

File Types in Bioinformatics

2023-02-07

Martin Dahlö
martin.dahlo@nbis.se

Enabler for Life Sciences

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)



- Overwhelming at first
- Overview
 - FASTA – reference sequences
 - FASTQ – reads in raw form
 - SAM – aligned reads
 - BAM – compressed SAM file
 - CRAM – even more compressed SAM file
 - GTF/GFF/BED – annotations

- Used for: nucleotide or peptide sequences
- Simple structure

> header
sequence

- Used for: nucleotide or peptide sequences
- Simple structure

```
> H.Sapiens chr17:135135135-1313566  
ACTCAGATCGGAATAGCATACGCATACTCAGATCGGAATAGCATACGCAT  
GGATAGCTCACGACACATGACACTACAGCCAGACTACACGACTACACGAT  
AAGGATATAGGACTACGACTAGCATCGACTAACTAGCTACATACG
```

```
>that random protein sequence i saw yesterday  
ARGAEBAEUIRGHAERGI AEUAEL LHGAEL GAHEGLAEJKRGNAERBI AE  
AEGHAELGIHAEGOUIAENGAEBARI OTYUGAEGHILAEHRGAEIRGYU  
AEHAELAEI OGAEGAERTBETHUETHIRTHJNRFS
```

- Just like FASTA, but with quality values
- Used for: raw data from sequencing (unaligned reads)

@ header
sequence
+
quality

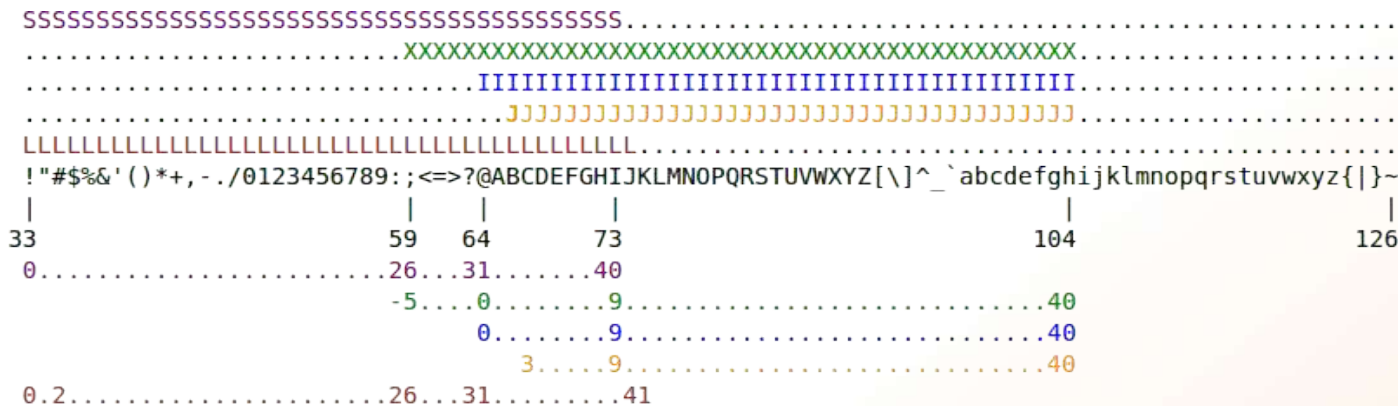
- Just like FASTA, but with quality values
- Used for: raw data from sequencing (unaligned reads)

```
@SEQ_001
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%++) (%%%! ''*(((((**%) .1***-+*'')) **55CC! ''*(D
@SEQ_002
GATTTGGGGTTCAAAGCAGTATTTGGGGTTCATTGGGGTTCATTGTTCAACTCACAGTTT
+
!''*(((((***+))%%>>CCCC%++ ((( (**).1***-+*'')) **55CCF>>>>>>C5
@SEQ_003
AAGCAGTATCGAGATTTGGGGTTCAAAGCAGTAT AAGCAGTATCGATAAATCCATTTGTT
+
!''*((((( (*! ''*((((( (**)(%%%) .1***-+*'')) **55CCF>>>>>>%%%) .1B5
```


- Quality 0-40
 - 40 = best
- ASCII encoded

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	\$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(72	48	H	104	68	h
9	09	Horizontal tab	41	29)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	
29	1D	Group separator	61	3D	=	93	5D]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F		127	7F	□

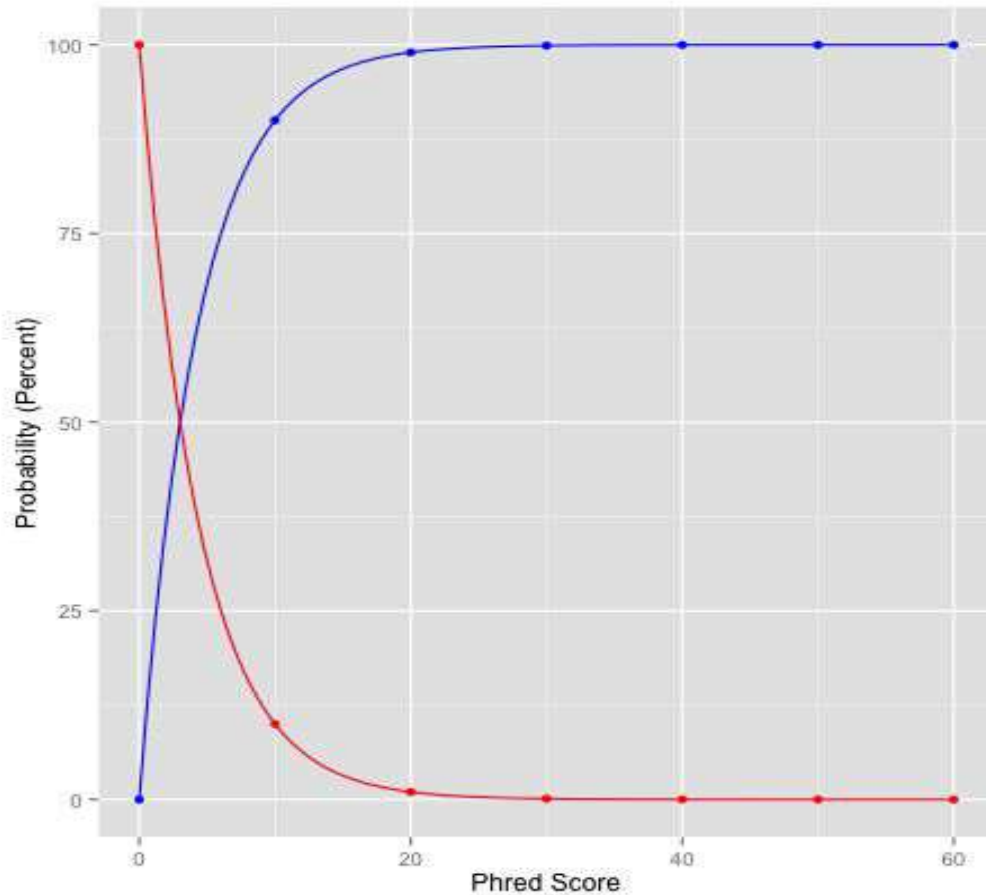
- Quality 0-40 (Illumina 1.8+ = 41)
 - 40 = best
- ASCII encoded



- S - Sanger Phred+33, raw reads typically (0, 40)
- X - Solexa Solexa+64, raw reads typically (-5, 40)
- I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
- J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
- L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

- Quality 0-40 (Illumina 1.8+ = 41)
 - 40 = best
- ASCII encoded

```
@SEQ_001
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!' '*((( (***) )%%%++) (%%%! ' '*((( (**%) .1***-+*' ')) **55CC!' '* (D
@SEQ_002
GATTTGGGGTTCAAAGCAGTATTTGGGGTTCATTGGGGTTCATTGTTCAACTCACAGTTT
+
!' '*((( (***) )%%>>CCCC%++ ((( (**).1***-+*' ')) **55CCF>>>>>>C5
@SEQ_003
AAGCAGTATCGAGATTTGGGGTTCAAAGCAGTAT AAGCAGTATCGATAAATCCATTTGTT
+
!' '*((( (*!' '*((( (**) (%%%) .1***-+*' ')) **55CCF>>>>>>%%%) .1B5
```



Functions
 ● Accuracy
 ● Error

Phred Quality Score	Error	Accuracy
10	$1/10 = 10\%$	90%
20	$1/100 = 1\%$	99%
30	$1/1000 = 0.1\%$	99.9%
40	$1/10000 = 0.01\%$	99.99%
50	$1/100000 = 0.001\%$	99.999%
60	$1/1000000 = 0.0001\%$	99.9999%

- Used for: aligned reads
- Lots of columns..

sequence_string.sam

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL> [<TAG>:<VTYPE>:<VALUE> [...]]

Field	Regular expression	Range	Description
QNAME	[[^] \t\n\r]+		Query pair NAME if paired; or Query NAME if unpaired ²
FLAG	[0-9]+	[0,2 ¹⁶ -1]	bitwise FLAG (Section 2.2.2)
RNAME	[[^] \t\n\r@=]+		Reference sequence NAME ³
POS	[0-9]+	[0,2 ²⁹ -1]	1-based leftmost POSition/coordinate of the clipped sequence
MAPQ	[0-9]+	[0,2 ⁸ -1]	MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect) ⁴
CIGAR	([0-9]+[MIDNSHP])+ *		extended CIGAR string
MRNM	[[^] \t\n\r@]+		Mate Reference sequence NaMe; “=” if the same as <RNAME> ³
MPOS	[0-9]+	[0,2 ²⁹ -1]	1-based leftmost Mate POSition of the clipped sequence
ISIZE	-?[0-9]+	[-2 ²⁹ ,2 ²⁹]	inferred Insert SIZE ⁵
SEQ	[acgtnACGTN.=]+ *		query SEQUENCE; “=” for a match to the reference; n/N/. for ambiguity; cases are not maintained ^{6,7}
QUAL	[!-~]+ *	[0,93]	query QUALity; ASCII-33 gives the Phred base quality ^{6,7}
TAG	[A-Z][A-Z0-9]		TAG
VTYPE	[AifZH]		Value TYPE
VALUE	[[^] \t\n\r]+		match <VTYPE> (space allowed)

- Used for: aligned reads
- Lots of columns..



Read name

Start position
bp chr

Sequence

Quality

- Binary SAM (compressed)
- 25% of the size
- SAMtools to convert
- .bai = BAM index

Contents

1	Linux Introduction	1
1.1	Connecting to UPPMAX	1
1.2	Getting a node of your own	2
1.3	Moving and Looking Around	3
1.4	Copying files needed for laboratory	6
1.5	Unpack Files	7
1.6	Copying and Moving Files	8
1.7	Deleting Files	11
1.8	Open files	13
1.9	Wildcards	15
1.10	Utility Commands	16
2	Advanced Linux	20
2.1	Ownership & Permissions	20
2.1.1	Owners	20
2.1.2	Permissions	20
2.1.3	Interpreting the permissions of files and directories	21
2.1.4	Editing Ownership & Permissions	23
2.1.5	Assignment	24
2.2	Symbolic links - Files	24
2.2.1	Assignment	25
2.3	Symbolic links - Directories	26
2.3.1	Assignment	27
2.4	Grep - Searching for text	27
2.4.1	Assignment	28
2.5	Piping	29
2.6	Word Count	30
2.6.1	Assignment	31
2.7	Extra material 1	31
2.8	Extra material 2	32
2.9	Extra material 3	32
3	UPPMAX Tutorial	34
3.1	Copying files needed for laboratory	34
3.2	Running a program	35
3.3	Modules	38
3.4	Submitting a job	38
3.5	Viewing the queue	39
3.6	Interactive	40
3.7	Extra, if you finish too fast	41

- Random order
- Have to sort before indexing



- Random order
- Have to sort before indexing



Unsorted BAM



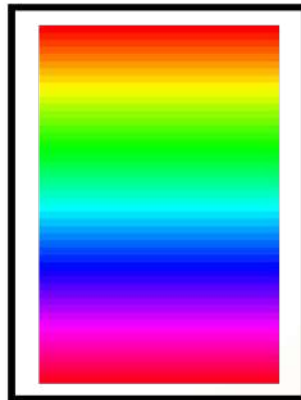
Unsorted BAM



samtools sort



Sorted BAM



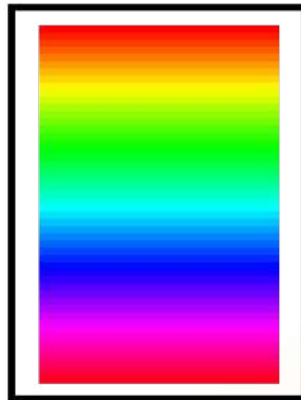
Unsorted BAM



samtools sort



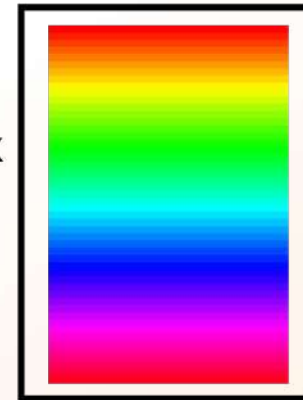
Sorted BAM



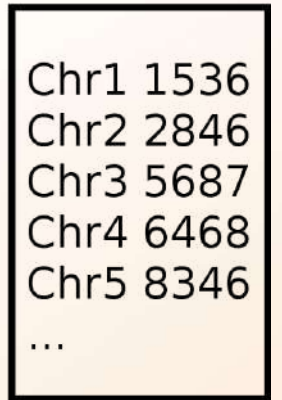
samtools index



Sorted BAM



BAM index



- Very complex format
- Used together with a reference genome

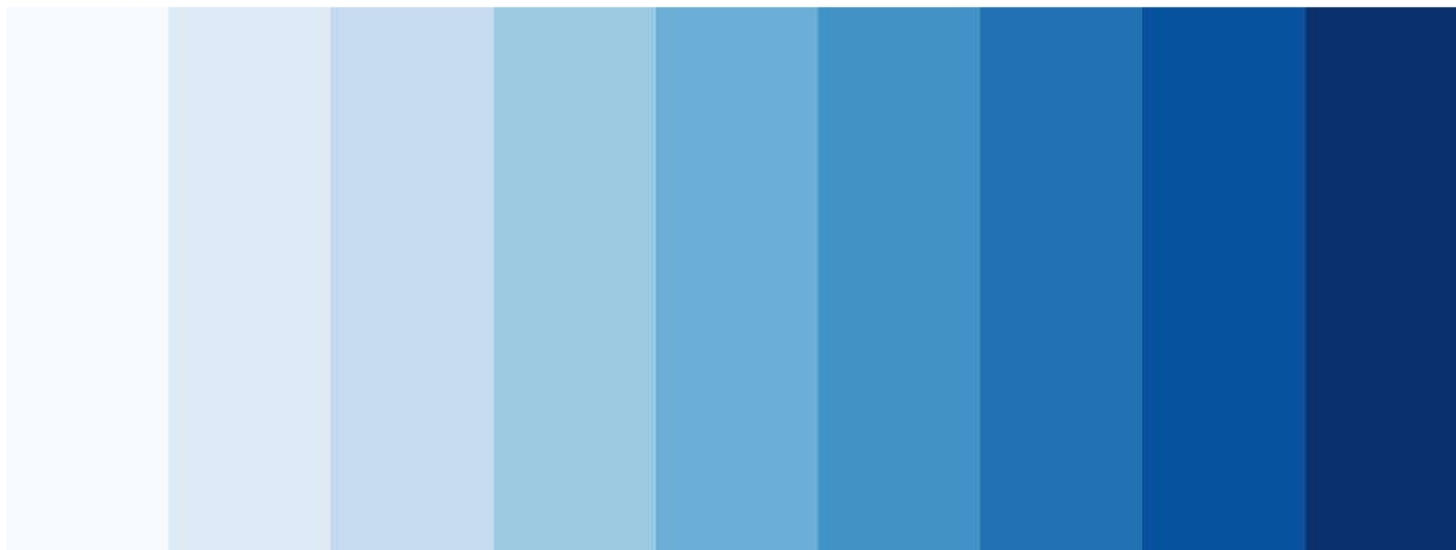
		AGGCTGAGTCACGACGTGTTGAGA	
Reads	TAGATCGAG	GGCTGAGTCACGACG	
	ATTCGGACGTAGATCGAG	GCTGAG	ACGTGTTGAGAGAGCCGTA
Ref:	ATTCGGACGTAGATCGACGCTGAGTCACGACGTGTTGTGAGAGCCGTAGAC		

- Quality scores?
- 3 modes:
 - Lossless
 - Binned
 - No quality



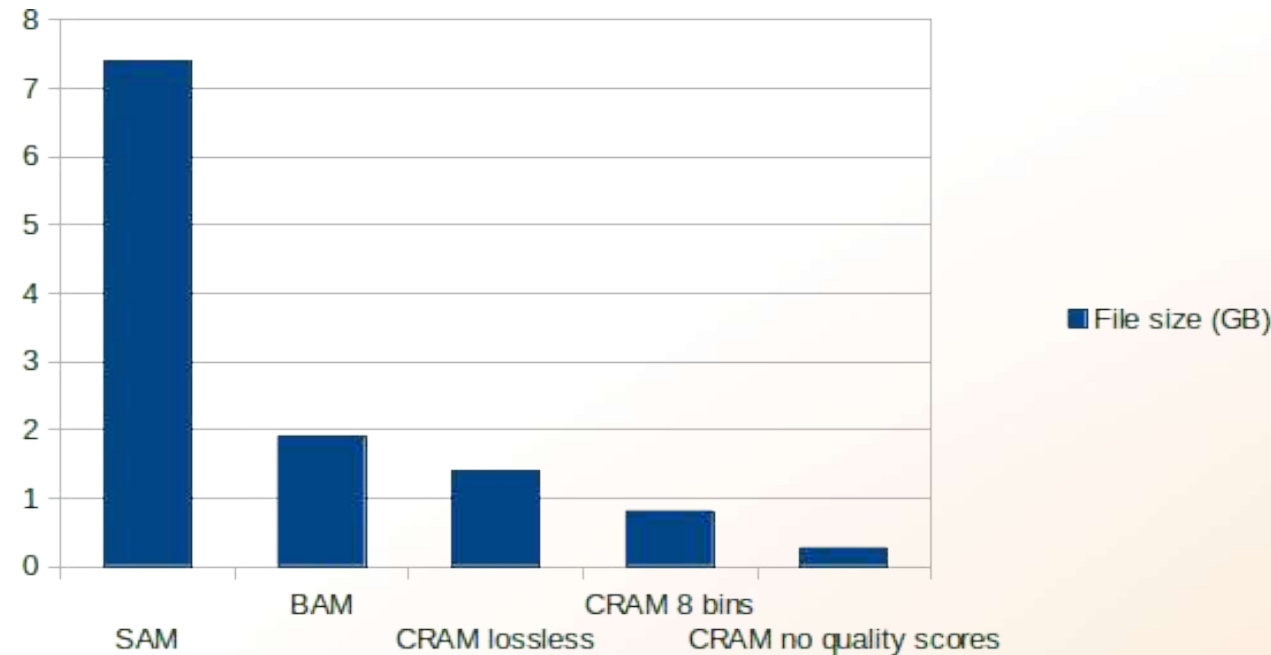
1 2 3 4 5 6 7 8 9 10 11 12 13 14 ... 32 33 34 35 36 37 38 39 40 41

1-5 6-10 11-15 16-20 21-25 26-30 31-35 35-40 41-45



=> Reducing the number of quality values increases shared blocks and improves compression.

- Quality scores?
- 3 modes:
 - Lossless
 - Binned
 - No quality



- Not widespread, yet

- Used for: annotations
- Column structure
- one line = one feature (match, exon, etc)

BED format:

- 3-12 columns
3 mandatory fields

+ 9 optional fields

chr	start	stop	extra info
chr1	213941196	213942363	
chr1	213942363	213943530	

BED format:

- optional fields

4. **name** - Label to be displayed under the feature, if turned on in "Configure this page".

5. **score** - A score between 0 and 1000.

6. **strand** - defined as + (forward) or - (reverse).

7. **thickStart** - coordinate at which to start drawing the feature as a solid rectangle

8. **thickEnd** - coordinate at which to stop drawing the feature as a solid rectangle

9. **itemRgb** - an RGB colour value (e.g. 0,0,255). Only used if there is a track line with the value of itemRgb set to "on" (case-insensitive).

10. **blockCount** - the number of sub-elements (e.g. exons) within the feature

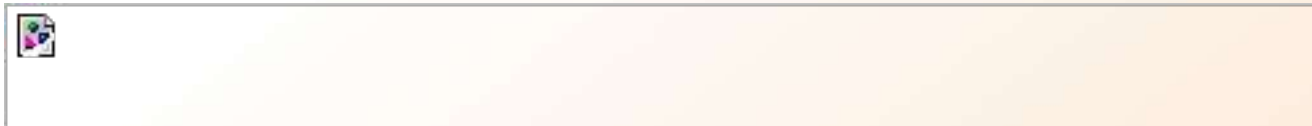
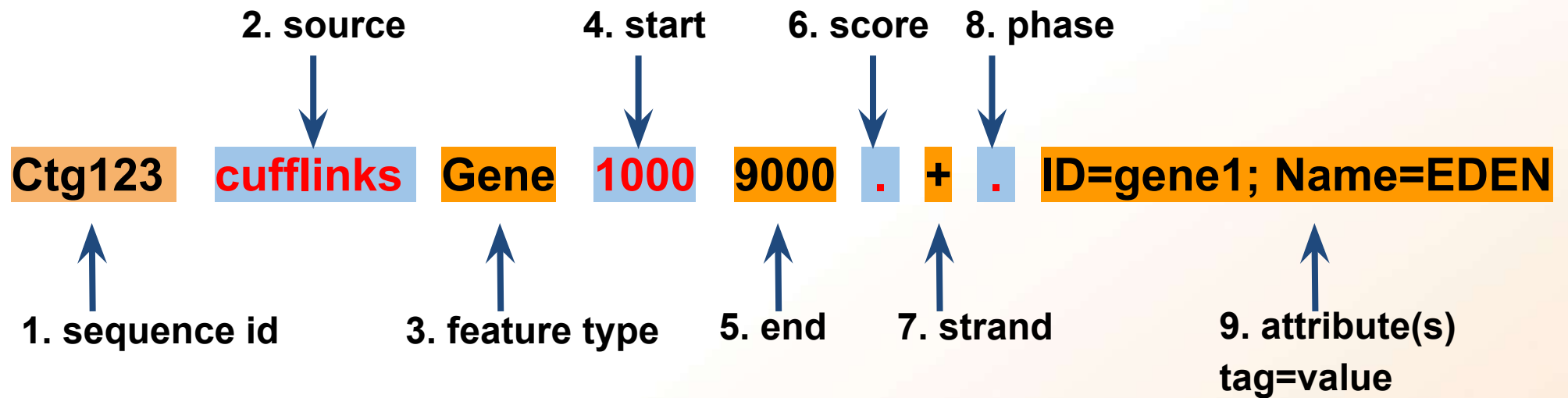
11. **blockSizes** - the size of these sub-elements

12. **blockStarts** - the start coordinate of each sub-element

```
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
```

GFF/GTF format:

- 9 columns



Laboratory time! (yet again)

<https://uppsala.instructure.com/courses/76870/pages/linux-3-file-types-in-bioinformatics-lab>