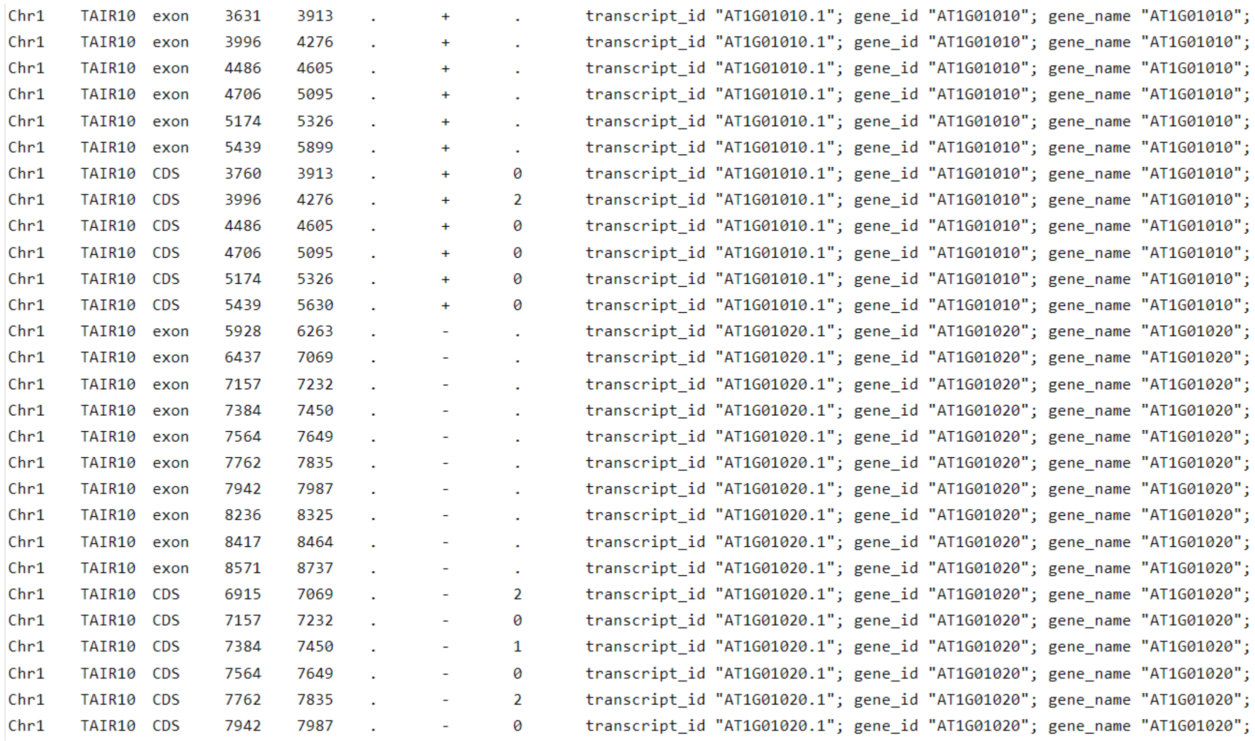
Inan Khan
Homework 4

Looking at the sam file it was not formatted correctly for viewing in most softwares. After some trial and error, I came to the conclusion that it would be best to convert the .sam file to a .bam file for easier viewing and manipulating. The software that I found that would be best for viewing this type of data was IGV, Internal Genomics Viewer.

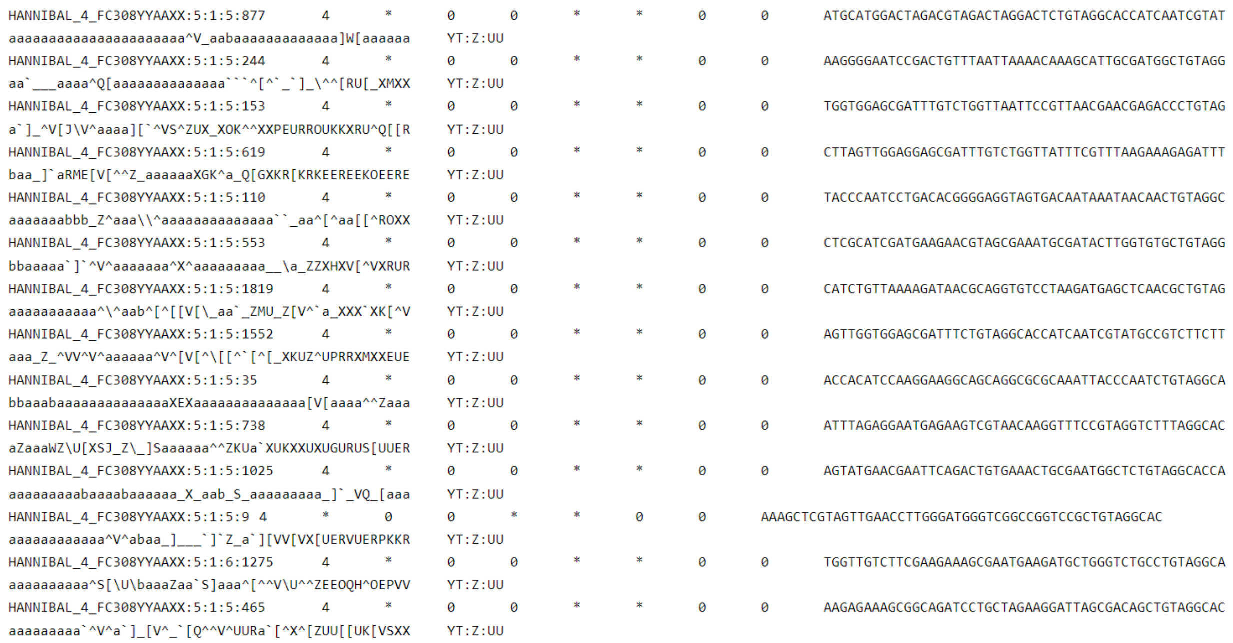
This was the view of the GTF file



```
Chr1	TAIR10	exon	3631	3913	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	3996	4276	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	4486	4605	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	4706	5095	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	5174	5326	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	5439	5899	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	3760	3913	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	3996	4276	.	+	2	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	4486	4605	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	4706	5095	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	5174	5326	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	5439	5630	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	5928	6263	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	6437	7069	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7157	7232	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7384	7450	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7564	7649	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7762	7835	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7942	7987	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	8236	8325	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	8417	8464	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	8571	8737	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	6915	7069	.	-	2	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7157	7232	.	-	0	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7384	7450	.	-	1	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7564	7649	.	-	0	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7762	7835	.	-	2	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7942	7987	.	-	0	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
```

As we can see the separate columns all represent different aspects of the gene in annotated form. This gene data was already pre-aligned and so viewing this we can more easily compare it to the sam file that was given.

Similarly if we take a look at the sam file we can see there there is some formatting needed to process this raw data. In its current state it is a bit of a mess and they give several methods in the documentation to do so. My method for turning this sam file into a bam file was using sam tools and then opening the resulting file in IGV.



```
HANNIBAL_4_FC308YYAAXX:5:1:5:877        4       *       0       0       *       *       0       0       ATGCATGGACTAGACGTAGACTAGGACTCTGTAGGCACCATCAATCGTAT
aaaaaaaaaaaaaaaaaaaaaa^V_aabaaaaaaaaaaaaaa]W[aaaaaa     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:244        4       *       0       0       *       *       0       0       AAGGGGAATCCGACTGTTTAATTAAAACAAAGCATTGCGATGGCTGTAGG
aa`___aaaa^Q[aaaaaaaaaaaaaa```^[^`_`]_\^^[RU[_XMXX     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:153        4       *       0       0       *       *       0       0       TGGTGGAGCGATTTGTCTGGTTAATTCCGTTAACGAACGAGACCCTGTAG
a`]_^V[J\V^aaaa][`^VS^ZUX_XOK^^XXPEURROUKKXRU^Q[[R     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:619        4       *       0       0       *       *       0       0       CTTAGTTGGAGGAGCGATTTGTCTGGTTATTTCGTTTAAGAAAGAGATTT
baa_]`aRME[V[^^Z_aaaaaaXGK^a_Q[GXKR[KRKEEREEKOEERE     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:110        4       *       0       0       *       *       0       0       TACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAACTGTAGGC
aaaaaaabbb_Z^aaa\\^aaaaaaaaaaaaaa``_aa^[^aa[[^ROXX     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:553        4       *       0       0       *       *       0       0       CTCGCATCGATGAAGAACGTAGCGAAATGCGATACTTGGTGTGCTGTAGG
bbaaaaa`]`^V^aaaaaaa^X^aaaaaaaaa__\a_ZZXHXV[^VXRUR     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:1819       4       *       0       0       *       *       0       0       CATCTGTTAAAAGATAACGCAGGTGTCCTAAGATGAGCTCAACGCTGTAG
aaaaaaaaaaa^\^aab^[^[[V[\_aa`_ZMU_Z[V^`a_XXX`XK[^V     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:1552       4       *       0       0       *       *       0       0       AGTTGGTGGAGCGATTTCTGTAGGCACCATCAATCGTATGCCGTCTTCTT
aaa_Z_^VV^V^aaaaaa^V^[V[^\[[^`[^[_XKUZ^UPRRXMXXEUE     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:35         4       *       0       0       *       *       0       0       ACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCTGTAGGCA
bbaaabaaaaaaaaaaaaaaaXEXaaaaaaaaaaaaaa[V[aaaa^^Zaaa     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:738        4       *       0       0       *       *       0       0       ATTTAGAGGAATGAGAAGTCGTAACAAGGTTTCCGTAGGTCTTTAGGCAC
aZaaaWZ\U[XSJ_Z\_]Saaaaaa^^ZKUa`XUKXXUXUGURUS[UUER     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:1025       4       *       0       0       *       *       0       0       AGTATGAACGAATTCAGACTGTGAAACTGCGAATGGCTCTGTAGGCACCA
aaaaaaaaabaaaabaaaaaa_X_aab_S_aaaaaaaaa_]`_VQ_[aaa     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:9 4        *       0       0       *       *       0       0       AAAGCTCGTAGTTGAACCTTGGGATGGGTCGGCCGGTCCGCTGTAGGCAC
aaaaaaaaaaaa^V^abaa_]___`]`Z_a`][VV[VX[UERVUERPKKR     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:6:1275       4       *       0       0       *       *       0       0       TGGTTGTCTTCGAAGAAAGCGAATGAAGATGCTGGGTCTGCCTGTAGGCA
aaaaaaaaaa^S[\U\baaaZaa`S]aaa^[^^V\U^^ZEEOQH^OEPVV     YT:Z:UU
HANNIBAL_4_FC308YYAAXX:5:1:5:465        4       *       0       0       *       *       0       0       AAGAGAAAGCGGCAGATCCTGCTAGAAGGATTAGCGACAGCTGTAGGCAC
aaaaaaaaa`^V^a`]_[V^_`[Q^^V^UURa`[^X^[ZUU[[UK[VSXX     YT:Z:UU
```

Lastly after having both of these files available and if we want to determine the number of reads that match a gene we would simply need to count how many start positions are between the start and stop coordinates for the exon. The start positions are found on the sample.sam file while the exon coordinates are found on arabidopsis file. The last thing I did to make the process more efficient was split the alignments found on the sample.sam based on chromosomes. If I am able to store the start positions for every chromosome and then use the quicksort function to sort the chromosomes then it becomes much easier count and compare individual matching entries

I noted several things when working with the sample data and manipulating it in IGV. The first thing I saw was that there were some entries and lines that did not have sequence alignment data and so those were left as blanks. The next thing I noticed when getting familiar with IGV was the fact that it does not accept sam files due to them being formatted incorrectly and needing an index. My previous attempt to work with the data before using IGV was trying to convert the file somehow to work with excel or google sheets. The reason why I believe this did not work out is that the column syntax is different in the sam file from what can be used in excel. If I was somehow able to temporarily remove the column headers/names, I feel as though this method may have been a possible alternative