Module 8

IGV for Data Visualization and Exploration

Goal: Utilize **Integrative Genomics Viewer (IGV)** to visualize predicted genes, transcript assembling and RNAseq read alignments

Background

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 2012.

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

8.1 Installing IGV

- ☐ Go to http://www.broadinstitute.org/software/igv/home
- ☐ On the left-hand, side click on "Downloads".
- ☐ Select and install the version appropriate for your system. Unless you have a specific reason to use an existing Java installation, you should select a version with Java included.

8.2 Prepare annotations

Input(s):	`magnaporthe_oryzae_70-15_8_single_contig.fasta`
	`Mo*accepted_hits_Chr7.bam`
	`cufflinks.gtf`
	`maker-annotations.gff3`
	or `maker-preview.gff` (if your **MAKER** run did not complete)

Output(s):	`Mo*accepted_hits_Chr7.bam`
	`Mo*accepted_hits_Chr7.bam.bai`
	Tracks and annotations visible in IGV

- ☐ Connect to your virtual machine.
- ☐ Change to the *rnaseq* directory.

We will only explore Chromosome 8.7 in the browser, so let's extract the alignment data for just that chromosome. This will significantly speed up the data transfer to our local machines.

- ☐ First, we need to index our bam files:

```
• for f in alignments/*.bam; do samtools index $f; done
```

- ☐ Check that the requisite index (.bai) files were created in the *alignments* directory.
- ☐ Now let's perform the extraction of just the Chromosome_8.7 records:

```
• for f in alignments/*.bam; do
  samtools view -b $f Chromosome_8.7 > ${f/hits/hits_Chr7}; done
```

- ☐ List the *alignments* directory (`ls -lrt`) to make sure six files with *_Chr7.bam* suffixes were created.
- ☐ Now, we need to index our Chromosome_8.7-specific bam files:

```
• for f in alignments/*Chr7.bam; do samtools index $f; done
```

- ☐ Again, list the directory (**ls -lrt**) to make six new (*.bai*) files were created.

8.3 Transfer annotations to local machine

- ☐ Use **scp** to transfer the following files from your VM to your local machine:

  ~/rnaseq/alignments/Mo_70-15_LC{1-3}_accepted_hits_Chr7.bam

  ~/rnaseq/alignments/Mo_FR13_IP{1-3}_accepted_hits_Chr7.bam

  ~/rnaseq/alignments/Mo_70-15_LC{1-3}_accepted_hits_Chr7.bam .bai

  ~/rnaseq/alignments/Mo_FR13_IP{1-3}_accepted_hits_Chr7.bam.bai

  ~/rnaseq/merged_asm/cufflinks.gtf

  ~/genes/maker/maker-annotations.gff3 **or** *maker-preview.gff* if your **MAKER** run did not complete.

  ~/genes/snap/magnaporthe_oryzae_70-15_8_single_contig.fasta

You can use the following command to transfer the .bam and .bai files en masse:

```
• scp myName@xx.xxx.xxx.xx:rnaseq/alignments/*Chr7* .
```

8.4 Navigating IGV

The best way to learn how to navigate a genome browser is simply to play with it by clicking on buttons/features and clicking and dragging in the navigation pane. However, to help you get going, you can refer to this video tutorial:

IGV | Data Navigation Basics

8.5 Load genome into the genome browser

- ☐ Now, open **IGV** on your local machine by double-clicking on the program's icon.
- ☐ In the menu, select **Genomes** > **Load Genome from File...**
- ☐ Browse to the folder where you secure-copied files from the server, and select *magnaporthe_oryzae_70-15_8_single_contig.fasta*.

8.6 Load MAKER annotations into the genome browser

- ☐ First, we will examine the genes that we predicted yesterday. Use **File** > **Load from File...** to import the *maker-annotations.gff3* file.
- ☐ Right-click on the *maker-annotations.gff3* name in the left-hand panel again and select "expanded" view. This will allow you to visualize potentially overlapping genes on the opposing DNA strands.
- ☐ Using several clicks of the "+" button at the top right, zoom in far enough to see the structures of the individual genes.

You will see that each feature is labeled with an identifier that tells you if it was predicted by **snap** and/or **augustus**, if it is a gene model created by **maker**, and/or matched a protein in the NCBI database. Features with an "XP" prefix exhibited matches to proteins at NCBI.

- ☐ When you are done, collapse the track by right-clicking on "maker-annotation.gff3" in the left panel and select "Collapsed."

8.7 Load transcript assemblies into the genome browser

- ☐ Now we will use **File** > **Load from File...** again to import the transcript assembly produced by the `perl Inherit_IDs.pl` command that we ran earlier.
- ☐ Using several clicks of the "+" button at the top right, zoom in far enough to see the individual transcript forms.
- ☐ Right click on the *cufflinks.gtf* name in the left-hand panel again and select "expanded" view. This will allow you to visualize alternative transcript forms identified by **cufflinks**, as well as overlapping transcripts on each strand of the DNA.

The **MAKER** run was informed by prior RNAseq data generated from fungus grown in liquid culture as well as from spores. The RNAseq data we used today incorporated new data from fungus growing *in planta*, and also contained MANY more reads from liquid grown cultures than were analyzed previously. Thus, it is possible that these new data might identify novel genes that escaped prediction by **SNAP** and **Augustus** and which were not picked up by the similarity searches performed by **MAKER**.

- ☐ Start exploring the single contig by clicking and dragging in one of the browser tracks, by dragging the red square that represents the current view window, or by clicking on the left and right arrowheads at each end of the window size indicator.

- ☐ Find a gene that is present in the *cufflinks.gtf* track but was not predicted by **MAKER**. Note that large genes are more likely to be valid genes than smaller ones which might be represented by just a handful of RNAseq reads. List the coordinate(s) of a few potentially novel gene(s):

- ☐ When you are done, collapse the track by right-clicking on "cufflinks.gtf" in the left panel and select "Collapsed."

8.8 Load RNAseq alignments into genome browser

- ☐ Next, we will load the .bam alignment files by selecting **File > Load from File...**
- ☐ Open all of the downloaded **accepted_hits_Chr7.bam* files.

Three plots should open up for each track. The first shows depth of RNAseq read coverage for each position on the chromosome; the second shows splice junctions; and the third shows alignments for each RNAseq read mapping to the chromosome region in question. Depending on the Zoom level, you may or may not see anything at first because information about the RNAseq read alignments only show up after you have zoomed in far enough. If necessary, zoom in so that you can start seeing the read alignments in the main window.

- ☐ Because we have so many reads, it may be necessary to "squish" the view so that more alignments can be displayed in each browser "track." Right-click over each of the filenames in the left-hand panel and select "squished" view in the menu that pops up. If necessary, close the "Coverage" tracks by clicking on the name in the left-hand panel, hitting the delete key, and confirming if prompted.

Note that the 70-15 RNAseq reads were much longer than those generated for FR13 (250 nt vs. 50 nt).

- ☐ Revisit the locations where your transcript assemblies found potentially novel genes and examine the RNAseq data tracks to determine which growth condition (liquid culture and/or *in planta*) produced the reads that supported the respective transcript assemblies.

  gene coordinate:_____; condition: _____

  gene coordinate:_____; condition: _____

  gene coordinate:_____; condition: _____

- ☐ Now, we'll manually examine the RNAseq data tracks for differential expression. For improved visualization, close the cufflinks and maker-annotations tracks and re-open when you need to see gene IDs.

- ☐ Navigate along the contig and find a gene that was expressed at a high level *in planta* but showed no expression in liquid culture:

  gene_id:_____

- ☐ Now find a gene that was expressed at a high level in liquid culture but showed no expression *in planta*:

  gene_id:_____