IGV and USCS Data Visualization

Maria Jose Rodriguez, Maria Fernanda Garcia and Maria Fernanda Requena

2023-03-01

P. Chabaudi Genome

We will use IGV to visualize data for Plasmodium chabaudi AS (P.chabaudi)

- P. Chabaudi is a malaria mouse pathogen.
- Using the files in: /mnt/Timina/bioinfoII/data/IGV/

We will visualize RNA-seq data.

Commands

```
#Copy files
cd /mnt/Timina/bioinfoII/data/
cp -r IGV/ /mnt/Timina/bioinfoII/user/practica_3
cd /mnt/Timina/bioinfoII/user/practica_3/IGV/
```

In the first step we copied the files nedded and we procedeed creating a Quality Control Report with fastqc:

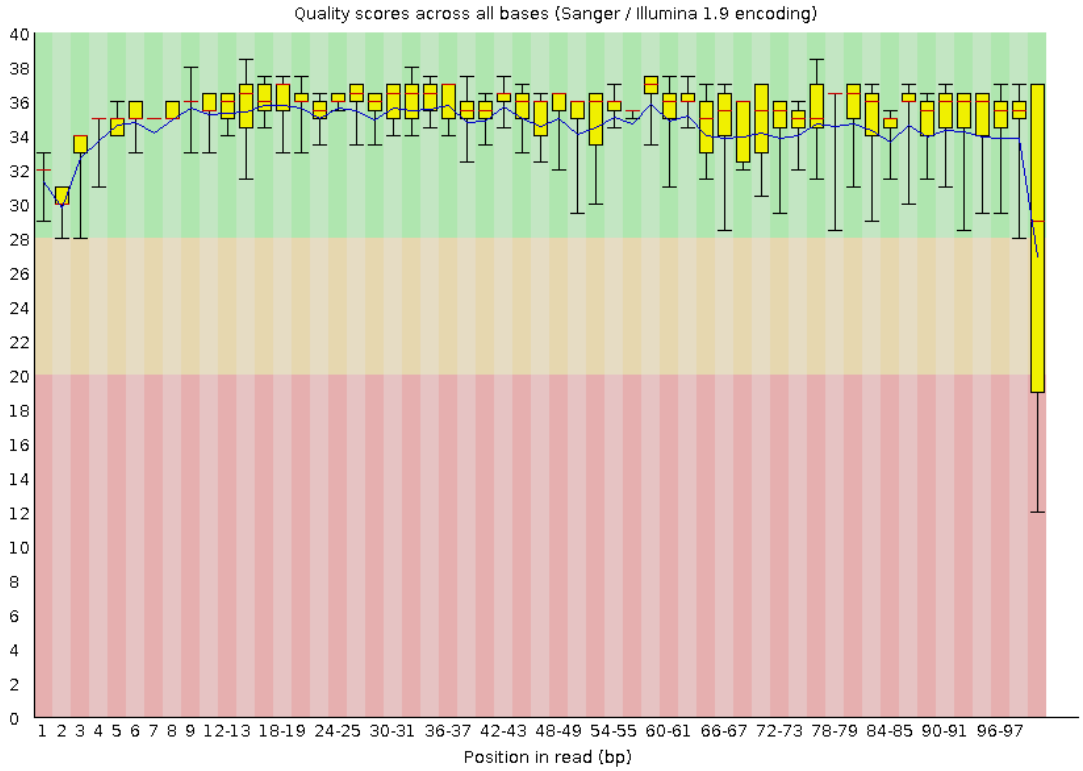
Quality Control

Commands

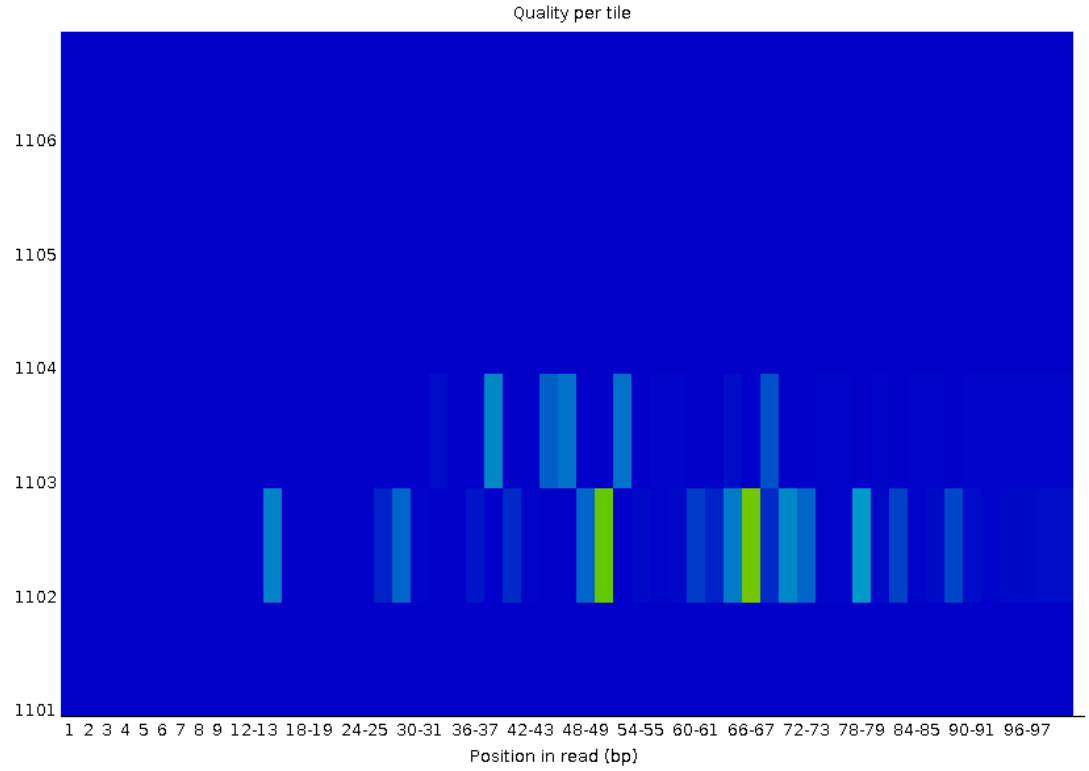
```
# Analize the data quality with fastqc command
fastqc MT1_1.fastq
fastqc MT1_2.fastq
fastqc MT2_1.fastq
fastqc MT2_2.fastq
```

Result

Per base sequence quality

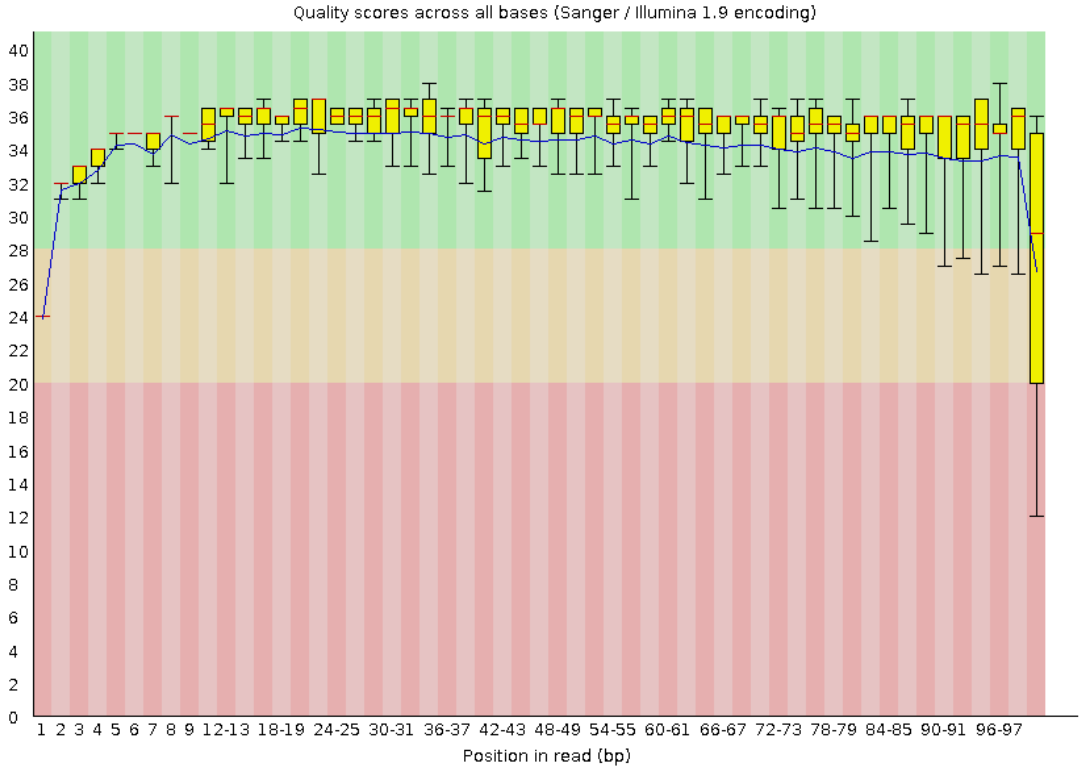


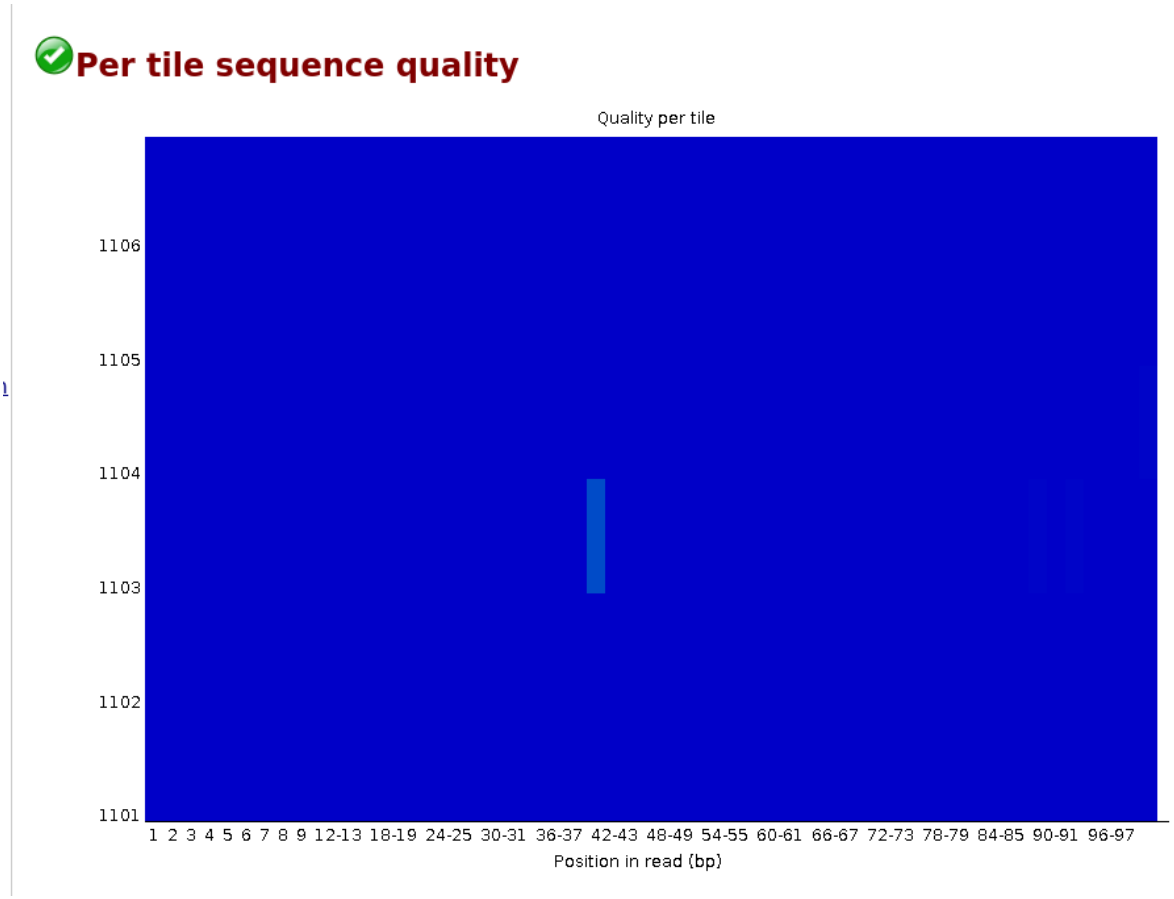
Per tile sequence quality



(MT1.1 images)

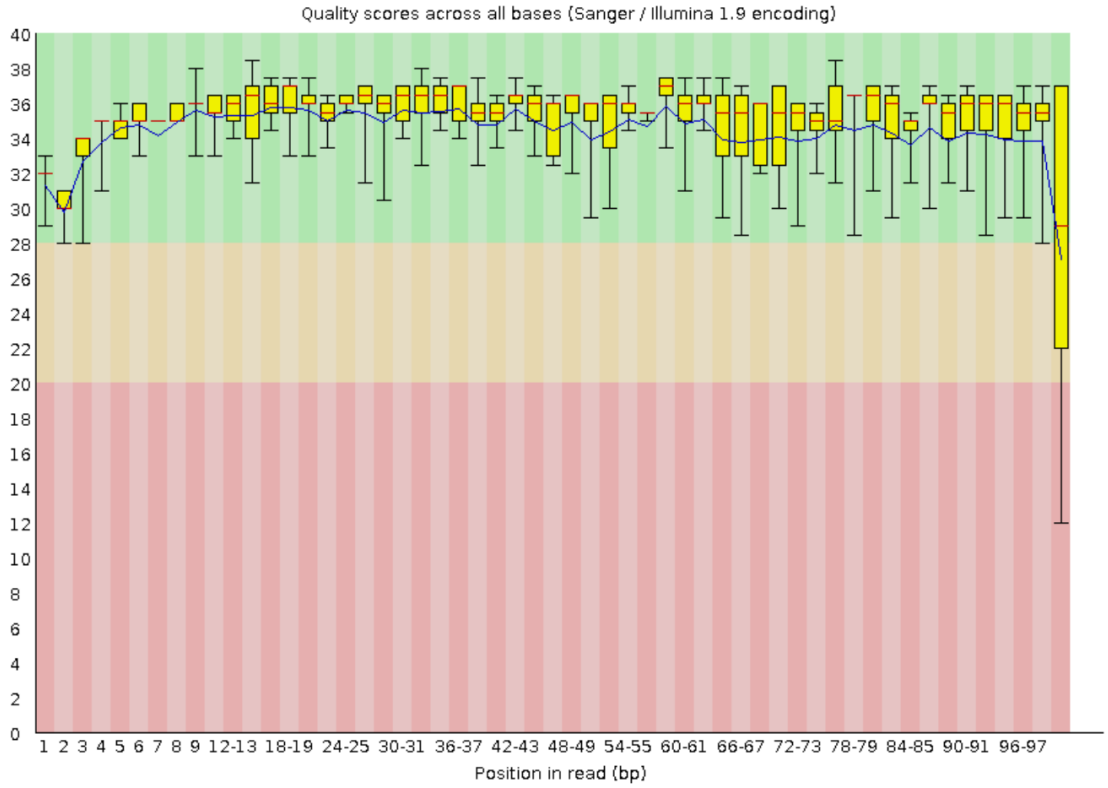
Per base sequence quality

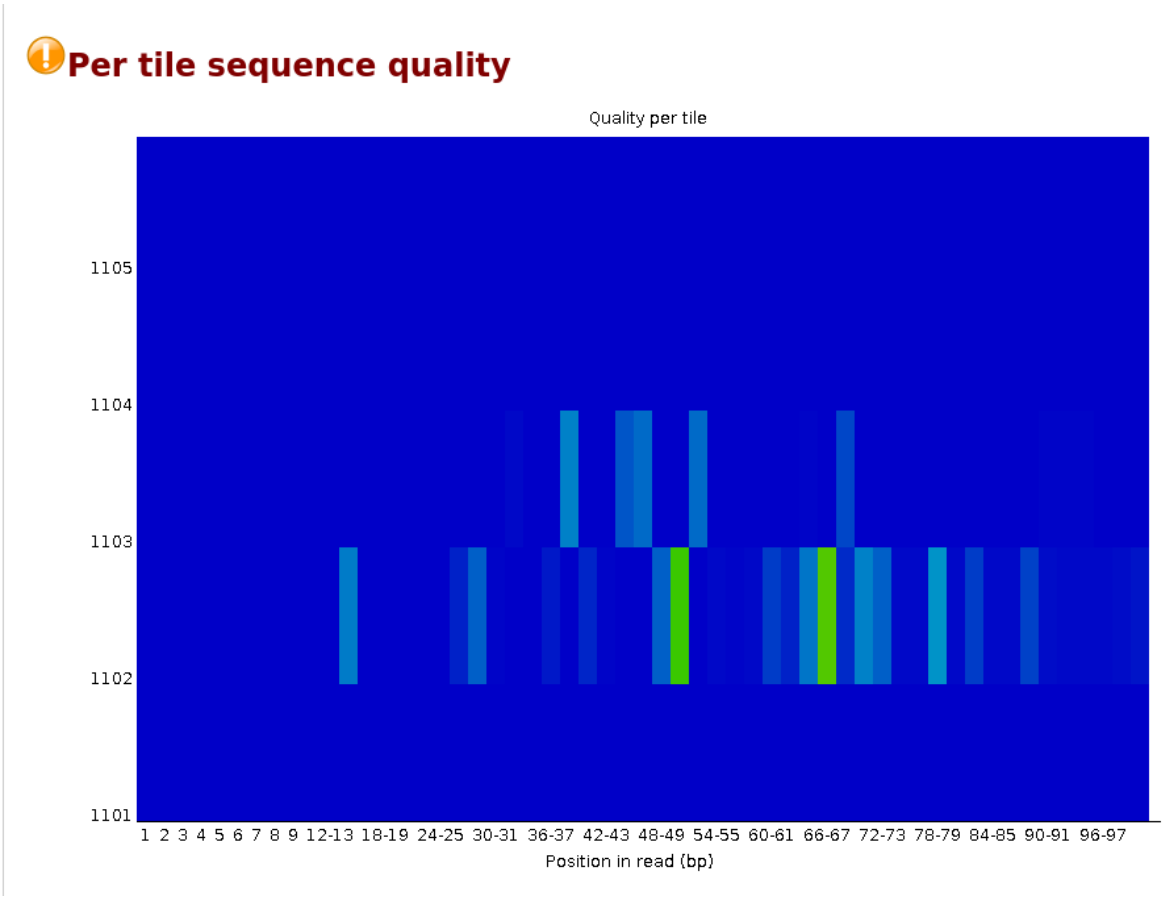



Per tile sequence quality
Quality per tile
1106
1105
1104
1103
1102
1101
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97
Position in read (bp)

(MT1.2 images)

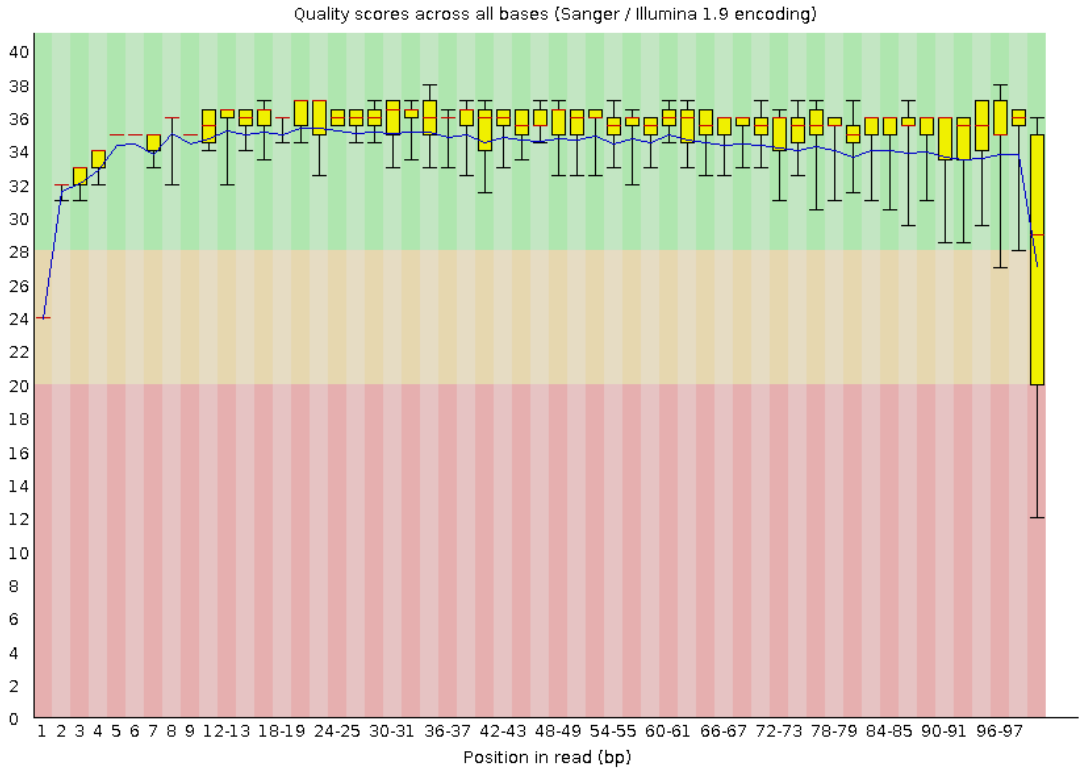
Per base sequence quality

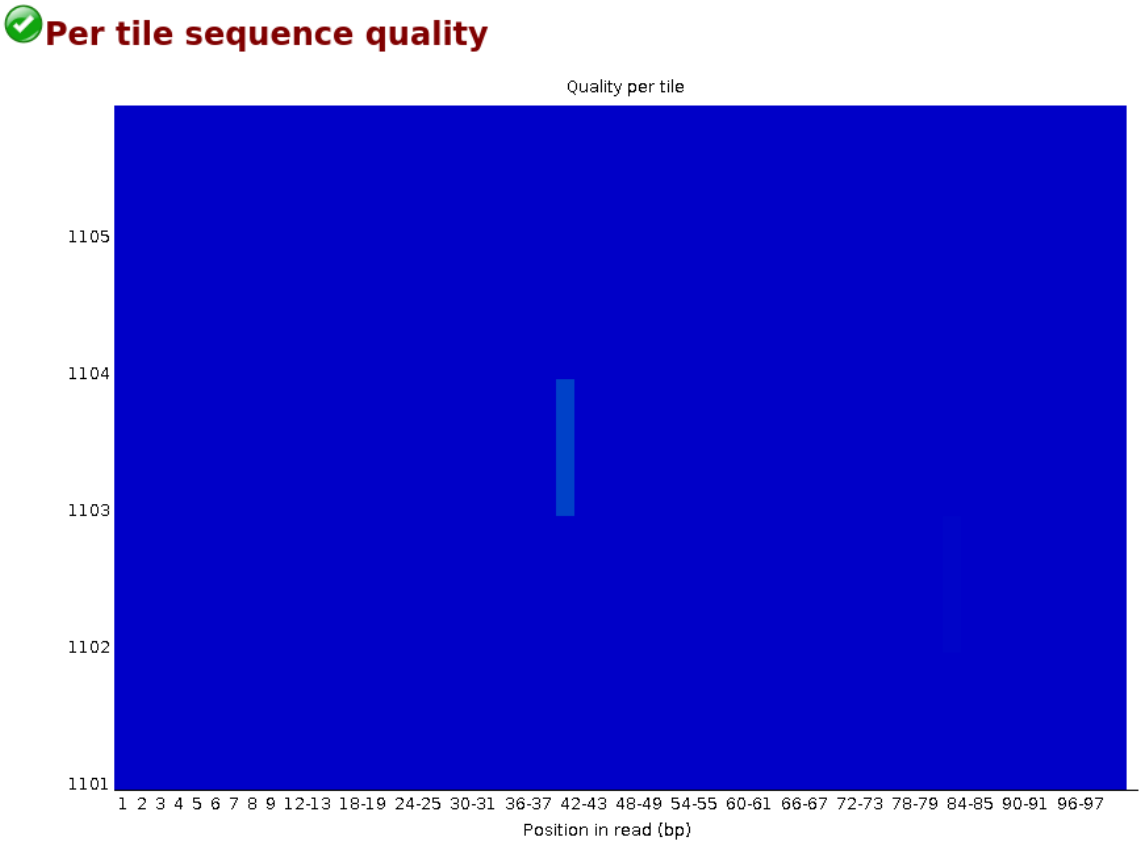



Per tile sequence quality
Quality per tile
1105
1104
1103
1102
1101
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97
Position in read (bp)

(MT2.1 images)

Per base sequence quality





(MT2.2 images)

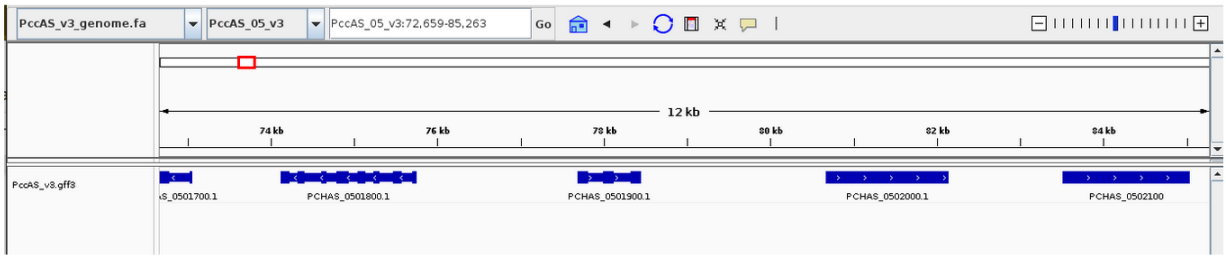
According to the data, these reads were sequenced with Ilumuna and so we hae a very good score in quality from all of the files of the mouse patogen. All of them (if we take into account the Per base Seequence Quality images) seem to be in the correct area of quality and the only message that reflects a non excelllent score is in the last postions of the sequence, which is a normal event in Ilimuna Sequencing because the primers start to stick off. NOw, if we take a look to the Per Tile Images we can see tiles have a good quality and those which are worst are still realiable. Thus, we can conlcude our 4 experimental data is trustworthy becasue of the sequence quality we acquire from the fastq report.

Using IGV

- Loading annotations for the genome

The anotation file is the PccAS_v3.gff3 in a General Feature Format (GFF) that is a tab-delimited text file which holds information of any feature that can be applied to a nucleic acid or protein sequence.

The IGV browser displays a graphic interphasse that allows us to see reads trough different positions of the genome and its characteristics. For example, it displays a view with the loci marked in blue and their names.



If we zoom in or out we can observe more characteristics of the genome and annotations done by IGV that describe better a location in the sequence, as well as a loci position.

- Loadign the alignment file of an RNA-seq experiment

We can't analyze the data in a .fastq format, first we have to index a genome reference and align the raw data

With the alignment we will obtain a .sam or .bam file that we can use

We have information from a paired end read, that is why two files per sample exist.

Aligning the data

We create the index for the reference genome and align the data using the command hisat2 to take into account isoforms.

Commands

```
# Create a directory to store the index mkdir index_h2
hisat2-build index_h2/PccAS_v3_genome.fa PccAS_v3_hisat2.idx

#Alignmemt

hisat2 --max-intronlen 10000 -x index_h2/PccAS_v3_hisat2.idx -1 MT1_1.fastq -2 MT1_2.fastq -S MT1.sam

hisat2 --max-intronlen 10000 -x index_h2/PccAS_v3_hisat2.idx -1 MT2_1.fastq -2 MT2_2.fastq -S MT2.sam
```

The second command corresponds to the First raw data alignment, we align both files in the command because they are paired end reads, obtain a .sam file. The thrid command is the second raw datra alignment.

- Use samtools view to convert .sam files with -b that indicates the output file is a .bam

```
samtools view -b MT1.sam > MT1.bam
samtools view -b MT2.sam > MT2.bam
```

We need to index the .bam files but first we must sort the alignments by leftmost coordinates with command -samtools sort-. The resulting files indexed will end in .bam.bai.

```
samtools sort MT1.bam -o MT1.sorted.bam
samtools sort MT1.bam -o MT1.sorted.bam
samtools index MT1.sorted.bam
samtools index MT2.sorted.bam
```

IGV Visualization

Commands

- Download the reference genome from LAVIS cluster to our laptops with rsync command.

```
rsync -rptuvl user@dna.lavis.unam.mx:/DIRECTORY/IGV/PccAS_v3_genome.fa
```

We run this command in our computer terminal to have access to the file *PccAS_v3_genome.fa* and donwload it.

- Download the sorted alignment files and their index

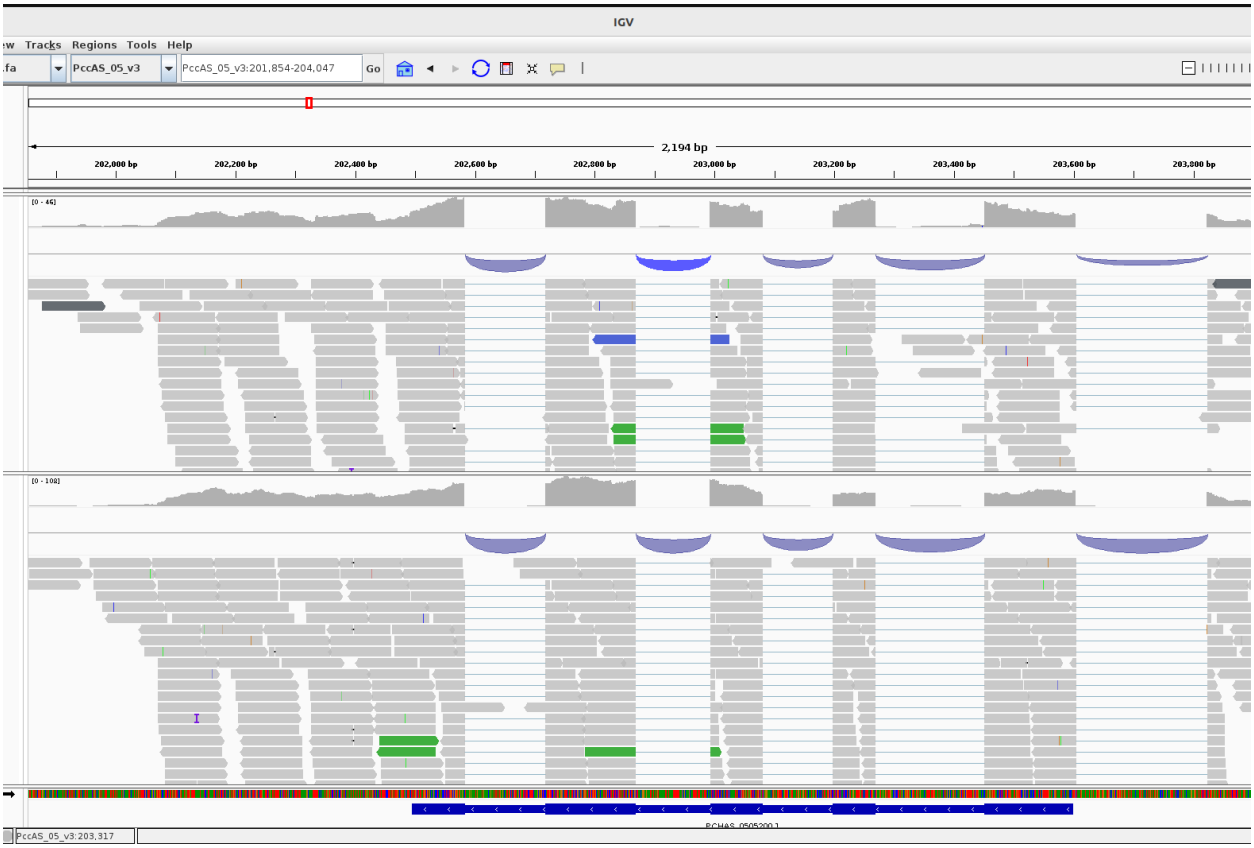
```
rsync -rptuvl user@dna.lavis.unam.mx:/DIRECTORY/IGV/MT1.sorted.bam .
rsync -rptuvl user@dna.lavis.unam.mx:/mnt/Timina/bioinfoII/user/practica_3/IGV/MT1.sorted.bam.bai .
rsync -rptuvl user@dna.lavis.unam.mx:/mnt/Timina/bioinfoII/user/practica_3/IGV/MT2.sorted.bam .
rsync -rptuvl user@dna.lavis.unam.mx:/mnt/Timina/bioinfoII/user/practica_3/IGV/MT2.sorted.bam.bai .
```

- Now we can load the BAM files to IGV.

After uploading to IGV both of the files (MT1.sorted.bam and MT2.sorted.bam) we can observe different locations in the genomes. If needed, we can export the figure seen with going to the *File* section in the app and clicking where it says *"Save images as .png"* or *"Save images as SVG"*

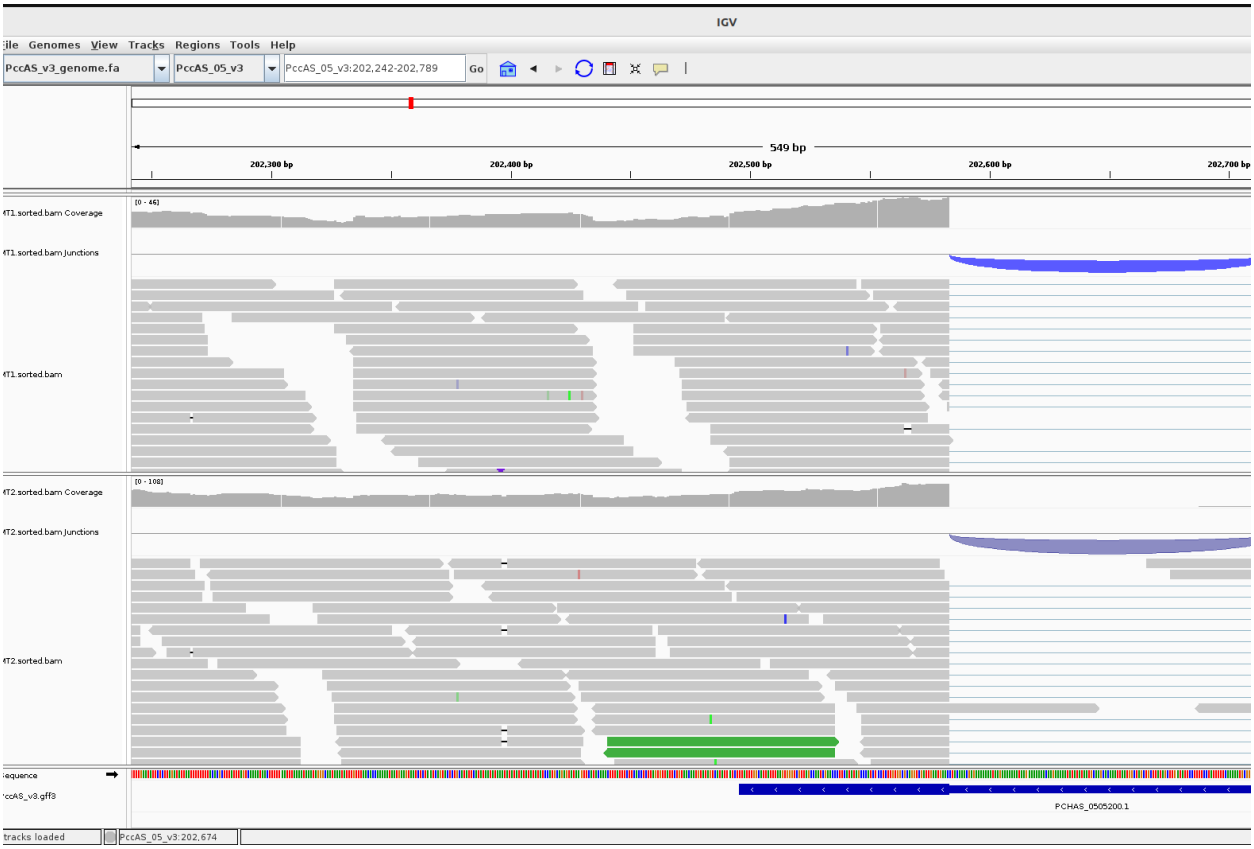
Visualize loci: PCHAS_0505200 and PCHAS_1409500

Loci: PCHAS_0505200



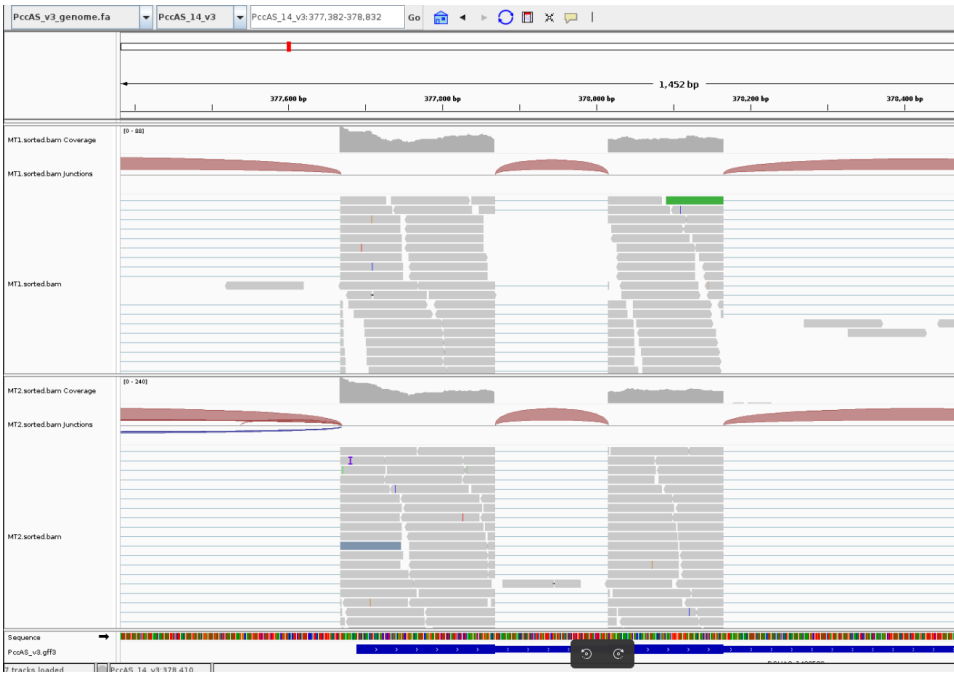
If we give a quick view in the position of the genome we can observe strands that correspond to the reads of the genome. As we can see, almost every read base in the strand is in gray color which describes that almost every base matches the reference genome. However, we can also visualize different colors, as well as aqua, dark blue and green in some reads. These 3 colors could correspond to different analysis, as well as Color alignment by pair orientation and Color alignment by insert size. In the case of this figure and in purpose of its analysis we decided to analyze it with the first option.

If we make a zoom to the reads we can distinguish more features and colors and analyze this data:



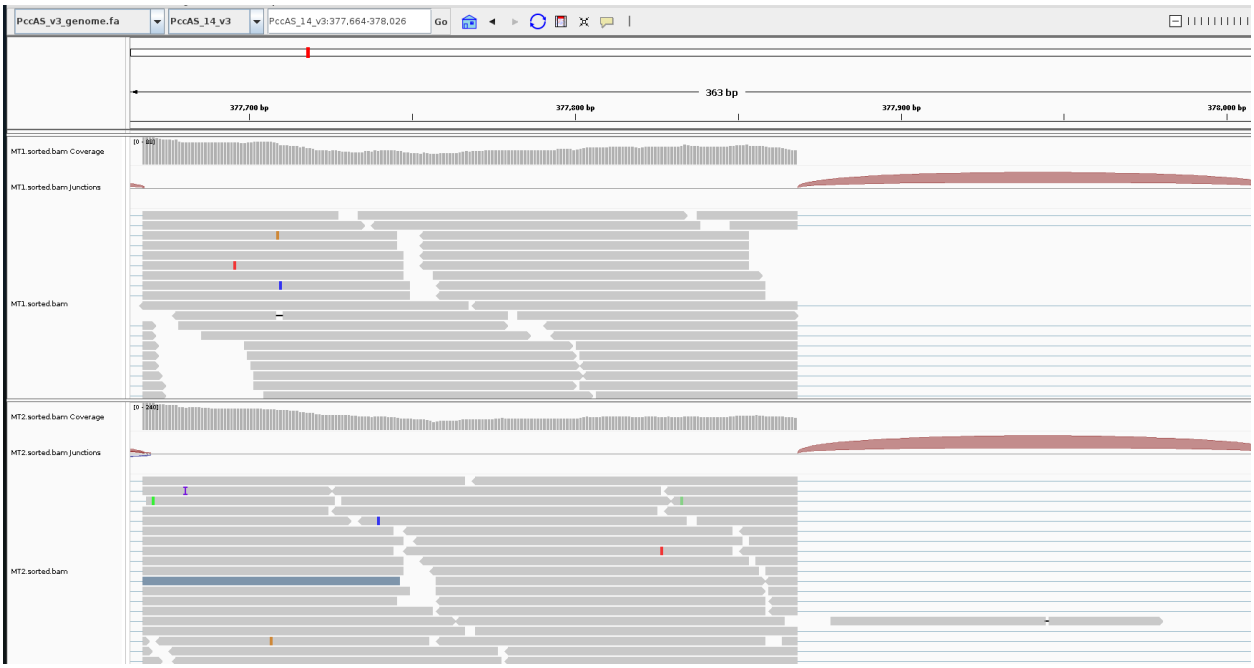
Color alignment by pair orientation This interpretation takes into account the orientation of the paired reads and it detects structural events as inversions, duplications and translocations. In our picture we can observe **blue segments** (as lines) that correspond to an inversion. Inversions are when a large section of DNA is reversed compared with the reference genome. When the inversion is in the paired end there is a lot of variance compared with the reference and the color aqua is displayed in IGV, which is the case of the picture above. Also, if we focus in the **green segments** we can infer a translocation on the same chromosome or a Tandem Duplication. In the same way, we can observe some kind of **blue arches** on top of the reading that corresponds to splice junctions. These arcs connect alignment blocks from a single read and its color blue references the - strand.

Loci PCHAS_1409500



Color alignment by pair orientation

Color alignment by insert size This type of interpretation helps us to recognize structural variants as well as deletions, insertions and inter-chromosomal rearrangements. As we can see in the picture, there are red blocks that analyzed correspond to a possible deletion and blue boxes that represent possible insertions.



Now, if we make a zoom in the IGV application we further visualize more fragments of colors in the lectures and even the letter I in purple color. This letter represents an insertion and deletions are marked with a black hyphen. We can both observe these symbols in the picture. Furthermore, we could also visualize red/pink splice junctions that reflected the + strand with a height and thickness proportional to the depth of read coverage.

Mouse Genome

We will work using the .sam file generated for the mouse data in the Genome_Aligment repository ([]). *m_al_mem.sam*

Commands

```
samtools view -b m_al_mem.sam > mouse.bam
```

Creation of bam from a sam format trough samtools view tool.

Result

```
mouse.bam
```

Sort so we can index:

```
#Sort
samtools sort mouse.bam -o mouse.sorted.bam

#Index
samtools index mouse.sorted.bam
```

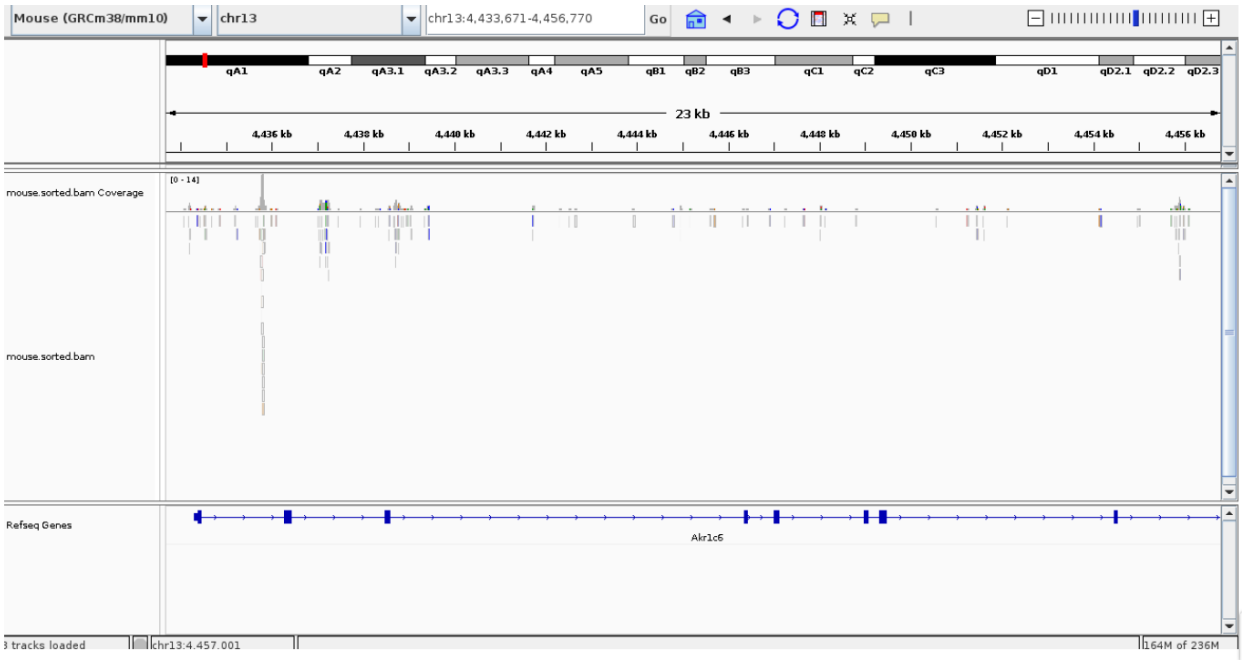
- Download sorted.bam and the index

```
rsync -rptuvl user@dna.lavis.unam.mx:/DIRECTORY/IGV/Mouse/mouse.sorted.bam .
rsync -rptuvl user@dna.lavis.unam.mx:/DIRECTORY/IGV/Mouse/mouse.sorted.bam.bai .
```

IGV Visualization

- Download the sorted alignment files and their index

```
rsync -rptuvl user@dna.lavis.unam.mx:/DIRECTORY/IGV/mouse.sorted.bam .
rsync -rptuvl user@dna.lavis.unam.mx:/DIRECTORY/IGV/mouse.sorted.bam.bai .
```



Usign IGV to visualize the data for mouse ChIP-seq we generated we can observe is a liver data set. We can visualize some interesting loci but we selected AKR1C6 because it is expressed in the liver, our reads align well and we have several variants but not too many to be a problem, it seems more like an error in the samples. Through this point of view we can not say much more of the analysis of the picture but that it has an expanded track which means there are no so much overlapping features as transcripts of the gene.

UCSC Genome Browser visualization

An URL must be used to access the data in this case.

We can upload our data formatted in bigbed, barChart, bigChain, bigGenePred, bigInteract, bigLolly, bigMaf, bigWig, BAM, barChar, VCF, BED, BED detail, BedGraph, broadPeak, CRAM,
GFF, GTF, hic, Interact, MAF, Narrow Peak, Personal Genome SNP, PSL, or WIG

The bigWig format is for display of dense, continuous data that will be displayed as a graph. The main advantage of the bigWig files is that only the portions of the files needed to display a particular region are transferred, so for large data sets bigWig is considerably faster than regular WIG files.

- Creating a BigWig using deepTools

```
bamCoverage -b reads.bam -o coverage.bw
```

Commands

- Required module deeptools

```
module load deeptools
```

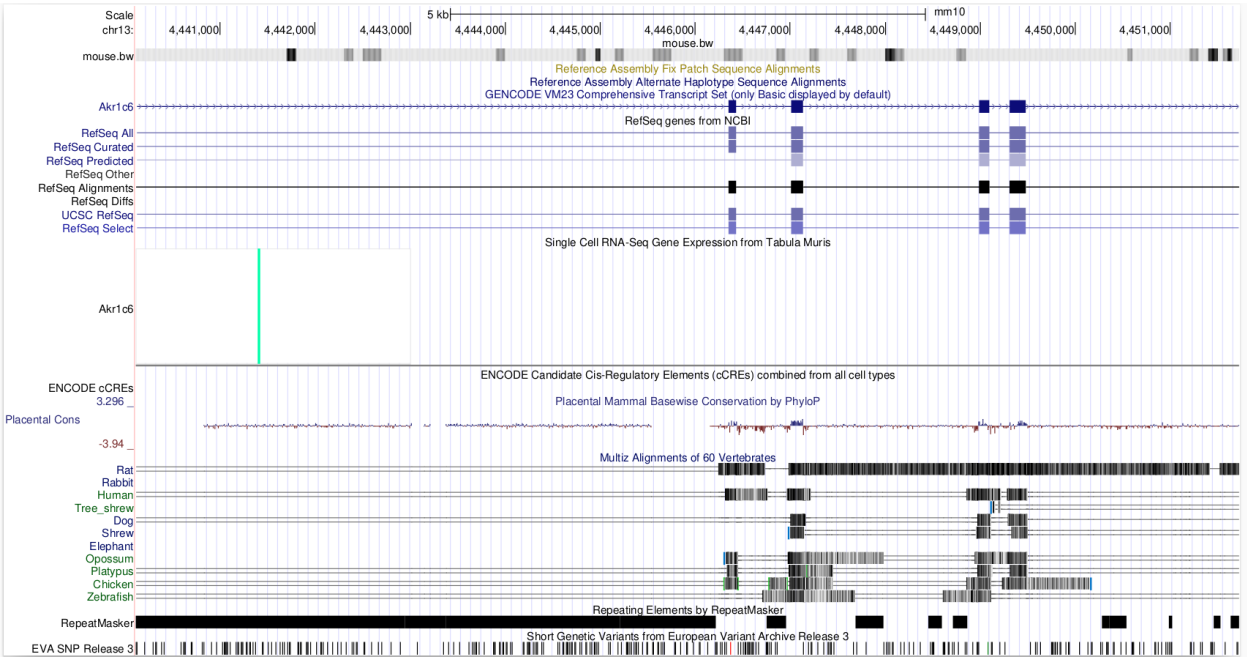
From deeptools we use the command -bamCoverage- to convert a .bam file into a .bw (bigwig) format. -b is to indicate that we will process a -bam file and -o is for the output file name.

```
bamCoverage -b mouse.sorted.bam -o mouse.bw
```

- Download the .bw file to upload it on CYVERSE.

```
rsync -rptuvl user@dna.lavis.unam.mx:/DIRECTORY/IGV/Mouse/mouse.bw .
```

Visualization



Visualization Notes

The browser provides a annotation track beneath the genome positions to help us visualize correlations between other information we would like to obtain. This traks include a display of data, genes, genes predictions, expressed sequence sequence evidence, regulation, mRNA, etc. We can see that the genome is in a horizontal orientation with the shortest arm of the chromosome in the left side.

Finally, it is also possible to upload data sets to compare our results obtained through the tools available in the traks options.

Bibliography

- (Akr1c6 aldo-keto reductase family 1, member C6 [Mus musculus (house mouse)] - Gene - NCBI, s/f) Akr1c6 aldo-keto reductase family 1, member C6 [Mus musculus (house mouse)] - Gene - NCBI. (s/f). Nih.gov. Recovered the 28 Feb 2023 from, National Library of Medicne
- (Genome Browser Training, s/f) Genome Browser Training. (s/f). Ucsc.edu. Recovered the 1st March 2023, from UCSC Help
- (Práctica 1 - Control de calidad con FastQC - b22carcaBMS, s/f) Práctica 1 - Control de calidad con FastQC - b22carcaBMS. (s/f). Google.com. (Genome Browser Training, s/f) Genome Browser Training. (s/f). Ucsc.edu. (Genome Browser Training, s/f) Recovered the 26 Feb 2023, from Práctica2
- Simplyy Publishing. (2019). Bigwig: Simple Blank Lined Notebook Journal. Independently Published. Recovered the 28 Feb 2023 from, IGV
- (IGV user guide, s/f) IGV user guide. (s/f). Broadinstitute.org. Recovered the 28 Feb 2023 from, IGV User Guide