Ensemble Solution and Evaluation of Genomic Annotation Softwares – Manual of Use

1. Pre-processing.

We can use GFF and GFF3 files as entry for this system, since they represents the final genomic annotation for the candidates softwares we want evaluate.

If you have a GFF3 file, use the script inventorySW1.pl;

If you have a GFF file, use the inventorySW2.pl.

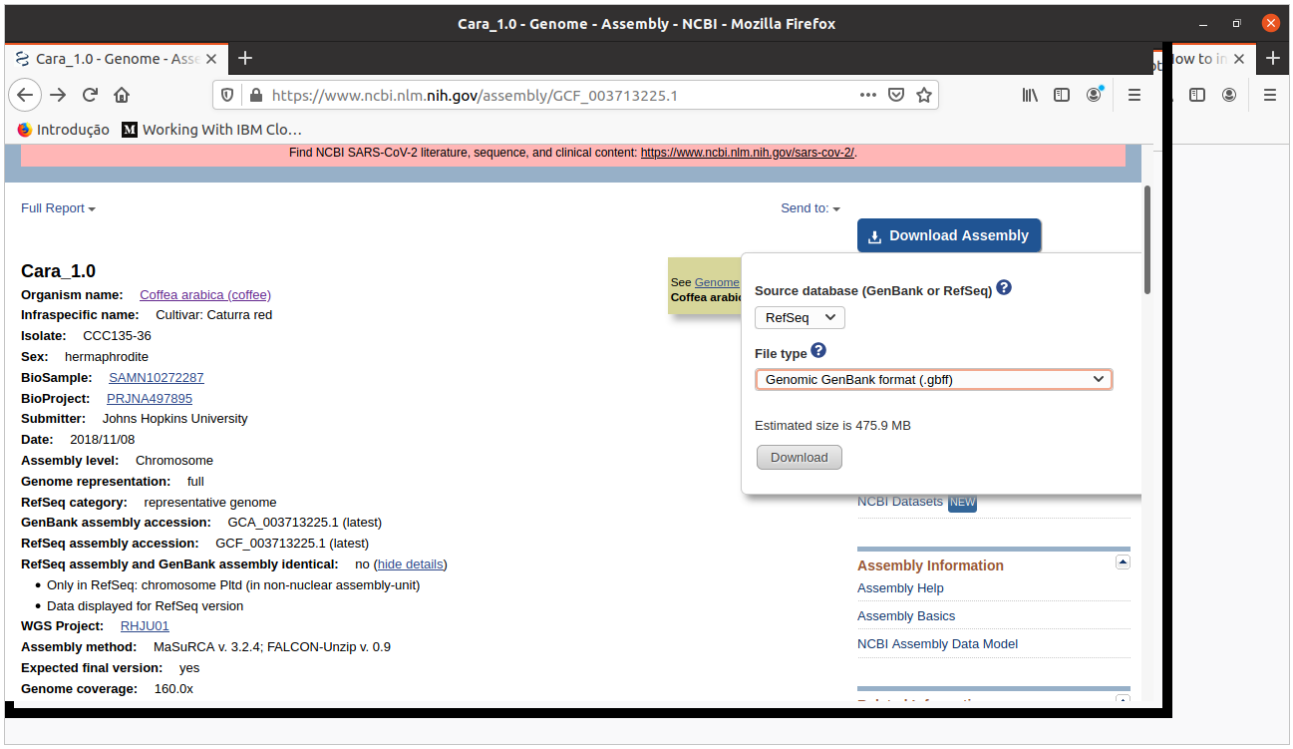
By running these scripts it is possible to generate text files having the list of annotated genes from each pipeline. The file names must be specified as arguments of the scripts. See:

```
perl ../inventorySW1.pl ./arabicaTirateima.gene_structures_post_PASA_updates.63989.gff3 gravaDestinoPASA.txt

perl ../inventorySW2.pl ./makerArabicaNCBI.gff3 gravaDestinoMAKER.txt
```

One property of Ensemble Solution is the generation of customized reports including information from a GenBank file, as the translation of CDS for each listed gene. At first download the specific .GBFF file of the specie used in the evaluated pipelines.

For example, in the NCBI web site we can find for *C. arabica* many file formats available:



2. Principal

Once we have both .txt and the .gbff files, we can run the main process:

```
perl ./ensemble.pl PASA ./gravaDestinoPASA.txt MAKER ./gravaDestinoMAKER.txt ./GCF_003713225.1_Cara_1.0_genomic.gbff
```

For this example, we are evaluating the softwares PASA and MAKER. Because of this their names must to be written beside of each list of genes text file. In the end of the command, write the GenBank file path.

This runs during 0m13,265s in a Penthium Core i3 processor with 4GB RAM and generates five reports and two graphs:

• onlySW1.txt: list of the annotated genes by the pipeline of software 1 (exclusively);

• onlySW2.txt: list of the annotated genes by the pipeline of software 2 (exclusively);

• intersec.txt: list of the annotated genes by both pipelines (intersection set);

• reportonlySW1.txt: GenBank information about the genes present exclusively in the pipeline of software 1;

• reportonlySW2.txt: GenBank information about the genes present exclusively in the pipeline of software 2;

• VennChart.png

• VennHistogram.png

3. Environment

Using Linux Ubuntu, for instance, some packages are needed to be installed, as:

- libbio-perl-perl
- libbio-perl-run-perl

The system also uses the modules:

- Venn::Chart;
- Array::Contains;
- Array::Utils;
- List::Uniq;
- Bio::Seq;
- Bio::SeqIO;

4. Contact

Written by Geraldo Cesar Cantelli

E-mail: geraldocesar.77@uol.com.br

PPGBIOINFO - UTFPR campus Cornélio Procópio – Brazil

Available on https://github.com/geraldocantelli/ensemble