**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**Dr. Arun Sethuraman**

**Lab 6: Genome Annotation**

Today, we will focus on an annotation pipeline that comprises (1) repeat masking, (2) ab initio gene prediction, (3) evidence-based refinement of the annotation, and (4) creating a genome browser instance. We will utilize the Galaxy Project for this purpose (just for ease of use, since a lot of the tools for annotation take way too long to install on JetStream), but please note that it’s just as easy to install these tools and then perform analyses. We will use the same zebrafish dataset from last week, since we also have RNAseq based evidence, and ways to build tracks from the already established annotation (GFF3) file.

**Exercise 0: Setting up your Galaxy history**

Switch to the correct history, and ensure that you have (1) Danio\_rerio.Zv9.66.dna.fa (note that this is only Chromosome 12), (2) 2cells\_1.fastq, 2cells\_2.fastq, 6hcells\_1.fastq, 6hcells\_2.fastq, (3) Danio\_rerio.Zv9.66.gtf.gz in your history. These should hopefully still be around, since you just completed a lab on RNAseq with these data. If not, please make sure to obtain these files (from Canvas), and upload them to your Galaxy page after logging in.

**Exercise 1: Repeat identification and masking**

We will use a popular tool called RepeatMasker for this purpose ([www.repeatmasker.org](http://www.repeatmasker.org)). Search for RepeatMasker in Galaxy, and this should pull up a window with options to run RepeatMasker. Thereon, click on “Upload Data”, and “Paste/Fetch Data”, and enter this URL: <https://www.dfam.org/releases/Dfam_3.6/families/Dfam.embl.gz> and hit “Start”. We will be using the Dfam repeat database (dfam.org) as a guide for annotating repeats in our genome. Note that this is a 2.1 Gb file, so it might take a bit of time. Be patient!

Once this is uploaded, you can now set up your RepeatMasker run - pick Danio\_rerio.Zv9.66.dna.fa as your Genomic DNA file, click on the folder icon to pick the Dfam file as your RepBase file. Thereon, select “Yes” to pick the species name from a list, and search for “Danio” in the Species. Go ahead and set the output to produce a GFF formatted annotation, then hit “Execute”.

**Exercise 2: Ab initio gene prediction with Augustus**

Now search for “Augustus” and pick the tool. Select your Danio\_rerio.Zv9.66.dna.fa as the “Genome Sequence”, use a predefined TrainingSet (Model Organism: Danio rerio), set the output format to be “GFF formatted”, select all outputs, and hit “Execute”.

**Exercise 3: Creating a Genome Browser instance**

Search for “JBrowse” in Tools, select “Use a genome from history” for the reference genome to be displayed, and then select your Danio\_rerio.Zv9.66.dna.fa file as your reference genome. Now create a new “Track Group”, give the category any name (I just used #date# and that fills up today’s date), thereon select the GFF3 tracks to display (RepeatMasker and Augustus outputs) in two separate Tracks. Thereon, you’re welcome to play around with advanced options prior to hitting “Execute”.

**Exercise 4: Using the MAKER pipeline**

In the interest of time, I will let you play with this - MAKER (Campbell et al., 2014), and subsequent versions of this tool are great annotation platforms that permit you to use a variety of evidences (e.g. EST, cDNA libraries, aligned cDNA, protein evidences from a previous round of annotation, etc) to perform evidence-based annotation (as against ab initio annotation). So go ahead and search for “Maker” in the Tools tab, and check out the different options that they have, and set up a Maker annotation for the Danio rerio FASTA.