Cloud Genome

Database Design Document

Analia Treviño-Flitton

University of Maryland Global Campus

DBST 651: 9040

Nov/03/2020

Table of Contents

Introduction 1

Overview 1

Literature Review 1

Assumptions/Constraints/Risks 2

Design Decisions 2

Detailed Database Design 13

Database Administration and Monitoring 62

Introduction

This database design report defines the framework and identifies specifications for the development of a genomic database for the use by private organizations in the biotechnology sector. The database is owned solely by Cloud Genome Inc., therefore this document will only be accessible to stakeholders and those directly involved in the database development lifecycle. This document is protected with company owned rights and those with access to it are legally bound to abide by a privacy clause.

Overview

Cloud Genome is a cloud-based database system that will be housed on private servers owned by Cloud Genome. The database will only be accessible through the Cloud Genome GUI and will not be stored on our client's local servers. Accordingly, we do not expect to encounter any interference or interactions with other database management systems that our clients may also utilize. The aim of this database is to provide organizations with the highest quality of data security while also supporting immediate multi-user data access without additional hardware requirements.

Literature Review

The next hurdle the field faces in the biotechnology era is related to one of the fundamentals to the scientific method, reproducibility. It has long been established that scientific work must be able to be verified by scientific peers so that they can repeat the exact process performed and yield corroborating results. Guseva, Batyrgazieva, Karetkin, & Menshutina (2019) have expressed a similar concern citing the lack of a systemized approach and inefficiently designed databases, has hindered the advancement and integration of therapeutics into mainstream medical care, specifically pro- and pre- biotics. While Guseva et al. (2019) has worked with clinical research databases like

PROBIO, they voice concern over the need for a database that also focuses on in-vitro and animal studies rather than just for data from field-trials. The team has suggested a remodel of a more current database, ODRAP and include in their recommendation a proposal for an integrated web application for the database. This would allow the database to be accessible by any internet access portal which would communicate directly with the database server. This would help create an open platform that is easy-to-use and easy-to-access and promotes the open-science methodology pushing scientific discoveries forward.

Assumptions and Constraints

- It is assumed all entries will be linked by an organism, genome, and at least one reference material, or they will not be allowed to be entered.
- The Genome table will never be incomplete.
- Each protein and gene is genome and organism-specific. There can be repeats of these values but they will remain exclusive to their organisms and genomes.
- Any further table additions to the database beyond this point will be optional to the client. However, the five original founding tables listed in this document are mandatory.

Design Decisions

Key Factors

Our design began with an Entity-Relationship diagram focusing on the biological data that would need to be stored. Cloud Genome is made up of five different inter-related entities. Ref_Literature is responsible for housing the reference literature the data has been published from, Organism holds organism specific data, Genome lists the actual genetic sequence and some metadata relating to it, Gene has information pertaining to each gene, and protein has information about the

proteins yielded from the genes. It was important to emphasize the usefulness of the data by eliminating data redundancy, however in order to uphold foreign key constraints, there are some repeats. Some biological data can share rather interesting similarities that can be hard to predict so we felt it was more important to preserve those linked associations.

Functional Design

The database will function on our own servers and will communicate with our client's machines via our GUI software. The GUI requires the internet in order to access the current up to date database. An offline version of the database can be stored in the GUI cache but this must be set up with the database administrator for access.

Security and Performance Decisions

Ensuring the safety of our client's data is a priority at Cloud Genome. In order to better guarantee the security of our database, our servers are completely private within our own network and housed in a non-descriptive location with additional security measures. We employ our own network engineers and security professionals to ensure there are no potential employee breaches and conduct regular systems analysis weekly and updates as deemed necessary.

Statement of Work

Overview

Cloud Genome Inc. will host a cloud-based data repository that provides customers private access to their genomic research findings. Our projected customers are biotech companies and academic organizations involved in the research and development of new compounds for therapeutic applications. The design and launch of the database is the aim of this project and the cornerstone to the success of Cloud Genome. Our client's data will be based on the research summaries created by their teams from their research findings based on the organisms under analysis, to include information about their genes and derivative proteins. For our customers, the utility of using our database is found with the immediate access to their data, swift sharing within their organizations, and ensured security of their sensitive information.

Purpose and Objectives

The database will be used as a repository for large organizations involved in different stages of the drug development lifecycle. Our platform will serve as a collaborative stream to the different departments within the client's organization to simultaneously provide access to all research, past and present. Our clients will be able to manage and add their genomic information to their organization's database with customizable permissions as a safeguard that limit departments who can make alterations to the database.

This would allow a research team to access the findings of another team while the same information is being accessed in a meeting with the CSO and COO, without worrying about accidental commits along the way. We enhance our customer's productivity by eliminating data redundancy, providing data security, and cutting hardware costs required to house the data on-site and employee expenses necessary for hardware upkeep.

Project Scope

Our finished product will require many layers of development including user interface design (UI) applications, network systems, all security measures, and server implementation. However,

this project will solely focus on the creation of a proper and efficiently designed database. This is to include an entity-relationship model of the database design, project management reports, a summary explaining the data definition language (DDL) and data manipulation language (DML) for the project, and sample structured query language (SQL) scripts illustrating the development approach.

In-Scope Work

- Entity-Relationship Database model
- Project Management Reports
- DDL and DML Explanation
- Sample SQL Scripts

Out-of-Scope Work

- Server implementation
- Network systems
- Security
- UI/UX Applications

Database Goals, Expectations, and Deliverables

The goal of this project is to develop a functional database that stores genomic information for clients. Upon completion, it is expected to have an entity-relationship model highlighting the database structure, a statement of work clarifying project specifications, a technical report summarizing the project, the functional database with proper primary keys, and the DDL and DML scripts used during the development.

Database Benefits

The database will benefit our clients by providing a system that offers a structured organization for their genomic data. They will be able to access their information from a system that has streamlined the useability of big data. Cloud Genome will maintain their data quality leading to more reliable and replicable research. Cloud access will enhance their productivity by allowing multiple

users to view their work simultaneously and our private servers ensure their data is protected with the most secure tools.

Project Hardware and Software Tools

Hardware

- Consumer-based Intel(R) Core i5-6200 U CPU @ 2.30Hz 2.40 GHz

  12. GB RAM 64 bit OS x64-based processor

Software

- Google Chrome Version 85.0.4183.102

Office Productivity Tools

- Microsoft Office 365 running on Windows 10

Diagram Tool

- ER Assistant 2.10 running on Windows 10

Database

- Oracle Database 12c Enterprise Edition Release 12.2.0.1.0 64bit running on a Linux VDA

Client Access Method

- UMGC Virtual Desktop Access Intel and Xeon running Linux

SQL Usage and Style

Adapted from Simon Holywell's SQL Style Guide, https://www.sqlstyle.guide/

Data Definition Language (DDL) is used to define the database schema and properties of the data, it uses keywords such as CREATE, RENAME, ALTER, DROP, COMMENT.

Data Manipulation Language (DML) is used to maintain data already in schema objects, it uses keywords such as SELECT, INSERT, CALL, UPDATE, MERGE, DELETE, LOCK TABLE.

General Guidelines

- Special characters will not be used only numbers, letters, and underscores
- All keywords will be written to the left and in upper-case
- Multiline comments will use " /* xyz */" format
- Single line comments will use "-- xyz" format
- snake_case will be used rather than CamelCase
- Equal signs will be surrounded by spaces for reading ease

Naming Protocol

- All names will be singular, no plural names will not be used
- All names must begin with a letter and cannot end with an underscore
- Underscores should be used in place of spaces
- Names cannot be reserved keywords
- Table and columns must have different names
- When tables are concatenated, the table must have a new name -- not the old names joining together

Create Syntax

- Tables must have a minimum of one key
- Default values must match the declared column value
- Do not use vendor-specific data types

- Keys must retain exclusivity and be relatively simple

Query Syntax

- Abbreviations should be avoided
- Spaces should be embraced for readability and to align code segments
- Joins should be indented further
- Additional queries will be on a new line with an indention

Benefits of the Cloud-Based Approach

A strong selling-point of our product is the long-term expense conservation to our customers by the utilization of our private cloud. The organizations we serve are centered on the advancement of scientific ventures for global health and environmental rehabilitation efforts, they do not want to lose focus by the additional responsibility that data management entails. This eliminates the need for our customers to allocate additional property to house servers, hardware expenses, and increased employees for maintenance.

By choosing private cloud access, Cloud Genome holds total control over all the hardware and infrastructure used for our customer's accounts. This allows us to have increased security on-site and online, the ability to scale-up when needed, and the option to expand to other cloud-based platforms. Cloud utilization is beneficial as it does not rely on one single server or machine to store your data, rather it spreads segments of data across many server hard drives.

Requirements Definition Document

Business Rules

- Each entry (protein, gene, genome, and reference literature) must have an organism associated with it.
- One organism can be linked to one or more referencing literature but each reference will only be linked to one organism and one genome.

- The organism and genome must be the same for all genes and proteins with foreign keys.
- One organism will have only one genome and it must be present.
- A single genome will be associated with only one organism.
- Many genes can arise from the same genome but the genes must be genome specific. While many genes are shared among species, they often perform specific tasks and produce specific proteins in each organism.
- Repeat genes are allowed but they will remain exclusive to organism and genome..
- There can be zero, one or many proteins from only one gene but a gene must be present.
- Many proteins can be produced from a genome but they are organism and genome specific. Repeated proteins are allowed but they will not be linked.

Entity and Attributes

Entity Name: Organism

Entity Description: The Organism table will hold information pertaining to the organism including the primary key of Org_ID and it's scientific name. There is one foreign key Ref_Literature_REF_ID, linking it with Ref_Literature and Genome tables. Every entry must have an Org_ID.

Main Attributes of Organism:

- ORG_ID ID (Primary Key): ID assigned to an organism for location
- Scientific_Name: Scientific name on record for the organism
- Org_Type: A quick reference for the type of organism
- Host: The host where the organism was identified from
- Lineage: The species lineage of the organism

Entity Name: Ref_Literature

Entity Description: Ref_Literature holds the information from the reference material the organism was sourced from. It includes information about the Journal material it is published in, it's PubMed ID, and the date it was published. The foreign key is the ORG_ID from Organism.

Main Attributes of Ref_Literature:

- REF_ID (Primary Key): The ID the article is assigned by the private organization
- PUBMED_ID: The ID the article is assigned by NCBI's PubMed
- Journal: The name of the journal the article is published in.
- Journal_Volume: The volume of the journal the article is featured in
- Article_Title: The title of the journal article
- Pub_Date: The date the article was published

Entity Name: Genome

Entity Description: Genome stores the actual DNA sequence, the length of the genome, a FASTA formatted sequence ID, and the GC content of the genome. The foreign keys are Gene_GENE_ID and Organism_ORG_ID.

Main Attributes of Genome:

- GENOME_ID (Primary Key): The ID to locate the genome from NCBI
- FASTA_ID: The ID to locate the FASTA format of the genome
- DNA_Seq: The actual DNA sequence of the genome
- DNA_Length: The length of the genome
- GC_Content: The median percentage of GC base pairs in the genome

Entity Name: Gene

Entity Description: The Gene table is made up on the gene information found in the genome of the specific organism. Each gene is classified based on type, symbol, a brief description, and the last update to the entry. The foreign keys are the primary keys from Genome and Protein.

Main Attributes of Gene:

- GENE_ID (Primary Key): The ID to locate the specific gene
- Gene_Type: Identifies what the gene is used for ie protein-coding
- Gene_Symbol: Lists the shorthand symbol for the gene
- Gene_Descripton: A quick reference
- Last_Update: Lists the date for the last edit to the gene information

Entity Name: Protein

Entity Description: The purpose of this entity is to hold the information of identified proteins derived from genes. It includes the protein family, the protein sequence, and the molecular weight of the protein. There is one foreign key linking protein with the gene it arose from.

Main Attributes of Protein:

- PROTEIN_ID (Primary Key): The ID to locate the specific protein
- Pfam: The protein domain the protein is classified as
- Pro_Seq: The actual amino acid sequence of the protein
- Region_Name: The name of the region where the protein is found
- Mol_Weight: The molecular weight of the protein

Relationship and Cardinality

- Relationship: Cited between Ref_Literature and Organism
- Cardinality: 1:M between Organism and Ref_Literature
- Business rule: Each reference literature can only be based on one organism.

- Relationship: Cited between Organism and Ref_Literature
- Cardinality: M:1 between Ref_Literature and Organism
- Business rule: There can only be one organism but many referencing literature.

- Relationship: Derived between Genome and Organism
- Cardinality: 1:1 between Genome and Organism
- Business rule: There can only be one genome per organism.

- Relationship: Derived between Organism and Genome
- Cardinality: 1:1 between Organism and Genome
- Business rule: Each organism must have only one genome.

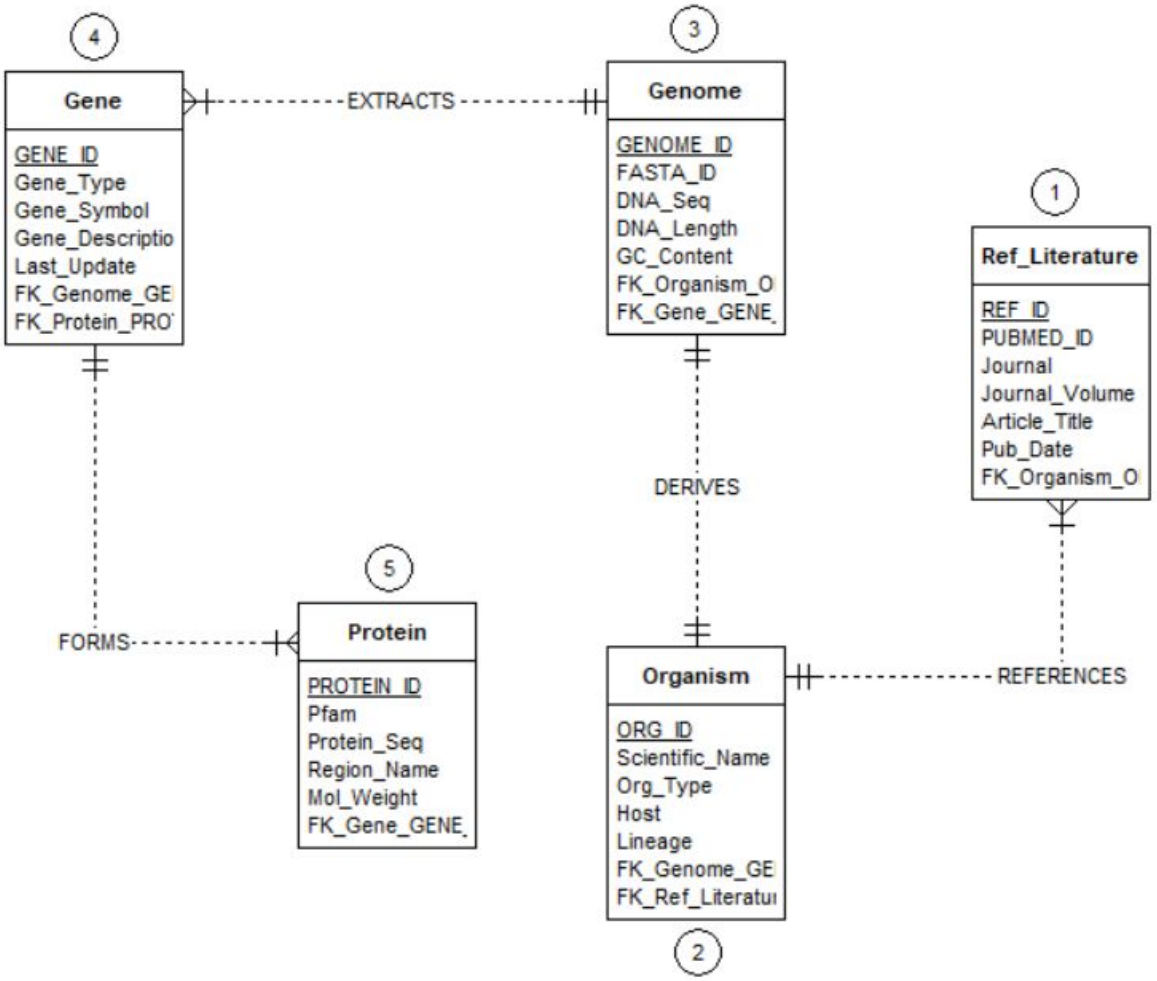
- Relationship: Extracted between Genome and Gene
- Cardinality: 1:M between Genome and Gene
- Business rule: Only one genome can be used that can have many genes.

- Relationship: Extracted between Gene and Genome
- Cardinality: M:1 between Gene and Genome
- Business rule: Many genes can arise from one genome. There will always be only one genome.

- Relationship: Formed between Gene and Protein
- Cardinality: 1:M between Gene and Protein
- Business rule: One gene can produce many proteins.

- Relationship: Formed between Protein and Gene
- Cardinality:: M:1 between Protein and Gene
- Business rule: Many proteins can arise from one gene.

Detailed Database Design



Design Justification

1 Ref_Literature: This table has the information needed to locate the literature source reporting the organism's background and it's genome. At least one literature reference is required, there can be more than one but it must only yield a single organism. The attributes include the journal name, volume, article title, date it was published, and two foreign keys. The primary key is the REF_ID it is assigned by the private organization. It has one foreign key that links it to the ORG_ID from the Organism table.

2 Organism: This table consists of specific information about the organism such as the host it was located in, the type of organism (eukaryote, prokaryote, virus, bacteria, etc.), the scientific name, and full species lineage. Each entry in the database must be based on only one organism, this is a mandatory

requirement. The primary key for this table is the ORG_ID for the specific organism. The two foreign keys link it with the Reference Literature it was sourced from and it's genome. The Organism and Genome tables must share the same ORG_ID, there will not be more than one.

3 Genome: This table consists of the full DNA sequence of the organism, the length of the sequence, and the percentage of GC base-pairs in the genome. You can also find the FASTA ID where the genome can be found in the FASTA format. The primary key is the GENOME_ID, a specific ID of letters and numbers assigned to the genome. Each organism is required to have only one genome and it must be present. The foreign keys link it with the organism it is derived from and the genes it holds.

4 Gene: This entity holds specific information for each gene found in the genome. The primary key is the GENE_ID, it also has information about the type of gene (coding or non-coding), the shorthand gene symbol, and a brief description of the gene. It also reports the last time the gene information was updated. The foreign keys link it with the Genome and Protein tables. Multiple genes can be derived from one genome but at least one gene must have been identified. Each gene must be genome and organism specific. No two different organisms or genomes will have the same GENE_ID.

5 Protein: The Protein table will be based on proteins produced from specific genes. The primary key is the PROTEIN_ID, it will be linked to the Gene table by the GENE_ID as a foreign key. It has protein specific information such as protein family (if known), the amino acids making up the protein sequence, the name of the region the protein is found, and the molecular weight of the protein. Many proteins can arise from a single gene but at least one gene is required. Some genes can have no protein identified. Each protein is gene and genome specific. Other organisms may have the same gene but they will have different GENE_IDs that will not be linked.

DDL Source Code

```
/* Analia Trevino-Flitton
DBST 651:9040
Fall 2020
Cloud Genome: DDL Script
*/

-----------------------------------------------------------------------------------------
-- Drop all objects in case they already exist
```

```
-------------------------------------------------------------------------------------------

DROP TABLE organism CASCADE CONSTRAINTS;
DROP TABLE ref_literature CASCADE CONSTRAINTS;
DROP TABLE genome CASCADE CONSTRAINTS;
DROP TABLE gene CASCADE CONSTRAINTS;
DROP TABLE protein CASCADE CONSTRAINTS;

DROP SEQUENCE seq_gen_ref;
DROP SEQUENCE seq_gen_org;
DROP SEQUENCE seq_gen_genome;
DROP SEQUENCE seq_gen_gene;
DROP SEQUENCE seq_gen_pro;

-------------------------------------------------------------------------------------------
-- Create tables for all objects with foreign key constraints
-------------------------------------------------------------------------------------------

CREATE TABLE ref_literature (
     ref_id int  NOT NULL CONSTRAINT PK_ref_id PRIMARY KEY,
     pubmed_id varchar2 (255),
     journal varchar2 (255) NOT NULL,
     journal_volume int,
     article_title varchar2 (255) NOT NULL,
     pub_date date NOT NULL,
     org_id varchar2 (255)
    );

CREATE TABLE organism (
     org_id varchar2 (255)  NOT NULL CONSTRAINT PK_org_id PRIMARY KEY,
     scientific_name varchar2 (255) NOT NULL,
     org_type varchar2 (255) NOT NULL,
     host varchar2 (255),
     lineage varchar2 (255) NOT NULL,
     ref_id int NOT NULL,
     genome_id varchar2 (255),

CONSTRAINT FK_org_ref_id
    FOREIGN KEY(ref_id)      REFERENCES ref_literature(ref_id),

CONSTRAINT FK_ref_org_id
     FOREIGN KEY(org_id) REFERENCES organism(org_id)
    );

CREATE TABLE genome (
     genome_id varchar2 (255) NOT NULL CONSTRAINT PK_genome_id PRIMARY KEY,
     fasta_id varchar2 (255),
     dna_seq varchar2 (255) NOT NULL,
     dna_length int NOT NULL,
     gc_content number,
     org_id varchar2 (255) NOT NULL,
     gene_id varchar2 (255),

CONSTRAINT FK_genome_org_id
    FOREIGN KEY(org_id)     REFERENCES organism(org_id),

CONSTRAINT FK_org_genome_id
     FOREIGN KEY(genome_id) REFERENCES genome(genome_id)
    );

CREATE TABLE gene (
     gene_id varchar2 (255)  NOT NULL CONSTRAINT PK_gene_id  PRIMARY KEY,
     gene_type varchar2 (255),
```

```
     gene_symbol varchar2 (255),
     gene_description varchar2 (255),
     last_update date,
     genome_id varchar2 (255) NOT NULL,
     protein_id varchar2 (255),

CONSTRAINT FK_gene_genome_id
    FOREIGN KEY(genome_id)      REFERENCES genome(genome_id),

CONSTRAINT FK_genome_gene_id
     FOREIGN KEY(gene_id) REFERENCES gene(gene_id)
    );

CREATE TABLE protein (
     protein_id varchar2 (255) NOT NULL CONSTRAINT  PK_protein_id PRIMARY KEY,
     pfam varchar2 (255),
     protein_seq varchar2 (255) NOT NULL,
     region_name varchar2 (255),
     mol_weight number (38),
     gene_id varchar2 (255) NOT NULL,

CONSTRAINT FK_pro_gene_id
    FOREIGN KEY(gene_id)        REFERENCES gene(gene_id),

CONSTRAINT FK_gene_protein_id
    FOREIGN KEY(protein_id) REFERENCES protein(protein_id)
    );

-------------------------------------------------------------------------------------------
-- Alter table to add audit columns
-------------------------------------------------------------------------------------------
ALTER TABLE ref_literature ADD(
     created_by varchar2 (30),
     date_created date,
     modified_by varchar2(30),
     date_modified date );

ALTER TABLE organism ADD(
     created_by varchar2 (30),
     date_created date,
     modified_by varchar2(30),
     date_modified date );

ALTER TABLE genome ADD(
     created_by varchar2 (30),
     date_created date,
     modified_by varchar2(30),
     date_modified date );

ALTER TABLE gene ADD(
     created_by varchar2 (30),
     date_created date,
     modified_by varchar2(30),
     date_modified date );

ALTER TABLE protein ADD(
     created_by varchar2 (30),
     date_created date,
     modified_by varchar2(30),
     date_modified date );

-------------------------------------------------------------------------------------------
/* Views for each table provide improved accessibility for use by specific departments
```

```
 or teams and eliminate the need to specify each individual attribute.
 Saves employees time with less coding and bug troubleshooting. */
-------------------------------------------------------------------------------------------

--Business Purpose: To provide a quick query for relevant reference literature information
CREATE OR REPLACE VIEW VW_ref_literature AS
     SELECT pubmed_id, article_title, journal, pub_date
     FROM ref_literature;

--Business Purpose: Provides a fast query to access the organism's information in a grouped
fashion
CREATE OR REPLACE VIEW VW_organism AS
     SELECT org_id, scientific_name, org_type, lineage
     FROM organism;

--Business Purpose: A simple query for many of the genome's attributes
CREATE OR REPLACE VIEW VW_genome AS
     SELECT genome_id, dna_seq, dna_length
     FROM genome;

--Business Purpose: A streamlined query to view all the gene's information at once
CREATE OR REPLACE VIEW VW_gene AS
     SELECT gene_id, gene_type, gene_symbol, gene_description
     FROM gene;

--Business Purpose: A single query to provide the protein attributes
CREATE OR REPLACE VIEW VW_protein AS
     SELECT protein_id, pfam, protein_seq, region_name
     FROM protein;

-------------------------------------------------------------------------------------------
-- Index for natural key, FK & frequently queried
-------------------------------------------------------------------------------------------

CREATE UNIQUE INDEX UX_org_sci_name ON organism(scientific_name);

CREATE UNIQUE INDEX  UX_genome_fasta_id ON genome(fasta_id);

CREATE UNIQUE INDEX  UX_pro_pro_seq ON protein(protein_seq);

CREATE UNIQUE INDEX  UX_ref_lit_article_title ON ref_literature(article_title);

CREATE UNIQUE INDEX  UX_ref_lit_pubmed_id ON ref_literature(pubmed_id);

-- Foreign Key Index
CREATE UNIQUE INDEX  UX_org_ref_id_FK ON organism(ref_id);

CREATE UNIQUE INDEX UX_ref_org_id_FK ON ref_literature(org_id);

CREATE UNIQUE INDEX UX_genome_org_id_FK ON genome(org_id);

CREATE UNIQUE INDEX UX_org_genome_id_FK ON organism(genome_id);

CREATE UNIQUE INDEX UX_gene_genome_id_FK ON gene(genome_id);

CREATE UNIQUE INDEX UX_genome_gene_id_FK ON genome(gene_id);

CREATE UNIQUE INDEX UX_gene_protein_id_FK ON gene(protein_id);

-------------------------------------------------------------------------------------------
-- Sequence generators for triggers
-------------------------------------------------------------------------------------------
```

```
CREATE SEQUENCE seq_gen_ref;

CREATE SEQUENCE seq_gen_org;

CREATE SEQUENCE seq_gen_genome;

CREATE SEQUENCE seq_gen_gene;

CREATE SEQUENCE seq_gen_pro;

-------------------------------------------------------------------------------------
-- Triggers
-------------------------------------------------------------------------------------

/* Business Purpose: To ensure each piece of reference literature has a unique corresponding
   reference ID (primary key) if one is not provided */

CREATE OR REPLACE TRIGGER ref_lit_TRG
     BEFORE INSERT OR UPDATE ON ref_literature
     FOR EACH ROW
     BEGIN
     IF :NEW.ref_id IS NULL THEN
     :NEW.ref_id := genseq_ref.NEXTVAL;
     END IF;

IF INSERTING THEN
     IF :NEW.created_by IS NULL THEN :NEW.created_by := USER; END IF;
     IF :NEW.date_created IS NULL THEN :NEW.date_created := SYSDATE; END IF;
     END IF;

IF INSERTING OR UPDATING THEN
     IF :NEW.modified_by IS NULL THEN :NEW.modified_by := USER; END IF;
     IF :NEW.date_modified IS NULL THEN :NEW.date_modified := SYSDATE; END IF;
    END IF;END;
     /

-- Business Purpose: Generates a required organism ID (primary key) if one is not listed to
ensure constraints are met

CREATE OR REPLACE TRIGGER org_TRG
     BEFORE INSERT OR UPDATE ON organism
     FOR EACH ROW
     BEGIN
     IF :NEW.org_id IS NULL THEN
     :NEW.org_id := genseq_org.NEXTVAL;
     END IF;

IF INSERTING THEN
     IF :NEW.created_by IS NULL THEN :NEW.created_by := USER; END IF;
     IF :NEW.date_created IS NULL THEN :NEW.date_created := SYSDATE; END IF;
     END IF;

IF INSERTING OR UPDATING THEN
     IF :NEW.modified_by IS NULL THEN :NEW.modified_by := USER; END IF;
     IF :NEW.date_modified IS NULL THEN :NEW.date_modified := SYSDATE; END IF;
    END IF;END;
     /

-- Business Purpose: Provides a random sequence for the gene ID ( primary key ) for the Gene
table if one is not provided

CREATE OR REPLACE TRIGGER gene_TRG
     BEFORE INSERT OR UPDATE ON gene
```

```
     FOR EACH ROW
     BEGIN
     IF :NEW.gene_id IS NULL THEN
     :NEW.gene_id := genseq_gene.NEXTVAL;
     END IF;

IF INSERTING THEN
     IF :NEW.created_by IS NULL THEN :NEW.created_by := USER; END IF;
     IF :NEW.date_created IS NULL THEN :NEW.date_created := SYSDATE; END IF;
     END IF;

IF INSERTING OR UPDATING THEN
     IF :NEW.modified_by IS NULL THEN :NEW.modified_by := USER; END IF;
     IF :NEW.date_modified IS NULL THEN :NEW.date_modified := SYSDATE; END IF;
    END IF;END;
     /

-- Business Purpose: Gives every protein a unique locator ID if one is not provided

CREATE OR REPLACE TRIGGER pro_TRG
     BEFORE INSERT OR UPDATE ON protein
     FOR EACH ROW
     BEGIN
     IF :NEW.protein_id IS NULL THEN
     :NEW.protein_id := genseq_pro.NEXTVAL;
     END IF;

IF INSERTING THEN
     IF :NEW.created_by IS NULL THEN :NEW.created_by := USER; END IF;
     IF :NEW.date_created IS NULL THEN :NEW.date_created := SYSDATE; END IF;
     END IF;

IF INSERTING OR UPDATING THEN
     IF :NEW.modified_by IS NULL THEN :NEW.modified_by := USER; END IF;
     IF :NEW.date_modified IS NULL THEN :NEW.date_modified := SYSDATE; END IF;
    END IF;END;
     /
```

DML Source Code

```
-------------------------------------------------------------------------------------------
/* Analia Trevino-Flitton
DBST 651:9040
Fall 2020
Cloud Genome: DML Script
*/


-- Populate all Tables
-------------------------------------------------------------------------------------------
-- 1

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (14755, 10482585, 'Nature', '18', 'A phylogenetically conserved hairpin-type 3
untranslated region pseudoknot functions in coronavirus RNA replication', TO_DATE('08-Oct-1999')
);

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
```

```
VALUES (78884, 15630477, 'PLoS Biol.', '3', 'The structure of a rigorously conserved RNA element
within the SARS virus genome', TO_DATE('18-Jan-2005') );

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (96657, 15680415, 'Virology', '332', 'Programmed ribosomal frameshifting in decoding the
SARS-CoV genome', TO_DATE('20-Feb-2005') );

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (74441, 32015508, 'Nature', '579', 'A new coronavirus associated with human respiratory
disease in China', TO_DATE('01-Mar-2020') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage)
VALUES (74441,'NC_045512', 'Orthocoronavirinae', 'Virus', 'Homo sapien', 'Viruses; Riboviria;
Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae;
Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related
coronavirus' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ('NC_045512.2', 'NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate
Wuhan-Hu-1, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
', 25699,  38, 'NC_045512' );

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'GU280_gp01', 'protein coding', 'ORF1ab', 'ORF1a polyprotein;ORF1ab polyprotein',
TO_DATE('04-Nov-2020'), 'NC_045512.2' );

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('YP_009725297.1', 'pfam11501', 'meslvpgfne kthvqlslpv lqvrdvlvrg fgdsveevls earqhlkdgt
cglvevekgv', 'Nsp1', 19644,'GU280_gp01' );

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('YP_009725300', 'pfam16348', 'rsdvllpltq ynrylalynk ykyfsgamdt tsyreaacch lakalndfsn',
'Corona_NSP4_C', 56184, 'GU280_gp01' );

-------------------------------------------------------------------------------
-- 2

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (59667, 26262818, 'ISME J.', '10', 'Deciphering the bat virome catalog to better
understand the ecological diversity of bat viruses and the bat origin of emerging infectious
diseases', TO_DATE('05-Mar-2016') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage)
VALUES (59667, 'NC_025217', 'Bat Hp-betacoronavirus/Zhejiang2013', 'Virus', 'Hipposideros
pratti', 'Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales;
Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Hibecovirus; Bat
Hp-betacoronavirus Zhejiang2013' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ( 'NC_025217.1', 'NC_025217.1 Bat Hp-betacoronavirus/Zhejiang2013, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
', 7325,  45, 'NC_025217' );

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'NA39_gp6','ribosomal_slippage', 'ORF1ab', 'ORF1ab polyprotein is cleaved to yield the
RNA-dependent RNA polymerase and other nonstructural proteins; polyprotein pp1ab',
TO_DATE('25-Aug-2020'), 'NC_025217.1');
```

```
INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('YP_009072438', 'pfam11501', 'kvrqlckll rgtkaltevi plteeaelel aenreilkep vhgvyydpsk
dliaeiqkqg', 'Nsp1', 17821, 'NA39_gp6' );

--------------------------------------------------------------------------------
-- 3

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (16344, 10073695, 'J Gen Virol.', '80', 'Characterization of the L gene and 5 trailer
region of Ebola virus', TO_DATE('10-Feb-1999') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage)
VALUES (16344,'NC_002549', 'Zaire ebolavirus', 'Virus', 'Homo sapien', 'Viruses; Riboviria;
Orthornavirae; Negarnaviricota; Haploviricotina; Monjiviricetes; Mononegavirales; Filoviridae;
Ebolavirus.' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ( 'NC_002549.1', 'NC_002549.1 Zaire ebolavirus isolate Ebola
virus/H.sapiens-tc/COD/1976/Yambuku-Mayinga, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
', 899,  36, 'NC_002549' );

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'ZEBOVgp1','protein coding', 'NP', 'nucleoprotein', TO_DATE('4-Jan-2020'),
'NC_002549.1');

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('NP_066243.1', 'pfam05505', 'kkekvyl awvpahkgig gneqvdklvs agirkvlfld gidkaqdeh',
'Nsp1', 83156, 'ZEBOVgp1' );

--------------------------------------------------------------------------------
-- 4

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (47781, 26862926, 'N Engl J med.', '374', 'Zika Virus Associated with Microcephaly',
TO_DATE('10-Mar-2016') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage)
VALUES (47781, 'NC_035889', 'Zika virus', 'Virus', 'Homo sapien', 'Viruses; Riboviria;
Orthornavirae; Kitrinoviricota; Flasuviricetes;Amarillovirales; Flaviviridae; Flavivirus' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ( 'NC_035889.1', 'NC_035889.1 Zika virus isolate ZIKV/H. sapiens/Brazil/Natal/2015,
complete genome', 'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG', 2777,  23, 'NC_035889'
);

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'CPG35_gp1','protein coding', 'POLY', 'polyprotein', TO_DATE('1-Aug-2020'),
'NC_035889.1');

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('YP_009428568', 'pfam01003', 'LRRVYNINGFDEVKPMALCALHYCEDCGMEMWCHSNFEEAYCPAEDKAEPGN',
'Flavi_capsid', 19096, 'CPG35_gp1' );

--------------------------------------------------------------------------------
-- 5

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
```

```
VALUES (77881, 9362478, 'EMBO J.', '16', 'Signal peptide fragments of preprolactin and HIV-1
p-gp160 interact with calmodulin', TO_DATE('17-Nov-1997') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage )
VALUES (77881,'NC_001802', 'Human immunodeficiency virus 1 (HIV-1)', 'Virus', 'Homo sapien', '
Viruses; Riboviria; Pararnavirae; Artverviricota; Revtraviricetes; Ortervirales; Retroviridae;
Orthoretrovirinae; Lentivirus' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ('NC_001802.1', 'NC_001802.1 Human immunodeficiency virus 1, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
', 8956,  38, 'NC_001802' );

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'HIV1gp1', 'protein coding', 'gag-pol', 'Gag-Pol', TO_DATE('27-Jun-2020'),
'NC_001802.1' );

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('NP_789740.1', 'pfam00077', 'qefgipy npqsqgvves mnkelkkiig qvrdqaehlk tavqmavfih
nfkrkggigg', 'RT_Rtv', 112754, 'HIV1gp1' );

--------------------------------------------------------------------------------
-- 6

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (24498, 3018124, 'N Engl J med.', '374', 'The complete DNA sequence of varicella-zoster
virus', TO_DATE('26-Sep-1986') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage)
VALUES (24498, 'NC_001348', 'Human alphaherpesvirus 3 (HHV-3)','Virus' , 'Homo sapien',
'Viruses; Duplodnaviria; Heunggongvirae; Peploviricota;Herviviricetes; Herpesvirales;
Herpesviridae; Alphaherpesvirinae; Varicellovirus.' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ( 'NC_001348.1' , 'NC_001348.1 Human herpesvirus 3, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG', 2787,  65, 'NC_001348'
);

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ('HHV3_gp01' ,'protein coding', 'ORF0', 'membrane protein UL56', TO_DATE('3-Mar-2019'),
'NC_001348.1');

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('YP_053044' , 'UL56 family' , 'hysrrp gtppvtltutcbbss psmddvatpi pylptyaeav adapppyrsr
eslvfsppl' , 'ORF0' , 39456 , 'HHV3_gp01' );

--------------------------------------------------------------------------------
-- 7

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (35512, 8805245, 'Curr Biol', '6', 'Metabolism and evolution of Haemophilus influenzae
deduced from a whole-genome comparison with Escherichia coli', TO_DATE('10-Mar-1996') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage )
VALUES (35512, 'NC_000907', 'Haemophilus influenzae', 'Bacteria', 'Homo sapien', 'Bacteria;
Proteobacteria; Gammaproteobacteria; Pasteurellales; Pasteurellaceae; Haemophilus' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
```

```
VALUES ( 'NC_000907.1', 'NC_000907.1 Haemophilus influenzae Rd KW20, complete sequence',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
', 65721,  24, 'NC_000907' );

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'aroG', 'protein coding', 'aroG', 'equivalog', TO_DATE('23-Jun-2020'), 'NC_000907.1' );

INSERT INTO protein (protein_id, protein_seq, mol_weight, gene_id)
VALUES ('NP_7849740.1', 'anddsdyytocdqvlppiallYYOOekypaseqaaalvkahniihgkddrllvvi', 38994, 'aroG'
);

--------------------------------------------------------------------------------
-- 8

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (02257, 28840828, 'Euro Surveill.', '22', 'Imported case of Middle East respiratory
syndrome coronavirus (MERS-CoV) infection from Oman to Thailand, June 2015',
TO_DATE('17-Aug-2017') );

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage )
VALUES (02257,'KT2254762', 'Middle East respiratory syndrome-related coronavirus (MERS-CoV)',
'Virus', 'Homo sapien', 'Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Merbecovirus.'
);

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ('KT225476.2', 'KT225476.2 Middle East respiratory syndrome coronavirus isolate
MERS-CoV/THA/CU/17_06_2015, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
', 21529,  42, 'KT2254762' );

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'SUD-M', 'protein coding', 'SUD-M', 'SUD-M', TO_DATE('15-Apr-2020'), 'KT225476.2' );

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('NP_789880.1', 'pfam1661', 'SVLACYNGRPYUTNTWEERBTAUADDIITGTFTDSFVVMRPNYTIKGSFLCGSCGS',
'Corona_S2', 25511, 'SUD-M' );

--------------------------------------------------------------------------------
-- 9

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES( 43699, 1326820, 'Virology', '190', 'Molecular cloning of a novel human papillomavirus
(type 60) from a plantar cyst with characteristic pathological changes', TO_DATE('01-Sep-1992')
);

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage )
VALUES (43699, 'NC_001693', 'Human papillomavirus type 60', 'Virus', 'Homo sapien', ' Viruses;
Monodnaviria; Shotokuvirae; Cossaviricota;Papovaviricetes; Zurhausenvirales;
Papillomaviridae;Firstpapillomavirinae; Gammapapillomavirus' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ( 'NC_001693.1', 'NC_001693.1 Human papillomavirus type 60, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG', 14879,  27, 'NC_001693'
);

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
```

```
VALUES ( 'E6_ght', 'protein coding', 'E6', 'transforming protein E6', TO_DATE('15-Aug-2018'),
'NC_001693.1' );

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('NP_043437', 'pfam00518', 'qmeedrfpt tvadycsefd iplkdlklkc vfcrfylteq qlaaf', 'E6',
16679, 'E6_ght' );

-------------------------------------------------------------------------------
-- 10

INSERT INTO ref_literature (ref_id, pubmed_id, journal, journal_volume, article_title, pub_date)
VALUES (89214, 2552166, 'J. Virol', '63', 'Human papillomavirus type 48', TO_DATE('12-Nov-1989')
);

INSERT INTO organism (ref_id, org_id, scientific_name, org_type, host, lineage )
VALUES (89214, 'NC_001690', 'Human papillomavirus type 48', 'Virus', 'Homo sapien', 'Viruses;
Monodnaviria; Shotokuvirae; Cossaviricota;Papovaviricetes; Zurhausenvirales;
Papillomaviridae;Firstpapillomavirinae; Gammapapillomavirus' );

INSERT INTO genome ( genome_id, fasta_id, dna_seq, dna_length, gc_content, org_id)
VALUES ('NC_001690.1', 'NC_001690.1 Human papillomavirus type 48, complete genome',
'ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG', 7100,  44, 'NC_001690'
);

INSERT INTO gene ( gene_id, gene_type, gene_symbol, gene_description, last_update, genome_id  )
VALUES ( 'E1', 'protein coding', 'E1', 'replication protein E1', TO_DATE('15-Aug-2018'),
'NC_001690.1' );

INSERT INTO protein (protein_id, pfam, protein_seq, region_name, mol_weight, gene_id)
VALUES ('NP_043418', 'pfam00524', 'qngaec elnsilrsnn iratvlckfk dkfgvsfnel', 'E1', 2744, 'E1' );

-------------------------------------------------------------------------------
-- All Data Dictionary
-------------------------------------------------------------------------------
SELECT TABLE_NAME FROM USER_TABLES;
SELECT OBJECT_NAME, STATUS, CREATED, LAST_DDL_TIME FROM USER_OBJECTS;
```

Query Source Code

```
/* Analia Trevino-Flitton
DBST 651:9040
Fall 2020
Cloud Genome: 20 SQL Statements- 8 Advanced Queries
*/
-----------------------------------------------------------------------------------------
/* Query 1: Select all columns and all rows from one table

   Business Purpose: This selects all the row information from the gene table. */
-----------------------------------------------------------------------------------------
SELECT *
FROM
     gene;

-----------------------------------------------------------------------------------------
/* Query 2: Select 5 columns and all rows from one table.

   Business Purpose: This provides information about all organisms currently in the
   database. */
-----------------------------------------------------------------------------------------
```

```
SELECT
     org_id, scientific_name, org_type, host, lineage
FROM
     organism;

-------------------------------------------------------------------------------------------
/* Query 3: Select all columns and all rows from one view.

   Business Purpose: This shows all the gene information in the gene view, it is a faster
   query than selecting specific gene information. */
-------------------------------------------------------------------------------------------
SELECT *
FROM
     VW_gene;

-------------------------------------------------------------------------------------------
/* Query 4: Using a join on 2 tables, select all columns and all rows from the tables
   without the use of a Cartesian product.

   Business Purpose: Joins the protein and gene tables. */
-------------------------------------------------------------------------------------------
SELECT *
FROM
     gene
LEFT JOIN protein ON gene.gene_id = protein.gene_id;

-------------------------------------------------------------------------------------------
/* Query 5: Select and order data retrieved from one table.

   Business Purpose: lists the proteins in the protein table in order of the highest
   molecular weight to the lowest. */
-------------------------------------------------------------------------------------------
SELECT *
FROM
     protein
ORDER BY
      mol_weight DESC;

-------------------------------------------------------------------------------------------
/* Query 6: Using a join on 3 tables, select 5 columns from the 3 tables. Use syntax that
   would limit the output to 10 rows.

   Business Purpose: This selects the an organism's org ID, scientific name, ref ID, the
   genome ID and it's genome length. */
-------------------------------------------------------------------------------------------
SELECT
     rl.ref_id,
     o.org_id, o.scientific_name,
     gm.genome_id, gm.dna_length
FROM
     ref_literature rl
JOIN organism o    ON o.ref_id = rl.ref_id
JOIN genome gm     ON gm.org_id = o.org_id
WHERE
     ROWNUM <= 10;

-------------------------------------------------------------------------------------------
/* Query 7: Select distinct rows using joins on 3 tables.

   Business Purpose: This selects distinctly different values from the gene symbol, protein
   family, protein ID, and the genome's fasta ID. */
-------------------------------------------------------------------------------------------
SELECT DISTINCT
```

```
     g.gene_symbol, p.pfam, p.protein_id, gm.fasta_id
FROM
     genome gm
JOIN gene g ON g.genome_id =  gm.genome_id
JOIN protein p ON p.gene_id = g.gene_id;

-------------------------------------------------------------------------------------------
/* Query 8: Use group by & having in a select statement using one or more tables.

   Business Purpose: Lists gene symbol, gene ID, protein associated with gene, protein
   family, and orders by lightest to heaviest molecular weight. */
-------------------------------------------------------------------------------------------
SELECT
     g.gene_symbol, g.gene_id, p.protein_id,p.pfam, p.mol_weight
FROM
     protein p JOIN gene g ON p.gene_id = g.gene_id
GROUP BY
     g.gene_symbol, g.gene_id, p.protein_id,p.pfam, p.mol_weight
HAVING
     p.mol_weight >= 2000
ORDER BY
      mol_weight ASC;

-------------------------------------------------------------------------------------------

/* Query 9: Use IN clause to select data from one or more tables.

   Business Purpose: Shows the reference ID, scientific name and organism host for organism's
   that have been published in the journals Nature and Virology. */
-------------------------------------------------------------------------------------------
SELECT
     rl.ref_id, rl.journal, o.scientific_name, o.host
FROM
     ref_literature rl
JOIN organism o ON o.ref_id = rl.ref_id

WHERE journal IN ('Nature', 'Virology');

-------------------------------------------------------------------------------------------
/* Query 10: Select Length of one column from one table (use Length function)

   Business Purpose: Shows the length of the journal titles */
-------------------------------------------------------------------------------------------
SELECT
     LENGTH(journal) AS "Journal Length", journal
FROM
     ref_literature;

-------------------------------------------------------------------------------------------

/* Query 11: use the SQL DELETE statement to delete one record from one table

  Business Purpose: This deletes the protein family with the value of pfam01003 */
-------------------------------------------------------------------------------------------
SELECT pfam FROM protein;
DELETE FROM
     protein
WHERE
     pfam = 'pfam01003';
SELECT pfam FROM protein;
COMMIT;
ROLLBACK;
```

```
-------------------------------------------------------------------------------------------
/* Query 12: use the SQL UPDATE statement to change some data

   Business Purpose: This updates all the organism's type to Prokaryote if data
   reclassification was to occur. */
-------------------------------------------------------------------------------------------
SELECT org_type FROM organism;
UPDATE
     organism
SET
     org_type = 'Prokaryote';

SELECT org_type FROM organism;
COMMIT;

ROLLBACK;

-------------------------------------------------------------------------------------------
-- 8 Advanced Queries

/* Query 13: Determine the count for literature published in 2020

   Business Purpose: This shows the most recent literature from the past year and displays
   the date it was published, the journal name, the article title, scientific name of the
   organism, the organism type and the count of references published in 2020. */
-------------------------------------------------------------------------------------------
SELECT
    rl.pub_date,  rl.journal, rl.article_title, o.scientific_name, o.org_type,

( SELECT COUNT(pub_date) FROM ref_literature
     WHERE pub_date > date '2020-01-01') AS "Journals Published IN 2020"

FROM
     ref_literature rl
JOIN ORGANISM o ON rl.ref_id = o.ref_id
WHERE pub_date > date '2020-01-01';

-------------------------------------------------------------------------------------------
/* Query 14: Display the molecular weights of proteins found in the Nsp1 region in
   ascending order

   Business Purpose: This shows the FASTA ID, gene ID, protein ID, protein family and
   molecular weight of proteins found in the Nsp1 region from lightest to heaviest. Protein
   molecular weight can be important when determining whether a property would be a good
   therapeutic candidate. */
-------------------------------------------------------------------------------------------
SELECT
    gm.fasta_id, g.gene_id, p.protein_id, p.pfam, P.mol_weight AS "Average Protein Weight in
Nsp1 Region"
FROM
    protein p
JOIN gene g ON g.gene_id = p.gene_id
INNER JOIN genome gm ON gm.genome_id = g.genome_id
WHERE region_name = 'Nsp1'
ORDER BY mol_weight ASC ;

-------------------------------------------------------------------------------------------

/* Query 15: List the gene symbols and descriptions of those genes updated before 2020

   Business Purpose: Shows the gene ID, gene symbol, gene description, FASTA ID for the
   genome, the protein family and protein ID of the genes that have been updated before
   2020. */
```

```
-------------------------------------------------------------------------------------------
SELECT
     g.last_update, g.gene_id, g.gene_symbol, g.gene_description,
     gm.fasta_id, p.protein_id, p.pfam
FROM
     gene g
JOIN genome gm     ON g.genome_id = gm.genome_id
INNER JOIN protein p  ON g.gene_id = p.gene_id
WHERE g.last_update < date '2020-01-01';

-------------------------------------------------------------------------------------------

/* Query 16: List the scientific name, DNA length, GC content, gene symbol and  protein
   family of the organisms with a protein coding gene type, where the DNA length is at least
   3000 and the GC content is no greater than 40%. Order by DNA length DESC/

  Business Purpose: This lists the organism's scientific name, the length of it's genome,
  the GC content, the gene symbol and protein family associated with the organism.   */
-------------------------------------------------------------------------------------------
SELECT o.scientific_name, gm.dna_length,gm.gc_content, g.gene_symbol, p.pfam
FROM
     organism o
LEFT JOIN genome gm
    ON gm.org_id = o.org_id
JOIN  gene g
     ON g.genome_id = gm.genome_id
JOIN protein p
     ON p.gene_id = g.gene_id
WHERE g.gene_type = 'protein coding'
    AND gm.dna_length > 3000
    AND gm.gc_content < 40
ORDER BY gm.dna_length DESC;

-------------------------------------------------------------------------------------------

/* Query 17:  Display the scientific name along with the average GC content and those with
   above average content.

   Business Purpose: High GC content has been correlated with the development of cancer in
   certain genes. We are listing organism's with abnormally high GC content. */
-------------------------------------------------------------------------------------------
 SELECT
     o.scientific_name, gm.gc_content, a.avg_content
FROM
     organism o
JOIN genome gm ON gm.org_id = o.org_id ,
     (SELECT  AVG( gc_content) as avg_content FROM genome) a
WHERE gm.gc_content > a.avg_content;

-------------------------------------------------------------------------------------------
/* Query 18:  List the gene that produces more than one protein.

   Business Purpose: This shows the scientific name of the organism, the gene symbol,
   protein ID, molecular weight of the proteins, and the protein family they belong to. */
-------------------------------------------------------------------------------------------
SELECT
 o.scientific_name, g.gene_symbol, p.protein_id, p.pfam, p.region_name, p.mol_weight
FROM
     gene g
JOIN protein p ON p.gene_id = g.gene_id
JOIN genome gm ON g.genome_id = gm.genome_id
JOIN organism o ON gm.org_id = o.org_id
```

```
WHERE g.gene_id = (SELECT G.GENE_ID
                    FROM gene g JOIN protein p
                    ON p.gene_id = g.gene_id
                    GROUP BY g.gene_id
                    HAVING COUNT(*) >1);

-------------------------------------------------------------------------------------------
/* Query 19: Find organisms and their proteins that share similar properties.

   Business Purpose: These proteins share the same protein families, gene symbols, and
   protein region names but they are from different genomes and not the same.  */
-------------------------------------------------------------------------------------------
SELECT
    g.genome_id, o.scientific_name, p.protein_id, p.mol_weight
FROM
     gene g
JOIN protein p ON p.gene_id = g.gene_id
JOIN genome gm ON g.genome_id = gm.genome_id
JOIN organism o ON gm.org_id = o.org_id

WHERE g.gene_symbol = (SELECT g.gene_symbol
                    FROM gene g JOIN protein p
                    ON p.gene_id = g.gene_id
                    GROUP BY g.gene_symbol
                    HAVING COUNT(*) >1)
AND
p.region_name = (SELECT  p.region_name
                    FROM  protein p
                    GROUP BY  p.region_name
                    HAVING COUNT(*) >1 )
AND
p.pfam = (SELECT  p.pfam
                    FROM  protein p
                    GROUP BY  p.pfam
                    HAVING COUNT(*) >1 ) ;

-------------------------------------------------------------------------------------------
/* Query 20: List information from all five tables where the host is 'Homo sapien' and the
   protein family is known but there are no repeated values.

   Business Purpose: Shows the reference article title, scientific name of the organism, the
   fasta ID, gene symbol associated with the genome and protein family of an organism found
   in Homo sapiens and with a known protein family. */
-------------------------------------------------------------------------------------------
SELECT DISTINCT
     rl.article_title, o.scientific_name, gm.fasta_id, g.gene_symbol, p.pfam

FROM
     ref_literature rl
JOIN organism o
     ON rl.ref_id = o.ref_id
JOIN genome gm
     ON gm.org_id = o.org_id
JOIN gene g
     ON g.genome_id = gm.genome_id
JOIN protein p
     ON p.gene_id = g.gene_id
WHERE o.host = 'Homo sapien'
AND p.pfam IS NOT NULL;
```

DDL Output

```
SQL> /* Analia Trevino-Flitton
SQL>DBST 651:9040
SQL>Fall 2020
SQL>Cloud Genome: DDL Script
SQL>*/
SQL>
SQL> SET ECHO OFF

Table ORGANISM dropped.

Table REF_LITERATURE dropped.

Table GENOME dropped.

Table GENE dropped.

Table PROTEIN dropped.

Sequence SEQ_GEN_REF dropped.

Sequence SEQ_GEN_ORG dropped.

Sequence SEQ_GEN_GENOME dropped.

Sequence SEQ_GEN_GENE dropped.

Sequence SEQ_GEN_PRO dropped.

Table REF_LITERATURE created.

Table ORGANISM created.

Table GENOME created.

Table GENE created.

Table PROTEIN created.

Table REF_LITERATURE altered.

Table ORGANISM altered.

Table GENOME altered.

Table GENE altered.

Table PROTEIN altered.

View VW_REF_LITERATURE created.

View VW_ORGANISM created.

View VW_GENOME created.

View VW_GENE created.

View VW_PROTEIN created.

INDEX UX_ORG_SCI_NAME created.

INDEX UX_GENOME_FASTA_ID created.
```

```
INDEX UX_PRO_PRO_SEQ created.

INDEX UX_REF_LIT_ARTICLE_TITLE created.

INDEX UX_REF_LIT_PUBMED_ID created.

INDEX UX_ORG_REF_ID_FK created.

INDEX UX_REF_ORG_ID_FK created.

INDEX UX_GENOME_ORG_ID_FK created.

INDEX UX_ORG_GENOME_ID_FK created.

INDEX UX_GENE_GENOME_ID_FK created.

INDEX UX_GENOME_GENE_ID_FK created.

INDEX UX_GENE_PROTEIN_ID_FK created.

Sequence SEQ_GEN_REF created.

Sequence SEQ_GEN_ORG created.

Sequence SEQ_GEN_GENOME created.

Sequence SEQ_GEN_GENE created.

Sequence SEQ_GEN_PRO created.

Trigger REF_LIT_TRG compiled

Trigger ORG_TRG compiled

Trigger GENE_TRG compiled

Trigger PRO_TRG compiled
```

DML Output

```
1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.
```

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

```
1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

1 row inserted.

TABLE_NAME
--------------------------------------------------------------------------------------------------
--------------------------------
GRADE
GRADE_TYPE
INSTRUCTOR
GRADE_CONVERSION
GRADE_TYPE_WEIGHT
SECTION
COURSE
ENROLLMENT
STUDENT
ZIPCODE
REF_LITERATURE

TABLE_NAME
--------------------------------------------------------------------------------------------------
--------------------------------
ORGANISM
GENOME
GENE
PROTEIN

15 rows selected.


OBJECT_NAME
STATUS  CREATED   LAST_DDL_
--------------------------------------------------------------------------------------------------
-------------------------------- ------- --------- ---------
GRADE
VALID   30-NOV-19 30-NOV-19
GRADE_TYPE
VALID   30-NOV-19 30-NOV-19
INSTRUCTOR
VALID   30-NOV-19 30-NOV-19
```

```
GRADE_CONVERSION
VALID   30-NOV-19 30-NOV-19
GRADE_TYPE_WEIGHT
VALID   30-NOV-19 30-NOV-19
SECTION
VALID   30-NOV-19 30-NOV-19
COURSE
VALID   30-NOV-19 30-NOV-19
ENROLLMENT
VALID   30-NOV-19 30-NOV-19
STUDENT
VALID   30-NOV-19 30-NOV-19
ZIPCODE
VALID   30-NOV-19 30-NOV-19
INST_ZIP_FK_I
VALID   30-NOV-19 30-NOV-19

OBJECT_NAME
STATUS  CREATED   LAST_DDL_
--------------------------------------------------------------------------------------------------------
-------------------------------- ------- --------- ---------
GR_GRTW_FK_I
VALID   30-NOV-19 30-NOV-19
GRTW_GRTYP_FK_I
VALID   30-NOV-19 30-NOV-19
SECT_CRSE_FK_I
VALID   30-NOV-19 30-NOV-19
SECT_INST_FK_I
VALID   30-NOV-19 30-NOV-19
CRSE_CRSE_FK_I
VALID   30-NOV-19 30-NOV-19
ENR_SECT_FK_I
VALID   30-NOV-19 30-NOV-19
STU_ZIP_FK_I
VALID   30-NOV-19 30-NOV-19
INST_PK
VALID   30-NOV-19 30-NOV-19
GR_PK
VALID   30-NOV-19 30-NOV-19
GRTYP_PK
VALID   30-NOV-19 30-NOV-19
GRCON_PK
VALID   30-NOV-19 30-NOV-19

OBJECT_NAME
STATUS  CREATED   LAST_DDL_
--------------------------------------------------------------------------------------------------------
-------------------------------- ------- --------- ---------
GRTW_PK
VALID   30-NOV-19 30-NOV-19
SECT_PK
VALID   30-NOV-19 30-NOV-19
CRSE_PK
VALID   30-NOV-19 30-NOV-19
ENR_PK
VALID   30-NOV-19 30-NOV-19
STU_PK
VALID   30-NOV-19 30-NOV-19
ZIP_PK
VALID   30-NOV-19 30-NOV-19
SECT_SECT2_UK
VALID   30-NOV-19 30-NOV-19
```

```
INSTRUCTOR_ID_SEQ
VALID   30-NOV-19 30-NOV-19
SECTION_ID_SEQ
VALID   30-NOV-19 30-NOV-19
STUDENT_ID_SEQ
VALID   30-NOV-19 30-NOV-19
COURSE_NO_SEQ
VALID   30-NOV-19 30-NOV-19

OBJECT_NAME
STATUS  CREATED   LAST_DDL_
------------------------------------------------------------------------------------------------------------
------------------------------- ------- --------- ---------
VW_REF_LITERATURE
VALID   12-OCT-20 03-NOV-20
VW_ORGANISM
VALID   12-OCT-20 03-NOV-20
VW_GENOME
VALID   12-OCT-20 03-NOV-20
VW_GENE
VALID   12-OCT-20 03-NOV-20
VW_PROTEIN
VALID   12-OCT-20 03-NOV-20
SEQ_REF_LIT_ID
VALID   12-OCT-20 12-OCT-20
SEQ_ORG_ID
VALID   12-OCT-20 12-OCT-20
SEQ_GENOME_ID
VALID   12-OCT-20 12-OCT-20
SEQ_GENE_ID
VALID   12-OCT-20 12-OCT-20
SEQ_PRO_ID
VALID   12-OCT-20 12-OCT-20
SEQ_GEN_ID
VALID   12-OCT-20 12-OCT-20

OBJECT_NAME
STATUS  CREATED   LAST_DDL_
------------------------------------------------------------------------------------------------------------
------------------------------- ------- --------- ---------
SEQ_REF_LIT_SUR
VALID   12-OCT-20 12-OCT-20
SEQ_ORG_SUR
VALID   12-OCT-20 12-OCT-20
SEQ_GENOME_SUR
VALID   12-OCT-20 12-OCT-20
SEQ_GENE_SUR
VALID   12-OCT-20 12-OCT-20
SEQ_PRO_SUR
VALID   12-OCT-20 12-OCT-20
SEQ_SUR_REF_LIT
VALID   12-OCT-20 12-OCT-20
SEQ_SUR_ORG
VALID   12-OCT-20 12-OCT-20
SEQ_SUR_GENOME
VALID   12-OCT-20 12-OCT-20
SEQ_SUR_GENE
VALID   12-OCT-20 12-OCT-20
SEQ_SUR_PRO
VALID   12-OCT-20 12-OCT-20
GSEQ_REF_LIT
VALID   12-OCT-20 12-OCT-20
```

```
OBJECT_NAME
STATUS  CREATED   LAST_DDL_
------------------------------------------------------------------------------------------------------------
------------------------------- ------- --------- ---------
GSEQ_ORG
VALID   12-OCT-20 12-OCT-20
GSEQ_GENOME
VALID   12-OCT-20 12-OCT-20
GSEQ_GENE
VALID   12-OCT-20 12-OCT-20
GSEQ_PRO
VALID   12-OCT-20 12-OCT-20
GENSEQ_REF
VALID   12-OCT-20 12-OCT-20
GENSEQ_ORG
VALID   12-OCT-20 12-OCT-20
GENSEQ_GENOME
VALID   12-OCT-20 12-OCT-20
GENSEQ_GENE
VALID   12-OCT-20 12-OCT-20
GENSEQ_PRO
VALID   12-OCT-20 12-OCT-20
SEQ_GEN_GENOME
VALID   03-NOV-20 03-NOV-20
SEQ_GEN_ORG
VALID   03-NOV-20 03-NOV-20

OBJECT_NAME
STATUS  CREATED   LAST_DDL_
------------------------------------------------------------------------------------------------------------
------------------------------- ------- --------- ---------
SEQ_GEN_GENE
VALID   03-NOV-20 03-NOV-20
SEQ_GEN_PRO
VALID   03-NOV-20 03-NOV-20
ORG_TRG
VALID   03-NOV-20 03-NOV-20
GENE_TRG
VALID   03-NOV-20 03-NOV-20
REF_LITERATURE
VALID   03-NOV-20 03-NOV-20
PK_REF_ID
VALID   03-NOV-20 03-NOV-20
ORGANISM
VALID   03-NOV-20 03-NOV-20
PK_ORG_ID
VALID   03-NOV-20 03-NOV-20
GENOME
VALID   03-NOV-20 03-NOV-20
REF_LIT_TRG
VALID   03-NOV-20 03-NOV-20
PRO_TRG
VALID   03-NOV-20 03-NOV-20

OBJECT_NAME
STATUS  CREATED   LAST_DDL_
------------------------------------------------------------------------------------------------------------
------------------------------- ------- --------- ---------
PK_GENOME_ID
VALID   03-NOV-20 03-NOV-20
GENE
VALID   03-NOV-20 03-NOV-20
```

```
PK_GENE_ID
VALID   03-NOV-20 03-NOV-20
PROTEIN
VALID   03-NOV-20 03-NOV-20
PK_PROTEIN_ID
VALID   03-NOV-20 03-NOV-20
UX_ORG_SCI_NAME
VALID   03-NOV-20 03-NOV-20
UX_GENOME_FASTA_ID
VALID   03-NOV-20 03-NOV-20
UX_PRO_PRO_SEQ
VALID   03-NOV-20 03-NOV-20
UX_REF_LIT_ARTICLE_TITLE
VALID   03-NOV-20 03-NOV-20
UX_REF_LIT_PUBMED_ID
VALID   03-NOV-20 03-NOV-20
UX_ORG_REF_ID_FK
VALID   03-NOV-20 03-NOV-20

OBJECT_NAME
STATUS  CREATED   LAST_DDL_
--------------------------------------------------------------------------------------------------
-------------------------------- ------- --------- ---------
UX_REF_ORG_ID_FK
VALID   03-NOV-20 03-NOV-20
UX_GENOME_ORG_ID_FK
VALID   03-NOV-20 03-NOV-20
UX_ORG_GENOME_ID_FK
VALID   03-NOV-20 03-NOV-20
UX_GENE_GENOME_ID_FK
VALID   03-NOV-20 03-NOV-20
UX_GENOME_GENE_ID_FK
VALID   03-NOV-20 03-NOV-20
UX_GENE_PROTEIN_ID_FK
VALID   03-NOV-20 03-NOV-20
SEQ_GEN_REF
VALID   03-NOV-20 03-NOV-20

95 rows selected.
```

Query Output

```
GENE_ID
GENE_TYPE
GENE_SYMBOL
GENE_DESCRIPTION
LAST_UPDA GENOME_ID
PROTEIN_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------
--------------------------------------------------------------------------------
```


-- ---------

------------------------------ --------- ------------------------------ ---------
GU280_gp01
protein coding
ORF1ab
ORF1a polyprotein;ORF1ab polyprotein
04-NOV-20 NC_045512.2
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
NA39_gp6
ribosomal_slippage
ORF1ab
ORF1ab polyprotein is cleaved to yield the RNA-dependent RNA polymerase and other nonstructural proteins; polyprotein pp1ab
25-AUG-20 NC_025217.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
ZEBOVgp1
protein coding
NP
nucleoprotein
04-JAN-20 NC_002549.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
CPG35_gp1
protein coding
POLY
polyprotein
01-AUG-20 NC_035889.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
HIV1gp1
protein coding
gag-pol
Gag-Pol
27-JUN-20 NC_001802.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
HHV3_gp01
protein coding
ORF0
membrane protein UL56
03-MAR-19 NC_001348.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
aroG
protein coding
aroG
equivalog
23-JUN-20 NC_000907.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

SUD-M
protein coding
SUD-M
SUD-M
15-APR-20 KT225476.2
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
E6_ght
protein coding
E6
transforming protein E6
15-AUG-18 NC_001693.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
E1
protein coding
E1
replication protein E1
15-AUG-18 NC_001690.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

10 rows selected.

ORG_ID
SCIENTIFIC_NAME
ORG_TYPE
HOST
LINEAGE

NC_045512
Orthocoronavirinae
Virus
Homo sapien
Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales;
Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe
acute respiratory syndrome-related coronavirus
NC_025217
Bat Hp-betacoronavirus/Zhejiang2013
Virus
Hipposideros pratti
Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales;

Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Hibecovirus; Bat Hp-betacoronavirus Zhejiang2013
NC_002549
Zaire ebolavirus
Virus
Homo sapien
Viruses; Riboviria; Orthornavirae; Negarnaviricota; Haploviricotina; Monjiviricetes; Mononegavirales; Filoviridae; Ebolavirus.
NC_035889
Zika virus
Virus
Homo sapien
Viruses; Riboviria; Orthornavirae; Kitrinoviricota; Flasuviricetes;Amarillovirales; Flaviviridae; Flavivirus
NC_001802
Human immunodeficiency virus 1 (HIV-1)
Virus
Homo sapien
Viruses; Riboviria; Pararnavirae; Artververiricota; Revtraviricetes; Ortervirales; Retroviridae; Orthoretrovirinae; Lentivirus
NC_001348
Human alphaherpesvirus 3 (HHV-3)
Virus
Homo sapien
Viruses; Duplodnaviria; Heunggongvirae; Peploviricota;Herviviricetes; Herpesvirales; Herpesviridae; Alphaherpesvirinae; Varicellovirus.
NC_000907
Haemophilus influenzae
Bacteria
Homo sapien
Bacteria; Proteobacteria; Gammaproteobacteria; Pasteurellales; Pasteurellaceae; Haemophilus
KT2254762
Middle East respiratory syndrome-related coronavirus (MERS-CoV)
Virus
Homo sapien
Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Merbecovirus.
NC_001693
Human papillomavirus type 60
Virus
Homo sapien
Viruses; Monodnaviria; Shotokuvirae; Cossaviricota;Papovaviricetes; Zurhausenvirales; Papillomaviridae;Firstpapillomavirinae; Gammapapillomavirus
NC_001690
Human papillomavirus type 48
Virus
Homo sapien
Viruses; Monodnaviria; Shotokuvirae; Cossaviricota;Papovaviricetes; Zurhausenvirales; Papillomaviridae;Firstpapillomavirinae; Gammapapillomavirus

10 rows selected.

GENE_ID
GENE_TYPE

GENE_SYMBOL
GENE_DESCRIPTION

GU280_gp01
protein coding
ORF1ab
ORF1a polyprotein;ORF1ab polyprotein
NA39_gp6
ribosomal_slippage
ORF1ab
ORF1ab polyprotein is cleaved to yield the RNA-dependent RNA polymerase and other nonstructural proteins; polyprotein pp1ab
ZEBOVgp1
protein coding
NP
nucleoprotein
CPG35_gp1
protein coding
POLY
polyprotein
HIV1gp1
protein coding
gag-pol
Gag-Pol
HHV3_gp01
protein coding
ORF0
membrane protein UL56
aroG
protein coding
aroG
equivalog
SUD-M
protein coding
SUD-M
SUD-M
E6_ght
protein coding
E6
transforming protein E6
E1
protein coding
E1
replication protein E1

```
10 rows selected.


GENE_ID
GENE_TYPE
GENE_SYMBOL
GENE_DESCRIPTION
LAST_UPDA GENOME_ID
PROTEIN_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
------------------------------------------------------------------------- ---------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------
PROTEIN_ID
PFAM
PROTEIN_SEQ
REGION_NAME
MOL_WEIGHT GENE_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
------------------------------------------------------------------------- ----------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------
GU280_gp01
protein coding
ORF1ab
```

```
ORF1a polyprotein;ORF1ab polyprotein
04-NOV-20 NC_045512.2
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

YP_009725297.1 pfam11501

meslvpgfne kthvqlslpv lqvrdvlvrg fgdsveevls earqhlkdgt cglvevekgv Nsp1
19644 GU280_gp01
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

GU280_gp01
protein coding
ORF1ab
ORF1a polyprotein;ORF1ab polyprotein
04-NOV-20 NC_045512.2
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

YP_009725300 pfam16348

rsdvllpltq ynrylalynk ykyfsgamdt tsyreaacch lakalndfsn Corona_NSP4_C
56184 GU280_gp01
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

NA39_gp6
ribosomal_slippage
ORF1ab
ORF1ab polyprotein is cleaved to yield the RNA-dependent RNA polymerase and other
nonstructural proteins; polyprotein pp1ab
25-AUG-20 NC_025217.1
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

GENE_ID
GENE_TYPE
GENE_SYMBOL
GENE_DESCRIPTION
LAST_UPDA GENOME_ID
PROTEIN_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------ ---------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------
PROTEIN_ID
PFAM
PROTEIN_SEQ
REGION_NAME
MOL_WEIGHT GENE_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------ ----------
------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------

YP_009072438 pfam11501

kvrqlckll rgtkaltevi plteeaelel aenreilkep vhgvyydpsk dliaeiqkqg Nsp1
17821 NA39_gp6
DBST_USER                      03-NOV-20 DBST_USER                    03-NOV-20

ZEBOVgp1
protein coding
NP
nucleoprotein
04-JAN-20 NC_002549.1
DBST_USER                      03-NOV-20 DBST_USER                    03-NOV-20

NP_066243.1 pfam05505

kkekvyl awvpahkgig gneqvdklvs agirkvlfld gidkaqdeh Nsp1
83156 ZEBOVgp1
DBST_USER                      03-NOV-20 DBST_USER                    03-NOV-20

CPG35_gp1
protein coding
POLY
polyprotein
01-AUG-20 NC_035889.1
DBST_USER                      03-NOV-20 DBST_USER                    03-NOV-20

YP_009428568 pfam01003

GENE_ID
GENE_TYPE
```

```
GENE_SYMBOL
GENE_DESCRIPTION
LAST_UPDA GENOME_ID
PROTEIN_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
------------------------------------------------------------------------- ---------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------
PROTEIN_ID
PFAM
PROTEIN_SEQ
REGION_NAME
MOL_WEIGHT GENE_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
------------------------------------------------------------------------- ----------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------

LRRVYNINGFDEVKPMALCALHYCEDCGMEMWCHSNFEEAYCPAEDKAEPGN Flavi_capsid
19096 CPG35_gp1
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

HIV1gp1
protein coding
gag-pol
Gag-Pol
```

27-JUN-20 NC_001802.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

NP_789740.1 pfam00077

qefgipy npqsqggvves mnkelkkiig qvrdqaehlk tavqmavfih nfkrkggigg RT_Rtv
112754 HIV1gp1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

HHV3_gp01
protein coding
ORF0
membrane protein UL56
03-MAR-19 NC_001348.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

YP_053044 UL56 family

hysrrp gtppvtltutcbbss psmddvatpi pylptyaeav adapppyrsr eslvfsppl ORF0
39456 HHV3_gp01
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

GENE_ID
GENE_TYPE
GENE_SYMBOL
GENE_DESCRIPTION
LAST_UPDA GENOME_ID
PROTEIN_ID
CREATED_BY DATE_CREA MODIFIED_BY DATE_MODI

--- ---

--- --------- ------------------------------ ---------
PROTEIN_ID
PFAM
PROTEIN_SEQ
REGION_NAME
MOL_WEIGHT GENE_ID
CREATED_BY DATE_CREA MODIFIED_BY DATE_MODI

```
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
------------------------------------------------------------------------- ----------
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------

aroG
protein coding
aroG
equivalog
23-JUN-20 NC_000907.1
DBST_USER                      03-NOV-20 DBST_USER                          03-NOV-20

NP_7849740.1

anddsdyytocdqvlppiallYYOOekypaseqaaalvkahniihgkddrllvvi
38994 aroG
DBST_USER                      03-NOV-20 DBST_USER                          03-NOV-20

SUD-M
protein coding
SUD-M
SUD-M
15-APR-20 KT225476.2
DBST_USER                      03-NOV-20 DBST_USER                          03-NOV-20

NP_789880.1 pfam1661

SVLACYNGRPYUTNTWEERBTAUADDIITGTFTDSFVVMRPNYTIKGSFLCGSCGS Corona_S2
25511 SUD-M
DBST_USER                      03-NOV-20 DBST_USER                          03-NOV-20


GENE_ID
GENE_TYPE
GENE_SYMBOL
GENE_DESCRIPTION
LAST_UPDA GENOME_ID
PROTEIN_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                        DATE_MODI
--------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------ ---------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------
PROTEIN_ID
PFAM
PROTEIN_SEQ
REGION_NAME
MOL_WEIGHT GENE_ID
CREATED_BY                     DATE_CREA MODIFIED_BY                    DATE_MODI
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------ ----------
------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
-------------------------------------------------------------------
------------------------------ --------- ------------------------------ ---------
E6_ght
protein coding
E6
transforming protein E6
15-AUG-18 NC_001693.1
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

NP_043437 pfam00518

qmeedrfpt tvadycsefd iplkdlklkc vfcrfylteq qlaaf E6
16679 E6_ght
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

E1
protein coding
E1
replication protein E1
```

15-AUG-18 NC_001690.1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

NP_043418 pfam00524

qngaec elnsilrsnn iratvlckfk dkfgvsfnel E1
2744 E1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20

11 rows selected.

PROTEIN_ID
PFAM
PROTEIN_SEQ
REGION_NAME
MOL_WEIGHT GENE_ID
CREATED_BY DATE_CREA MODIFIED_BY DATE_MODI

NP_789740.1
pfam00077
qefgipy npqsqgvves mnkelkkiig qvrdqaehlk tavqmavfih nfkrkggigg
RT_Rtv
112754 HIV1gp1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
NP_066243.1
pfam05505
kkekvyl awvpahkgig gneqvdklvs agirkvlfld gidkaqdeh
Nsp1
83156 ZEBOVgp1
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
YP_009725300
pfam16348
rsdvllpltq ynrylalynk ykyfsgamdt tsyreaacch lakalndfsn
Corona_NSP4_C
56184 GU280_gp01
DBST_USER 03-NOV-20 DBST_USER 03-NOV-20
YP_053044
UL56 family

```
hysrrp gtppvtltutcbbss psmddvatpi pylptyaeav adapppyrsr eslvfsppl
ORF0
39456 HHV3_gp01
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20
NP_7849740.1
anddsdyytocdqvlppiallYYOOekypaseqaaalvkahniihgkddrllvvi
38994 aroG
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20
NP_789880.1
pfam1661
SVLACYNGRPYUTNTWEERBTAUADDIITGTFTDSFVVMRPNYTIKGSFLCGSCGS
Corona_S2
25511 SUD-M
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20
YP_009725297.1
pfam11501
meslvpgfne kthvqlslpv lqvrdvlvrg fgdsveevls earqhlkdgt cglvevekgv
Nsp1
19644 GU280_gp01
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20
YP_009428568
pfam01003
LRRVYNINGFDEVKPMALCALHYCEDCGMEMWCHSNFEEAYCPAEDKAEPGN
Flavi_capsid
19096 CPG35_gp1
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20
YP_009072438
pfam11501
kvrqlckll rgtkaltevi plteeaelel aenreilkep vhgvyydpsk dliaeiqkqg
Nsp1
17821 NA39_gp6
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20
NP_043437
pfam00518
qmeedrfpt tvadycsefd iplkdlklkc vfcrfylteq qlaaf
E6
16679 E6_ght
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20
NP_043418
pfam00524
qngaec elnsilrsnn iratvlckfk dkfgvsfnel
E1
2744 E1
DBST_USER                      03-NOV-20 DBST_USER                      03-NOV-20

11 rows selected.


    REF_ID ORG_ID
SCIENTIFIC_NAME
GENOME_ID
DNA_LENGTH
----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------
---------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------
--------------------------------------------------------------------------
---------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------
-------------------------------------------------------------------------- ----------
     74441 NC_045512
Orthocoronavirinae
NC_045512.2
25699
     59667 NC_025217
Bat Hp-betacoronavirus/Zhejiang2013
NC_025217.1
7325
     16344 NC_002549
Zaire ebolavirus
NC_002549.1
899
     47781 NC_035889
Zika virus
NC_035889.1
2777
     77881 NC_001802
Human immunodeficiency virus 1 (HIV-1)
NC_001802.1
8956
     24498 NC_001348
Human alphaherpesvirus 3 (HHV-3)
NC_001348.1
2787
     35512 NC_000907
Haemophilus influenzae
NC_000907.1
65721
      2257 KT2254762
Middle East respiratory syndrome-related coronavirus (MERS-CoV)
KT225476.2
21529
     43699 NC_001693
Human papillomavirus type 60
NC_001693.1
14879
     89214 NC_001690
Human papillomavirus type 48
NC_001690.1
7100

10 rows selected.


GENE_SYMBOL
PFAM
PROTEIN_ID
FASTA_ID
```


ORF1ab
pfam11501
YP_009072438
NC_025217.1 Bat Hp-betacoronavirus/Zhejiang2013, complete genome
aroG
NP_7849740.1
NC_000907.1 Haemophilus influenzae Rd KW20, complete sequence
gag-pol
pfam00077
NP_789740.1
NC_001802.1 Human immunodeficiency virus 1, complete genome
ORF1ab
pfam16348
YP_009725300
NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ORF0
UL56 family
YP_053044
NC_001348.1 Human herpesvirus 3, complete genome
E6
pfam00518
NP_043437
NC_001693.1 Human papillomavirus type 60, complete genome
ORF1ab
pfam11501
YP_009725297.1
NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
POLY
pfam01003
YP_009428568
NC_035889.1 Zika virus isolate ZIKV/H. sapiens/Brazil/Natal/2015, complete genome
SUD-M
pfam1661
NP_789880.1
KT225476.2 Middle East respiratory syndrome coronavirus isolate MERS-CoV/THA/CU/17_06_2015, complete genome
E1
pfam00524
NP_043418
NC_001690.1 Human papillomavirus type 48, complete genome

NP
pfam05505
NP_066243.1
NC_002549.1 Zaire ebolavirus isolate Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Mayinga, complete genome

11 rows selected.

GENE_SYMBOL
GENE_ID
PROTEIN_ID
PFAM
MOL_WEIGHT

E1
E1
NP_043418
pfam00524
2744
E6
E6_ght
NP_043437
pfam00518
16679
ORF1ab
NA39_gp6
YP_009072438
pfam11501
17821
POLY
CPG35_gp1
YP_009428568
pfam01003
19096
ORF1ab
GU280_gp01
YP_009725297.1
pfam11501
19644
SUD-M
SUD-M
NP_789880.1

pfam1661
25511
aroG
aroG
NP_7849740.1
38994
ORF0
HHV3_gp01
YP_053044
UL56 family
39456
ORF1ab
GU280_gp01
YP_009725300
pfam16348
56184
NP
ZEBOVgp1
NP_066243.1
pfam05505
83156
gag-pol
HIV1gp1
NP_789740.1
pfam00077
112754

11 rows selected.

```
    REF_ID JOURNAL
SCIENTIFIC_NAME
HOST
----------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------
     74441 Nature
Orthocoronavirinae
Homo sapien
     43699 Virology
Human papillomavirus type 60
Homo sapien


Journal Length JOURNAL
--------------
--------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
             6 Nature
            10 PLoS Biol.
             8 Virology
             6 Nature
             7 ISME J.
            12 J Gen Virol.
            13 N Engl J med.
             7 EMBO J.
            13 N Engl J med.
             9 Curr Biol
            14 Euro Surveill.

Journal Length JOURNAL
--------------
--------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
             8 Virology
             8 J. Virol

13 rows selected.


PFAM
--------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
pfam11501
pfam16348
pfam11501
pfam05505
pfam01003
pfam00077
UL56 family

pfam1661
pfam00518
pfam00524

11 rows selected.


1 row deleted.


PFAM
--------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
pfam11501
pfam16348
pfam11501
pfam05505
```

```
pfam00077
UL56 family

pfam1661
pfam00518
pfam00524

10 rows selected.


Commit complete.


Rollback complete.


ORG_TYPE
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------
Virus
Virus
Virus
Virus
Virus
Virus
Bacteria
Virus
Virus
Virus

10 rows selected.


10 rows updated.


ORG_TYPE
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------------------------------------------------------------------------
Prokaryote
Prokaryote
Prokaryote
Prokaryote
Prokaryote
Prokaryote
Prokaryote
Prokaryote
Prokaryote
Prokaryote

10 rows selected.
```

```
Commit complete.


Rollback complete.


PUB_DATE  JOURNAL
ARTICLE_TITLE
SCIENTIFIC_NAME
ORG_TYPE
Journals Published IN 2020
---------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
--------------------------
01-MAR-20 Nature
A new coronavirus associated with human respiratory disease in China
Orthocoronavirinae
Prokaryote
1


FASTA_ID
GENE_ID
PROTEIN_ID
PFAM
Average Protein Weight in Nsp1 Region
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------
-------------------------------------------------------------------------
-------------------------------------
NC_025217.1 Bat Hp-betacoronavirus/Zhejiang2013, complete genome
NA39_gp6
YP_009072438
pfam11501
17821
```

NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
GU280_gp01
YP_009725297.1
pfam11501
19644
NC_002549.1 Zaire ebolavirus isolate Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Mayinga, complete genome
ZEBOVgp1
NP_066243.1
pfam05505
83156

LAST_UPDA GENE_ID
GENE_SYMBOL
GENE_DESCRIPTION
FASTA_ID
PROTEIN_ID
PFAM

03-MAR-19 HHV3_gp01
ORF0
membrane protein UL56
NC_001348.1 Human herpesvirus 3, complete genome
YP_053044
UL56 family
15-AUG-18 E6_ght
E6
transforming protein E6
NC_001693.1 Human papillomavirus type 60, complete genome
NP_043437
pfam00518
15-AUG-18 E1
E1
replication protein E1
NC_001690.1 Human papillomavirus type 48, complete genome

NP_043418
pfam00524

SCIENTIFIC_NAME
DNA_LENGTH GC_CONTENT GENE_SYMBOL
PFAM

Haemophilus influenzae
65721 24 aroG
Orthocoronavirinae
25699 38 ORF1ab
pfam11501
Orthocoronavirinae
25699 38 ORF1ab
pfam16348
Human papillomavirus type 60
14879 27 E6
pfam00518
Human immunodeficiency virus 1 (HIV-1)
8956 38 gag-pol
pfam00077

SCIENTIFIC_NAME
GC_CONTENT AVG_CONTENT

Bat Hp-betacoronavirus/Zhejiang2013
45 38.2
Human alphaherpesvirus 3 (HHV-3)
65 38.2
Middle East respiratory syndrome-related coronavirus (MERS-CoV)
42 38.2
Human papillomavirus type 48
44 38.2

SCIENTIFIC_NAME
GENE_SYMBOL
PROTEIN_ID
PFAM
REGION_NAME
MOL_WEIGHT

--- ----------
Orthocoronavirinae
ORF1ab
YP_009725297.1
pfam11501
Nsp1
19644
Orthocoronavirinae
ORF1ab
YP_009725300
pfam16348
Corona_NSP4_C
56184

GENOME_ID
SCIENTIFIC_NAME
PROTEIN_ID
MOL_WEIGHT

--- ----------
NC_045512.2
Orthocoronavirinae
YP_009725297.1
19644
NC_025217.1
Bat Hp-betacoronavirus/Zhejiang2013
YP_009072438
17821

ARTICLE_TITLE
SCIENTIFIC_NAME
FASTA_ID

GENE_SYMBOL
PFAM

A new coronavirus associated with human respiratory disease in China
Orthocoronavirinae
NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ORF1ab
pfam11501
Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin
Human immunodeficiency virus 1 (HIV-1)
NC_001802.1 Human immunodeficiency virus 1, complete genome
gag-pol
pfam00077
Imported case of Middle East respiratory syndrome coronavirus (MERS-CoV) infection from Oman to Thailand, June 2015
Middle East respiratory syndrome-related coronavirus (MERS-CoV)
KT225476.2 Middle East respiratory syndrome coronavirus isolate MERS-CoV/THA/CU/17_06_2015, complete genome
SUD-M
pfam1661
Human papillomavirus type 48
Human papillomavirus type 48
NC_001690.1 Human papillomavirus type 48, complete genome
E1
pfam00524
A new coronavirus associated with human respiratory disease in China
Orthocoronavirinae
NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ORF1ab
pfam16348
The complete DNA sequence of varicella-zoster virus
Human alphaherpesvirus 3 (HHV-3)
NC_001348.1 Human herpesvirus 3, complete genome
ORF0
UL56 family
Molecular cloning of a novel human papillomavirus (type 60) from a plantar cyst with characteristic pathological changes
Human papillomavirus type 60
NC_001693.1 Human papillomavirus type 60, complete genome

```
E6
pfam00518
Characterization of the L gene and 5 trailer region of Ebola virus
Zaire ebolavirus
NC_002549.1 Zaire ebolavirus isolate Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Mayinga,
complete genome
NP
pfam05505

8 rows selected.
```

Database Administration and Monitoring

Roles and Responsibilities

There will be three main roles at Cloud Genome, the Database Administrator, Security Administrator, and System Administrator. The Security Administrator and System Administrator will work alongside each other collectively, to oversee the database. The Database Administrator will also work with them but in a more independent role focused on serving as a liaison between clients and Cloud Genome.

System Information

The database to be used will be Oracle Database 12c Enterprise Edition Release 12.2.0.1.0 64-bit running on a Linux VDA. The minimum software required is i5-6200 U CPU @ 2.30Hz 2.40 GHz with12. GB RAM 64 bit OS x64-based processor. All other software requirements are up to the client.

Date Formats

The database will support date formats as integers or strings from text files, csv files, and other files from string or integer format.

Backup and Recovery

Cloud Genome is to be backed up every eight hours with a full system reboot every night.

References

DeBarros, A. (2018). *Practical SQL.* No Starch Press.

Guseva, E., Batyrgazieva, D., Karetkin, B., & Menshutina, N. (2019). Development of a User Web-Interface for Working with the Information Database in the Field on Biotechnology: Prebiotics, Probiotics and Their Activity. *Proceedings of the International Multidisciplinary Scientific GeoConference SGEM*, *19*(1), 651.

Holywell, S. (n.d). *SQL Style Guide*. https://www.sqlstyle.guide/