Genome and File Formats

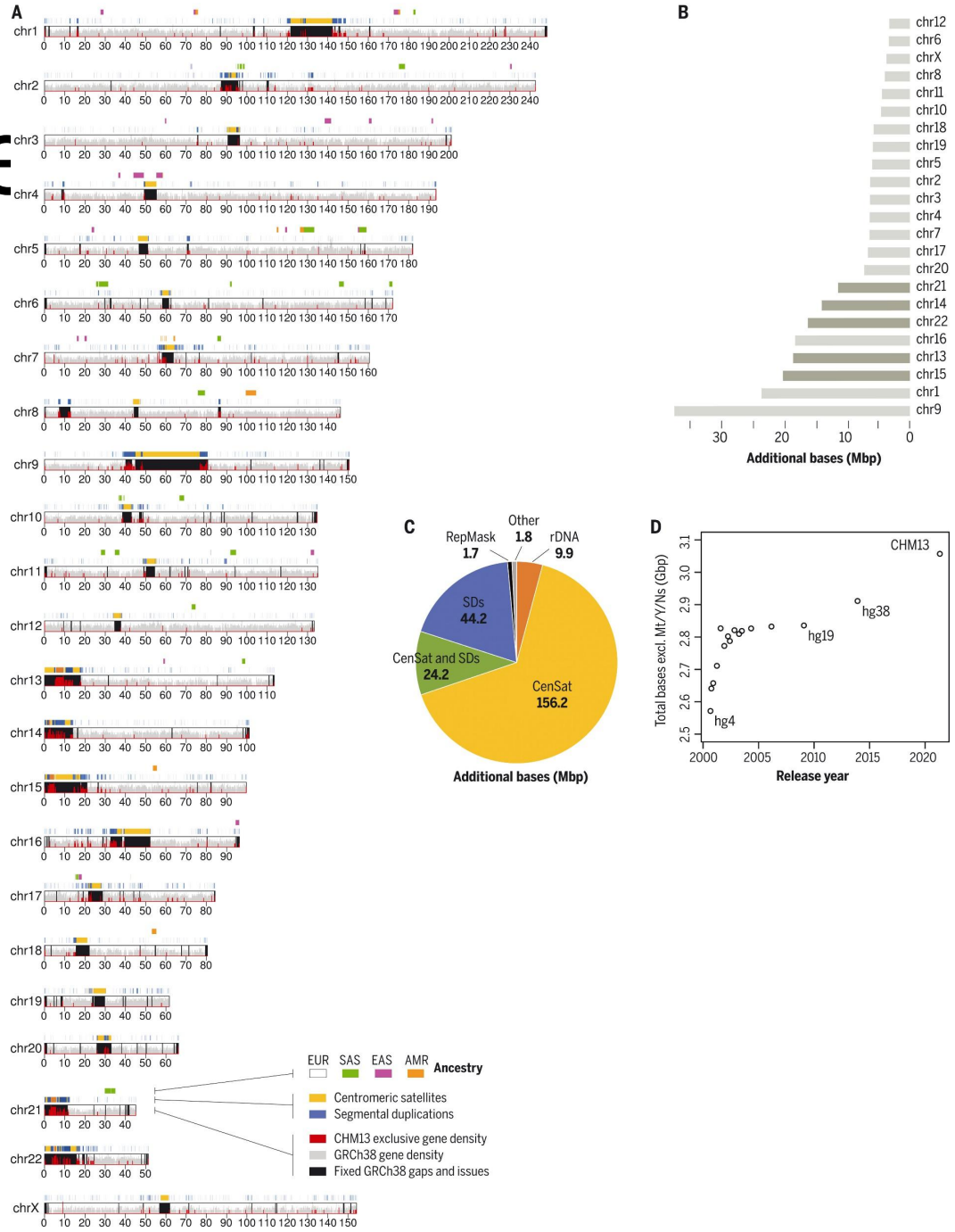
Phuc Loi Luu, PhD
Loi.lp@pacificinformatics.com.vn
5/8/2023

Human Reference genome

https://genome.ucsc.edu/

Fasta format

```
Header   >VIT_201s0011g03530.1
Sequence AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
         GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
Header   >VIT_201s0011g03540.1
Sequence CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
         AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
Header   >VIT_201s0011g03550.1
Sequence CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
         GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

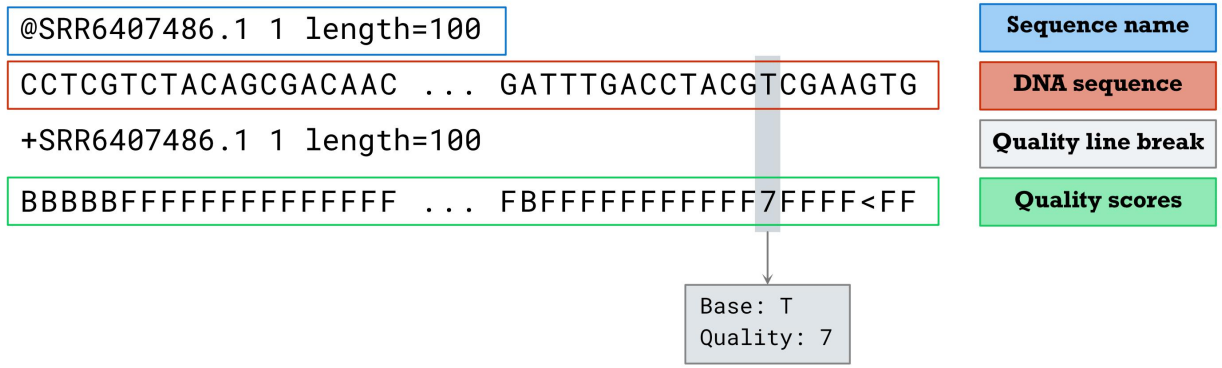


https://www.science.org/doi/10.1126/science.abj6987

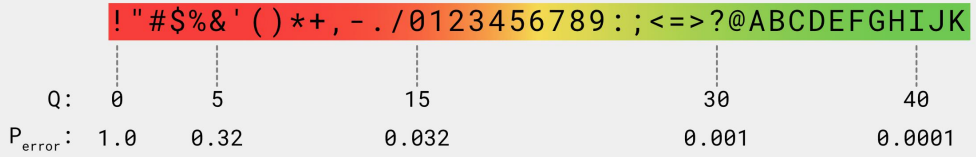
Fastq file

FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFF7FFFF<FF
```



Quality scores as ASCII characters:



$$Q = -10\log_{10} P_{error}$$

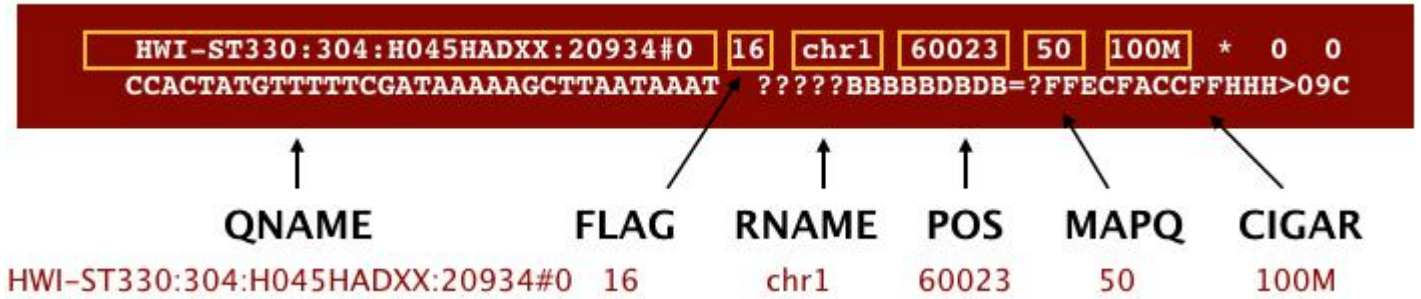
Human Gene Annotation

https://www.gencodegenes.org/

GTF format

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr4	protein_coding	CDS	24053	24477	.	+	0	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	24053	24477	.	+	.	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	24979	25153	.	+	1	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	24979	25153	.	+	.	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	25218	25450	.	+	0	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	25218	25450	.	+	.	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	CDS	25501	25618	.	+	1	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	exon	25501	25621	.	+	.	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	protein_coding	stop_codon	25619	25621	.	+	0	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p_
chr4	pseudogene	exon	26994	27101	.	-	.	exon_number "7"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	27167	27349	.	-	.	exon_number "6"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	28371	28609	.	-	.	exon_number "5"; gene_id "FBgn0052011"; gene_name "CR32011";

BAM Format




```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGAT *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGG   *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCT     * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCA     *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC         * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGC       * NM:i:1
```

Gene expression count matrix

samples: want to see if differences across condition are significant (w.r.t. biological and technical variation)

features (e.g. genes)

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

Gene expression count matrix (1)

Each column is a sample

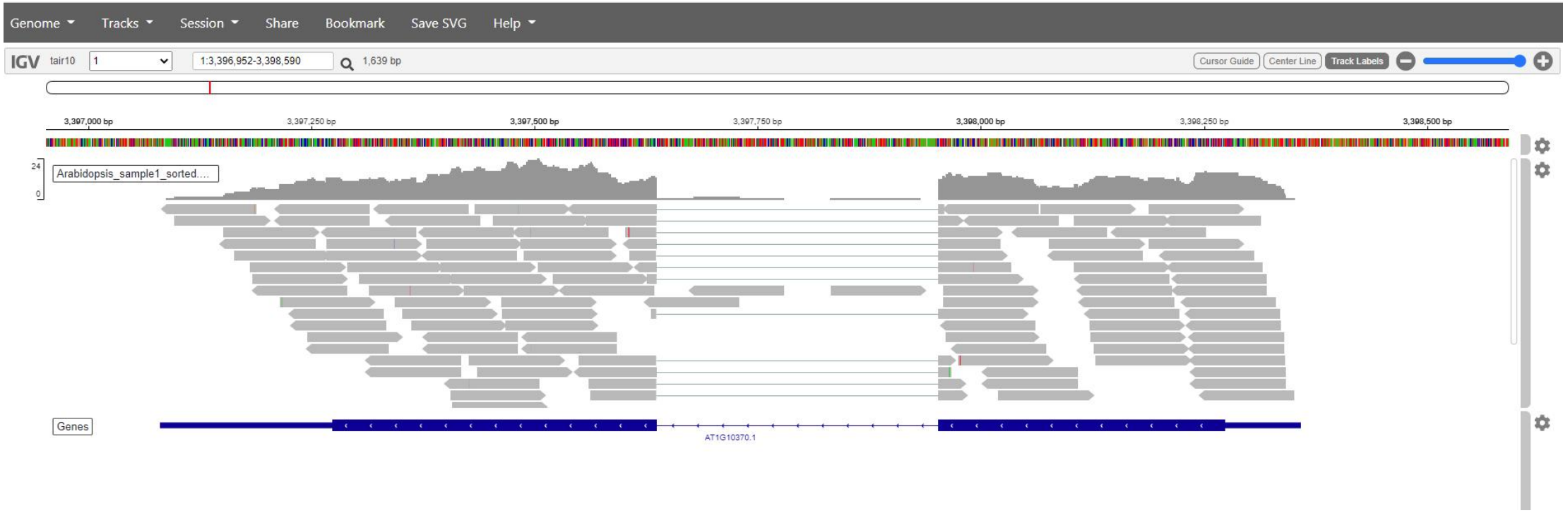
Each row is a gene

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

Gene expression count matrix (2)

EntrezGeneID	Length	MCL1-DG_BC2CTUACXX_ACTTGA_L002_R1	MCL1-DH_BC2CTUACXX_CAGATC_L002_R1	MCL1-DI_BC2CTUACXX_ACAGTG_L002_R1	MCL1-DJ_BC2CTUACX
497097	3634	438	300	65	237
100503874	3259	1	0	1	1
100038431	1634	0	0	0	0
19888	9747	1	1	0	0
20671	3130	106	182	82	105
27395	4203	309	234	337	300
18777	2433	652	515	948	935
100503730	799	0	1	0	0
21399	2847	1604	1495	1721	1317
58175	2241	4	2	14	4
108664	1976	769	752	1062	987

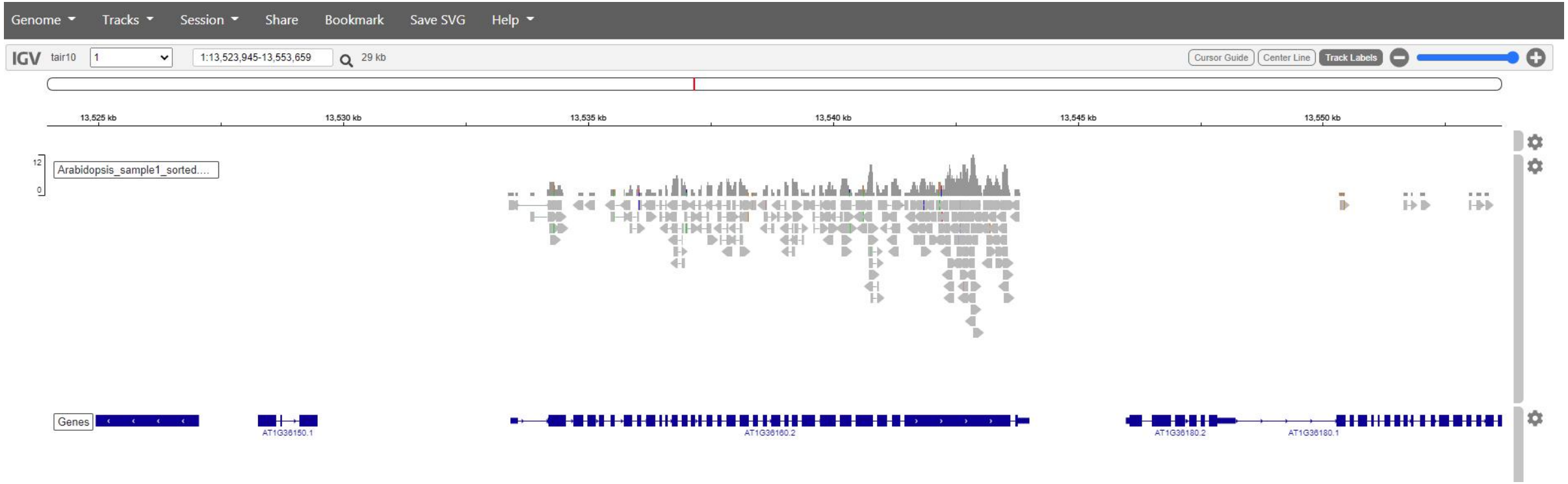
BAM in IGV (1)


Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
tair10
1
1:3,396,952-3,398,590
1,639 bp
Cursor Guide
Center Line
Track Labels
3,397,000 bp
3,397,250 bp
3,397,500 bp
3,397,750 bp
3,398,000 bp
3,398,250 bp
3,398,500 bp
24
0
Arabidopsis_sample1_sorted....
Genes
AT1G10370.1

BAM in IGV (2)


Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
tair10
1
1:5,514,244-5,516,099
1,856 bp
Cursor Guide
Center Line
Track Labels
5,514,500 bp
5,514,750 bp
5,515,000 bp
5,515,250 bp
5,515,500 bp
5,515,750 bp
5,516,000 bp
31
0
Arabidopsis_sample1_sorted....
Genes
AT1G16080.1

BAM in IGV (3)


Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
tair10
1
1:13,523,945-13,553,659
29 kb
Cursor Guide
Center Line
Track Labels
13,525 kb
13,530 kb
13,535 kb
13,540 kb
13,545 kb
13,550 kb
12
0
Arabidopsis_sample1_sorted....
Genes
AT1G36150.1
AT1G36160.2
AT1G36180.2
AT1G36180.1