Quick introduction to genomic file types
Preliminary quality control (lab)

File types overview

- Fasta/fasta qual
- Fastq
- SAM

Text files

- BAM
- sff

Binary files

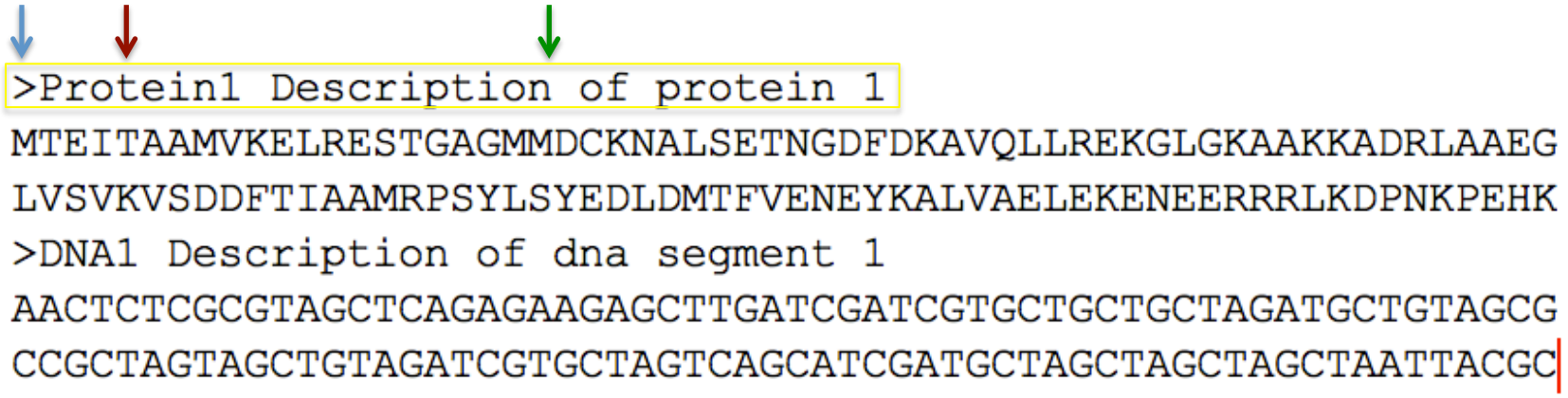
- …
- …

http://www.molecularevolution.org/resources/fileformats

Fasta

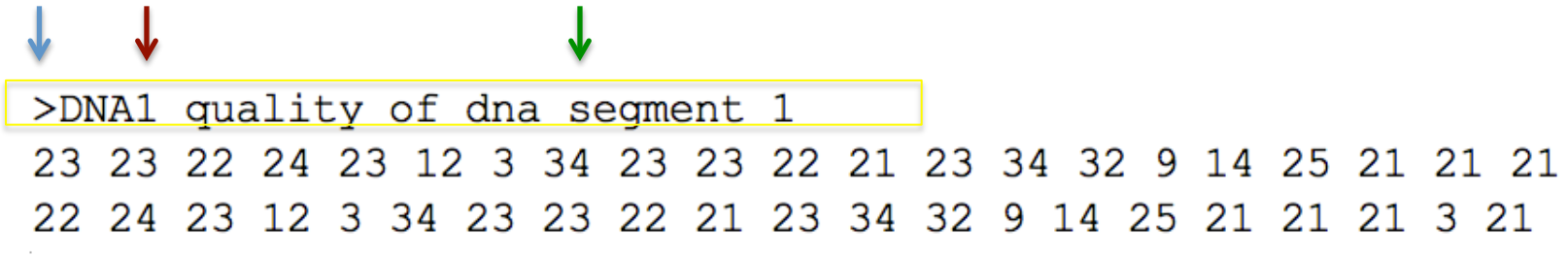
- Most basic file format for nucleotide or amino-acid sequences
- Each sequence requires:
  - A single description line (shouldn't exceed 80 characters):
  - Starts with "**>**"
  - Followed by **sequence ID** and a space
  - More information (**description**)
  - The sequence, over one or several lines

  (the number of characters per line is generally 70 or 80, but it doesn't matter)



Fasta qual

- Fasta-like quality format
- Always paired with a fasta file (sequences with same IDs, same order)
- Description line as in fasta format
- Qualities: a number for each base in the corresponding fasta, separated by spaces
- Can be gzip-ped and used as such by some programs



Quality - Phred scores

- Most common representation of qualities
- Related to the probability of errors (*P)* in a particular base

$$Q = -10\log_{10} P$$

$$P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

- Solexa runs < 1.3 use a different calculation:
  - Equivalent for high quality
  - Different for low quality (negative values of Q allowed)

FastQ

- A more compact format to store sequence and qualities
- Normally on 4 lines:
  - "@" followed by the sequence ID
  - Sequence
  - "+"
  - The quality score

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAA
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>
```

Example taken from Wikipedia

FastQ (contd.)

- **Quality score**:
  - ASCII encoding of phred scores
  - Sanger has one scale
  - Illumina has 3 differents

- Can be gzip-ped and used as such by some programs

ASCII decimal codes

Dec	Symbol	Dec	Symbol	Dec	Symbol
32	Space	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46		78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r

FastQ – quality values

- Solexa picked different quality definition and ranges over time, all different from Sanger values
- Ask your sequence provider!
- Guessing by getting the range of all values in all/many reads (not foolproof)

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
```

```
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)
```

Example taken from Wikipedia

SAM/BAM

- SAM (Sequence Alignment/Map) format is the alignment of sequences (e.g. reads) to a reference sequence (e.g. genome)
  - Simple to read and parse (text, tab-delimited)
  - Flexible (possibility to add custom fields)
  - Compact in file size
  - Can store paired-end information

- Reference document: http://samtools.sourceforge.net/SAM1.pdf
- **BAM** is a binary (more compact) representation of SAM

SAM/BAM (cont.)

- Structure: two sections:

  - Header: lines starting with **@**, two letters, then several key:value pairs. The keys are again two letters. Contains information about the **reference sequence (SQ)**, the libraries used (**read groups, RG**), etc…

  -Sequences: one line for each read, with the following fields (among others)

- Query (pair) name
- Reference name
- Position
- Mapping quality
- CIGAR string
- Seq and quality
- Tag:type:value fields

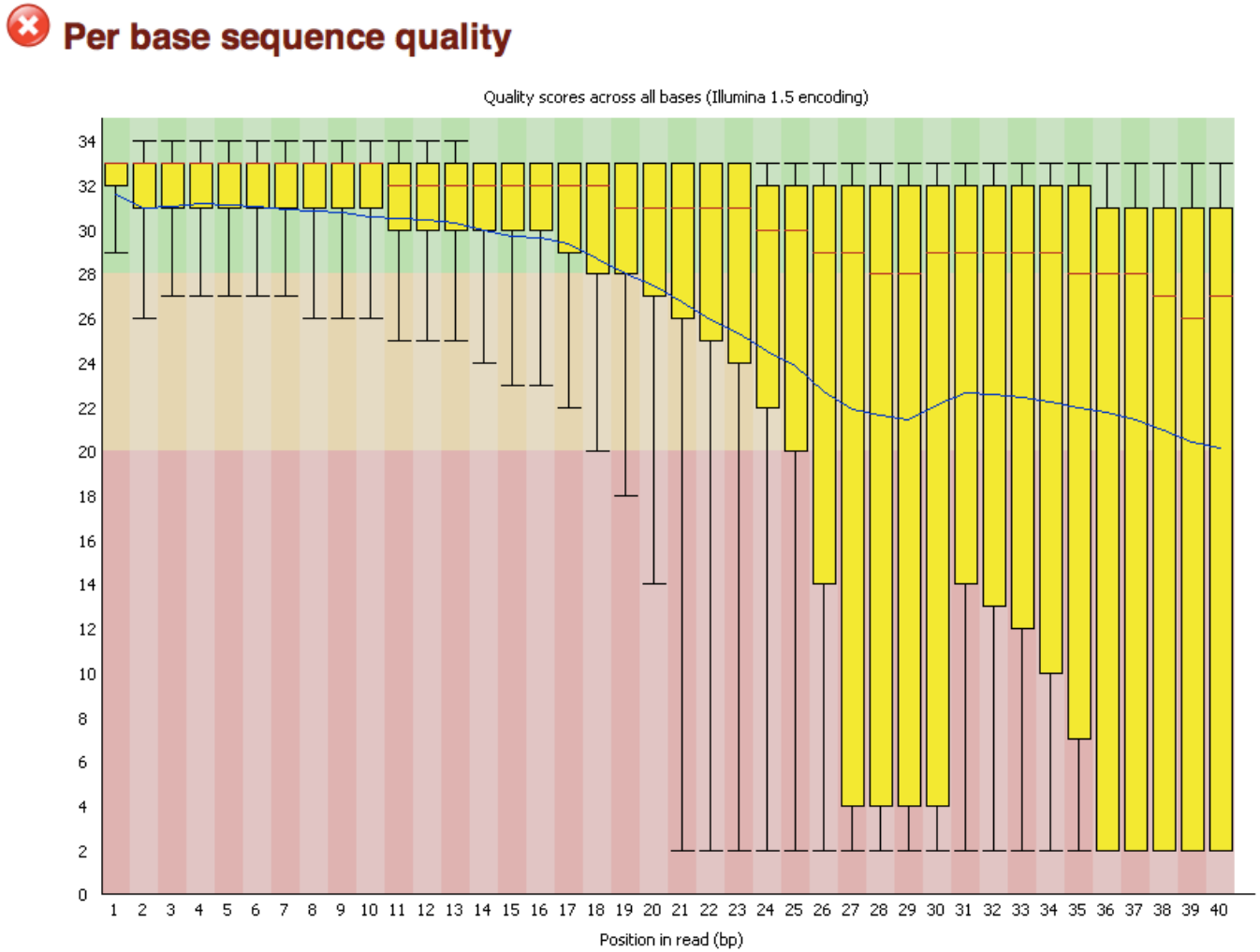
```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 1
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<
       MF:i:18 RG:Z:L2
```

sff

- Binary format provided by 454 sequencing
- Contains
  - A header with information on the run (name, key sequence, number of reads, etc.)
  - For each read:
    - Name, length of the read
    - Clipping information (quality and adaptor)
    - Numeric representation of the flowgrams (454 equivalent to chromatograms)
    - Base sequence called from flowgrams
    - Qualities

Why QC is important

- bad illumina example



Exercise: preliminary quality control of raw sequences

- number of sequences, length, average, distribution
- fasta/fastq conversion
- fastq statistics
- fasta quality chart/boxplot
- nucleotide distribution
- clipping/trimming reads

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

454 flowgram