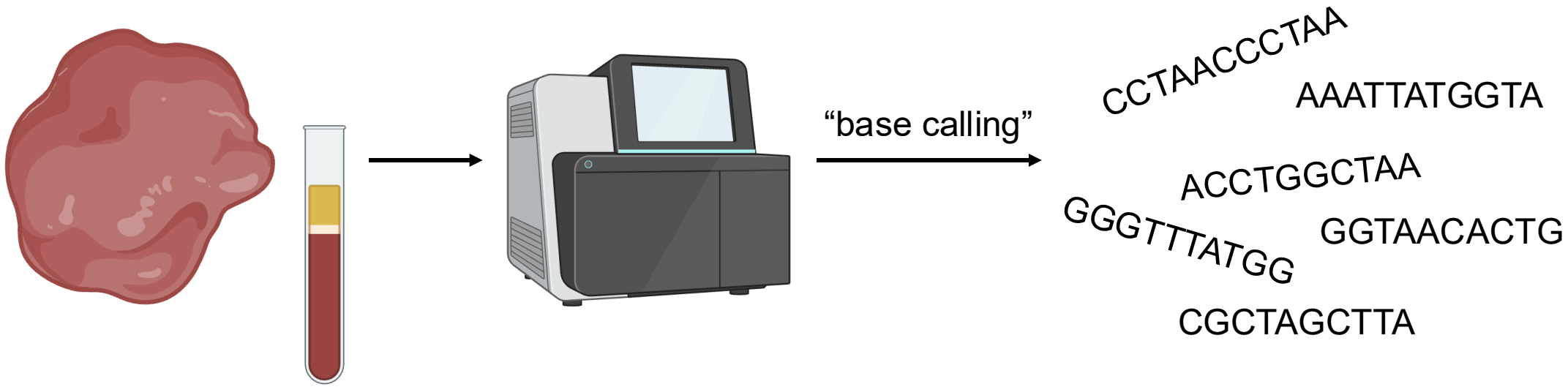
What are the inputs to a bioinformatics analysis?

Clinical Cancer Genomics
12 May 2025

Learning objectives and lecture agenda

- Prepare for the analysis of DNA-sequencing data during lectures and exercises on Wednesday
- Learn about the components of a bioinformatics analysis other than the code and software – **what are the inputs?**
  - Become familiar with some of the main file types
- Learn the basics of the Integrative Genomics Viewer (IGV)



Overview of Illumina sequencing: https://www.youtube.com/watch?v=fCd6B5HRaZ8

Sequencing instruments like the Illumina sequencer produce millions of reads that are stored in **FASTQ** files: 1 file for single-end, 2 files for paired-end sequencing.

Example

Paired-end sequencing of a **DNA fragment**.

In paired-end sequencing, a DNA fragment is sequenced from its both ends resulting in two sequencing reads: *read1* and *read2*. These are stored in `2` FASTQ files.

Note that the DNA fragment to be sequenced is also referred to as an *insert*.

Remember that an Illumina sequencer sequences each of the strands separately and one after another. The illustration above is only meant to show from where each of the reads is originating.

This is *read1* obtained from the DNA fragment above (stored in the 1st FASTQ file).

```
@HWI-ST898:563:C4CK5ACXX:5:1212:14521:85648/1
AGGTTAATGCTTTCTCTCTA.....
+
=?>=@AAADCC@@BBBBBB@.....
```


$$Q = -10*\log_{10}(P_{err})$$

This is *read2* obtained from the same DNA fragment (stored in the 2nd FASTQ file).

```
@HWI-ST898:563:C4CK5ACXX:5:1212:14521:85648/2
TTTCTCCTAAGATCTCAGTG.....
+
>><@BBBC@@C@@BBBBC?C.....
```

$$Q = -10 * \log_{10}(P_{err})$$

Probability of a base being wrong = 0.001 (P_{err})

$-10 * \log_{10}(0.001) = 30$

Symbol	Phred Quality Score	Probability of Incorrect Base Call
<	27	0.002
=	28	0.001
>	29	0.001
?	30	0.001
@	31	0.0008
A	32	0.0006

Quality control

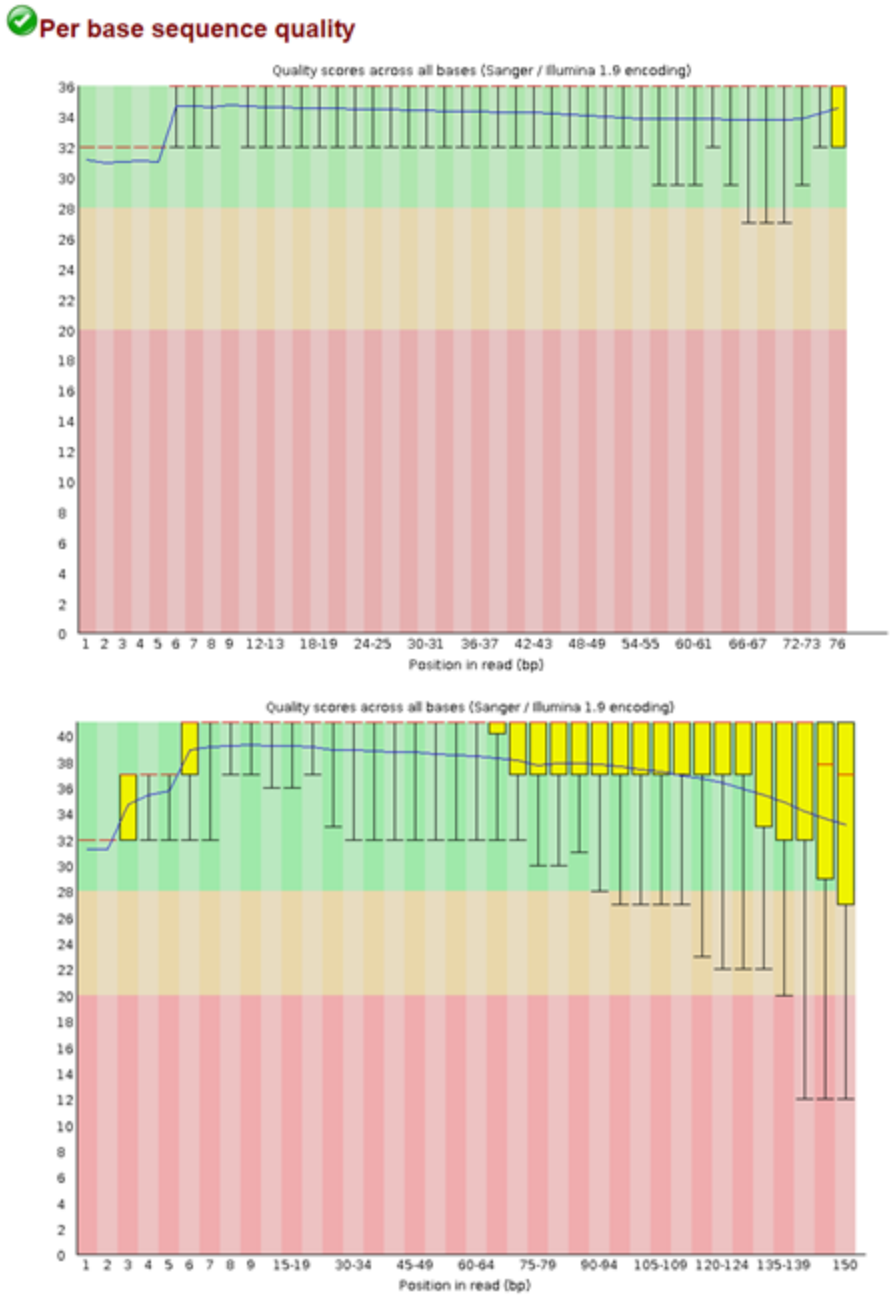
A typical step after obtaining FASTQ files is to perform **quality control** using tools such as FastQC.

This analysis tells us if we need to take further actions before proceeding with downstream analyses.

The figure at the bottom right shows closer to the end of reads, quality score drops.

Trimming low-scoring end bases can enhance downstream analysis outcomes:

- e.g. using bioinformatics tools **Trimmomatic, Cutadapt, TrimGalore**.



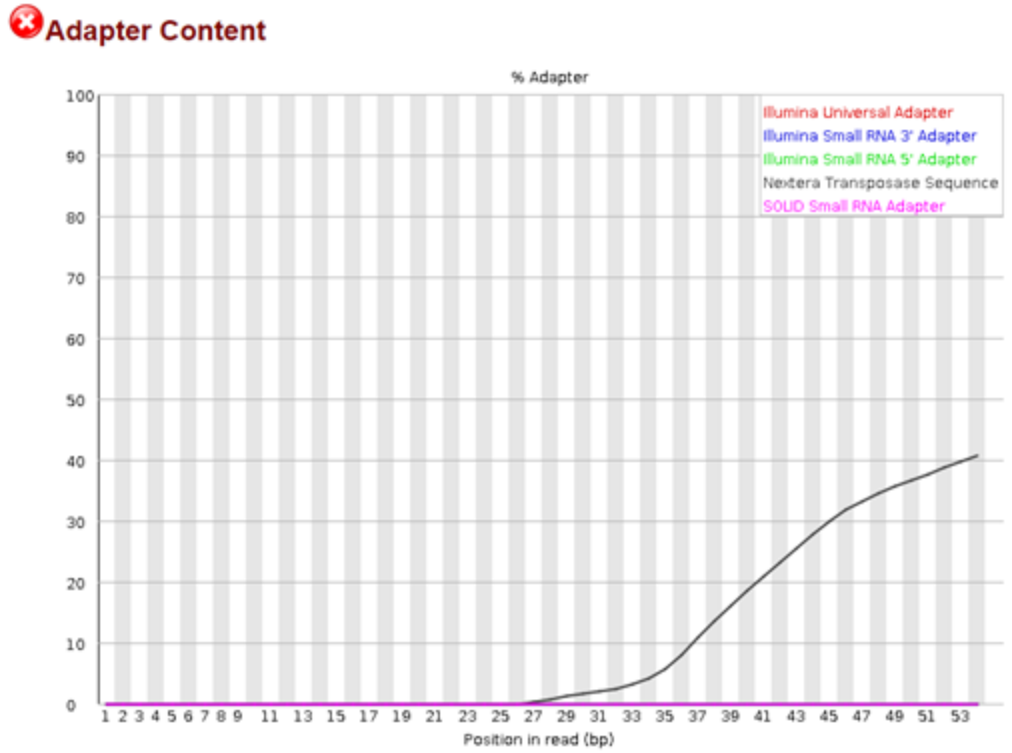
The top right figure shows adapter contamination closer to the end of reads.

This is caused by adapter read-through.

- In paired-end sequencing, this may occur when the DNA fragment length is shorter than the read length.

Trimming the adapter sequences enhances downstream analysis outcomes:

- e.g. using bioinformatics tools **Trimmomatic, Cutadapt, TrimGalore**.

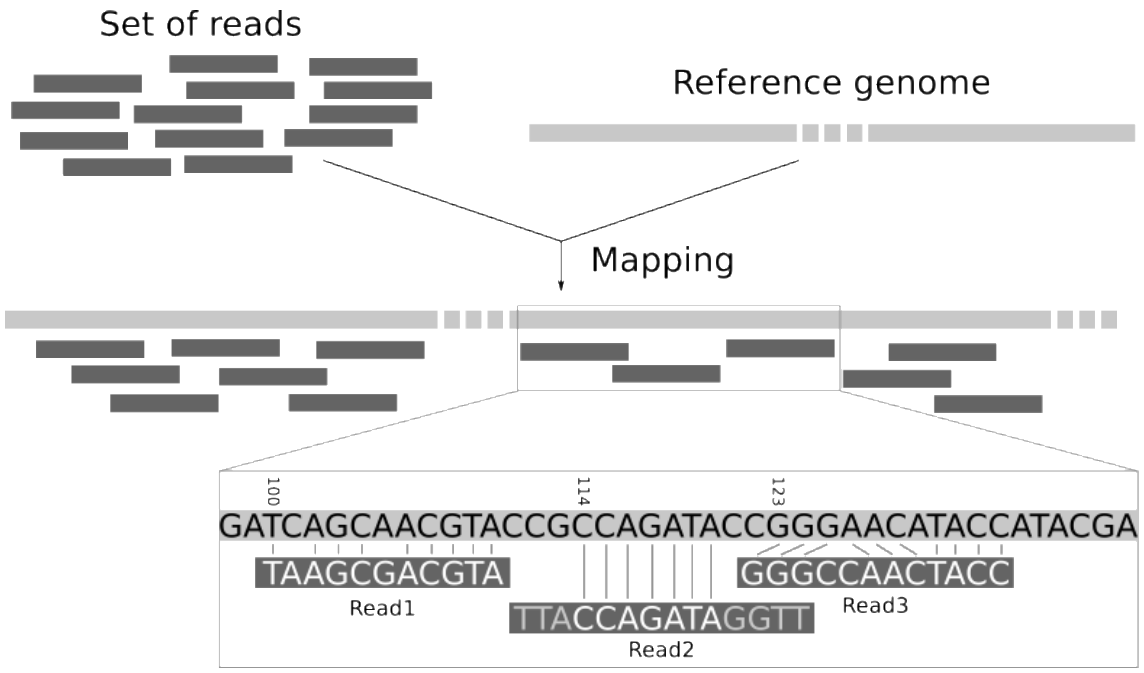


Example

Adapter read-through

```
   ADAPTER----------------------read1------------------------->
                                                        ADAPTER
                                                        ⇩⇩⇩⇩⇩
5′ NNNNNNNAGGTTAATGCTTTCTCT...........CACTGAGATCTTAGGAGAAANNNNNNN 3′
       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3′ NNNNNNNTCCAATTACGAAAGAGA...........GTGACTCTAGAATCCTCTTTNNNNNNN 5′
     <-------------------------read2--------------------
     ⇧⇧⇧⇧⇧
```

Sequence alignment



To sequence DNA, we have to fragment it into smaller pieces and therefore lose the location information (i.e. where in the genome they belong).

Sequence alignment or mapping is the process of finding the genomic locations of the sequencing reads obtained from the sequencing of DNA fragments.

Example

Sequence alignment.
We have a reference genome, and a set of sequencing reads. We want to align these reads to the genomic region they have originated from.

```
CTAGAGCGTGGCCCGGAGCTGCCCTTTCCTCTTCGGTGAAGTTTTTAAAAGCTGCTGCGA # reference

CCGGAGCTGCCCTTTCCTCTTCGGTGA
CCTTACCTCTTCGGTGAAGTTTTTAAA
GAGCTGCCCTTTCCTCTTCGGTGAAGT
TGCCCTTACCTCTTCGGTGAAGTTTTT
CCCGGAGCTGCCCTTTCCTCTTCGGTG
TACCTCTTCGGTGAAGTTTTTAAAAGC
```

After alignment.

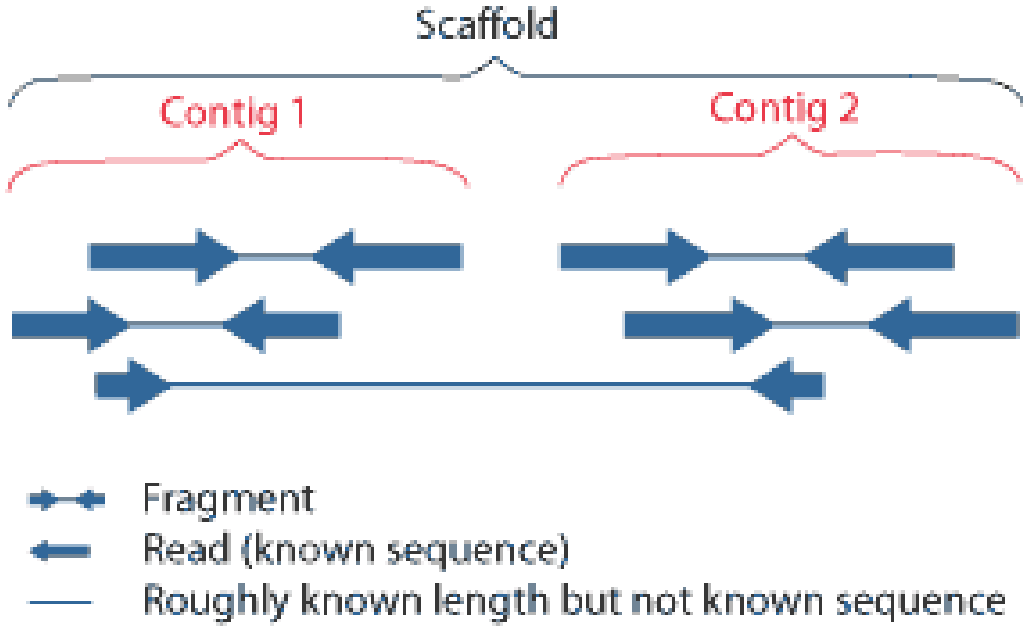
```
                                ⇓
00000000011111111112222222222333333333344444444445555555556 # position-10s
12345678901234567890123456789012345678901234567890123456789 # position-1s
CTAGAGCGTGGCCCGGAGCTGCCCTTTCCTCTTCGGTGAAGTTTTTAAAAGCTGCTGCGA # reference
            CCGGAGCTGCCCTTTCCTCTTCGGTGA
                      CCTTACCTCTTCGGTGAAGTTTTTAAA
                GAGCTGCCCTTTCCTCTTCGGTGAAGT
                    TGCCCTTACCTCTTCGGTGAAGTTTTT
           CCCGGAGCTGCCCTTTCCTCTTCGGTG
                         TACCTCTTCGGTGAAGTTTTTAAAAGC
                                ⇑
```

🧩 In a real situation, the **reference is billions of nucleotides long** and we have **hundreds of millions of sequencing reads** each are **e.g. 100 nucleotides long**.

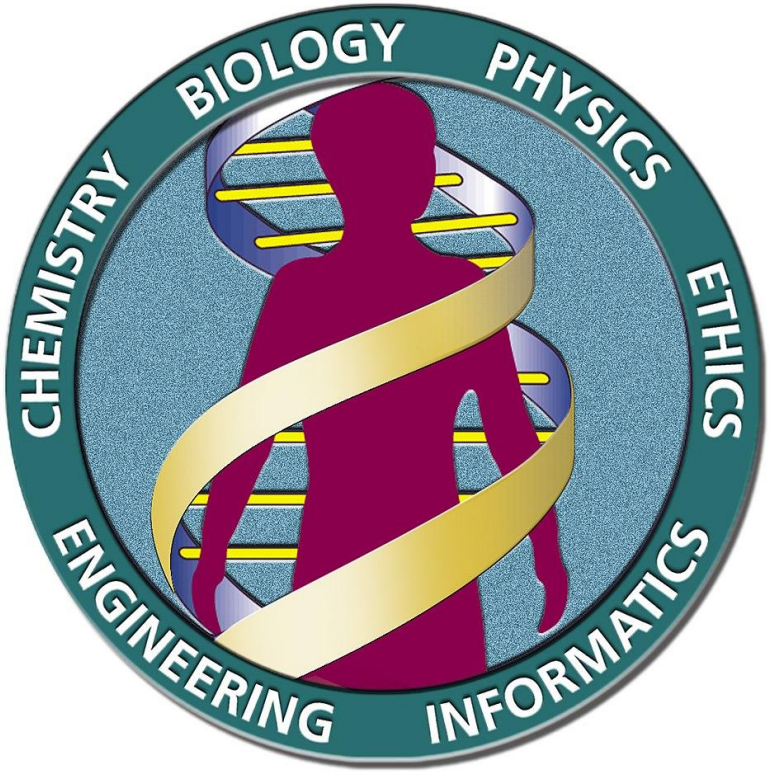
❓ Incidentally, what is happening at the position `27` marked by the arrow?

We need a reference human genome

- An attempt at a complete representation of the nucleotide sequence of an individual genome.
- When we sequence new samples, we can map the reads to the reference rather than doing a new assembly each time.



We need a reference human genome



Human Genome Project

1990 - 2003

- The Human Genome Project genome sequence accounted for over 90% of the human genome in 2003
  - … but it contained some **gaps** and **errors**

- Why are there gaps?
  - **Difficult-to-sequence regions,** e.g. repetitive regions
  - **Sequencing errors** and **low sequence read coverage** at certain loci make it difficult to reliably reconstruct the genome, thus resulting in gaps

- Efforts by the **Genome Reference Consortium (GRC)** and recently by the **Telomere to Telomere (T2T)** consortium have sought to fill in the remaining gaps.

- There are further efforts by the **Human Pangenome Reference Consortium (HPRC)** and others to *"better represent the human diversity"* in the human genome sequence.

Major releases:

2006 → hg18 / NCBI36

2009 → hg19 / NCBI37 / GRCh37

2013 → hg38 / GRCh38

2022 → T2T-CHM13

Minor releases or patches:

GRCh38.p13

Need to convert genome coordinates from one assembly to another (e.g. from hg19 to hg38)? → LiftOver

The FASTA file

```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTA
GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGG
```

Aligned sequences

The alignment results to a reference genome are stored in a standard file format called **Sequence Alignment/Map (SAM).**

- This file is text-based, human readable, and takes up a lot of space.
- The **binary alignment map (BAM)** is a binary file that stores similar information in a compressed form and thus uses less space, but it is not human-readable.

```
13894   HS2000-905_68:3:1307:14091:6825 137 chr2    92045101    254 28M1D72M    *   0   0
        ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
        CCCFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDBB
        BC:Z:0  XD:Z:11T16^A$5A1C45A18  SM:i:328    AS:i:0
13895   HS2000-905_68:1:1305:12812:167908   147 chr2    92045105    254 100M    =   92044908    -297
        TCAAAGAGTGGGACCCCTGAACCTGACCCTGACCCCTGACCCTGATCCCTAACCTCTGACCCTGACCCCTAACCCCTGACCCTAACCCTAACCCCTAACC
        CDDDCCDDDBDBBDDDDCCCCDDDCCDDDDB?DEEEEC@FFFFHGHGIGDC=IIIJIHGJJJHEDJJJIGF?IJJIIIHJJIGFCJJHHHFHFFFDD=@B
        AM:i:0  BC:Z:0  XD:Z:A3CT1TCA1AGTGGGAACC1TGAC4A14C8C12A13A18    SM:i:0  AS:i:370
13896   HS2000-905_68:2:2107:9712:70649 163 chr2    92045106    254 100M    =   92045307    301
        CAACTATCAGAGGGGGAACCCTGACCCCTAACCCCTGACCCTGACCCCTAACCCCTGACCCTGAGCACTAACCCCTGACCATAACCCTAACCTCCAACCC
        ?8?1BBDB>DDFAG61EBCDB)?;?B):@FAB886(<3=)=8=C>@(-;57(.6=??3(;;(,=(555@5::9A8?8A######################
        BC:Z:0  XD:Z:12T51C27C1T5   SM:i:346    AS:i:797
```

```
13894   HS2000-905_68:3:1307:14091:6825 137 chr2    92045101    254 28M1D72M    *   0   0
        ATAGACAACTAACAGAGTGGGAACCCTGCCCCTGAACCCTGACCCTGACCCCTAACCCCTGACCCTGACCACTAACCCCTGGCCATAACCCTAACCCCTA
        CCCFFFFFHHHHHJJJJFHIGIJJJJIJJJJJJJJJJJJJIIJJJJIIJJHIIJJIIJJHHHHHFFFFCECDECDDDDDDDDDDDDADDDDDDDDDDBB
        BC:Z:0  XD:Z:11T16^A$5A1C45A18  SM:i:328    AS:i:0
13895   HS2000-905_68:1:1305:12812:167908   147 chr2    92045105    254 100M    =   92044908    -297
        TCAAAGAGTGGGACCCCTGAACCTGACCCTGACCCCTGACCCTGATCCCTAACCTCTGACCCTGACCCCTAACCCCTGACCCTAACCCTAACCCCTAACC
        CDDDCCDDDBDBBDDDDCCCCDDDCCDDDDB?DEEEEC@FFFFHGHGIGDC=IIIJIHGJJJHEDJJJIGF?IJJIIIHJJIGFCJJHHHFHFFFDD=@B
        AM:i:0  BC:Z:0  XD:Z:A3CT1TCA1AGTGGGAACC1TGAC4A14C8C12A13A18    SM:i:0  AS:i:370
13896   HS2000-905_68:2:2107:9712:70649 163 chr2    92045106    254 100M    =   92045307    301
        CAACTATCAGAGGGGGAACCCTGACCCCTAACCCCTGACCCTGACCCCTAACCCCTGACCCTGAGCACTAACCCCTGACCATAACCCTAACCTCCAACCC
        ?8?1BBDB>DDFAG61EBCDB)?;?B):@FAB886(<3=)=8=C>@(-;57(.6=??3(;;(,=(555@5::9A8?8A######################
        BC:Z:0  XD:Z:12T51C27C1T5   SM:i:346    AS:i:797
```

Common genomic data file formats

- **FASTQ** – raw sequence data (with qualities)
- **FASTA** – sequences (DNA, RNA, protein)
- **SAM/BAM** – aligned sequence data
- **BED** – other genome features
- **GTF** – gene/transcript annotations
- **VCF** – variant calls (individual, multi-individual)
- **MAF** – aggregated variant information (project, population)

- Plus, many other (potentially custom) data formats output by specialized tools

Post-alignment, sequencing reads can be summarized over/within genomic intervals.

Browser Extensible Data (BED) file format - Genomic intervals and information.

```
1   chr7    127471196  127472363
2   chr7    127472363  127473530
3   chr7    127473530  127474697
4   chr7    127474697  127475864
5   chr7    127475864  127477031
6   chr7    127477031  127478198
7   chr7    127478198  127479365
8   chr7    127479365  127480532
9   chr7    127480532  127481699
```

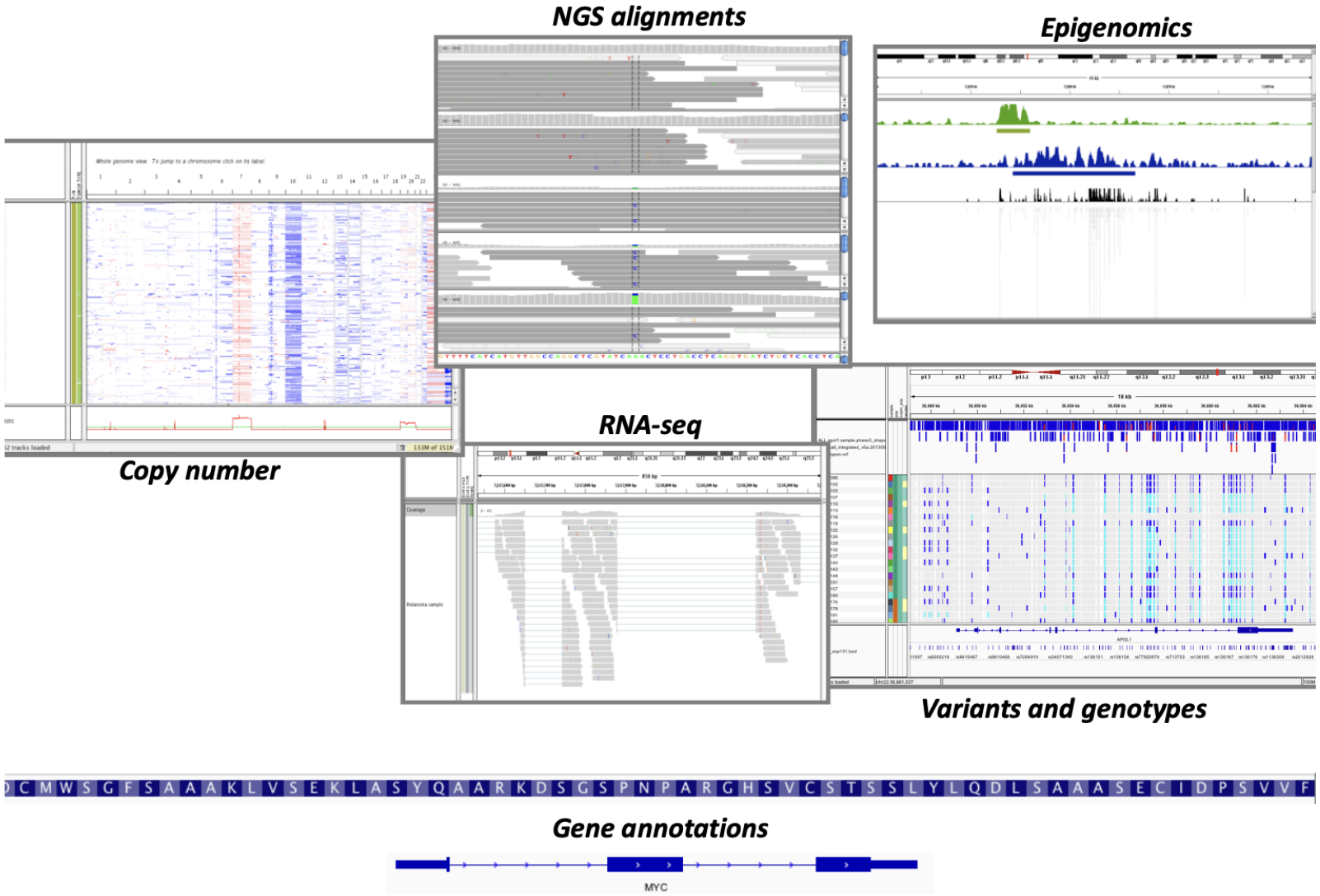
We may also need some annotations of our genome
… for example, where are certain genomic features located?

Gene Transfer Format (GTF) – Information on gene structure

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG00000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG00000169..

Integrative Genomics Viewer (IGV)

A visualization tool to simultaneously integrate and analyze multiple types of genomic data.



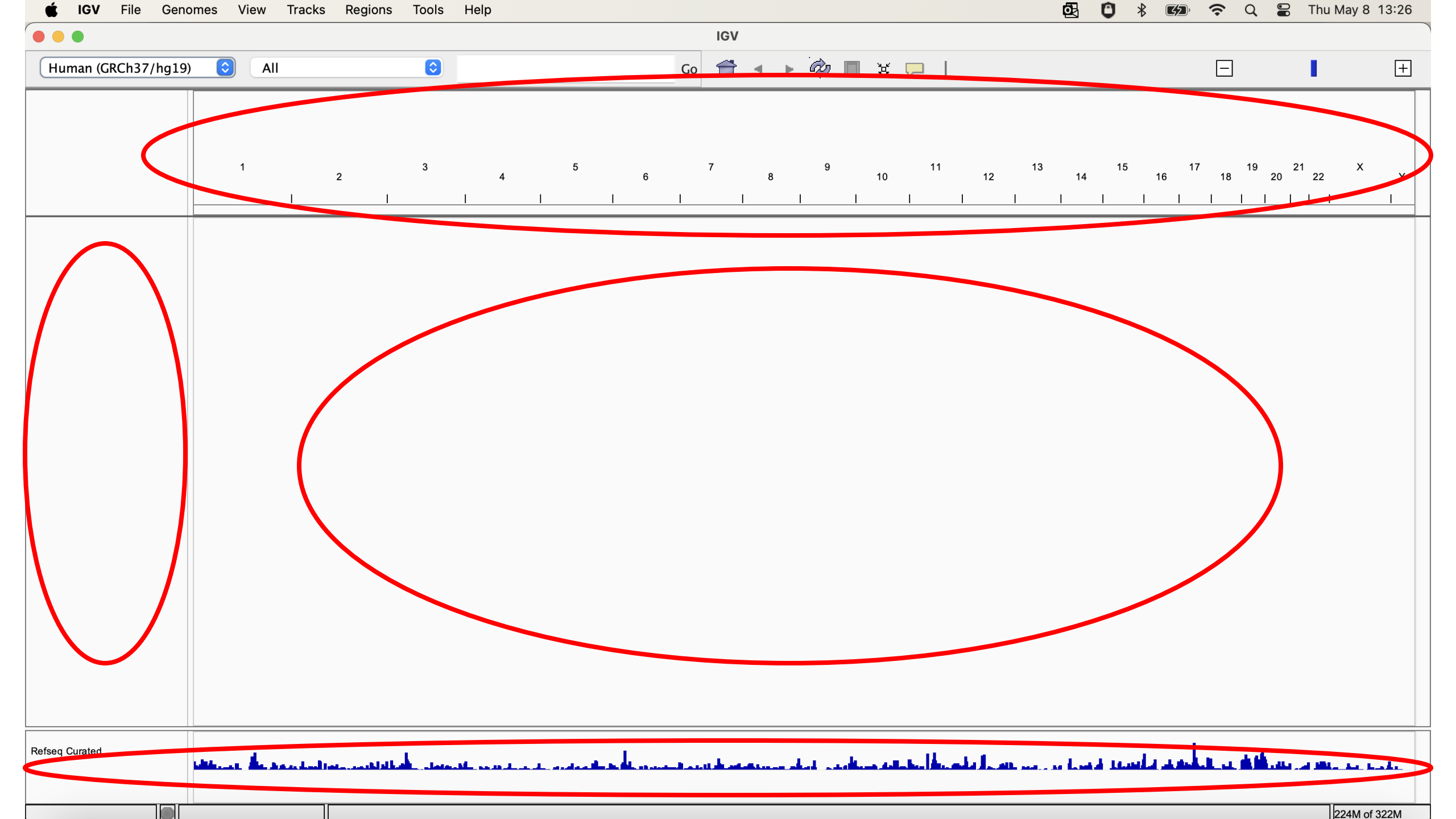
Slide source: Broad Institute

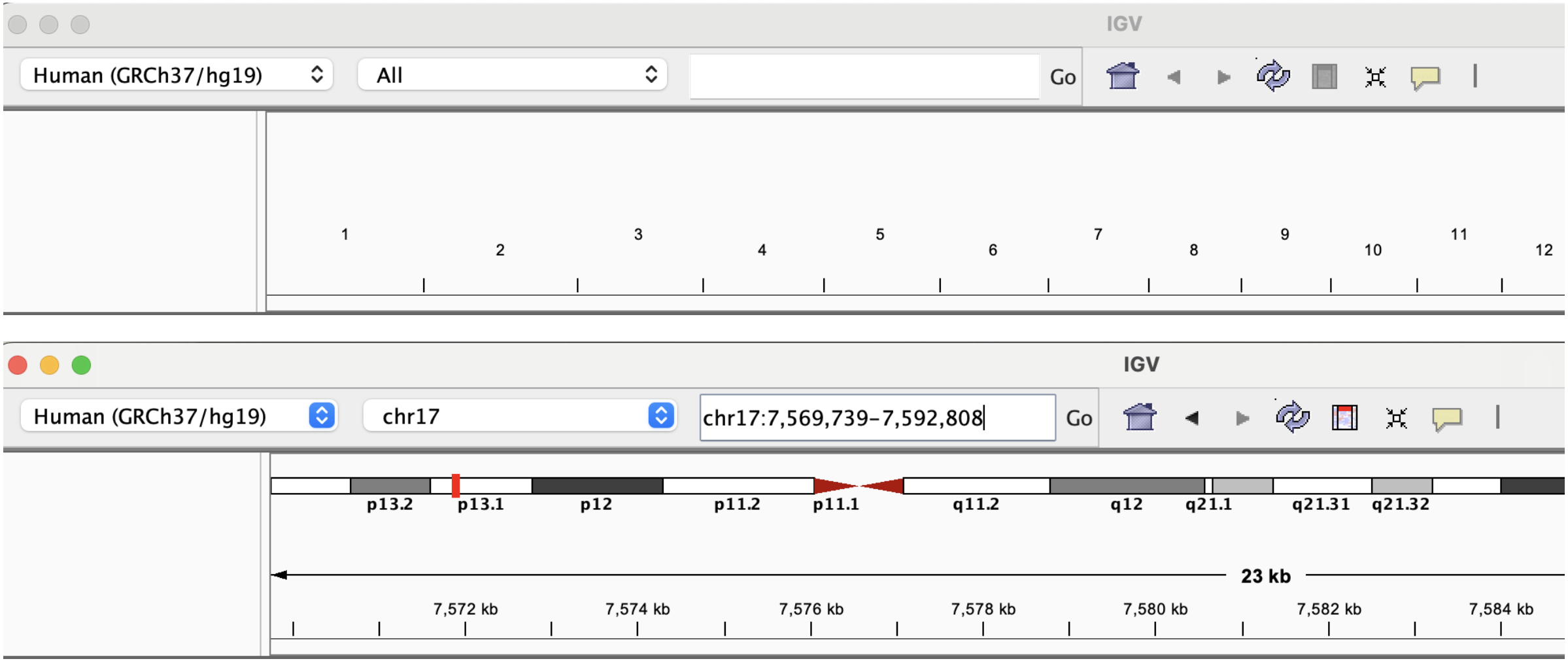
IGV supports many different file formats

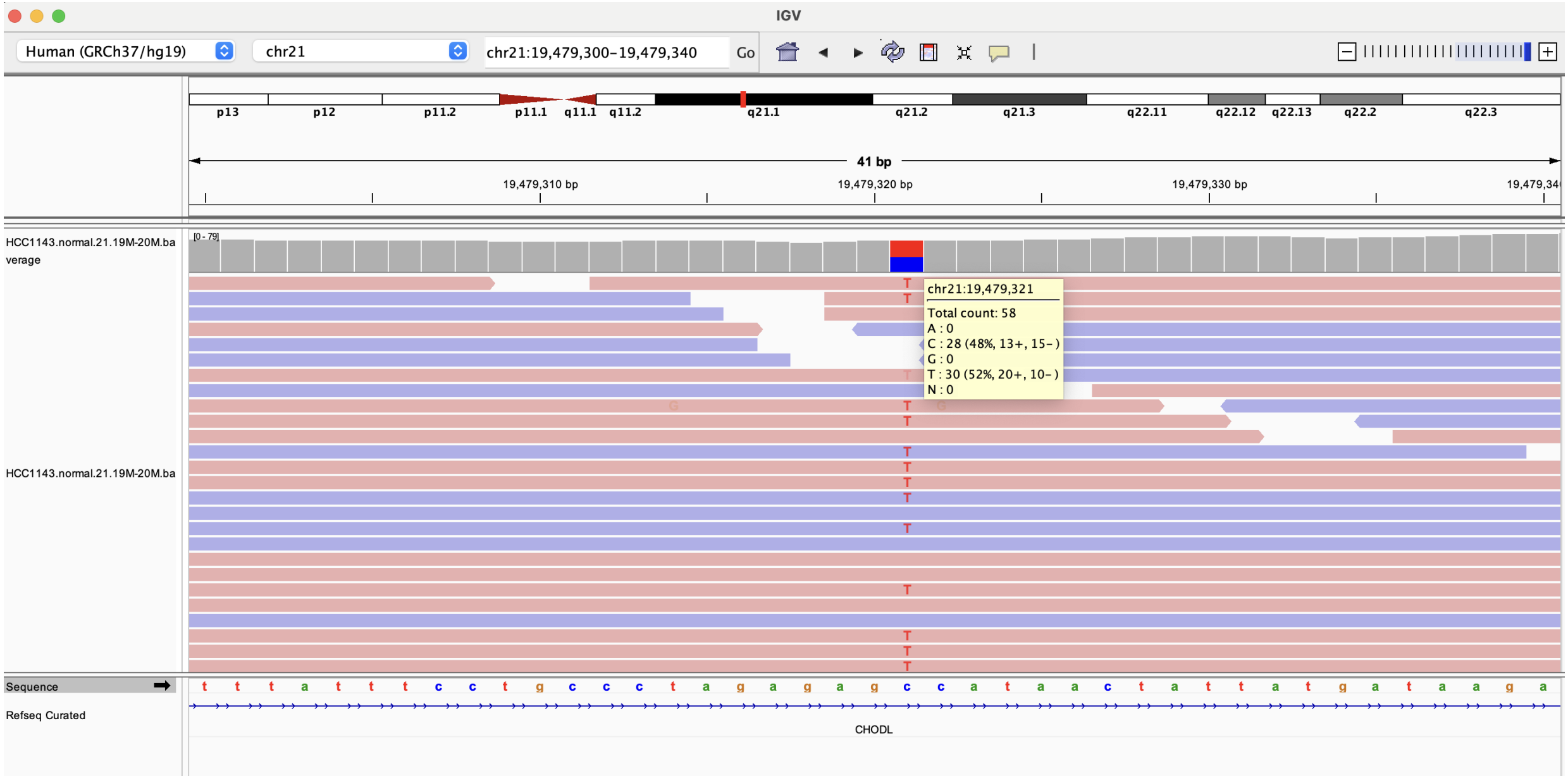
- BAM
- BED
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- broadPeak
- CBS
- CN
- Custom File Formats
- Cytoband
- FASTA
- GCT
- genePred
- GFF/GTF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF (Multiple Alignment Format)
- MAF (Mutation Annotation Format)
- Merged BAM File
- MUT
- narrowPeak
- PSL
- SAM
- Sample Info (Attributes) file
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- VCF
- WIG
- chrom.sizes

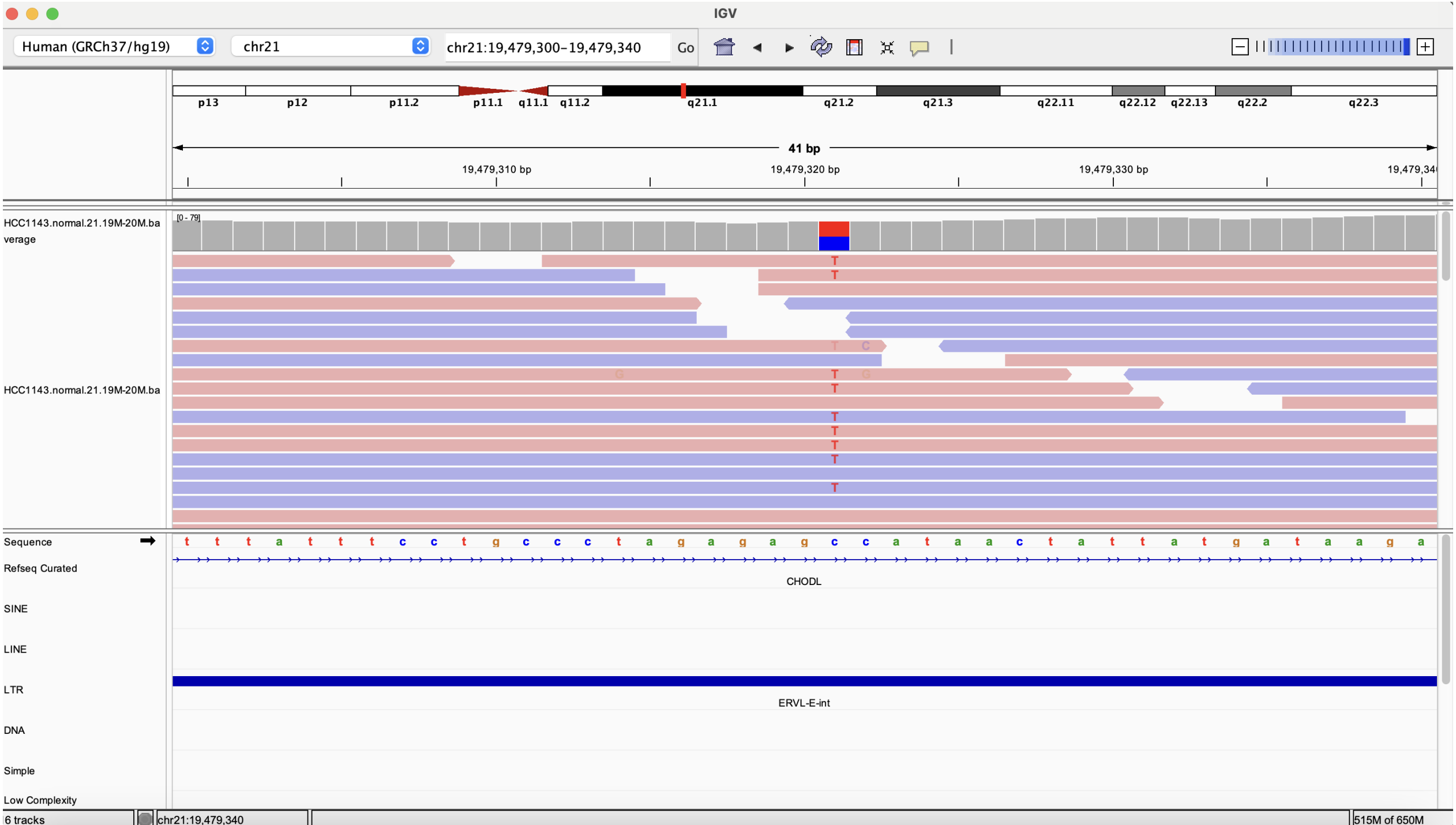
For more info see: *www.broadinstitute.org/igv/FileFormats*

Slide source: Broad Institute


IGV File Genomes View Tracks Regions Tools Help
Thu May 8 13:26
IGV
Human (GRCh37/hg19)
All
Go
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Refseq Curated
224M of 322M


IGV
Human (GRCh37/hg19)
All
Go
1
2
3
4
5
6
7
8
9
10
11
12
IGV
Human (GRCh37/hg19)
chr17
chr17:7,569,739-7,592,808
Go
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
23 kb
7,572 kb
7,574 kb
7,576 kb
7,578 kb
7,580 kb
7,582 kb
7,584 kb


IGV
Human (GRCh37/hg19)
chr21
chr21:19,479,300-19,479,340
Go
p13
p12
p11.2
p11.1
q11.1
q11.2
q21.1
q21.2
q21.3
q22.11
q22.12
q22.13
q22.2
q22.3
41 bp
19,479,310 bp
19,479,320 bp
19,479,330 bp
HCC1143.normal.21.19M-20M.ba
verage
[0 - 79]
chr21:19,479,321
Total count: 58
A : 0
C : 28 (48%, 13+, 15-)
G : 0
T : 30 (52%, 20+, 10-)
N : 0
HCC1143.normal.21.19M-20M.ba
Sequence
Refseq Curated
CHODL


IGV
Human (GRCh37/hg19)
chr21
chr21:19,479,300-19,479,340
Go
p13
p12
p11.2
p11.1
q11.1
q11.2
q21.1
q21.2
q21.3
q22.11
q22.12
q22.13
q22.2
q22.3
41 bp
19,479,310 bp
19,479,320 bp
19,479,330 bp
HCC1143.normal.21.19M-20M.bam Coverage
[0 - 79]
HCC1143.normal.21.19M-20M.bam
Sequence
Refseq Curated
CHODL
SINE
LINE
LTR
ERVL-E-int
DNA
Simple
Low Complexity
6 tracks
chr21:19,479,340
515M of 650M

… now all we need is some tools and software!

- In **Exercise Set 1**, we will be inspecting some of the files and file types discussed in this lecture
- We will also work on becoming familiar with the **Integrative Genomics Viewer (IGV)** software, which we will use many times during later Exercise Sets
- Goal: prepare for analysis of DNA-sequencing data in Exercise Set 2 (starting Wednesday)

Exercise Set 1 overview

1. Overview of the computing environment
2. Parts of a bioinformatics analysis: reference genome, annotation files, and raw data files
3. Introduction to IGV

Optional:

- Programming in R for Bioinformatics


KAROLINSKA INSTITUTET
ANNO 1810
Karolinska
Institutet