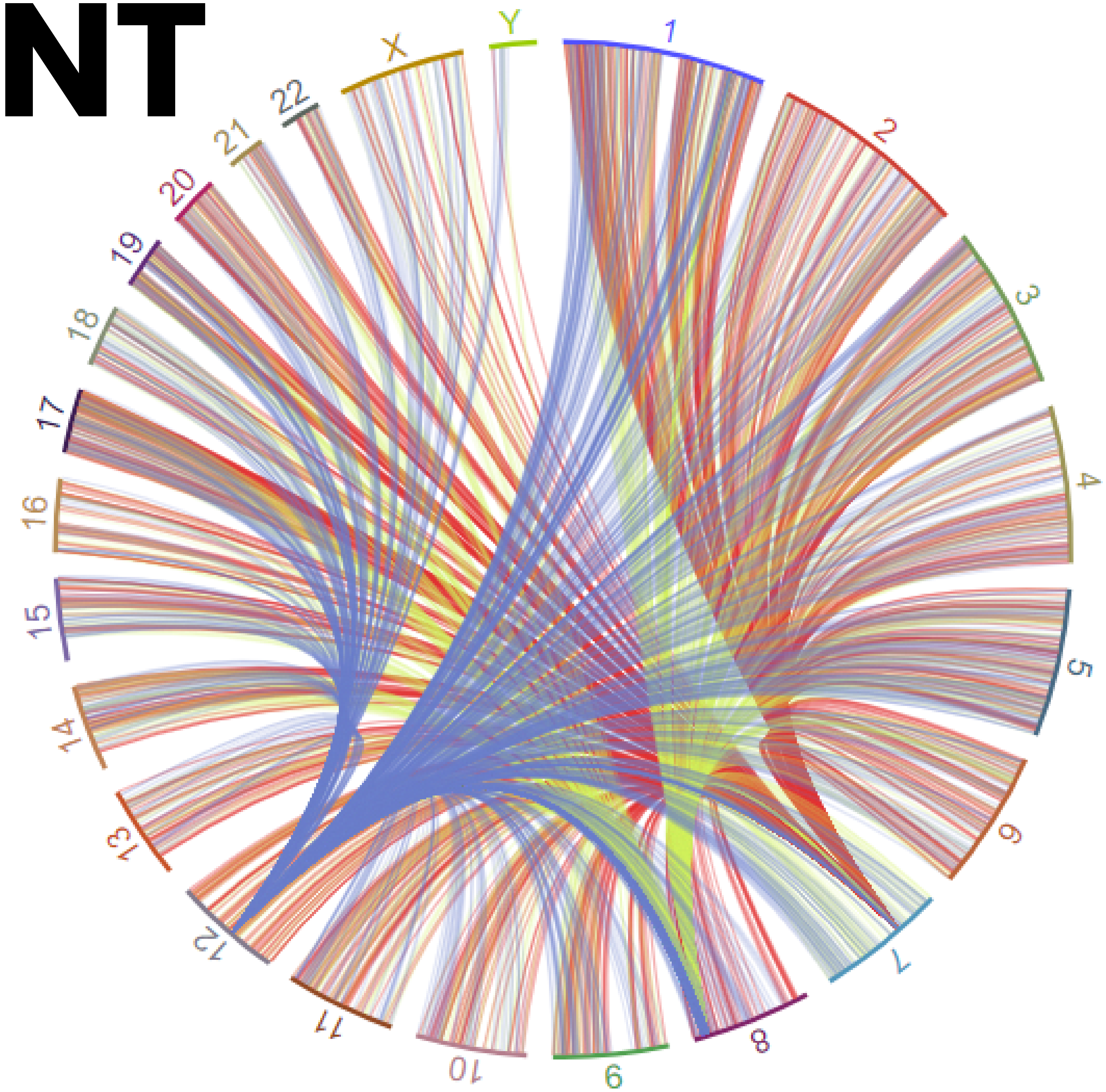
PRESENTATION 15 :

GENOME ALIGNMENT & VISUALISATION TOOLS

BWA, BOWTIE, SAMTools, Output Formats & IGV

PRESENTED BY:
103 -Bhatia Aayushi
108 -Kadam Shruti
109 -Kashyap AnjaniKumar
127 -Rawat Manasi



SOURCE: ChIA-Pet - Wei Lab GSM3553630 Interactions in Circular View (https://igv.org/app/)

A SHORT RECAP

-NGS PIPELINE-

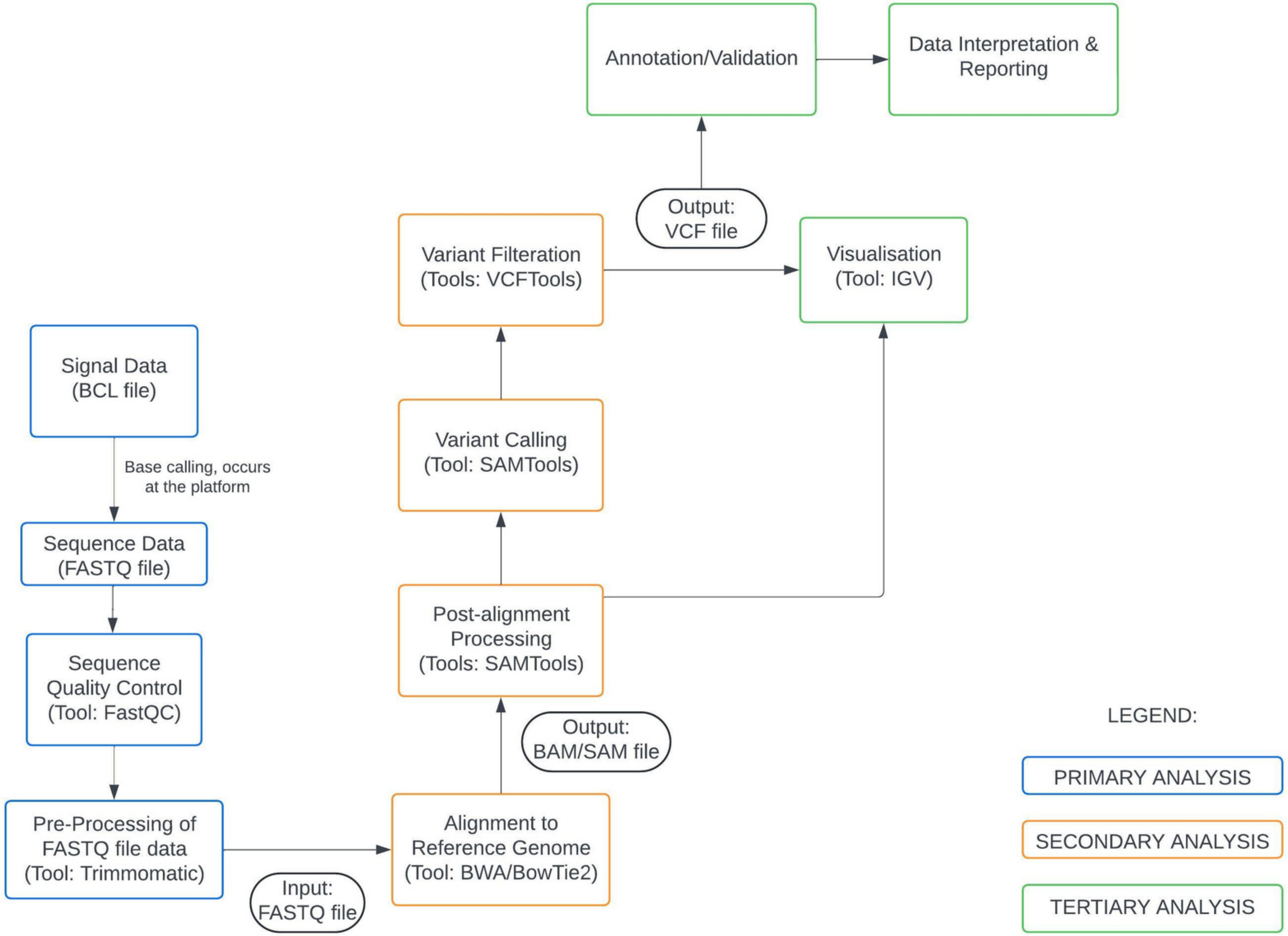
WHAT IS A PIPELINE?

A set of connected algorithms, executed in a predefined order to process and analyze NGS data, is collectively referred to as a pipeline.

RECAP:

NGS PIPELINE:

Our focus will be on the tools used in secondary analysis for aligning the reads to a reference genome and variant calling. As well as a genome browser used for visualization of the alignment and other interactions.



Signal Data (BCL file)

Base calling, occurs at the platform

Sequence Data (FASTQ file)

Sequence Quality Control (Tool: FastQC)

Pre-Processing of FASTQ file data (Tool: Trimmomatic)

Input: FASTQ file

Alignment to Reference Genome (Tool: BWA/BowTie2)

Output: BAM/SAM file

Post-alignment Processing (Tools: SAMTools)

Variant Calling (Tool: SAMTools)

Variant Filteration (Tools: VCFTools)

Output: VCF file

Visualisation (Tool: IGV)

Annotation/Validation

Data Interpretation & Reporting

LEGEND:

PRIMARY ANALYSIS

SECONDARY ANALYSIS

TERTIARY ANALYSIS

RECAP:

NGS PIPELINE:

From a glance, the standard NGS data analysis pipeline looks complicated, however, it can be simplified and broken down into three individual sections:

1. **Primary Analysis:** Involves conversion of signal data file (BCL) into sequence data file (FASTQ) and pre-processing of the FASTQ files via trimming, filtering and demultiplexing. This step uses tools like FASTQC and Trimmomatic.
2. **Secondary Analysis:** Involves mapping the data to a reference genome or to each other (de-novo assembly) to align the sequences. Software like BWA, BOWTIE2 are used and FASTQ files are converted into BAM & SAM files. Secondary analysis also includes post-alignment processing and variant calling for which software like SAMTools can be used.
3. **Tertiary Analysis:** Includes the downstream analysis wherein the genome is visualized using IGV, genome/variants are annotated and the results are interpreted.

USED IN NGS

-ALIGNMENT TOOLS-

Once high-quality sequence reads have been generated, the next step in the data analysis workflow is to align the reads against a reference genome (using tools like BWA, BOWTIE) or perform a de novo assembly(using tools like Velvet, SPAdes) and then call any variants detected(using SAMTools).

Our focus is on performing read alignment using a reference genome and study the tools used to do the same.

ALIGNMENT TOOLS:

- The aim of sequence alignment is to find the genomic location where a read originates from and determine how many reads aligned to that position.
- Traditional alignment methods such as BLAST cannot be used for large data sets generated by NGS as they are too memory consuming.
- Alignment tools require FASTQ files for input and output data into a SAM or BAM file.
- Before the aligned reads can be used for variant calling analysis, there are additional post-alignment processing steps that ensure only the highest quality reads are retained in the BAM file.
- Some of these steps include: Duplicate removal, Base quality score recalibration

-BWA-

BURROWS WHEELER ALIGNER

Next gen sequencing experiments result in millions of relatively small reads that must be mapped to a reference genome. Since using the standard alignment algorithms is unfeasible for such large numbers of reads, a lot of effort has been put into developing methods which are fast and are relatively memory-efficient.

ALIGNMENT TOOLS:

BWA

- BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.
- It is based on a backward search with Burrows-Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps.
- It consists of three algorithms:
    1. BWA-backtrack
    2. BWA-SW
    3. BWA-MEM
- BWA supports both base space reads and color space reads.
- BWA can be accessed via command line as well as via Galaxy Server.

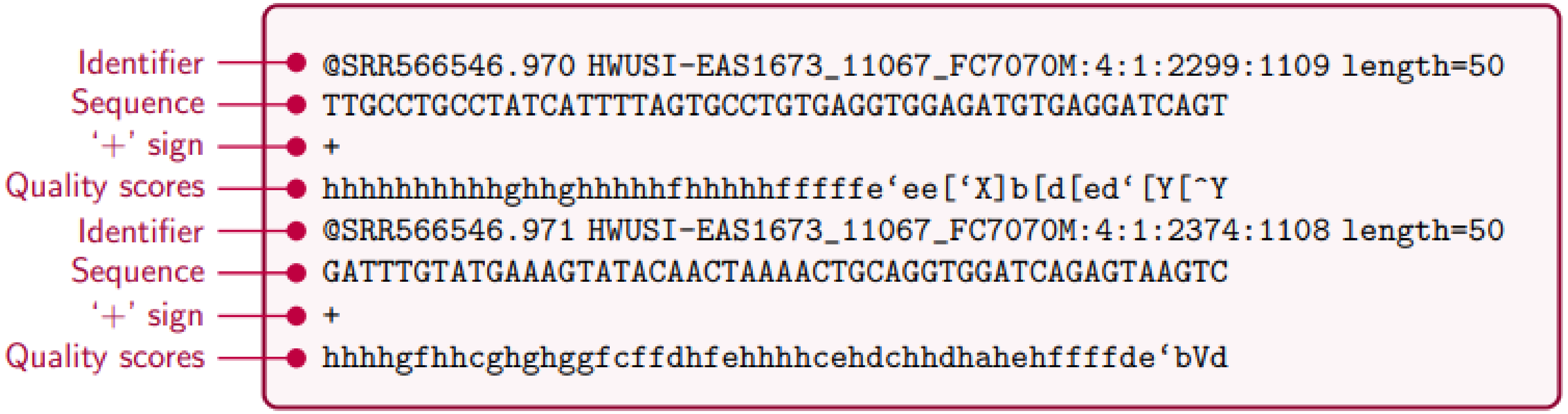
INPUT & OUTPUT FILES:

INPUT FILE:

BWA accepts files in FASTQ format.

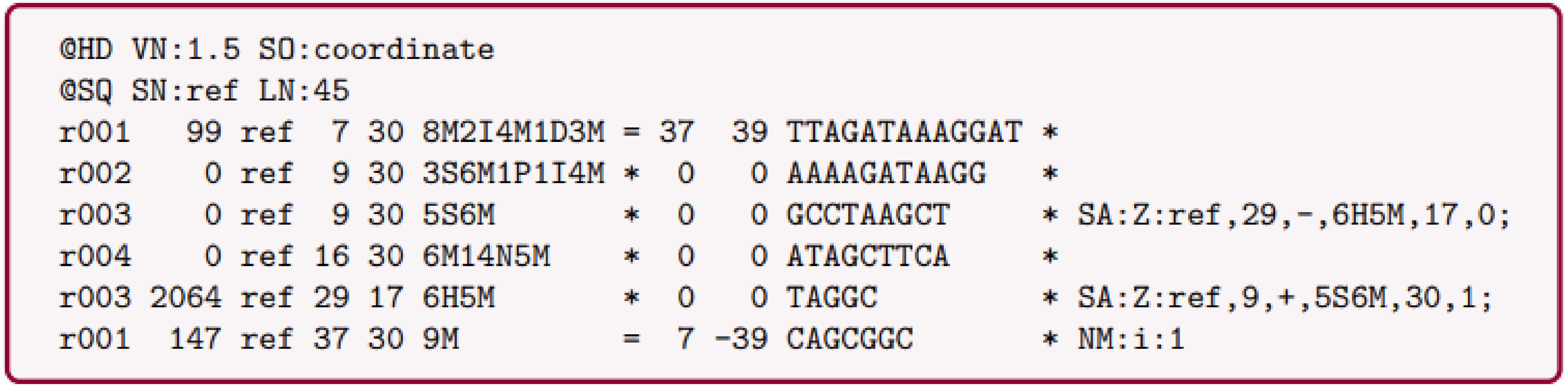
OUTPUT FILE:

The output is generally in SAM (.sam) file format which can be converted into a BAM (.bam) file format.



```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

A) FASTQ file 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
```

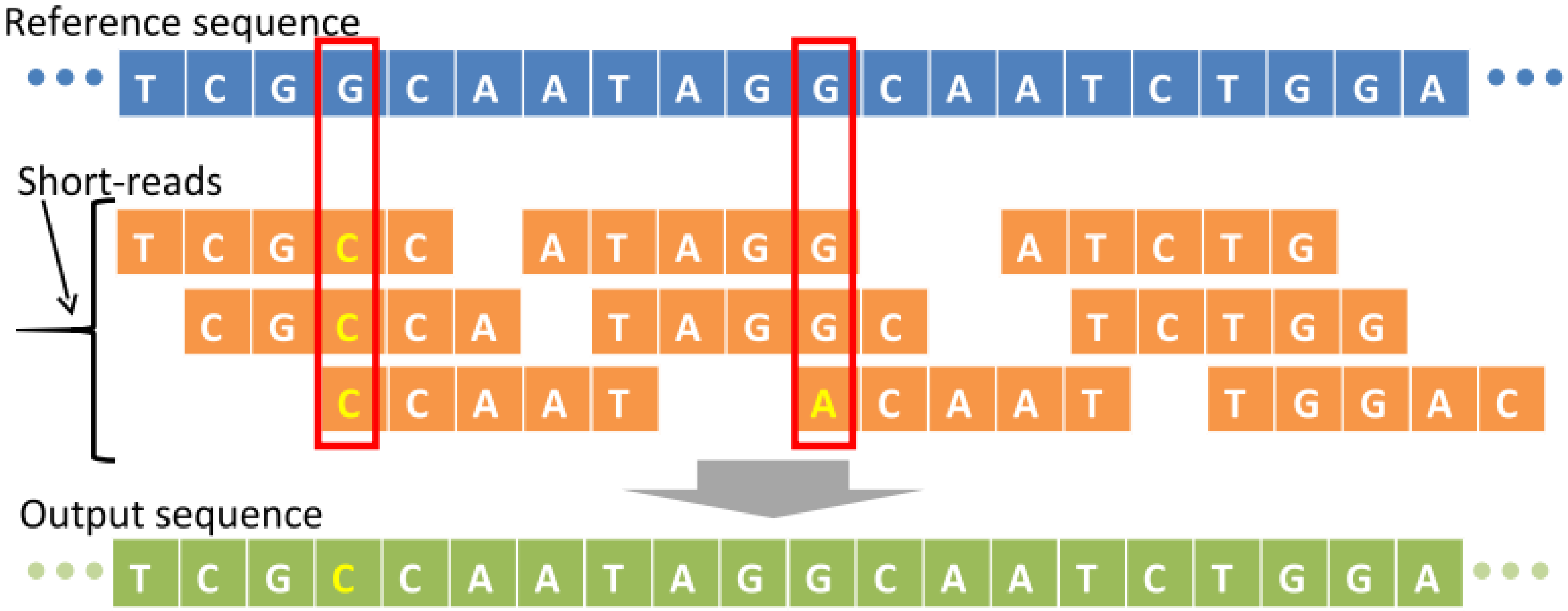
B) SAM file format

Image Source: 10.3390/info7040056

ALGORITHMS:

Feature	**BWA Backtrack**	**BWA SW**	**BWA MEM**
Read Length Suitability	Up to 100bp	70bp to 1Mbp	70bp to 1Mbp
Error Tolerance	Lower	High	High
Speed	Faster	Slower	Fastest
Use	Short, clean Illumina reads	Long reads with potential errors	General purpose, good balance of speed and accuracy

BWA-MEM:


Reference sequence
T C G G C A A T A G G C A A T C T G G A
Short-reads
T C G C C
A T A G G
A T C T G
C G C C A
T A G G C
T C T G G
C C A A T
A C A A T
T G G A C
Output sequence
T C G C C A A T A G G C A A T C T G G A

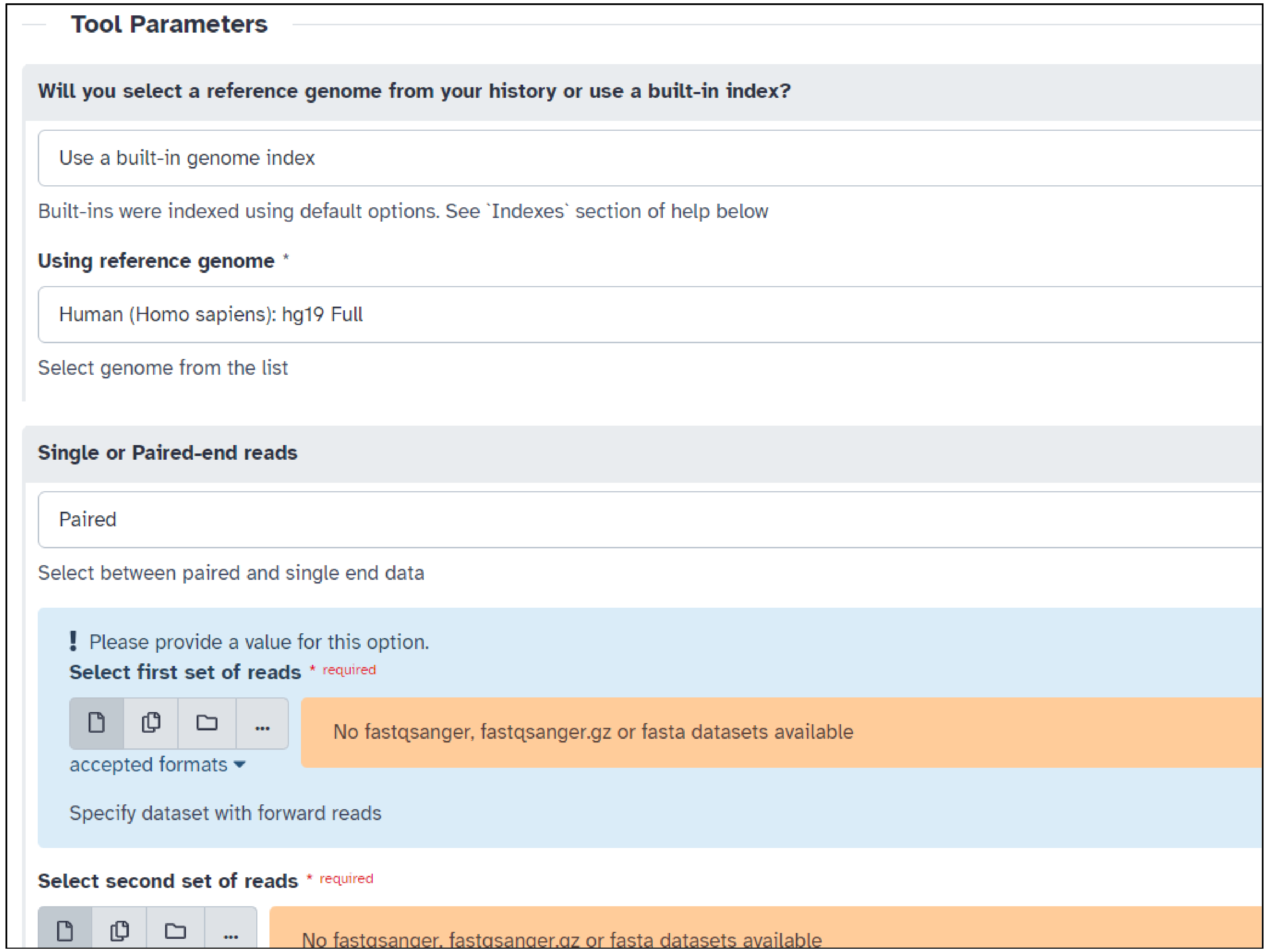
Short Read Mapping | Image Source: doi:10.1109/embc.2013.6609584

- BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate.
- The first step is to transform the genome which is done via the index algorithm and will usually take a few hours (Indexing the genome), it then detects exact matches between the read and the reference (seeding), and then chains them into local or global alignments(extension), it then assigns scores to the alignments based on matches, mismatches, indels. It then filters alignments below a certain quality threshold (scoring).

Map with BWA-MEM: Galaxy Server

- BWA-MEM is an alignment algorithm for aligning sequence reads or long query sequences against a large reference genome such as human. It automatically chooses between local and end-to-end alignments, supports paired-end reads and performs chimeric alignment.
- The first step of transforming the genome is already done in the server for most genomes to avoid wasting time indexing the genome.
- Galaxy server provides output in BAM file format; BAM offers advantages over SAM, the Galaxy server likely converts the BWA output (SAM) to BAM format automatically for users. This eliminates the need for users to perform the conversion themselves and provides the alignments in a more efficient format for downstream analysis.

Map with BWA-MEM: Galaxy Server



Tool Parameters

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome *

Human (Homo sapiens): hg19 Full

Select genome from the list

Single or Paired-end reads

Paired

Select between paired and single end data

! Please provide a value for this option.

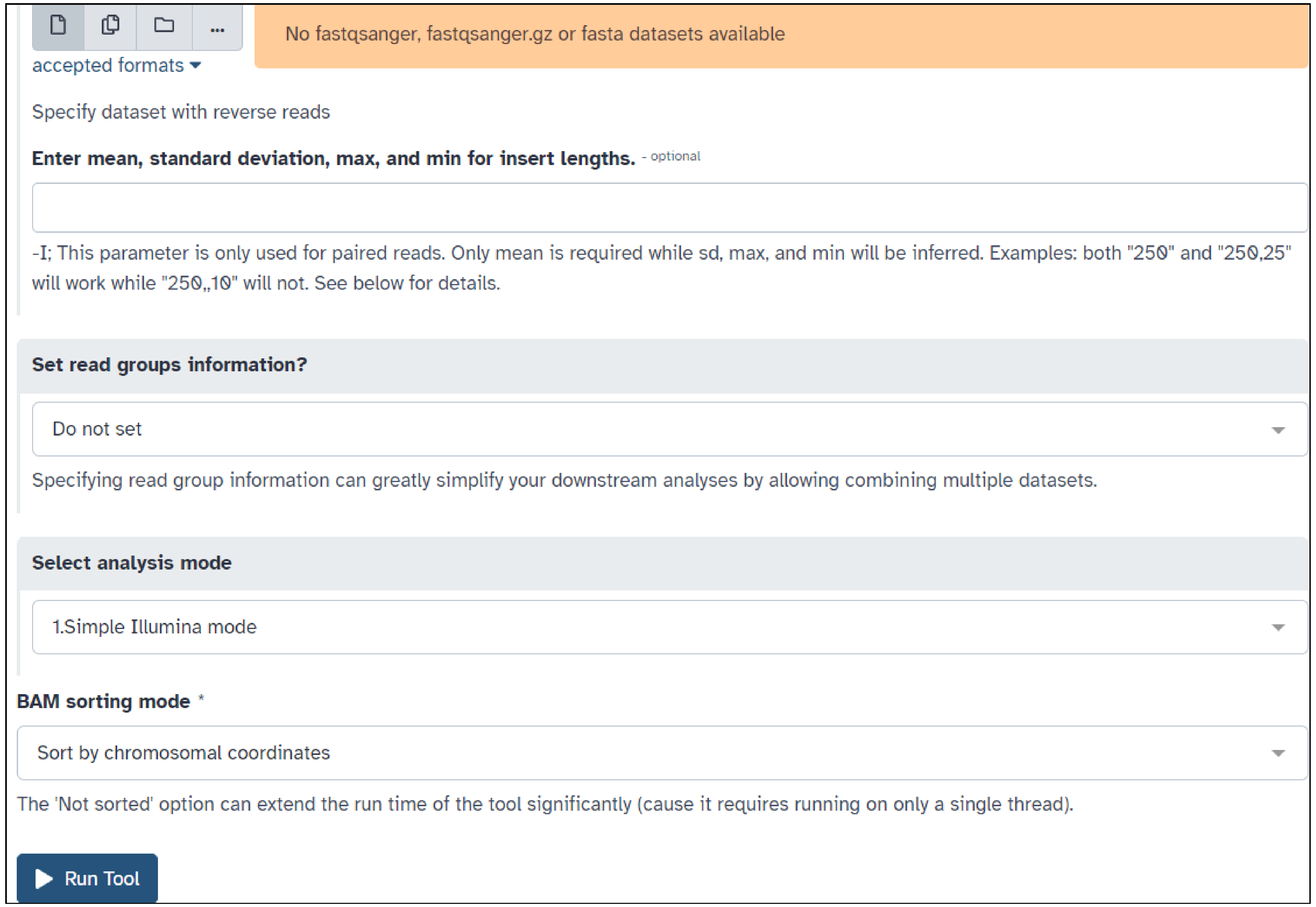
Select first set of reads * required

No fastqsanger, fastqsanger.gz or fasta datasets available

Specify dataset with forward reads

Select second set of reads * required

No fastqsanger, fastqsanger.gz or fasta datasets available



No fastqsanger, fastqsanger.gz or fasta datasets available

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths. - optional

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1.Simple Illumina mode

BAM sorting mode *

Sort by chromosomal coordinates

The 'Not sorted' option can extend the run time of the tool significantly (cause it requires running on only a single thread).

Tool Source: https://usegalaxy.eu/root?tool_id=toolshed.g2.bx.psu.edu/repos/devteam/bwa/bwa_mem/0.7.17.2

-BOWTIE-

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s of characters to relatively long (e.g. mammalian) genomes.

ALIGNMENT TOOLS:

BOWTIE:

- Bowtie is a tool used to align short sequence reads to a reference genome.
- It accepts the file in a FASTA/FASTQ format.
- The output of the aligned reads is generated in BAM/SAM file formats.

BowTie 1 vs. BowTie 2:

- Bowtie 2 allows alignments to overlap ambiguous characters (e.g. 'NNN..') in the reference whereas Bowtie 1 does not.
- Bowtie 2 is generally faster and more sensitive for reads longer than about 50 bp. It also uses less memory than Bowtie 1.
- Although for relatively short reads (less than 50 bp) Bowtie 1 is sometimes faster and/or more sensitive.
- Bowtie 2 does not align color space reads. (We use trimmomatic)

How does it work?

Step 1: Extracting Seeds

- Read is chopped into small piece, reverse component is also taken into consideration

Step 2: Quick Alignment using Index

- Uses FM-Index to map reads to reference genome, no gaps/insertions are allowed

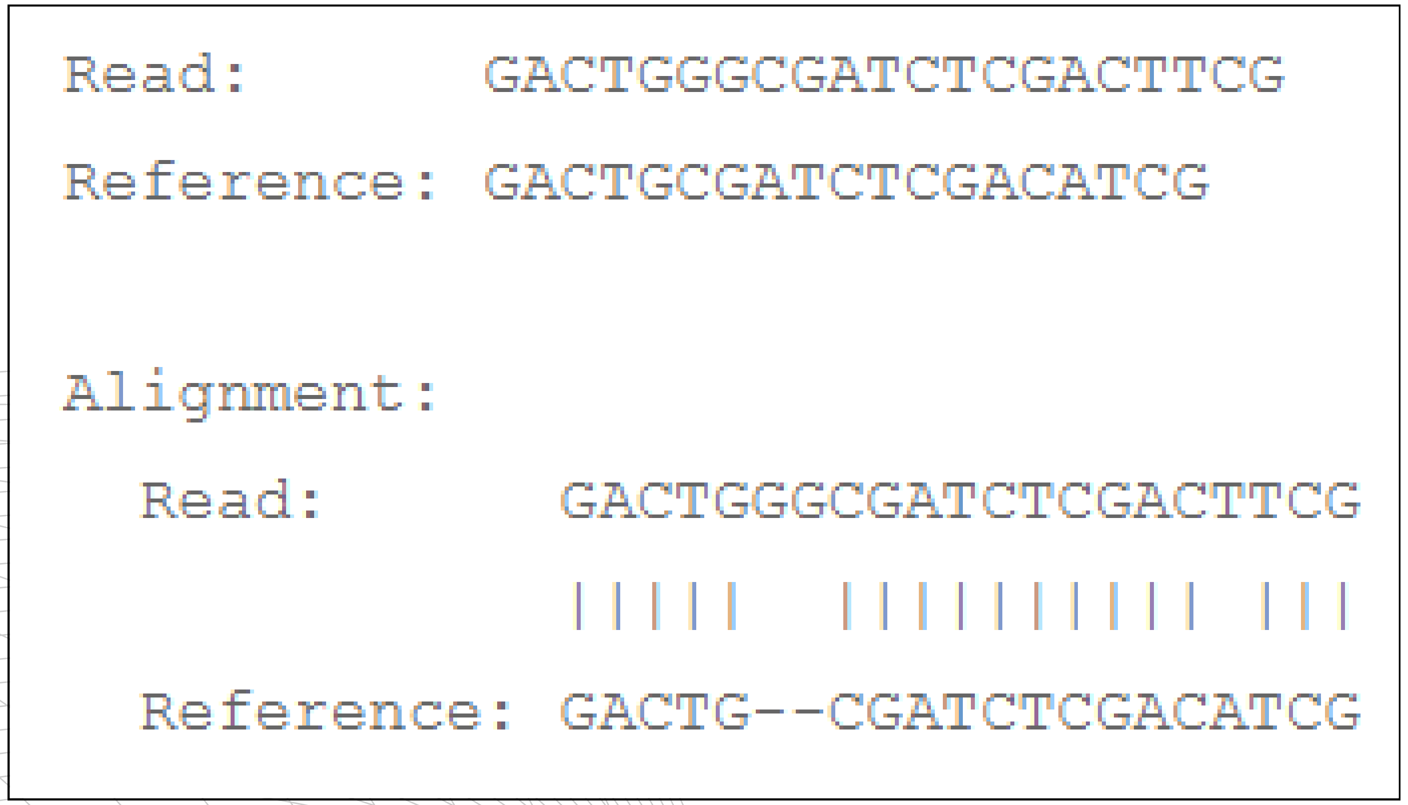
Step 3: Prioritizing Seed Alignments

- As not all seeds are equally good, we prioritize it based on many factors such as Location of mismatches, Number of mismatches

Step 4: Extending Seeds into Full Alignments

- After prioritizing seed alignments, the seed is extended upwards as well as downwards onto the reference genome, gaps between seeds is filled and the short sequence read is aligned with the reference genome

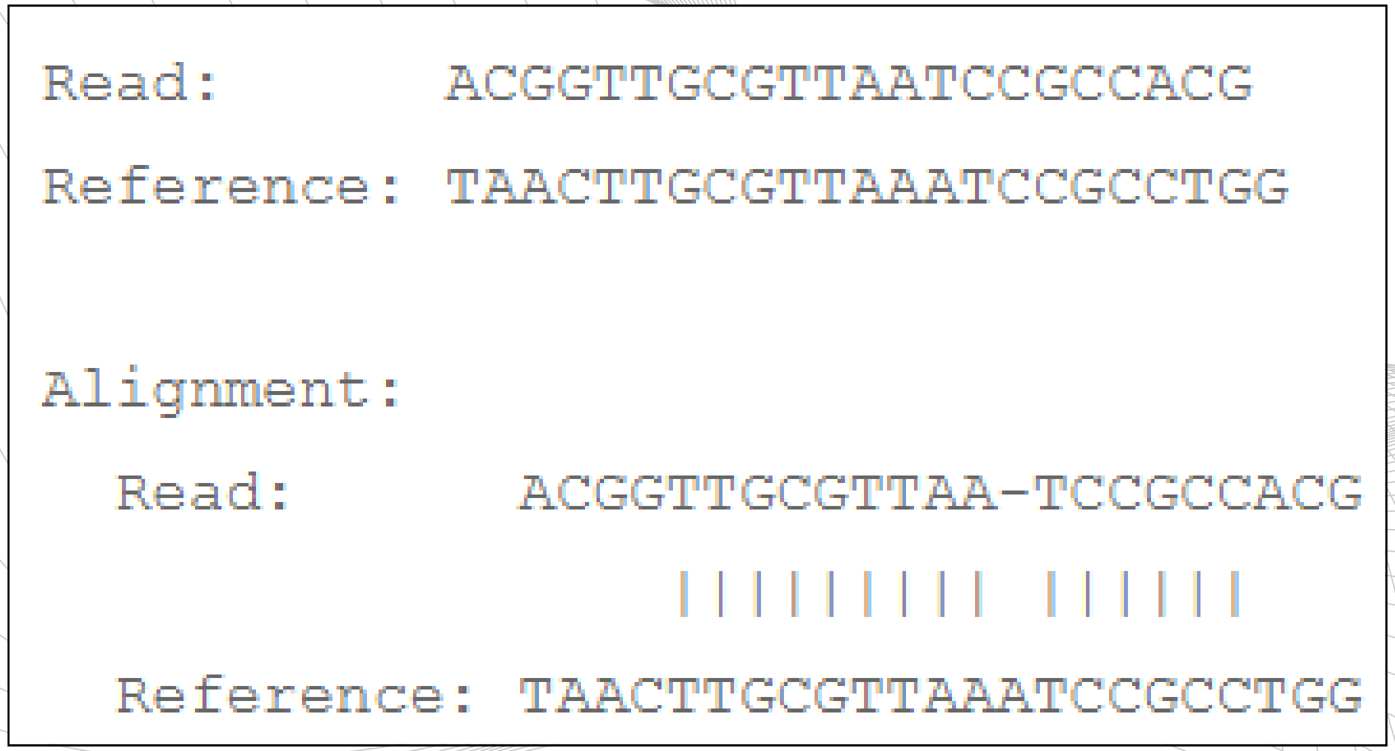
Global Alignment using BOWTIE 2:



```
Read:      GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG


Alignment:
  Read:      GACTGGGCGATCTCGACTTCG
             |||||  |||||||||| |||
  Reference: GACTG--CGATCTCGACATCG
```

Global alignment mapping

Local alignment using BOWTIE 2:



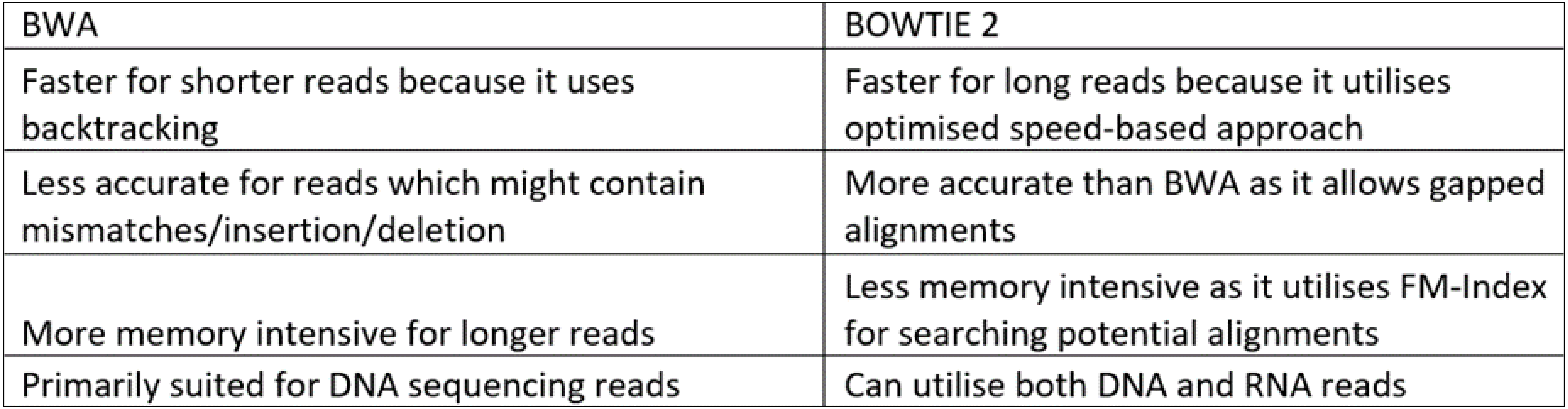
```
Read:      ACGGTTGCGTTAATCCGCCACG
Reference: TAACTTGCGTTAAATCCGCCTGG


Alignment:
  Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
  Reference: TAACTTGCGTTAAATCCGCCTGG
```

Local alignment mapping

Image Source: https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

WHEN TO USE BWA & BOWTIE 2:

- We use BWA when we want **faster results for shorter reads** and when **computational speed is low.**
- We use Bowtie 2 when **we need to analyze both DNA and RNA** sequencing data and when **we need to analyze longer reads**.



BWA	BOWTIE 2
Faster for shorter reads because it uses backtracking	Faster for long reads because it utilises optimised speed-based approach
Less accurate for reads which might contain mismatches/insertion/deletion	More accurate than BWA as it allows gapped alignments
More memory intensive for longer reads	Less memory intensive as it utilises FM-Index for searching potential alignments
Primarily suited for DNA sequencing reads	Can utilise both DNA and RNA reads

-SAMTools-

After reads have been aligned and processed, the next step in the pipeline is to identify differences observed between the selected reference genome and the newly sequenced reads. One such tool used is discussed further.

SAMTools:

Functions of SAM Tools:

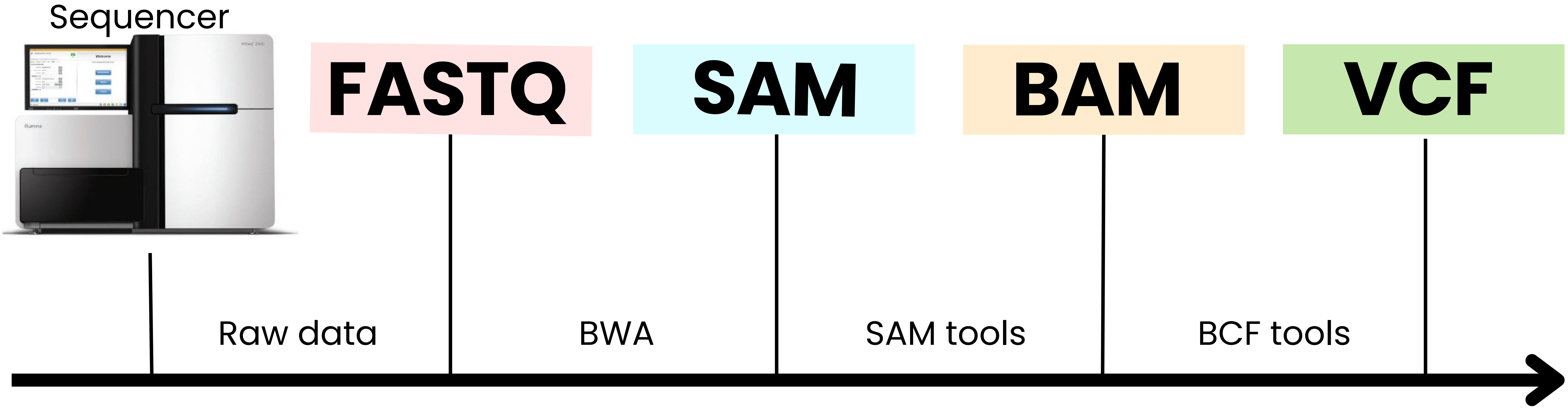
1. Variant calling: While there are dedicated tools for variant calling, Samtools offers basic functionalities to identify potential single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) from alignments.
2. Analysis Tasks: Samtools offers a text-based viewer for alignments, allowing researchers to inspect how reads map to the reference genome.
3. Sorting and Indexing: BAM files are large and can be slow to search through linearly. Samtools can sort BAM files based on their genomic location and create indexes for faster retrieval of specific regions.
4. Statistical Analysis: Samtools provides tools to calculate statistics from BAM files, such as coverage depth (how many times each base in the genome is sequenced) and base quality scores.

USED IN NGS

-OUTPUT FORMATS-

The output formats in an NGS pipeline vary depending on the specific step within the pipeline. The output format in the pre-processing step was FASTQ but as we move further in the analysis the format changes to SAM/BAM/CRAM etc. The alignment tools discussed previously generate files the formats discussed further

OUTPUT FORMATS:SAM/BAM/CRAM



Sequencer

FASTQ

SAM

BAM

VCF

Raw data

BWA

SAM tools

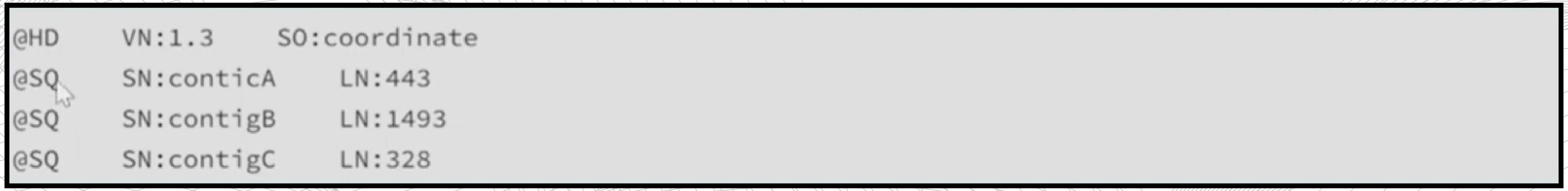
BCF tools

OUTPUT FORMATS:

1. **File Format:** SAM is a text-based format, while BAM and CRAM are binary formats.
2. **Compression:** BAM is a compressed version of SAM, while CRAM is a compressed format that uses reference-based compression, which makes it more efficient in terms of storage space compared to BAM.
3. **Data Storage:** CRAM records use bits as well as bytes for data storage, allowing for more efficient data storage compared to BAM. Data series that do not require binary coding can be stored separately in external blocks with some other compression applied to them independently.
4. **File Identifier:** Each CRAM file starts with a fixed length (26 bytes) definition with the file id CRAM file identifier (e.g. file name or SHA1 checksum)

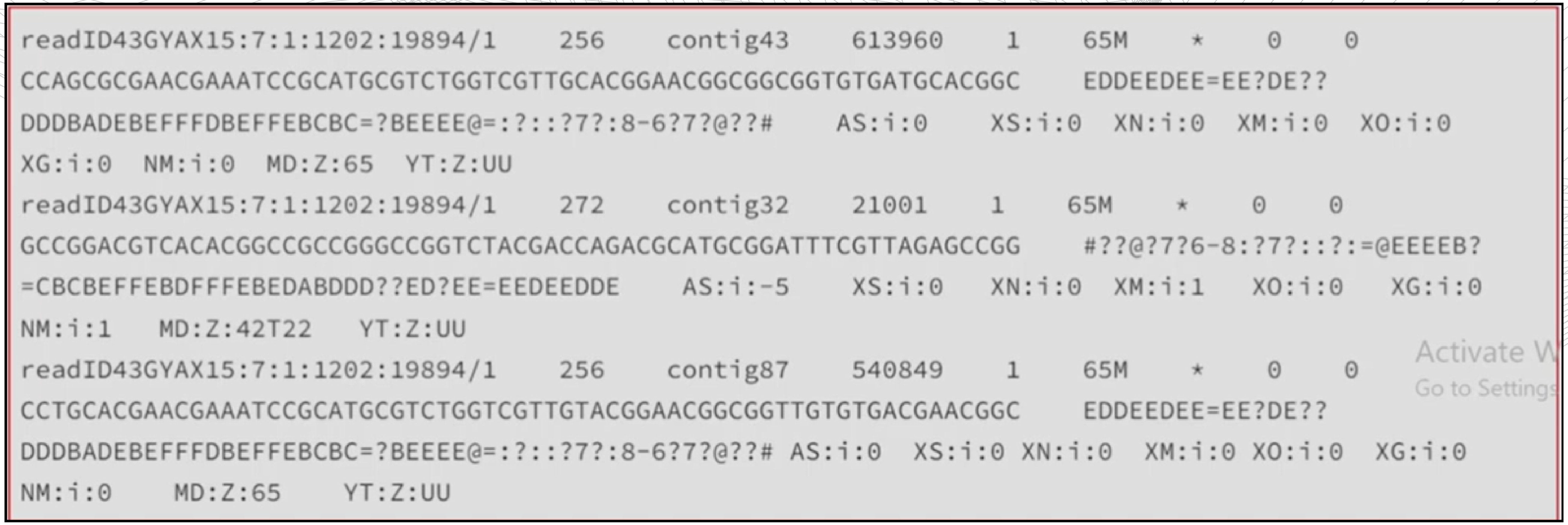
OUTPUT FORMATS:
Eg:- SAM FILE FORMAT

HEADER



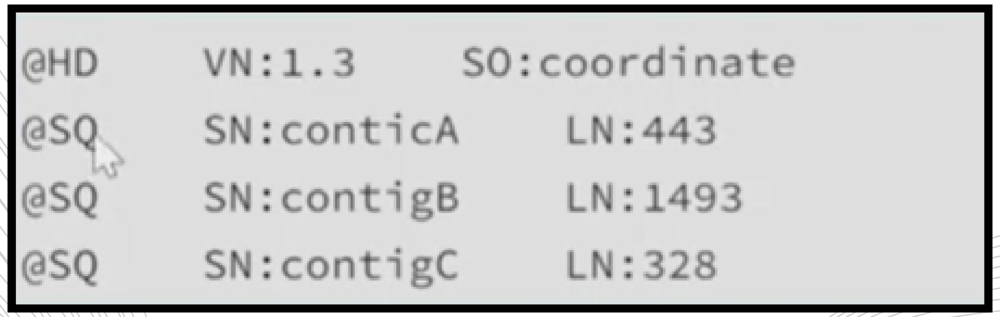
```
@HD    VN:1.3    SO:coordinate
@SQ    SN:conticA    LN:443
@SQ    SN:contigB    LN:1493
@SQ    SN:contigC    LN:328
```

ALIGNMENT



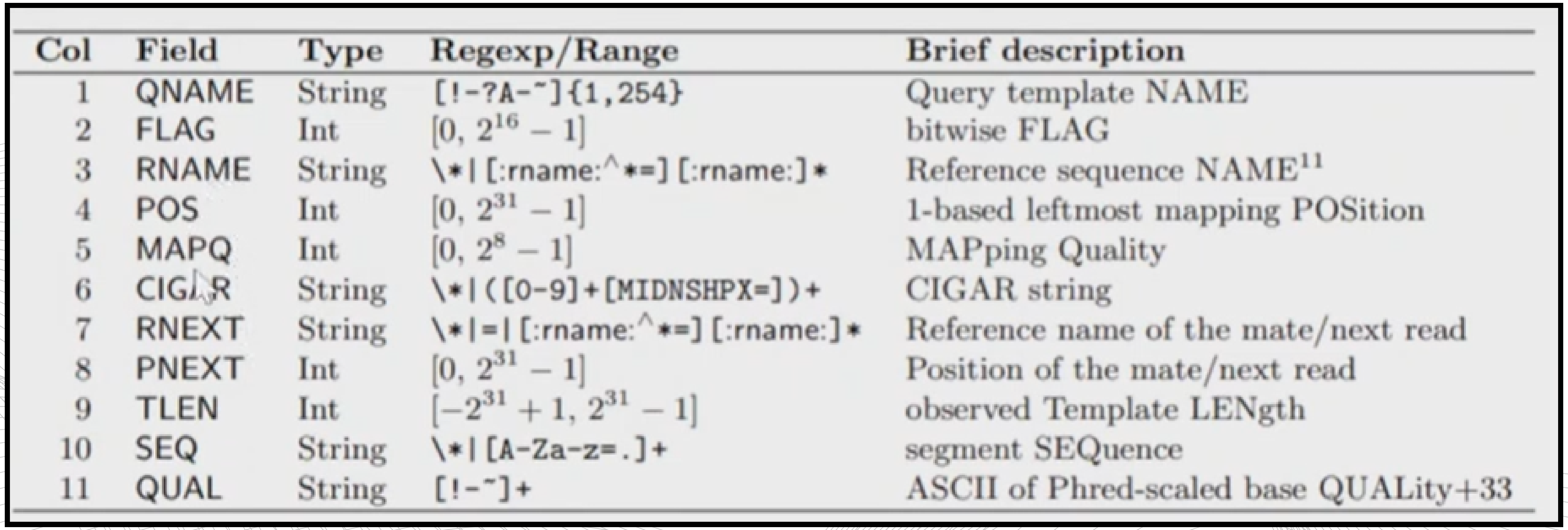
```
readID43GYAX15:7:1:1202:19894/1    256    contig43    613960    1    65M    *    0    0
CCAGCGCGAACGAAATCCGCATGCGTCTGGTCGTTGCACGGAACGGCGGCGGTGTGATGCACGGC    EDDEEDEE=EE?DE??
DDDBADEBEFFFDBEFFEBCBC=?BEEEE@=:?::?7?:8-6?7?@??#    AS:i:0    XS:i:0   XN:i:0   XM:i:0   XO:i:0
XG:i:0   NM:i:0   MD:Z:65   YT:Z:UU
readID43GYAX15:7:1:1202:19894/1    272    contig32    21001    1    65M    *    0    0
GCCGGACGTCACACGGCCGCCGGGCCGGTCTACGACCAGACGCATGCGGATTTCGTTAGAGCCGG    #??@?7?6-8:?7?::?:=@EEEEB?
=CBCBEFFEBDFFFEBEDABDDD??ED?EE=EEDEEDDE    AS:i:-5    XS:i:0    XN:i:0   XM:i:1    XO:i:0    XG:i:0
NM:i:1    MD:Z:42T22    YT:Z:UU
readID43GYAX15:7:1:1202:19894/1    256    contig87    540849    1    65M    *    0    0
CCTGCACGAACGAAATCCGCATGCGTCTGGTCGTTGTACGGAACGGCGGTTGTGTGACGAACGGC    EDDEEDEE=EE?DE??
DDDBADEBEFFFDBEFFEBCBC=?BEEEE@=:?::?7?:8-6?7?@??# AS:i:0   XS:i:0 XN:i:0   XM:i:0 XO:i:0   XG:i:0
NM:i:0    MD:Z:65    YT:Z:UU
```

Activate W
Go to Settings



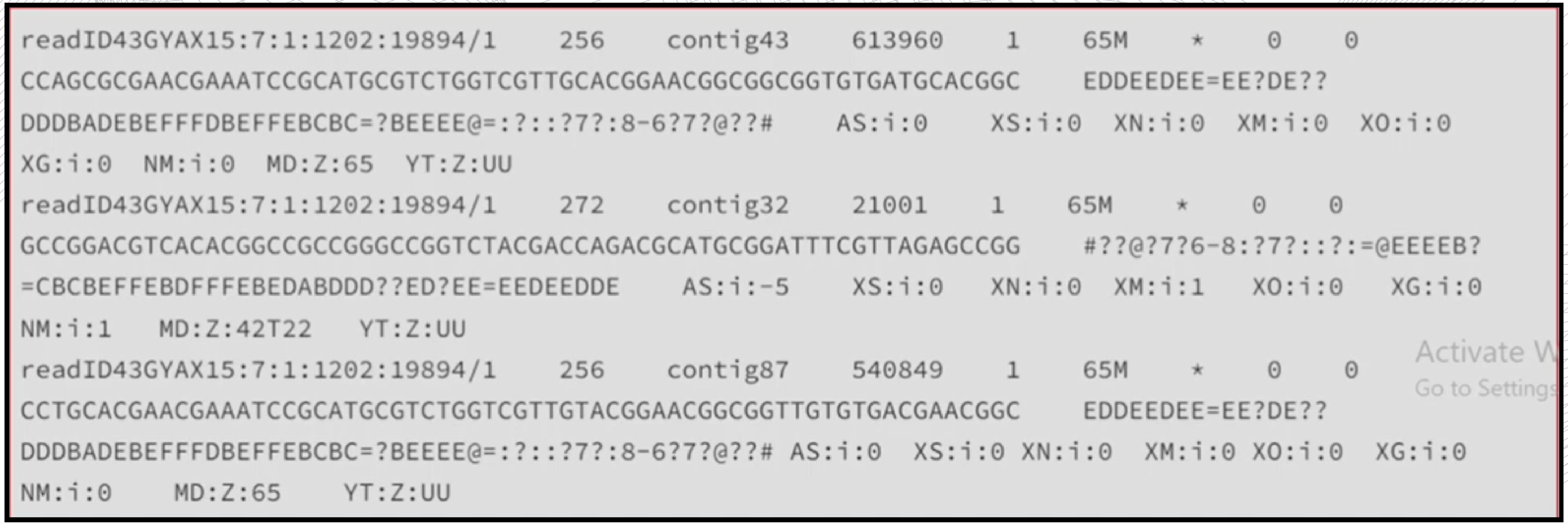
```
@HD     VN:1.3     SO:coordinate
@SQ     SN:conticA      LN:443
@SQ     SN:contigB      LN:1493
@SQ     SN:contigC      LN:328
```

HEADER



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	*|[:rname:^*=][:rname:]*	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[:rname:^*=][:rname:]*	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

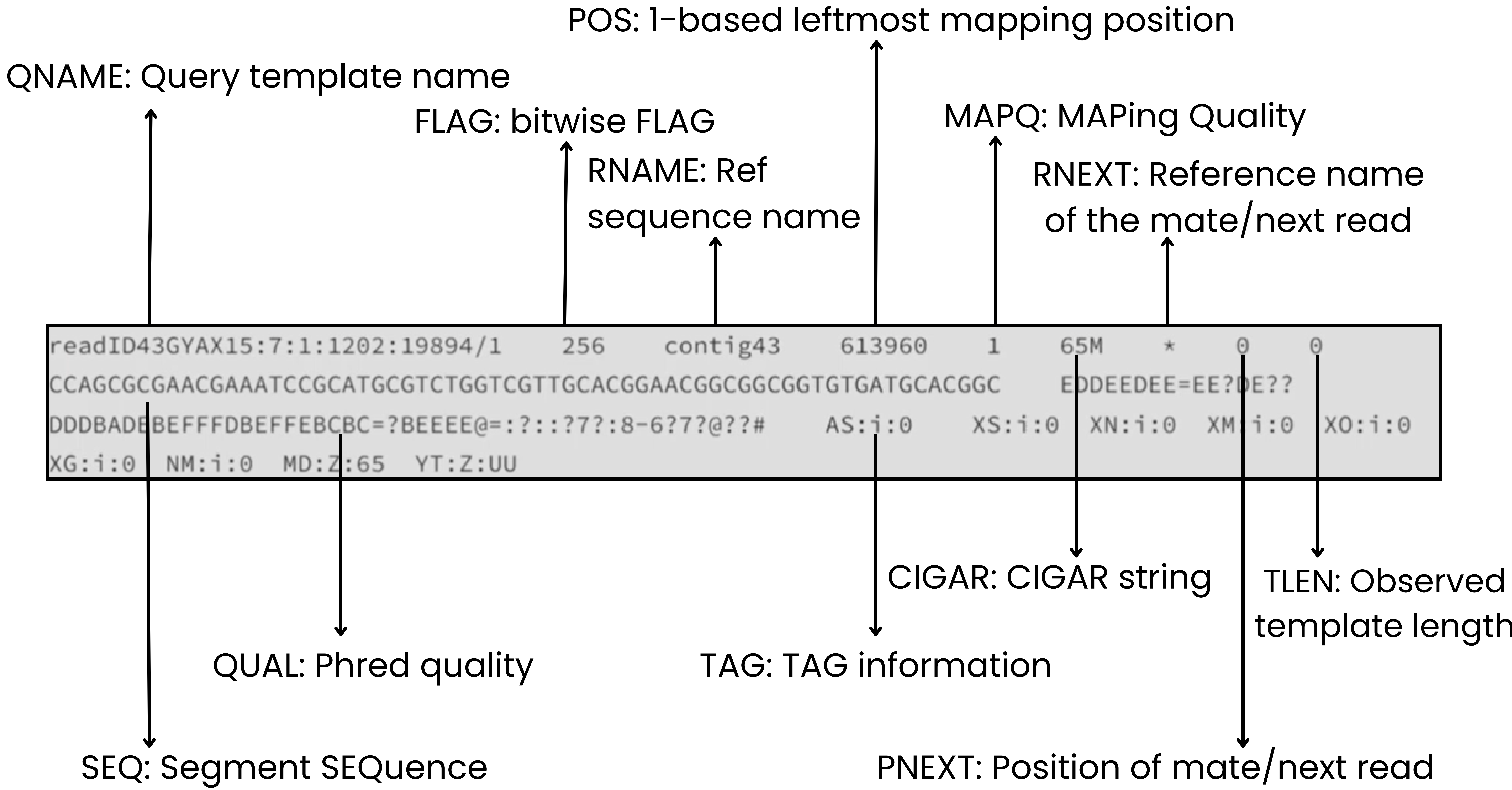
FIELD description

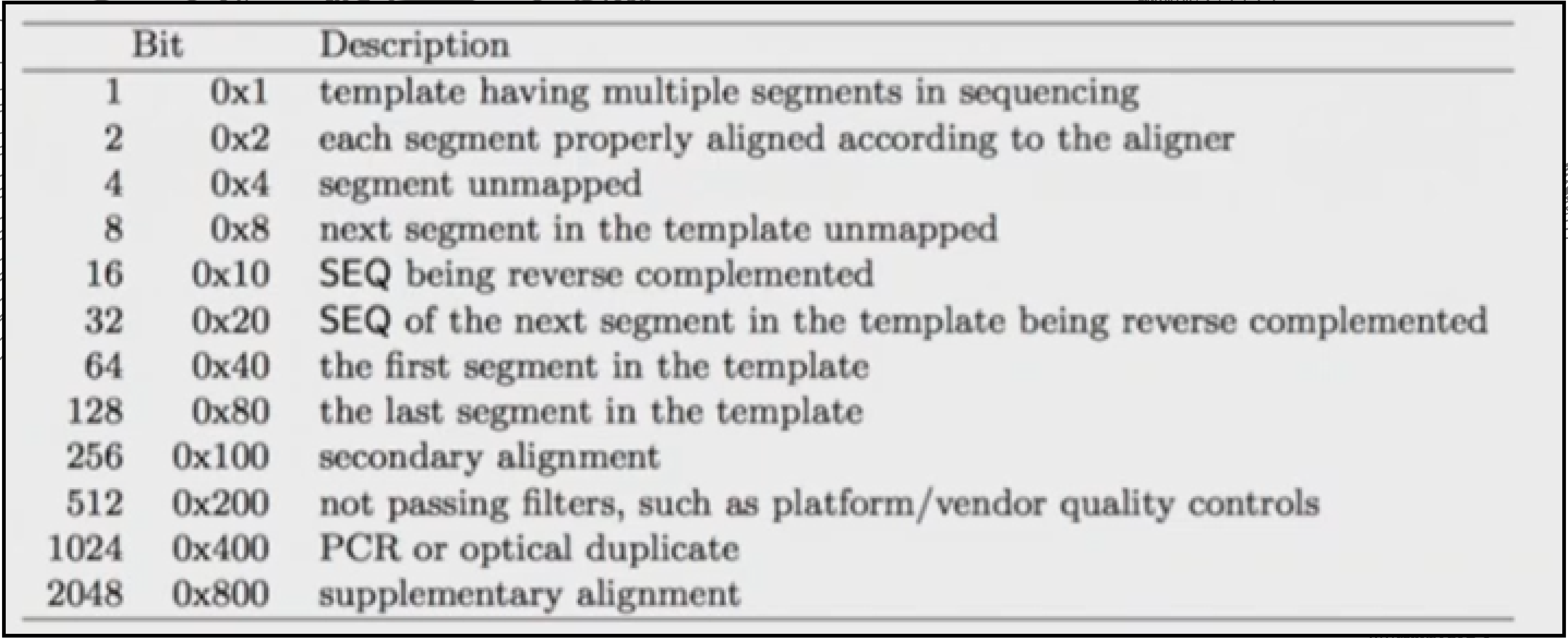


```
readID43GYAX15:7:1:1202:19894/1    256    contig43    613960    1    65M    *    0    0
CCAGCGCGAACGAAATCCGCATGCGTCTGGTCGTTGCACGGAACGGCGGCGGTGTGATGCACGGC    EDDEEDEE=EE?DE??
DDDBADEBEFFFDBEFFEBCBC=?BEEEE@=:?::?7?:8-6?7?@??#    AS:i:0    XS:i:0   XN:i:0   XM:i:0   XO:i:0
XG:i:0   NM:i:0   MD:Z:65   YT:Z:UU
readID43GYAX15:7:1:1202:19894/1    272    contig32    21001    1    65M    *    0    0
GCCGGACGTCACACGGCCGCCGGGCCGGTCTACGACCAGACGCATGCGGATTTCGTTAGAGCCGG    #??@?7?6-8:?7?::?:=@EEEEB?
=CBCBEFFEBDFFFEBEDABDDD??ED?EE=EEDEEDDE    AS:i:-5    XS:i:0    XN:i:0   XM:i:1    XO:i:0    XG:i:0
NM:i:1    MD:Z:42T22    YT:Z:UU
readID43GYAX15:7:1:1202:19894/1    256    contig87    540849    1    65M    *    0    0
CCTGCACGAACGAAATCCGCATGCGTCTGGTCGTTGTACGGAACGGCGGTTGTGTGACGAACGGC    EDDEEDEE=EE?DE??
DDDBADEBEFFFDBEFFEBCBC=?BEEEE@=:?::?7?:8-6?7?@??# AS:i:0   XS:i:0 XN:i:0   XM:i:0 XO:i:0   XG:i:0
NM:i:0    MD:Z:65    YT:Z:UU
```

Activate W
Go to Settings

ALIGNMENT

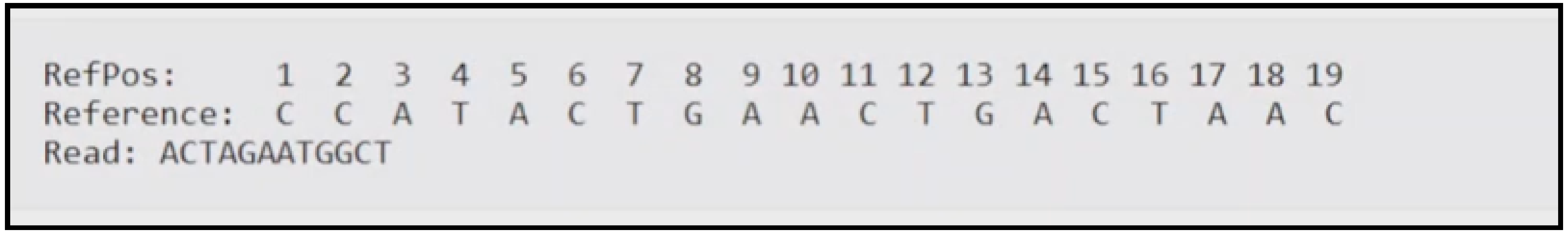




Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

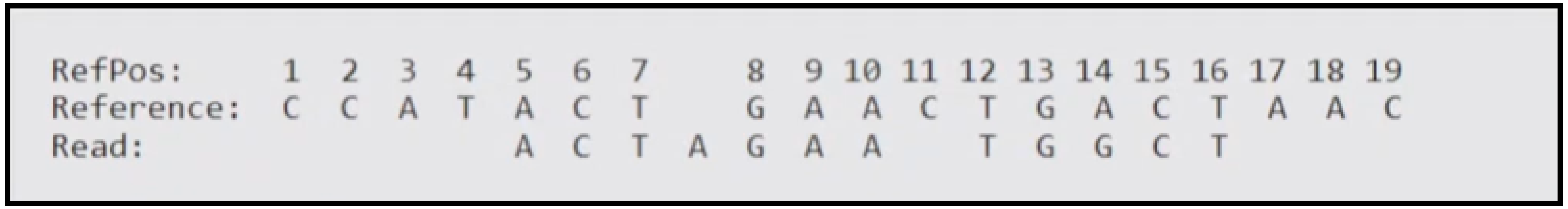
bitwise FLAG description

COMPARISON



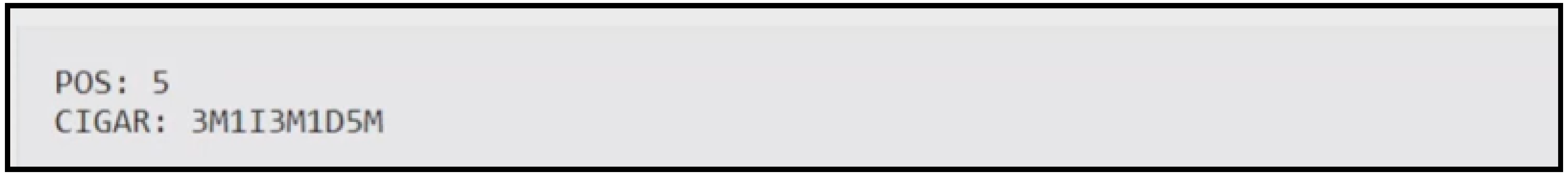
```
RefPos:     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

Aligning these two

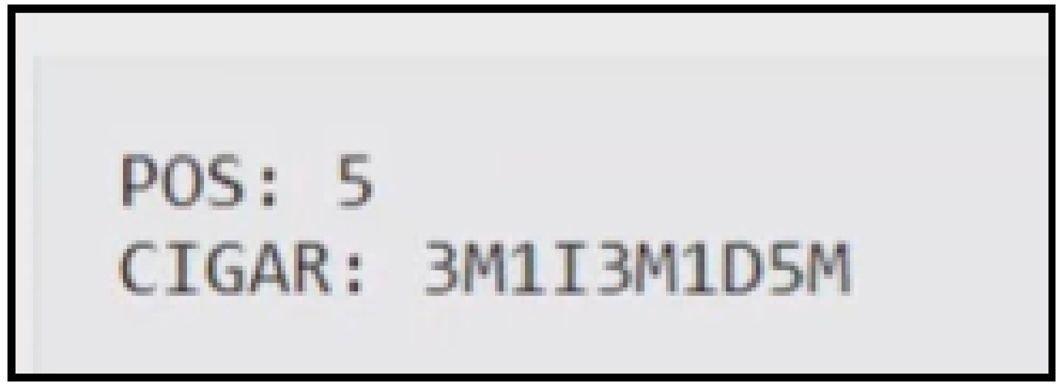


```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                   A  C  T  A  G  A  A     T  G  G  C  T
```

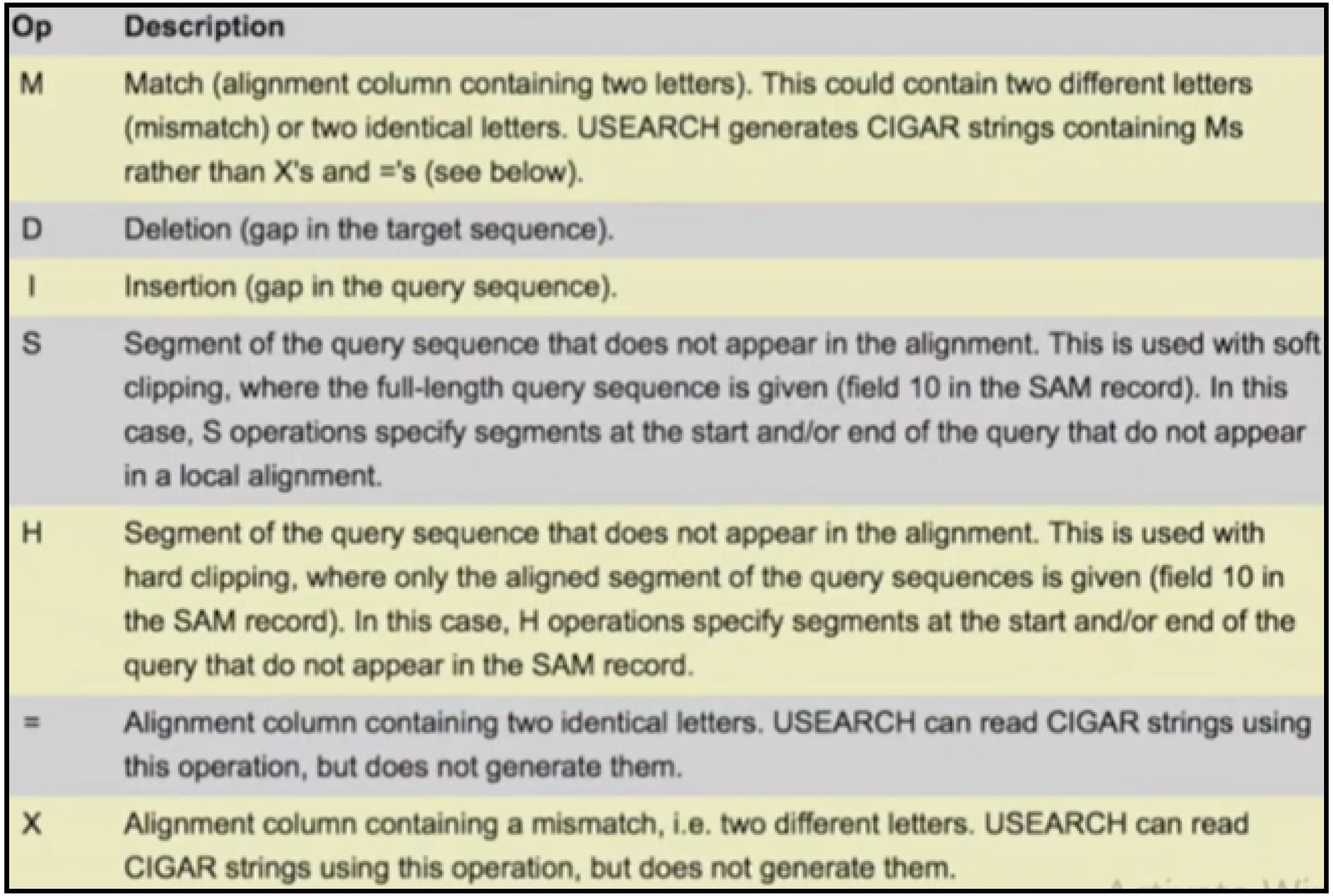
With the alignment above you get:



```
POS: 5
CIGAR: 3M1I3M1D5M
```



```
POS: 5
CIGAR: 3M1I3M1D5M
```



Op	Description
M	Match (alignment column containing two letters). This could contain two different letters (mismatch) or two identical letters. USEARCH generates CIGAR strings containing Ms rather than X's and ='s (see below).
D	Deletion (gap in the target sequence).
I	Insertion (gap in the query sequence).
S	Segment of the query sequence that does not appear in the alignment. This is used with soft clipping, where the full-length query sequence is given (field 10 in the SAM record). In this case, S operations specify segments at the start and/or end of the query that do not appear in a local alignment.
H	Segment of the query sequence that does not appear in the alignment. This is used with hard clipping, where only the aligned segment of the query sequences is given (field 10 in the SAM record). In this case, H operations specify segments at the start and/or end of the query that do not appear in the SAM record.
=	Alignment column containing two identical letters. USEARCH can read CIGAR strings using this operation, but does not generate them.
X	Alignment column containing a mismatch, i.e. two different letters. USEARCH can read CIGAR strings using this operation, but does not generate them.

CIGAR description

-IGV-

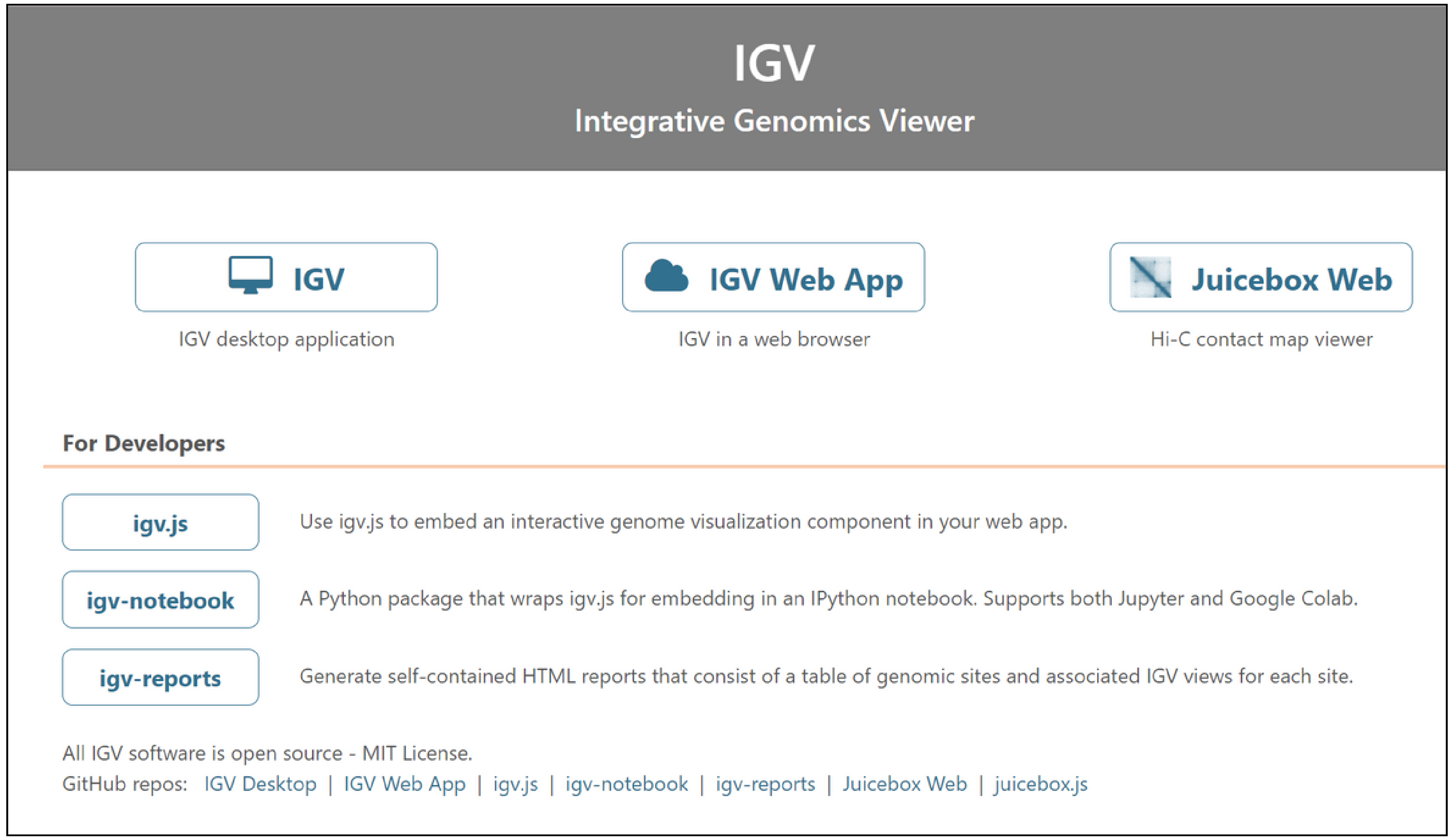
INTEGRATIVE GENOMICS VIEWER

The downstream analysis of the pipeline involves visualization of the aligned reads to a reference genome and plays a crucial role in interpreting the results as it allows researchers to visually inspect the alignment of reads, identified variants, and other genomic features.

VISUALISATION TOOL:

Integrative Genomics Viewer

- Visualizing tool for exploring and analyzing large genome datasets
- The Broad Institute of MIT and Harvard developed IGV, and Illumina modified it
- Work with prebuilt genome or import a gene
- View a variety of data such as expression data, NGS alignment, microarray, epigenomics, RNA-seq, genomic annotations etc.
- Can run locally on machine or on website
- IGV is a JAVA application



IGV
Integrative Genomics Viewer

IGV desktop application
IGV in a web browser
Hi-C contact map viewer

For Developers

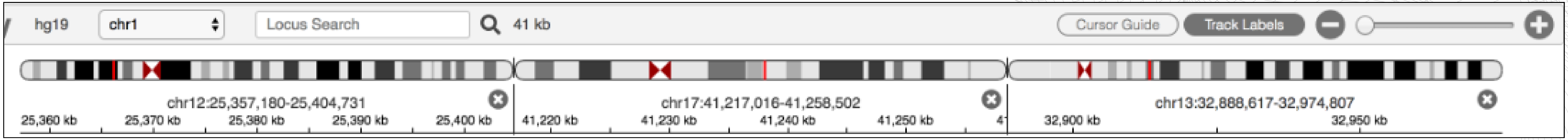
igv.js	Use igv.js to embed an interactive genome visualization component in your web app.
igv-notebook	A Python package that wraps igv.js for embedding in an IPython notebook. Supports both Jupyter and Google Colab.
igv-reports	Generate self-contained HTML reports that consist of a table of genomic sites and associated IGV views for each site.

All IGV software is open source - MIT License.
GitHub repos: IGV Desktop | IGV Web App | igv.js | igv-notebook | igv-reports | Juicebox Web | juicebox.js

Homepage of IGV
(Source: https://igv.org/)

Searching Patterns:

1. Chromosome based
2. Coordinate based (e.g. chr17:41,195,312-41,278,500)
3. Gene based (e.g. BRCA1)
4. Multiple loci (e.g. BRCA2 BRCA2 KRAS)


hg19
chr1
Locus Search
41 kb
Cursor Guide
Track Labels
chr12:25,357,180-25,404,731
25,360 kb
25,370 kb
25,380 kb
25,390 kb
25,400 kb
chr17:41,217,016-41,258,502
41,220 kb
41,230 kb
41,240 kb
41,250 kb
chr13:32,888,617-32,974,807
32,900 kb
32,950 kb

Searching Patterns

File Formats:

IGV supports a variety of genomic file formats, which can be divided into three categories:

1. Non-indexed (GFF,BED,WIG)
2. Indexed (BAM)
3. Multiresolution (TDF)

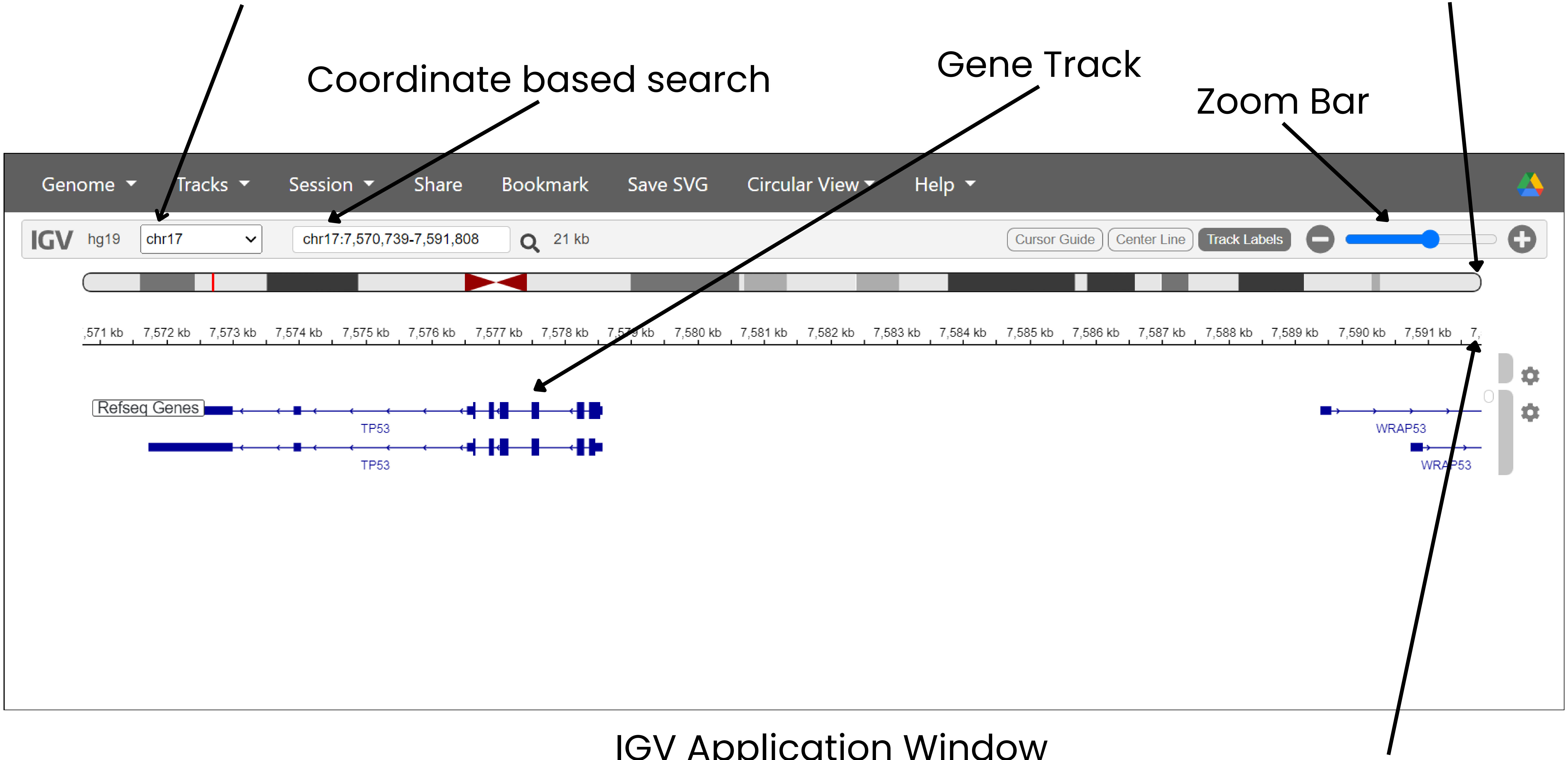
Chromosome-based search

Coordinate based search

Gene Track

Zoom Bar

Chromosome Ideogram



Genome
Tracks
Session
Share
Bookmark
Save SVG
Circular View
Help
IGV
Cursor Guide
Center Line
Track Labels
,571 kb
7,572 kb
7,573 kb
7,574 kb
7,575 kb
7,576 kb
7,577 kb
7,578 kb
7,579 kb
7,580 kb
7,581 kb
7,582 kb
7,583 kb
7,584 kb
7,585 kb
7,586 kb
7,587 kb
7,588 kb
7,589 kb
7,590 kb
7,591 kb
7,
Refseq Genes
TP53
TP53
WRAP53
WRAP53

hg19

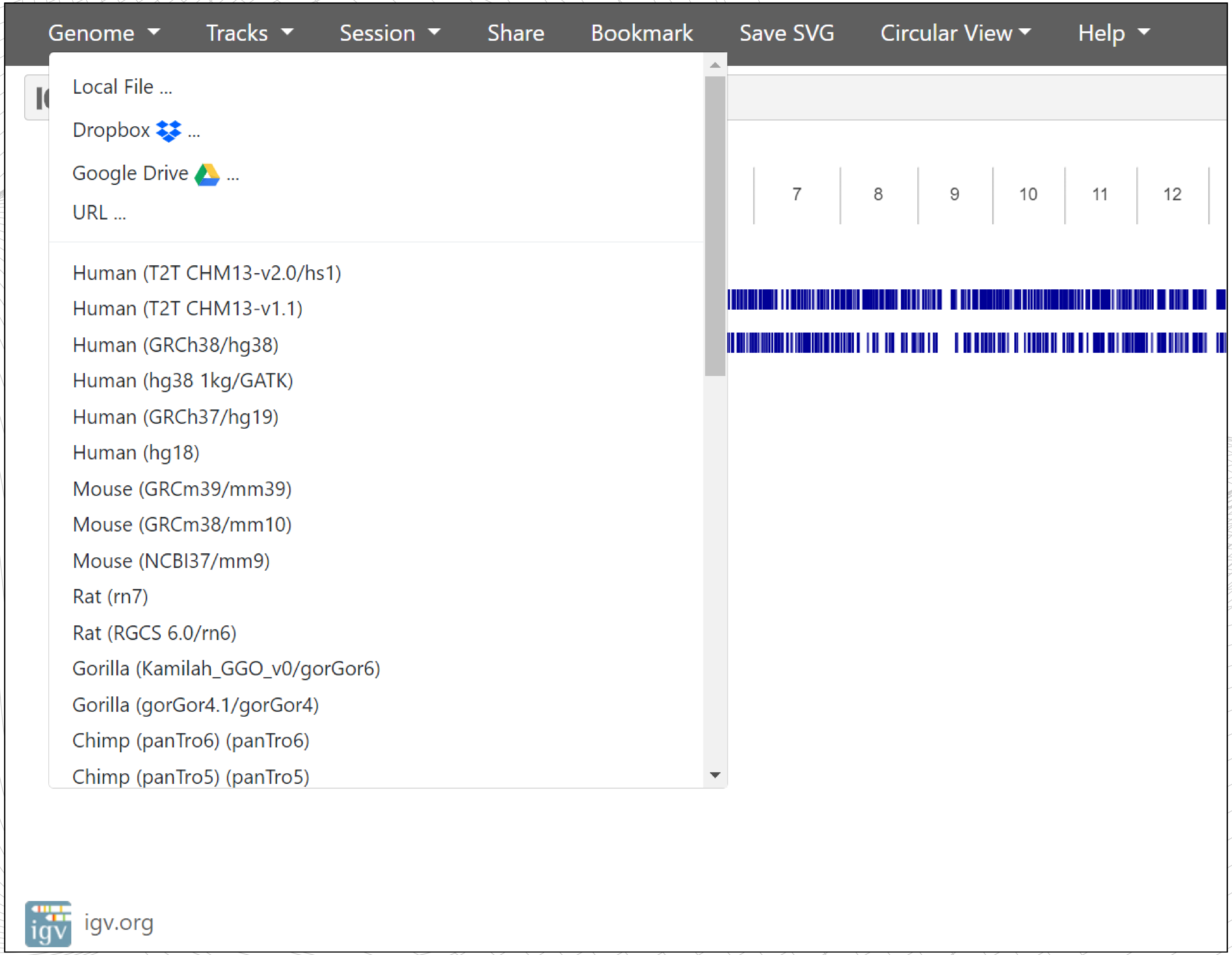
chr17

chr17:7,570,739-7,591,808

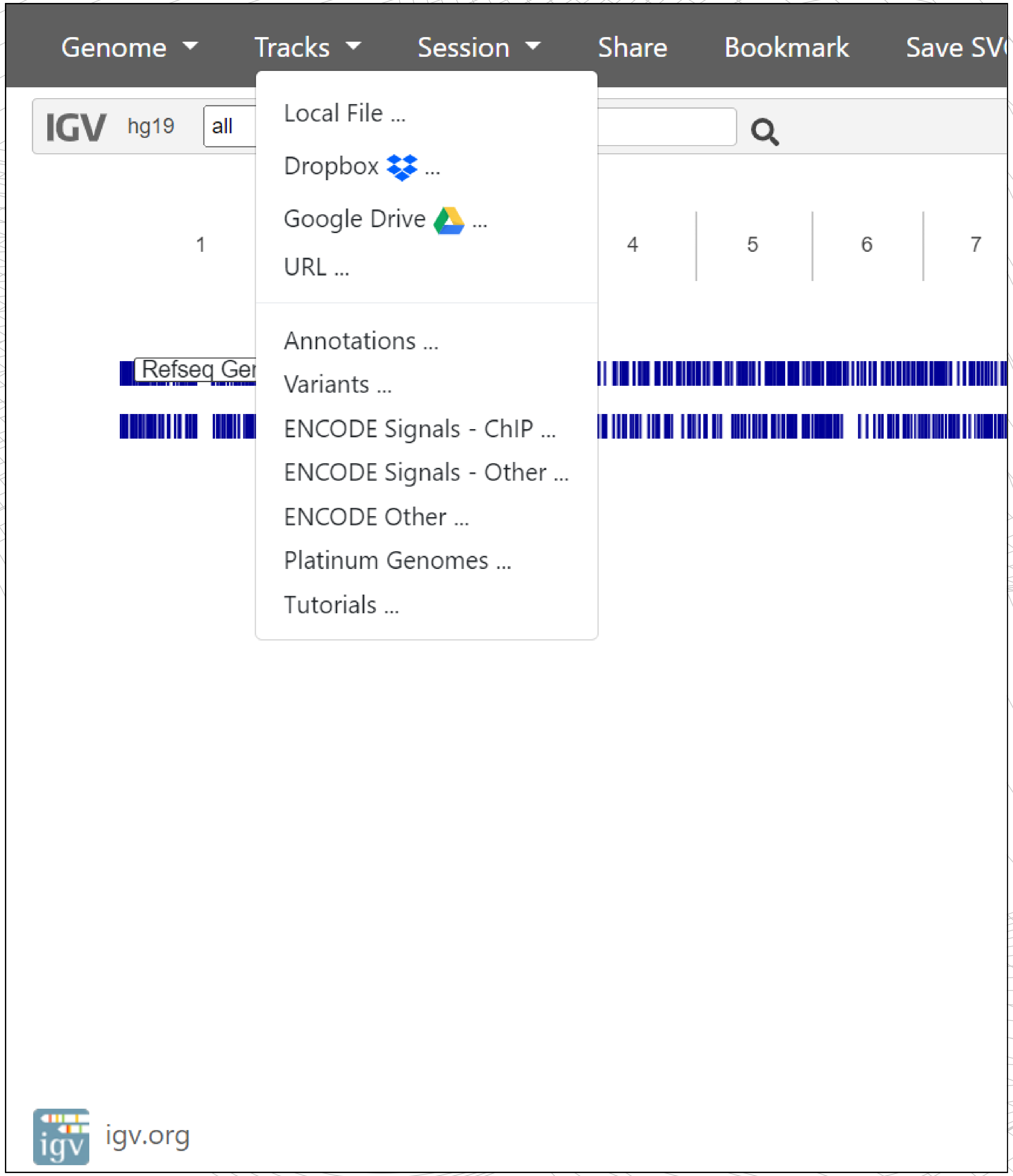
21 kb

IGV Application Window

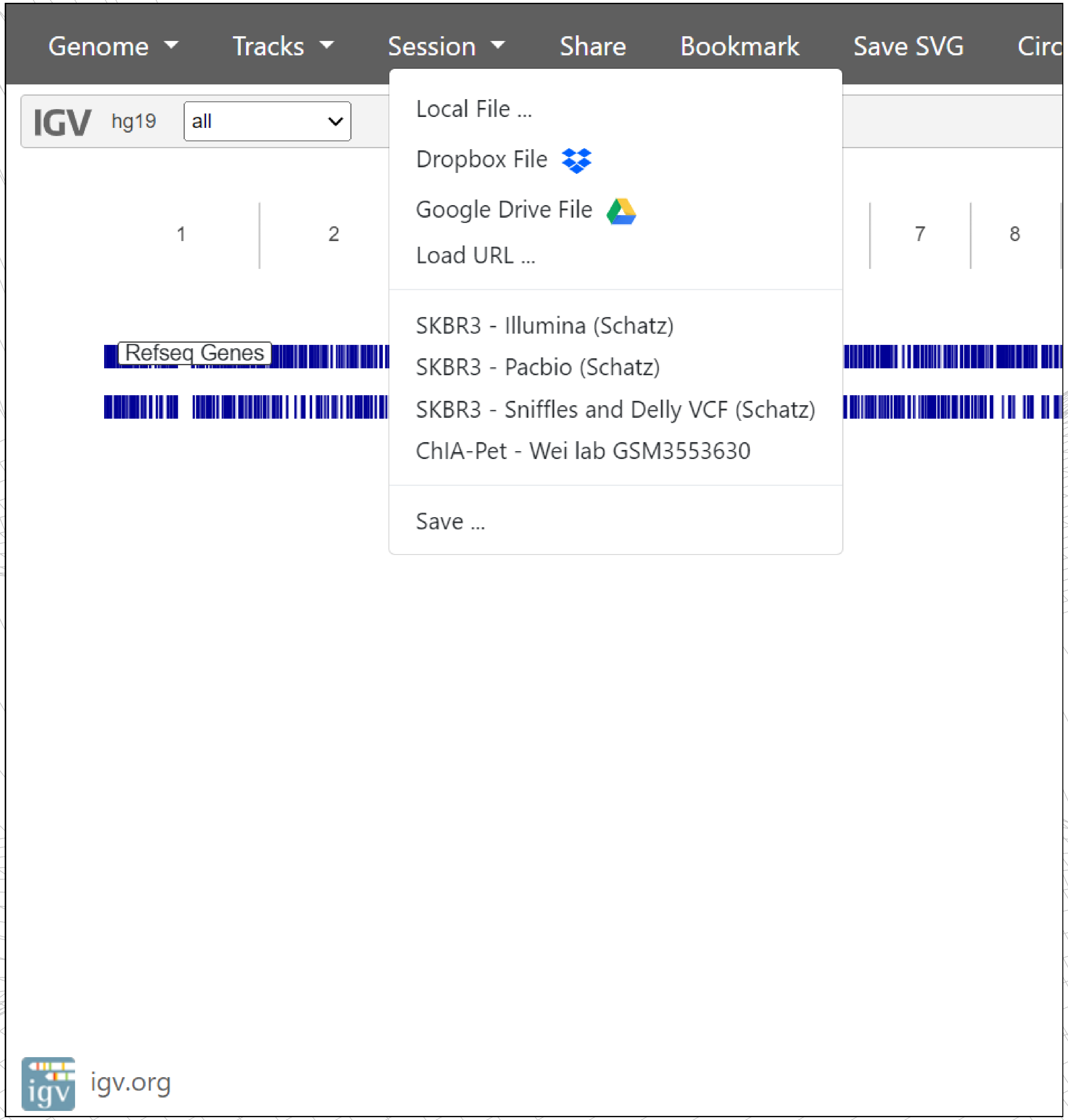
Genomic Coordinates


Genome
Tracks
Session
Share
Bookmark
Save SVG
Circular View
Help
Local File ...
Dropbox ...
Google Drive ...
URL ...
Human (T2T CHM13-v2.0/hs1)
Human (T2T CHM13-v1.1)
Human (GRCh38/hg38)
Human (hg38 1kg/GATK)
Human (GRCh37/hg19)
Human (hg18)
Mouse (GRCm39/mm39)
Mouse (GRCm38/mm10)
Mouse (NCBI37/mm9)
Rat (rn7)
Rat (RGCS 6.0/rn6)
Gorilla (Kamilah_GGO_v0/gorGor6)
Gorilla (gorGor4.1/gorGor4)
Chimp (panTro6) (panTro6)
Chimp (panTro5) (panTro5)
7
8
9
10
11
12
igv.org

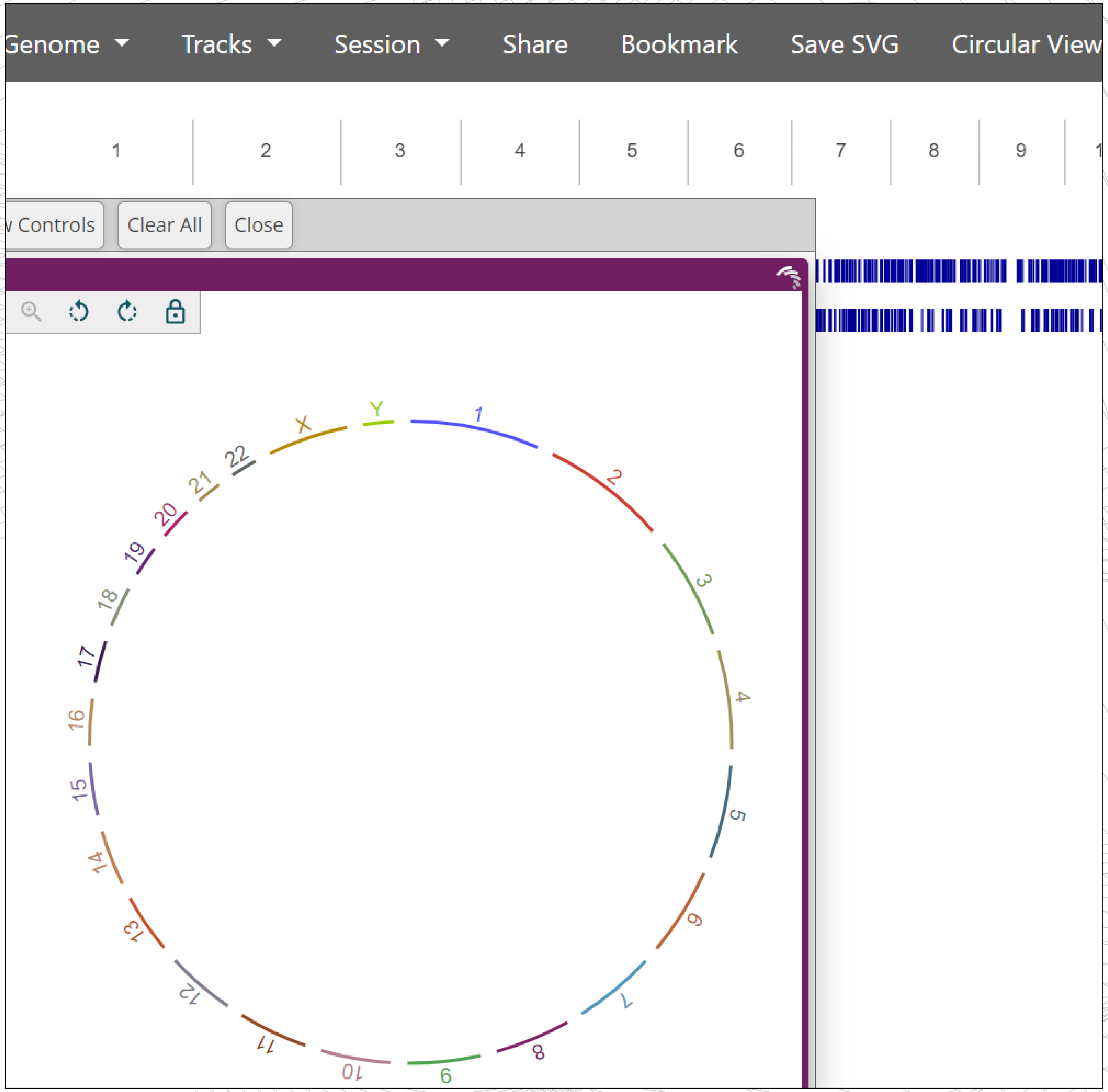
Genomes Feature


Genome
Tracks
Session
Share
Bookmark
Save SV
IGV
hg19
all
Local File ...
Dropbox ...
Google Drive ...
URL ...
Annotations ...
Variants ...
ENCODE Signals - ChIP ...
ENCODE Signals - Other ...
ENCODE Other ...
Platinum Genomes ...
Tutorials ...
1
4
5
6
7
Refseq Ger
igv
igv.org

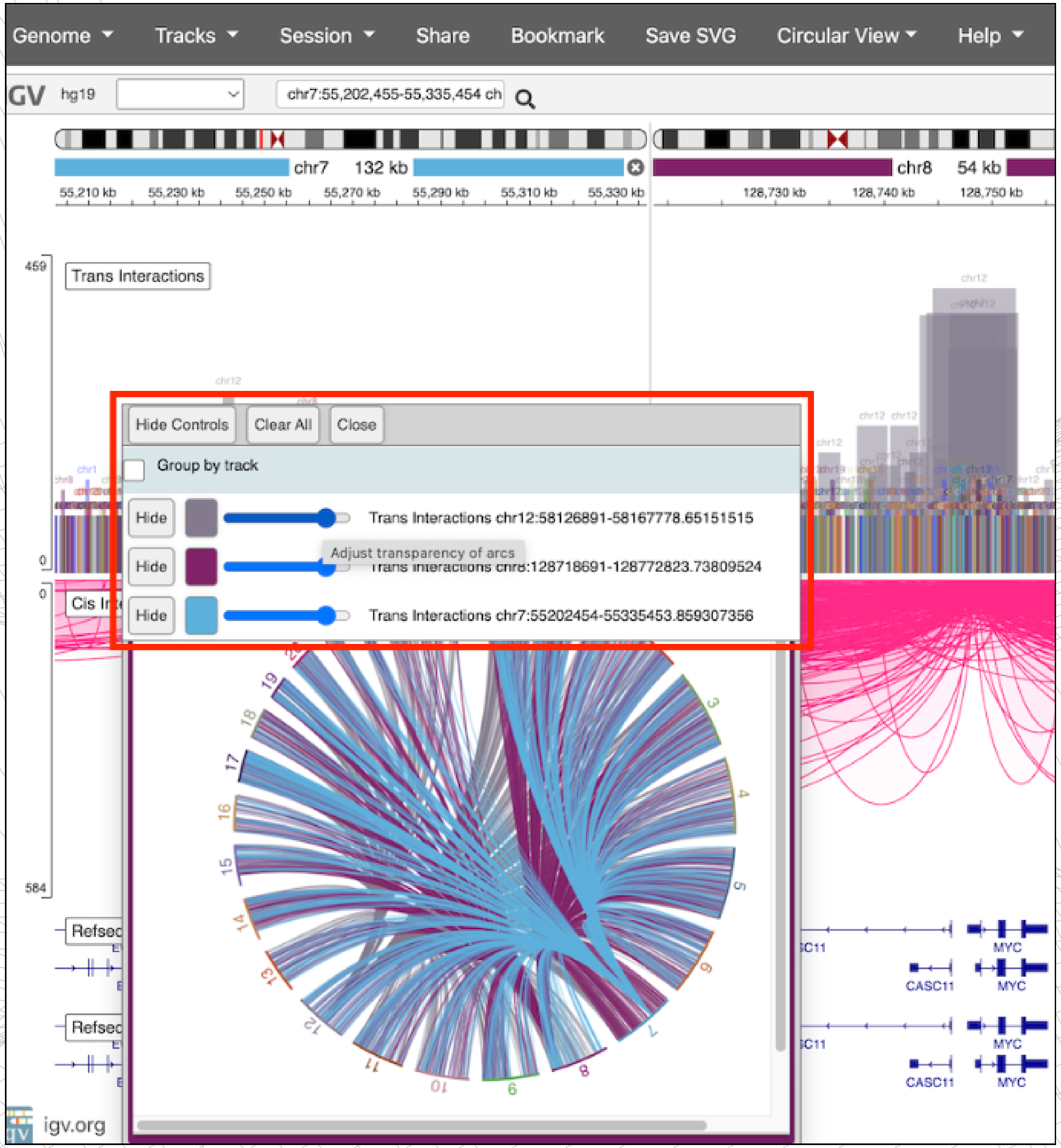
Tracks Feature


Genome
Tracks
Session
Share
Bookmark
Save SVG
Circ
IGV
hg19
all
Local File ...
Dropbox File
Google Drive File
Load URL ...
SKBR3 - Illumina (Schatz)
SKBR3 - Pacbio (Schatz)
SKBR3 - Sniffles and Delly VCF (Schatz)
ChIA-Pet - Wei lab GSM3553630
Save ...
1
2
7
8
Refseq Genes
igv
igv.org

Session Feature


Genome
Tracks
Session
Share
Bookmark
Save SVG
Circular View
1
2
3
4
5
6
7
8
9
Controls
Clear All
Close
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

Circular View


Genome
Tracks
Session
Share
Bookmark
Save SVG
Circular View
Help
hg19
chr7:55,202,455-55,335,454
chr7
132 kb
chr8
54 kb
Trans Interactions
Cis In
Refseq
Hide Controls
Clear All
Close
MYC
CASC11
igv.org

☐ Group by track

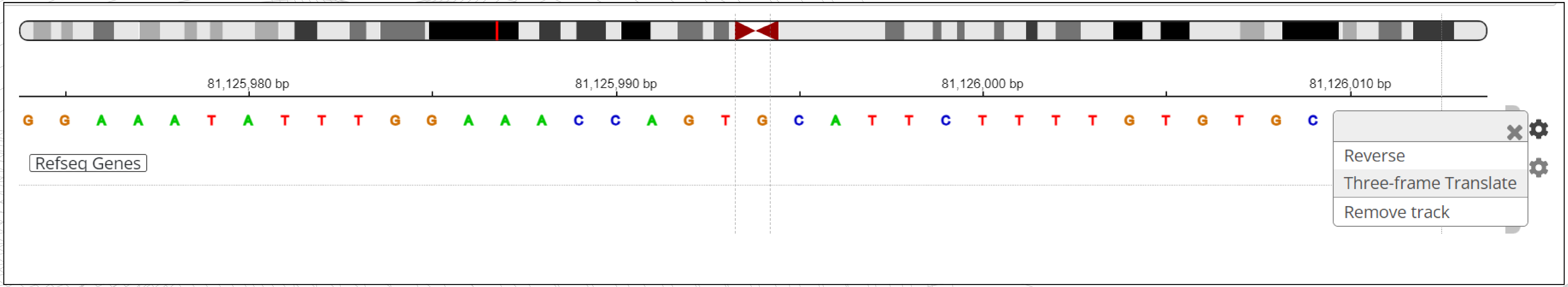
Hide ▨ Trans Interactions chr12:58126891-58167778.65151515

Adjust transparency of arcs

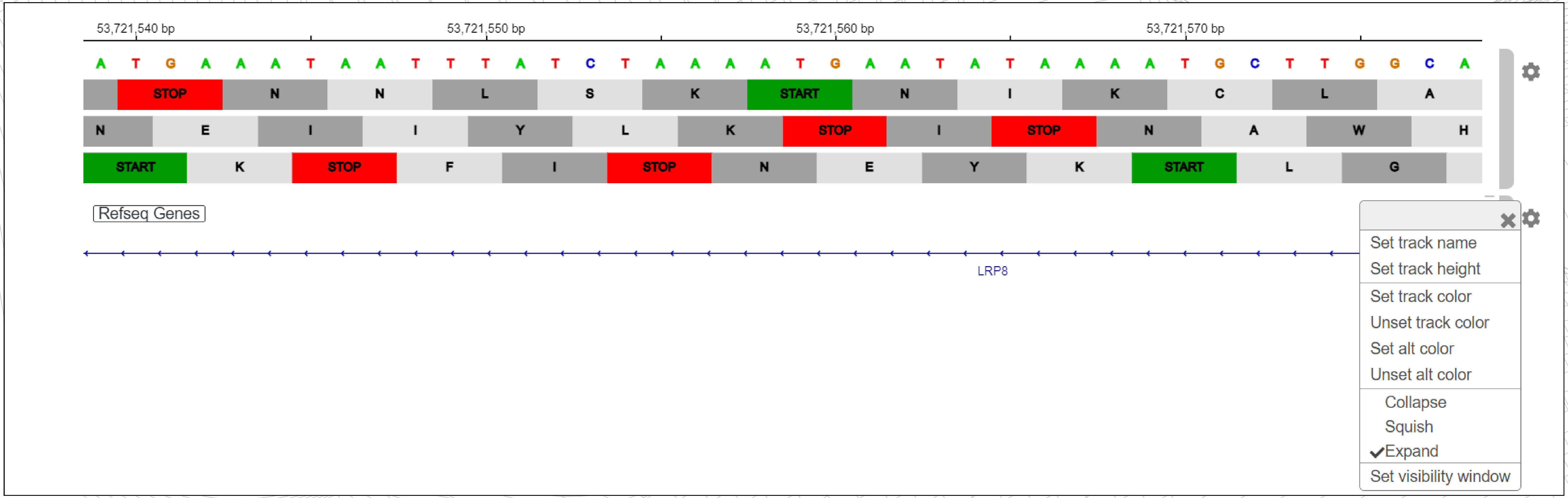
Hide ▨ Trans Interactions chr8:128718691-128772823.73809524

Hide ▨ Trans Interactions chr7:55202454-55335453.859307356

Trans-Interaction of 3 chromosomes


81,125,980 bp
81,125,990 bp
81,126,000 bp
81,126,010 bp
G G A A A T A T T T G G A A A C C A G T G C A T T C T T T T G T G T G C
Refseq Genes
Reverse
Three-frame Translate
Remove track

Translation Options


53,721,540 bp
53,721,550 bp
53,721,560 bp
53,721,570 bp
A T G A A A T A A T T T A T C T A A A A T G A A T A T A A A A T G C T T G G C A
STOP N N L S K START N I K C L A
N E I I Y L K STOP I STOP N A W H
START K STOP F I STOP N E Y K START L G
Refseq Genes
LRP8
Set track name
Set track height
Set track color
Unset track color
Set alt color
Unset alt color
Collapse
Squish
✓Expand
Set visibility window

Translation Settings

ADVANTAGES	DISADVANTAGES
1.Versatile Data Visualization	1. Complexity of Interpretation
2. Interactive Exploration	2. Resource Intensive
3. Customization Options	3. Potential for Misinterpretation

APPLICATIONS

1.Visualizing Sequence Alignments

2. Exploring Gene Expressions

3. Studying Chromatin Structure

4. Comparative Genomics

5. Investigating Epigenetic Modifications

REFERENCES:

NGS Pipeline:

1. McFadyen, R. (2020, August 3). Intro to NGS Data Analysis Workflow. Diagnostech. https://diagnostech.co.za/intro-to-ngs-data-analysis-workflow/

BWA:

1. Bwa.1. (n.d.). https://bio-bwa.sourceforge.net/bwa.shtml
2. Difference between BWA-backtrack and BWA-MEM. (n.d.). Bioinformatics Stack Exchange. https://bioinformatics.stackexchange.com/questions/15/difference-between-bwa-backtrack-and-bwa-mem
3. Dunning, M. (n.d.). Understanding alignments. https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2017/Day1/Session5-alignedReads.html
4. Galaxy. (n.d.). https://usegalaxy.eu/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fdevteam%2Fbwa%2Fbwa_mem%2F0.7.17.2&version=latest
5. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

REFERENCES:

Output file formats:

1. Bioinfo Byte. (2023, February 11). Genomic file formats: BAM / SAM file format [Video]. YouTube. https://www.youtube.com/watch?v=1-hwgtKX8qU
2. Bioinformagician. (2022, May 13). Understanding Bioinformatics File Formats: SAM/BAM [Video]. YouTube. https://www.youtube.com/watch?v=vh6l9aPFUgU
3. GA4GH. (2019, March 25). CRAM: The Data Storage Standard of Genomics [Video]. YouTube. https://www.youtube.com/watch?v=mMr26JG6SCI
4. Katharine ME. (2022, June 7). FASTQ, BAM, and VCF file formats easily explained - A must watch if you have had a DNA test [Video]. YouTube. https://www.youtube.com/watch?v=y4KqVfCdLo

IGV:

1. Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017
2. Nusrat, S., Harbig, T., & Gehlenborg, N. (2019). Tasks, techniques, and tools for genomic data visualization. Computer Graphics Forum (Print), 38(3), 781–805. https://doi.org/10.1111/cgf.13727
3. About IGV-Web. (n.d.). IGV-Web Documentation. https://igvteam.github.io/igv-webapp/

REFERENCES:

BOWTIE:

Bowtie 2: Manual. (n.d.). https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

SAMTOOLS:

samtools(1) manual page. (n.d.). http://www.htslib.org/doc/samtools.html

THANK YOU

FROM SHOMU’S BIOLOGY 2.0