Erika Maldonado-Rosado

27 March 2019

Lab 7

BINF 6203

**Introduction**

For this lab we need to map short sequences using three different types of software to help us get the end results. All the analyses come back to mapping and their outcomes depends on the accuracy of the mapping process. However, non-uniquely mapping reads are counted, and on the quality of the reference annotation used to interpret the mapped reads. In the Methods, I will go over all the tools used for this lab.

**Method**

In this lab I used Ubuntu has my Linux terminal to use bowtie2, samtools, and bcftools. Bowetie2 is a read mapping and analysis software, which can be used directly or in conjuction with other packages that add on features and post-alignment analysis. Samtools is one of three repositories which we used for reading, writing, editing, indexing, and viewing SAM/BAM/CRAM format. Lastly, BCFtools, which is another separate repository offered by samtools, was used for reading and writing BCF2/VCF/gVCF files and calling, filtering, and summarizing SNP and short indel sequence variants. For this lab we first got 12 chloroplast sequence that was stored in the NCBI database but was distributed by UNC: Charlotte with the help of Jennifer Weller using an Ion Torrent PGM instrument. We brought the sequences onto our terminal using Fastq-dump, and trimmed it using Trimmomatic. The untrimmed and trimmed sequences were viewed on FastQC, which was on my windows browser. After all of our sequences are trimmed and back on our terminal, we then used Bowties to index the reference genome, then we use that index to align with one of our unpaired sequence alignments to create a sam file. We then convert our sam to a bam using the view -b, then sort that bam file using the sort command. After we sort, we then index that sorted file. We then use bcftools to convert it from a bam to a vcf format. We then move it from our terminal to our windows screen to view our vcf files on IGV. Sadly, I was unable to do this and only viewed the bam files on IGV, but I did view each vcf file to see how it aligned with the reference genome as you will see below.

**Results**

**How does the genomic divergence between these cultivars show up in terms of individual sequence variations in the chloroplast genome?**

Each sequence covers or uncovers a new area that aligns with the reference chloroplast genome.

**Are variations between the genomes primarily SNPs, or small insertions or deletions? We’ll go over the VCF file format in class so you can see how to tell.**

From what I saw in the file was that there was insertions or duplications. There was some areas where they noticed duplications and deleted it, but then that means there is a missing SNP, this in turn can tell us why there are small gaps in the genome.

**How are the individual SNPs and/or indels dispersed in the chloroplast genome? Do they disproportionately affect particular genes or regions?**

They are dispersed with, but there seems to be gaps throughout the sequences at different positions.

**Can you see any evidence of heterogeneity in the chloroplast genome sequence in these samples (i.e. more than one variant at the same site, with respect to the reference genome?)**

I saw small differences compared to the reference genome, but not enough to question its origin.

**CODE**

fastq-dump -A SRR17637\*\*

cp SRR17637\*\*.fastq /mnt/c/Users/erika/Onedrive/Desktop

java -jar ~/Trimmomatic-0.36/trimmomatic-0.36.jar SE -threads 3 SRR17637\*\*.fastq SRR7\*\*.fastq LEADING:20 TRAILING:20 CROP:**250-350** **HEADCROP:5**

**\*\*\*For the Crop was the range that I used for all the files, On file 773 was the only one that I had to used headcrop due to one of the sequences falling below the average.**

bowtie2 -x NC\_007898 -U SRR7\*\*.fastq -S SRR7\*\*.sam

samtools view -b SRR7\*\*.sam > SRR7\*\*.bam | samtools sort SRR7\*\*.bam > SRR7\*\*.sort.bam

samtools index SRR7\*\*.sort.bam

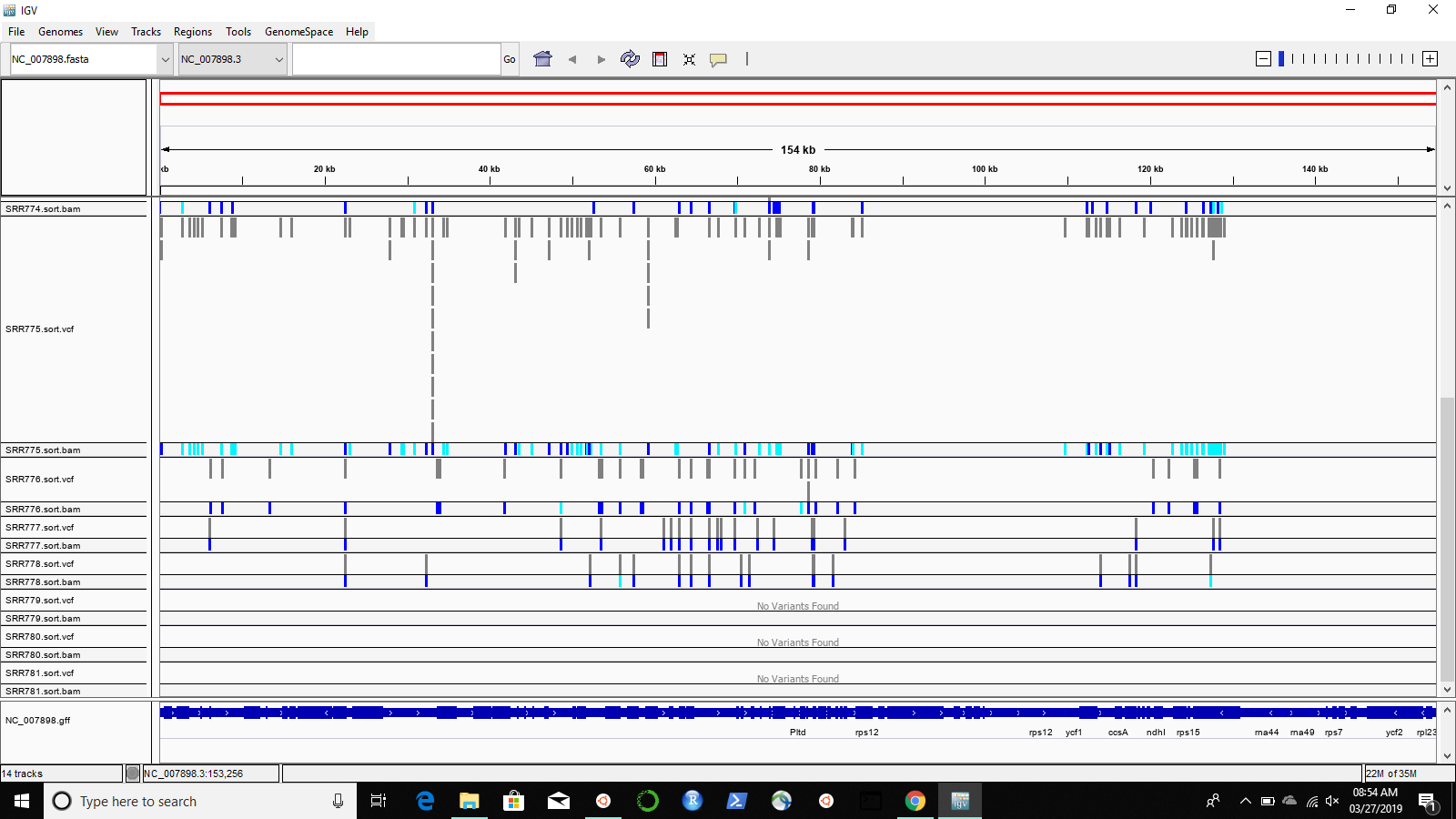
bcftools mpileup -f NC\_007898.fasta SRR7\*\*.sort.bam | bcftools call -c -v > SRR781.sort.vcf

cp SRR7\*\*.sort.vcf /mnt/c/Users/erika/Onedrive/Desktop

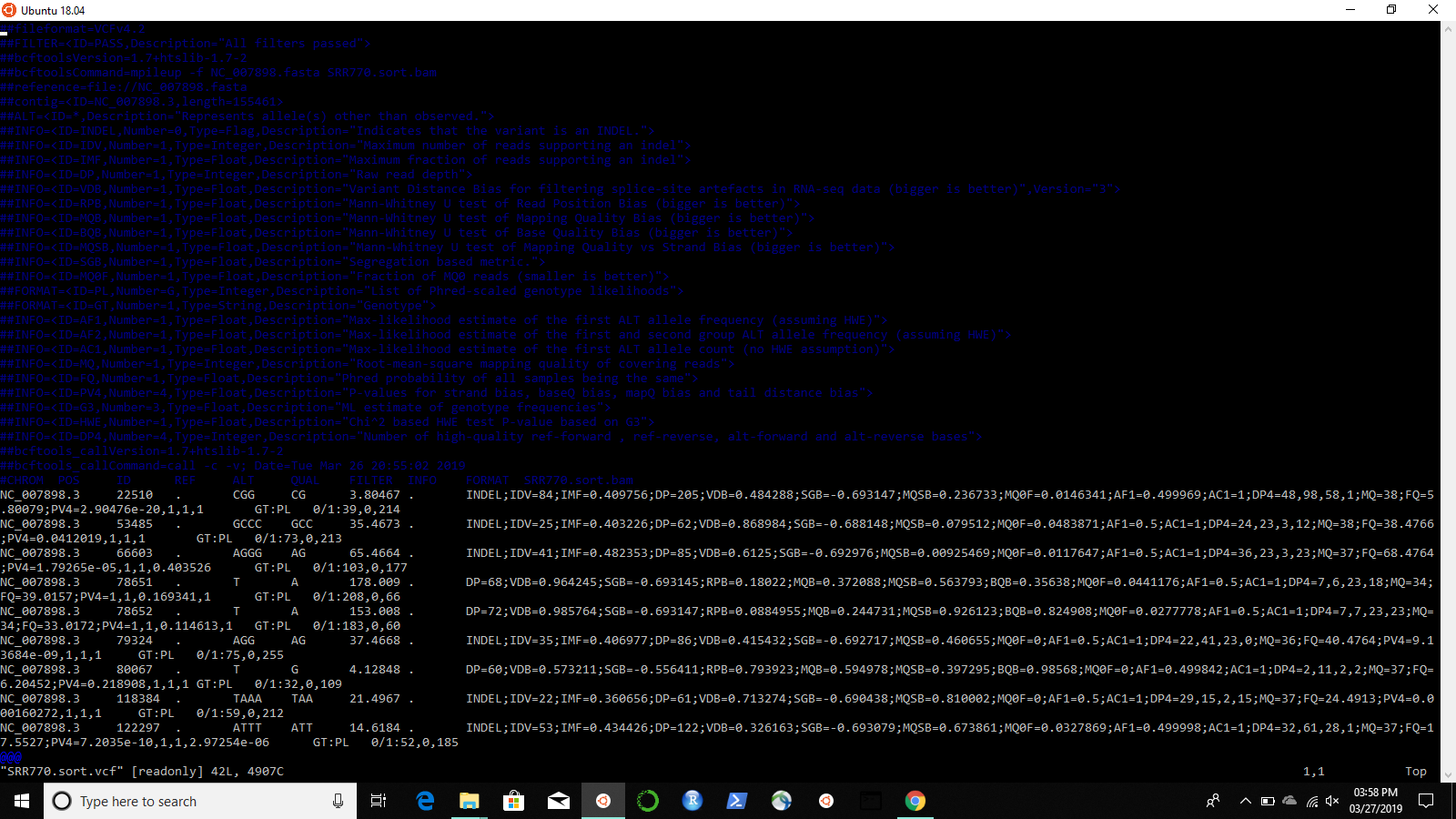
**IGV Screenshots**

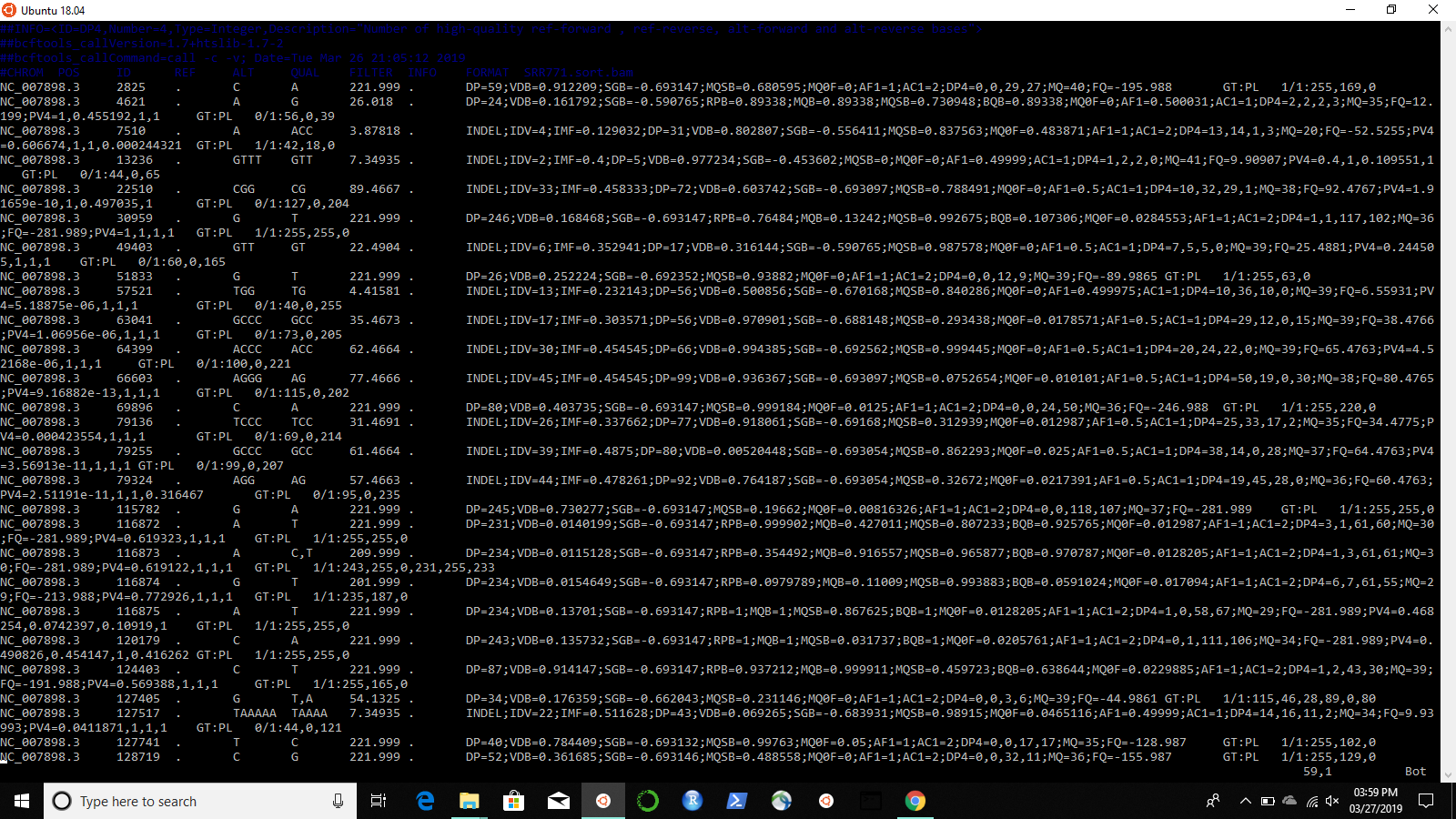
**Image 1&2**

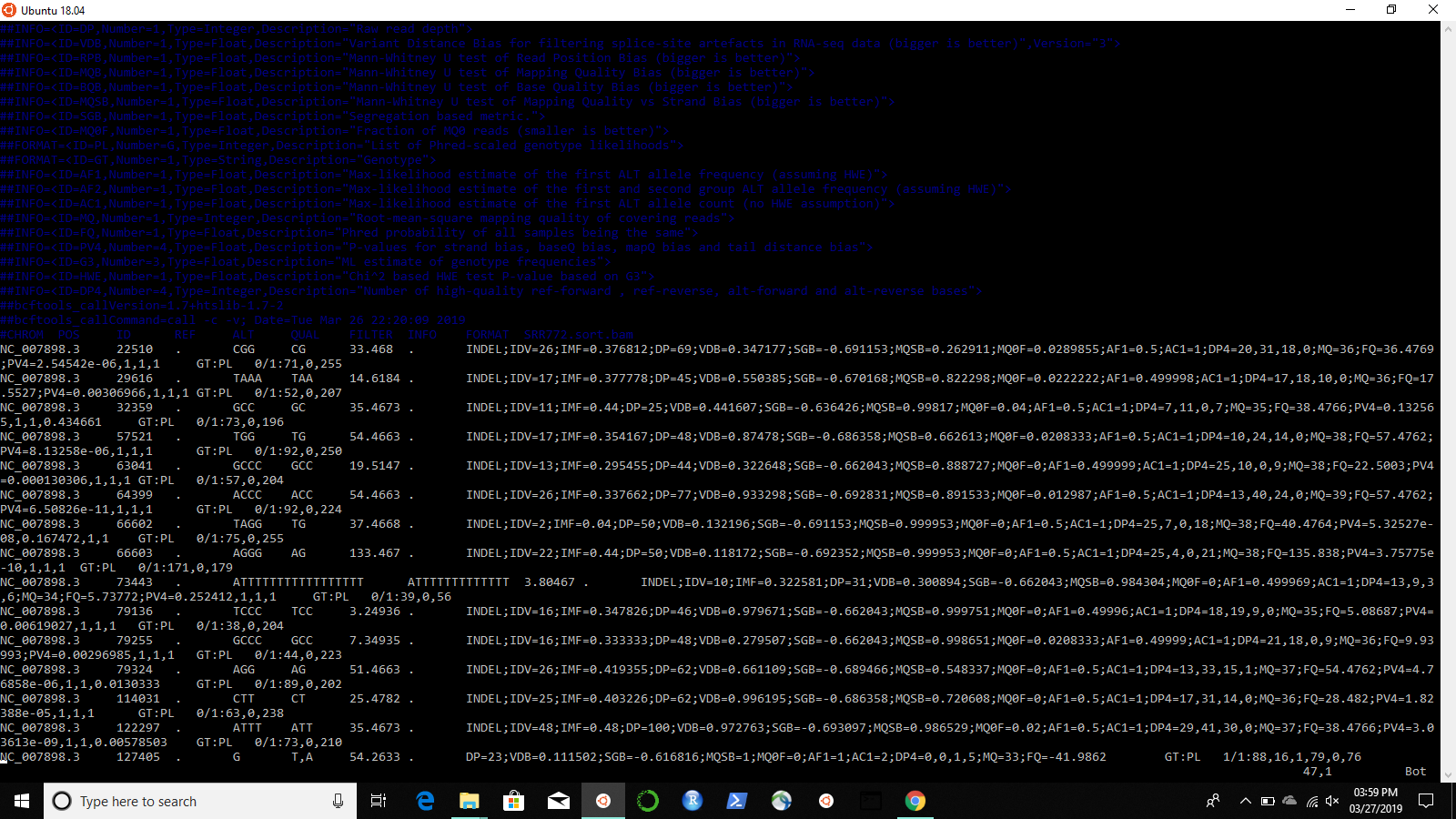
****

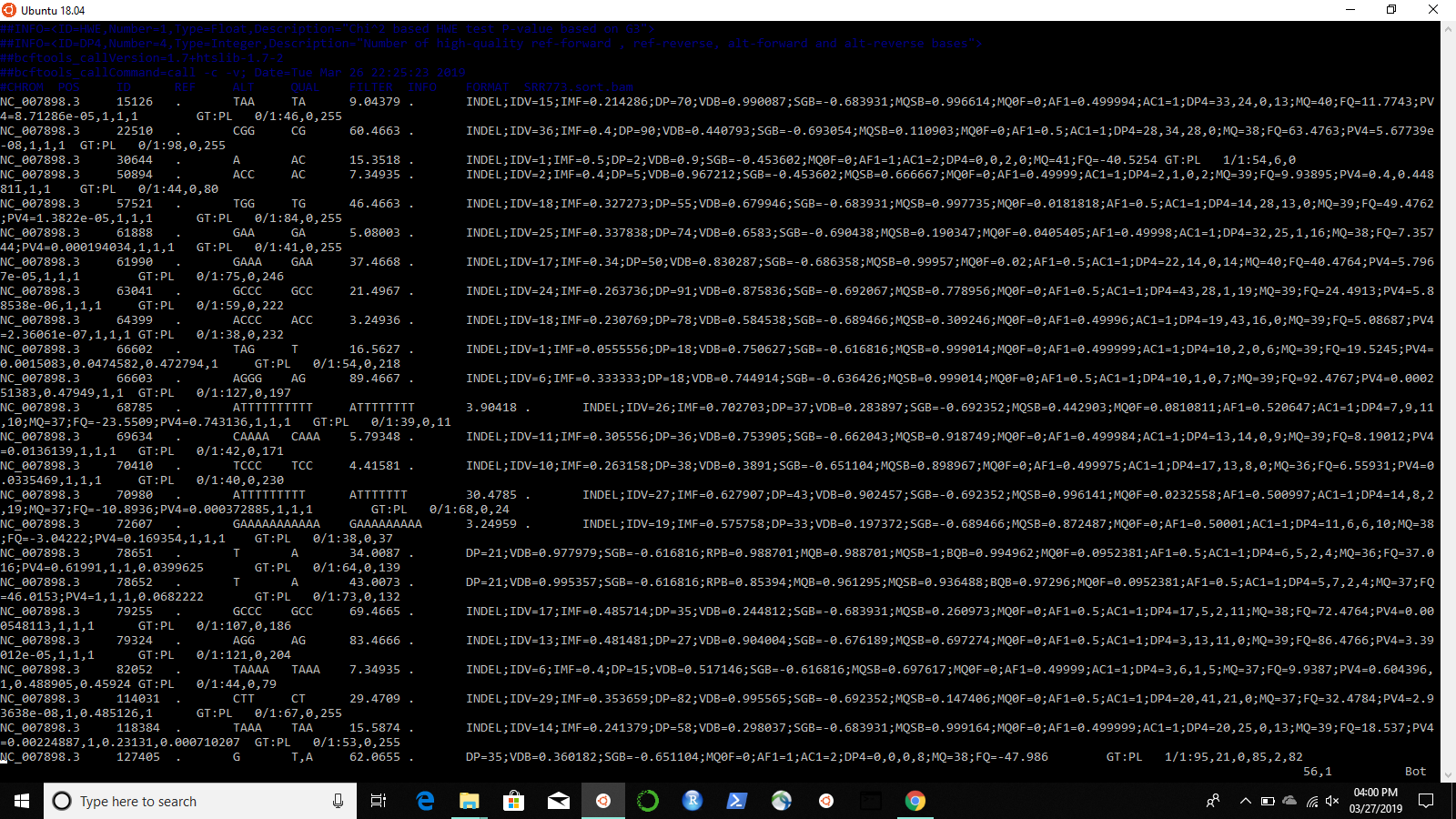
****

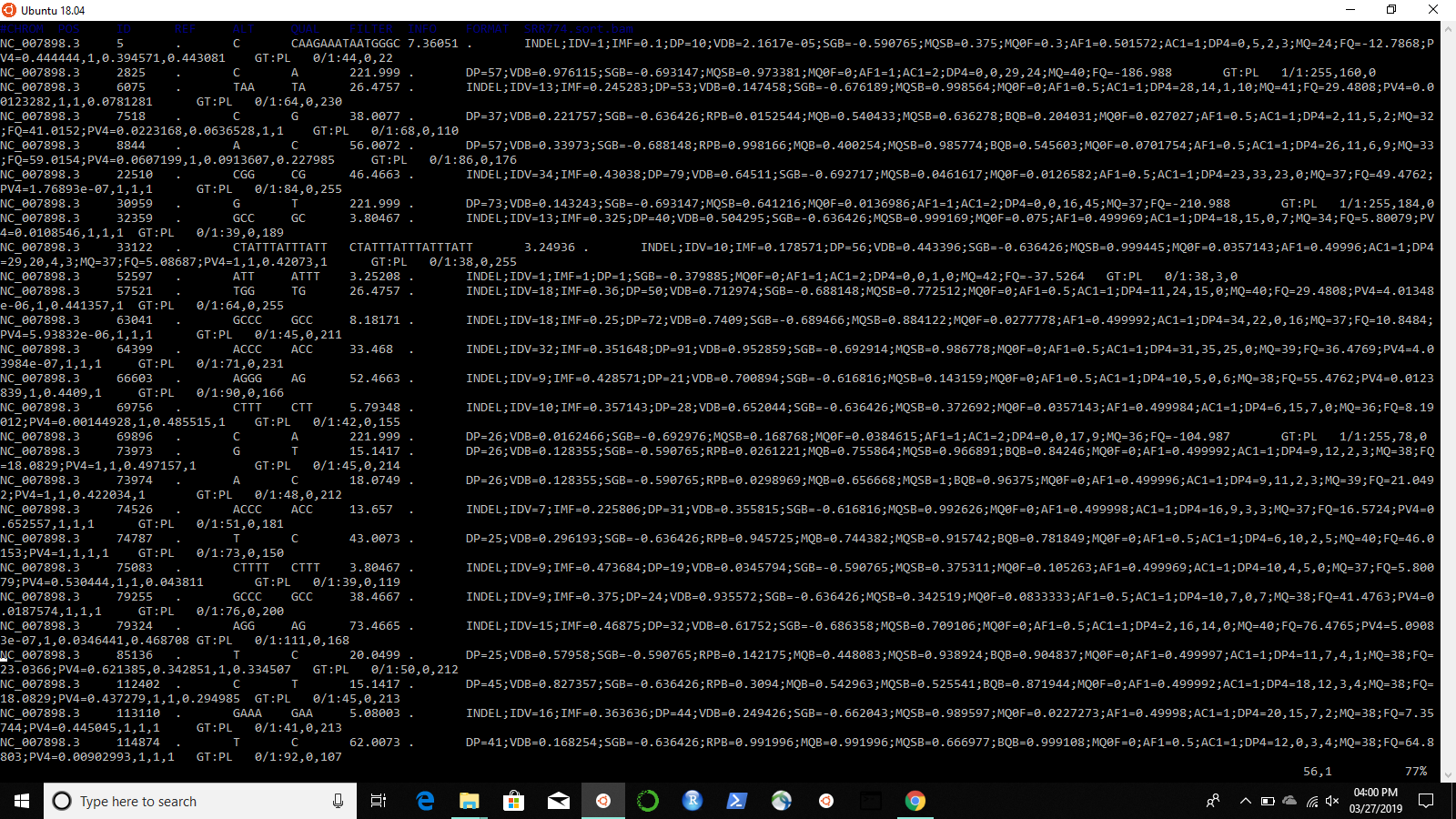
**VCF Files**

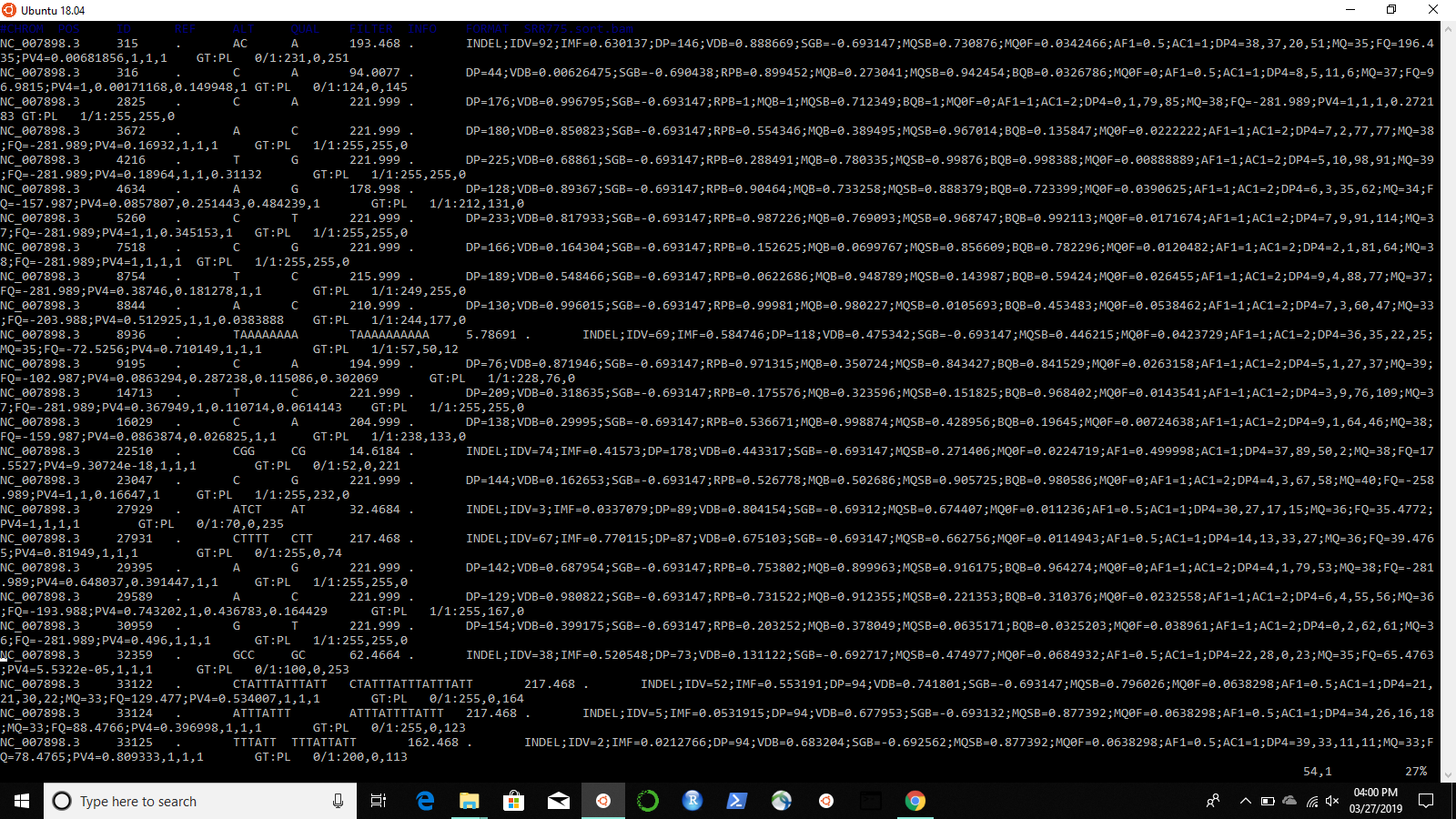
**Image 3: SRR770**

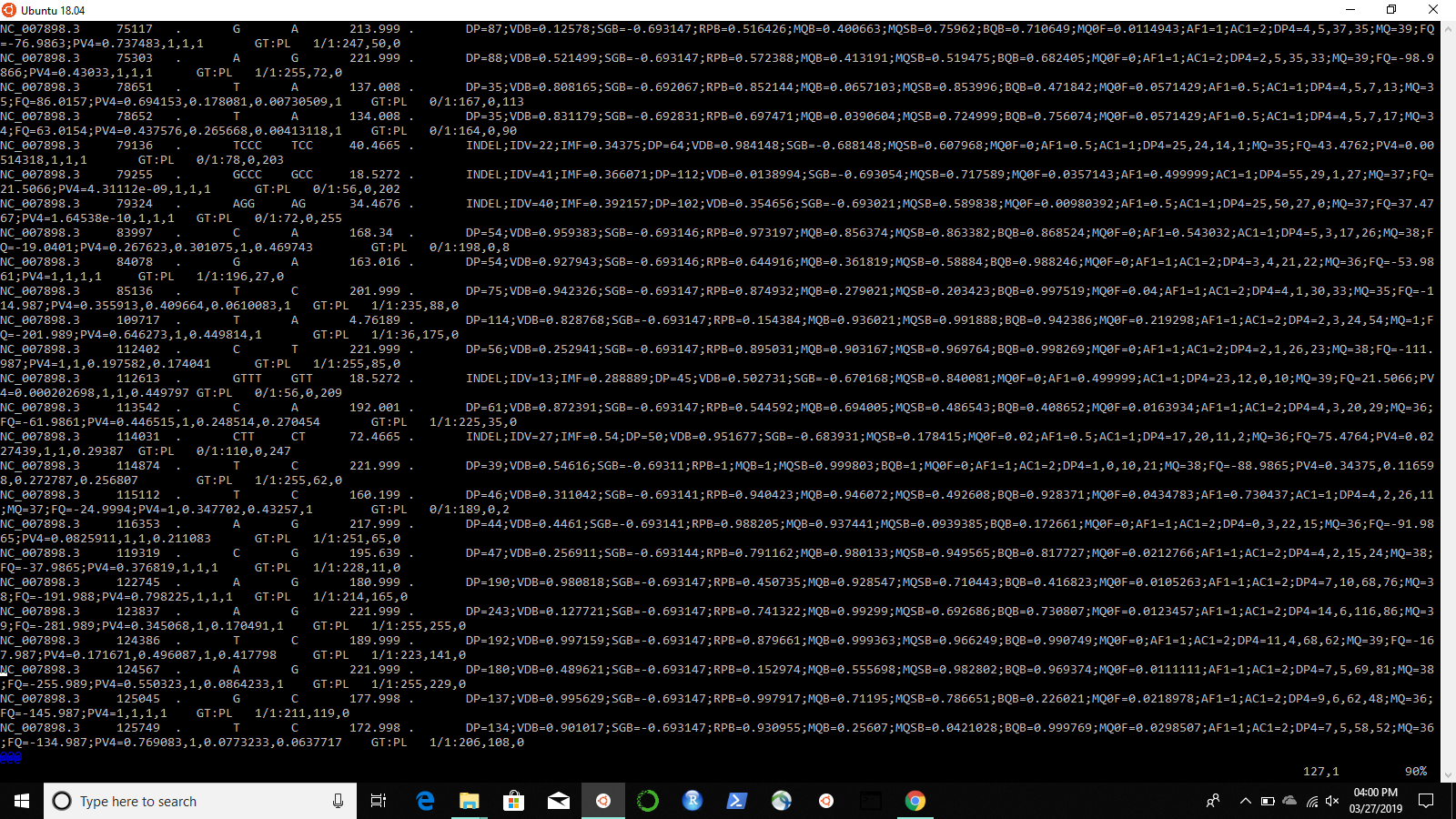
**Image 4: SRR771**

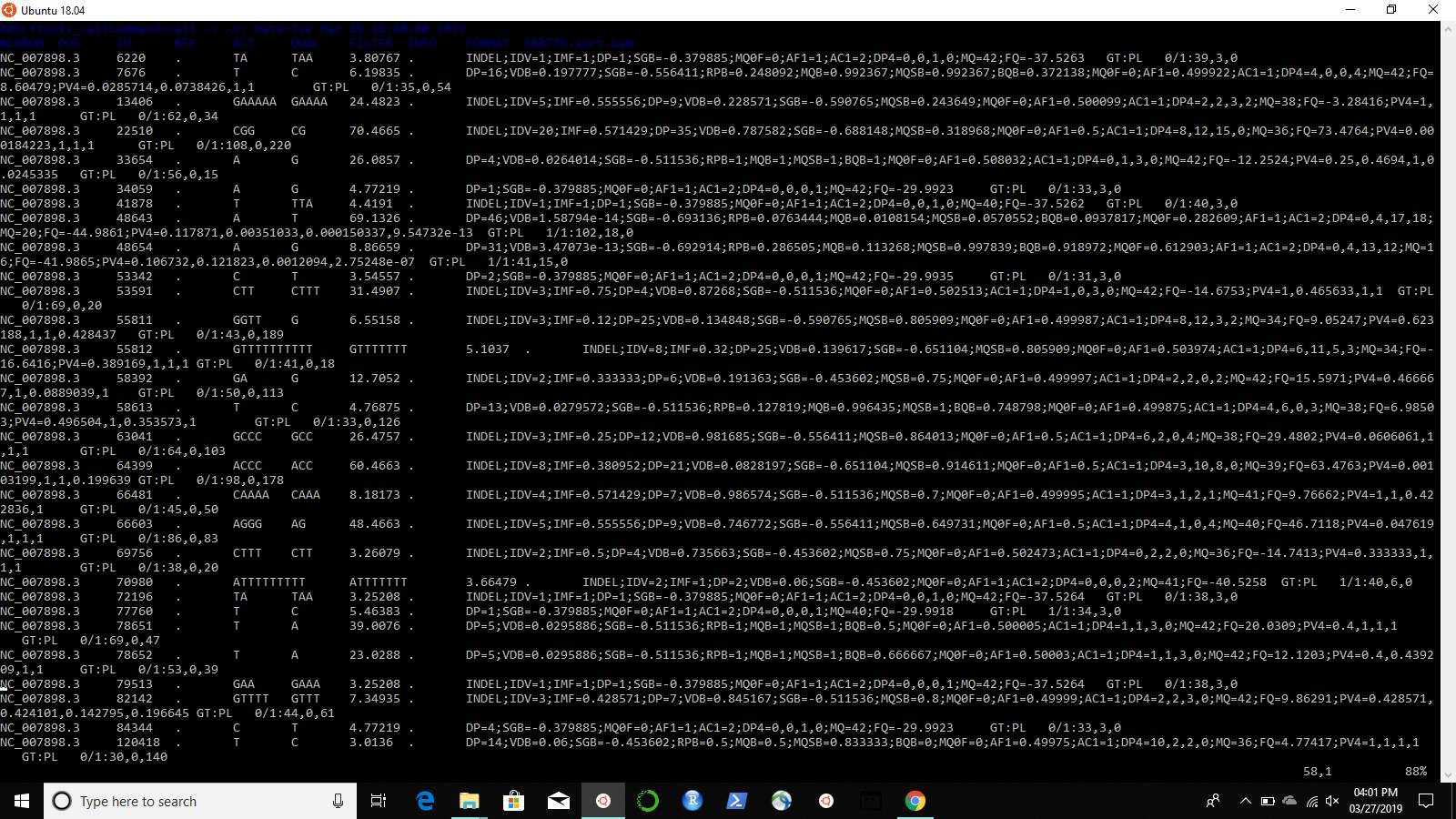
**Image 5: SRR772**

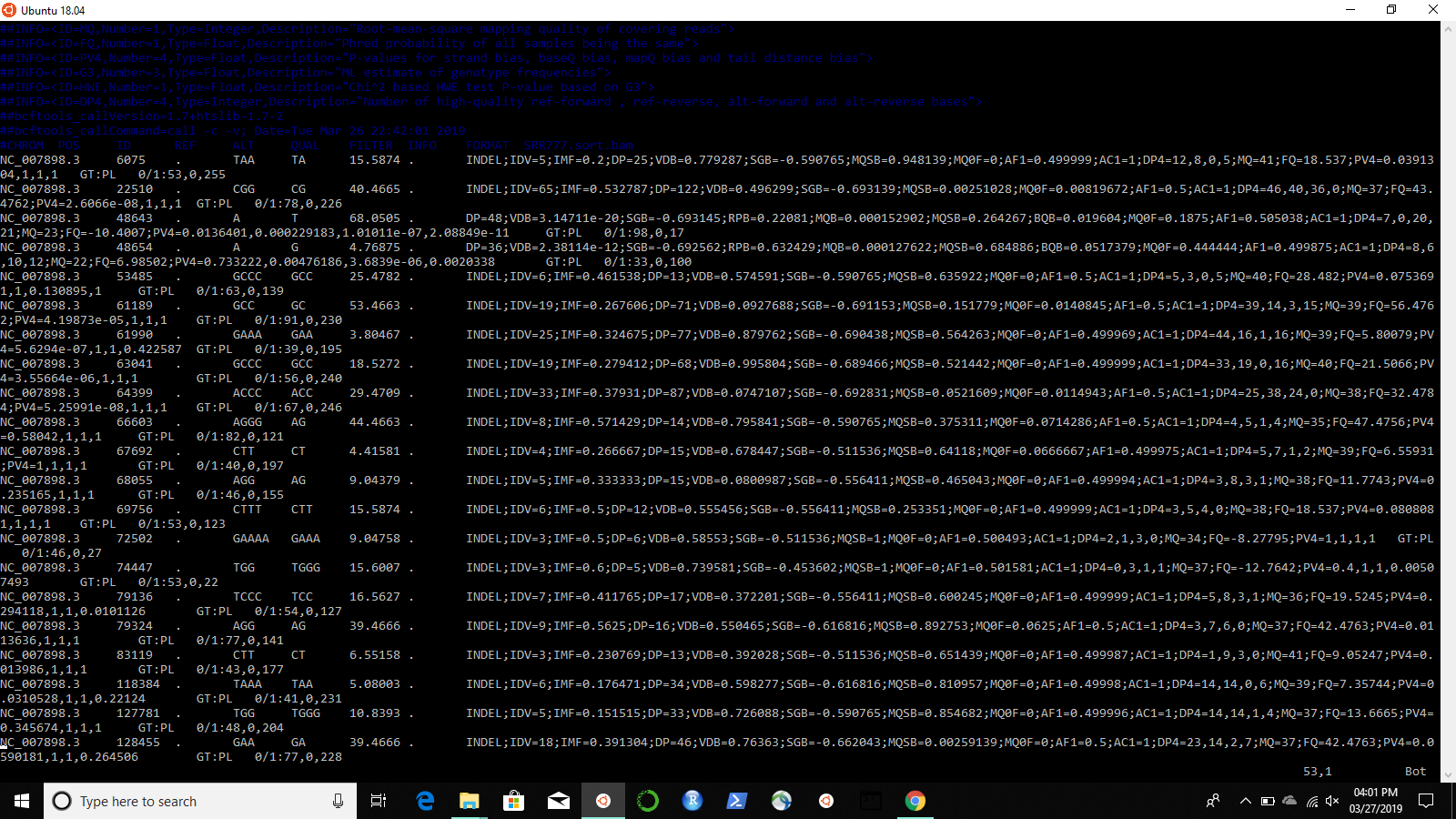
**Image 6: SRR773**

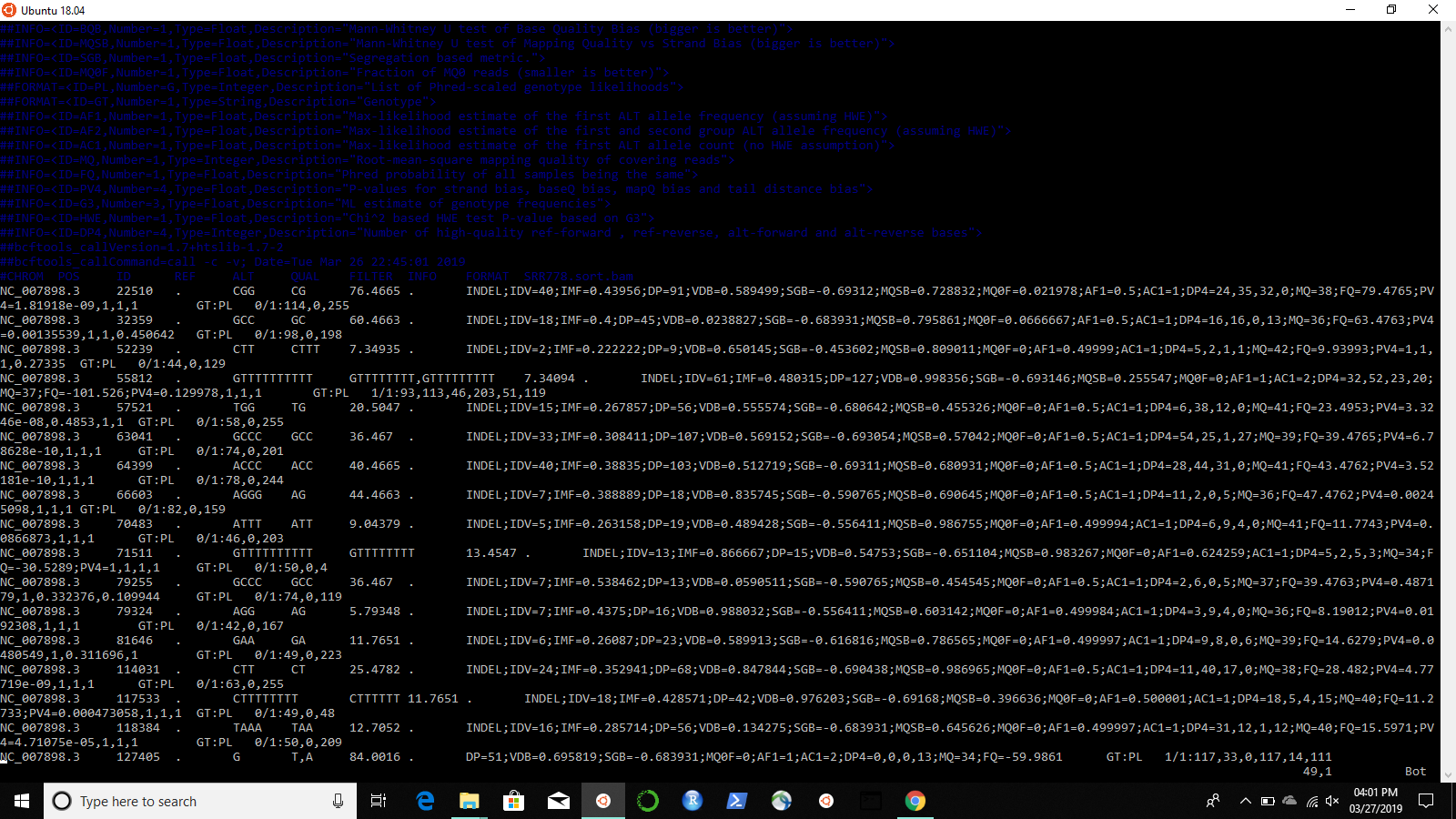
**Image 7: SRR774**

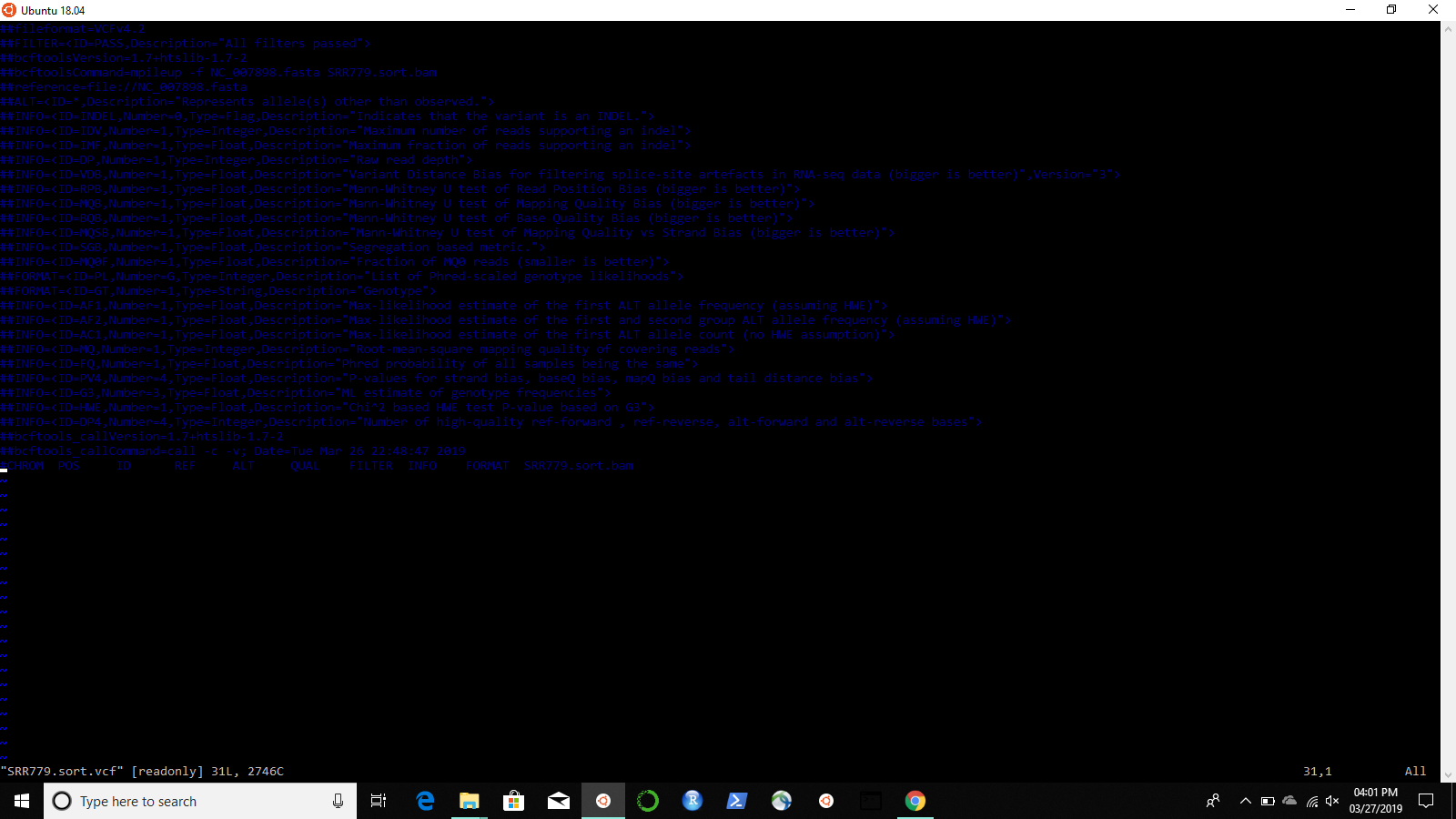
**Image 8&9: SRR775**

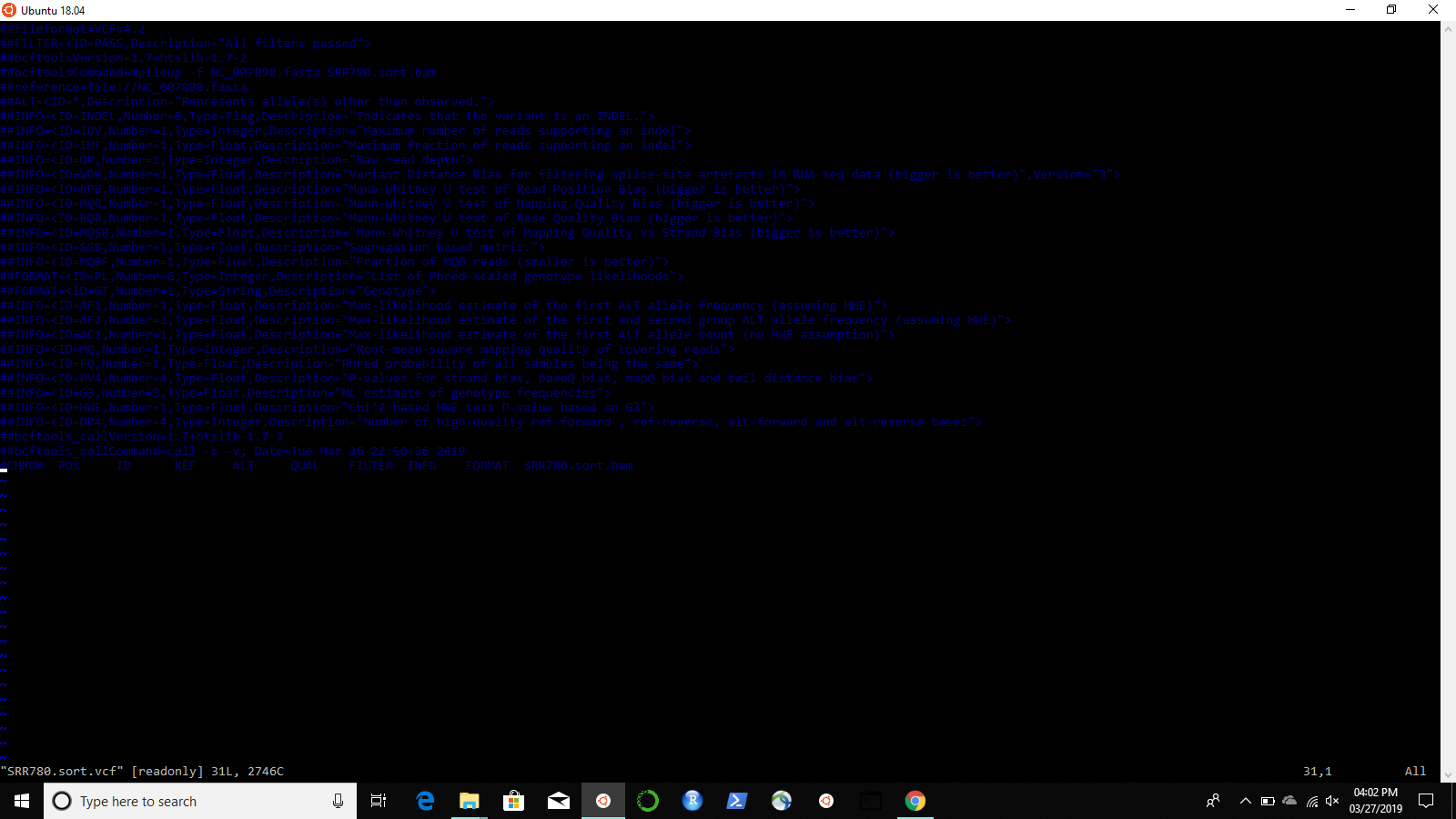
****

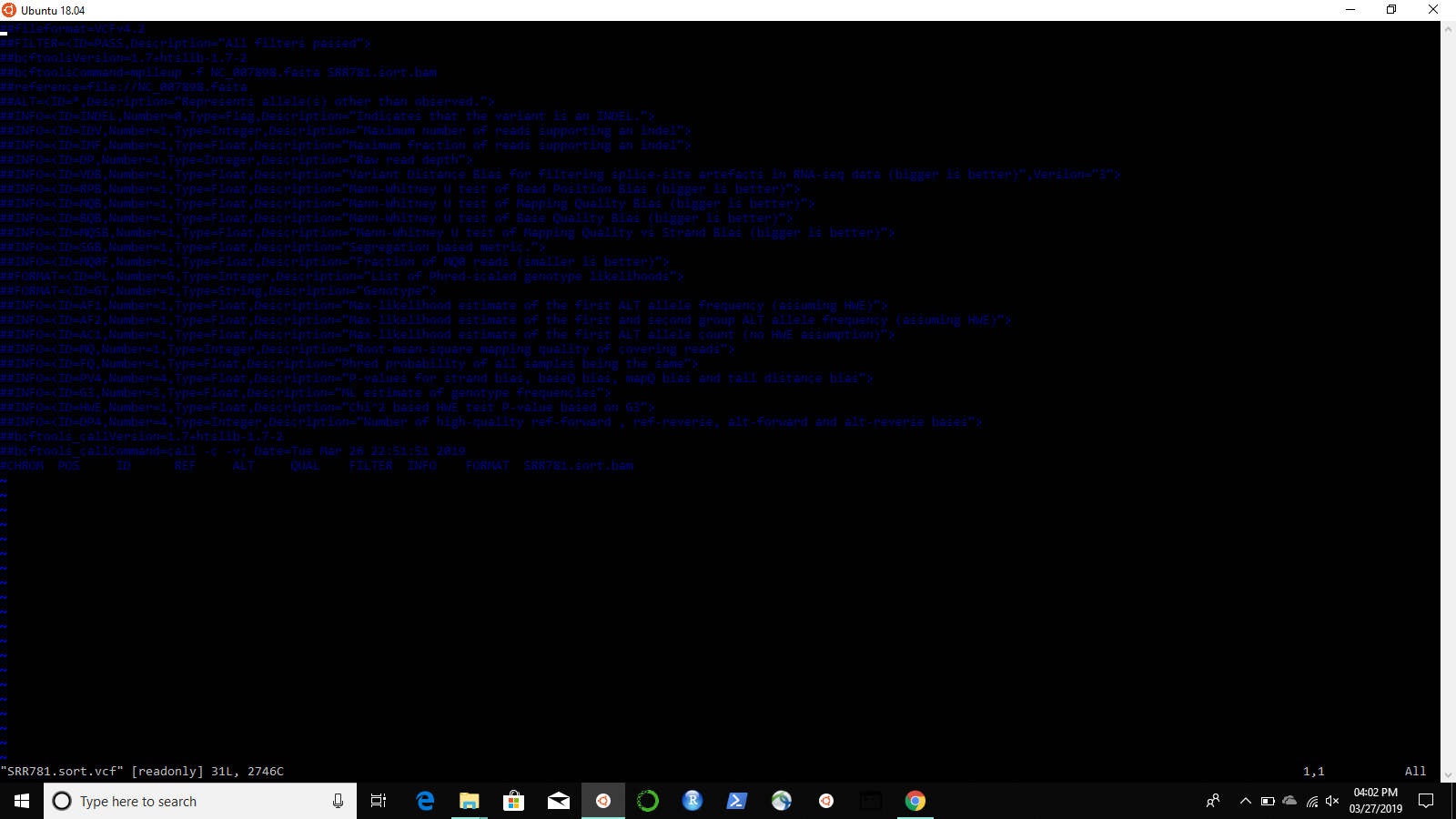
**Image 9: SRR776**

**Image 10: SRR777**

**Image 11: SRR778**

**Image 12: SRR779**

**Image 13: SRR780**

**Image 14: SRR781**