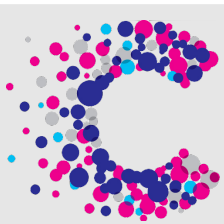
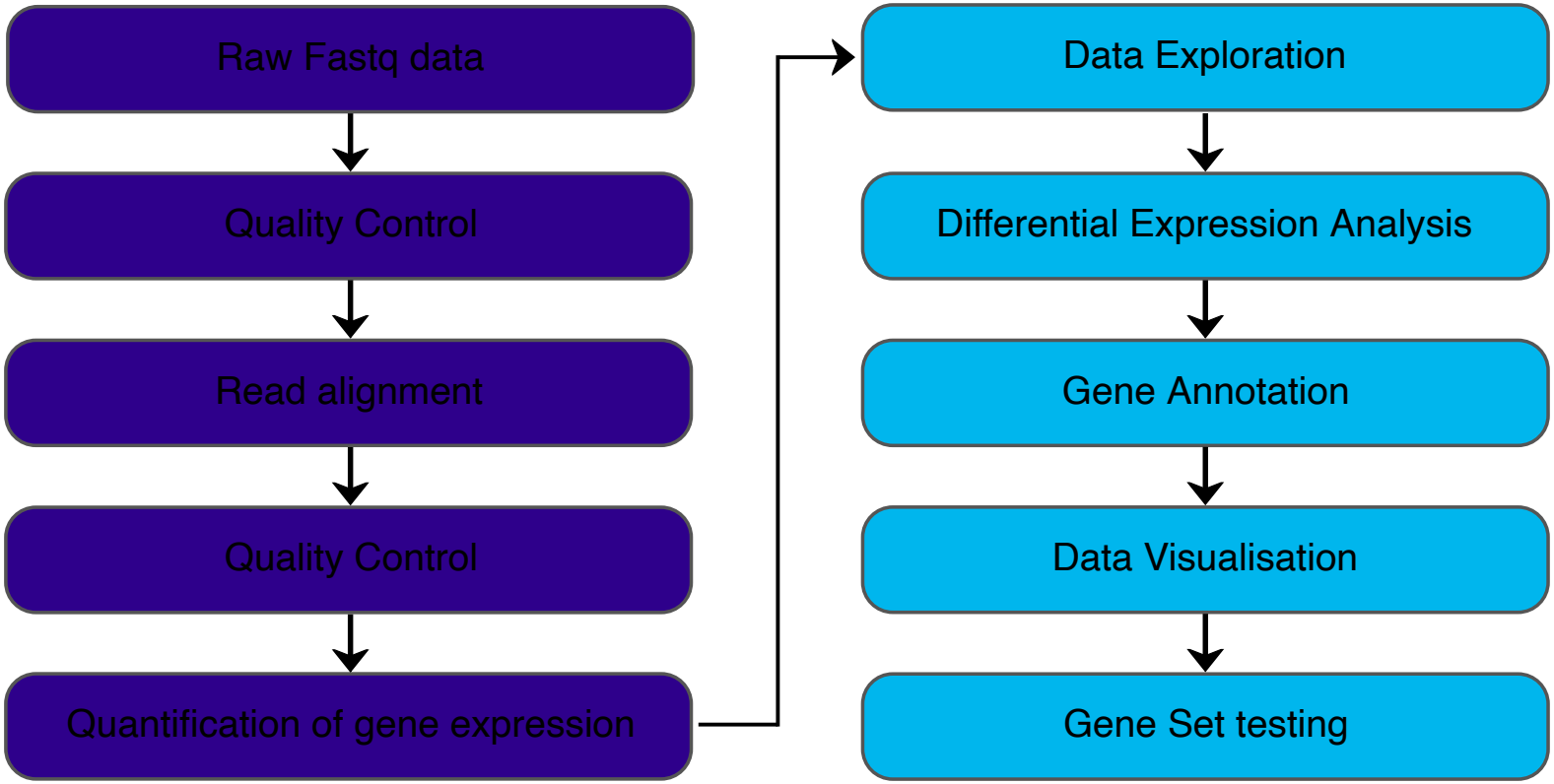


CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Short Read Alignment

April 2021

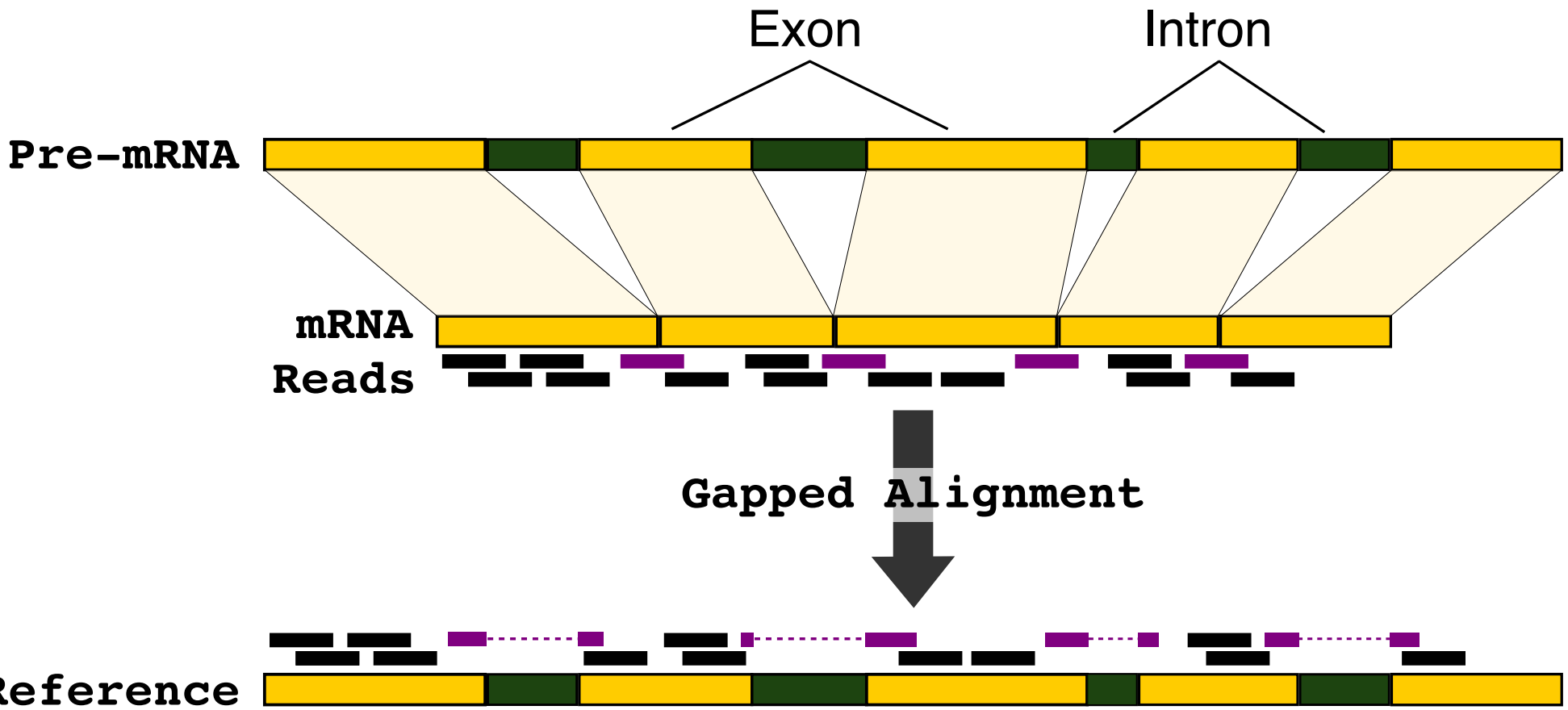
Differential Gene Expression Analysis Workflow



Alignment

AIM: Given a reference sequence and a set of short reads, align each read to the reference sequence finding the most likely origin of the read sequence.

Alignment - Gap aware alignment



Aligners: STAR, HISAT2

SAM format

Sequence **A**lignment/**M**ap (SAM) format is the standard format for files containing aligned reads.

Definition of the format is available at https://samtools.github.io/hts-specs/SAMv1.pdf.

Two main parts:

- Header
  - contains meta data (source of the reads, reference genome, aligner, etc.)
  - header lines start with “@”
  - header fields have standardized two-letter codes
- Alignment section
  - 1 line for each alignment
  - contains details of alignment position, mapping, base quality etc.
  - 11 required fields, but other content may vary depending on aligner and other tools used to create the file

SAM format - header

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:1    LN:195471971
@SQ     SN:10   LN:130694993
@SQ     SN:11   LN:122082543
@SQ     SN:12   LN:120129022
```

……
……

```
@SQ     SN:JH584292.1   LN:14945
@SQ     SN:JH584295.1   LN:1976
@PG     ID:hisat2       PN:hisat2       VN:2.1.0
        CL:"/home/sawle01/Software/hisat2-2.1.0/hisat2-
align-s --wrapper basic-0 -x references/hisat_index/mmu
.GRCm38 -S bam/MCL1.DL.sam -p 7 -U /tmp/1264.unp"
```

SAM format - alignment

```
SRR7657883.sra.4486068  163     1       3207176 60      142M6121N8M     =       3207227 6220
    CTCCTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGG
ATTAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGA  AAFFFAJFJJJFFFFJJJJJFFJJJJJJJFJJJJ
JJFJJJAFFJJJJJJFAJAFJFJJJJFJJJJFFJJJJJJJJFJJJJFJJJJJJJFJ7AJJJJJAJFAFJFFFJFFFJ<J<A<F-<AJ77A<FJJJ
F-7-<FFJ<FJ--F<<F<JA7      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.24078254 99      1       3207179 60      139M6121N11M    =       3213440 290
    CTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGGAT
TAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGA  AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJF      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
   YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.5094794  163     1       3207181 60      43M1D93M6121N14M        =       3213440
 288    TTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCCTGATTGCAAATTACAAGTCAAGAATTTGCAAGATTGAGGTGTCTGTTGGA
TTAATTAACTGAAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGAGTG  -AAFFJJFJJFFJF<AAFFJJJJAF77FJ
-F<FFJ--A7FFJFFF-F-FJ-FJ<JJF-AJFJFJJJJJ<FAFJ-AAA<A-FJJJFA-<7FA<JJ77F--FJJA7FF<-7-AFFJJA-7FA77AF
JJ<A---A-7--7-<F-7-7---<7<      AS:i:-17        XN:i:0  XM:i:3  XO:i:1  XG:i:1  NM:i:4      MD:
Z:35T7^A30A23C52     YS:i:0  YT:Z:CP XS:A:-  NH:i:1
```

SAM format - alignment

```
SRR7657883.sra.4486068 163     1       3207176 60      142M6121N8M     =       3207227 6220
    CTCCTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGG
ATTAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGA  AAFFFAJFJJJFFFFJJJJJFFJJJJJJJFJJJJ
JJFJJJAFFJJJJJJFAJAFJFJJJJFJJJJFFJJJJJJJJFJJJJFJJJJJJJFJ7AJJJJJAJFAFJFFFJFFFJ<J<A<F-<AJ77A<FJJJ
F-7-<FFJ<FJ--F<<F<JA7      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.24078254 99      1       3207179 60      139M6121N11M    =       3213440 290
     CTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGGAT
TAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGA  AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJF      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
   YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.5094794 163     1       3207181 60      43M1D93M6121N14M        =       3213440
 288    TTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCCTGATTGCAAATTACAAGTCAAGAATTTGCAAGATTGAGGTGTCTGTTGGA
TTAATTAACTGAAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGAGTG  -AAFFJJFJJFFJF<AAFFJJJJAF77FJ
-F<FFJ--A7FFJFFF-F-FJ-FJ<JJF-AJFJFJJJJJ<FAFJ-AAA<A-FJJJFA-<7FA<JJ77F--FJJA7FF<-7-AFFJJA-7FA77AF
JJ<A---A-7--7-<F-7-7---<7<      AS:i:-17        XN:i:0  XM:i:3  XO:i:1  XG:i:1  NM:i:4      MD:
Z:35T7^A30A23C52     YS:i:0  YT:Z:CP XS:A:-  NH:i:1
```

SAM format - alignment

```
SRR7657883.sra.4486068 163     1       3207176 60      142M6121N8M     =       3207227 6220
    CTCCTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGG
ATTAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGA  AAFFFAJFJJJFFFFJJJJJFFJJJJJJJFJJJJ
JJFJJJAFFJJJJJJFAJAFJFJJJJFJJJJFFJJJJJJJJFJJJJFJJJJJJJFJ7AJJJJJAJFAFJFFFJFFFJ<J<A<F-<AJ77A<FJJJ
F-7-<FFJ<FJ--F<<F<JA7      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.24078254 99      1       3207179 60      139M6121N11M    =       3213440 290
     CTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGGAT
TAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGA  AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJF      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
   YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.5094794 163     1       3207181 60      43M1D93M6121N14M        =       3213440
 288    TTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCCTGATTGCAAATTACAAGTCAAGAATTTGCAAGATTGAGGTGTCTGTTGGA
TTAATTAACTGAAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGAGTG  -AAFFJJFJJFFJF<AAFFJJJJAF77FJ
-F<FFJ--A7FFJFFF-F-FJ-FJ<JJF-AJFJFJJJJJ<FAFJ-AAA<A-FJJJFA-<7FA<JJ77F--FJJA7FF<-7-AFFJJA-7FA77AF
JJ<A---A-7--7-<F-7-7---<7<      AS:i:-17        XN:i:0  XM:i:3  XO:i:1  XG:i:1  NM:i:4     MD:
Z:35T7^A30A23C52     YS:i:0  YT:Z:CP XS:A:-  NH:i:1
```

SAM format - alignment

```
SRR7657883.sra.4486068 163     1       3207176 60      142M6121N8M     =       3207227 6220
    CTCCTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGG
ATTAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGA AAFFFAJFJJJFFFFJJJJJFFJJJJJJJFJJJJ
JJFJJJAFFJJJJJJFAJAFJFJJJJFJJJJFFJJJJJJJJFJJJJFJJJJJJJFJ7AJJJJJAJFAFJFFFJFFFJ<J<A<F-<AJ77A<FJJJ
F-7-<FFJ<FJ--F<<F<JA7      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.24078254 99      1       3207179 60      139M6121N11M    =       3213440 290
     CTTTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCTTGATTGCAAAATTACAAGTCAAGAATTTGCAAGATTGAAGTGTCTGTTGGAT
TAATTAACTGCAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGA AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJF      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:150   YS:i:0
   YT:Z:CP XS:A:-  NH:i:1
SRR7657883.sra.5094794 163     1       3207181 60      43M1D93M6121N14M        =       3213440
 288    TTCCCATTAATTGATTCATGTTCTCTTCTAGTAGCCTGATTGCAAATTACAAGTCAAGAATTTGCAAGATTGAGGTGTCTGTTGGA
TTAATTAACTGAAATTCATCTCCAGTAAAATTTGGTAAGTTCCAATGTTTATGAAAGAAGAGTG -AAFFJJFJJFFJF<AAFFJJJJAF77FJ
-F<FFJ--A7FFJFFF-F-FJ-FJ<JJF-AJFJFJJJJJ<FAFJ-AAA<A-FJJJFA-<7FA<JJ77F--FJJA7FF<-7-AFFJJA-7FA77AF
JJ<A---A-7--7-<F-7-7---<7<      AS:i:-17        XN:i:0  XM:i:3  XO:i:1  XG:i:1  NM:i:4     MD:
Z:35T7^A30A23C52     YS:i:0  YT:Z:CP XS:A:-  NH:i:1
```

SAM format - alignment

QNAME	SRR7657883.sra.4486068
FLAG	163
RNAME	1
POS	3207176
MAPQ	60
CIGAR	142M6121N8M
RNEXT	=
PNEXT	3207227
TLEN	6220
SEQ	CTCCTTTCCCATTAATTGATTCATGTTCTCTTCTA...
QUAL	AAFFFAJFJJJFFFFJJJJJFFJJJJJJJFJJJJJ...
	AS:i:0
	XN:i:0
	XM:i:0
	XO:i:0
	XG:i:0
	NM:i:0
	MD:Z:150
	YS:i:0
	YT:Z:CP
	XS:A:-
	NH:i:1

SAM format - alignment

QNAME	SRR7657883.sra.4486068
FLAG	163
RNAME	1
POS	3207176
MAPQ	60
CIGAR	142M6121N8M
RNEXT	=
PNEXT	3207227
TLEN	6220
SEQ	CTCCTTTCCCATTAATTGATTCATGT
QUAL	AAFFFAJFJJJFFFFJJJJJFFJJJJ
	AS:i:0
	XN:i:0
	XM:i:0
	XO:i:0
	XG:i:0
	NM:i:0
	MD:Z:150
	YS:i:0
	YT:Z:CP
	XS:A:-
	NH:i:1

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Explain SAM flags

HISAT2

Fast and good performance in published benchmark tests

First need to generate an index for the reference genome with the `hisat2-build` command

Indexing is where all the work takes place and so is computationally intensive

Then we can align reads to the genome with `hisat2`

Practical

1. Create an index to the genome with `HISAT2`
2. Align reads to the genome with `HISAT2` and store outcome in a SAM file
3. Convert the SAM file (human readable text) to BAM (binary) with `samtools`
4. Index the BAM file with `samtools`