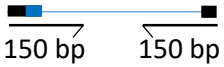
Mapping to a reference genome

1. Quality check and trimming raw reads

paired-end Illumina sequencing

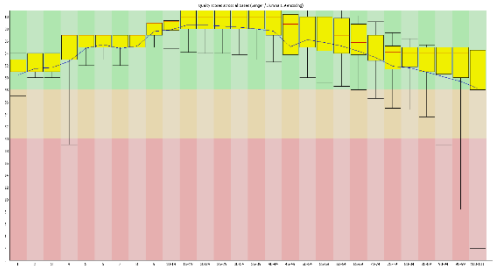
Up to 20 billion read pairs



Forward reads

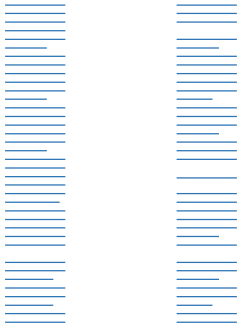
Reverse reads

Fastqc quality plots

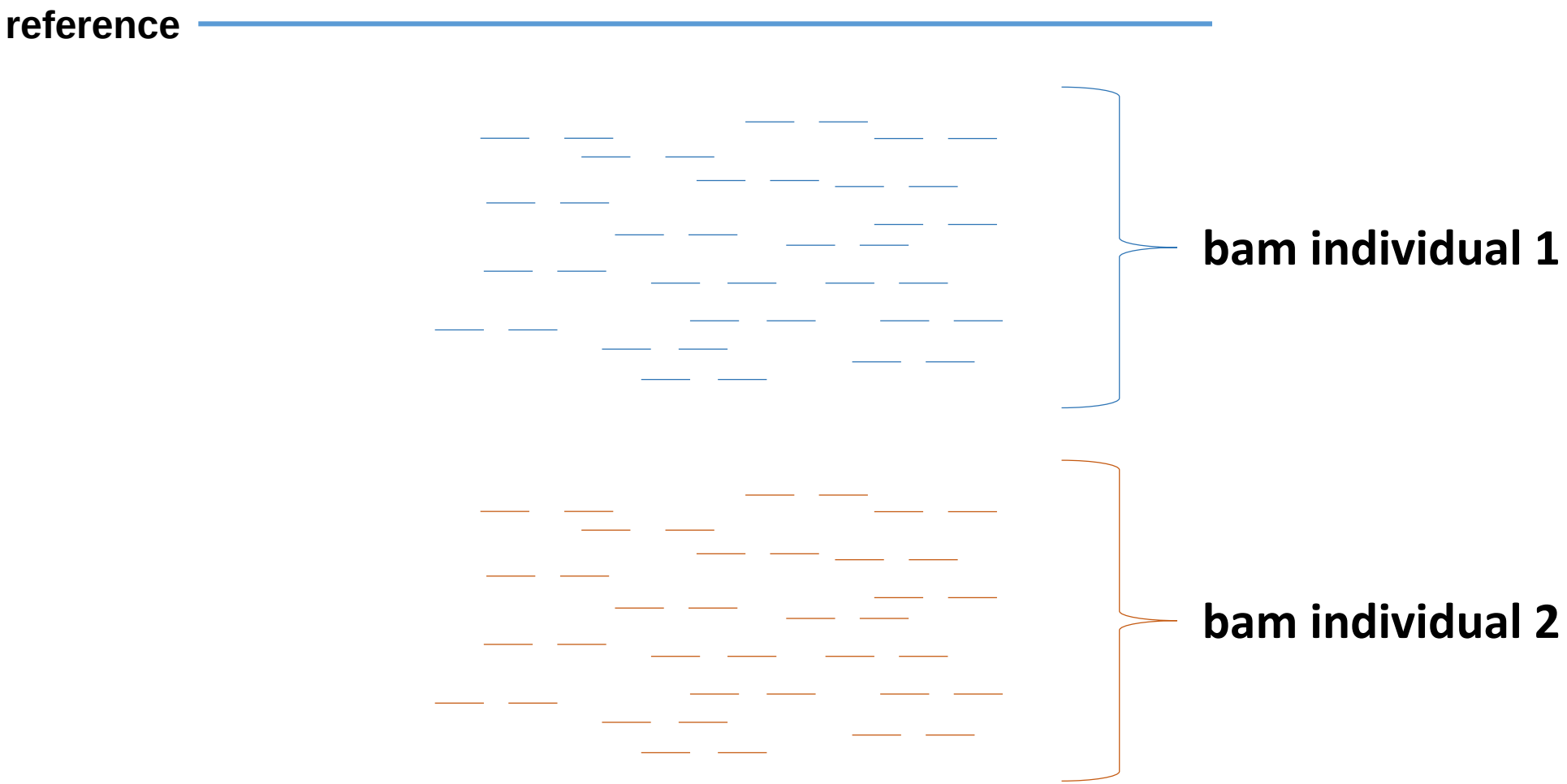


fastp

Trimmed and filtered reads of good quality



2. Alignment to the reference genome



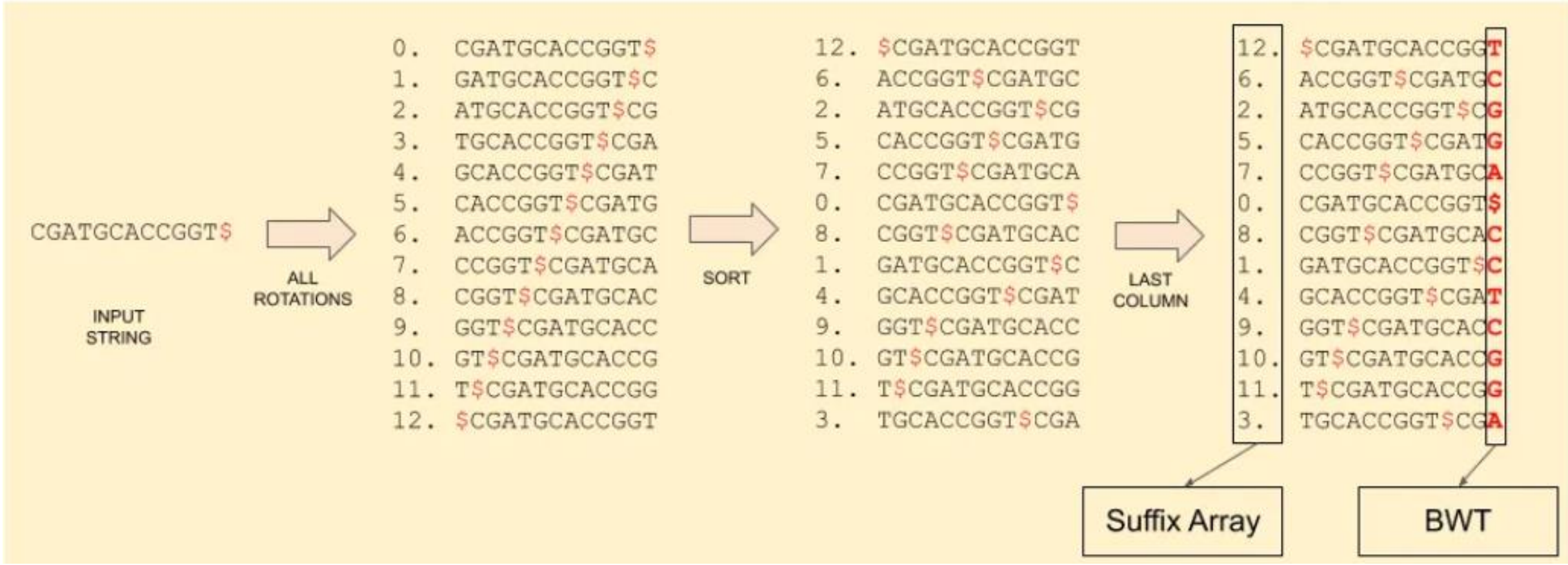
Tools to align short read data

bwa-mem2 (Burroughs-Wheeler alignment – maximum exact matches)

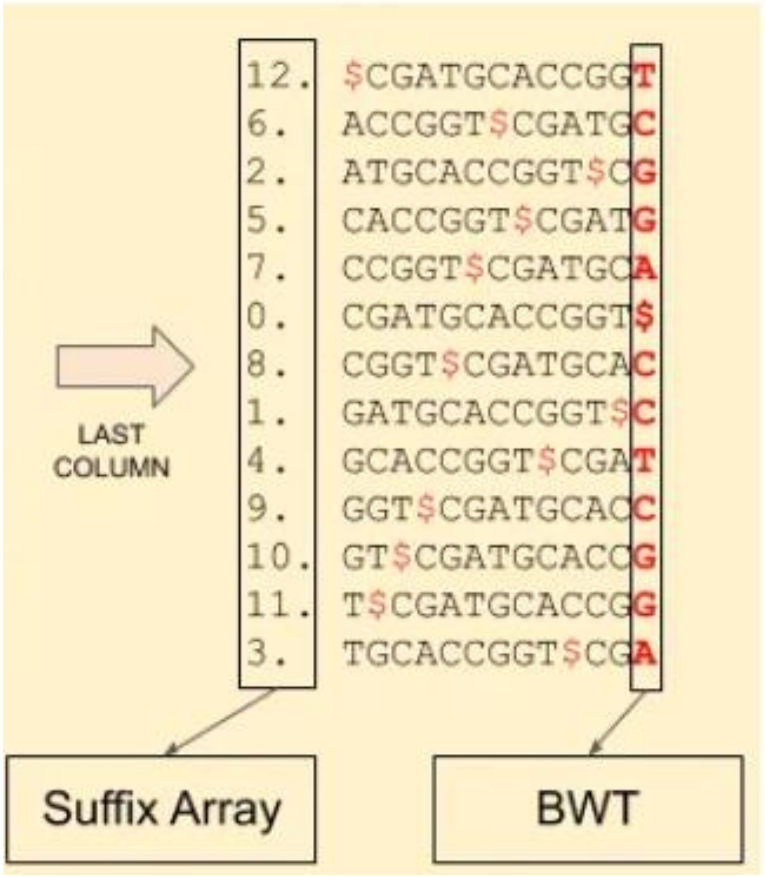
Bowtie2

Burroughs Wheeler Transform

(also used in file compression, e.g. bzip, figures from https://medium.com/@mr-easy)

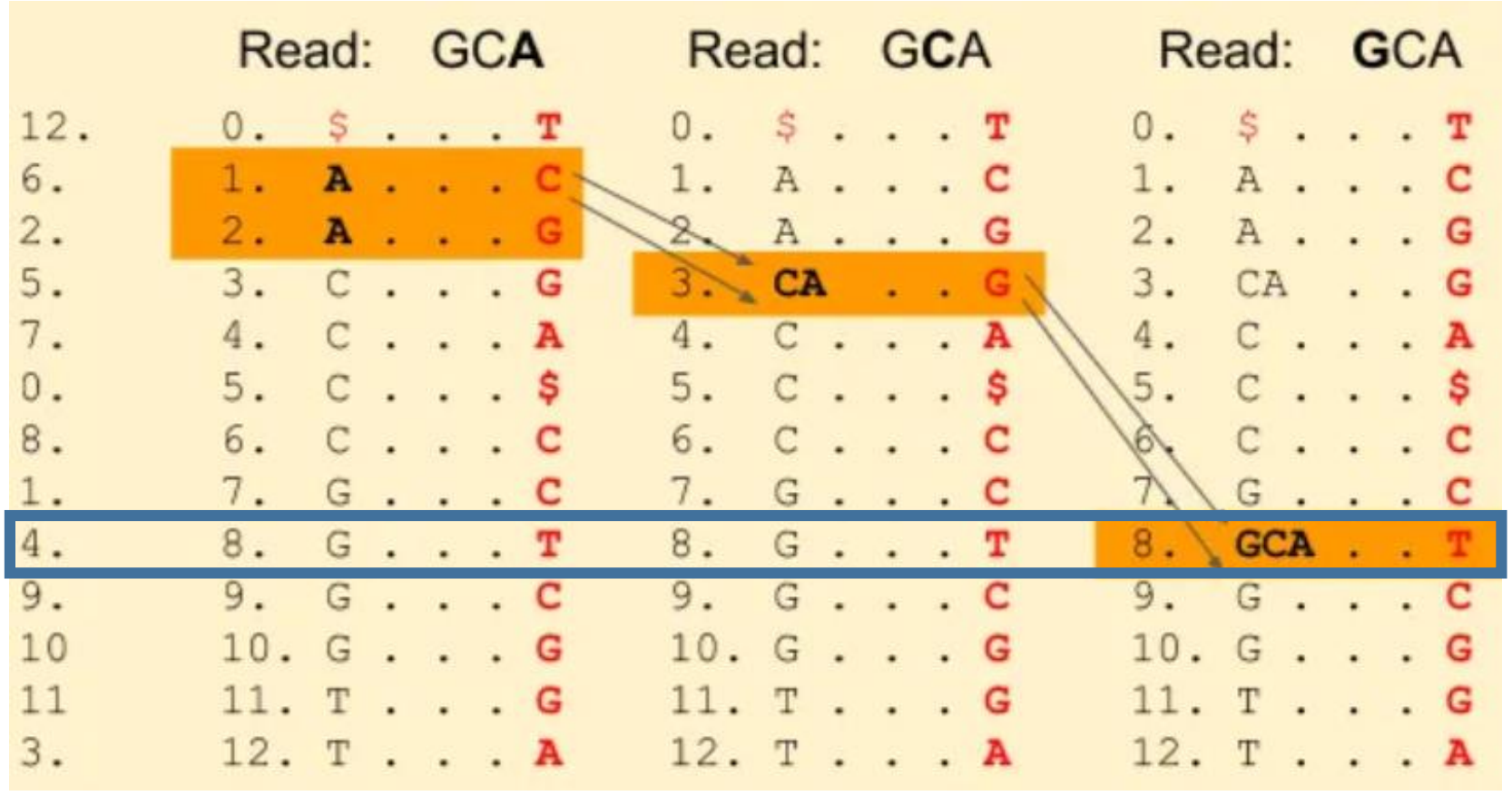


Burroughs Wheeler Alignment



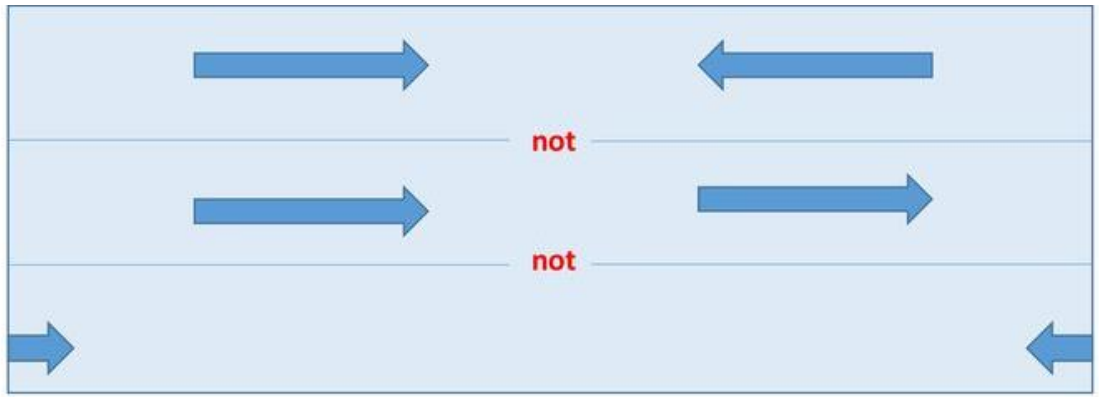
CGATGCACCGGT$

Read: GCA



Aligning reads is complicated by:

- Imperfect match to the reference due to mutations or sequencing errors, or errors in the reference genome
- Multiple positions where the read could match (repeated regions)
- Low quality of the read
- With paired-end reads: only one read maps or the other read maps on a different chromosome



Alignment tools such as bwa-mem2 are able to handle all of these complications and give information on the mapping quality.

Regions that are repeated in the genome make it very difficult to map reads

Reference genome

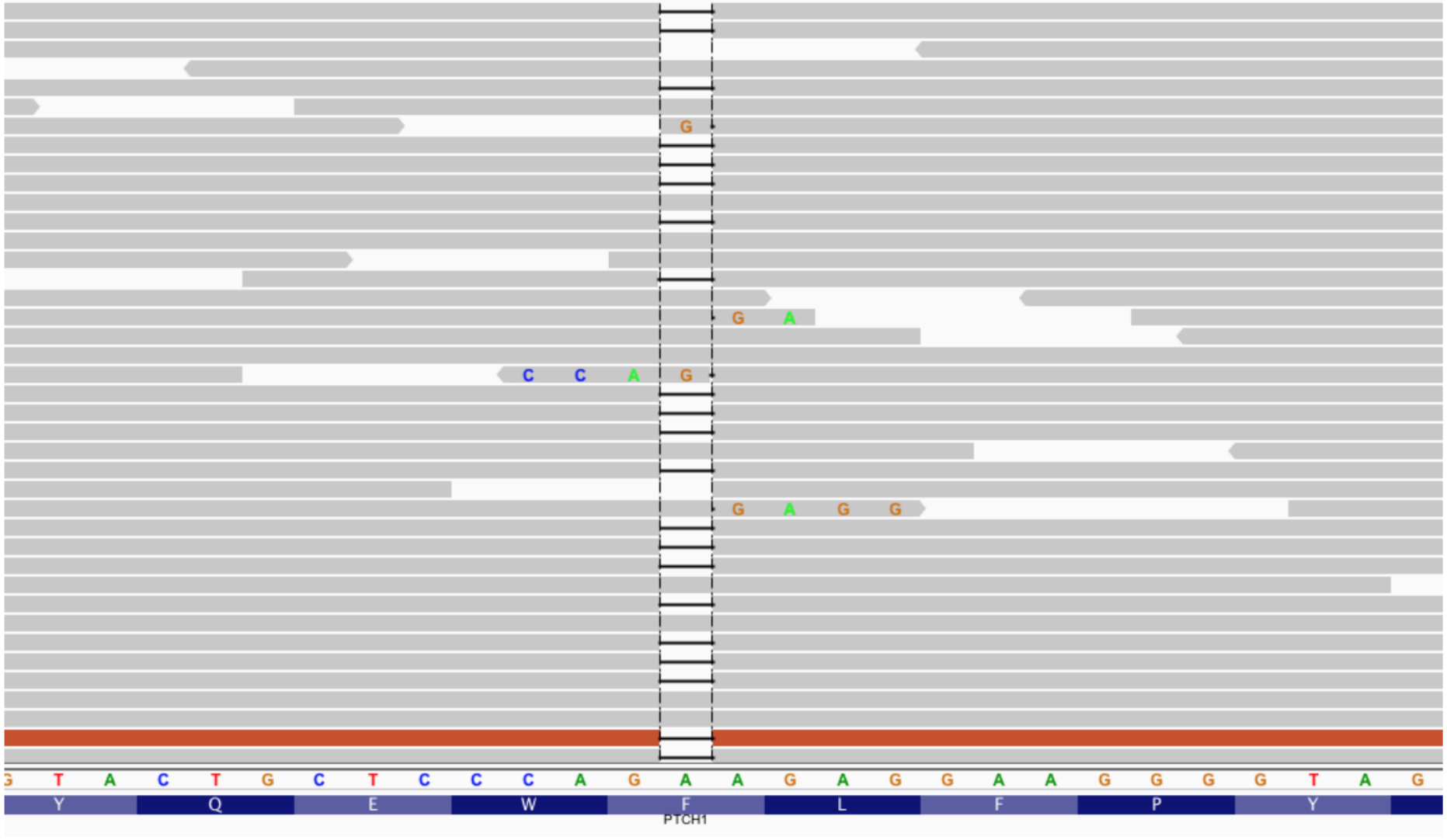
Read to be aligned to the reference

Regions that are repeated in the genome make it very difficult to map reads

Reference genome

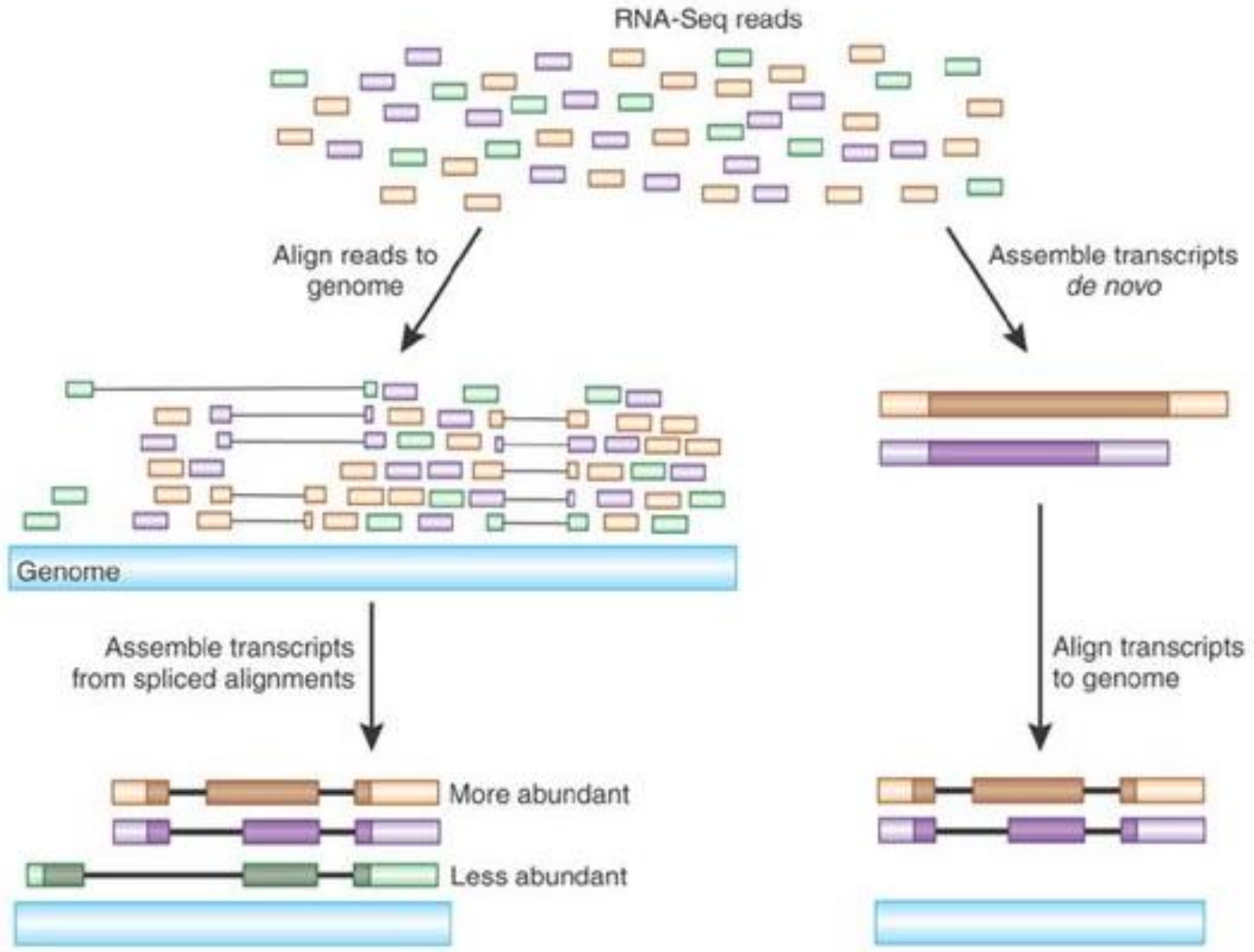
Read to be aligned to the reference

Indels: deletions and insertions



Aligning other type of data: RNA data

Tools:
e.g.
STAR
HISAT2



Haas et al. 2010

PacBio mapping

Karin Näsvall

Sequel IIe Instrument

Generate Sequencing Data

Data Transfer/SMRT Link Server

SMRT Link GUI

Access using a web browser

Sequel IIe System output files

```
<your_specified_output_directory>/r64009_20200825_221039/1_A01/
|-- m64009_200825_222052.baz2bam_1.log
|-- m64009_200825_222052.ccs.log
|-- m64009_200825_222052.ccs_reports.json
|-- m64009_200825_222052.ccs_reports.txt
|-- m64009_200825_222052.consensusreadset.xml
|-- m64009_200825_222052.reads.bam
|-- m64009_200825_222052.reads.bam.pbi
|-- m64009_200825_222052.sts.xml
|-- m64009_200825_222052.transferdone
|-- m64009_200825_222052.zmw_metrics.json.gz
```

Automatic HiFi reads generation (Export Reads)

```
hifi_reads.fastq.gz- FASTQ file containing HiFi Reads
hifi_reads.fasta.gz- FASTA file containing HiFi Reads
hifi_reads.bam - BAM file containing HiFi Reads
```

HiFi reads QV > 20

https://pacbiofileformats.readthedocs.io/en/13.0/

PacBio

https://pacbiofileformats.readthedocs.io/en/13.0/BAM.html#hifi-reads

- `$ samtools view -H /pacbio/m84093_240426_124306_s1.ccs.bc2033.rmdup.bam |head -n2`
- `@HD VN:1.6 SO:coordinate pb:5.0.0`
- `@RG ID:f56a67e5/0--0 PL:PACBIO DS:READTYPE=CCS;Ipd:CodecV1=ip;PulseWidth:CodecV1=pw;BINDINGKIT=102-739-100;SEQUENCINGKIT=102-118-800;BASECALLERVERSION=5.0;FRAMERATEHZ=100.000000;BarcodeFile=/lustre/scratch123/tol/resources/barcodes/PacBio_ULI_adapter.fasta;BarcodeHash=cf95303a081e62fbaedf29888b16fdb7;BarcodeCount=1;BarcodeMode=Symmetric;BarcodeQuality=Score LB:TRAC-2-8589 PU:m84093_240426_124306_s1 SM:Meier_Genomes13637824 PM:REVIO BC:AAGCAGTGGTATCAACGCAGAGTACT CM:R/P1-C1/5.0-25M`

PacBio

https://pacbiofileformats.readthedocs.io/en/13.0/BAM.html#hifi-reads

- $ samtools view /pacbio/m84093_240426_124306_s1.ccs.bc2033.bam |head -n2
- m84093_240426_124306_s1/261297081/ccs 4 * 0 255 * * 0 0
GTTTGGTTATATTGTTCATAGACACCTACGGTACACGAATATAATAAACAAATAATATTAAATTGTTATACATATTTAAATATAAATTAATATTGTAGCAACTATTAACTAACTAGTTCA
CTTCCTTCAACAGTTTAAATCAAGGATTCCTTAACTATTTTTTACCAAAACGTACTGATACCAAAGATTGTTTTCTAATGATTTTGATGATATGGATGTGATTTGACTACATTAATGCTA
CTGCTAGAACTTTACGATATCCCTCACATCACCTCGTCAAAATGCTTTGTTTAAATTGTTAAAAAAATTCTTTGTATTTAAAGTGAGTAATCCTTAAGTAAACAGTGTTTCAAAAGTTTA
AAATTTATTTTATACGTGAAATGTCGTATTATCGATGGCATTATTAGTTTCCAGCATATTTATAAAAGCCTTTGTGACGCAATCACCGAGCGGTGAAAATATGATTTGAGACGCTTCCTG
AAATTGAAAAAAGAAACCCTCTATTGTTTTTAGATGATGTTTATTTTCTGCATTACTTTCTTAAAATTAATATAATTTATACTATACAAATGTGACATTTCAATTGCGATACGAGTTTAA
GAGGAGAACATCGATTACTGAAGTTATACAAACTTCTCCACAATAGATAACATTAACTTTATATTTTAAACTTTAAGATAGTTCTTAAAAGTGTGAAAAGATGTAAAAAATGGTACATTT
GAAGAAACCTTTATTATGAAAGAGTTTAAAAGGTATCGGTATTTTCAAAAAGATTTTACATCTTTCTAATGCGAAATGTTAAAAATCTTAAAAATATTTTTTAGCTTATTTAACGAATAA
GAAGAGAAAATATTTATTAAAATTCATATTTGATATATTTATATTATTGATTTTATTTATTCCTTTTTAAATAGTTTTAATTTAGTTCGATTTCGAAATCTCTTTTATTGTTAAATGTTT
CATCTGGAGACATTTACAGTCCAATAGTTTCCAAAAGAAATTTTAAAACGCTGTTTCGAAATTATCTTTCGTGATTATGAACTAGTTAAAAATTCCACTCAGTTAATGGTTTAAAATTTT
ATGTATTGCAAGAAAAGATAATAAAAGCTTGAAATAAAATCATTTTATCAGTATTTTGCATACCTATAAATATTTGAAGGTA
- ----
- IIIIIIIIIIIIIDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIDIIIIIIIIIIIII7III
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII<IIIIIIDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIDIII
IIIIIIIIIIIIIIIIIIII+IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII+III7IIIIIIIIIIIIII+IIIIIIIIIIIIIIDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIID
IIIIIIIIIIIIIIII2IIIIIIIIIIIIIIDI2IIIIIIIIIIIIIIIIIIIIII7IIIIIDIIIIIIIIIIIIIIIII
- ac:B:i,8,0,10,0 bl:Z:CGACTCGTATGAGTAT bt:Z:ATACTCATACGAGTCG bx:B:i,16,16 ec:f:7.5012 ma:i:0 np:i:7 gl:Z:IIIIIIIIIIIIIIII
qt:Z:IIIIIIIIIIIIIIII rq:f:0.999815 sn:B:f,17.2518,25.3332,6.55462,11.2707 we:i:8638323 ws:i:40959 zm:i:261297081
qs:i:26 qe:i:17132 bc:B:S,0,0 bq:i:100 cx:i:12
ls:B:C,134,162,98,99,164,48,45,45,48,162,98,113,100,164,108,101,97,100,130,162,113,108,186,73,73,73,73,73,73,73,73,73,73
,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,162,115,113,186,65,65,71,67,65,71,84,71,71,84,65,84,67,65,65,67,71,67,6
5,71,65,71,84,65,67,84,166,110,101,115,116,101,100,192,164,112,53,109,99,0,165,116,114,97,105,108,130,162,113,108,186,73
,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,73,162,115,113,186,65,71,84,65,67,84,67,84,71,6
7,71,84,84,71,65,84,65,67,67,65,67,84,71,67,84,84 RG:Z:f56a67e5/0--0

PacBio

```
cat m84093_240426_124306_s1.ccs.bc2033.stats
_
A = 7108650862 (34.9%), C = 3058767823 (15.0%), G = 3061827434 (15.0%),
T = 7127816130 (35.0%), CpG = 1003156910 (4.9%)
sum = 20357062249, n = 2218131, mean = 9177.57438537219, largest =
30301, smallest = 135
```

PacBio mapping – Minimap2

Minimap2 https://github.com/lh3/minimap2

Uses **minimizers** – short sequences of length –k [default 15 bases]

1. Extracts minimizers from the reference (target) and index them
2. Match each minimizer in the query sequence against the reference set of minimizers
3. Sorts the position of each minimizer after position
4. Make a chain of the minimizers
5. Repeat for all query sequences

```
Minimap2 –ax map-hifi target.fa query.fa > output.sam
```

```
Align PacBio high-fidelity (HiFi) reads to a reference genome (-k19 -w19 -U50,500 -g10k -A1 -B4 -O6,26 -E2,1 -s200)
```

PacBio

- Preprocessing
  - Filtering adapters (blastn)
- Alignment
  - MINIMAP2
- Alignment post-processing
  - Statistics

https://pipelines.tol.sanger.ac.uk/readmapping