## June 2018

# An Introduction to Deep-Sequencing Data Analysis

## Exercise #3: Analyzing RNA-Seq data

***Dena Leshkowitz & Ester Feldmesser***

### Introduction

We will analyze RNA-Seq data from an experiment performed in Arabidopsis. The data was downloaded from the public domain, see the following link for more details: <https://www.ncbi.nlm.nih.gov/pubmed/26084880>.

In the previous exercise we ran FASTQC and STAR. In this exercise we will observe the outputs produced by them. If you did not finish running STAR please do so now.

Answers to Exercise 2 are found here

<http://dors.weizmann.ac.il/course/course2018/AnswersForExercise2.docx>

### Instructions

1. FASTQC results

View the reports you produced and decide if the sequences are of sufficient quality.

The correct way was to first examine the fastqc output and then if needed trim the sequences before running STAR. However, due to time limitations we did not observe the fastqc outputs before running STAR.

1. STAR results

STAR produces several outputs as you can see if you list the files in your folder.

* 1. The files that end with “Log.final.out” contain summary mapping statistics and they are very useful for quality control.

**Question: How many reads did each sample have and how many were mapped uniquely?**

**To view the output,** type:

(the “\*” is termed wildcard and means any character):

*more \*Log.final.out*

* 1. The file ReadsPerGene.out.tab contains STAR read counts per gene. In the output, the first 4 lines are general statistics.
  2. Then there are 4 columns which correspond to different library protocols strand options:
     1. column 1: gene ID
     2. column 2: counts for unstranded RNA-seq
     3. column 3: counts for the 1st read strand aligned with RNA (htseq-count option -s yes)
     4. column 4: counts for the 2nd read strand aligned with RNA (htseq-count option –s reverse)

**Question: How many genes are represented in the file?** **Hint**- each line except the first lines in the files are counts per gene, you should use a Linux command for the answer (look at supplementary section)

*module load samtools*

*module load python/bio-2.7.13*

*samtools index*

The last command will return the instructions for the command

Type the command to send the job to the cluster:

*bsub -q bio-guest -R "rusage[mem=10000]" -J sam1660397 -e ./samtools1660397.err -o ./samtools1660397.out "samtools index SRR1660397Aligned.sortedByCoord.out.bam"*

Once the run finishes (job not running any more), the file ./samtools1660397.out should contain a line: "Successfully completed" and you should have an index file created. Examine this using the command ls.

*ls -l \*bai*

You should be able to see the file, as seen below.

-rwxr-x--- 1 class1 testing-wx-grp 256608 Jun 11 11:28 SRR1660397Aligned.sortedByCoord.out.bam.bai

* 1. View Alignment on a genome browser

Open IGV (see exercise 1 or supplementary section)

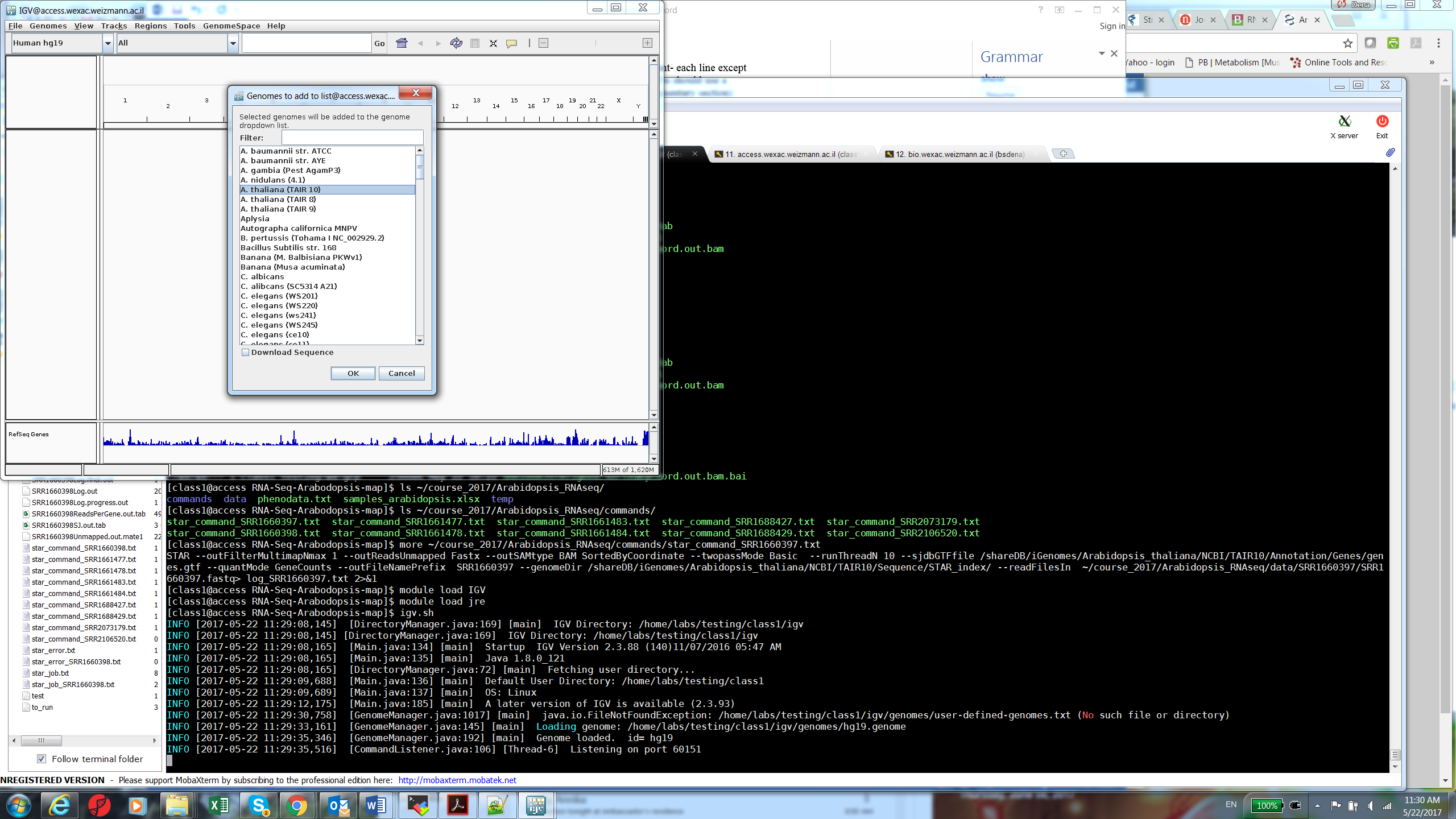
*module load IGV*

*module load jre*

*igv-10g.sh*

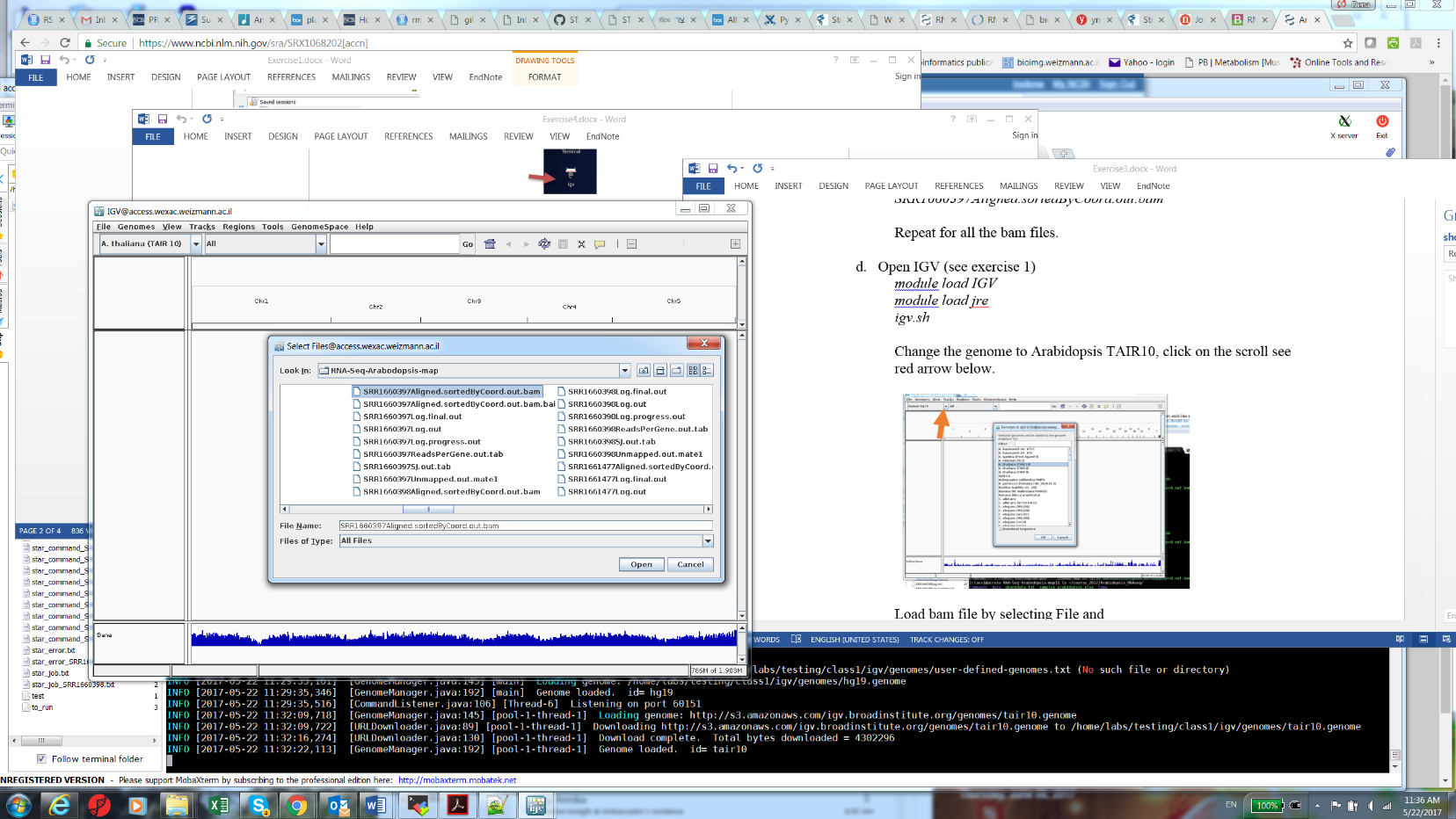
It will take a few minutes to open the IGV browser.

Change the genome to Arabidopsis TAIR10, click on the scroll see red arrow below

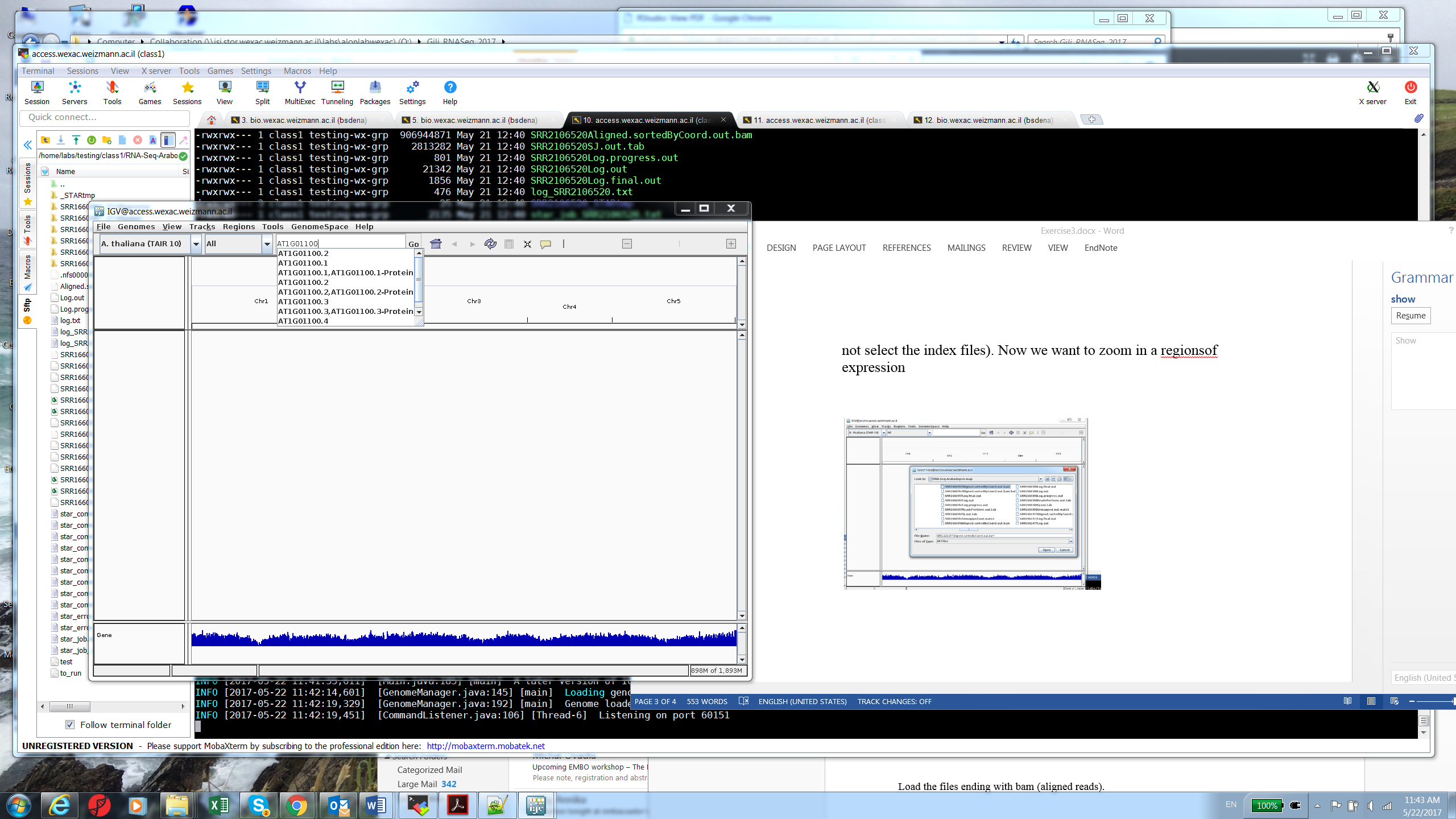


Load a bam file by selecting it using the File menu.

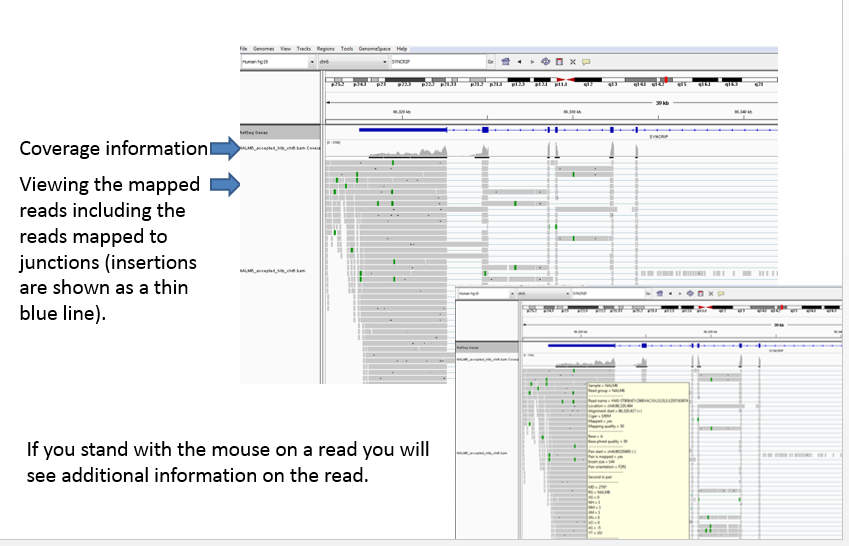
Note: select only a bam file that has also an index file (\*.bai). Do not select the index files.



Now we want to zoom into a region of expression such as gene AT1G01100. Enter the gene name – see arrow below.



Following is an explanation of the features of IGV:



**Question: Use the CIGAR information of the reads spanning the last exons of AT1G01100 (by placing the mouse on such a read) to find the last intron size in bases.**

* 1. Run indexing on all bam files.

Good job. You completed this exercise.

### Supplementary

### STAR manual

### <https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf>

### IGV user guide

### http://software.broadinstitute.org/software/igv/userguide

### Basic Linux commands:

man (command) ...... shows help on a specific command

ls ................. show directory, in alphabetical order

logout ............. logs off system

mkdir .............. make a directory

rmdir .............. remove directory (rm -r to delete folders with files)

rm ................. remove files

cd ................. change current directory

more .............. views a file, pausing every screenful

grep ............... search for a string in a file

head ............... show the first few lines of a file

tail ............... show the last few lines of a file

df ................. shows disk space available on the system

du ................. shows how much disk space is being used up by folders

chmod .............. changes permissions on a file

cut ............... print selected parts of lines

cp ............... copy file

mv ............... move file

wc –l .............. print the number of lines

sort ............. sort lines of text files