NGS Course

Session 1

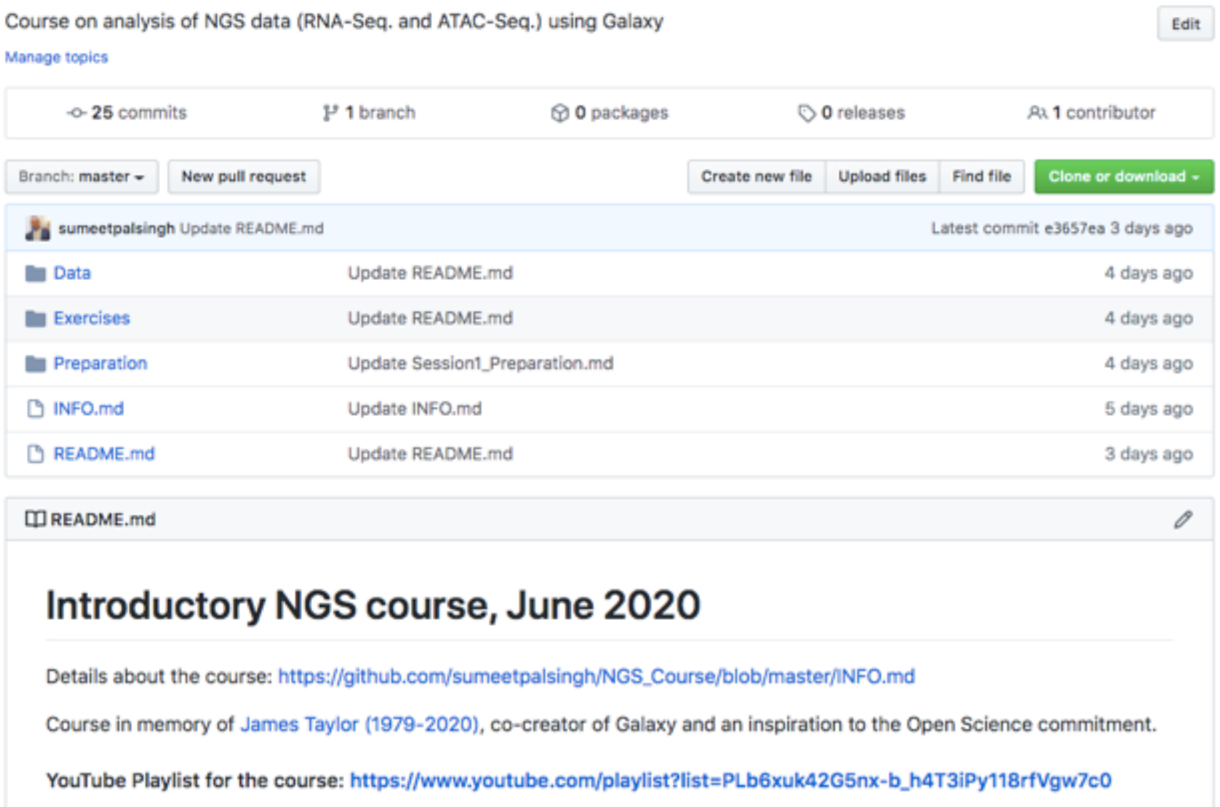
Sumeet Pal Singh and Yura Song

NGS Course

- Session 1 (05 June)
  File types in NGS (fastq, sam / bam, genome index and gtf / gff3)
  Mapping Fastq files
- Session 2 (18 June)
  Designing and saving workflow in Galaxy
  RNA-Seq. analysis
  Controlling for covariates in RNA-Seq. analysis
- Session 3 (25 June)
  ATAC-Seq. pipeline
  Interrogating ATAC-Seq. data for peaks, tf binding sites, enriched motifs

Course Repository

- https://github.com/sumeetpalsingh/NGS_Course



Course on analysis of NGS data (RNA-Seq. and ATAC-Seq.) using Galaxy

sumeetpalsingh Update README.md — Latest commit e3657ea 3 days ago

Data	Update README.md	4 days ago
Exercises	Update README.md	4 days ago
Preparation	Update Session1_Preparation.md	4 days ago
INFO.md	Update INFO.md	5 days ago
README.md	Update README.md	3 days ago

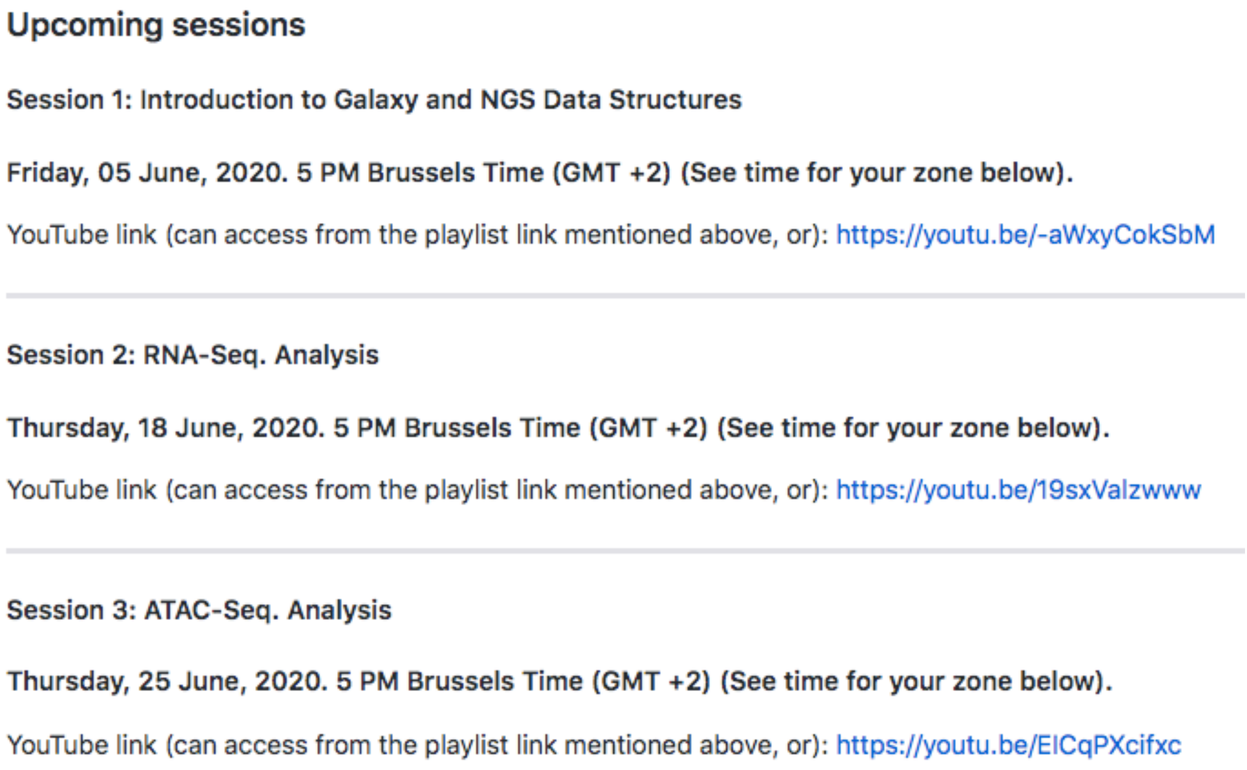
README.md

Introductory NGS course, June 2020

Details about the course: https://github.com/sumeetpalsingh/NGS_Course/blob/master/INFO.md

Course in memory of James Taylor (1979-2020), co-creator of Galaxy and an inspiration to the Open Science commitment.

YouTube Playlist for the course: https://www.youtube.com/playlist?list=PLb6xuk42G5nx-b_h4T3iPy118rfVgw7c0



Upcoming sessions

Session 1: Introduction to Galaxy and NGS Data Structures

Friday, 05 June, 2020. 5 PM Brussels Time (GMT +2) (See time for your zone below).

YouTube link (can access from the playlist link mentioned above, or): https://youtu.be/-aWxyCokSbM

Session 2: RNA-Seq. Analysis

Thursday, 18 June, 2020. 5 PM Brussels Time (GMT +2) (See time for your zone below).

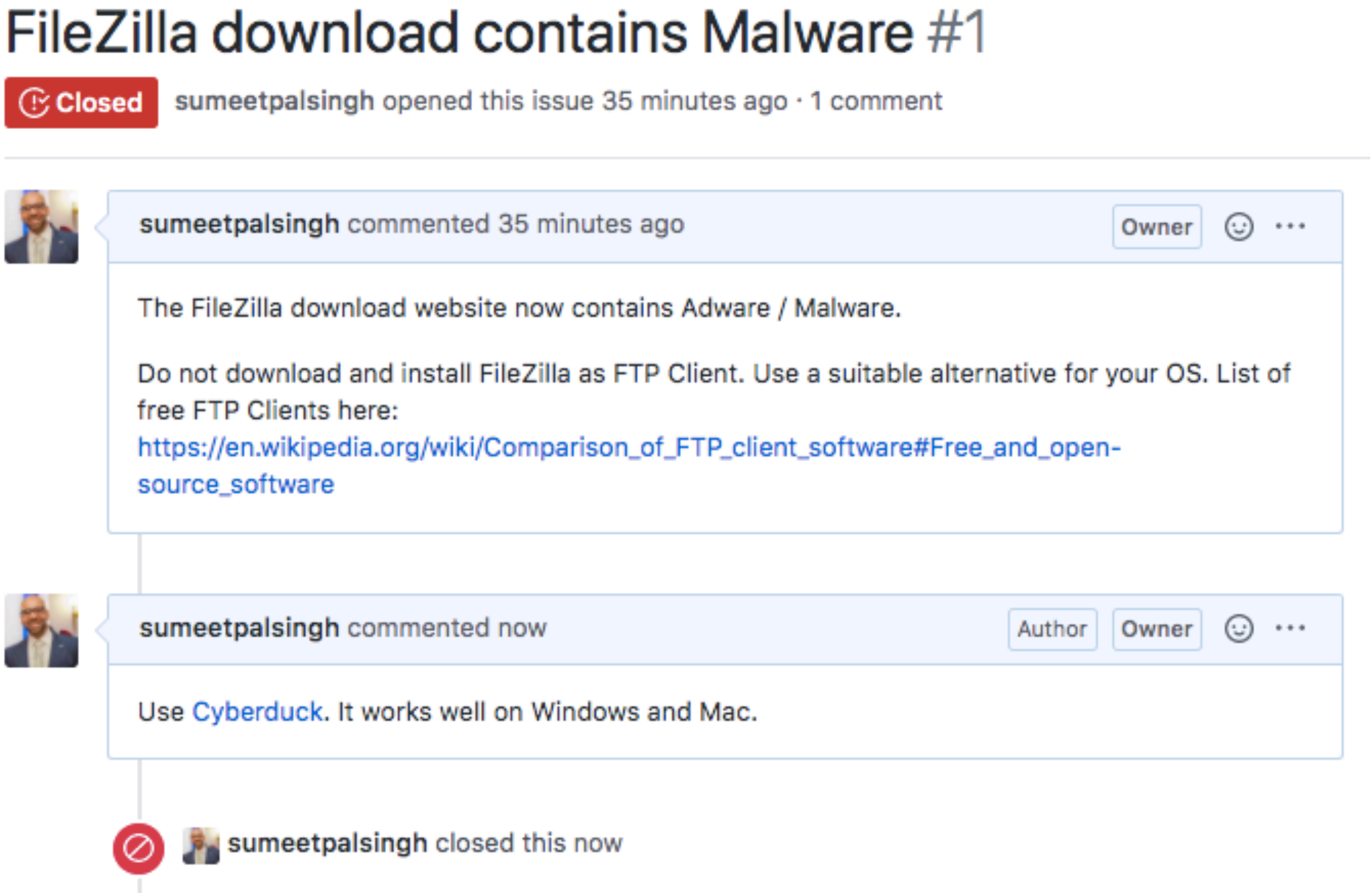
YouTube link (can access from the playlist link mentioned above, or): https://youtu.be/19sxValzwww

Session 3: ATAC-Seq. Analysis

Thursday, 25 June, 2020. 5 PM Brussels Time (GMT +2) (See time for your zone below).

YouTube link (can access from the playlist link mentioned above, or): https://youtu.be/EICqPXcifxc

Post Issues on Github



FileZilla download contains Malware #1

Closed sumeetpalsingh opened this issue 35 minutes ago · 1 comment

sumeetpalsingh commented 35 minutes ago Owner

The FileZilla download website now contains Adware / Malware.

Do not download and install FileZilla as FTP Client. Use a suitable alternative for your OS. List of free FTP Clients here: https://en.wikipedia.org/wiki/Comparison_of_FTP_client_software#Free_and_open-source_software

sumeetpalsingh commented now Author Owner

Use Cyberduck. It works well on Windows and Mac.

sumeetpalsingh closed this now

What to expect from the course

- Using Galaxy

- Allow you to go from Raw Data to Analysis

- Teach the steps and associated tools for analysis pipeline

- Independent to work with NGS-Data

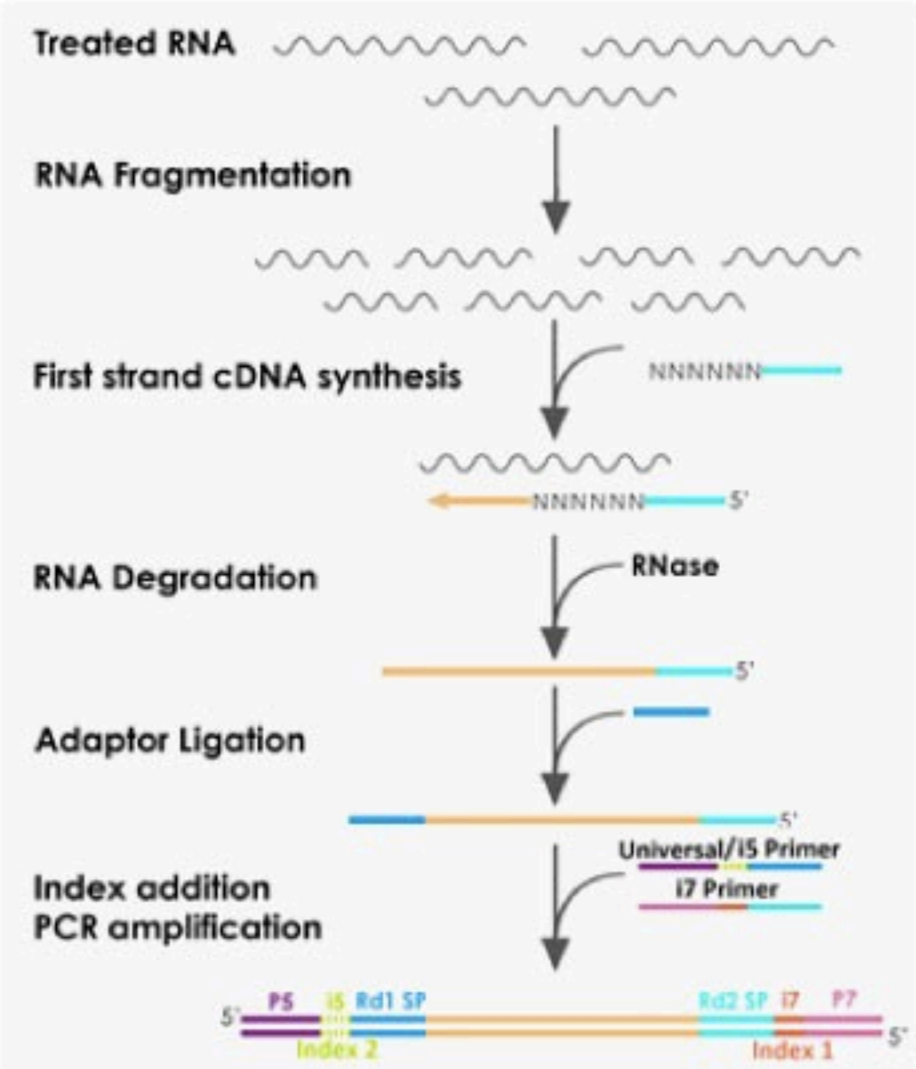
- Develop Galaxy Workflows for new analysis / pipeline

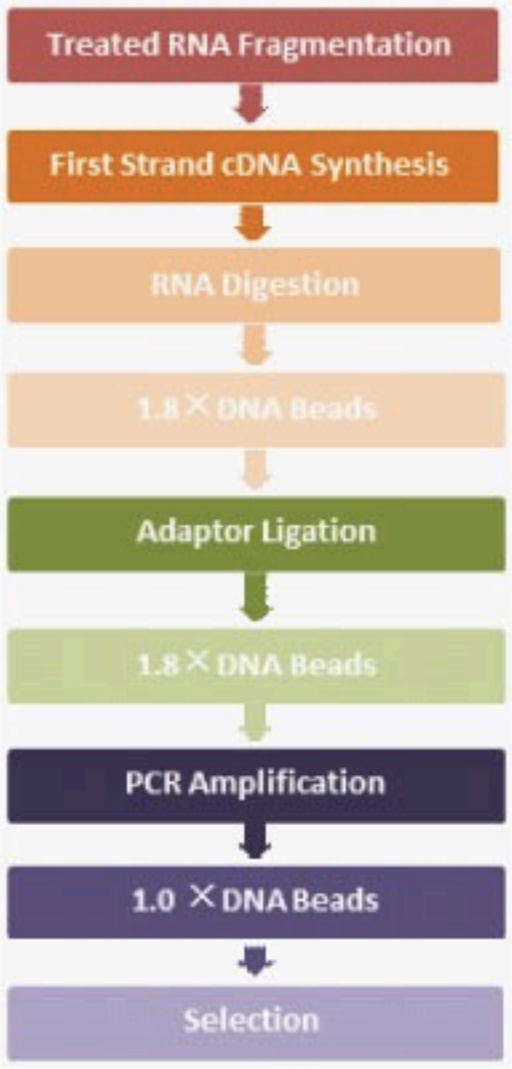
What the course is not about

- Does not cover tools / pipelines not present in Galaxy
- Does not cover executing the tools in Shell Script / HPC
- Does not cover adding features to tools that are not implemented in Galaxy
- Does not cover working with non model-organisms

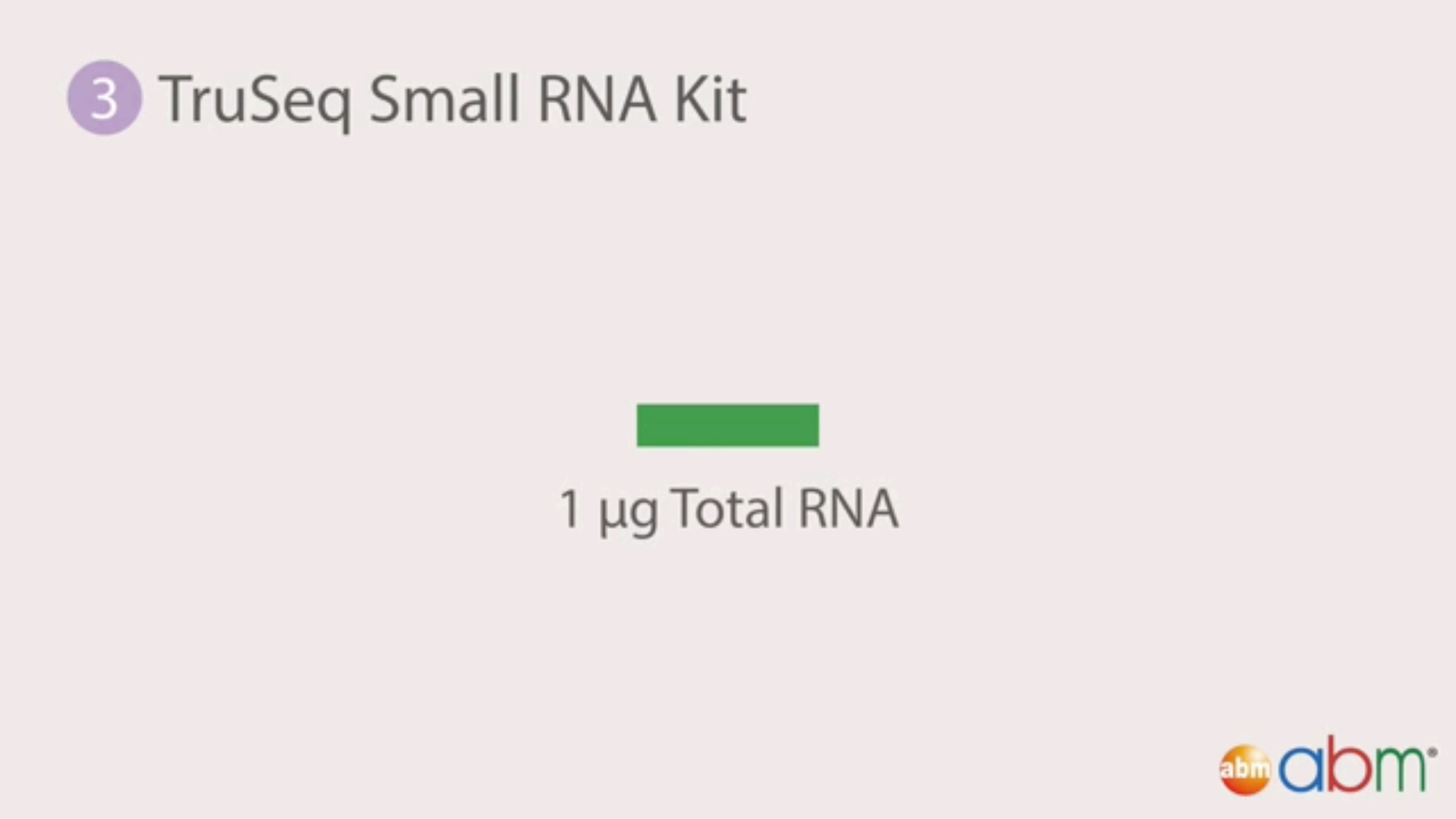
- Only covers bulk RNA-Seq. and bulk ATAC-Seq. analysis (not single-cell) made using Illumina instrument

RNA-Seq.


Treated RNA
RNA Fragmentation
First strand cDNA synthesis
NNNNNN
NNNNNN
5'
RNA Degradation
RNase
5'
Adaptor Ligation
5'
Universal/i5 Primer
i7 Primer
Index addition
PCR amplification
5'
P5
i5
Rd1 SP
Rd2 SP
i7
P7
5'
Index 2
Index 1


Treated RNA Fragmentation
First Strand cDNA Synthesis
RNA Digestion
1.8 × DNA Beads
Adaptor Ligation
1.8 × DNA Beads
PCR Amplification
1.0 × DNA Beads
Selection

RNA-Seq. Library Preparation

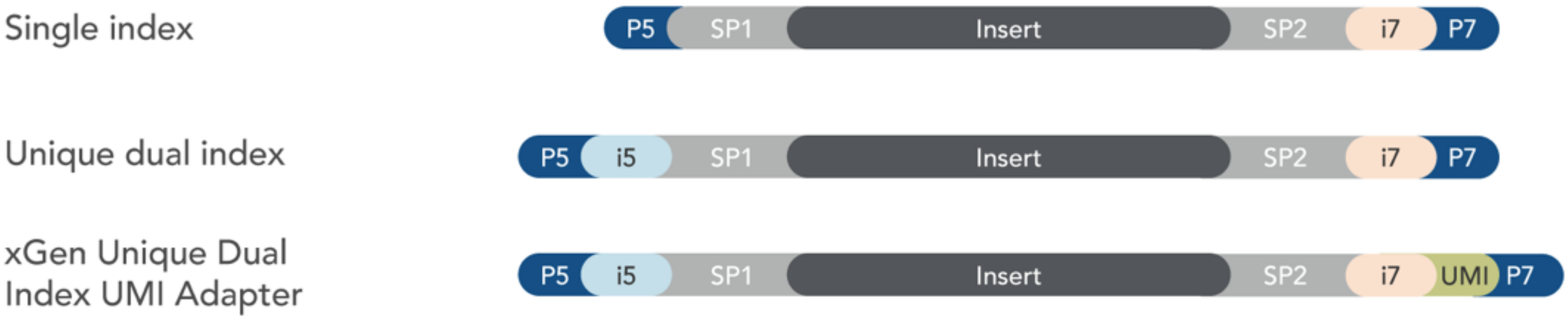


3 TruSeq Small RNA Kit

1 µg Total RNA

https://www.youtube.com/watch?v=-kTcFZxP6kM

RNA-Seq. Library

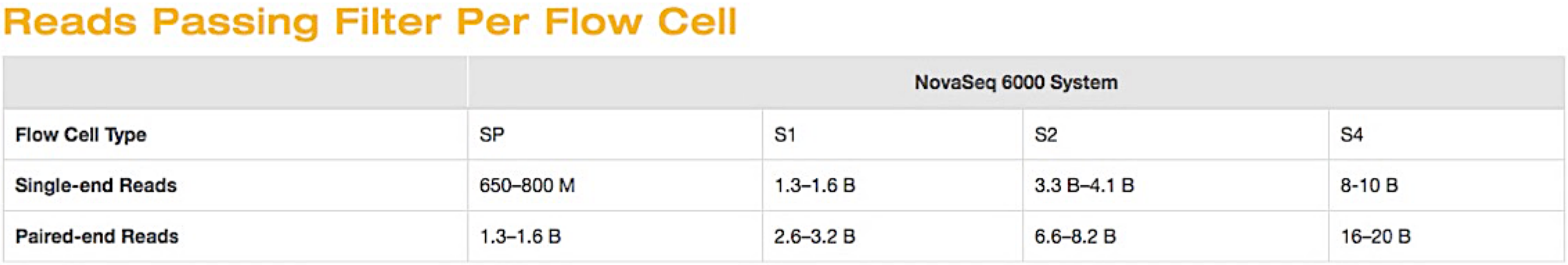


- 🟦 **Flow cell binding sequence:** Platform-specific sequences for library binding to instrument
- ⬜ **Sequencing primer sites:** Binding sites for general sequencing primers
- 🟦🟧 **Sample indexes:** Short sequences specific to a given sample library
- 🟩 **Molecular index/barcode:** Short sequence used to uniquely tag each molecule in a given sample library
- ⬛ **Insert:** Target DNA or RNA fragment from a given sample library

Flow cell-based Sequencing

https://www.youtube.com/watch?v=womKfikWlxM

Flow cell capacity



Reads Passing Filter Per Flow Cell

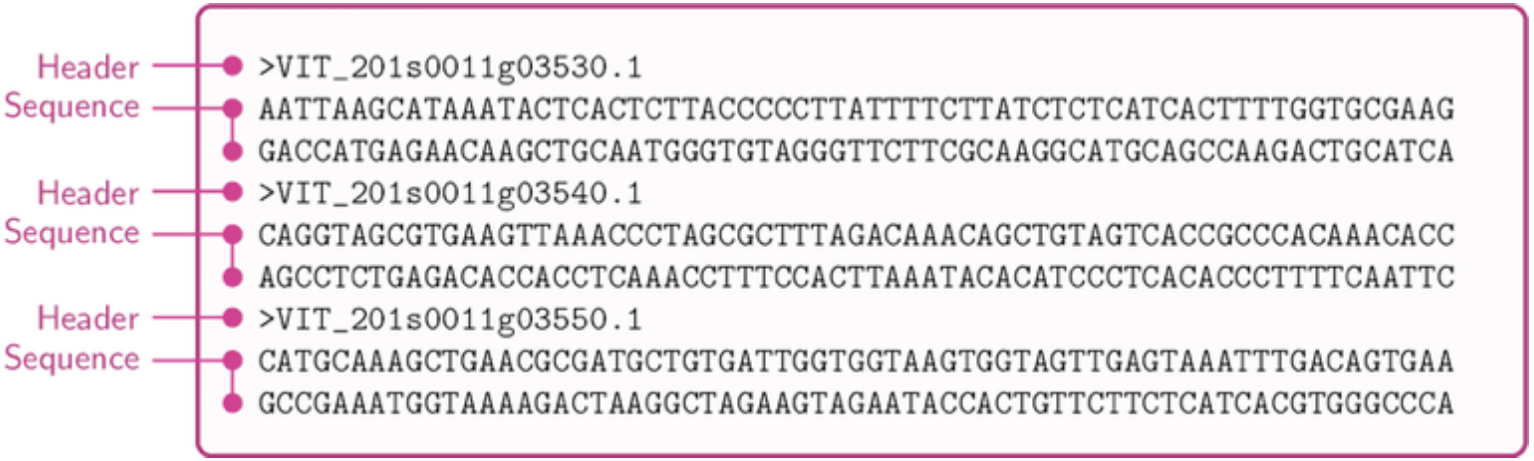
	NovaSeq 6000 System			
Flow Cell Type	SP	S1	S2	S4
Single-end Reads	650–800 M	1.3–1.6 B	3.3 B–4.1 B	8-10 B
Paired-end Reads	1.3–1.6 B	2.6–3.2 B	6.6–8.2 B	16–20 B

For a regular RNA / ATAC-Seq. sample: 10 – 50 M

Raw Data

- Fastq format: Fasta format with quality scores

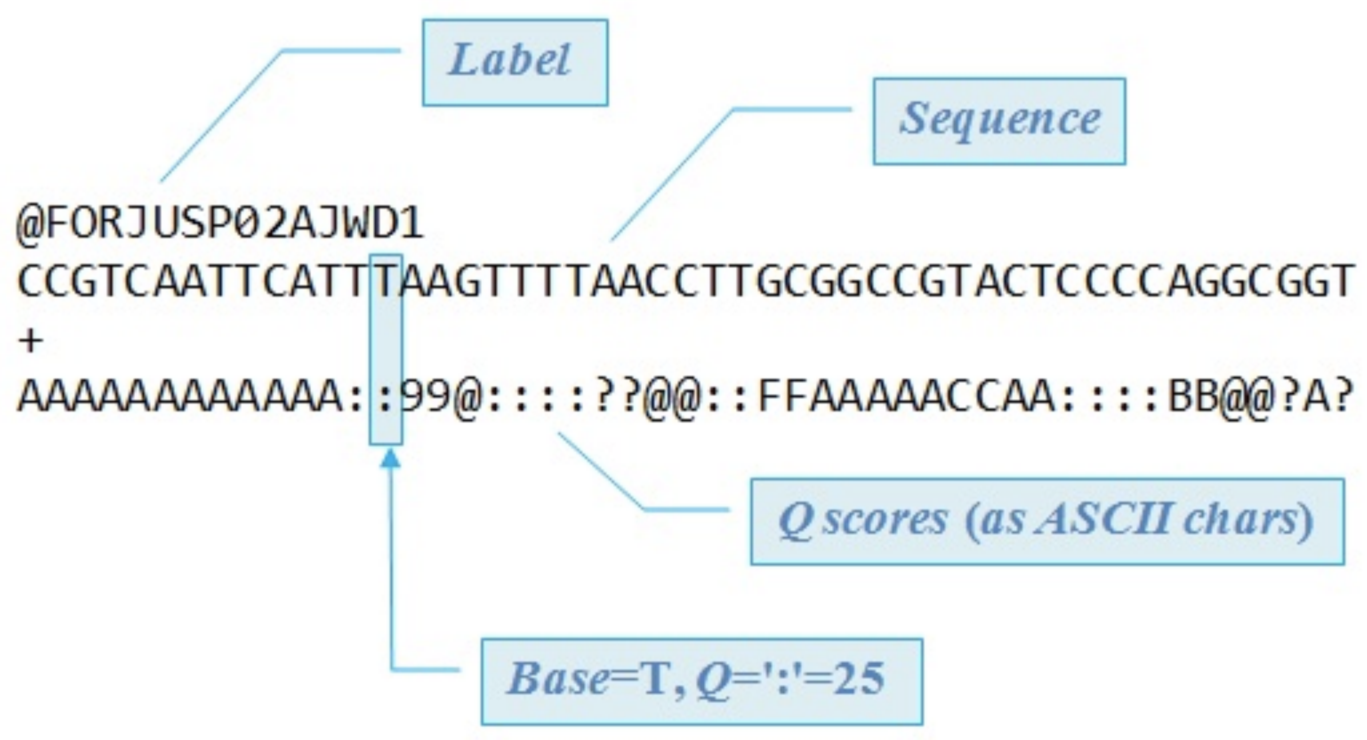
Fasta



```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

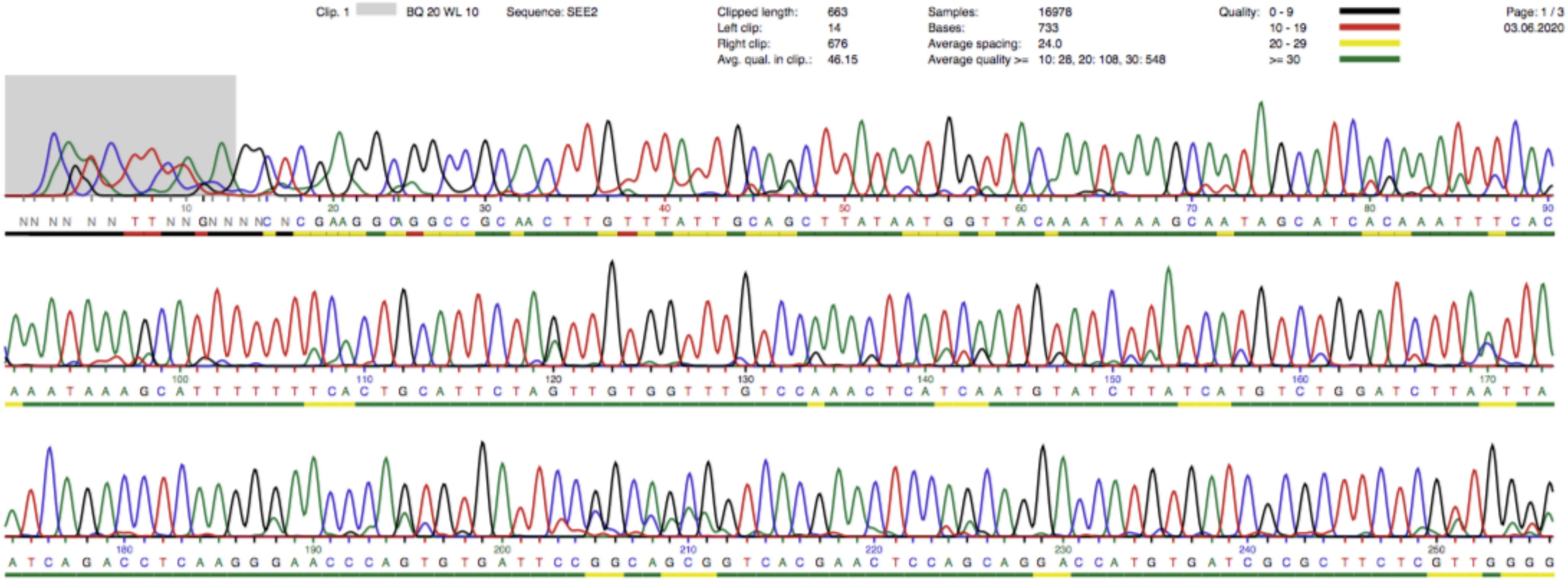
Raw Data

- Fastq format: Fasta format with quality scores



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Quality Score



Clip. 1 ▨ BQ 20 WL 10 Sequence: SEE2

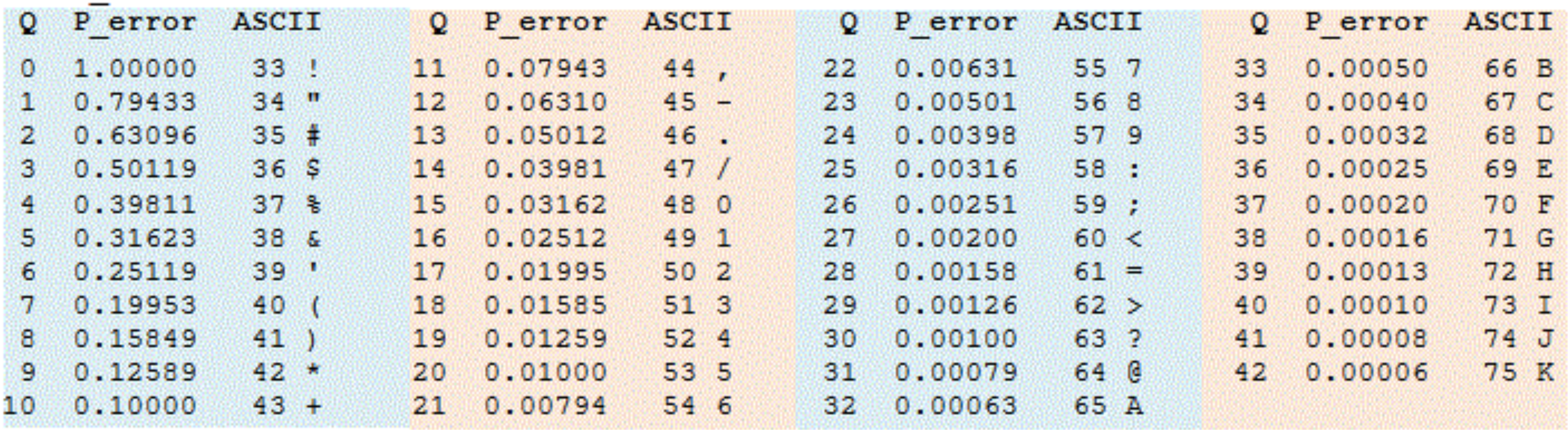
Clipped length: 663
Left clip: 14
Right clip: 676
Avg. qual. in clip.: 46.15

Samples: 16978
Bases: 733
Average spacing: 24.0
Average quality >= 10: 28, 20: 108, 30: 548

Quality: 0 - 9 ■
10 - 19 🟥
20 - 29 🟨
>= 30 🟩

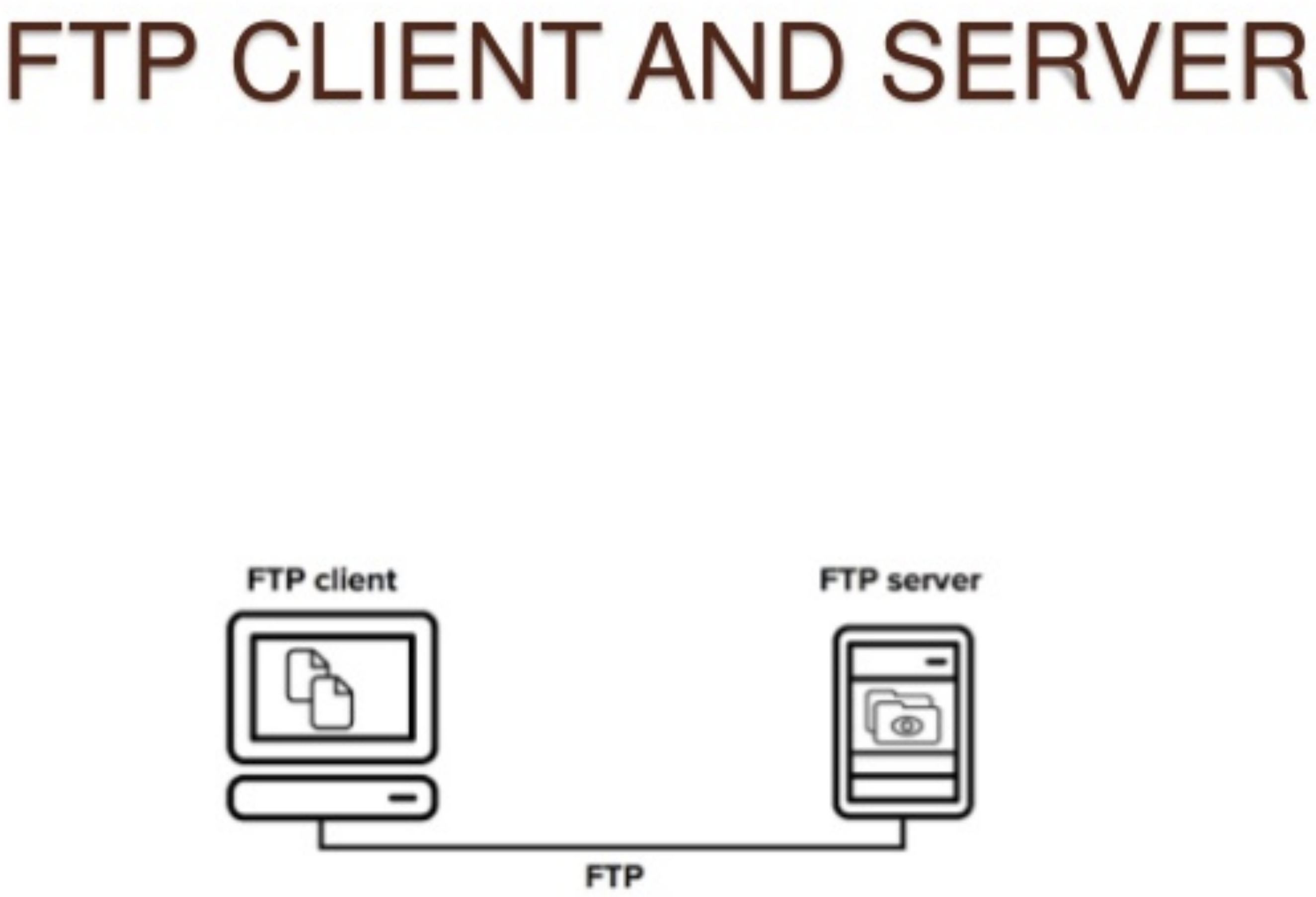
Page: 1 / 3
03.06.2020

Fastq Quality Scores



Q	P_error	ASCII
0	1.00000	33 !
1	0.79433	34 "
2	0.63096	35 #
3	0.50119	36 $
4	0.39811	37 %
5	0.31623	38 &
6	0.25119	39 '
7	0.19953	40 (
8	0.15849	41)
9	0.12589	42 *
10	0.10000	43 +
11	0.07943	44 ,
12	0.06310	45 -
13	0.05012	46 .
14	0.03981	47 /
15	0.03162	48 0
16	0.02512	49 1
17	0.01995	50 2
18	0.01585	51 3
19	0.01259	52 4
20	0.01000	53 5
21	0.00794	54 6
22	0.00631	55 7
23	0.00501	56 8
24	0.00398	57 9
25	0.00316	58 :
26	0.00251	59 ;
27	0.00200	60 <
28	0.00158	61 =
29	0.00126	62 >
30	0.00100	63 ?
31	0.00079	64 @
32	0.00063	65 A
33	0.00050	66 B
34	0.00040	67 C
35	0.00032	68 D
36	0.00025	69 E
37	0.00020	70 F
38	0.00016	71 G
39	0.00013	72 H
40	0.00010	73 I
41	0.00008	74 J
42	0.00006	75 K

Using FTP Client to Transfer Files to FTP Server



FTP CLIENT AND SERVER

Connect to FTP Server

- For galaxy.org: https://galaxyproject.org/ftp-upload/
  FTP Server: usegalaxy.org

- For galaxy.eu: https://galaxyproject.eu/ftp/
  FTP Server: galaxy.uni-freiburg.de

- For galaxy.au: https://usegalaxy-au.github.io/posts/2019/03/18/new-ftp-upload-url/
  FTP Server: usegalaxy.org.au

Genome Index

- Genome Index <-> Genome
- Dictionary <-> Words

—see also

def·i·nite·ly /ˈdefɪ... certainly: Max knew that he with no cha

being wrong; certainly: Max knew that he had def

been wrong about Diana. \ "It's not worth that muc

"No, definitely not!" —see OF COURSE (USAGE)

(USAGE)

def·i·ni·tion /ˌdefɪˈnɪʃən/ n 1 [C] a phrase o

that says exactly what a word, phrase, or ide

definition in a dictionary \ [+ of] No one has

with a satisfactory definition of terrorism.

nition if something has a particular qu

nition, it must have that quality because al

type have it: A message that cannot be see

definition, not effective. 3 [U] the degre

thing such as a picture, sound etc is c

nition The photograph lacks definitio

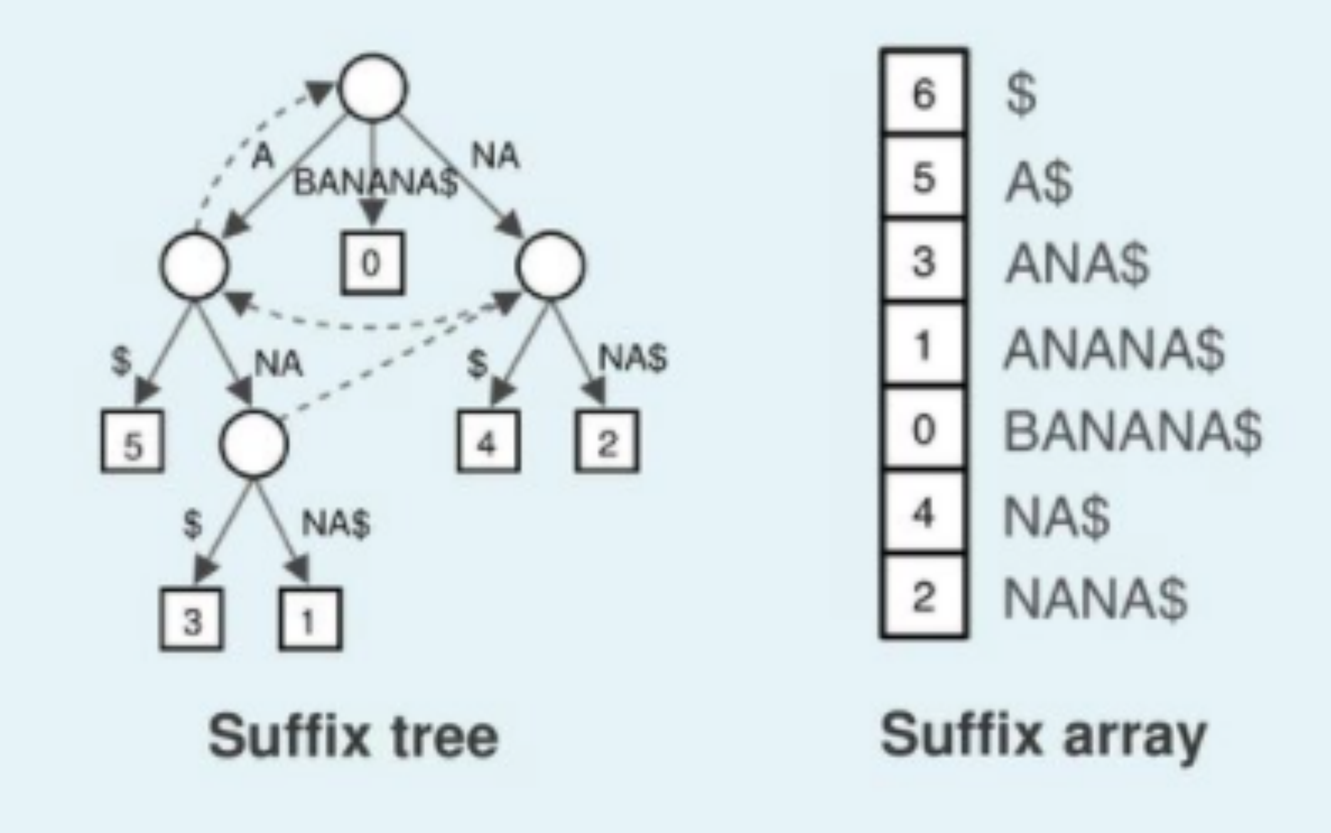
de·fin·i·tive /dɪˈfɪnətɪv/ adj 1 [usu

finitive book, study of something

ever produced and cann

finitive book on t

Genome Index



Suffix tree

6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

Suffix array

Genome Index is unique to every mapping tool

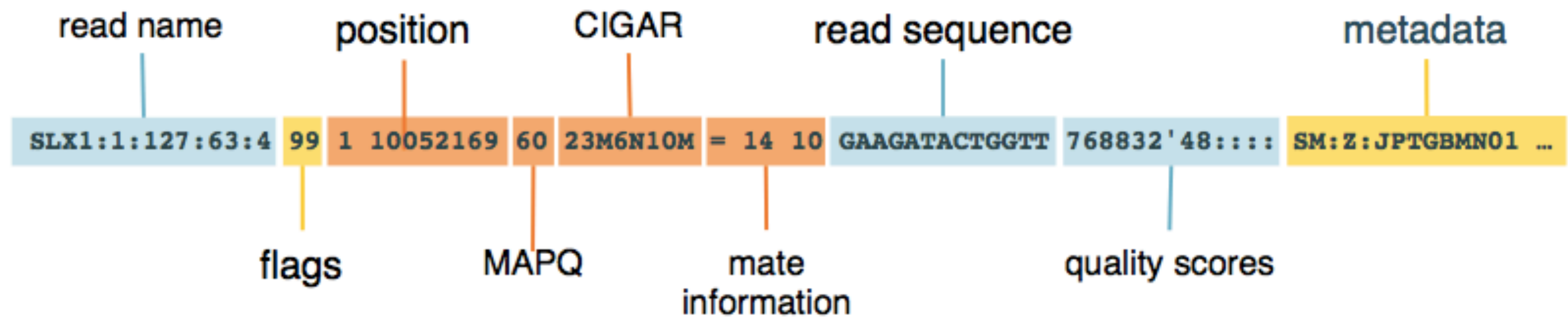
Mapping Tools

- Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- BWA: http://bio-bwa.sourceforge.net/
- Hisat2: https://ccb.jhu.edu/software/hisat2/manual.shtml (Normal Laptop)
- STAR: https://github.com/alexdobin/STAR (High RAM requirements: Human ~32 Gb)

SAM file format (Alignment Formats)

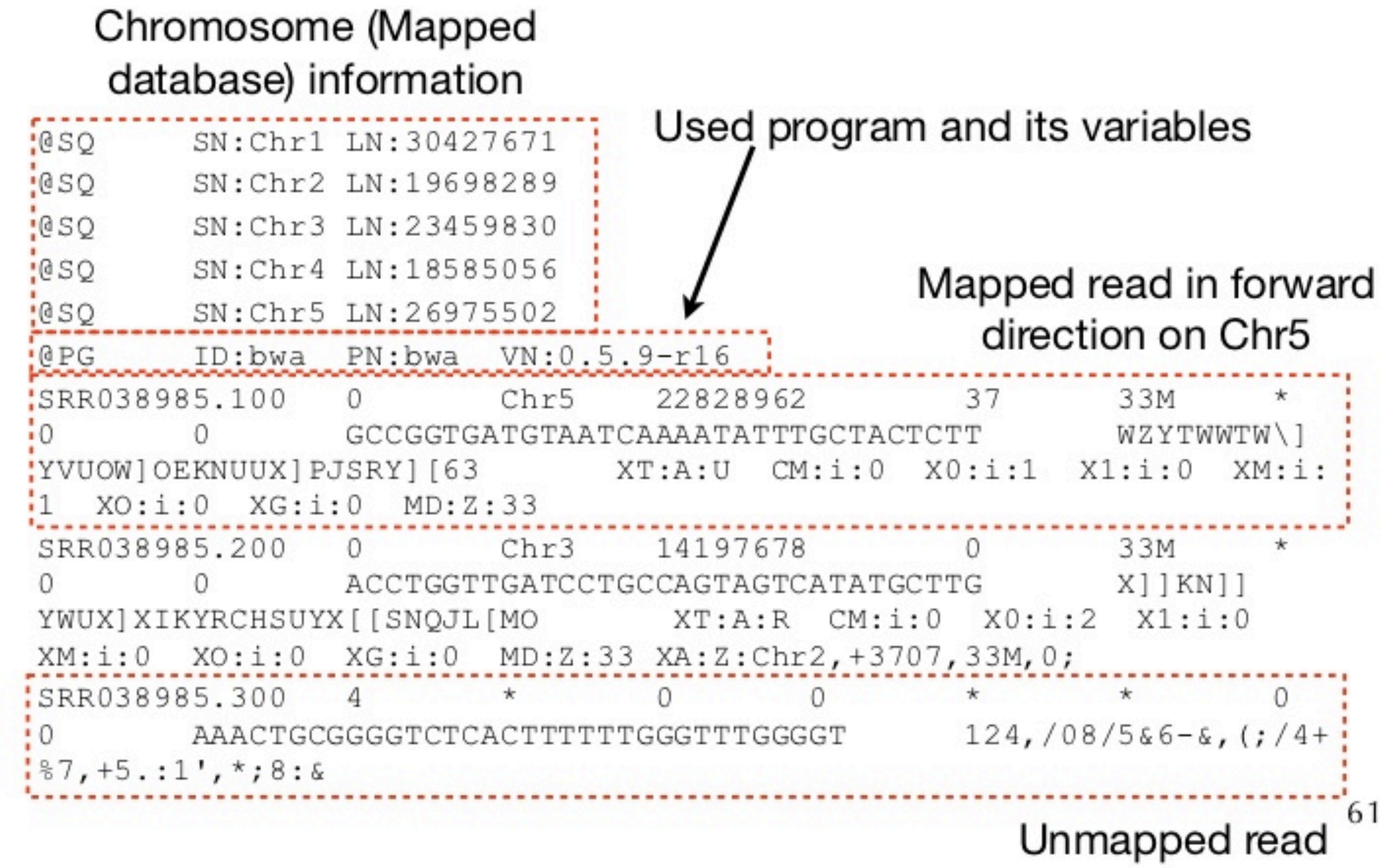
- SAM – Sequence Alignment / Map Format
- Plain Text (Human Readable)
- Contains
  Quality Scores, Sequence info (Fastq) +
  Alignment Info + MetaData

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …
```

SAM Format Example



Chromosome (Mapped database) information

Used program and its variables

Mapped read in forward direction on Chr5

```
@SQ      SN:Chr1 LN:30427671
@SQ      SN:Chr2 LN:19698289
@SQ      SN:Chr3 LN:23459830
@SQ      SN:Chr4 LN:18585056
@SQ      SN:Chr5 LN:26975502
@PG      ID:bwa  PN:bwa  VN:0.5.9-r16
SRR038985.100   0       Chr5    22828962        37      33M     *
0       0       GCCGGTGATGTAATCAAAATATTTGCTACTCTT       WZYTWWTW\]
YVUOW]OEKNUUX]PJSRY][63       XT:A:U  CM:i:0  X0:i:1  X1:i:0  XM:i:
1  XO:i:0  XG:i:0  MD:Z:33
SRR038985.200   0       Chr3    14197678        0       33M     *
0       0       ACCTGGTTGATCCTGCCAGTAGTCATATGCTTG       X]]KN]]
YWUX]XIKYRCHSUYX[[SNQJL[MO       XT:A:R  CM:i:0  X0:i:2  X1:i:0
XM:i:0  XO:i:0  XG:i:0  MD:Z:33 XA:Z:Chr2,+3707,33M,0;
SRR038985.300   4       *       0       0       *       *       0
0       AAACTGCGGGGTCTCACTTTTTTGGGTTTGGGGT      124,/08/5&6-&,(;/4+
%7,+5.:1',*;8:&
```

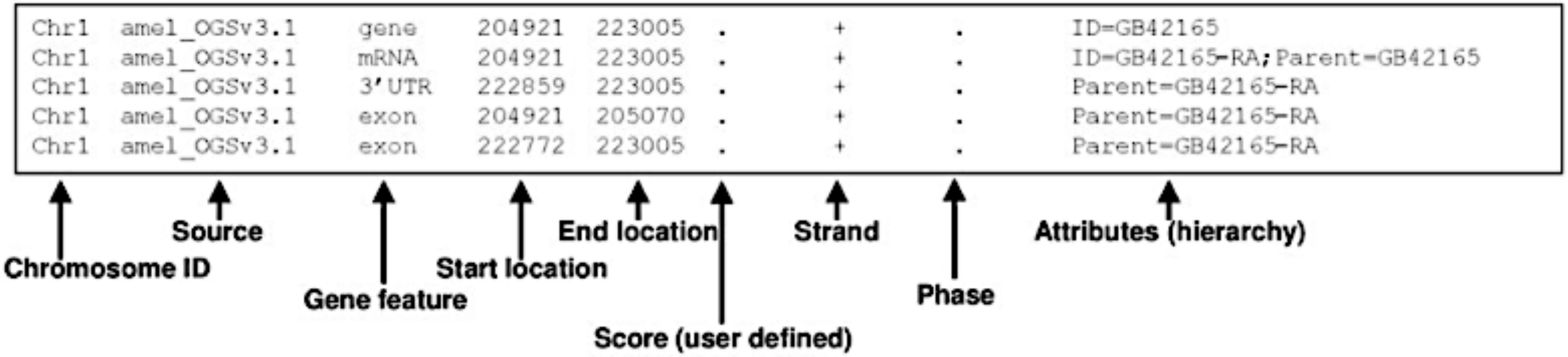
61

Unmapped read

BAM File Format (Alignment Format)

- BAM: BZGF compressed SAM Format
- Not human readable
- ~ 1 / 5 size of SAM

gtf / gff3 files

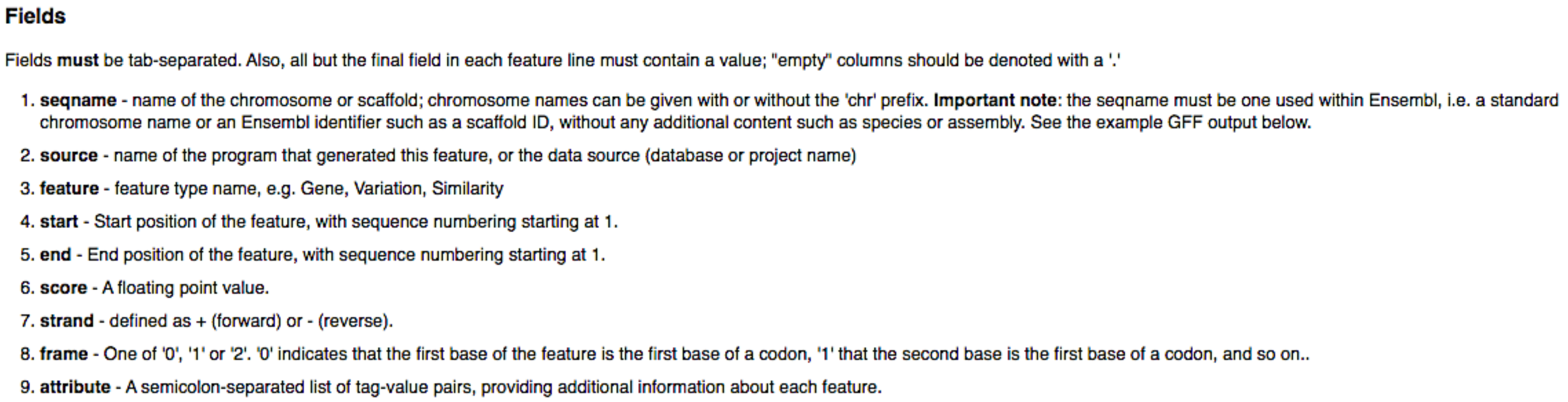


```
Chr1  amel_OGSv3.1   gene    204921  223005  .    +    .    ID=GB42165
Chr1  amel_OGSv3.1   mRNA    204921  223005  .    +    .    ID=GB42165-RA;Parent=GB42165
Chr1  amel_OGSv3.1   3'UTR   222859  223005  .    +    .    Parent=GB42165-RA
Chr1  amel_OGSv3.1   exon    204921  205070  .    +    .    Parent=GB42165-RA
Chr1  amel_OGSv3.1   exon    222772  223005  .    +    .    Parent=GB42165-RA
```

- Be aware of the format being used and its compatibility!
- Some tools will only work with gtf / gff3
- Ensembl GTF is NOT the same as UCSC GTF (even for the same assembly)

GTF / GFF3 Fields

- https://www.ensembl.org/info/website/upload/gff.html

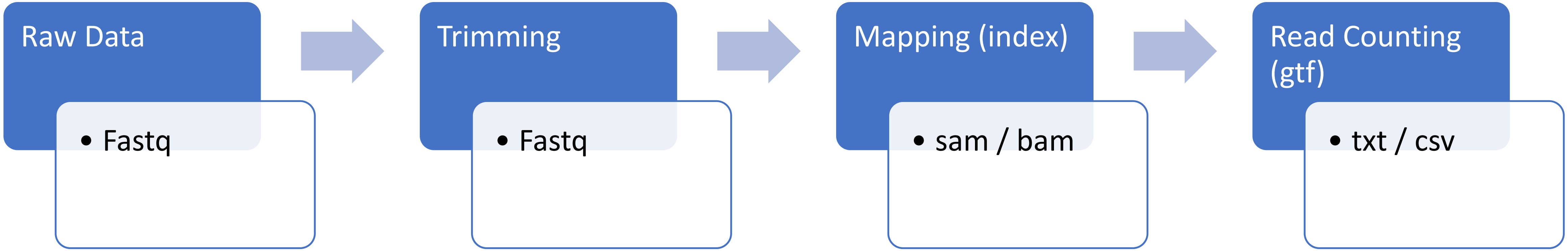


Fields

Fields **must** be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Summary



Raw Data

- Fastq

Trimming

- Fastq

Mapping (index)

- sam / bam

Read Counting (gtf)

- txt / csv