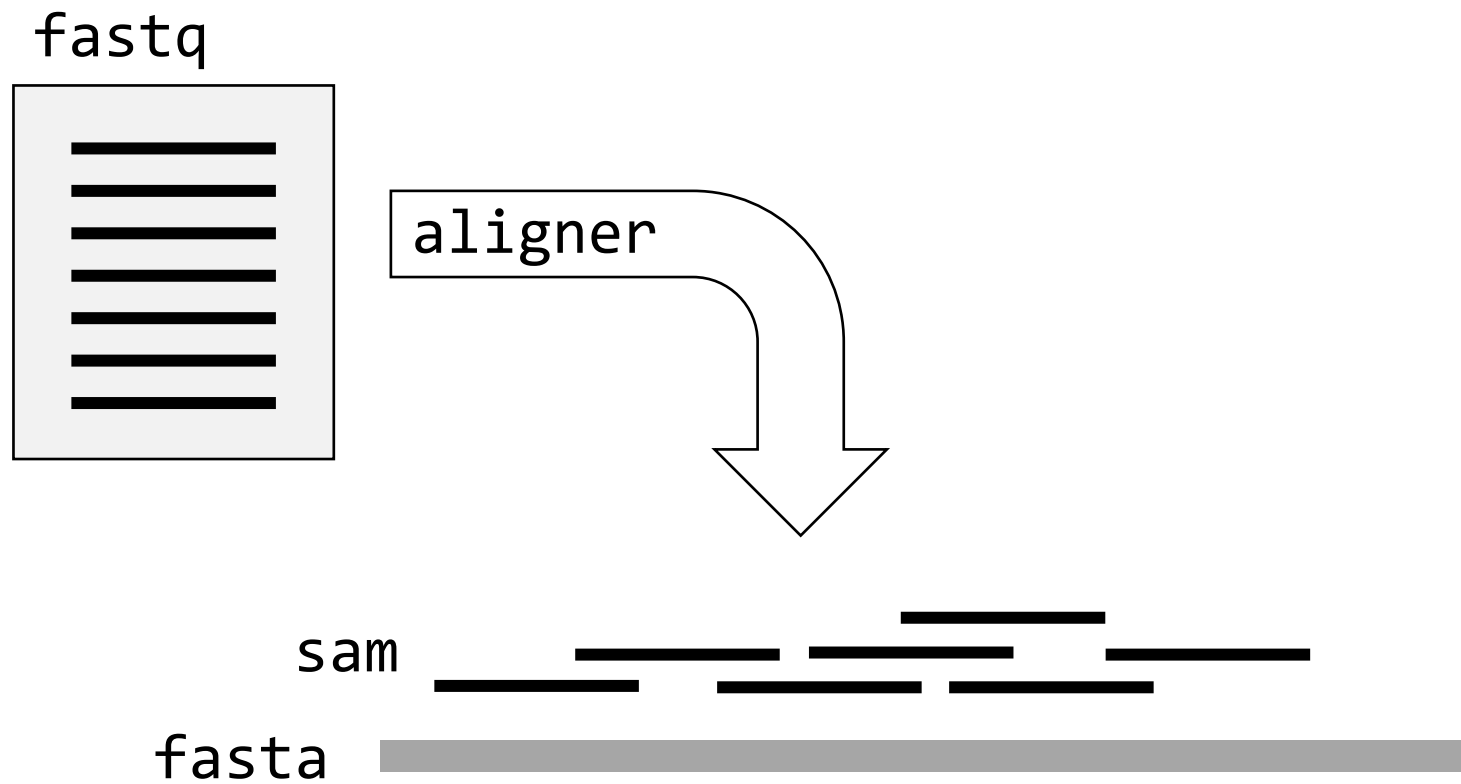
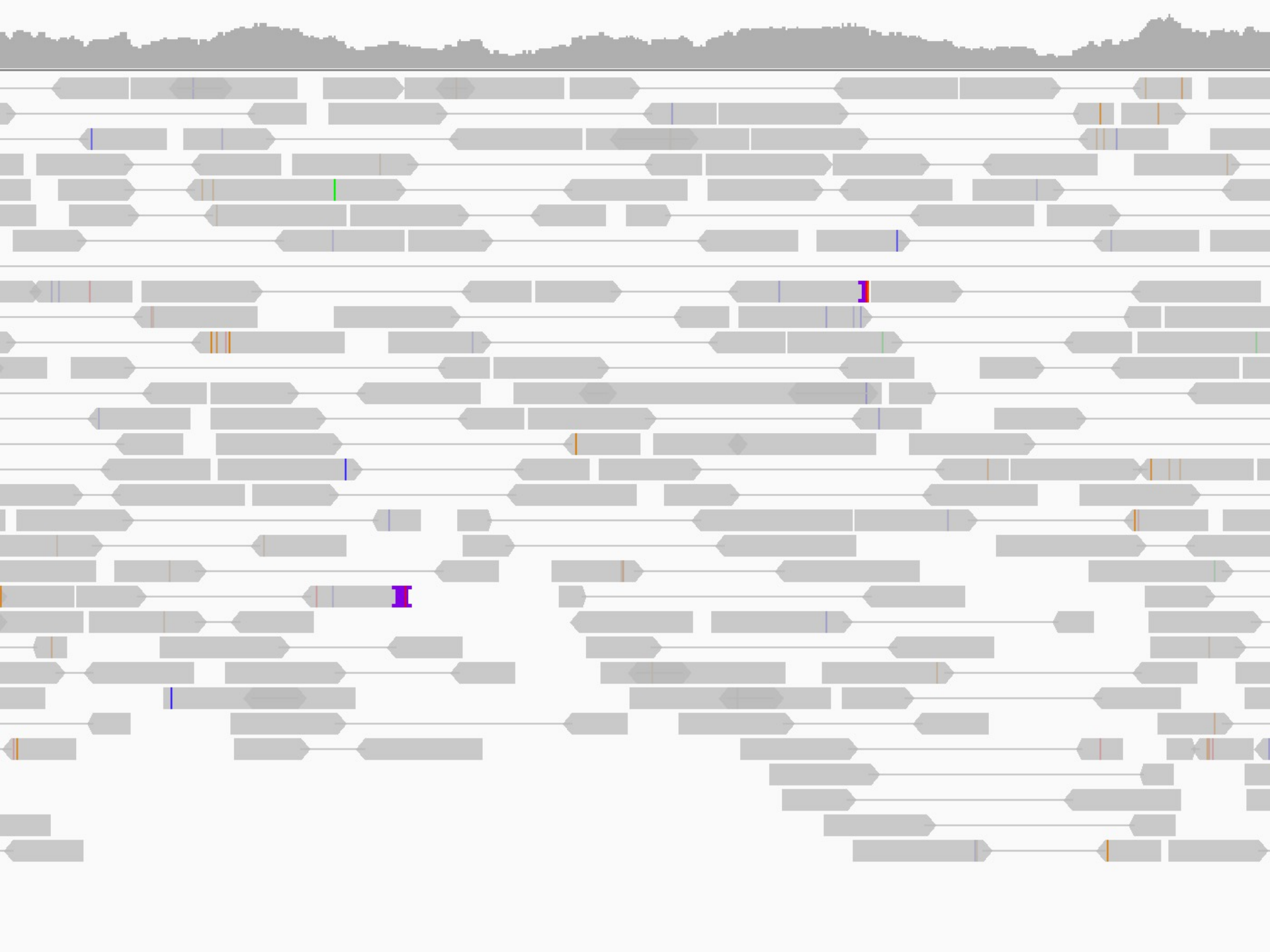


NGS - quality control, alignment, visualisation

Read alignment





How do aligners work?

Aim: find **substrings** in **large string**



Typically:

- Millions of substrings (reads)
- In string of tens of millions of characters (genome)

Indexing

Aim: generate a 'phonebook' for fast searches

Reference: TAATA\$

↑
EOF

suffix array

0	T	A	A	T	A	\$
1	A	A	T	A	\$	
2	A	T	A	\$		
3	T	A	\$			
4	A	\$				
5	\$					

→
sort


5	\$					
4	A	\$				
1	A	A	T	A	\$	
2	A	T	A	\$		
3	T	A	\$			
0	T	A	A	T	A	\$

Querying

Reference: TAATA\$

Query: ATA

Can use binary search:



5	\$					
4	A	\$				
1	A	A	T	A	\$	
2	A	T	A	\$		
3	T	A	\$			
0	T	A	A	T	A	\$

Indexing and querying

- Suffix array: large, same sequence stored multiple times
- BWT: only **first** and **last** columns are stored -> still enables fast querying

suffix array

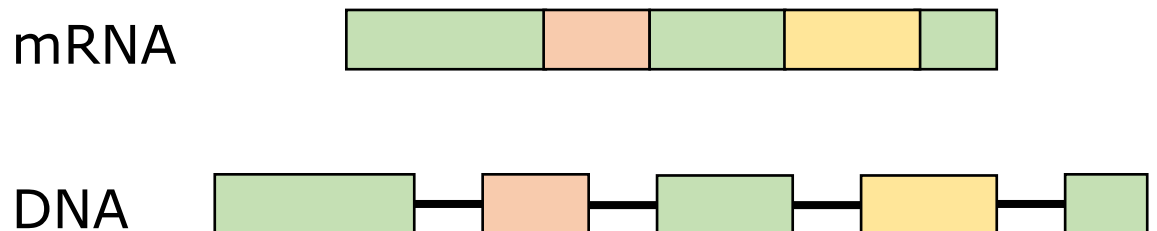
5	\$					
4	A	\$				
1	A	A	T	A	\$	
2	A	T	A	\$		
3	T	A	\$			
0	T	A	A	T	A	\$

Burrows-**W**heeler **T**ransformation

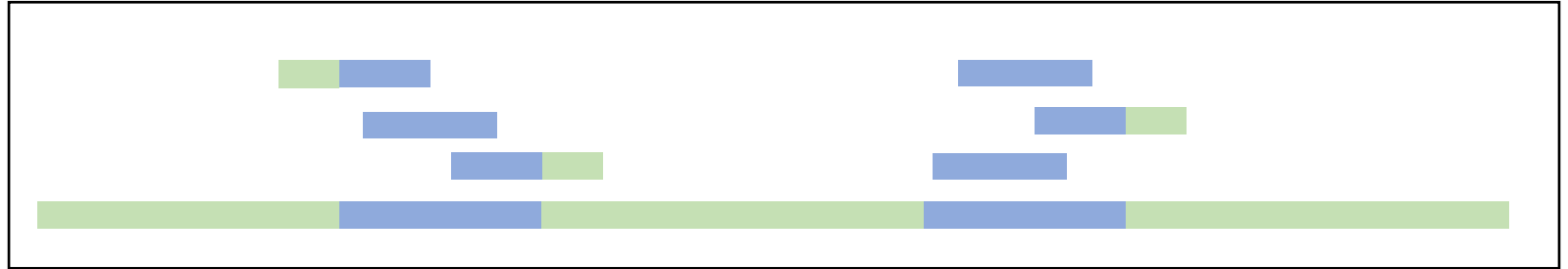
\$	T	A	A	T	A
A	\$	T	A	A	T
A	A	T	A	\$	T
A	T	A	\$	T	A
T	A	\$	T	A	A
T	A	A	T	A	\$

Software

- Basic alignment:
 - bowtie2 (Burrows-Wheeler transformation)
 - bwa-mem (Burrows-Wheeler transformation)
- Splice-aware (RNA-seq):
 - hisat2
 - STAR
- Long reads + short reads + splice-aware:
 - minimap2



Mapping quality



$MAPQ$

$= -10\log_{10} \Pr\{\text{mapping position is wrong}\}$

$$-10\log_{10} (0.01) = 20$$

$$-10\log_{10} (0.5) = 3$$