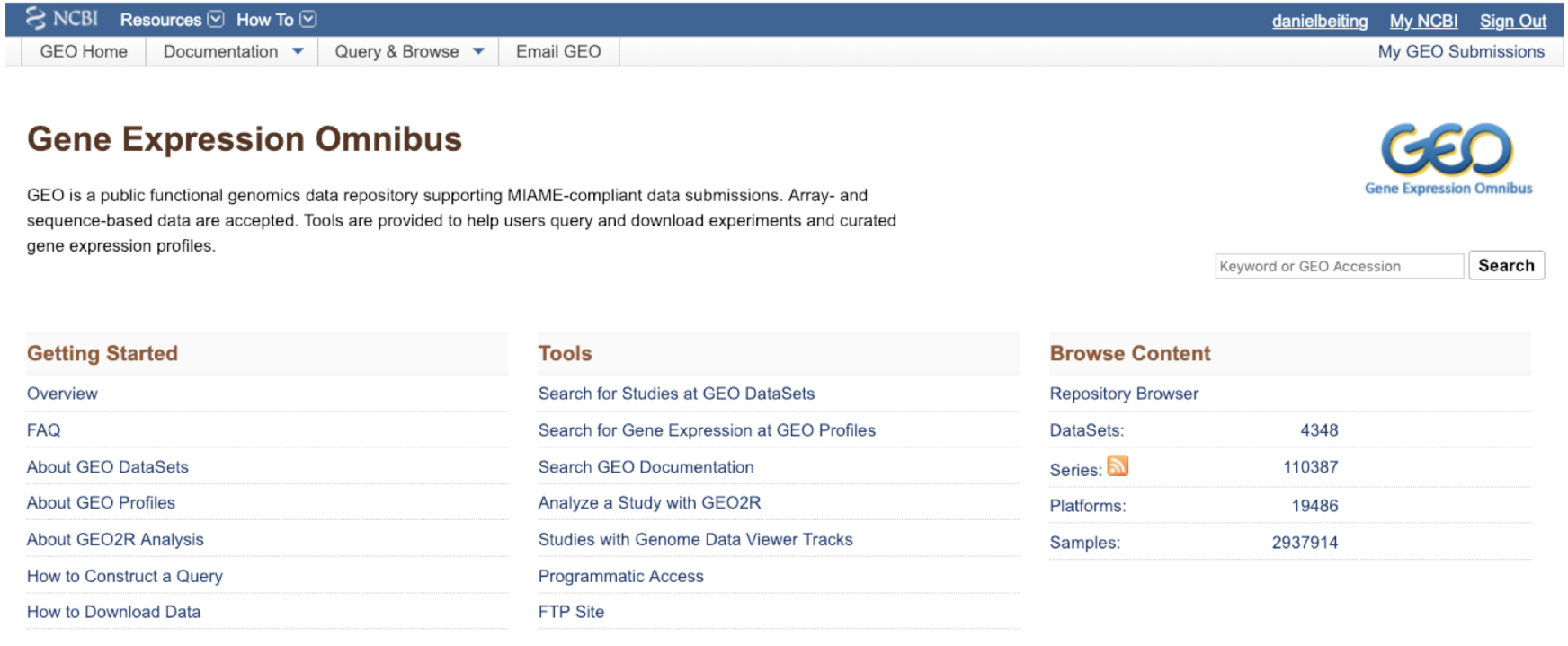
Home to all sequence data

Use Case

- You'd use SRA if you're looking to access raw sequencing data to perform custom downstream analyses, such as:
  - read alignment,
  - transcript assembly, or
  - variant calling.

Gene Expression Omnibus (GEO)

Home to Gene Expression Experiments



https://www.ncbi.nlm.nih.gov/geo/

Gene Expression Omnibus (GEO)

Home to Gene Expression Experiments

1. **Focus**:
    - Stores processed and analyzed data related to gene expression, including transcriptomics, epigenomics, and functional genomics studies.
    - Datasets from various high-throughput methods like microarrays and RNA-Seq.
2. **Data Type**:
    - Processed and normalized data (e.g., gene expression matrices).
    - Metadata about the experiment, including sample descriptions, experimental conditions, and methods.

Gene Expression Omnibus (GEO)

Home to Gene Expression Experiments

3. **Purpose**:
    - To serve as a repository for data that has already been processed and is ready for interpretation.
    - Facilitates data reuse for comparative analyses, hypothesis generation, or training machine-learning models.
4. **Examples of Data**:
    - Differential gene expression data.
    - Normalized read counts.
    - Expression profiles of genes under specific conditions.

Gene Expression Omnibus (GEO)

Home to Gene Expression Experiments

Use Case

- You'd use GEO if you're interested in accessing ready-to-use gene expression datasets or other processed genomic information to:
  - explore patterns,
  - validate findings, or
  - run secondary analyses.

The Sequence Read Archive (SRA)

Accessing public data on SRA

Some aligners can access data from SRA

Hisat2

```
hisat2 --sra-acc SRR2145310
```

Bowtie2

```
bowtie2 --sra-acc SRR2145310
```

The Sequence Read Archive (SRA)

Command line tools for SRA (SRA Toolkit)

fastq-dump

```
fastq-dump
```

fasterq-dump

```
fasterq-dump SRR2145310 -e 4
```

They do not get metadata (sample info)

https://hpc.nih.gov/apps/sratoolkit.html

The Sequence Read Archive (SRA)

Command line tools for SRA (SRA Toolkit)

A message from the developer:

> It is not unusual for users to get errors while downloading SRA data with prefetch, fasterq-dump, or hisat2, because many people are constantly downloading data and the servers can get overwhelmed.

The Sequence Read Archive (SRA)

Getting data and metadata from SRA

grabseqs

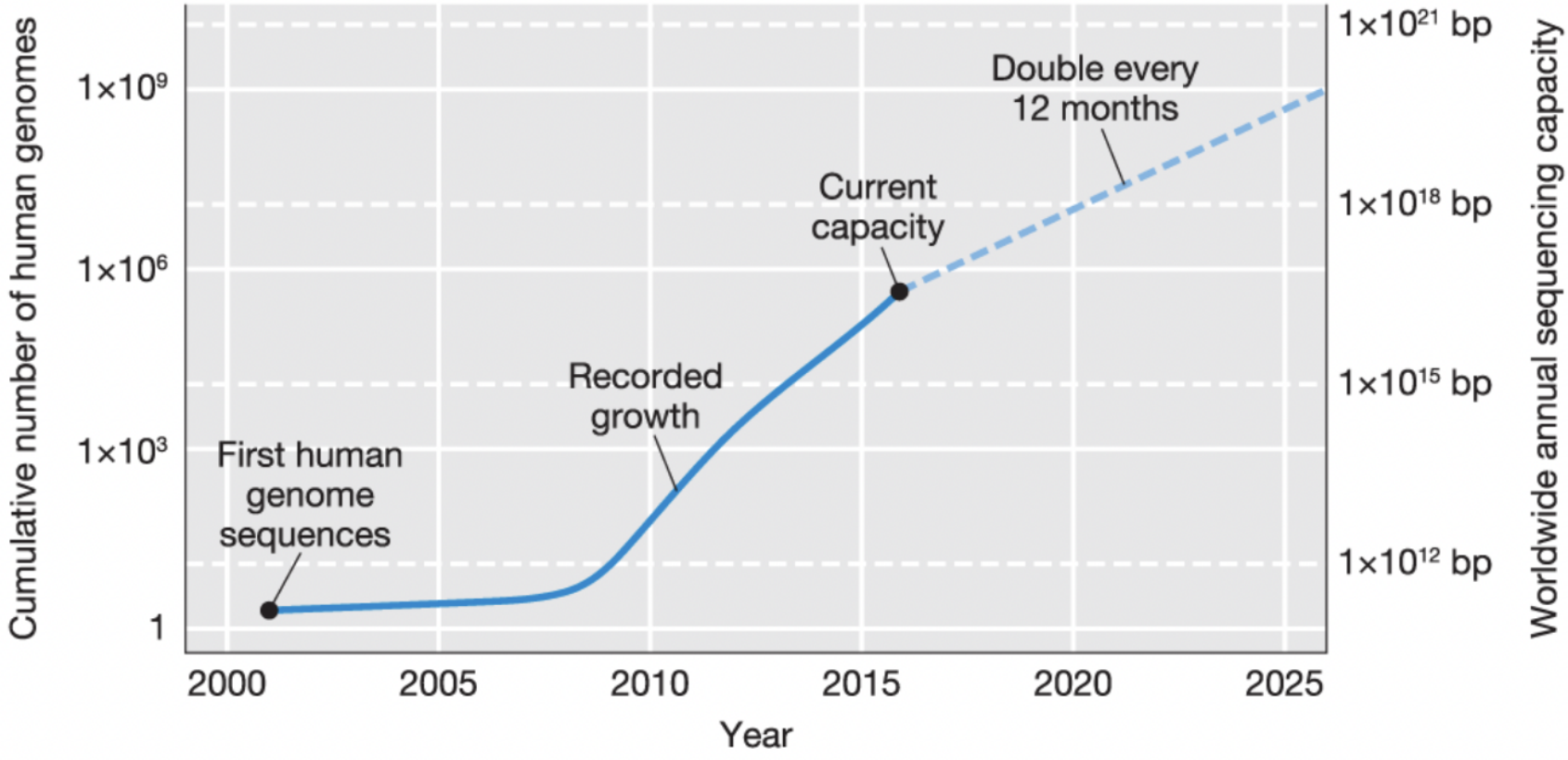
```
grabseqs sra -t 4 -m metadata.csv SRR8668755
```

> A utility for easy downloading of reads from next-gen sequencing repositories like NCBI SRA

https://github.com/louiejtaylor/grabseqs

The Sequence Read Archive (SRA)

Size of data / computer resources



Solution: community-driven projects

Public datasets

Genotype-Tissue Expression (GTEx) Project

GTEx

The Adult GTEx project is a comprehensive resource of WGS, RNA-Seq, and QTL data from samples collected from 54 non-diseased tissue sites across ~1000 adult individuals.

Explore »

dGTEx

The Developmental GTEx (dGTEx) project is a new effort to study development-specific genetic effects on gene expression and to establish a new data analysis and tissue biobank resource.

**Data Not Yet Available*

Explore »

NHPdGTEx
Non-Human Primate Developmental Genotype-Tissue Expression

The Non-Human Primate Developmental GTEx (NHP-dGTEX) project is a complement to dGTEx in 2 translational non-human primate model species: the rhesus macaque and common marmoset.

**Data Not Yet Available*

Explore »

https://www.gtexportal.org/

Public datasets

Genotype-Tissue Expression (GTEx) Project

> The Genotype-Tissue Expression (GTEx) Portal is a comprehensive public resource for researchers studying tissue and cell-specific gene expression and regulation across individuals, development, and species, with data from 3 NIH projects.

> A public resource to study tissue-specific gene expression and regulation. Samples from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq.

Public datasets

The Cancer Genome Atlas (TCGA) project

> 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.

An official website of the United States government

NIH NATIONAL CANCER INSTITUTE
Center for Cancer Genomics

Search

Research Access Data Funding News & Events About CCG Contacts & Help

Home > Research > Genome Sequencing > The Cancer Genome Atlas Program (TCGA)

TCGA

- Program History
- TCGA Cancers Selected for Study
- Publications by TCGA
- Using TCGA

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.

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

Public datasets

Other examples

- Encyclopedia of DNA Elements (ENCODE) project
- Flybase
- Vectorbase

and many others

Downloading public datasets

Installing SRA Toolkit with Conda

Step 1: Install Conda (if not already installed)

- If you don't have Conda installed:
  - Download and install Miniconda or Anaconda.

Installing software using CONDA

SRA Toolkit

1a. Unix: Create a new environment `sra`:

```
conda create --name sra
conda activate sra
```

1b. MacOS: Create a new environment `sra`:

```
CONDA_SUBDIR=osx-64 conda create -n sra
conda activate sra
conda env config vars set CONDA_SUBDIR=osx-64
```

Slide 2: Installing SRA Toolkit

Step 2: Install SRA Toolkit

Run the following command in your terminal:

```
conda install -c bioconda sra-tools
```

- This installs the SRA Toolkit from the Bioconda channel.

Verifying Installation

Verify Installation

1. Check if the SRA Toolkit is installed correctly:

```
fastq-dump --version
```

2. The version of the SRA Toolkit should be displayed if the installation was successful.
3. Create a new folder `sra` (`~/sra`)

```
mkdir sra
cd sra
```

Optional Configuration

Configure SRA Toolkit

1. Set up the SRA Toolkit to use a specific download directory:

```
vdb-config --interactive
```

2. Using the interactive menu:
    - Configure the default download location for `.sra` files.
    - Save and exit.

Key Commands Overview

Commonly Used Commands in SRA Toolkit

- **`prefetch`** : Download `.sra` files from SRA.

```
prefetch <accession>
```

- **`fastq-dump`** : Convert `.sra` files to FASTQ format.

```
fastq-dump <accession>
```

- **`fasterq-dump`** : Faster alternative for converting `.sra` to FASTQ.

```
fasterq-dump <accession>
```

Example Usage

Example: Downloading RNA-Seq Data

1. Use `prefetch` to download data for BioProject **PRJNA608454**:

```
prefetch PRJNA608454
```

2. Convert `.sra` files to FASTQ format:

```
fasterq-dump --split-files SRX7792860
```

Repeat for SRX7792859, SRX7792858, SRX7792857