WEBLEM 9
Introduction of Genome Database

- ## Genome Database:

The Genome Database (GDB) is a public repository of data on human genes, clones, STSs, polymorphisms and maps. GDB entries are highly cross-linked to each other, to literature citations and to entries in other databases, including the sequence databases, OMIM, and the Mouse Genome Database. Mapping data from large genome centers and smaller mapping efforts are added to GDB on an ongoing basis. The database can be searched by a variety of methods, ranging from keyword searches to complex queries. Major functionality extensions in the last year include the ongoing computation of integrated human genome maps, called Comprehensive Maps, and the use of those maps to support positional queries and graphic displays. The capabilities of the GDB map viewer (Mapview) have been extended to include map printing and the graphical display of ad hoc query results. The HUGO Nomenclature Committee continues to curate the proposed and official gene symbols and related data in collaboration with GDB. As genome research shifts its emphasis from mapping to sequencing and functional analysis, the scope of the GDB schema is being extended. We are in the process of adding representations of gene function and expression, and improving our representation of human polymorphism and mutation.

1. Genome Databases in NCBI:

The Genome database contains sequence and map data from the whole genomes of over 1000 species or strains. The genomes represent both completely sequenced genomes and those with sequencing in-progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

2. GOLD:

Genomes OnLine Database (GOLD) is an open-access repository of genome and metagenome sequencing projects with their associated metadata. Login-free access is provided to a growing catalogue of manually curated public projects from all over the world. Starting from whole genome sequencing of cultured bacteria, uncultured single cells, complex eukaryotes or metagenome sequencing of environmental samples, the world of sequencing is advancing at a rapid pace. In overcoming the limitations of culturing microbial isolates, improved sequencing and analysis methods have broadened our understanding of the microbial world. Metagenome- assembled genomes (MAGs) and single-amplified genomes (SAGs) (1,2) are two such examples of genomes of uncultivated organisms that have recently expanded our knowledge about microorganisms. In any sequencing project, metadata or accessory information about the sample being sequenced plays a key role.In order to interpret raw sequence data and make accurate scientific predictions, the presence of a wide variety of metadata is extremely important. For example, comparative analyses of the microbial community of soil samples from multiple locations would be very difficult if the geographic location information or soil type is not recorded. On the other hand, the scientific value of the same sequence will likely increase several-fold if detailed environmental and physicochemical properties of the soil are collected.

WEBLEM 9/a
(URL: https://www.ncbi.nlm.nih.gov/genome/)

Aim:
To study the Staphylococcus epidermidis in NCBI genome browser

Introduction:
The Genome database contains sequence and map data from the whole genomes of over 1000 species or strains. The genomes represent both completely sequenced genomes and those with sequencing in-progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

Staphylococci are common bacterial colonizers of the skin and mucous membranes of humans and other mammals. *S. epidermidis* in particular is the most frequently isolated species from human epithelia. It colonizes predominantly the axillae, head, and nares. Analysis of the *S. epidermidis* genome indicated that the species is well equipped with genes assumed to provide protection from the harsh conditions encountered in its natural habitat. For example, to cope with extremes of salt concentration and osmotic pressure, *S. epidermidis* has eight sodium ion/proton exchangers and six transport systems for osmoprotectants.

Methodology:
1. Open The homepage of Genome in NBCI
2. Enter the query in Search bar.
3. Open any one of the result page.
4. Interpret the Result.

Observation:
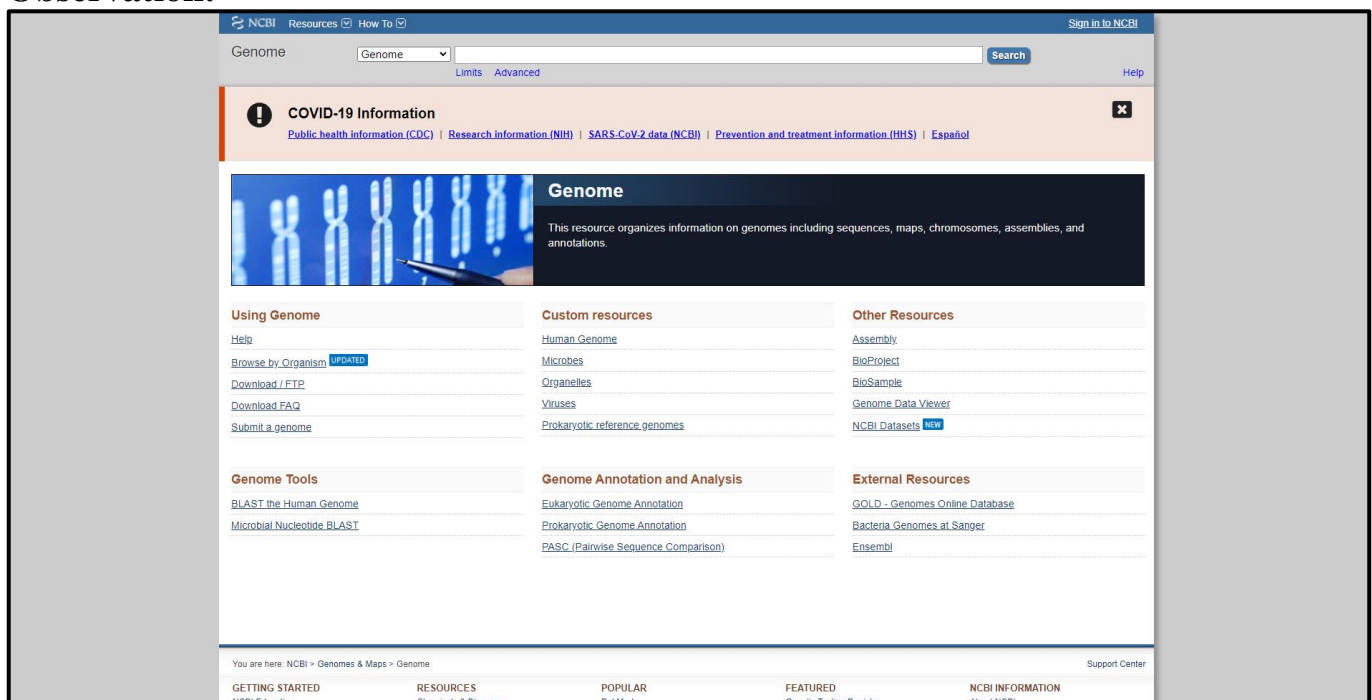
Fig1. Homepage of Genome in NCBI
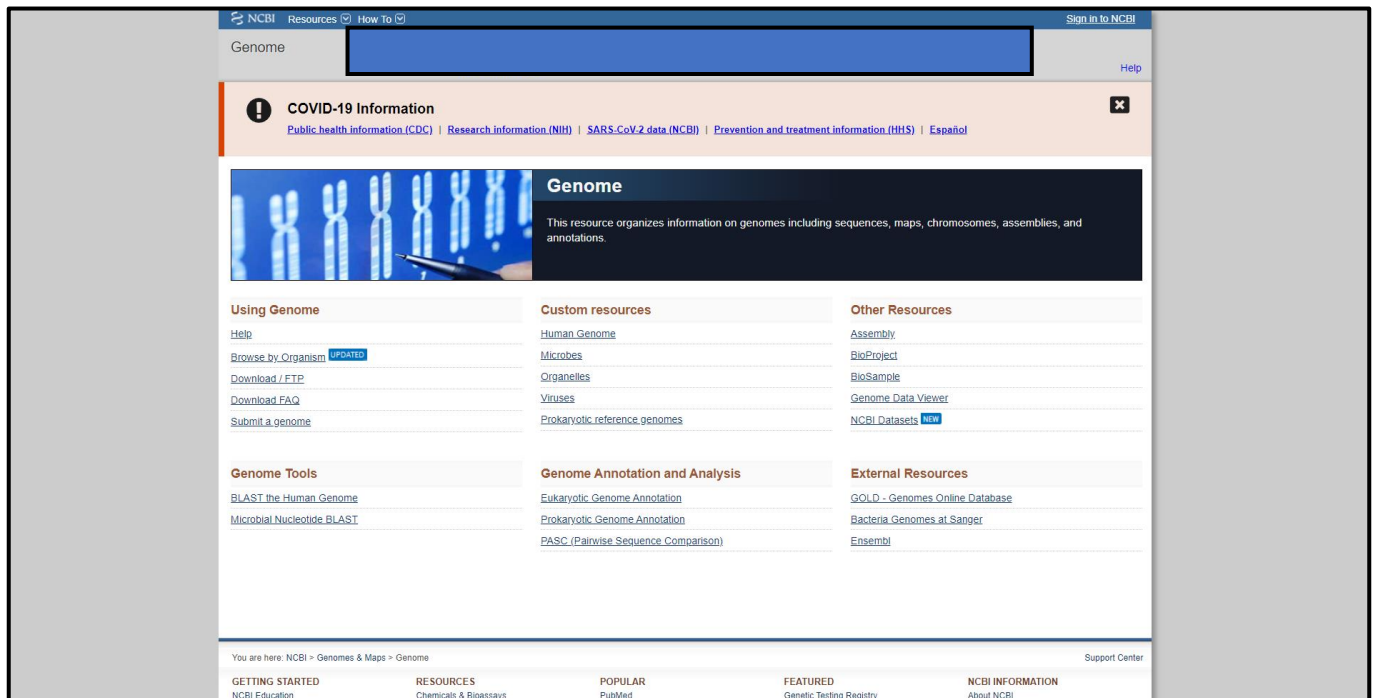
Fig2. Search page of Genome in NCBI with query Staphylococcus epidermidis
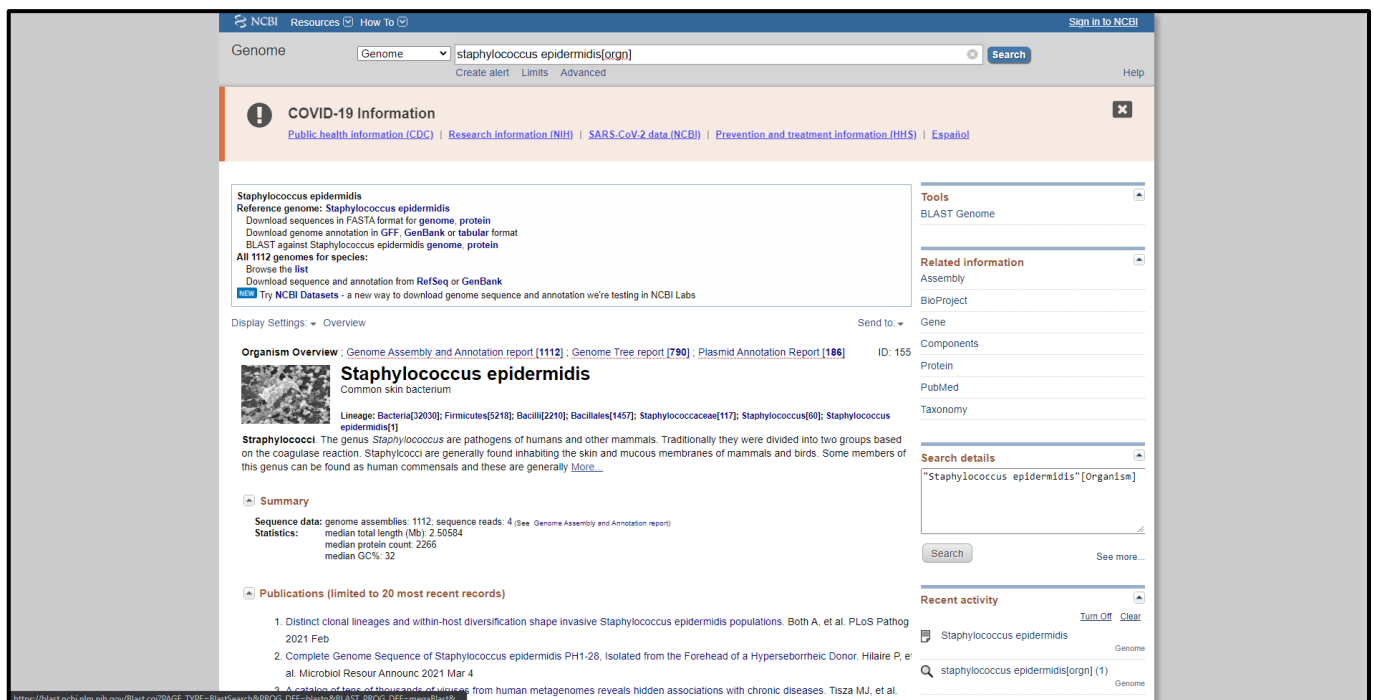
Fig3. Result page of Genome in NCBI for query Staphylococcus epidermidis.

Fig4. Genome annotation Table with query Staphylococcus epidermidis

Fig5. Result page for query "Staphylococcus epidermidis" showing publications and genome information for reference and representative genomes.

Fig6. Reference sequence information for query "Staphylococcus epidermidis" in NCBI GenBank database.
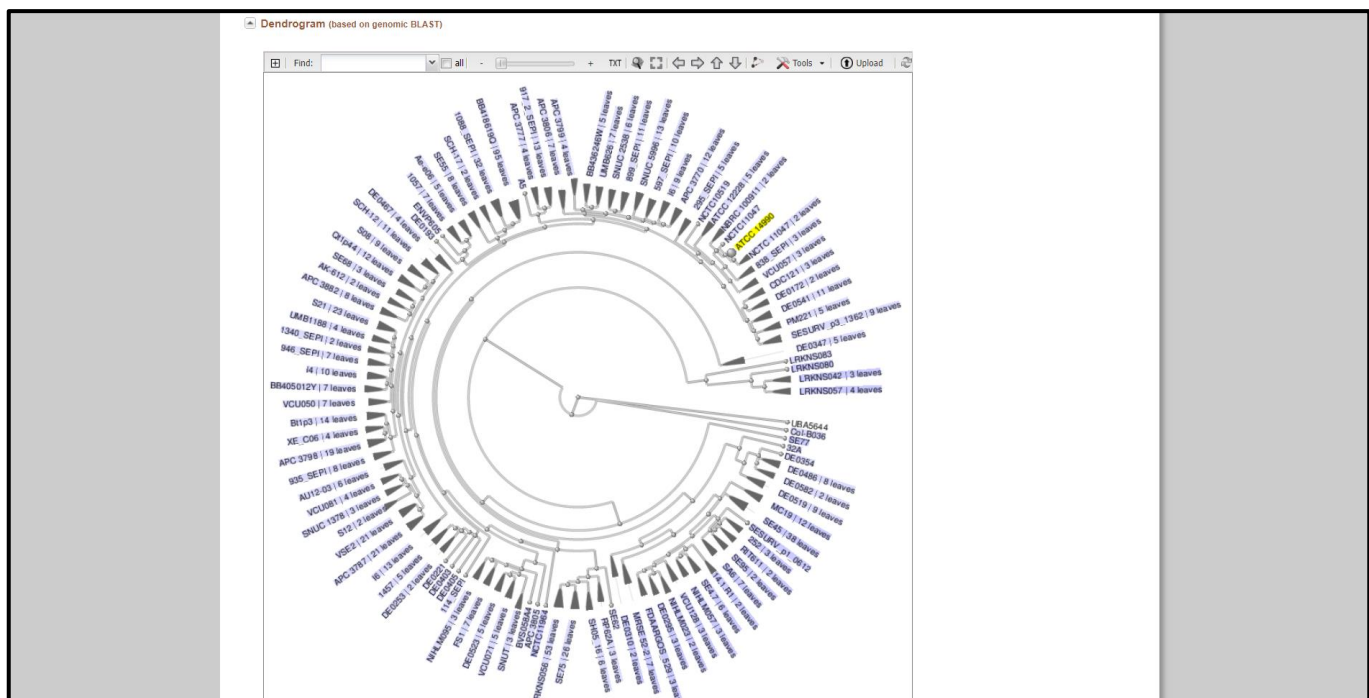
Fig7. Result page for query "Staphylococcus epidermidis" showing dendrogram on genomic BLAST

Fig8. Genome assembly and annotation report for query "Staphylococcus epidermidis"

Result:

The genome information for *"Staphylococcus epidermidis"* was retrieved using the NCBI genome browser. 2216 genome annotation results were observed. Genome assembly and annotation report gave 1112 results. Genome information for reference and representative genomes and dendrogram information was also observed.

Conclusion:

NCBI genome browser can be used to study genome information on various organisms (highly effective for bacteria and other microorganisms). It provides sequences, maps, chromosomes, assemblies and annotations.

References:

1. Genome List - Genome - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/genome/browse/#!/proteins/155/570875|Staphylococcus epidermidis/
2. Genome List - Genome - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/genome/browse/#!/prokaryotes/155/
3. Home - Genome - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/genome/
4. Otto, M. (2009, August). Staphylococcus epidermidis--the 'accidental' pathogen. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2807625/
5. Staphylococcus epidermidis (ID 155). (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/genome/?term=Staphylococcus epidermidis/
6. Staphylococcus epidermidis (ID 155). (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/genome/?term=Staphylococcus epidermidis/
7. Staphylococcus epidermidis strain ATCC 14990 chromosome, complete geno - Nucleotide - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP035288.1

WEBLEM 9/b
GOLD
(URL: https://gold.jgi.doe.gov/)

Genomes OnLine Database (GOLD) is an open-access repository of genome and metagenome sequencing projects with their associated metadata. Login-free access is provided to a growing catalogue of manually curated public projects from all over the world. Starting from whole genome sequencing of cultured bacteria, uncultured single cells, complex eukaryotes or metagenome sequencing of environmental samples, the world of sequencing is advancing at a rapid pace. In overcoming the limitations of culturing microbial isolates, improved sequencing and analysis methods have broadened our understanding of the microbial world. Metagenome-assembled genomes (MAGs) and single-amplified genomes (SAGs) are two such examples of genomes of uncultivated organisms that have recently expanded our knowledge about microorganisms.

There are three different ways in which projects and their associated metadata are entered into GOLD:

- Projects that are added by individual GOLD users.
- Projects for samples sequenced at the Joint Genome Institute (JGI) as part of the JGI User Programs and
- Projects sourced from public databases such as the NCBI GenBank and Sequence Read Archive (SRA).

Well-defined genome and metagenome projects in GOLD are a prerequisite for submitting sequence data to the Integrated Microbial Genomes with Microbiome (IMG/M) system for annotation and analysis. GOLD and IMG work closely to provide a seamless experience to our broad user base of scientists, students, and policy makers as well as novice science enthusiasts. Users can browse through several metadata fields in GOLD and take advantage of the multiple filters and advanced search tools to identify a subset of genomes or metagenomes of interest. They can then use the IMG identifiers from GOLD and directly go to the IMG user interface and conduct further research and comparative analyses.

GOLD Organization and Current Status:
GOLD is organized in a four-level hierarchical system to describe the overall proposal, samples or organisms studied, sequencing projects undertaken and their analysis process. These levels include Study, Organism or Biosample, Sequencing Project (SP), and Analysis Project (AP). All four levels, along with their complex list of metadata fields and controlled vocabulary (CV) terms, are connected to each other in a lucid framework to enhance scientific discovery.

A brief description of each of these four levels is provided below:

Study: The overall research objectives and goals are captured in a Study, which lies at the helm of the four-level organization structure. Studies can vary in the type of samples collected. Subsequently, a single Study may have several Sequencing and Analysis Projects that differ in their methodology and application such as Whole Genome Sequencing (WGS) and analysis, metagenome analysis or a combination of both.

Organism/Biosample: A GOLD Organism or Biosample contains descriptive information about the biological or physical material that is being sequenced. Any living entity such as bacteria, archaea, fungi, virus, plant or animal may constitute an Organism in GOLD. An Organism may be cultured, uncultured

(such as single cells) or even bioinformatically predicted (such as Metagenome-Assembled Genomes or MAGs). The most important metadata feature of a GOLD Organism is its Taxonomic assignment including phylum, genus, species, strain etc. To promote interconnectedness with other databases, all GOLD Organisms have the NCBI Taxonomy ID.

Sequencing Project: The process of generating sequencing data from a Biosample or Organism is described in a Sequencing Project (SP). GOLD currently has 15 different types of SPs, from which whole-genome sequencing (WGS) and metagenome are most commonly used. The input material for an SP can either be DNA or RNA corresponding to a genome or transcriptome project, respectively. This material can come fromeither an organism, in the case of WGS and transcriptomes, or from a Biosample, in the case of metagenomesand metatranscriptomes.

Analysis project: The information about the processing of sequence data and its analysis is captured in a GOLD Analysis Project (AP). A GOLD AP is required in order to submit the corresponding sequence data to IMG for annotation. There are 13 different types of Analysis Projects that can be defined in the current version, namely: Genome, Metagenome, Metagenome—Cell Enrichment, Metagenome—Single Particle Sort, Metatranscriptome, Single Cell (Unscreened), Single Cell (Screened), Metagenome-Assembled Genome, as well as five different types of Combined Assembly.

BROWSING AND SEARCHING GOLD:

Public data in GOLD are freely accessible through the website (https://gold.jgi.doe.gov). The homepage provides key statistics on different entities, along with links to related metadata.

Search: The Search tab has two options: Advanced Search and Metadata Search. While there are different ways to search the database, one of the most popular and commonly used features is the 'Advanced Search.' It provides a quick and easy way to search across all the different levels using several metadata fields.

Distribution Graphs: The Distribution Graphs section gives a top-down view of different fields in which the user receives a snapshot of the metadata scale and their diversity. A brief summary of the different types of Sequencing Projects for genome and metagenome projects and their sequencing statuses are displayed as piecharts in the Distribution Graphs section. Additionally, a user can look at the phylogenetic distribution of projects both as pie charts and in hyperlinked expandable table formats. Finally, it also shows the breakdown of Biosamples by each of the five different Ecosystem levels.

Biogeographical Metadata: The Biogeographical Metadata section shows the isolation location of Biosamples and Organisms as placeholders on interactive maps.

Ecosystem classification: GOLD Ecosystem Classification paths describe the surroundings from which an environmental sample or an organism is collected. This five-level hierarchical classification system shown below, was originally described by Ivanova et. al. in a paper titled "A call for standardized classification of metagenome projects."

To facilitate the annotation and classification of a large number of diverse environmental samples and organisms we need a simple, versatile and adaptable classification system. GOLD ecosystem classification provides a simple but powerful framework to describe samples from environmental, host associated as well as engineered or built environments.

This classification system is not a comprehensive list of all possible paths from any given environment,

host or engineered setting. Instead it is primarily driven by the samples we curate. Within this framework we periodically review and revise these paths based on existing as well as the new samples we encounter.

The Five Levels of the classification system are as follows:

Ecosystem -> Ecosystem Category -> Ecosystem Type -> Ecosystem Subtype -> Specific Ecosystem

SRA Explorer: NCBI Sequence Read Archive (SRA) stores a collection of sequence data from all forms of life including metagenomic samples. The NCBI SRA Explorer provides a way to search SRA data. Using a combination of free text fields (such as organism name or SRA Experiment Title) and fields with predefined or controlled vocabulary based values (such as sequencing technologies or library strategies), the different types of data available in SRA can be reviewed. The SRA Explorer also has a filter where users can see which SRA data are currently present in GOLD.

Statistics: The Statistics tab contains graphical descriptions of several Sequencing Project metadata fields. This ranges from the phylogenetic distribution and relevance of bacterial genome projects, breakdown of projects by sequencing centers or the yearly growth of sequencing projects by organism domain.

Downloads: The Downloads page gives users the ability to access different types of data, ranging from a list of several controlled vocabulary terms, pubic metagenome and metatranscriptome projects from SRA that were annotated in IMG or a data dump of GOLD-containing select fields from various database entities. These files can be downloaded in Excel format and the underlying data is updated daily.

GOLD database thus can be used to retrieve genome and metagenome sequencing projects with their associated metadata information. It consists of organized four-level hierarchical system which includes study, organism or biosample, sequencing project, and analysis project and it provides various browsing and searching resources the makes it easier for the users to access information.

REFERENCES:
1. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J., Lee, J., . . . Reddy,T.
    a. (2020). Genomes OnLine Database (GOLD) v.8: Overview and updates. Nucleic Acids Research, 49(D1). doi:10.1093/nar/gkaa983
2. JGI gold: Home. JGI GOLD | Home. (n.d.). Retrieved October 20, 2021, from https://gold.jgi.doe.gov/