



CEPLAS

Cluster of Excellence on Plant Sciences

From fastq to quantified transcripts

CEPLAS RNA-Seq Workshop 2022





Demultiplexing

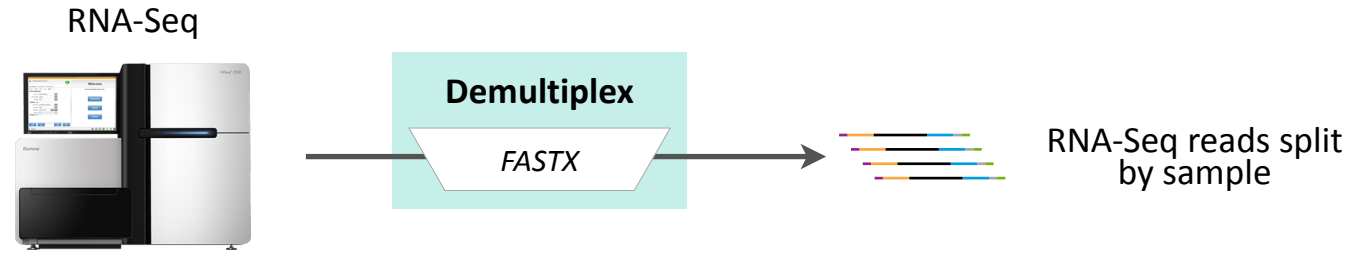
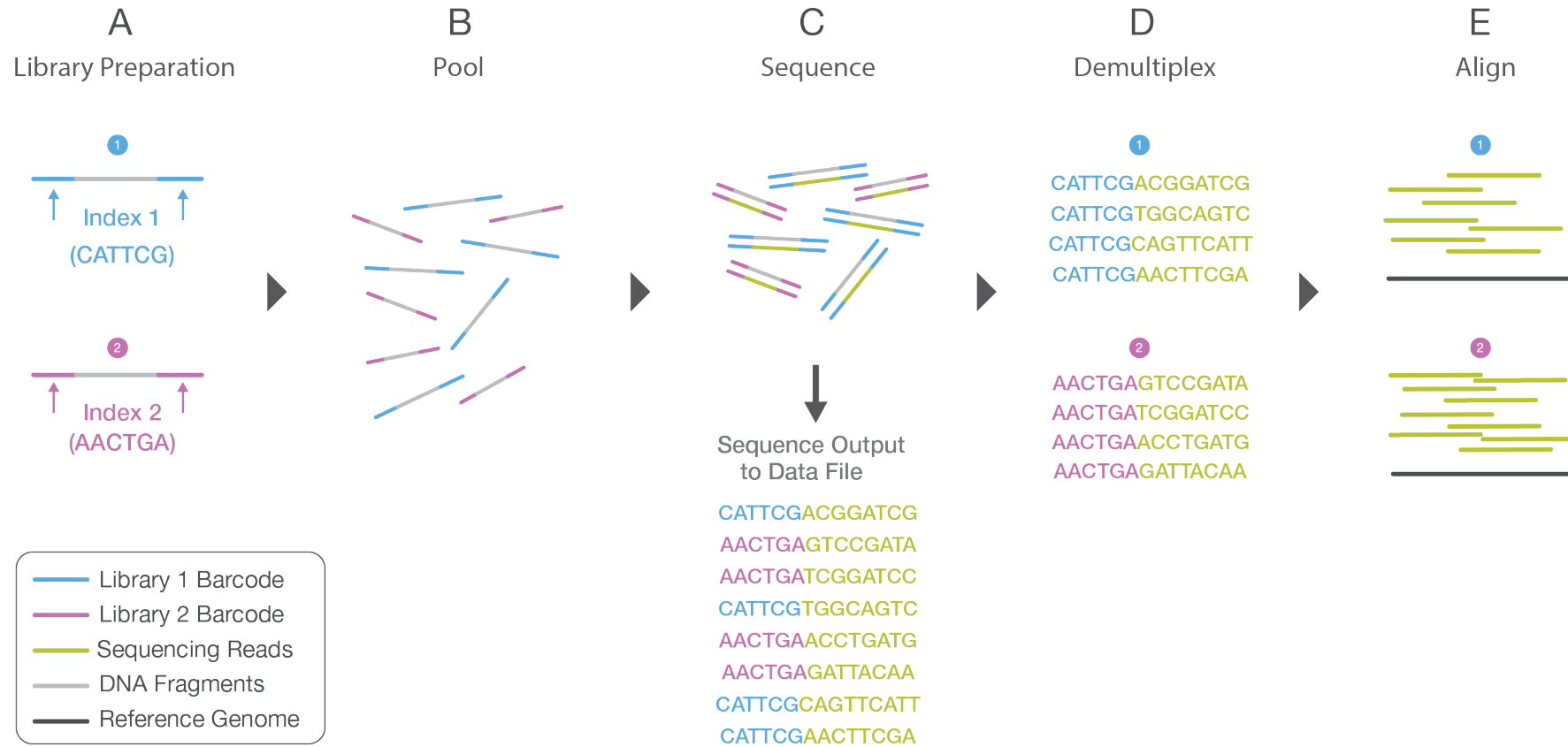


Photo: www.illumina.com





Demultiplexing





Demultiplexing Stats

Lane	Sample	Barcode sequence	PF Clusters	% of the lane	Yield (Mbases)	% PF Clusters	% >= Q30 bases	Mean Quality Score
7	311	GTCCGC	34,383,904	8.6	5,192	100	93.92	38.91
7	312	GTGAAA	29,615,024	7.41	4,472	100	94.42	39.04
7	313	GTGGCC	33,503,347	8.38	5,059	100	95.98	39.47
7	314	GTTTCG	31,257,959	7.82	4,720	100	95.13	39.24
7	315	CGTACG	30,104,699	7.53	4,546	100	96.17	39.51
7	316	GAGTGG	38,777,553	9.7	5,855	100	88.83	37.53
7	317	ACTGAT	37,039,274	9.27	5,593	100	95.71	39.4
7	318	ATTCCT	35,883,841	8.98	5,418	100	87.71	37.24
7	319	ATCACG	24,474,812	6.12	3,696	100	94.61	39.09
7	320	CGATGT	35,103,812	8.78	5,301	100	94.99	39.2
7	321	TTAGGC	27,689,677	6.93	4,181	100	96.18	39.52
7	322	TGACCA	26,135,742	6.54	3,946	100	95.34	39.3
7	Undetermined	unknown	15,708,905	3.93	2,372	15.86	93.87	38.84
8	323	ACAGTG	32,627,470	7.99	4,927	100	96.32	39.57
8	324	GCCAAT	36,955,093	9.05	5,580	100	96.28	39.56
8	325	CAGATC	31,784,966	7.78	4,800	100	95.69	39.4
8	326	ACTTGA	29,892,017	7.32	4,514	100	95.58	39.37
8	327	GATCAG	30,164,205	7.38	4,555	100	95.81	39.44
8	328	TAGCTT	32,620,324	7.99	4,926	100	94.82	39.17
8	329	GGCTAC	32,216,852	7.89	4,865	100	96.34	39.58
8	330	CTTGTA	33,811,657	8.28	5,106	100	94.96	39.2
8	331	AGTCAA	31,813,716	7.79	4,804	100	93.87	38.91
8	332	AGTTCC	35,091,218	8.59	5,299	100	96.45	39.61
8	333	ATGTCA	31,647,609	7.75	4,779	100	95.5	39.36
8	334	CCGTCC	29,550,632	7.23	4,462	100	95.34	39.31
8	Undetermined	unknown	20,307,719	4.97	3,066	21.42	94.64	39.03





“Your data is ready”

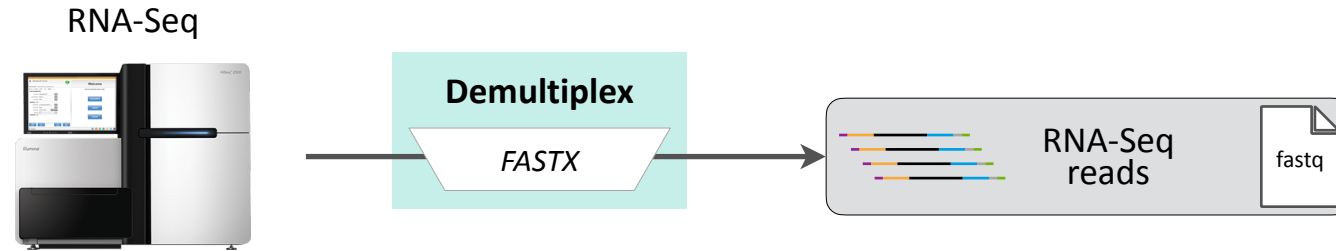


Photo: www.illumina.com





The .fastq format

```
@HWI-ST737:111:8164GABXX:1:1101:1367:2206 1:N:0:CGATGT
CGGTAGATAGCAGATGCAGTCAAGTAACGGCCATGACGAGGGTCAGCAGCACACATCATGTTCTTTGCGTCCCACATTTGTTGTGTGAGGTCTGGAACAGT
+
CCCCFFFFHHHGHGIIJJJHIIJJJIJIGIHIJBGGGIGGGI;FHDCECC(..@D3=;3=7?>;C>),,3>@C@8@D;@CA?(+2:C(99>9?#####
@HWI-ST737:111:8164GABXX:1:1101:1486:2206 1:N:0:CGATGT
CCCCAAATAGAACAAATATCCCTTCTAAAAATCCCATTTTAAATGGTGGGTTTCGGAAGATTTGCAGCAATAAACACAAATTTGTTGCTAGATTTACATATCT
+
CCCCFFFFHHHHHJJJJJJJJJJJJJJJJJJIIIGIJJJJJJJJIIJIFHIJ?DFHJBCGG>GHC:CH3(77AEE;?(;?=C>>C;CC=;;AAC:AA#####
@HWI-ST737:111:8164GABXX:1:1101:1419:2214 1:N:0:CGATGT
GGATTTTTTCACCTATCTTGCAGTTTGAACAGGACCCTGTTTCAGATTCTTGATGCTTTGTTGCCATTGTATTTGAACAGTCAGATCTTGAGGTCTTTACAGG
+
@C@FDDFFHHHHHJJJJJJJJJJJJJJJJJJGIIJJJIHGGIGJJJJJJGJIIJJGGGGIGIIGHIGIJ<DGH@)@CHE=DDHHGHAE>CEHECCF@;ACEEA####
@HWI-ST737:111:8164GABXX:1:1101:1255:2232 1:N:0:CGATGT
CACTGGATTTCACTGTCCAATTCTCGAATTTAAAGGTTTCACTTTTCAACCCTAAACTTTTCAGCTATTTTCTTTCCAATCTCCTTAGCAATCACATGTATA
+
@?@FFEADBDDHIGDBFH@FHFHGHIFGGHH@FGGHFGGDFHIIGIIBHAFHIIGHGGHGFGI4BCHIII>@@AG@;@ECEECEHHD@CDFEA>CCC>3>;
@HWI-ST737:111:8164GABXX:1:1101:1416:2232 1:N:0:CGATGT
CCGAAGGGCATCAGCATAGGAGCTCATATCGGTTAAGATCACAAGGACATGTTTCCCACATTCATATGCCAAATATTCTGCTGTGGTGAGAGCAATACGAG
+
<<?DDDDFGFDHHDHIIIIIGAHGHIIEIFEHIF:GHHB>BDFH3F?FDHGGEEH;=@GDDCAEIGDA3?EAEFB?C?DDCCCC@CB>A-:??>9:5:?89
@HWI-ST737:111:8164GABXX:1:1101:1357:2236 1:N:0:CGATGT
CGACGAAATCACATCCTCAGGTACGCTTCAACTACACGTCTGCAACTTCCTCAACGTGTTAATCGAGAGCAAACGCATCGATATGATCAAGGAGATCATTA
+
@@?BDD>DHHDBHDDHH?FBFHGEHGDDHGIIGIFBFFAGHGDEG3B8=FHC>;DEHACEH756?2??CCCB?:>?@BCBBBC>:(>:4<8<>@3>:3
@HWI-ST737:111:8164GABXX:1:1101:1390:2249 1:Y:0:CGATGT
CTCATCTTTCACTGTTAAAGCCGAGGGTGTGAGCAAGAAGCTCTGCACCCTAGTTCCAAAAGACTCAACTGATGATTCGTTTGACGTACACCCTTGATGCAA
+
; <<:??>1>><=9?@?<5=@#####
```



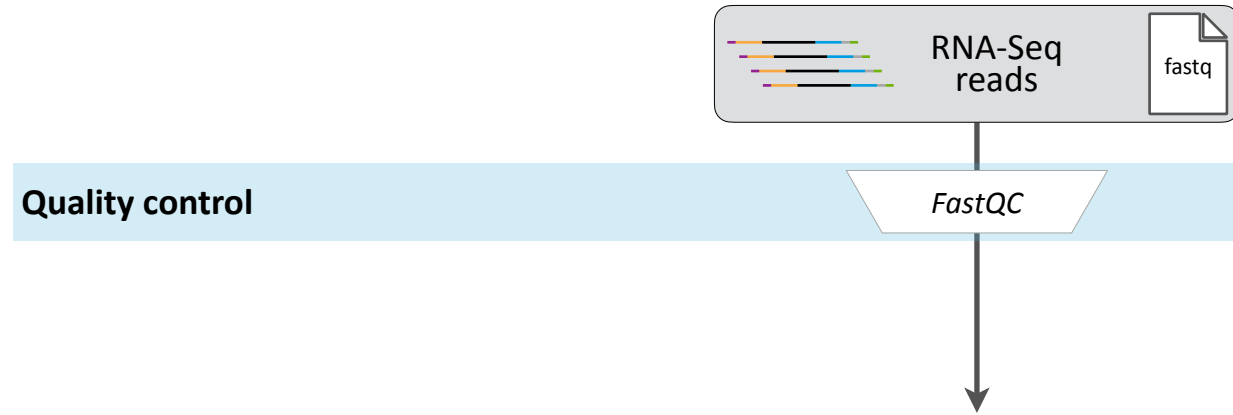
The .fastq format

- File name convention
 - <sample name>_<barcode sequence>_L<lane (0-padded to 3 digits)>_R<read number>_<set number (0-padded to 3 digits)>.fastq.gz
 - e.g. NA10831_ATCACG_L002_R1_001.fastq.gz
- The sequence identifier
 - @<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos><read>:<is filtered>:<control number>:<index sequence>
 - e.g. @HWI-ST737:111:8164GABXX:1:1101:1367:2206 1:N:0:CGATGT





Quality control – FastQC



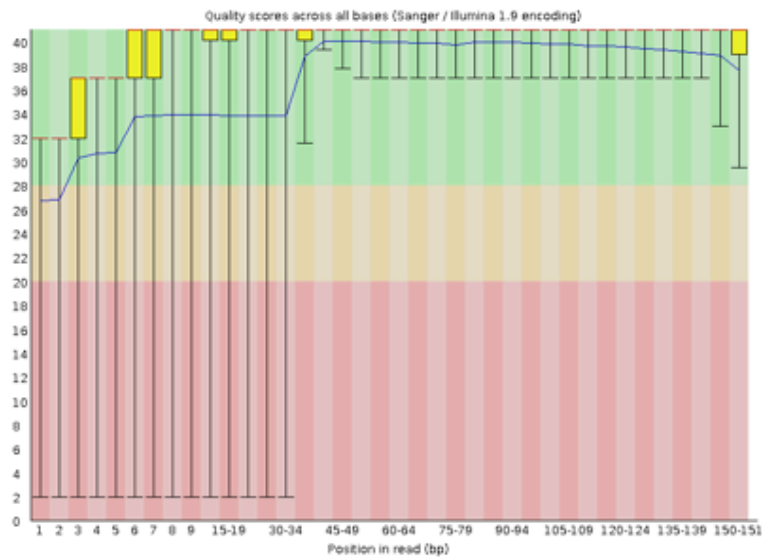


Quality control – FastQC

✓ Basic Statistics

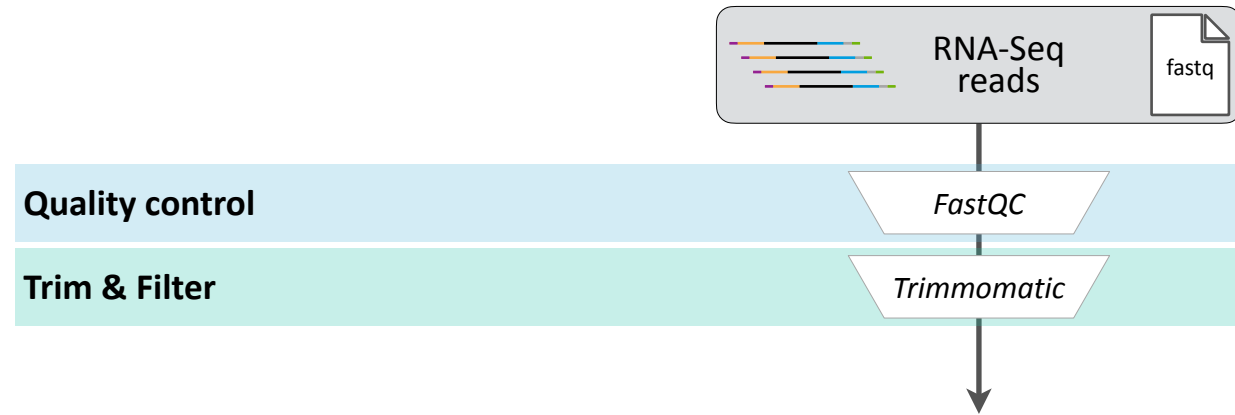
Measure	Value
Filename	316_S78_L007_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	38777553
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	49

✓ Per base sequence quality





Read trimming





Read trimming

- Trimmomatic removes
 - PCR primers
 - Adapter sequences
 - Low-quality bases

P5 Rd1 SP cDNA Insert Rd2 SP Index P7





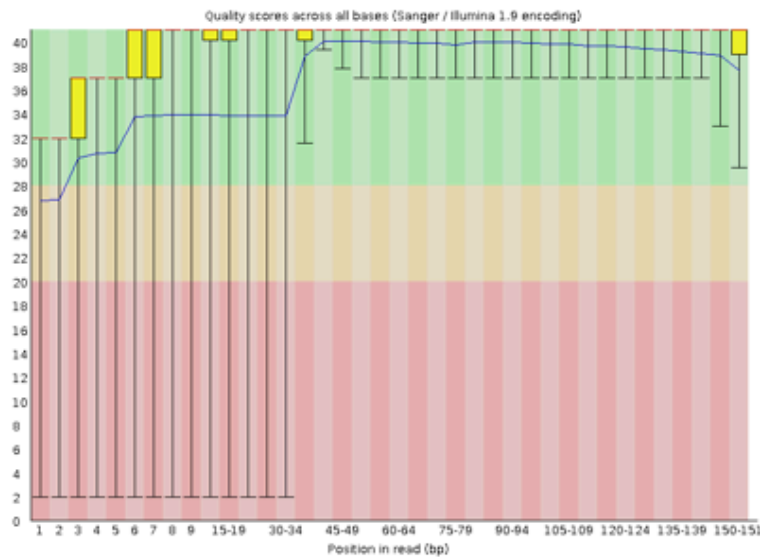
Quality control – FastQC

Before trimming

Basic Statistics

Measure	Value
Filename	316_S78_L007_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	38777553
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	49

Per base sequence quality

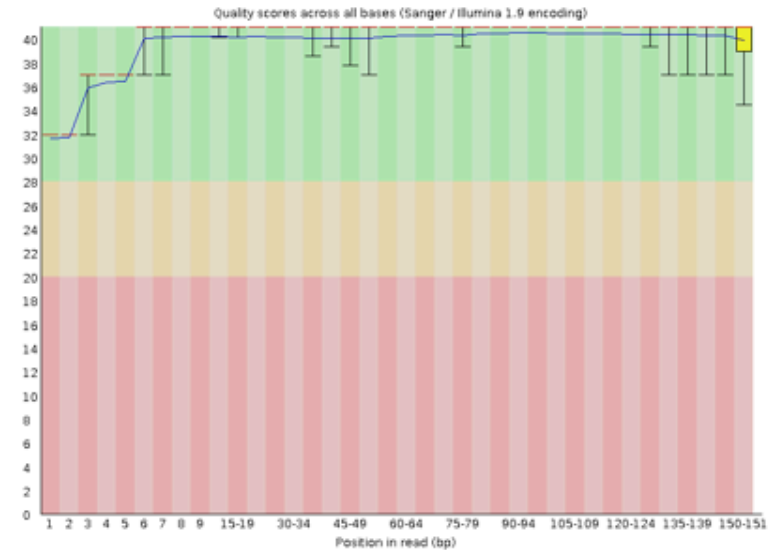


After trimming

Basic Statistics

Measure	Value
Filename	316_S78_L007_R1_001.trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	32365938
Sequences flagged as poor quality	0
Sequence length	36-151
%GC	49

Per base sequence quality



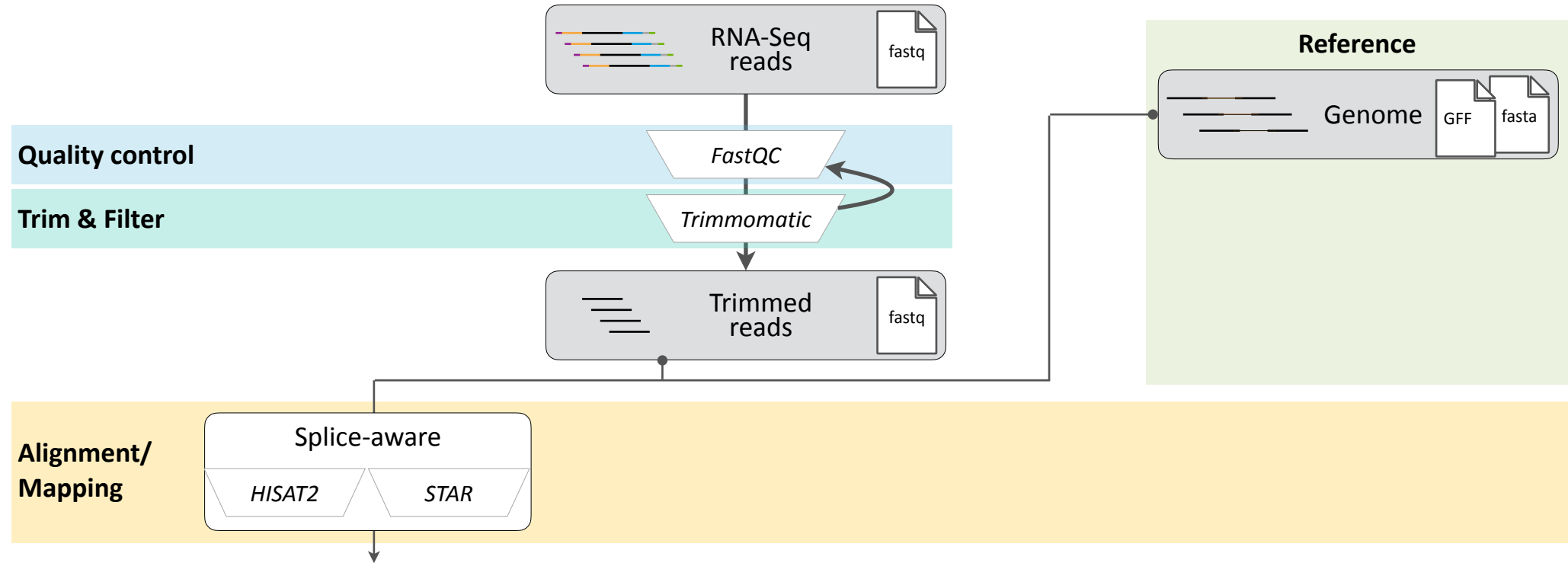


Read trimming



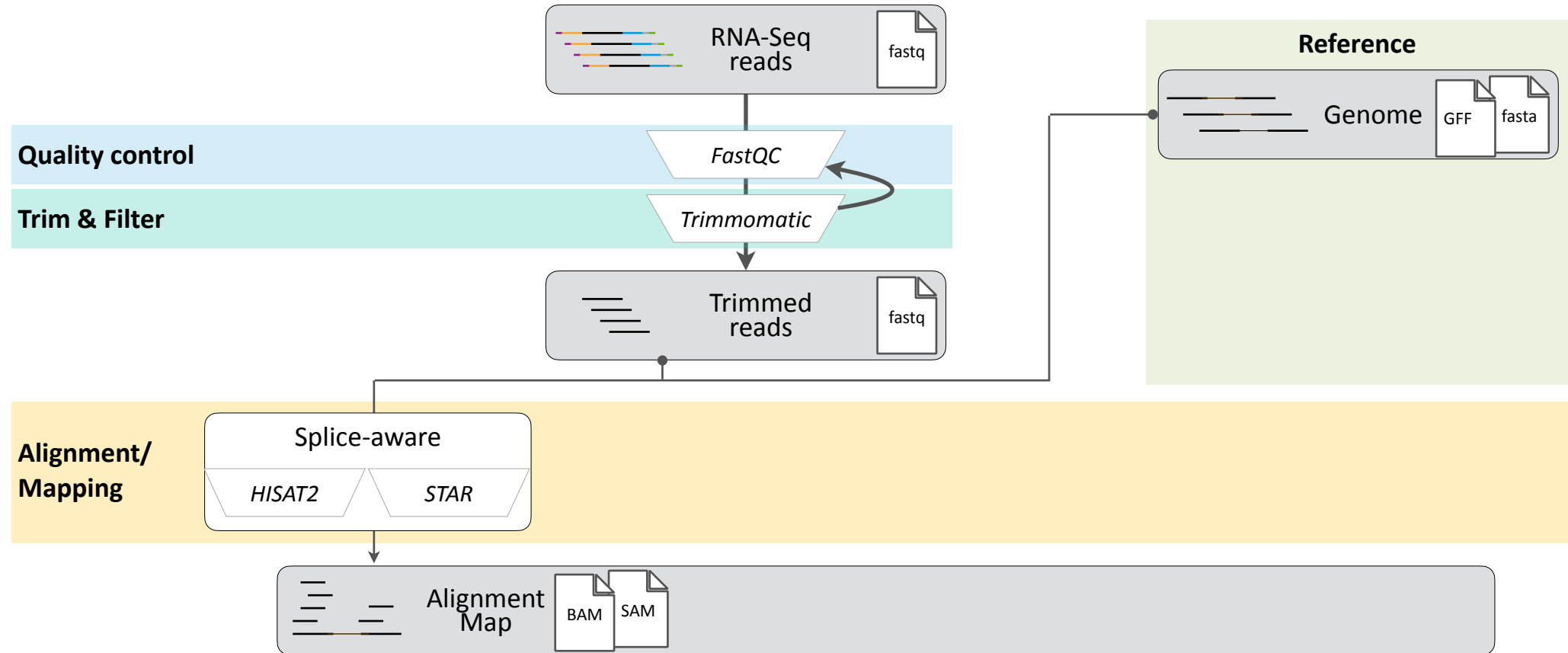


Alignment



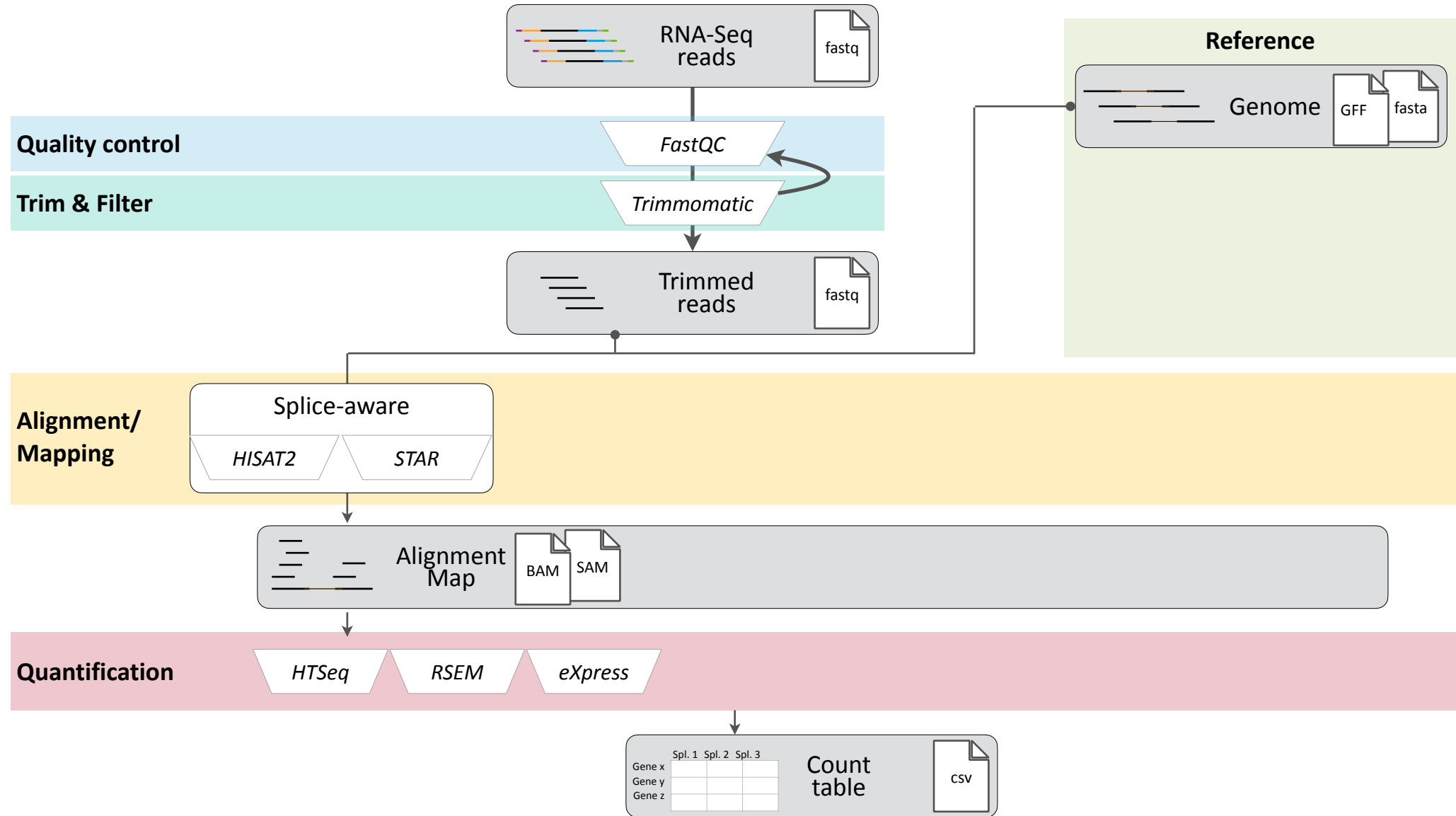


Alignment



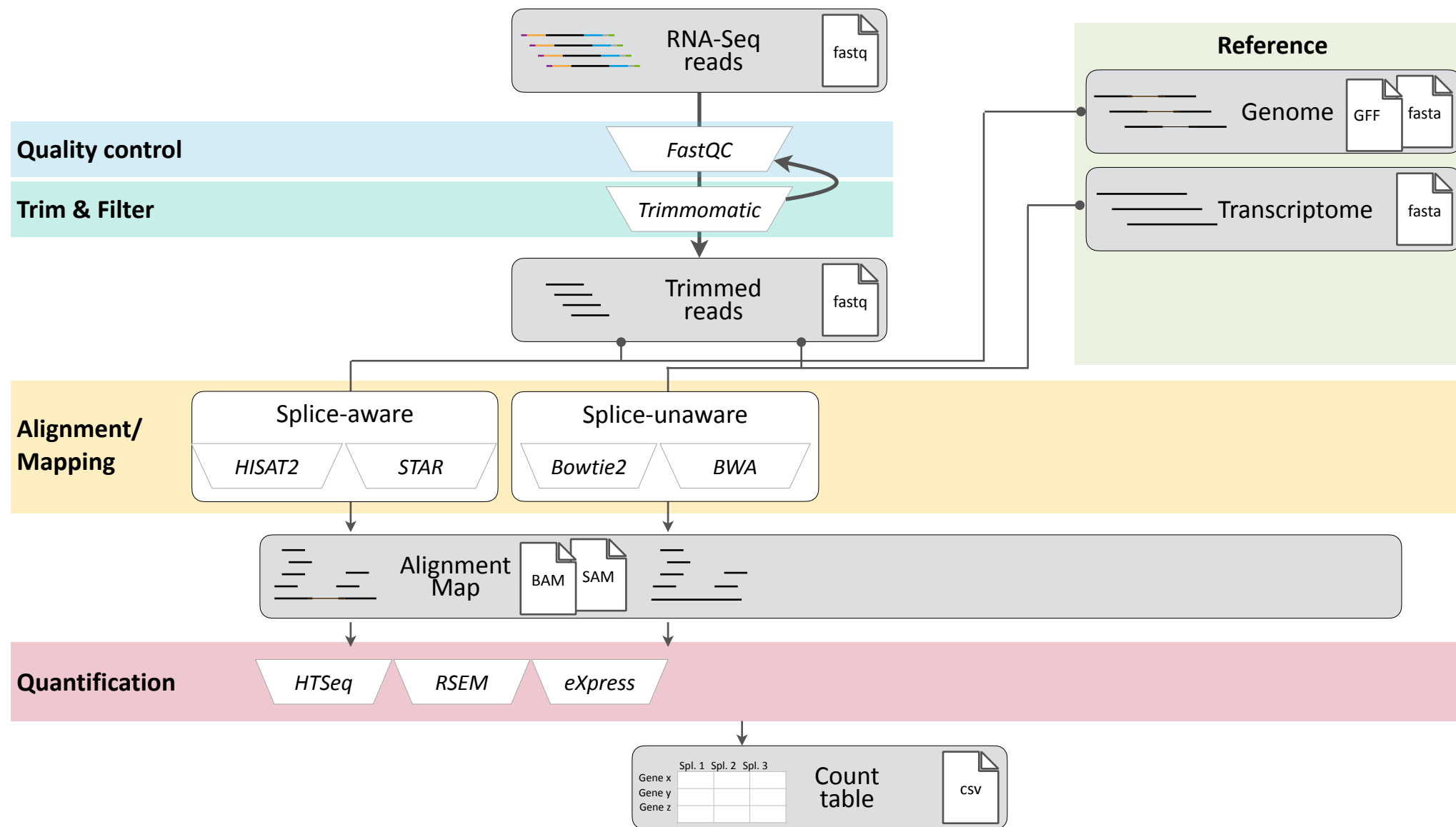


Alignment & Quantification



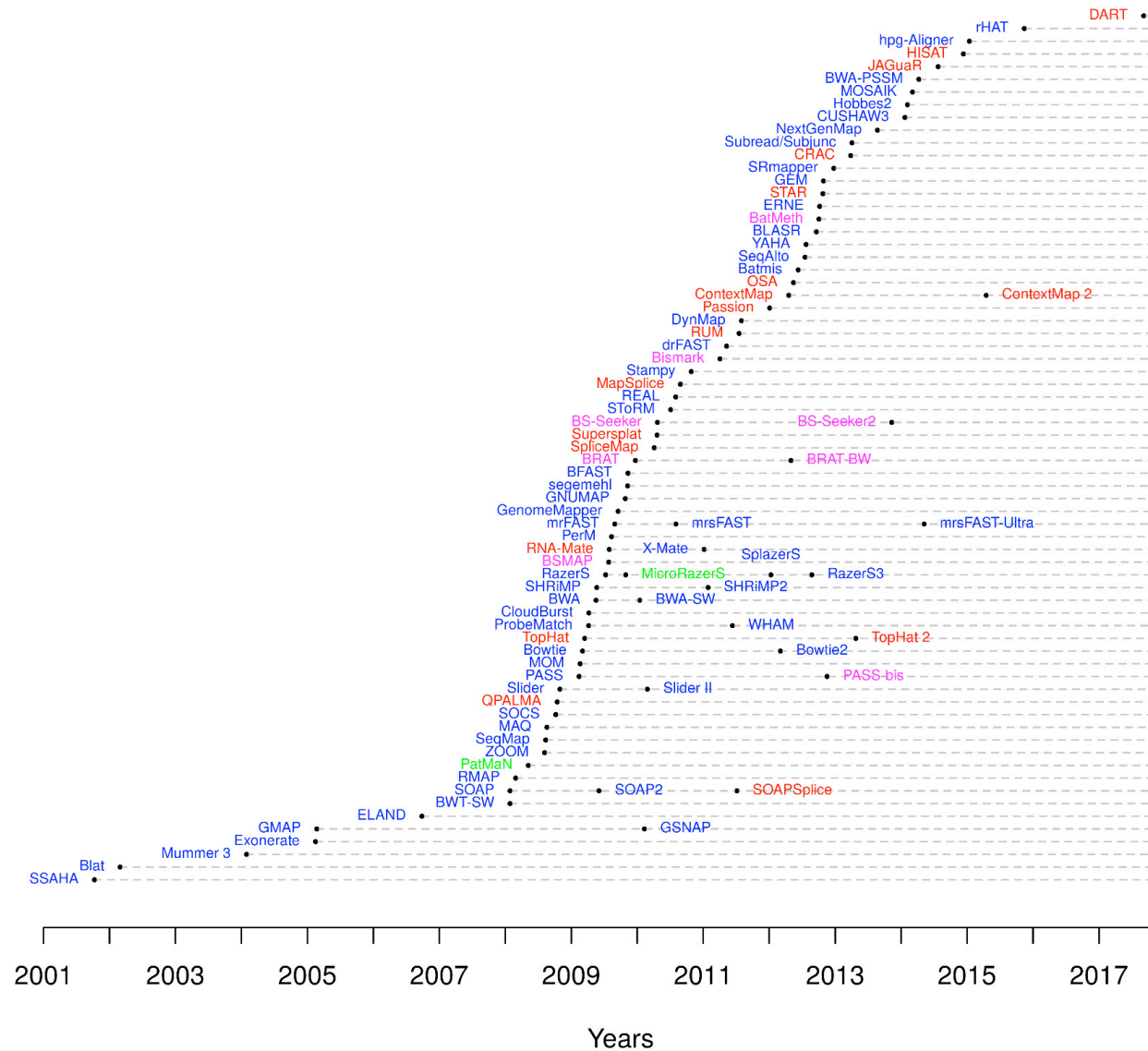


Alignment & Quantification





Read mappers



https://www.ebi.ac.uk/~nf/hts_mappers/





Alignment & Quantification

Filtered reads

```

TAGGAAAGATTAGAGGGAAATGTC
TAGGAAAGATTAGAGGGAAATGTC
TAGGAAAGATTAGAGGGAAATGTC
ACCCAATGAGCCCTACCGTAATCT
ACCCAATGAGCCCTACCGTAATCT
GGACAGAGGAGTATCTACAATAGTA
GGACAGAGGAGTATCTACAATAGTA
GGACAGAGGAGTATCTACAATAGTA
GGACAGAGGAGTATCTACAATAGTA
GGACAGAGGAGTATCTACAATAGTA
TTGATTAACTATTCTGCTGCACAG
TTGATTAACTATTCTGCTGCACAG
TTGATTAACTATTCTGCTGCACAG
TTGATTAACTATTCTGCTGCACAG
TTGATTAACTATTCTGCTGCACAG
TTGATTAACTATTCTGCTGCACAG

```

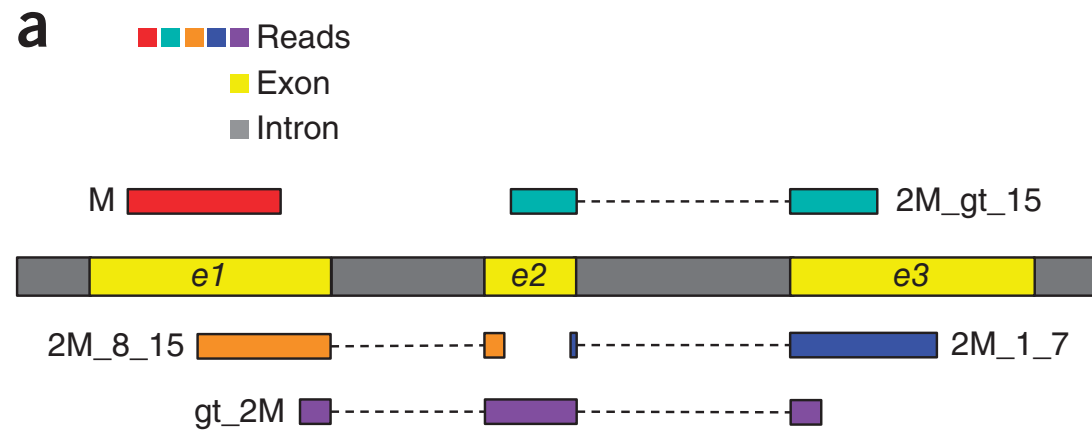
Reference genome /
transcriptome

```

CGAGGATTACACGTGTAGACGCAGTGAGAAAGTAGGA

```

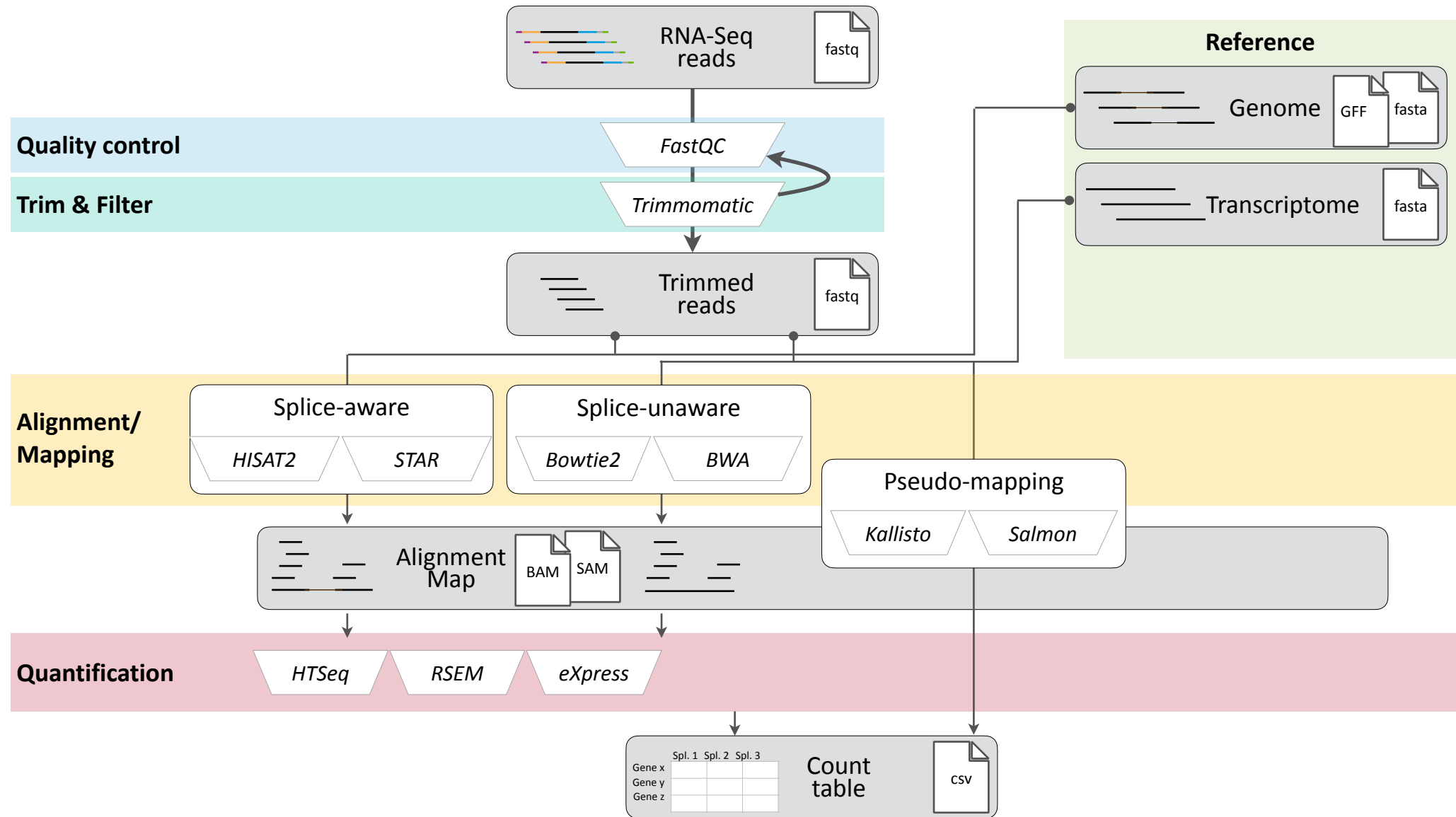




Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 12: 357–360.



Pseudo-mapping (e.g. Kallisto)



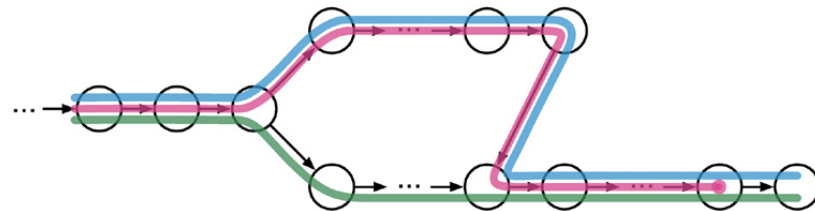


Pseudo-mapping (e.g. Kallisto)

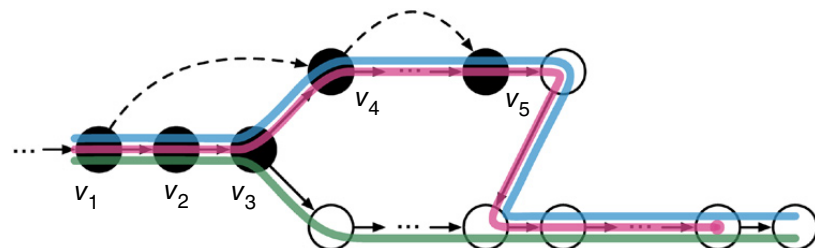
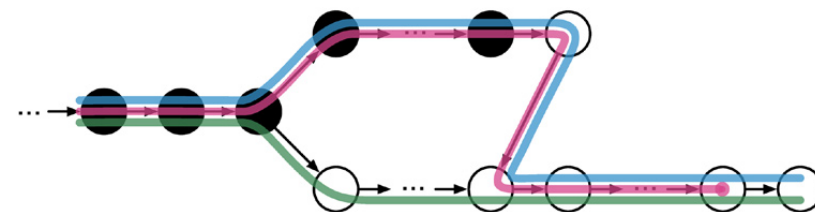


RNA-Seq read

Overlapping transcripts



Index: Transcriptome de Bruijn Graph (T-DBG)



Skip redundant k -mers

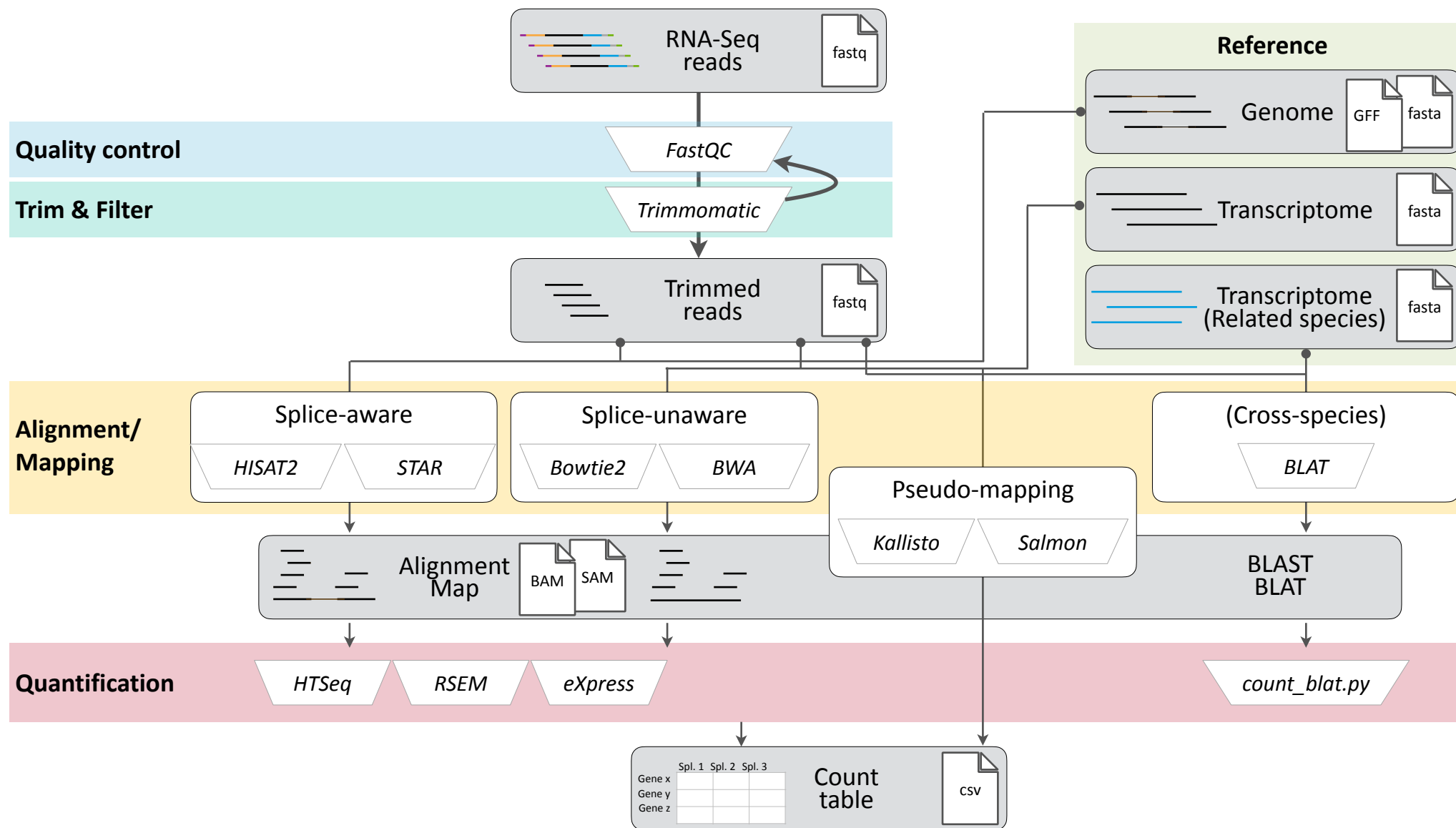
$$\begin{matrix} \text{blue} \\ \text{pink} \\ \text{green} \end{matrix} \cap \begin{matrix} \text{blue} \\ \text{pink} \\ \text{green} \end{matrix} \cap \begin{matrix} \text{blue} \\ \text{pink} \\ \text{green} \end{matrix} = \begin{matrix} \text{blue} \\ \text{pink} \\ \text{green} \end{matrix}$$

$v_1 \quad v_4 \quad v_5$

Bray, N.L. et al. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34: 525–527.



BLAST / BLAT





Reader section 6: Example short read RNAseq analysis

