SAM/BAM Format

Pliki SAM (Sequence Aligment/Map) - to pliki wynikowe programów przyrównujących odczyty (w formacie FASTQ) do referencji (genom referencyjny).

Pliki SAM:

- pliki tekstowe
- rozdzielone znakami tabulacji
- informacja o sekwencji (tak jak plik FASTQ), ale zawierają dodatkowe informacje (o mapowaniu każdego odczytu)
- zbudowane z dwóch sekcji
    - nagłówek, sekcja opcjonalna, rozpoczyna się od @
    - aligment, sekcja obligatoryjna, informacje o aligmencie sekwencji, 11 pól (kolumn)
- dokładna specyfikacja na stronie projektu

Pliki BAM

- to skompresowane pliki SAM (ta sama informacja, ale znacznie mniejszy rozmiar)
- pliki BAM często są indeksowane (dodatkowy plik z rozszerzeniem bai)
- wiele narzędzi potrafi korzystać z plików BAM, nie ma potrzeby ich dekompresji do formatu SAM przed wykorzystaniem

Podstawowe narzędzie do pracy z plikami SAM/BAM to program `samtools`.

```
samtools --version

samtools 1.9-216-g4c82d2c
Using htslib 1.9-455-g3f140ad
Copyright (C) 2019 Genome Research Ltd.
```

Informacje zawarte w części nagłówkowej

- W tej części przechowywane są informacje o źródle danych, sekwencji referencyjnej, metodzie aligmentu itp.
- Każda sekcja rozpoczyna się symbolem @ a następnie dwuliterowe oznaczenie rokordu, następnie mamy dwuliterowy tag i wartość
    - @HD Linijka nagłówka
        - VN: wersja formatu
        - SO: kolejność sortowania
    - SQ Słownik sekwencji referencyjnej
        - SN: nazwa sekwencji referencyjnej
        - LN: długość sekwencji referencyjnej
        - SP: gatunek
    - Grupa odczytu
        - ID: identyfikator
        - nazwa centrum sekwencji

        - nazwa próbki
    - Program
        - PN: nazwa programu
        - VN: wersja programu

Informacje zawarte w części nagłówkowej pozwalają na re-procesowanie danych

- Przykładowy nagłówek

```
@HD VN:1.0  SO:unsorted
@PG ID:hisat2   PN:hisat2   VN:2.1.0    CL:"/bioapp/hisat2/hisat2-align-s --
wrapper basic-0 -q --phred33 --dpad 0 --gbar 99999999 --mp 1,1 --np 1 --n-ceil
L,0,0.15 --no-mixed --no-discordant --rna-strandness RF -p 38 -k 10 -x
/home/bartek/Lupin_seq/reference/GCA_001865875.1_LupAngTanjil_v1.0_genomic.fna -
S BAM/UP1.sam -1 clean_data1/UP1_1.PE.fastq -2 clean_data1/UP1_2.PE.fastq"
@PG ID:samtools PN:samtools PP:hisat2   VN:1.9-216-g4c82d2c CL:samtools view -H
UP1.bam
```

Informacje zawarte w części aligment

- Kolumny (pola)

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[10]
4	POS	Int	$[0, 2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

- FLAG - jak najwięcej informacji o mapowaniu w postaci jednej liczby

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

- Znaczenie poszczególnych flag (liczb) można sprawdzić na stronie.
- CIGAR string

Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

- Przykładowy alignement

```
A00718:124:HW5VNDSXX:4:1101:9733:1031   83  CM007367.1  6766116 60  151M    =
6765056 -339
CCGTGCATAAGCCTGCCCCAAAAGGTTATCGAGGGCCTAGAGTTAAAAAGAATCTTTTTAGCCCTCCTCCCCCTCGGTAT
TGATGGAAGTAGTTATCTTCTTGATCACATGTTCAAATGGTGTCTTCTTCCCTATCTCAGATCAATCAGNA
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF#F AS:i:-2
ZS:i:-3 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:32A116C1   YS:i:-1 YT:Z:CP
XS:A:+  NH:i:1
A00718:124:HW5VNDSXX:4:1101:9733:1031   163 CM007367.1  6765056 60  11M872N140M
=   6766116 339
TCAAGCTTTAGGGAGTGAAGTGAATGTTGCTGGAAAGCTTGTAAAGGAAGAACTTGCAAAGACAAACCTCAAAGTTCAAG
ATGATGAAAAAGCCAAATCTAAAGGATTTTTTGATCAGAAGCCAATTTTTAAGTGGCCATTTATCAAACCT
FFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF,FFFFFFFFFFFF:,FFFFFF,FFFFFFF
FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFF,F:FFFFFFFF:F:F AS:i:-1
ZS:i:-13    XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:138G12 YS:i:-2 YT:Z:CP
XS:A:+  NH:i:1
```

Do kompresji plików SAM w pliki BAM możemy wykorzystać program `samtools`

```
samtools view -Sb -@ 2 plik.sam > plik.bam
```

Indeksowanie plików BAM również wykonujemy za pomocą programu `samtools`. W pierwszej kolejności musimy posortować plik BAM

```
samtools sort plik.bam -o plik.sorted.bam
```

Następnie możemy go zindeksować

```
samtools index plik.sorted.bam
```

Ćwiczenie 1

Proszę utworzyć pliki BAM, dla plików SAM (z poprzednich ćwiczeń). Następnie proszę dla utworzonych plików BAM utworzyć pliki indeksu bam.bai

Ćwiczenie 2

Proszę podać id odczytów mapujących się do dwóch różnych eksonów (pliki UP)