Welcome back to the practice session on the analysis of DNA sequencing data. We will be continuing with the same data set we were working with in the previous segment. In Galaxy everything should be saved in your history unless you explicitly delete the file.

The next stage involves mapping the reads to the genome. Since these sequences are DNA sequencing, we can use an aligner that aligns the entire read. We do not need to use an aligner that is splicing aware as is the case for RNA-Seq. We will use BWA, a popular aligner used by the 1000 genome project to align the sequences to the genome. On the left side of the page under genomics analysis find the link called mapping. Under mapping find map with bwa mem - medium and long reads greater than 100 base pairs against a reference genome, and click on the link. A page listing all the options will be displayed. Choose use a built in index using reference genome. Please choose hg19. For single or paired end reads, please choose single. And that our dataset should pop-up. The rest you may leave as running as default, and then please click execute.

Depending on the size of the input file and the computer speed and in this in this case the activity on this Galaxy server, this process can take several minutes, hours or even days. In our example we've taken a small subset of the data so it should be finished within a few minutes.

The alignment program has completed mapping the reads. As we can see from the green in the job number nine here on the right-hand side. Before we take a look at what a BAM/SAM file looks like, let's get a feel for how the alignment program performed overall. For this purpose we will use the flagstat program. It is part of the Samtools package located on the left hand side under Samtools and understand BAM, you will find flagstat. Flagstat reports the statistics on a bam file. As you can see, Galaxy automatically loaded the correct file as the bam file. We now click on execute...and it is now added to our history. Flagstat reports the number of reads that were successfully aligned. The Flagstat has completed and as we can see, 99% of the reads have mapped to the genome.

Let's take a look at the bwa alignment results. The bwa program returns a bam or a SAM file. The SAM file stands for sequence alignment map format, is a generic nucleotide alignment format that describes the alignment of sequencing reads to a reference. The SAM file is a human readable tab delimited file that can be compressed and converted into a BAM file, a binary alignment map format.

Let's now look at the results from the BWA-MEM file by clicking on the eye. The SAM/BAM files typically contain a header section and an alignment section. Each row represents a single read alignment. The header section includes information about the alignment, and the program that generated it. All lines in the header section are tab delimited and begin with an '@' symbol followed by some tag value pair where the tag is a two letter string that defines the content and the format of the value. For example, the header contains information about the reference sequences including the name; for example, here chromosome 12, and the length of the sequence.

OK, let's take a look now at the alignment section listed below. In the SAM format each alignment line typically represents the linear alignment of a read. Each line consists of 11 mandatory fields and one or more tab separated optional fields. The first field is the query name, as found for the read in the Fastq file. In a SAM file a read may occupy multiple alignment lines either when the alignment is chimeric or when multiple mappings are given. The second field is a bitwise flag that describes the alignment. The third field contains the reference name. In this case we will see the name of the chromosome the read has been aligned to. The fourth field is the leftmost mapping position of the alignment. The fifth field is the called the mapping quality or the MAPQ which is basically the probability that a read is aligned in the wrong place. For example, reads mapping in a repetitive region will usually get a low mapping quality. The sixth field contains the cigar string, which gives detailed information about the alignment. Here we see a 148 match, means the entire sequence matched without any deletions or insertions. The seventh, eighth and ninth fields are used in the case of paired-end read alignments. The tenth field contains a sequence of the aligned read. The eleventh field... there we go... that's the tenth field. The eleventh field contains the per base quality scores for the aligned read, as found in the FastQ file. The twelfth field contains the optional fields. More details on all these fields can be found in the SAM specification file located at the following URL:
https://samtools.github.io/hts-specs/SAMv1.pdf

The SAM or BAM file is a textual format of the alignment data. In order to be able to visualize the data, Galaxy offers some built-in visualization options. The links are located below the BAM file and allow for direct access. There are several options; one is the IGV viewer, and it is recommended to get familiar with it, but during the current practice session we will stick to the UCSC main browser, as you should already be familiar with it from a previous session.

And it also, in addition, it is located totally on the web and it doesn't require any local installation. So click on the link next to the display at UCSC main.

As an example, we will look at the known variants in the CYP2C19 gene that causes a known mutation in this subject. A single base pair, a G to A transition at nucleotide 681 in the 5th exon of the gene, creates an aberrant splice site. The change alters the reading frame of the mRNA starting with the 215th amino acid and produces a premature stop codon 20 amino acids downstream, resulting in a truncated non-functional protein. The SNP is located in on chromosome 10. Here you can see the whole gene, and you can see the track, the BAM format track. Right now it's in... If you right-click you'll see in dense format. We can instead you squish if we want to get to be able to see the actual reads that have undergone alignment.

And now we're looking about two hundred thousand base pairs. So I'm here we're talking about... these are all the reads that are covering this location. OK. And this is the gene that we're interested in, the CYP2C19. OK. And we're interested in this specific mutation that causes the aberrant splice site. So we'll actually zoom into the area which is on chromosome 10. It's 96,541,616. So we'll look in the surrounding area to 96,541,630... So there it should be right in the middle.

So we're looking at now a window of thirty one base pairs where the middle base pair should actually be the variant that we're interested in, and as you can see the reference is a G and yet here, you will have to just trust me, about 50% of the sites are A and in this subjects genome contains a heterozygote allele at this location instead of being a homozygote for G, and that's what causes the mutation in this subject.

So, in the next segment we'll see how to detect these types of variants on a genome-wide scale and see you next time.