**Plant Pathogenomics program – April 2017 – Norwich, UK**

**Tuesday, April 4th 2017**

**Annotation and comparative genomics of *Magnaporthe oryzae* (rice blast) genomes**

By Dr. Ronnie de Jonge

In this exercise, we are going to reproduce parts of the analyses published in Chen et al. 2013 & Dong et al. 2015. We will go through a set of commonly deployed steps to obtain annotated genome sequences, extract protein sequences, perform parts of automated annotation, (e.g. to identify carbohydrate-active enzymes (CAZYs) and biosynthetic gene clusters (BGCs)) if time permits), and determine sets of genes encoding potentially secreted proteins, representing candidate effector molecules. Finally, we will proceed to compare the individual genome assemblies (by means of whole-genome alignment), and their respective coding potential (with emphasis on candidate effector molecules).

**Input data**

We will work with the reference *Magnaporthe oryzae* strain 70-15 (PMID 15846337). This was one of the first fungal genomes sequenced and has since undergone substantial improvement. The current assembly version is MG8. In addition, we have two field isolates from Chen et al. (PMID 24341723) and one from Dong et al. (PMID 25837042).

Reference strain: (Dean et al. 2005, Nature)

* *M. oryzae* 70-15

*Strains we are going to compare with:*

Chen et al. 2013 (BMC Genomics)

* *M. oryzae* FJ81278
* *M. oryzae* HN19311

Dong et al. 2015 (PLoS Pathogens)

* *M. oryzae* 98-06

Data files and genome sequences are here:

*##* [*https://drive.google.com/drive/folders/0B3E0PNqPeoThb3pDbEtzV09kSlU*](https://drive.google.com/drive/folders/0B3E0PNqPeoThb3pDbEtzV09kSlU)

**Exercise**

Part 1a. Genome annotation

The first step in genome analyses concerns the identification and refinement of gene models. It strongly affects downstream analyses. We will experiment with a widely-used *ab initio* gene finder, i.e. Augustus, first developed in 2004 by Stanke (PMID: 15215400). Augustus uses genome-specific parameters that require optimization to work best. Fortunately, specific *Magnaporthe* parameters are known and incorporated in the common Augustus species set (i.e. magnaporthe\_grisea). We will use this set. In the absence of such set, one needs to either train one her/himself or use the parameters of a self-chosen nearest alternative. To demonstrate the impact of correct parameter choice/optimisation, we will run similar predictions using another set of parameters (*Aspergillus fumigatus*).

*## open running\_augustus.txt to proceed with this step*

Once we have gene predictions (in GFF3 format) we will load them in IGV (Integrative Genome Viewer) and compare them to each other and to aligned RNA-Seq data to get a feel for accuracy and associated problems.

*## open visualizing\_genemodels.txt to proceed with this step*

Finally, I’ll perform a short demonstration of the abilities to modify and update gene models in WebApollo, a community annotation platform for genome sequences.

Part 1b. Protein/gene annotation

Now that we have obtained gene predictions, assessed their quality we will proceed to perform semi-automated functional annotation. A commonly used tool for this purpose is BLAST2GO (PMID: 16081474), but here we will experiment with a more quick-n-dirty, command-line driven annotation. To this end, we first need to obtain protein sequences (using the Augustus gene predictions, and the underlying genome sequences). Secondly we will perform 1) protein blast against the uniport/swissprot database 2) protein domain identification using hmmsearch in combination with the Pfam-A database of HMM models (prepare the data to do so, run for small test set, but we will use pre-calculated data here), 3) determine protein localization by signal peptide prediction. If time permits, we can 4) determine carbohydrate-active enzymes using dbCAN HMM (PMID: 22645317) search and 5) detect secondary metabolite clusters using SMURF (webserver, requiring specific genome edits).

*## open protein\_annotation.txt to proceed with this step*

Part 2. Comparing genome annotations / protein coding-gene models (presence/absence polymorphisms)

In this exercise, we will use two methods to perform gene catalogue comparisons. Firstly (Part2) we will use the protein annotations to derive presence/absence variation, and secondly (Part3) we will use whole-genome alignments to determine presence/absence variation of genes. In this part, Part2, we will determine protein families (using blast all-v-all (pre-run) followed by orthoMCL clustering(pre-run)) and extract unique proteins (families). Subsequently we can explore annotations and if time permits assess domain (Pfam) enrichment.

*## open compare\_protein\_annotations.txt to proceed with this step*

Part 3. Comparing genome sequences, by means of whole-genome alignment

Now that we have established genomes and gene models lets proceed with the third part of this exercise, i.e. ‘comparative genomics´. We will be looking at genome-wide alignments a broad overview perspective and on the gene level (particularly presence/absence polymorphism). First we’ll generate whole genome alignments using MUMMER (PMID: 14759262). Depending on the expected similarity between species and/or strains a nucleotide-based alignment (*nucmer*) or a translated nucleotide-based alignment (*promer*) can be used. In this exercise, we will limit ourselves to *nucmer* alignments (considering relatively high similarity among strains).

*## open compare\_genomes.txt to proceed with this step*