Phuong Anh Dinh

Advisor: Dr. Saville

07/14/2020

**FURSCA END OF SUMMER REPORT**

**Introduction**

My 2020 Summer FURSCA consisted of two different projects, which are:

1. Transcription Start Sites (TSSs) annotation of genes in contig 29 of *Drosophila ananassae*.
2. Variant calling on genome sequencing data and downstream analysis using GEMINI and R for interactive human-diseases-related SNPs report.

The first project was part of GEP, which stands for Genomics Education Partnerships – “a nationwide collaboration of 100+ institutions that integrates active learning into the undergraduate curriculum through Course-based Undergraduate Research Experiences (CUREs) centered in bioinformatics and genomics”[[1]](#footnote-1). TSS annotation attempts to look for the nucleotide at which transcription begins based on a variety of available evidence. This project, hence, is expected to offer a closer look at transcription and better our understanding on this part of gene regulation. In this project, I worked specifically with genes in contig29 which is a part of the chromosome in *Drosophila ananassae*.

With the rapid development of sequencing technology, the cost associated with genome sequencing has decreased significantly over the past decade. In the near future, people may be able to obtain their own DNA sequence at a reasonable and affordable price and use such information for personalized medicine, treatment, or informative diseases risks report. My second project seek to process the genome sequencing data, find single nucleotide polymorphisms (SNPs) and its association with diseases (if known), and produce an interactive human-diseases-related SNPs report for users.

**Methods & Results**

1. **Transcription Start Sites (TSSs) annotation of genes in contig 29 of *Drosophila ananassae***



Figure 1: Full view of contig29 in *D.ananassae* in the genome browser.

The annotation followed the standard workflow provided by GEP which included the categorization of TSS in *Drosophila melanogaster* and the search for a TSS search range in *Drosophila ananassae*.

In the first place, isoforms with unique TSS will be identified, and each of them will be categorized as either peaked, intermediate, broad, or insufficient evidence based on the number of TSS and DNase Hypersensitive Sites (DHS) identified.

 Moving on to annotating TSSs in target species, the first transcribed exon of specific isoform will be obtained from gene record finder and blast against the contig sequence of *D. ananassae.*



Figure 2: Blast result of first transcribed exon in one isoform vs contig sequence of *D.ananassae*.

As indicated in the module, if the blast result is usable with a low enough E-value, the TSS search range can be defined by taking the coordinate of first transcribed exon +/- 300. For example, the search range indicated from figure 2 will be 233,062 – 233,662.

The genome browser will then be used to look into detail in this region and find supporting evidence such as RNA-Seq data and RAMPAGE result. However, in most of the case, the blast result had high E-value, and the coordinate can’t be used to define a TSS search region. Other evidences in genome browser would be used to define a search region instead. Examples of those evidences include RAMPAGE, CAMPAGE, RNA Pol II chip seq, RNA seq & Top Hat Junctions, conservation blocks, etc.

The final step in TSS annotation was to look for core promoter motif to support the TSS search range found. Some of the common Drosophila motifs and its relative position compared to the TSS can be found in table 1 below. It is worth mentioning that some promoters do not contain any motif at all.

Similar analyses were performed on 12 isoforms of 3 different genes in contig29. In the end, data for both CDS[[2]](#footnote-2) and TSS annotation were combined to fill in the final report for contig29.



Table 1: Some of the common Drosophila motifs and its relative position compared to TSS.

1. **Variant calling on genome sequencing data and downstream analysis using GEMINI and R for interactive human-diseases-related SNPs report.**

Small sample of human genome sequencing data was obtained from Galaxy training material[[3]](#footnote-3). The data was processed using FastQC and Trimmomatic for filtering out low-quality bases, BWA-MEM for mapping against reference genome, RmDup for post-alignment cleanup, Freebayes & bcftoolnorms for variant calling, and Snpeff for annotation. All these steps were outlined in Galaxy Variant Calling Training. The vcf file generated from Snpeff was then loaded into GEMINI – a package that integrates annotation from many different sources such as ClinVar, dbSNP, ENCODE, 1000 Genomes, etc – to create an accessible database. SQL commands were then used to query for specific information in question about SNPs. Some of the important information about each SNP that can be retrieved from GEMINI include chromosome number (chrom), type of variants (type), variant sub-type (sub\_type), gene of highly affected transcript (gene), consequence of the most severely affected transcript (impact), etc.



Table 2: An example of result obtained from GEMINI query

The table was then saved and loaded into R for further analysis. Using R Shiny package, an app was written to produce an interactive report where users can customize and easily access their desired information.

So far, the app has 2 tabs.

The Raw\_Data tab contained the raw information achieved from GEMINI query. Each line corresponds with one SNP and its related information such as chromosome number, name of mostly affected gene, type of effect, clinical significance, amino acid change, etc. Users can customize to see interested columns only. The key word search box allow search for all SNPs that are associated with a disease of interest. One drawback of this search box was that the entered keyword must match 100% with the phrase or part of the phrase in the clinvar\_disease\_name column.



Figure 3: Raw\_Data tab of the application



Figure 4: Unique\_Data tab of the application.

The second tab – Unique\_Data – contained filtered information based on unique disease name in the clinvar\_disease\_name column of the original table. Here, the number of SNPs with low severity, medium severity, and high severity associated with unique disease name were reported. Search box at the end of each column allows for quick search. For example, name of interested disease can be entered in the box under clinvar\_disease\_name column, and the table will filter out rows with matching key word.

**Future Direction**

 I plan to address some current problems and add more functions to the interactive report of the second project. Some aspects that I will work on include:

1. Improving the search function on Raw\_Data tab.
2. Providing explanation on information contained in each column in Raw\_Data tab.
3. Adding a visualization tab where information on Unique\_Data tab can be visualized through graphs.

**Acknowledgement**

 I would like to express my deepest thank for the Bruce A., ’53 and Peggy Kresge, 53’ Endowed Science Fellows Fund for making this research opportunity available, my advisor for guiding me through this project, and the FURSCA Committee for an awesome summer research experience!

1. https://gep.wustl.edu [↑](#footnote-ref-1)
2. The coding region (CDS) annotation data was obtained from Dr. Saville’s Spring 2020 bioinformatics class. [↑](#footnote-ref-2)
3. https://galaxyproject.github.io/training-material/topics/variant-analysis/tutorials/exome-seq/tutorial.html [↑](#footnote-ref-3)