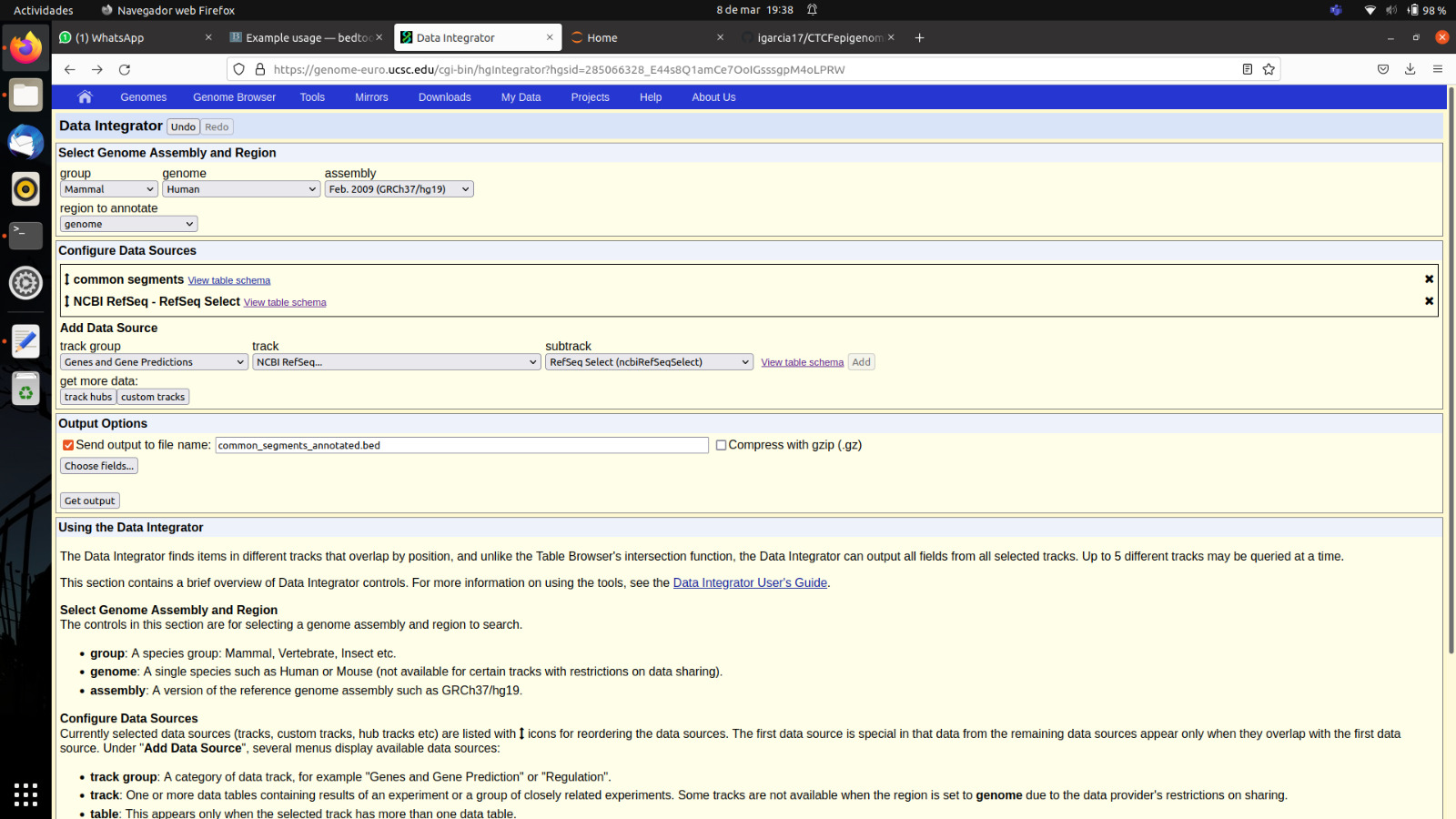
**2. Annotation**

First we annotated the segments with UCSC Genome Browser Data integrator tool.

The settings where as follow:



We realised with the following commands that the obtained bed file had more lines.

wc -l common\_segments\_annotated.bed

8066

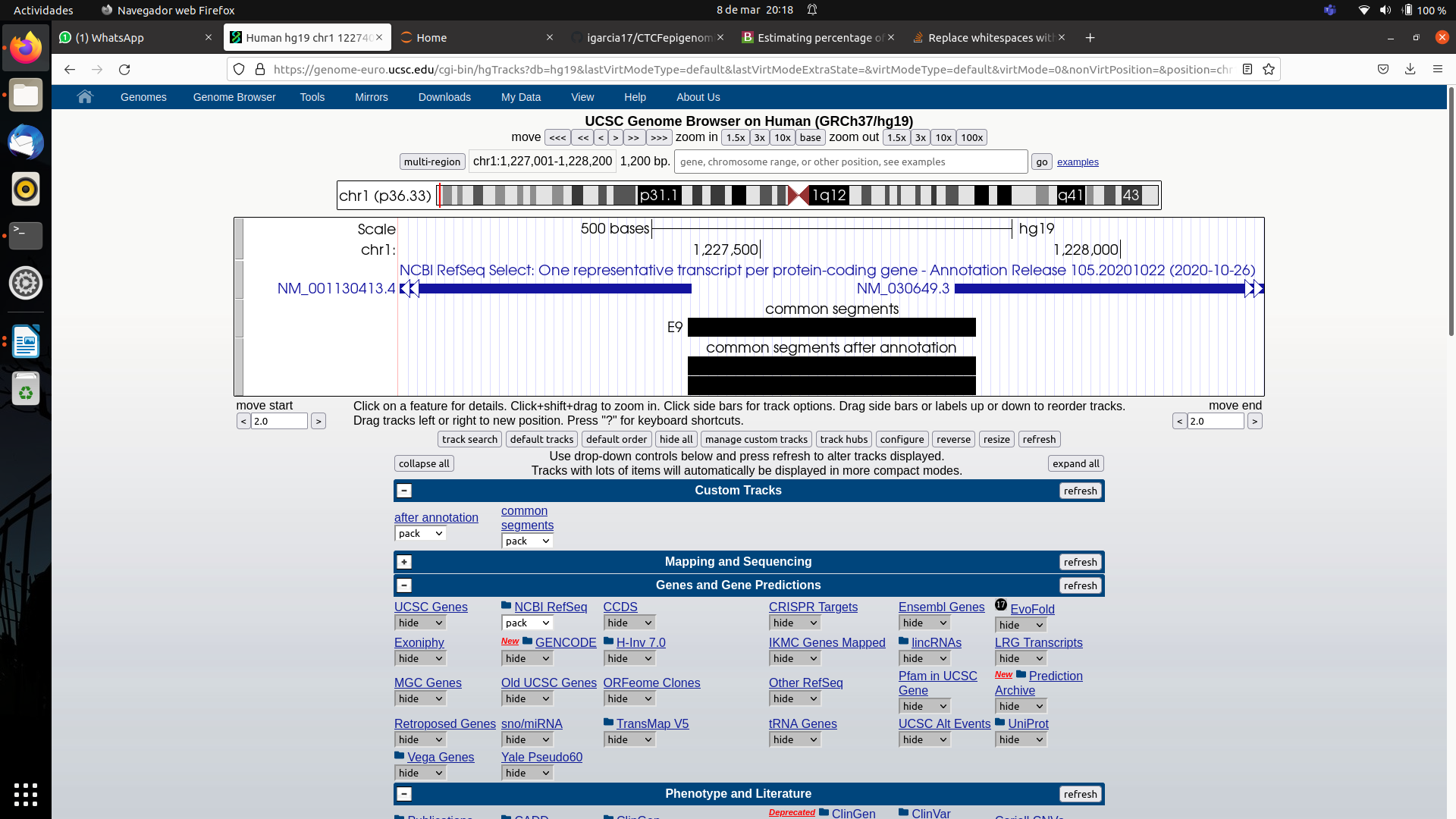
wc -l common\_segments.bed

8022

This is because some segments are in overlapping regions of genes, such as the case of chr1 1227400 1227800. For this image, the annotated file was modified to be read by UCSC.

A single Select transcript is chosen as representative for each protein-coding gene

Same command as for Ex 3 to get percentage

covered exons:

80035

1.522%

Complete genes:

2089126

39.73%

|  |  |
| --- | --- |
| (2) Percentage of coding genes with CTCF segments | 39.73% |
| (2) Percentage of exons of coding genes with CTCF segments | 1.52% |
| (2) Percentage of introns of coding genes with CTCF segments | 35.70% |
| (2) Percentage of 3 UTR of coding genes with CTCF segments | 2.36% |
| (2) Percentage of 5 UTR of coding genes with CTCF segments | 0.14% |
| (2) Percentage of upstream 200 bp of coding genes with CTCF segments | 0.22% |
| (3) Percentage of overlapping with DNaseI | 25.88% |
| (6) Percentage of overlapping with hypermethylated regions | 80.87% |
| (6) Percentage of overlapping with hypomethylated regions | 2.55% |

**(3)**

First: let’s know how many bases are there in the common segments file:

cat my\_results/common\_segments.bed | awk -F'\t' 'BEGIN{SUM=0}{ SUM+=$3-$2 }END{print SUM}'

5257400

Then we are going to intersect the DNAse peaks with this file. To do so, we directly use the intersection option of bedtools.

bedtools intersect -a my\_results/common\_segments.bed -b input/CD14\_monocytesDukeDNaseSeq.pk > my\_results/covered\_DNAsepeaks.bed

And calculate the covered bases again:

cat my\_results/covered\_DNAsepeaks.bed | awk -F'\t' 'BEGIN{SUM=0}{ SUM+=$3-$2 }END{print SUM}'

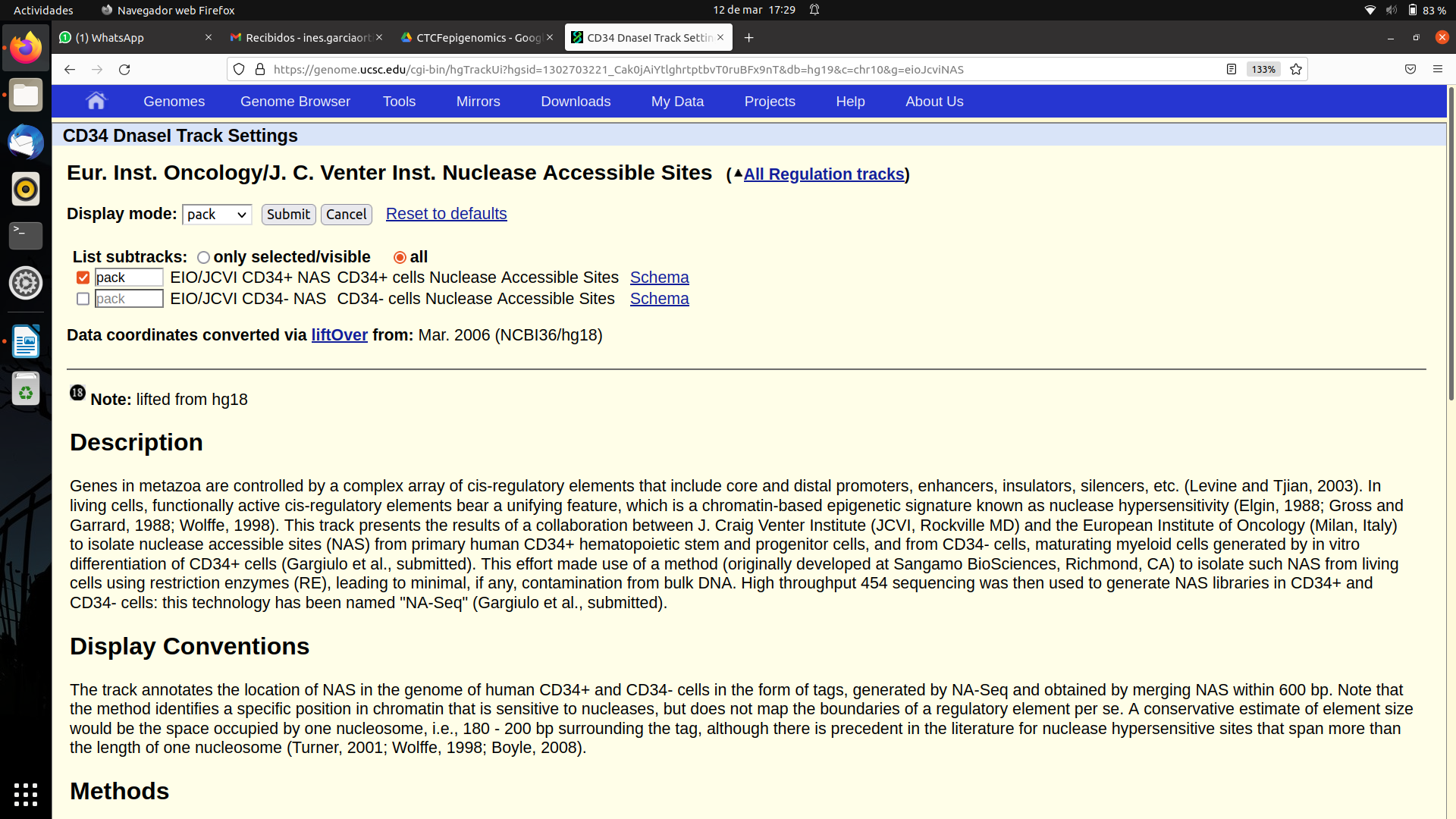
1360517

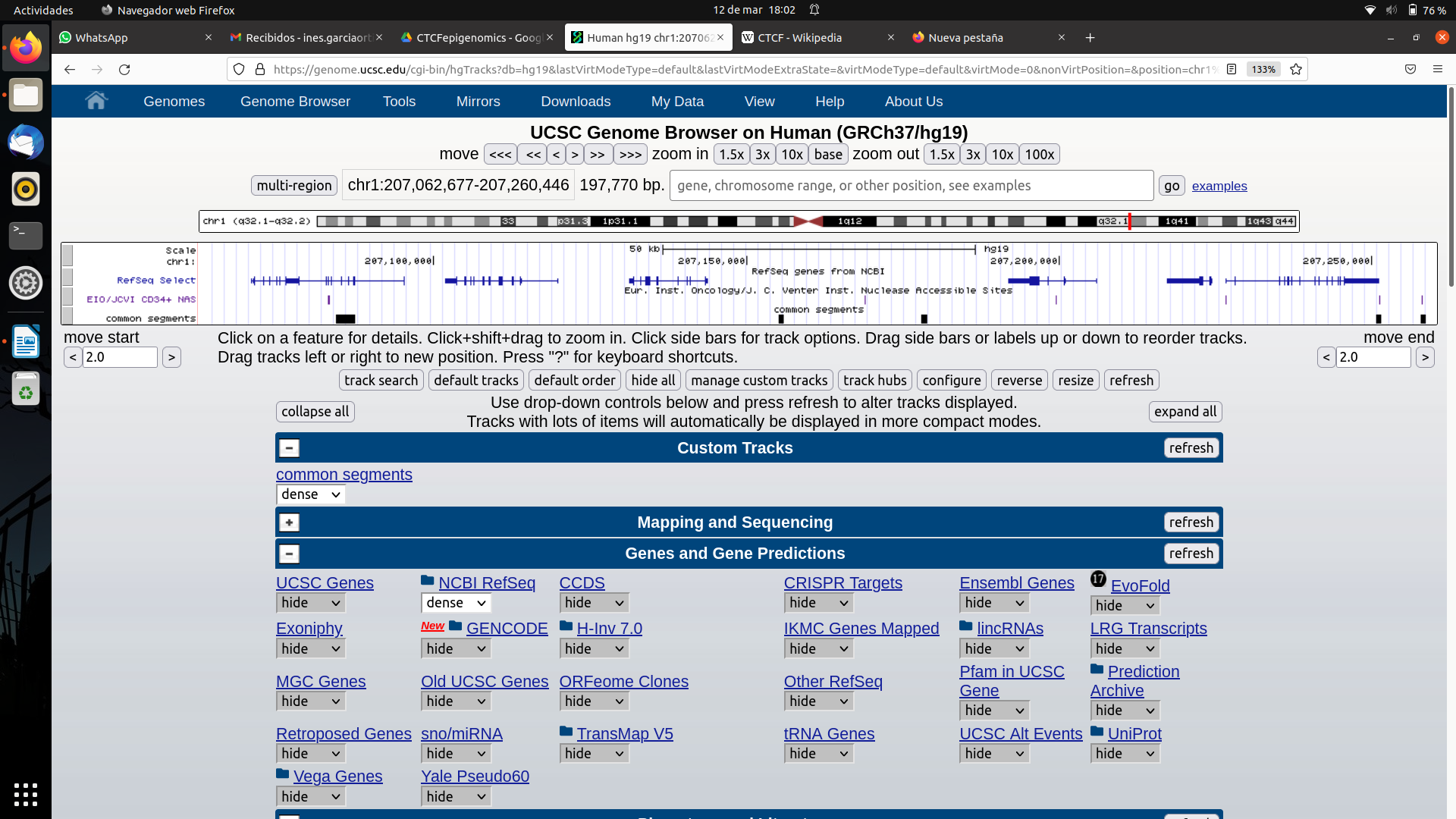
We do the division between both and the result is that 25.88% of the segments overlap with a DNAseI peak.

(4)

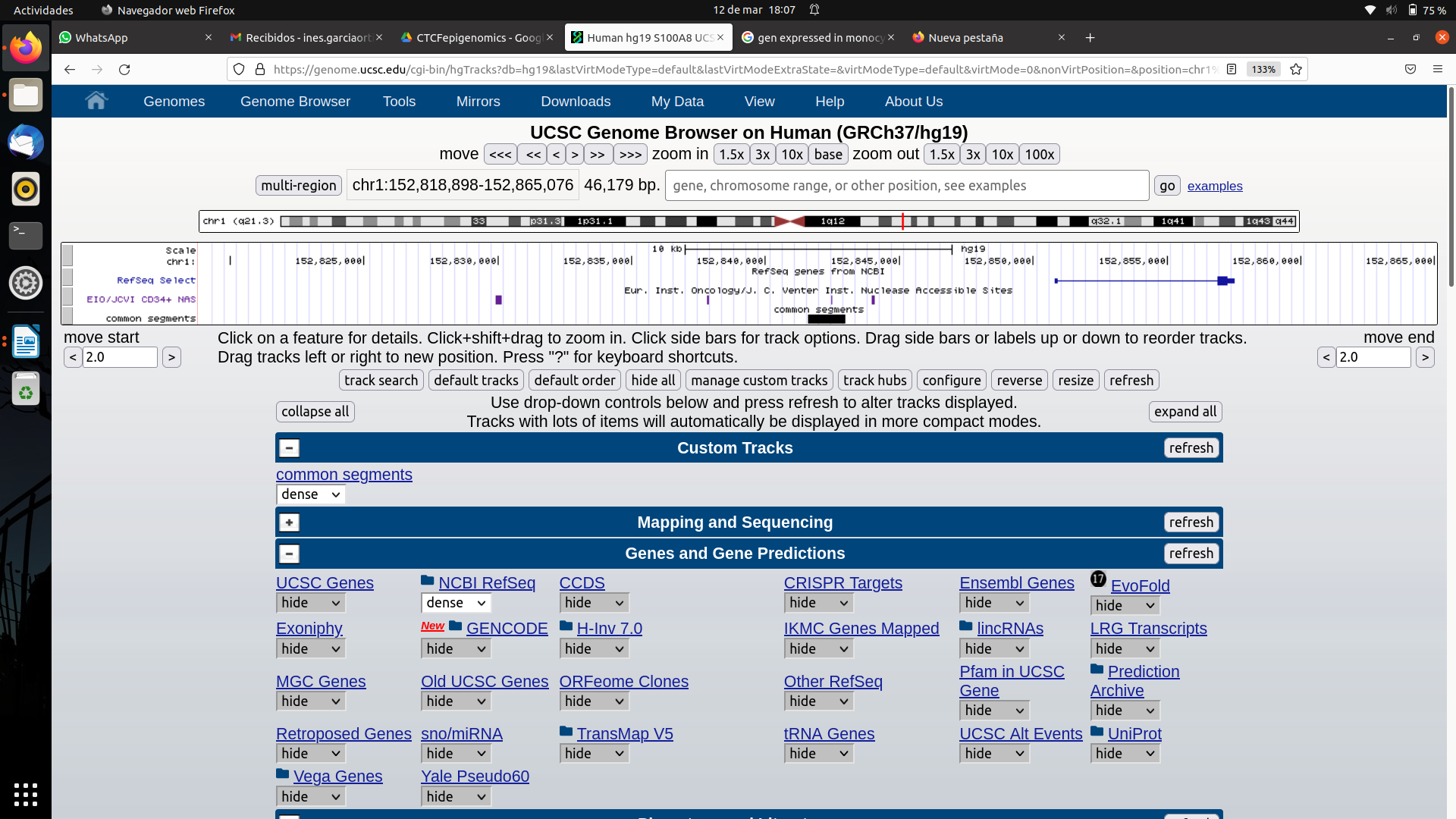
4) Visualize (and show) a region of the genome in the UCSC browser where you see at least one of your segments (upload the track generated by ChromHMM) and the DNaseI in all ENCODE blood cell types.

CD34+ cells (enriched in hematopoietic stem and progenitor cells)





En la imagen se ven casos comunes de CTCF.



In this other case we see an overlapping between DNAseI and CTCF; it is near S**100A8, highly expressed in monocytes (segun google).**

**(6)**

For percentage of overlapping with hypermethylated regions:

4251502

80.87%

With hypomethylated regions?

cat hypoBED |awk -F'\t' 'BEGIN{SUM=0}{ SUM+=$3-$2 }END{print SUM}'

134285

2.55%