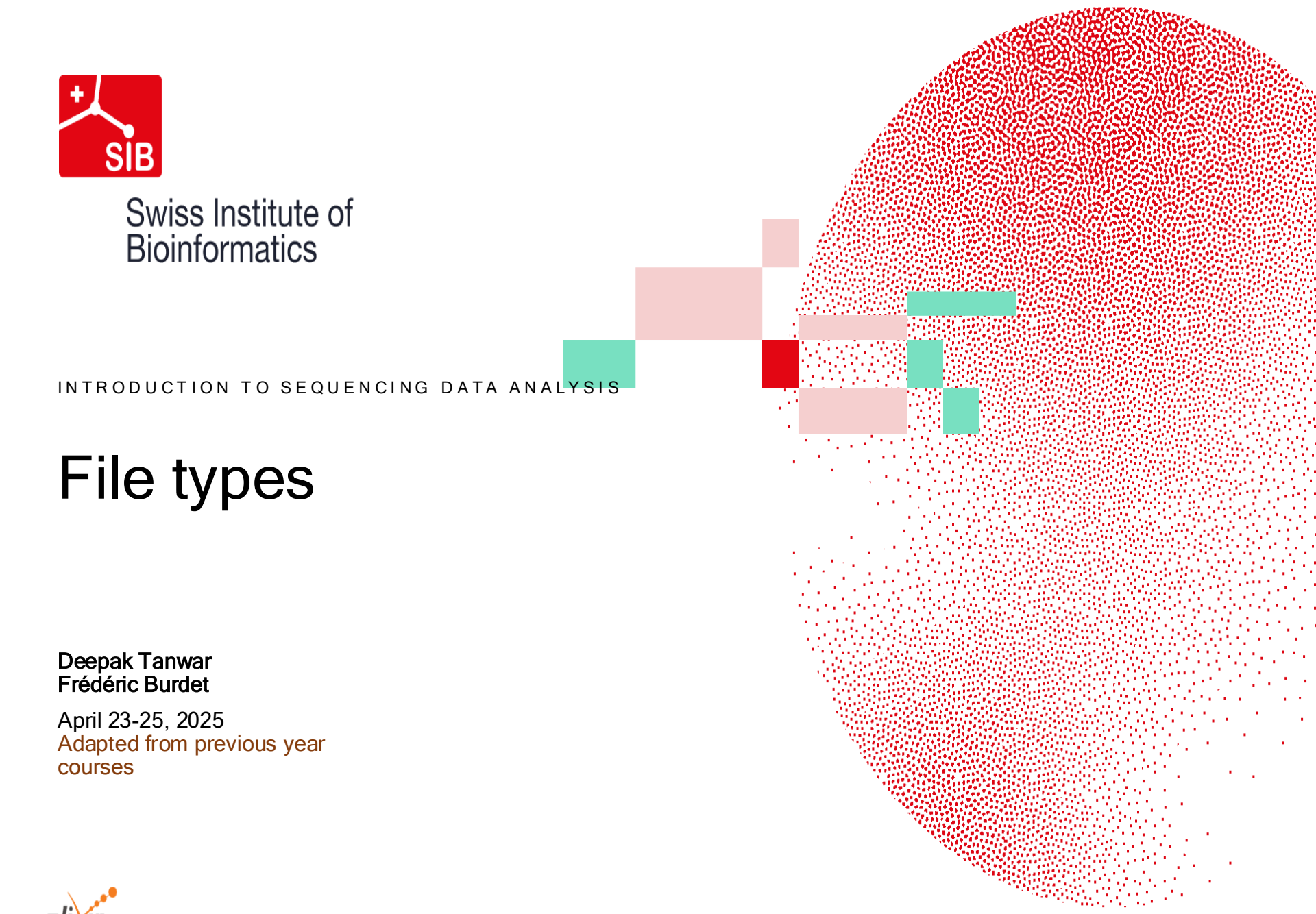


Swiss Institute of Bioinformatics

INTRODUCTION TO SEQUENCING DATA ANALYSIS

# File types

**Deepak Tanwar**
**Frédéric Burdet**

April 23-25, 2025
Adapted from previous year courses

elixir SWITZERLAND

# Frequently used file types

| | |
|---|---|
| fasta | sequences |
| fastq | reads |
| sam/bam | alignments |
| bed | regions |
| gff | annotations |
| vcf | variants |

Broad overview of file types:
https://genome-euro.ucsc.edu/FAQ/FAQformat.html

# fasta

- Plain sequence: *.fasta or *.fa
- Nucleotides or amino acids (proteins)
- Useful command:

```
grep -c "^>" sequence.fasta
```

sequence.fasta

```
>sequence title1
ATCGTATCTATCGTATCT
GGTTTATCGTATCT
>sequence title2
ATGATGACGT
```

# fastq files

reads.fastq

```
@D00283R:66:CC611ANXX:4:2311:2596:2330 1:N:0:TCCGGAG
ACTCTACGCTCAATAAAGATTTCTGATACGGCTCCTGAAATGCAGAATGAGT
+
B/<<<B<FFFFFFFFFBBFFFBFFFFBFFFF/FFFFFFFF/BFFFFFBFFF
```

title, starts with @

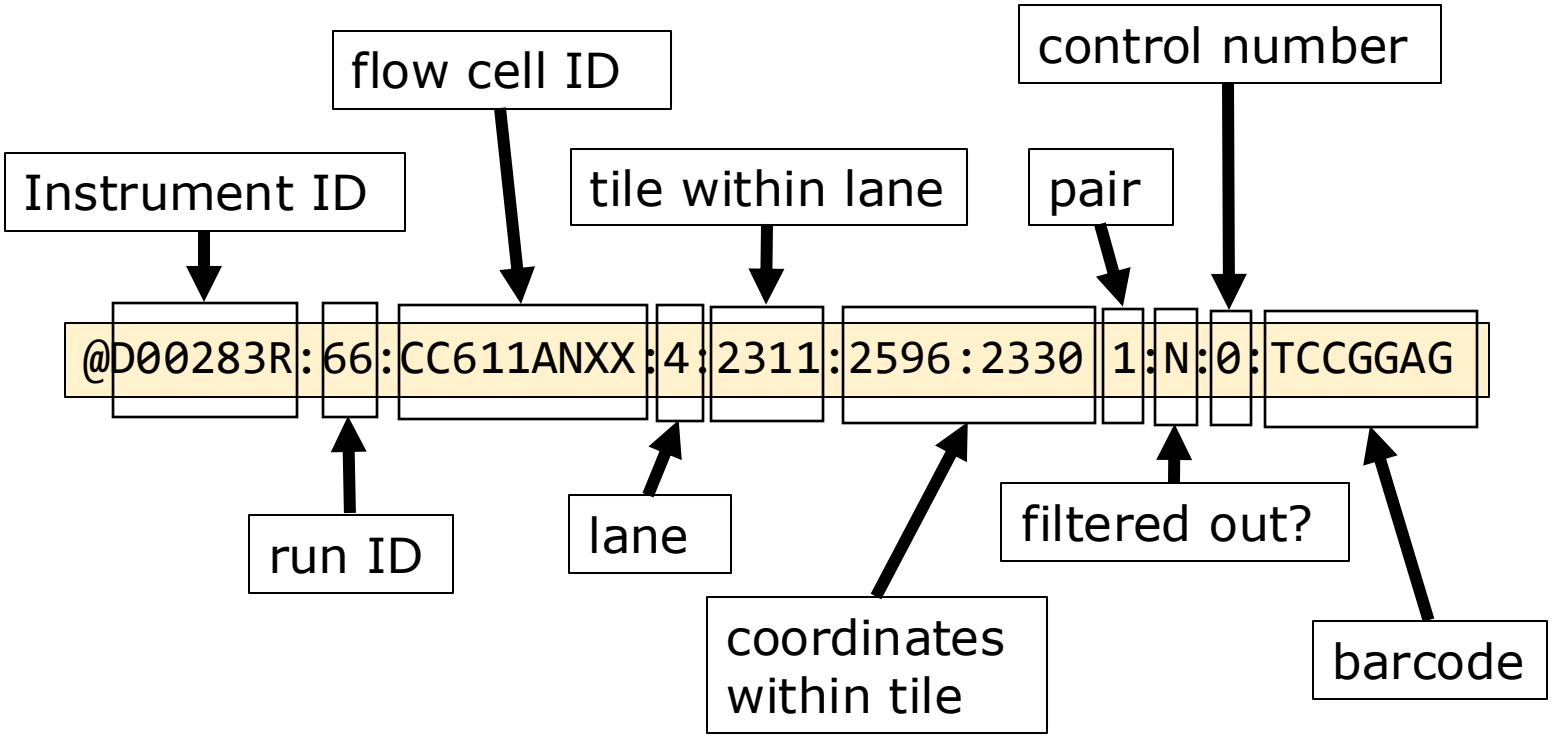
nucleotide sequence

optional description

base quality

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                              |    |        |
0.2.....................26...31........41
```

# fastq header



# Quiz Question 9

# sam

sequence alignment format

Aim: alignments

# sam header

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:U00096.3     LN:4641652
@PG     ID:bowtie2      PN:bowtie2      VN:2.4.1
CL: "/opt/miniconda3/envs/ngs/bin/bowtie2-align-s \
--wrapper basic-0 \
-x /home/ubuntu/ecoli/ref_genome//ecoli-strK12-MG1655.fasta \
-1 /home/ubuntu/ecoli/trimmed_data/paired_trimmed_SRR519926_1.fastq \
-2 / home/ubuntu/ecoli/trimmed_data/paired_trimmed_SRR519926_2.fastq"
```

# sam header – human data

```
@HD     VN:1.4  SO:coordinate
@SQ     SN:chr1 LN:248956422
@SQ     SN:chr2 LN:242193529
@SQ     SN:chr3 LN:198295559
@SQ     SN:chr4 LN:190214555
@SQ     SN:chr5 LN:181538259
@SQ     SN:chr6 LN:170805979
@SQ     SN:chr7 LN:159345973
@SQ     SN:chr8 LN:145138636
@SQ     SN:chr9 LN:138394717
@SQ     SN:chr10        LN:133797422
@SQ     SN:chr11        LN:135086622
@SQ     SN:chr12        LN:133275309
@SQ     SN:chr13        LN:114364328
@SQ     SN:chr14        LN:107043718
@SQ     SN:chr15        LN:101991189
@SQ     SN:chr16        LN:90338345
@SQ     SN:chr17        LN:83257441
@SQ     SN:chr18        LN:80373285
@SQ     SN:chr19        LN:58617616
@SQ     SN:chr20        LN:64444167
@SQ     SN:chr21        LN:46709983
@SQ     SN:chr22        LN:50818468
@SQ     SN:chrX LN:156040895
@SQ     SN:chrY LN:57227415
@SQ     SN:chrM LN:16569
```

# One aligned read from the E.coli dataset

@D00283R:66:CC611ANXX… 81 U00096.3 107 24 6M2I2M2I239M = 116 261
GCTCTTCCGATCTTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTCCA
CAACATCCATGAAACGCATTAGCACCACCATTACCCCCACCATCACCATTACCACAGGTAACGGTGCGGGCTGCCGC
GTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACA
ACCATGCGAGTGTTGAAGTT
###@<0>?4(A>:(4:>>:@>CB@>:>4@@>>8+@CCC:@4(>:3<0.7DCC>C>>::>>@8DCACC@::CCA::((+(A@4::
4C>:>><<2@AACCA@><9?:+AC@>)&50&?:3CC>:>4C@@::9?@:@CDBBB?BDDDDDDDDDBDDDDCCCC
DDDDCACCDDDDDDDDDDDDDCC>DDDDDDDDDDDDDEDDFHJJJJJIJJIJJJJJJJJJJJIJIJJJJIHFJJJJJJHHHHHFFFFF
CCC
AS:i:-44 XN:i:0 XM:i:7 XO:i:2 XG:i:4 NM:i:11 MD:Z:0T0T1A4C63A37A37A98 YS:i:-57 YT:Z:DP

| SAM column | example |
|---|---|
| read name | @D00283R:66:CC61… |
| flag | 81 |
| reference | U00096.3 |
| start position | 107 |
| mapping quality | 24 |
| CIGAR string | 6M2I2M2I239M |
| reference name mate is mapped | = |
| start position mate | 116 |
| fragment length | 261 |
| sequence | GCTCTTCCGA |
| base quality | ###@<0>?4( |
| optional | AS:i:-44 |
| optional | XN:i:0 |

# Tags explanation

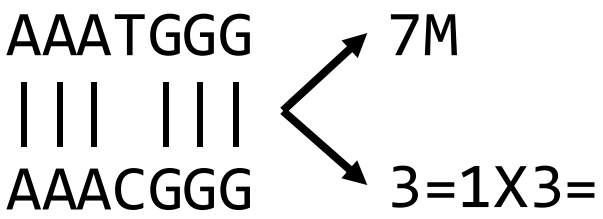
| Tag | Meaning | Value | Notes |
|---|---|---|---|
| AS | Alignment score | -44 | Lower = worse |
| XN | Ambiguous bases in ref (e.g. Ns) | 0 | None |
| XM | Mismatches | 7 | Align-specific |
| XO | Gap opens | 2 | 2 insertions |
| XG | Gap bases total | 4 | 2 insertions of 2 |
| NM | Edit distance | 11 | 7 mismatches + 4 indel bases |
| MD | Reference mismatches | 0T0T1A4... | Ref-level |
| YS | Mate's alignment score | -57 | Worse than this read |
| YT | Alignment type | DP | Duplicate Pair |

# CIGAR strings

Concise Idiosyncratic Gapped Alignment Report

| Op | BAM | Description |
|---|---|---|
| M | 0 | alignment match (can be a sequence match or mismatch) |
| I | 1 | insertion to the reference |
| D | 2 | deletion from the reference |
| N | 3 | skipped region from the reference |
| S | 4 | soft clipping (clipped sequences present in SEQ) |
| H | 5 | hard clipping (clipped sequences NOT present in SEQ) |
| P | 6 | padding (silent deletion from padded reference) |
| = | 7 | sequence match |
| X | 8 | sequence mismatch |

Almost never used




https://www.biostars.org/p/17043/
https://www.biostars.org/p/354942/
https://samtools.github.io/hts-specs/SAMv1.pdf

# Quiz Question 10

# sam flags

| 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 |

https://samtools.github.io/hts-specs/SAMv1.pdf

| | read paired? | properly aligned? | unmapped? | mate unmapped? | flag |
|---|---|---|---|---|---|
| **read1** | 1 | 1 | 0 | 0 | 3 |
| **read2** | 1 | 0 | 0 | 1 | 9 |
| **read3** | 1 | 0 | 1 | 1 | 13 |

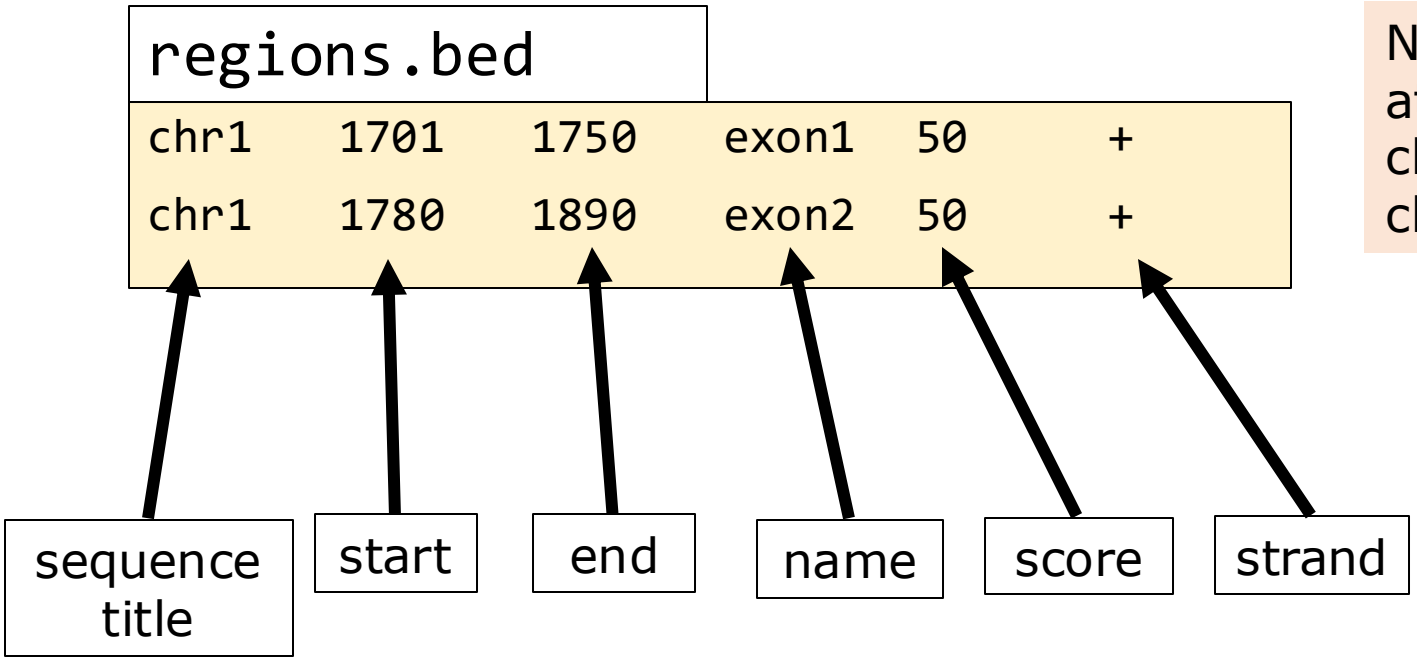
| 1 | 2 | 4 | 8 |
|---|---|---|---|

# Quiz Question 11

# bed

Browser Extensible Data

Aim: specify regions




Numbering starts at 0!
chr1:1702-1750
chr1:1781-1890

# Simple DNA Sequence Example

- Sequence: ACTGAT

| | | | | | | |
|---|---|---|---|---|---|---|
| 0-based Position: | 0 | 1 | 2 | 3 | 4 | 5 |
| 1-based Position: | 1 | 2 | 3 | 4 | 5 | 6 |
| Base: | A | **C** | **T** | **G** | A | T |

# The 0-based System

- Used by: BED, BigBed, Python, C, etc.

- The first nucleotide (A) is at position 0
- Intervals are defined as [start, end) → start included, **end excluded**

- Example:
- A BED interval with start = 1, end = 4 → positions

  1, 2, 3 → C, T, G

# The 1-based System

- Used by: GTF/GFF, VCF, SAM/BAM

- The first nucleotide (A) is at position 1
- Intervals are defined as [start, end] → start and **end included**

- Example:
- A GTF interval with start = 2, end = 4 → positions

  2, 3, 4 → C, T, G

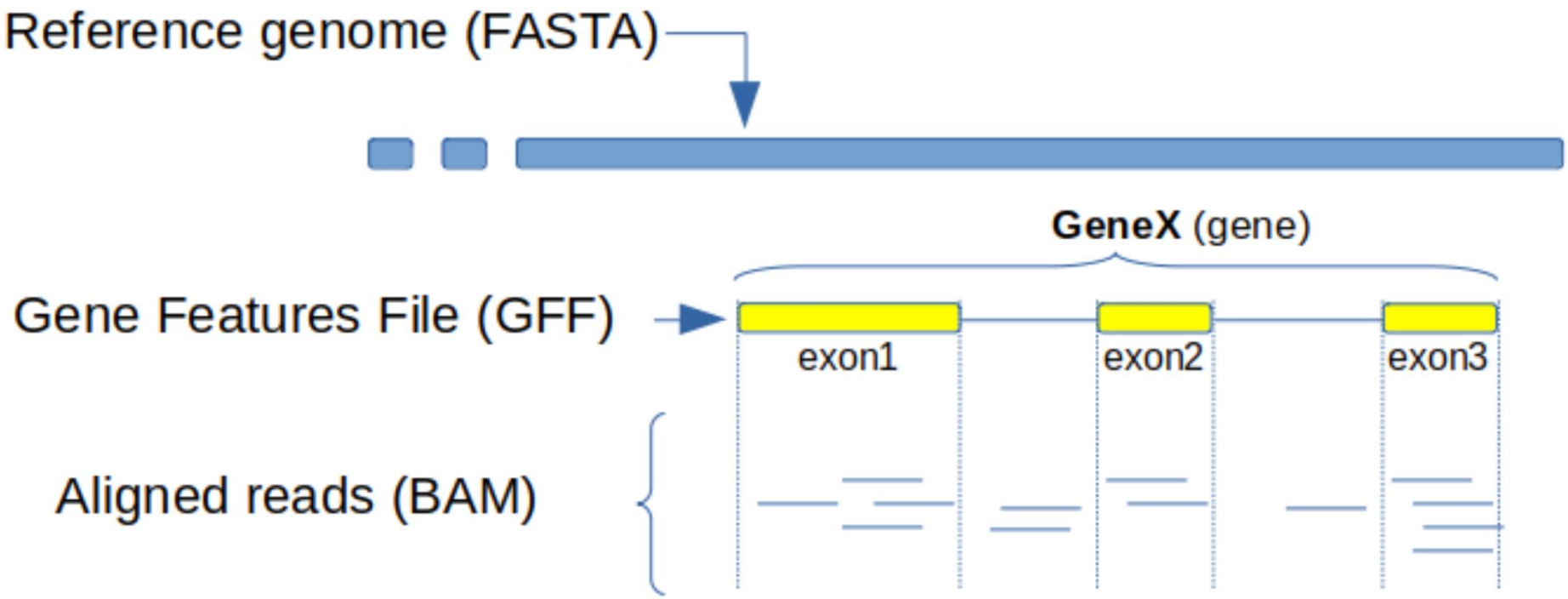
# Crucial Difference

- Example (C-T-G)
- BED 0 [start, end) start = 1, end = 4
- GTF 1 [start, end] start = 2, end = 4

- Same bases, different coordinates!

Info summarized by chatGPT

# gff




From https://bioinfogp.cnb.csic.es/tools/seqnjoy/help/

# gff

General Feature Format

Aim: annotation

| seq name | source | feature | start | end | score | strand | frame | attributes |
|---|---|---|---|---|---|---|---|---|
| 1 | ensembl | mRNA | 339070 | 346959 | . | - | . | ID=…; |
| 1 | ensembl | exon | 339070 | 339312 | . | - | . | Parent=; … |
| 1 | ensembl | CDS | 339070 | 339312 | . | - | 0 | ID=;… |

```
##gff-version 3
#description: evidence-based annotation of the human genome (GRCh38), version 45 (Ensembl 111)
#provider: GENCODE
#contact: gencode-help@ebi.ac.uk
#format: gff3
#date: 2023-09-19
##sequence-region chr1 1 248956422
chr1 HAVANA gene 11869 14409 . + . ID=ENSG00000290825.1;gene_id=ENSG00000290825.1;gene_type=lncRNA
;gene_name=DDX11L2;level=2;tag=overlaps_pseudogene
chr1 HAVANA transcript 11869 14409 . + . ID=ENST00000456328.2;Parent=ENSG00000290825.1;gene_id=ENSG
00000290825.1;transcript_id=ENST00000456328.2;gene_type=lncRNA;gene_name=DDX11L2;transcript_type=lncRNA;transcript_
name=DDX11L2-
202;level=2;transcript_support_level=1;tag=basic,Ensembl_canonical;havana_transcript=OTTHUMT00000362751.1
chr1 HAVANA exon 11869 12227 . + . ID=exon:ENST00000456328.2:1;Parent=ENST00000456328.2;gene_id=ENS
G00000290825.1;transcript_id=ENST00000456328.2;gene_type=lncRNA;gene_name=DDX11L2;transcript_type=lncRNA;transcrip
t_name=DDX11L2-
202;exon_number=1;exon_id=ENSE00002234944.1;level=2;transcript_support_level=1;tag=basic,Ensembl_canonical;havana_tra
nscript=OTTHUMT00000362751.1
chr1 HAVANA exon 12613 12721 . + . ID=exon:ENST00000456328.2:2;Parent=ENST00000456328.2;gene_id=ENS
G00000290825.1;transcript_id=ENST00000456328.2;gene_type=lncRNA;gene_name=DDX11L2;transcript_type=lncRNA;transcrip
t_name=DDX11L2-
202;exon_number=2;exon_id=ENSE00003582793.1;level=2;transcript_support_level=1;tag=basic,Ensembl_canonical;havana_tra
nscript=OTTHUMT00000362751.1
chr1 HAVANA exon 13221 14409 . + . ID=exon:ENST00000456328.2:3;Parent=ENST00000456328.2;gene_id=ENS
G00000290825.1;transcript_id=ENST00000456328.2;gene_type=lncRNA;gene_name=DDX11L2;transcript_type=lncRNA;transcrip
t_name=DDX11L2-
202;exon_number=3;exon_id=ENSE00002312635.1;level=2;transcript_support_level=1;tag=basic,Ensembl_canonical;havana_tra
nscript=OTTHUMT00000362751.1
```

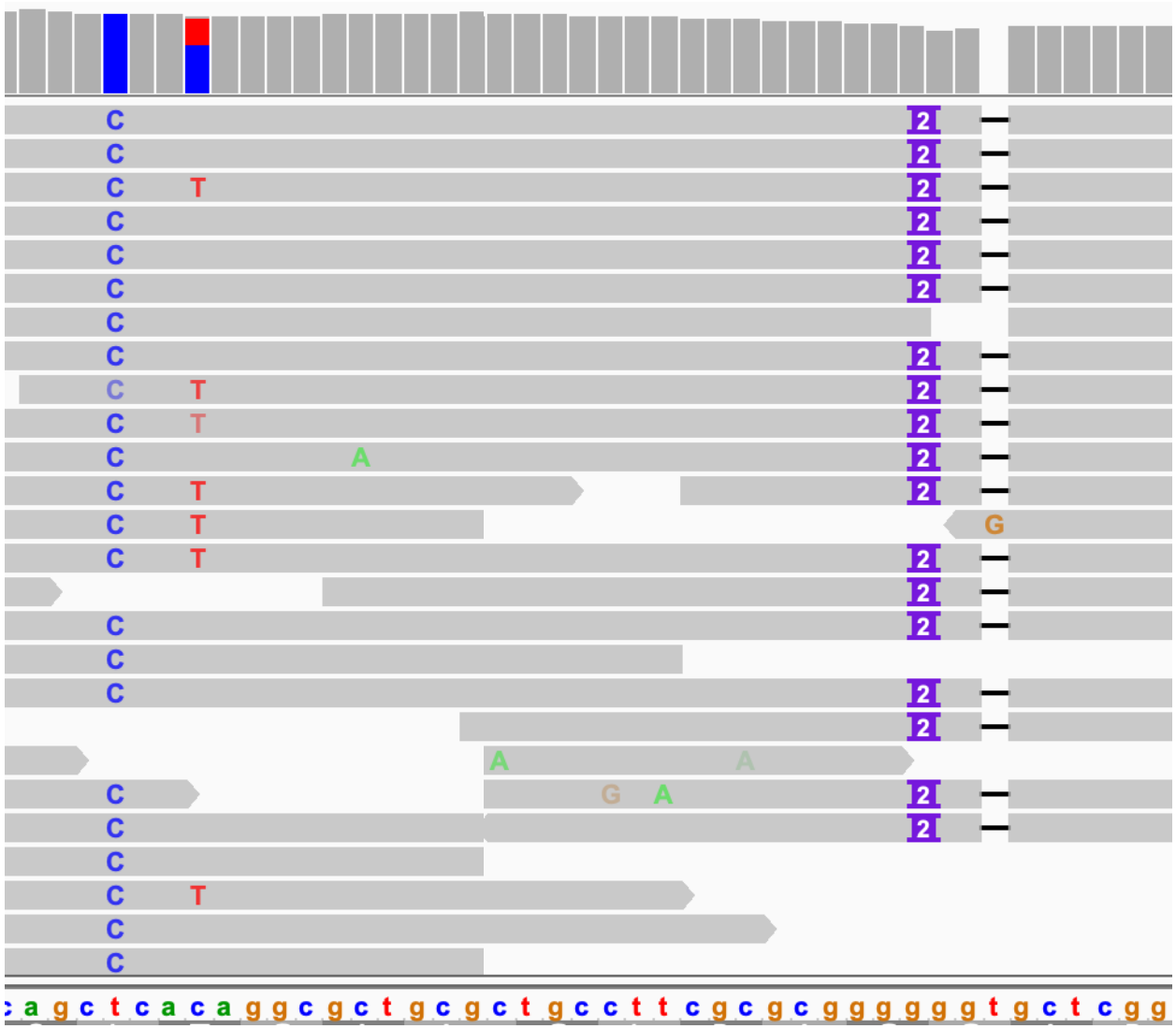
# vcf

## Variant Call Format

## Aim: variants

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

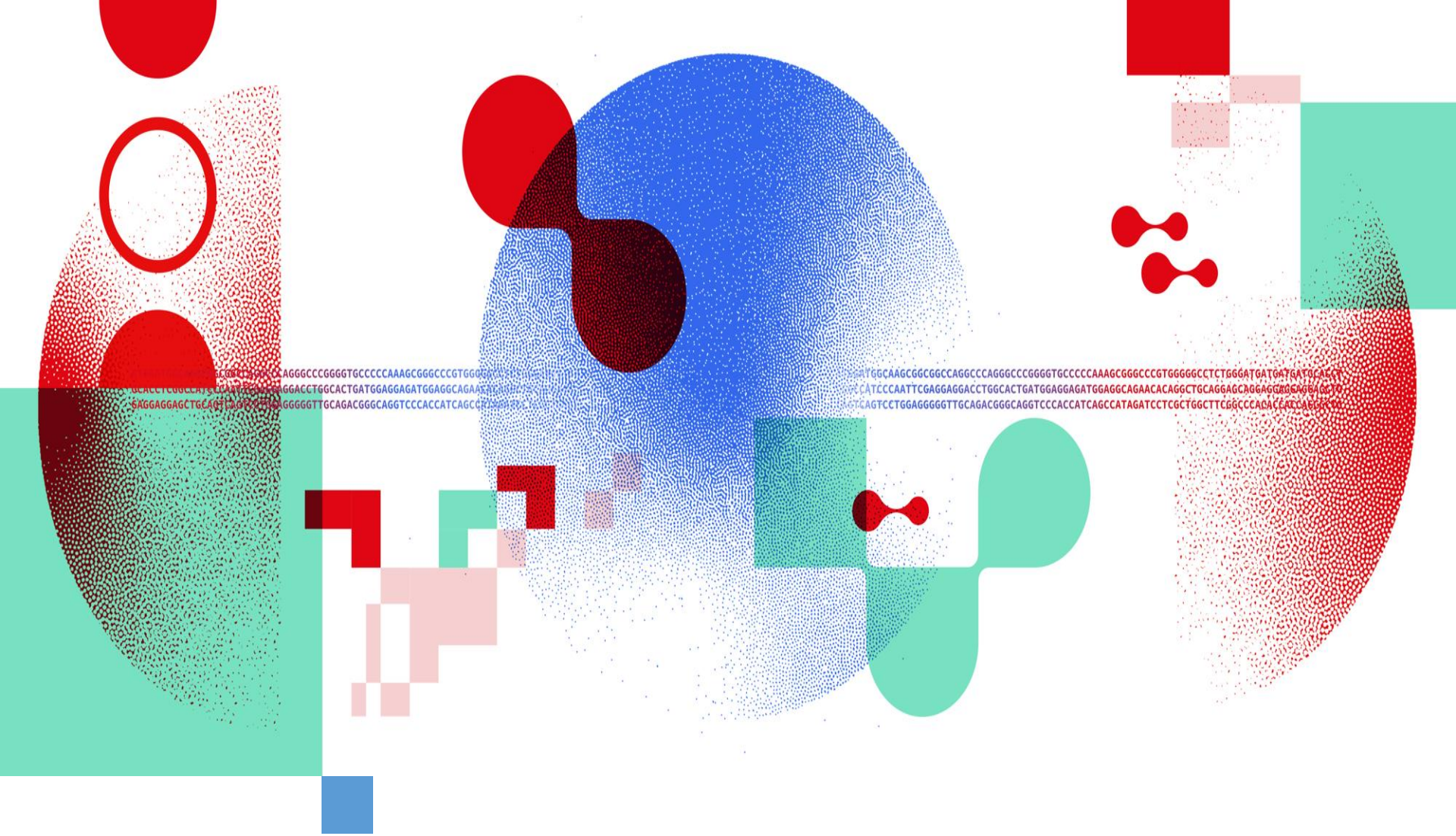
SIB



# vcf

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF  ALT    QUAL FILTER  INFO                              FORMAT      NA00001         NA00002
20     14370    rs6054257  G    A      29   PASS    NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51  1|0:48:8:51,51
20     17330    .          T    A      3    q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50  0|1:3:5:65,3
20     1110696  rs6040355  A    G,T    67   PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27  2|1:2:0:18,2
20     1230237  .          T    .      47   PASS    NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60  0|0:48:4:51,51
20     1234567  microsat1  GTC  G,GTCT 50   PASS    NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4        0/2:17:2
```

samples



# Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss