12/03/2021

Lily Hofman and Alex Habegger

Gene Annotation of Zebrafish Genome Report

Zebrafish, *danio rerio,* are small tropical freshwater fish. They are great model organisms because they are small and can be easily housed and maintained in laboratory environments, they have very similar organ structures to humans except for lungs, and can be made transparent by using different chemicals to make their stripes disappear. Zebrafish are important because they are used in a variety of studies like studies on genetic research, cardiovascular activity, human metabolic activity, and even brain research. We created a tool to perform a gene annotation of two different releases from zebrafish chromosome 12; gene release 100 and gene release 104. Gene release 100 was released in March of 2020 and release 104 was released in March of 2021.

When creating our tool, we used GitLab to create a project where we could store and save our important files necessary for our project. After we chose our species, we began by doing research on the 25 chromosomes in the zebrafish genome. We chose to analyze chromosome 12 release 100 and release 104. We then downloaded the GFF3 files for chrosmoms 12 release 100 and 104 from zebrafish genome information from Ensembl.org. We then used a parser to get this information into python. This parser is called Parser.py and the primary function was to divide each line that was not a comment into separate elements that were stored in an array. Then we used this array to print out SQL insert queries using the element array for each line. Next we input the SQL insert queries into a SQL database to create tables to organize the data from our downloaded release files. This information formed the foundation of our project and is what is displayed in the user interface.

We designed a program that creates two tables of annotations for each release. The program grabs information from our SQL databases and places it within the corresponding tables: Release Information and Gene Categories. For each release, our gene category table examined the different transcription factors of the gene release including the different types of RNA genes, their size and whether they are coding, noncoding, or translating or non translating. The categories of exons describe parts of a gene sequence that could code for proteins during translation, and categories of CDS which are coding regions of the gene. There is a category called biological regions describing the number of active regions on the release. There are also categories representing pseudogenes and pseudogene transcripts which are nonfunctional genes in a region that indicate genes that could be functional in different releases.

Our database answers the question of finding the associated transcript numbers for each gene category though it is not included within our tool. In order to find the information for release 100, the proper SQL query would be ‘SELECT Distinct Transcript\_ID FROM Transcripts100 WHERE Type = “gene category”;’. Also to find the information for release 104, the proper SQL query would be ‘SELECT Distinct Transcript\_ID FROM Transcripts104 WHERE Type = “gene category”;’. Both of these SQL queries would grab the transcript ID from the database of a specified gene category.

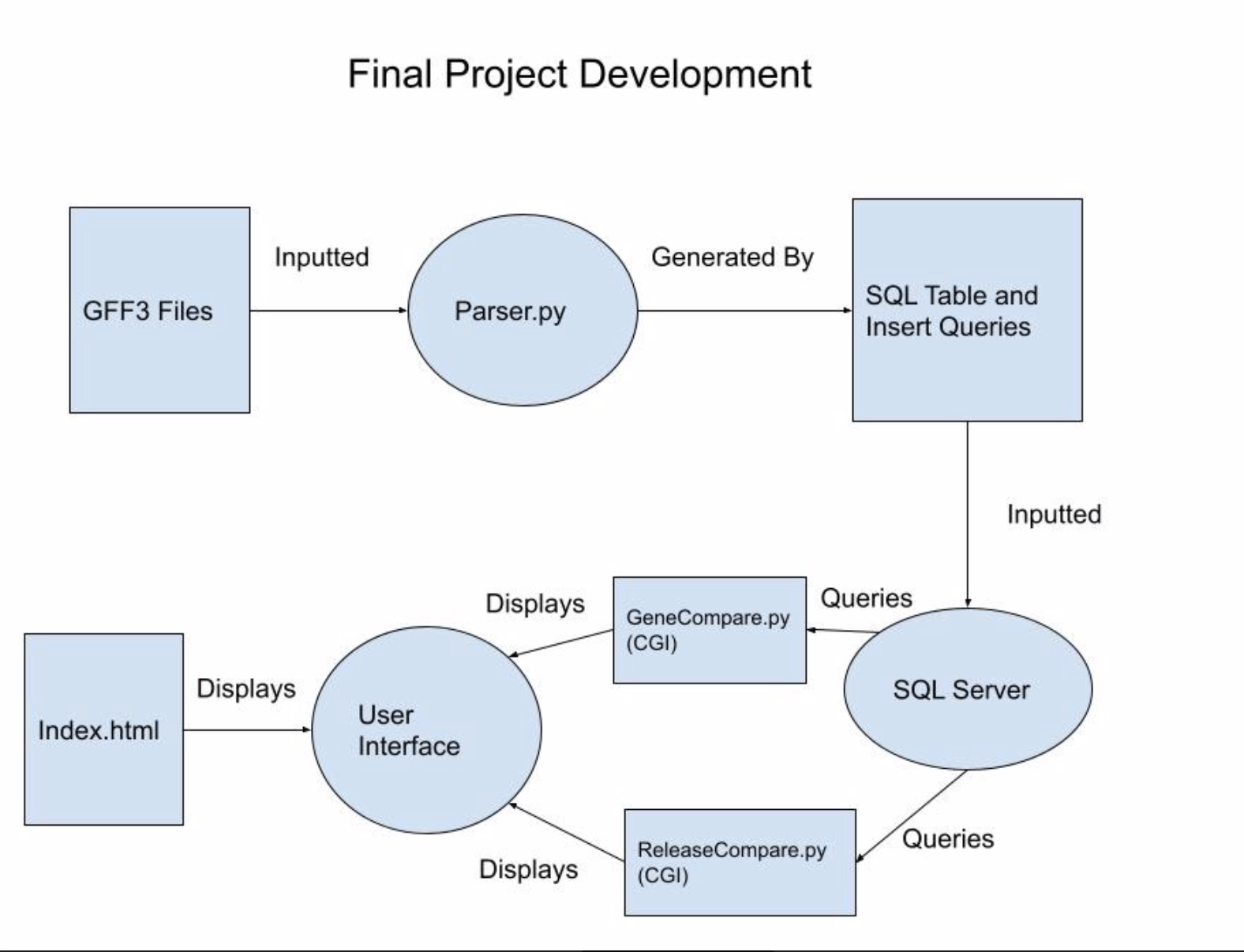
For release 100 and 104, the release information table has three columns displaying their release number, their chromosome number that the release came from and the unique transcript number in that gene release. The gene category information table displays the number of mRNA type genes, exons, CDS, the number of biological regions, non coding RNA genes, the number of long non coding RNA genes, the number of small nucleolar RNA, the number of pseudogenes, pseudogene transcripts, and the number of microRNA genes.

Our tool creates two tables of annotations for each of the two releases grabbing information from our SQL databases. The tables created were release information and gene category information. For the gene release information table release 100 was on chromosome number 12 and the unique transcript number was 1834. For release 104 this information was exactly the same as the information for release 100. Also, there were 13 columns describing the different gene categories. For release 100 there were 1492 mRNA type gene, 1139 three prime untranslated regions, 17169 exons, 15378 coding sequences (CDS), 1687 five prime untranslated regions, 2445 biological regions, 160 non coding RNA type genes, 243 long non coding RNA type gene, 81 ribosomal RNA type genes, 7 small nucleolar RNA type genes, 4 pseudogenes, 4 pseudogene transcripts and 7 microRNA type genes. For release 104 these values were exactly the same for each column of the table, this can be seen in table 1. All in all we found that there were 1834 unique transcripts annotated for both release 100 and release 104.

For the gene comparison, our tool connects to an SQL database which is used to grab the information about a gene ID that a user inputs. The corresponding SQL tables are Transcripts100 and Transcripts104 each grabbing from their respective releases.For a specific inputted gene, we found that between releases 100 and release 104 there were no difference values displayed in any of the categories in each table. For example, for inputted gene ENSDARG00000032578 there was no difference between the two releases. They are both on chromosome 12, the annotation provider for both releases is ensembl havana, the transcript type for both is gene, they both have a start position of position 45799982 and an end position of 45876387, they both have the same score and phase, both have negative strand directions, both have a gene length of 76405 and exon numbers of 74. This can be seen below in table 2.

Our work is an important tool for further research on genetic diseases and disabilities. Our program takes two gene releases from the same chromosome and compares the numbers of specific gene factors and transcription factors. This tool could help support researchers studying how various regions of genes on the same chromosomes may function or be affected by certain factors in different ways. In the case of our releases, the gene release information was exactly the same for release 100 and release 104. There was no difference in unique Transcript IDs or Gene IDs between the two releases in the release information table. There was no difference between the two releases for each column value in the gene category information table. The lack of differences between gene release 100 and gene release 104 is likely a result of the one year difference between the two releases;gene release 100 was released in 2020 and gene release 104 released in 2021. It is important to keep the gene annotation information updated so that the researchers, as well as the general public, has access to relevant and up to date information when studying genomes of different species.

How we created our tool flow chart

This information could provide insight into how different gene releases may function exactly the same 

Our web interface

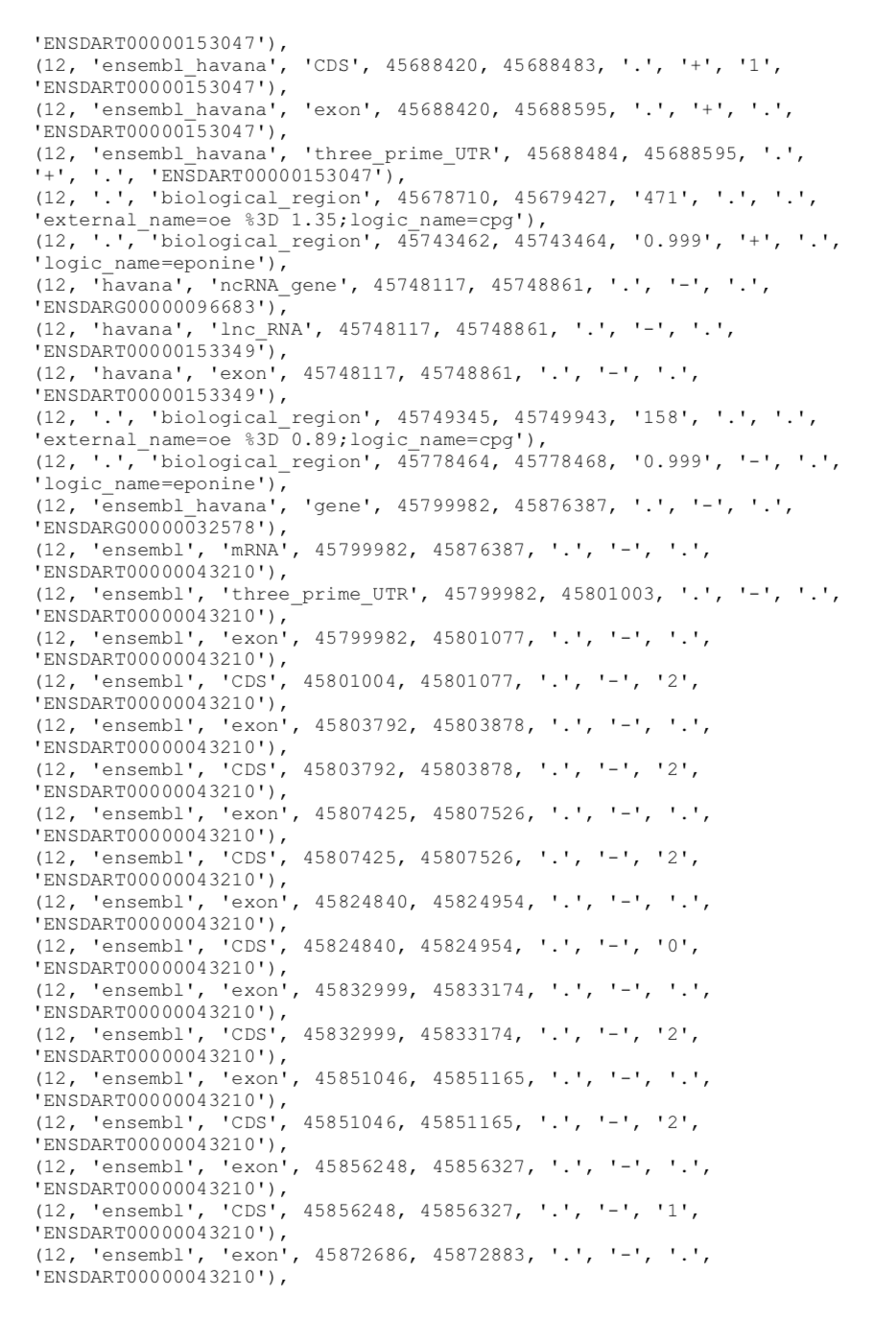




Table 1: Compares gene release information and gene category information between release 100 and release 104



Table 2: Compares gene annotation information for an inputted gene between release 100 and release 104



Example of SQL dump raw data from transcript 100