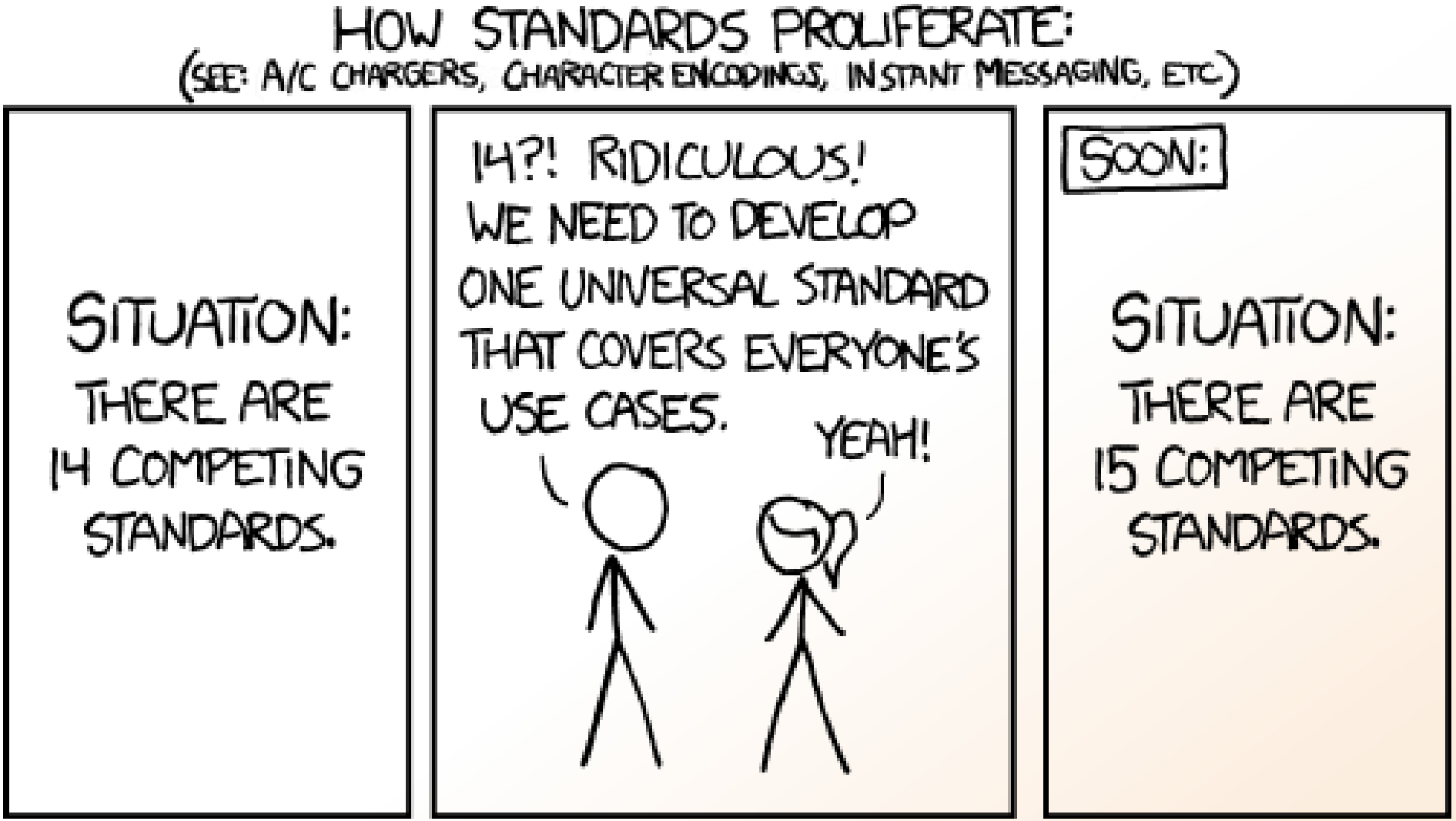
SciLifeLab
UPPSALA

File Types in Bioinformatics

150519
Martin Dahlö
martin.dahlo@scilifelab.uu.se

Enabler for Life Sciences

SciLifeLab
UPPSALA



http://xkcd.com

SciLifeLab
UPPSALA

- 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

FASTA

- Used for: nucleotide or peptide sequences
- Simple structure

> header

sequence

SciLifeLab UPPSALA

FASTA

- Used for: nucleotide or peptide sequences
- Simple structure

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

>that random protein sequence i saw yesterday
ARGAEBAEUIRGHAERGIAEUAEILHGAEIGAHEGLAEJKRGNAERBIAE
AEGHAELGIHAEGOUIAENGAEBAERIOTYUGAEGHILAEHRGAEIRGYU
AEHAEHAEIOGAEGAERTBETHUETHIRTHJNRFS
```

SciLifeLab UPPSALA

FASTQ

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

@ header

sequence

+

quality

SciLifeLab UPPSALA

FASTQ

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

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

SciLifeLab UPPSALA

FASTQ

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

SciLifeLab
UPPSALA

FASTQ

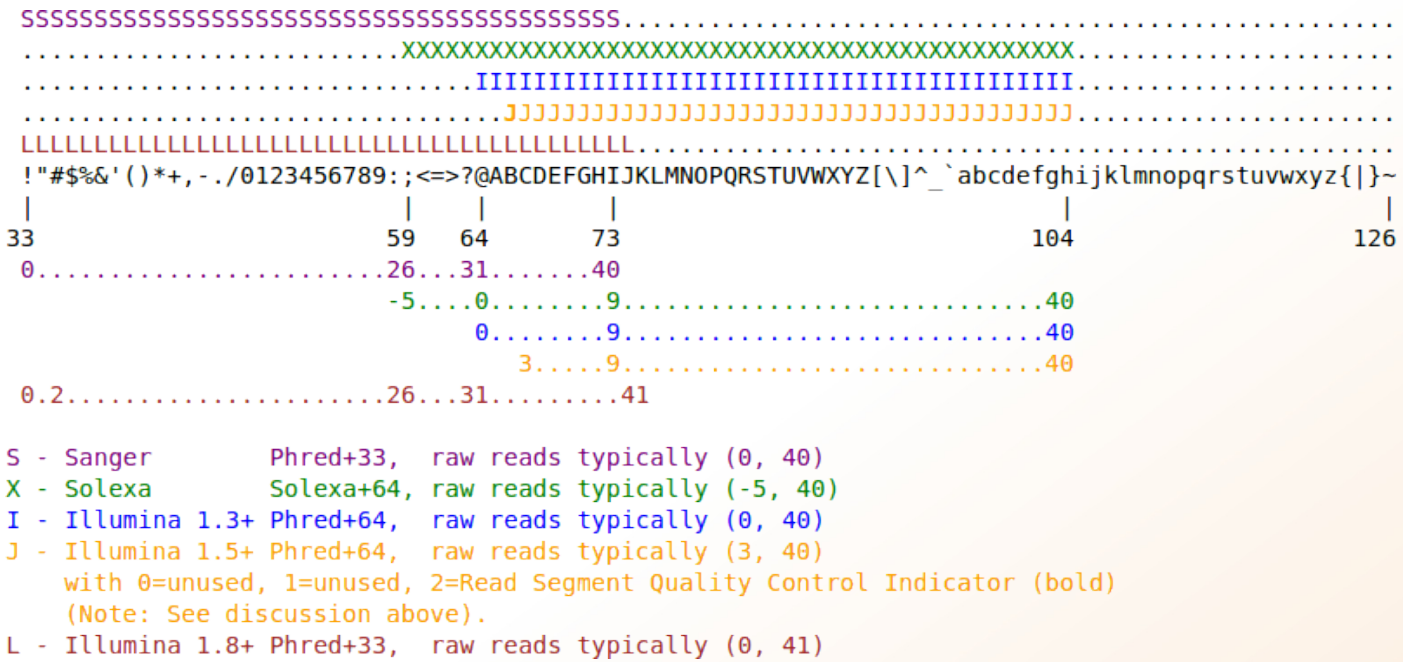
- Quality 0-40
  - 40 = best
- ASCII encoded

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	□

SciLifeLab UPPSALA

FASTQ

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



SciLifeLab UPPSALA

FASTQ

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

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

SciLifeLab UPPSALA

SAM

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

SciLifeLab UPPSALA

SAM

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)

SAM

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

```
@SQ     SN:31   LN:39895921
@PG     ID:bwa  PN:bwa  VN:0.7.8-r455   CL:bwa samse -f 02_sample.fq.sam /sw/data/uppnex/reference/Canis_familiaris/CanFam3/program_files/bwa/chr.31.fa 01_sample.fq.sai sample.fq
read_001        0       chr31   26546617        37      150M    *       0       0       AAAGGCTATTTCCACCT       )%%>((***+))%%>)%%>     XT:A:U  NM:i:0  XO:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:150
read_002        0       chr31   26546617        37      150M    *       0       0       AGGAGAAAGGCAGATCG       '*((!''*((((***+))%     XT:A:U  NM:i:0  XO:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:150
read_003        0       chr31   26546617        37      150M    *       0       0       AAAGGAGGCTAACGTTT       )%%>!''*((*+))%%((*     XT:A:U  NM:i:0  XO:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:150
read_004        0       chr31   26546617        37      150M    *       0       0       AGGCCATGACATCATCT       *((((***+))%%>)%%>%     XT:A:U  NM:i:0  XO:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:150
read_005        0       chr31   26546617        37      150M    *       0       0       TAGCAGAGCTATTTCAT       ((**!''*((*+))%%>AD     XT:A:U  NM:i:0  XO:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:150
```

Start position
bp chr

Sequence

Quality

Read name

SciLifeLab UPPSALA

BAM

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

Contents

SciLifeLab UPPSALA

BAM

- Random order
- Have to sort before indexing

SciLifeLab
UPPSALA

BAM

- Random order
- Have to sort before indexing

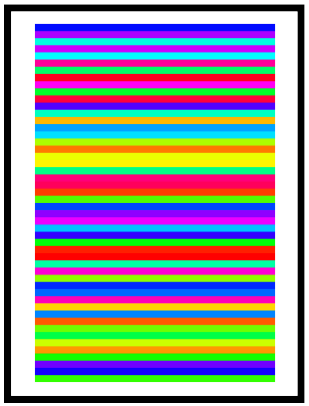
SciLifeLab UPPSALA

BAM

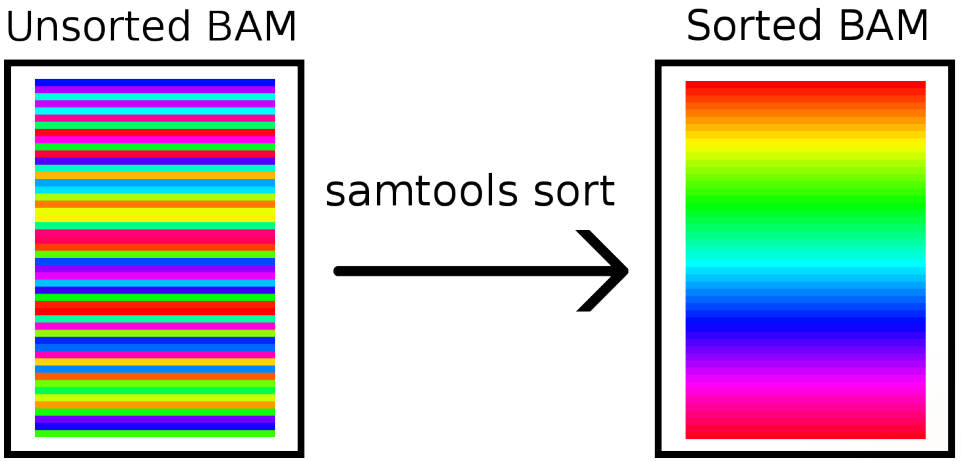
- Random order
- Have to sort before indexing

BAM

Unsorted BAM

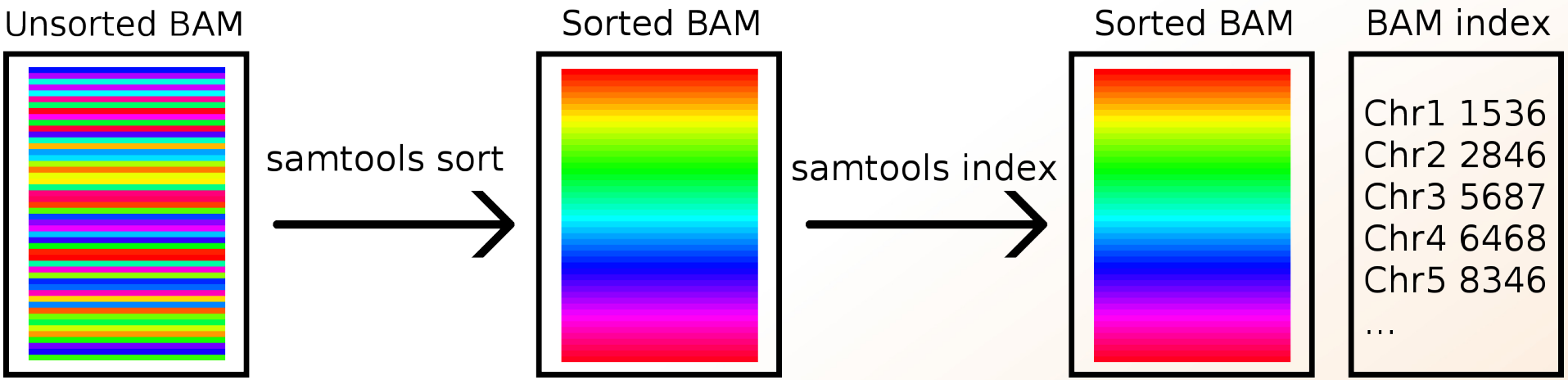


BAM



SciLifeLab UPPSALA

BAM



SciLifeLab UPPSALA

CRAM

- Very complex format
- Used together with a reference genome

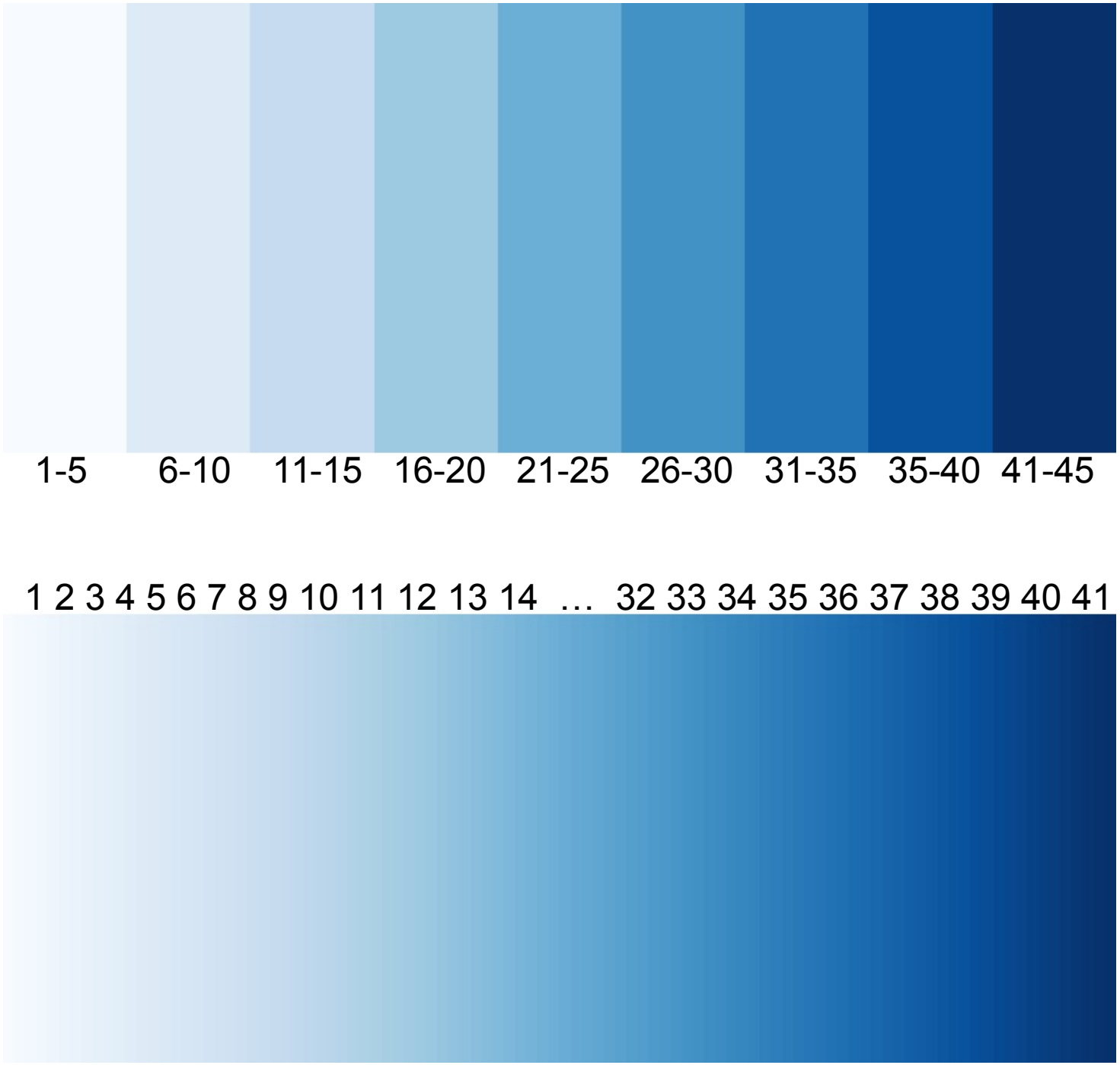
```
                                    AGGCTGAGTCACGACGTGTTGAGA
Reads                     TAGATCGAGGCTGAGTCACGACG
          ATTCGGACGTAGATCGAGGCTGAG          ACGTGTTGAGAGAGCCGTA

Ref:      ATTCGGACGTAGATCGACGCTGAGTCACGACGTGTTGTGAGAGCCGTAGAC
```

SciLifeLab UPPSALA

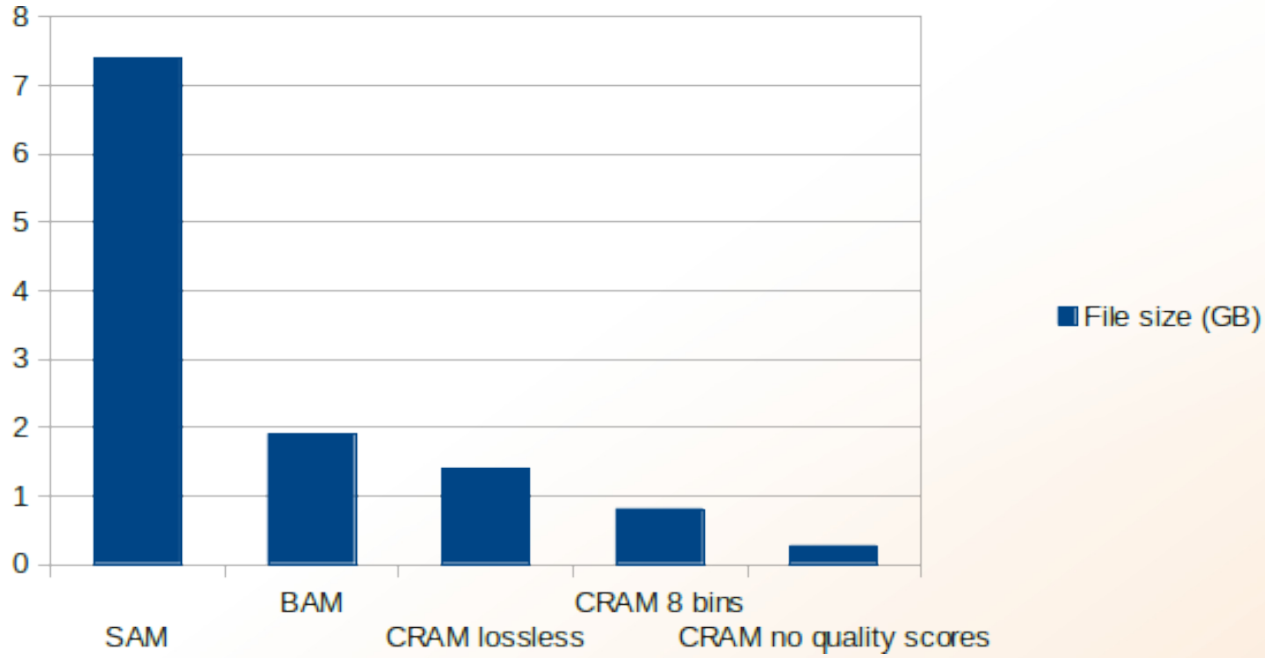
CRAM

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


1-5
6-10
11-15
16-20
21-25
26-30
31-35
35-40
41-45
1 2 3 4 5 6 7 8 9 10 11 12 13 14 … 32 33 34 35 36 37 38 39 40 41

CRAM

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



- Not widespread, yet

SciLifeLab UPPSALA

GTF/GFF/BED

- Used for: annotations
- Simple structure

- Usually:

  chr start stop extra info

SciLifeLab UPPSALA

GTF/GFF/BED

- Used for: annotations
- Simple structure

- Usually:

  chr start stop extra info

- BED

```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

GTF/GFF/BED

- Used for: annotations
- Simple structure

- Usually:

  chr start stop extra info

- GFF

```
chr22  TeleGene enhancer  10000000  10001000  500 +  .  touch1
chr22  TeleGene promoter  10010000  10010100  900 +  .  touch1
chr22  TeleGene promoter  10020000  10025000  800 -  .  touch2
```

- Laboratory time! (yet again)