

Integrative Genomics Viewer (IGV)

IGV is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. Datasets supported by IGV include NGS sequence data, array-based data and genomic annotations.

1. Parsing files for IGV input

First, download the prepared files via secure copy onto your Desktop as you will use IGV from your local computer:

```
# You should not be on logrus for this command.
# cd into Desktop first before you do the following.

scp -P 44111 -r <username>@gateway.training.ncgr.org:/home/elavelle/\
IGV_files ./
```

Downloading the entire reference genome and full set of alignments can take a long time and considerably slow down IGV when loaded. The "IGV_files" directory contains only data for the first chromosome. As the files transfer, we'll walkthrough the process to parsing down the original files within logrus.

1.1 Genome parsing

```
# Navigate to the location of the human reference genome

cd ~/DGE_Virtual/human_reference
```

We need to activate the "samtools" environment:

```
# samtools is part of the conda environment designated "bio" on our server

source activate samtools
```

We'll make use of this software package to subset our genome file:

```
# will automatically create an index for the original in order to create the output

samtools faidx GRCh38.p12.genome.fa chr1 > GRCh38.p12.chr1.fa
```

1.2 Annotation parsing

While we're in this directory, let's subset the annotation file also. This is a little trickier, but the command should look mostly familiar:

```
awk '$1 ~ "^#" || $1 == "chr1" {print $0}' \  
gencode.v29.chr_patch_hapl_scaff.annotation.gff3 > \  
gencode.v29.chr1.annotation.gff3
```

- The ~ within the awk command indicates the following is a regular expression.
- The ^ within the parentheses matches the beginning of a line.
- The || is equivalent to "or"
- Whereas numbers following \$ denote the field, \$0 is a stand-in for the whole row

1.3 Alignment parsing

Move up one directory, then into the directory containing the alignments generated by HISAT: Headers and chromosome alignments can be parsed in a similar fashion to what was one with the genome annotation file:

```
awk '$1 ~ "^@" || $3 == "chr1" {print $0}' \  
759_7-p5-2.sam > \  
759_7-p5-2.chr1
```

```
awk '$1 ~ "^@" || $3 == "chr1" {print $0}' \  
Scram_1-3.sam > \  
Scram_1-3.chr1
```

1.3 Alignment sorting

We'll use samtools again to sort the .sam files and convert them to .bam files:

```
# When this is finished, the precursor files can and should be deleted  
for file in *.chr1; do samtools sort -@ 2 ${file} -o ${file}.sorted.bam; done
```

1.4 Alignment indexing

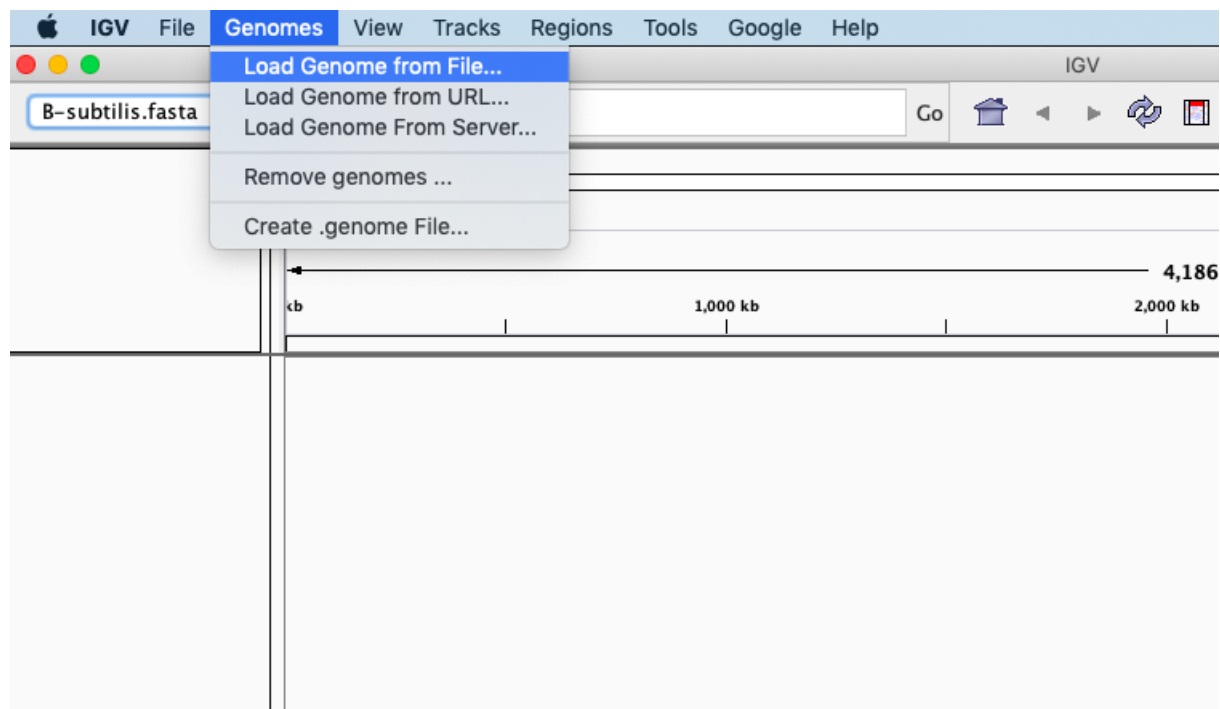
Lastly, we want to index the .bam files:

```
for file in *.sorted.bam; do samtools index ${file}; done
```

Now we have all the files needed for IGV (that you first downloaded from logrus). There are four different types of files we made: a .fa genome file, a .gff annotation file, .bam alignment files and their associated .bai index files.

2. Loading files into IGV

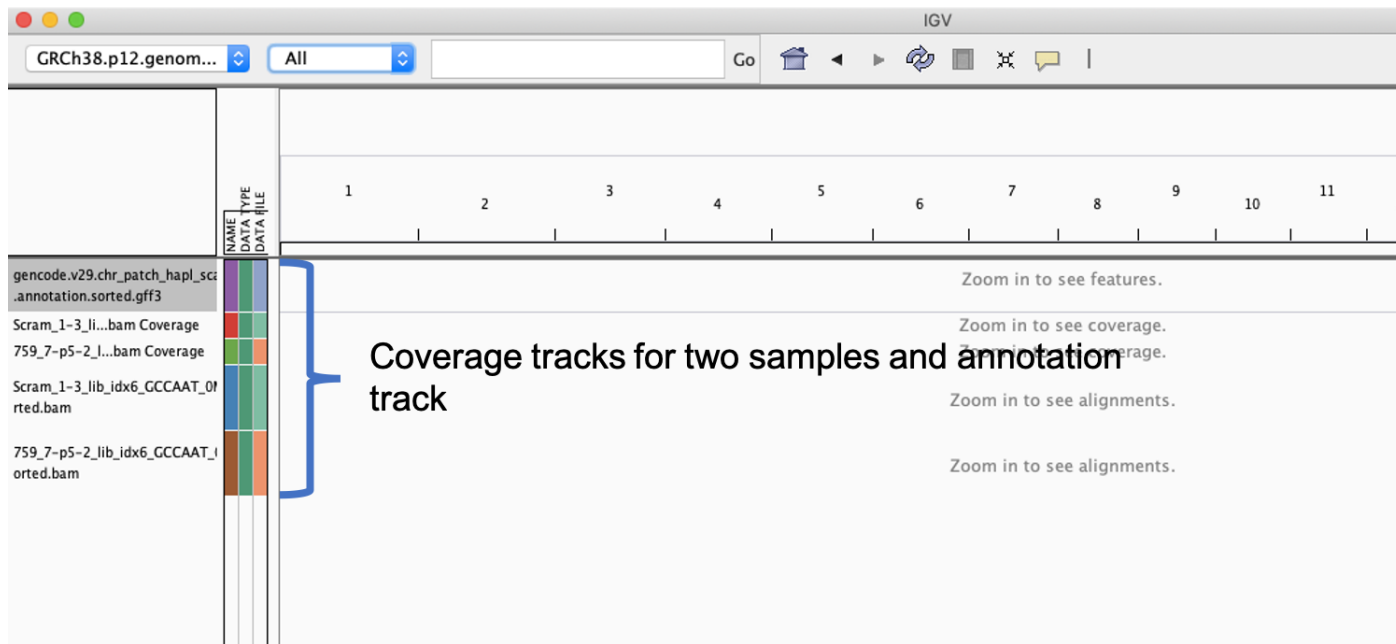
Once you have opened IGV, first upload the reference fasta file from your Desktop (GRCh38.p12.chr1.fa). IGV will have many preloaded genomes to choose from. Your organism of interest may be available, but you have to ensure the version used for your analysis matches the preloaded version of genome in IGV database.



You should now see the appropriate filename in the scroll-down menu. Next comes the annotation file, which must be sorted prior to loading. Use Tools → Run igvtools to do this.

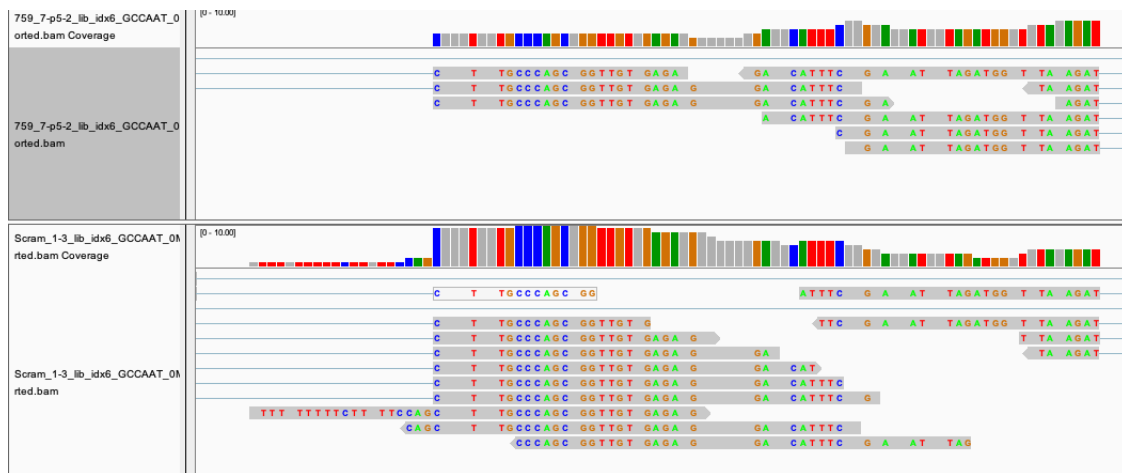
Now we can load the sorted .gff file via File → Load from File and agree to the prompt to have the file indexed.

Files can also be loaded by simply dragging them into the interface. Try this with the .bam files. The .bai files are required as well, but it won't explicitly demand them; as long as they share a directory with the alignment files, they will be read.

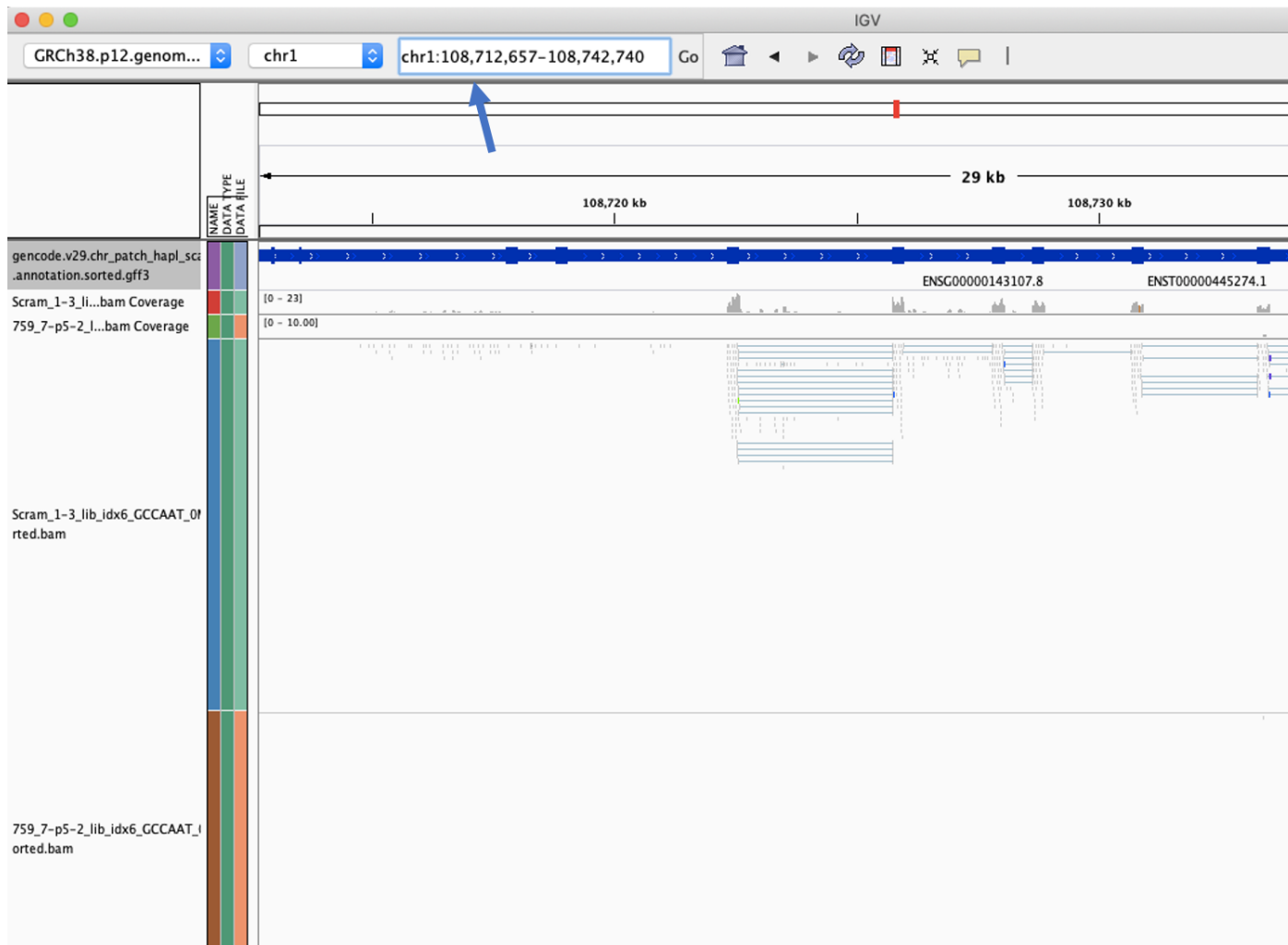


3. Navigating in IGV

Try navigating with some of the zoom options. Look close enough in and you can view the reads on a nucleotide scale.



Let's focus in on a gene with a very low padj value, FNDC7. Type in the annotation ID (ENSG00000143107.8) or chromosome indexes (chr 1:108712657-108742740) to jump to that location. We can find these by searching the annotation file.



Exercise:

Investigate read pile-up for another significant DEG on chromosome 1, LPTM5

5. Using IGV to view variants

There is a subfolder in the directory we downloaded titled "bacterium files". Try loading the files therein into IGV. It should look like as below. You can select a chromosome from the drop-down menu and view details of variants by hovering over the bars.



More tutorials:

Investigating alternative splicing patterns using IGV

Advanced alternative splicing and sashimi plots

Genome viewing