Topics Discussed in Meeting: DNA Sequencing Project

Meeting 1

April 9, 2025 | Wednesday

Goal: To build a classification ML model that will be trained on Cancerous & Non-Cancerous Human DNA Sequences.

Starter Resources (Shared by our Professor)

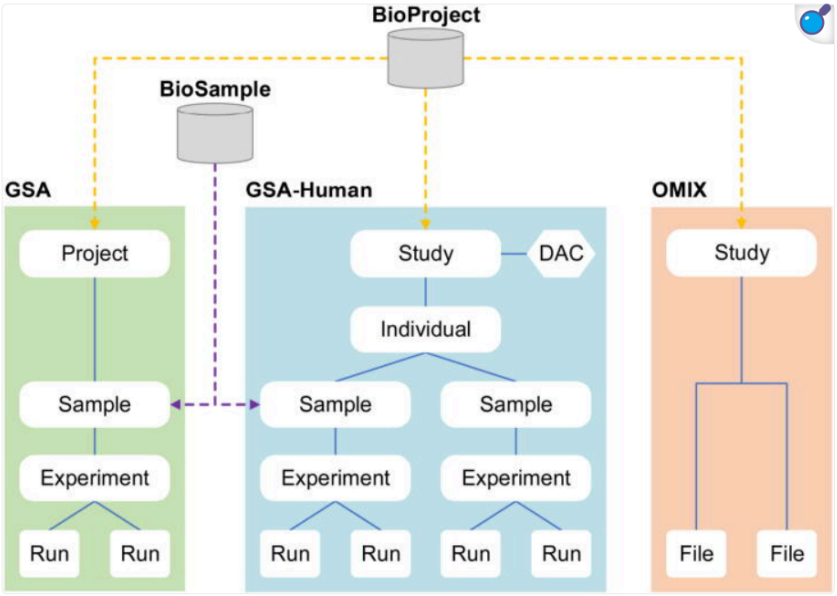
What is DNA & How Does It Work?

What is a Gene?

Genome Sequence Archive (GSA)

The data model is as follows:

Figure 1.



Data model of the GSA family. BioProject and BioSample are two independent meta-information databases, acting as an organizational framework to provide centralized access to descriptive metadata about research projects and samples, respectively. GSA-Human is for archiving human genetic data and OMIX is for various types of data (that are unsuitable for GSA/GSA-Human).

here we present the GSA family by expanding into a set of resources for raw data archive with different purposes, namely,

1. GSA (https://ngdc.cncb.ac.cn/gsa/),
2. GSA for Human (GSA-Human, https://ngdc.cncb.ac.cn/gsa-human/)
3. Open Archive for Miscellaneous Data (OMIX, https://ngdc.cncb.ac.cn/omix/).

Compared with the 2017 version, GSA has been significantly updated in data model, online functionalities, and web interfaces. GSA-Human, as a new partner of GSA, is a data repository specialized in human genetics-related data with controlled access and security. OMIX, as a critical complement to the two resources mentioned above, is an open archive for miscellaneous data. Together, all these resources form a family of resources dedicated to archiving explosive data with diverse types, accepting data submissions from all over the world, and providing free open access to all publicly available data in support of worldwide research activities.

Dataset Source 1

Whole exome sequencing for primary breast cancer and lymph node metastasis samples (SRR5177930)

So currently we are looking for datasets of cancerous patients (breast cancer) and their sequences. The website for Data Retrieval is as follows:

Dataset Source 1: Cancerous Sequence

This is the most suitable dataset that we need for our project.

Based on this run information:
Run Information (The ~NCBI Link is here)

We go to this website:
~NCBI Link - SRA: Sequence Read Archive Download

1. *We will try to retrieve the dataset using this GitHub Repo:*
   Retrieval Using Node JS Wrapper: SRA Toolkit
2. *And the next step would be to read this dataset in Python using the FastQ Library:*
   FastQ Library

Why FastQ
The DNA sequences need to be converted into a format python can understand.

Dataset Source 2

Search ongoing - This dataset shall consist of a sequence of non-cancerous patients.

Search completed on May 30, 2025: the data source is as follows:

DNA Exome Seq of Homo sapiens: Matched Normal Breast tissue (SRR6269879)
Run Information 2

The best method to find the dataset would be:

- Go to the ENA (European Nucleotide Archive)
- 2-3 keywords (e.g tissue type, genomic, nature of tissue etc)
- Then click on the suitable experiment and copy paste the Run accession in GSA.