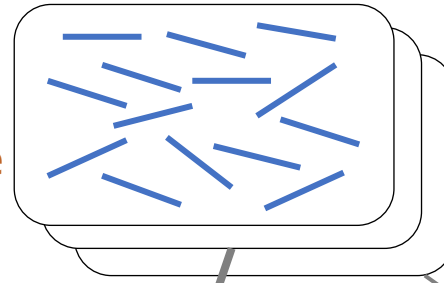# Short Read Mapping & Data Format

Genome (Reference) sequence

format (Sequence)

```
>chr
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAA
TTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAG
AGTACACAACATCCATGAAACGCATTAGCACCACC
```
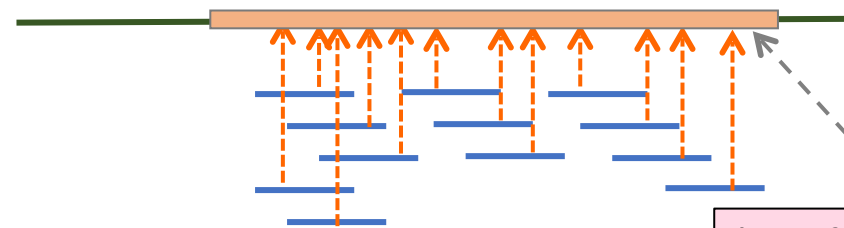
Short read sequence

format (Sequence + Quality)

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
ATCCGGCTGGCGCACCGACCTATGTTCCGGGCGAATACAAGCTGGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CACCGTGTAGTACCAGCATCCTGCGTACAATCAGCAATCCCAGTCCTCCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCFFDFDFHDFFHIIIEGIHJJJJGFHGGHGGHGGIIJDGIJHHGGGHIH
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CAGGACATCGCCTTTGATCGGTTCAGACTTCGGACCAACCTGCATTTTCAG
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CCCFFFFDFAFHFHIJGHIJIJJIJJHEHIIJGHIFEHIIA@FIFHGGIIGI
```

Sample

Mapping to a reference seq.

Quality control
Adapter removal

```
chr  RefSeq  start_codon  190  192  1.000  +  .  gene_id "b0001"; transcript_id "b0001";
chr  RefSeq  CDS          190  252  1.000  +  0  gene_id "b0001"; transcript_id "b0001";
chr  RefSeq  stop_codon   253  255  1.000  +  .  gene_id "b0001"; transcript_id "b0001";
chr  RefSeq  exon         190  255  1.000  +  .  gene_id "b0001"; transcript_id "b0001";
```

format（Result of mapping）

```
@HD          VN:1.0       SO:unsorted
@SQ          SN:chr       LN:4639675
@PG          ID:bowtie2   PN:bowtie2   VN:2.2.4   CL:"/bio/bin/bowtie2-alig
SRR1515276.40          0   chr   4423609   42   51M   *   0   0   GGAATTCCTCACTGCCA
SRR1515276.158        16   chr   501700    42   51M   *   0   0   ACGCACCGAGTGCAAAG
SRR1515276.212         4   *     0         0    *     *   0   0   GGCCGCTTTCAGCGTGT
SRR1515276.319         0   chr   2922768   42   51M   *   0   0   GCTTAAGTTGATTAAGG
SRR1515276.367        16   chr   2753873   42   51M   *   0   0   GCGTGTCCGTCCGCAGC
SRR1515276.411         0   chr   3440721   42   51M   *   0   0   ACGGCATAATTTCTTGA
SRR1515276.434         0   chr   4198737   42   51M   *   0   0   GCGCGGTACGCATCTGG
```

format (Gene annotation)

Binary digitization

format (Result of mapping)