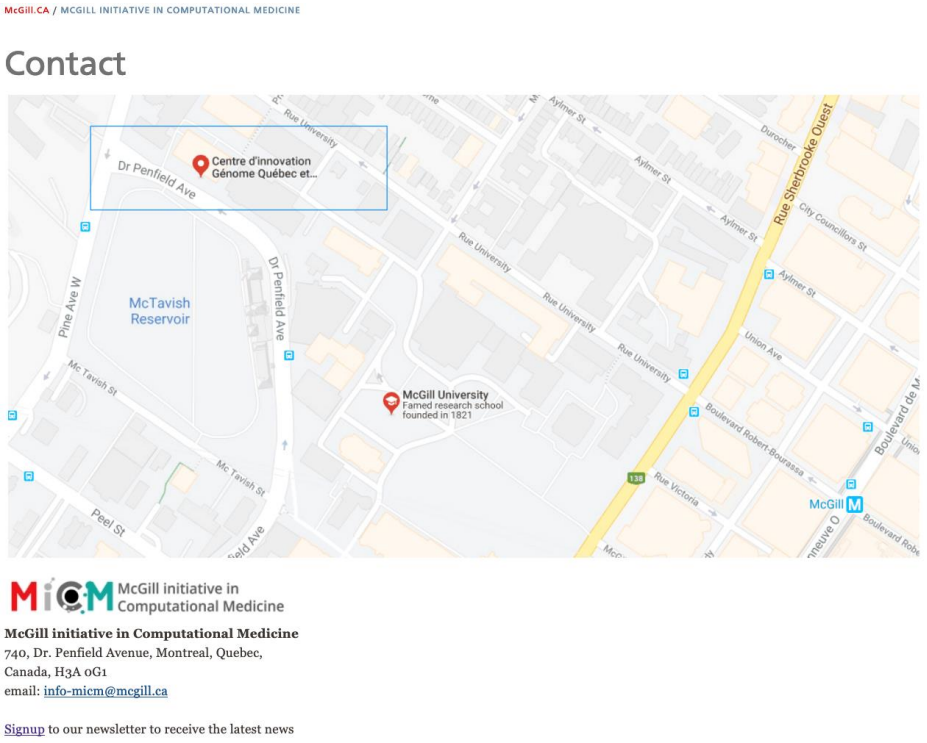
Introduction to RNA-seq formats

December 1st, 2022

Audrey Baguette
Rached Alkallas
Georgette Femerling

Mission : aims to deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery



https://www.mcgill.ca/micm

Overview

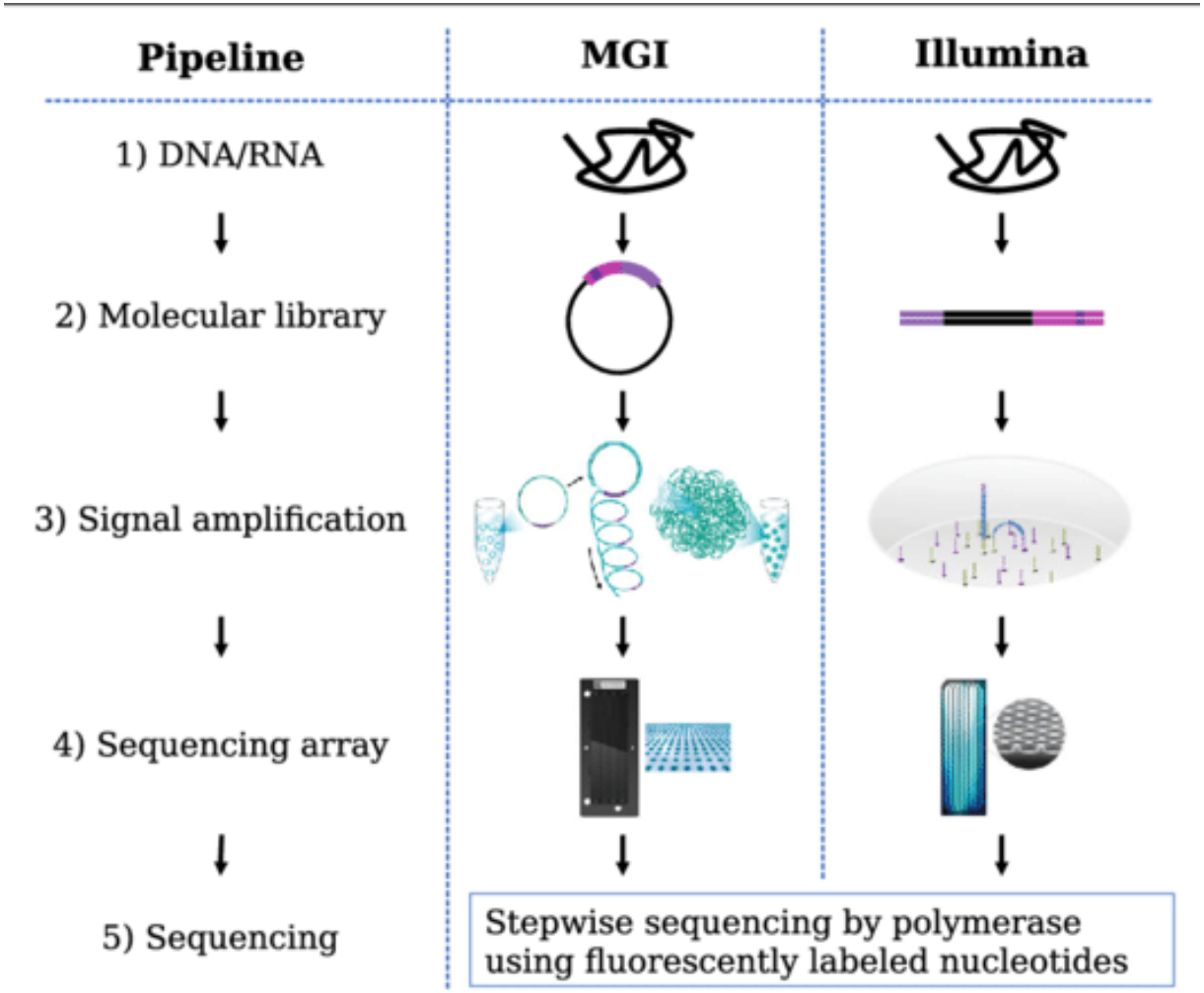
- Introduction (10 min)
  - NGS platforms
  - Single End vs Paired End sequencing
  - Stranded vs unstranded library prep
  - Adapters and PCR
- Raw sequence files: fasta and fastq (15 min)
  - Fasta vs fastq: what is the difference?
  - Decoding fastq quality scores
  - Hands on: Cutting a read at Q30 (5 min)

Overview

- Aligning reads (25 min)
  - How to choose the reference?
  - Downloading a reference genome
  - SAM vs BAM format
  - Hands on: converting between formats (5 min)
- Files for genomic regions analysis (30 min)
  - Wig and bigWig
  - bedGraph
  - Bed and bigBed
  - Gtf
  - Formats along the genome
  - Liftover to change reference
  - Hands on: Lifting genes with the liftover tool (5 min)

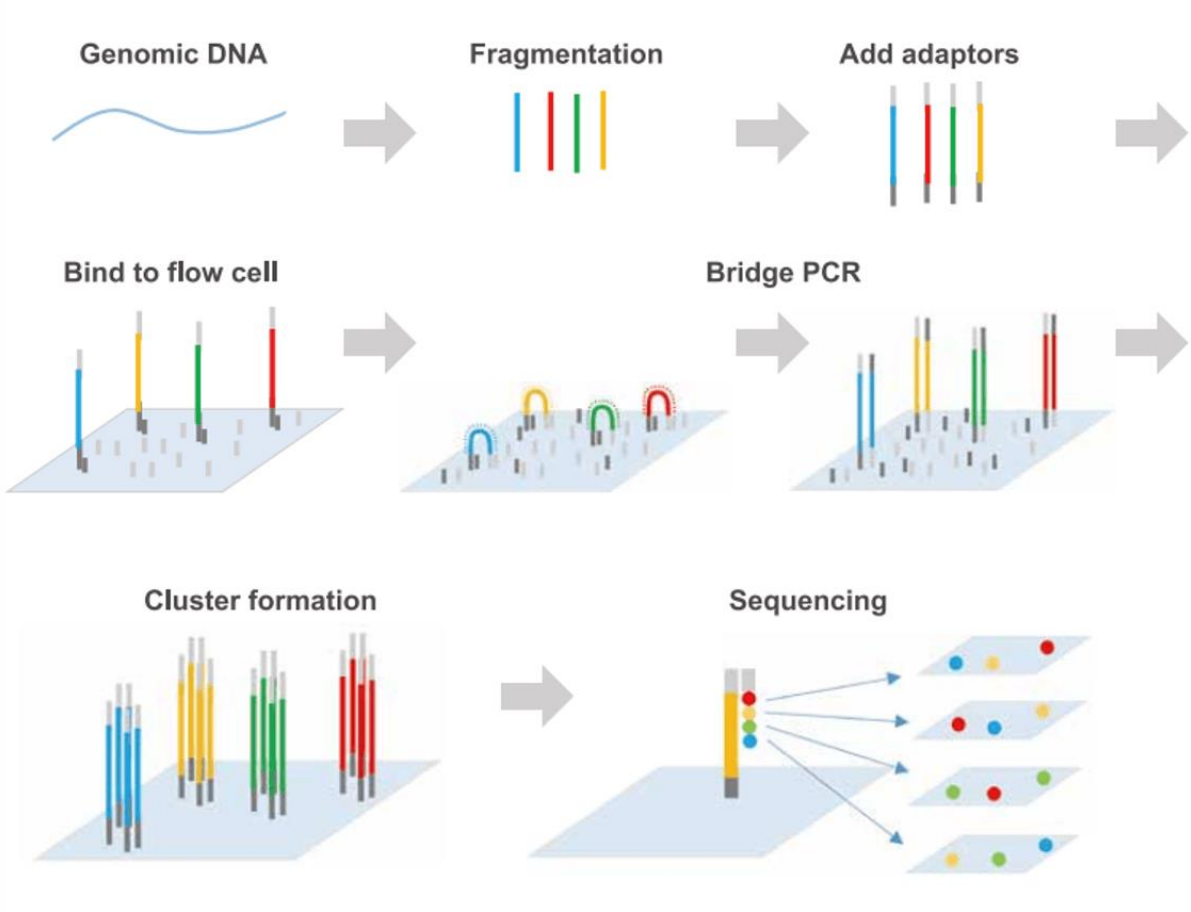
Introduction

NGS platforms



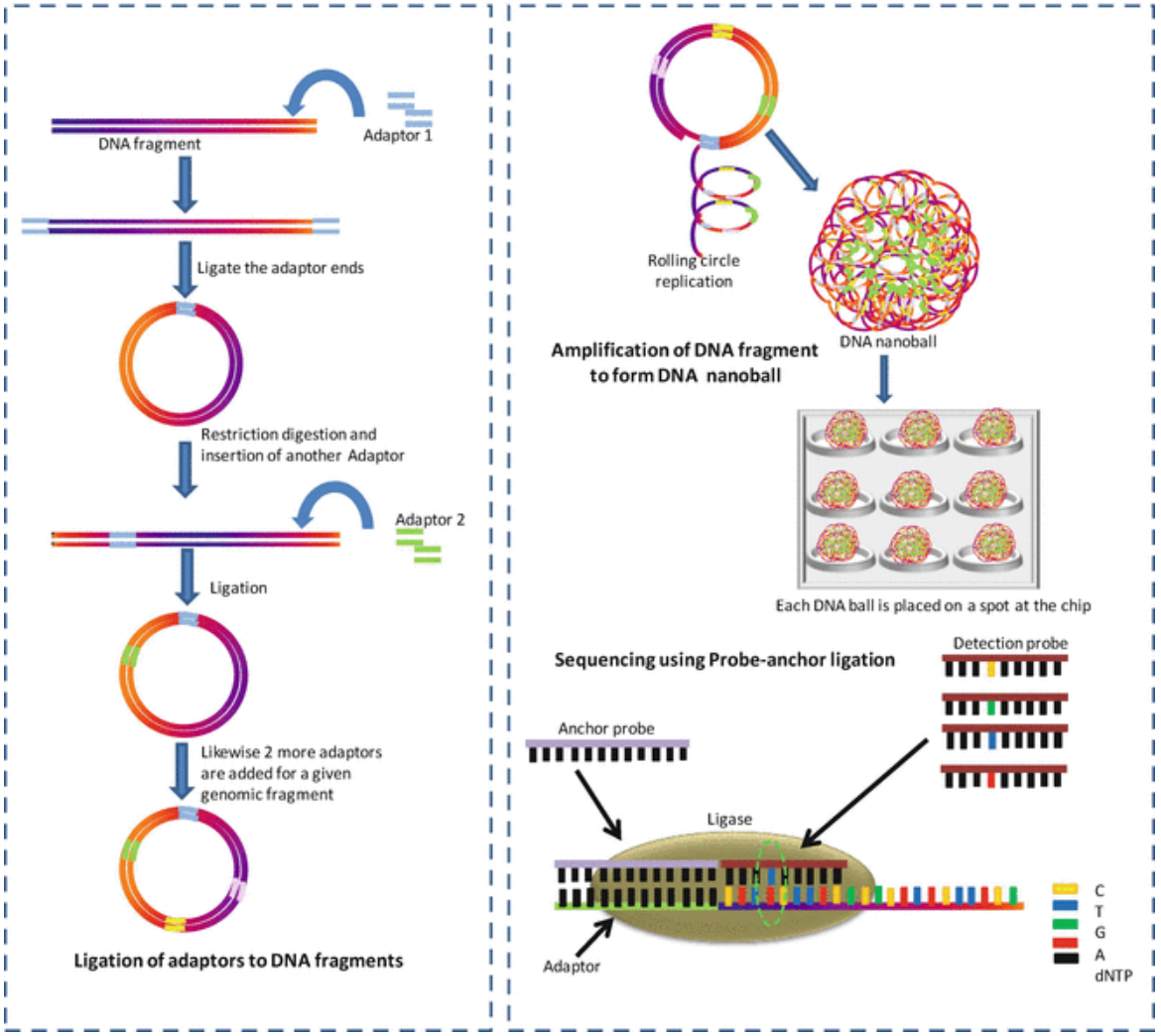
https://www.researchgate.net/figure/Technical-comparison-of-DNBSeq-and-Illumina-platforms_fig1_334652496

Illumina



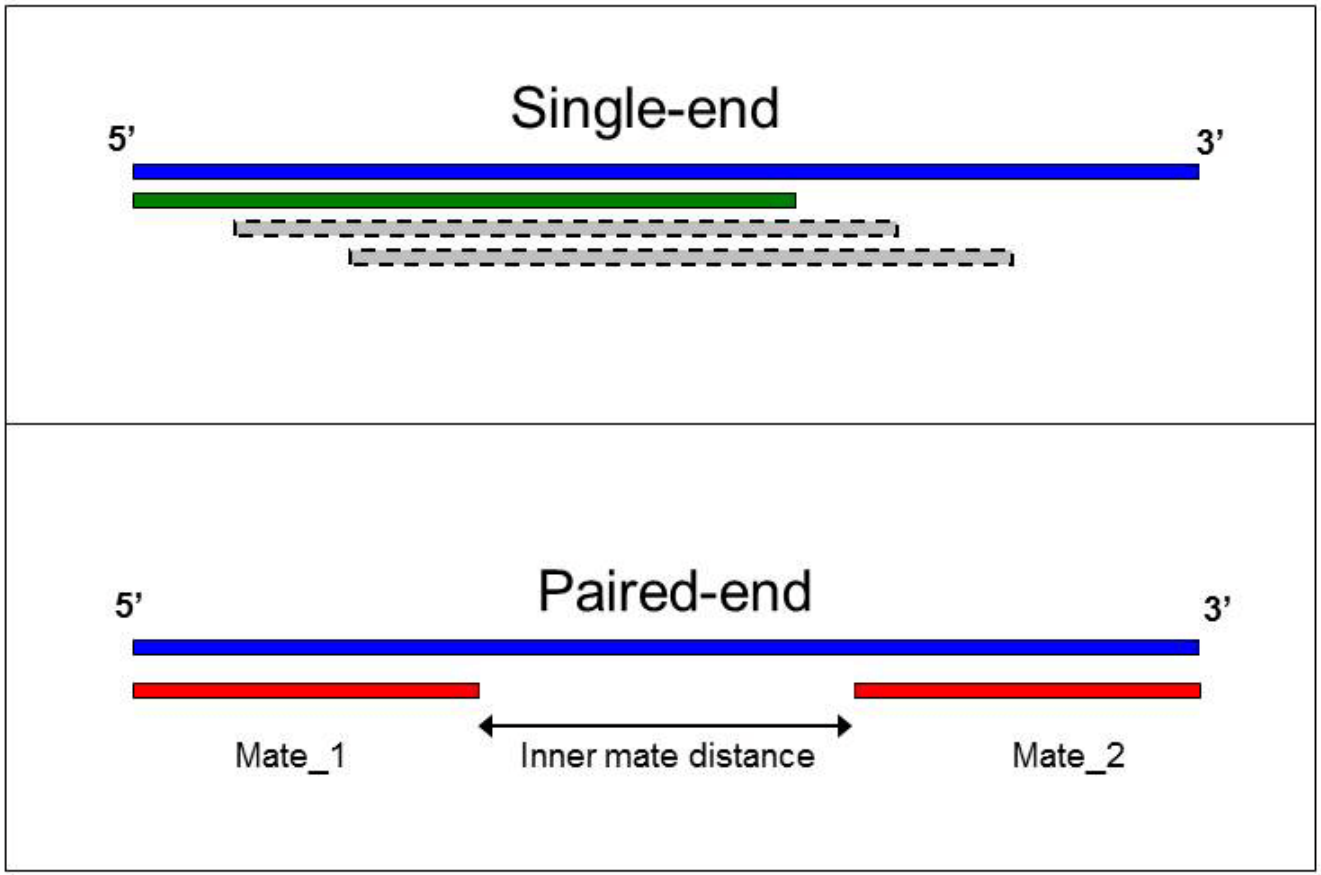
https://blogs.iu.edu/ncgas/2021/09/17/introduction-to-illumina-sequencing/

MGI (nanoball)



https://link.springer.com/protocol/10.1007/978-1-4939-6622-6_1

Single End vs Paired End sequencing



https://www.researchgate.net/figure/Sequencage-single-end-et-paired-end-Dans-le-premier-les-fragments-sont-sequences-a_fig9_305320151

Stranded vs unstranded library prep

2 reads align in a region but with a different orientation

-> same transcript, but generated during reverse transcription?

-> two overlapping genes

Solution: have different adapters for 3’ and 5’

Raw sequence files: fasta and fastq

Fasta VS fastq: what is the difference?

FASTA:

- Text file
1. Name of the sequence, generally starts with '>'
2. Sequence
- May contain nucleotides or amino acids
- May have a related index file (.fai)

```
cat dna.fasta
cat 1HV4
```

.fa
.fasta
.txt
Ø

.gz (.fa.gz, .txt.gz, …)

Fasta VS fastq: what is the difference?

FASTQ:

- Text file
- 4 lines per sequence (read)
  1. Name of the sequence, starts with '@'
  2. Sequence
  3. Optional description, starts with '+'
  4. Quality scores

.fq
.fastq
.txt
Ø

.gz (.fq.gz, .fastq.gz, …)

```
head left_ventricle_34m_100_rep1_R1.fastq
head left_ventricle_34m_100_rep1_R2.fastq
```

Decoding fastq quality scores

$$Q_{phred} = -10log_{10}(p)$$

p: probability of a base to be wrong

Q_{phred}: Phred quality score

Q_{phred} + 33 -> ASCII code of symbol

Examples:

p = 0.05 -> Q_{phred} = 13 -> ASCII code = 46 -> symbol = .

symbol = ? -> ASCII code = 63 -> Q_{phred} = 30 -> *p* = 0.001

Hands on (5 min)
Cutting a read at Q33

Where would we cut (the beginning and end of) the first 3 reads of `left_ventricle_34m_100_rep1_R1.fastq`

with a Q-score of 33?

Hints:

Show the first 3 reads with

`head –n 12 left_ventricle_34m_100_rep1_R1.fastq`

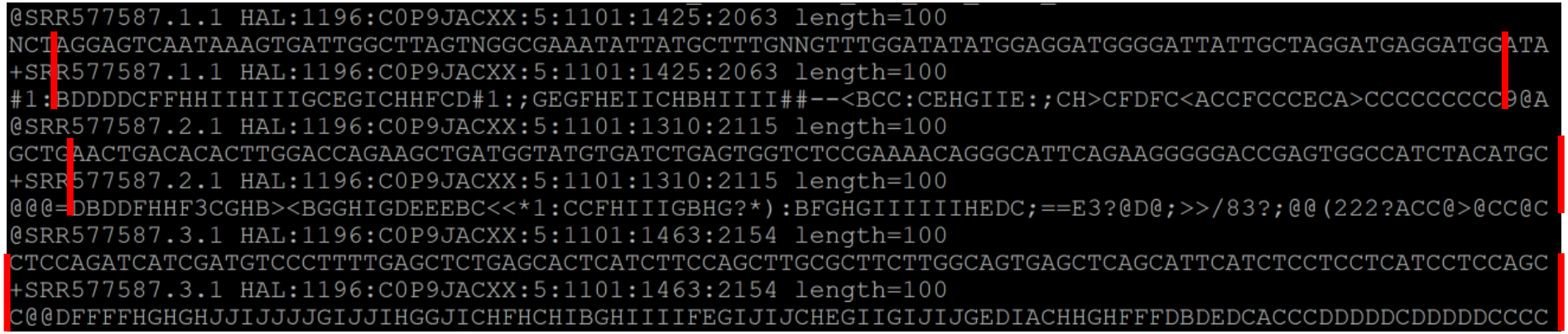
Find the ASCII scores at

https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

Hands on
Cutting a read at Q33

ANSWER

Q33: B -> we remove all starting and ending bases until re reach a B (or above)



Aligning reads

How to choose the reference?

- What is a reference?
  - Genomic coordinates
  - Complete
  - Multiple chromosomes and unresolved contigs
  - Haploid

- Different references
  - Organism (mouse, human, …)
  - Consortium (GRCm, GRCh, mm, hg, …)
  - Version (mm9, mm10, hg19, hg38)

How to choose the reference?

- Elements influencing the choice
  - Completeness
  - Quality of the assembly
  - Reproducibility

Downloading a reference genome

- UCSC Genome Browser

https://hgdownload.soe.ucsc.edu/downloads.html

- Ensembl

https://useast.ensembl.org/

SAM vs BAM format

SAM

.sam

- Aligned reads
- Human readable
- Big file
- Header contains all chromosomes, contigs, etc. and their lengths + the command(s) used to create the file

```
more left_ventricle_34m_chr11.sam
```

SAM vs BAM format

BAM

- Aligned reads
- Binary file
- Smaller file than SAM

.bam

```
ls -lh left_ventricle_34m_chr11.bam
ls -lh left_ventricle_34m_chr11.sam
```

```
[aubag1@workshop2021a Data]$ ls -lh left_ventricle_34m_chr11.*
-rw-r----- 1 aubag1 aubag1 344M Nov 23 18:16 left_ventricle_34m_chr11.bam
-rw-r----- 1 aubag1 aubag1 1.6G Nov 24 11:35 left_ventricle_34m_chr11.sam
```

CIGAR, MAPQ and Sam flags

SAM mandatory fields

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[11]
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

https://samtools.github.io/hts-specs/SAMv1.pdf

CIGAR – alignment details

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

SRR577589.525003.1 355 chr11 5253278 3 100M = 5253347 16217
CAGTGGTATCTGGAGGACAGGGCACTGGCCACTCCAGTCACCATCTTCTGCCAGGAAGCCTGCACCTCAGGGGTGAATTCTTTGCCGA
AATGGATTGCCA
BCCDFFEDHHHHHJJJJJJJJJJJJJJJJJIJIJIJJIJJJJJJJJJJJIJJJJJJJIGIIJJIJHHHHHDFF>>ACEEDEEDDDDDDBDDDDDDDDDDD NH:i:2
HI:i:2 AS:i:197 NM:i:0 MD:Z:100
SRR577589.525003.1 147 chr11 5253347 3 59M886N41M = 5253278 -1055
GGGGTGAATTCTTTGCCGAAATGGATTGCCAAAACGGTCACCAGCACATTTCCCAGGAGCTTGAAGTTCTCAGGATCCACATGCAGCT
TGTCACAGTGCA
DBDDDDCDDDDDDDDDEEEFFFFFFHHHHHJJIGHJJJJIHJIIIIIJIHFJJJIJJIJJJJJJJJJJJJIIIIJJJJHJJJJJJJJHHHHHHFFFFFCCCNH:i:2 HI:i:1
AS:i:198 NM:i:0 MD:Z:100

https://samtools.github.io/hts-specs/SAMv1.pdf

MAPQ – alignment quality

$$MAPQ = -10log_{10}(p)$$

If 255: mapping quality unavailable

SRR577589.525003.1 355 chr11 5253278 3 100M = 5253347 16217
CAGTGGTATCTGGAGGACAGGGCACTGGCCACTCCAGTCACCATCTTCTGCCAGGAAGCCTGCACCTCAGGGGTGAATTCTTTGCCGA
AATGGATTGCCA
BCCDFFEDHHHHHJJJJJJJJJJJJJJJJJIJIJIJJIJJJJJJJJJJJJIJJJJJJJIGIIJJIJHHHHHDFF>>ACEEDEEDDDDDDBDDDDDDDDDDD NH:i:2
HI:i:2 AS:i:197 NM:i:0 MD:Z:100
SRR577589.525003.1 147 chr11 5253347 3 59M886N41M = 5253278 -1055
GGGGTGAATTCTTTGCCGAAATGGATTGCCAAAACGGTCACCAGCACATTTCCCAGGAGCTTGAAGTTCTCAGGATCCACATGCAGCT
TGTCACAGTGCA
DBDDDDCDDDDDDDDDEEEFFFFFFHHHHJJIGHJJJJIHJIIIIIJIHFJJJJIJJIJJJJJJJJJJJJIIIIJJJJHJJJJJJJJHHHHHHFFFFFCCCNH:i:2 HI:i:1
AS:i:198 NM:i:0 MD:Z:100

MAPQ = 3 -> P(mismapping) = 10^(3/-10) = 0.5

FLAGS

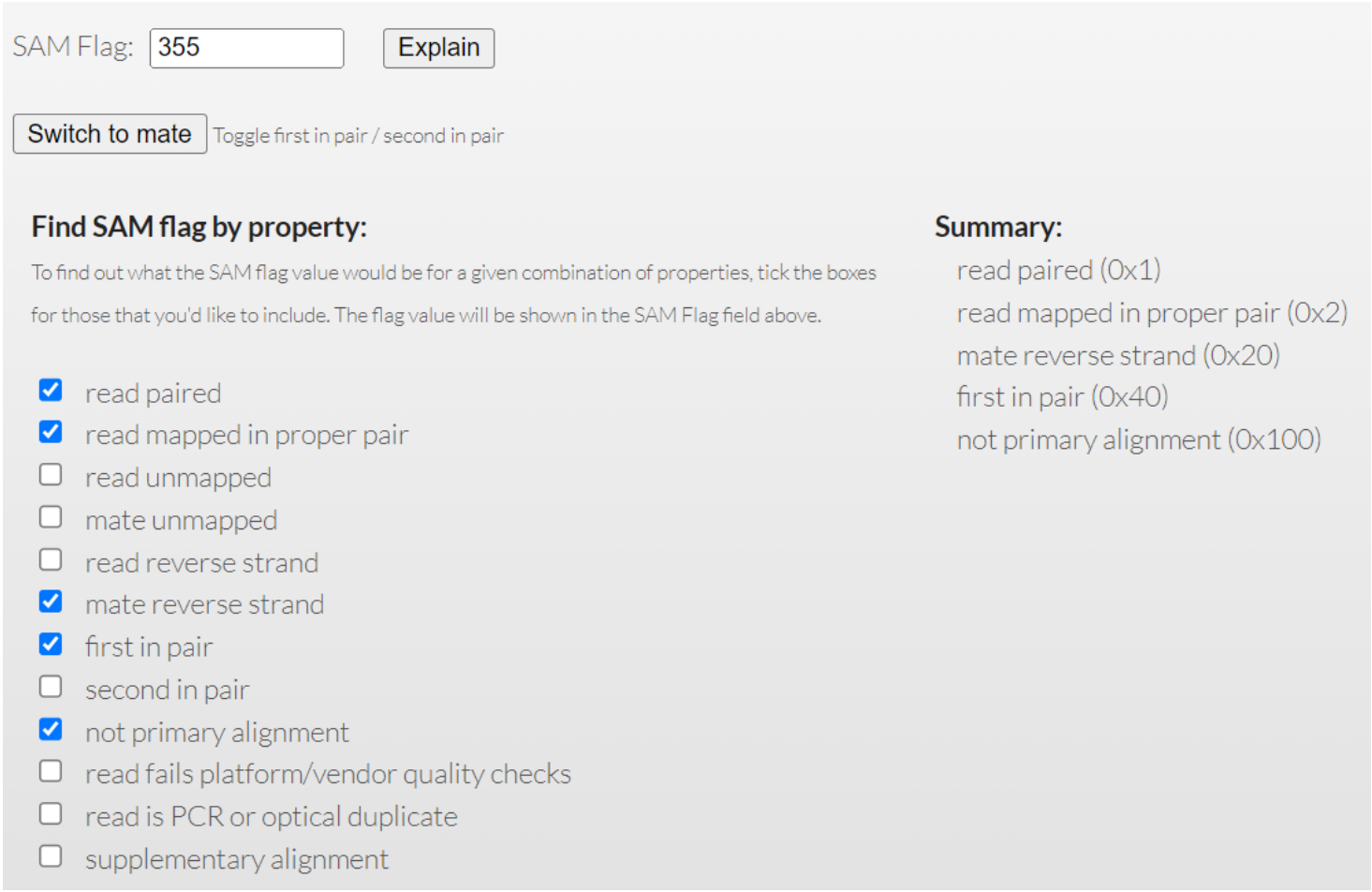
Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
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

SRR577589.525003.1 355 chr11 5253278 3 100M = 5253347 16217
CAGTGGTATCTGGAGGACAGGGCACTGGCCACTCCAGTCACCATCTTCTGCCAGGAAGCCTGCACCTCAGGGGTGAATTCTTTGCCGA
AATGGATTGCCA
BCCDFFEDHHHHHJJJJJJJJJJJJJJJJJIJIJIJJIJJJJJJJJJJJIJJJJJJJIGIIJJIJHHHHHDFF>>ACEEDEEDDDDDDBDDDDDDDDDDD NH:i:2
HI:i:2 AS:i:197 NM:i:0 MD:Z:100
SRR577589.525003.1 147 chr11 5253347 3 59M886N41M = 5253278 -1055
GGGGTGAATTCTTTGCCGAAATGGATTGCCAAAACGGTCACCAGCACATTTCCCAGGAGCTTGAAGTTCTCAGGATCCACATGCAGCT
TGTCACAGTGCA
DBDDDDCDDDDDDDDDEEEFFFFFFHHHHHJJIGHJJJJIHJIIIIIJIHFJJJJIJJIJJJJJJJJJJJJIIIIJJJJHJJJJJJJJHHHHHHFFFFFCCCNH:i:2 HI:i:1
AS:i:198 NM:i:0 MD:Z:100

https://samtools.github.io/hts-specs/SAMv1.pdf

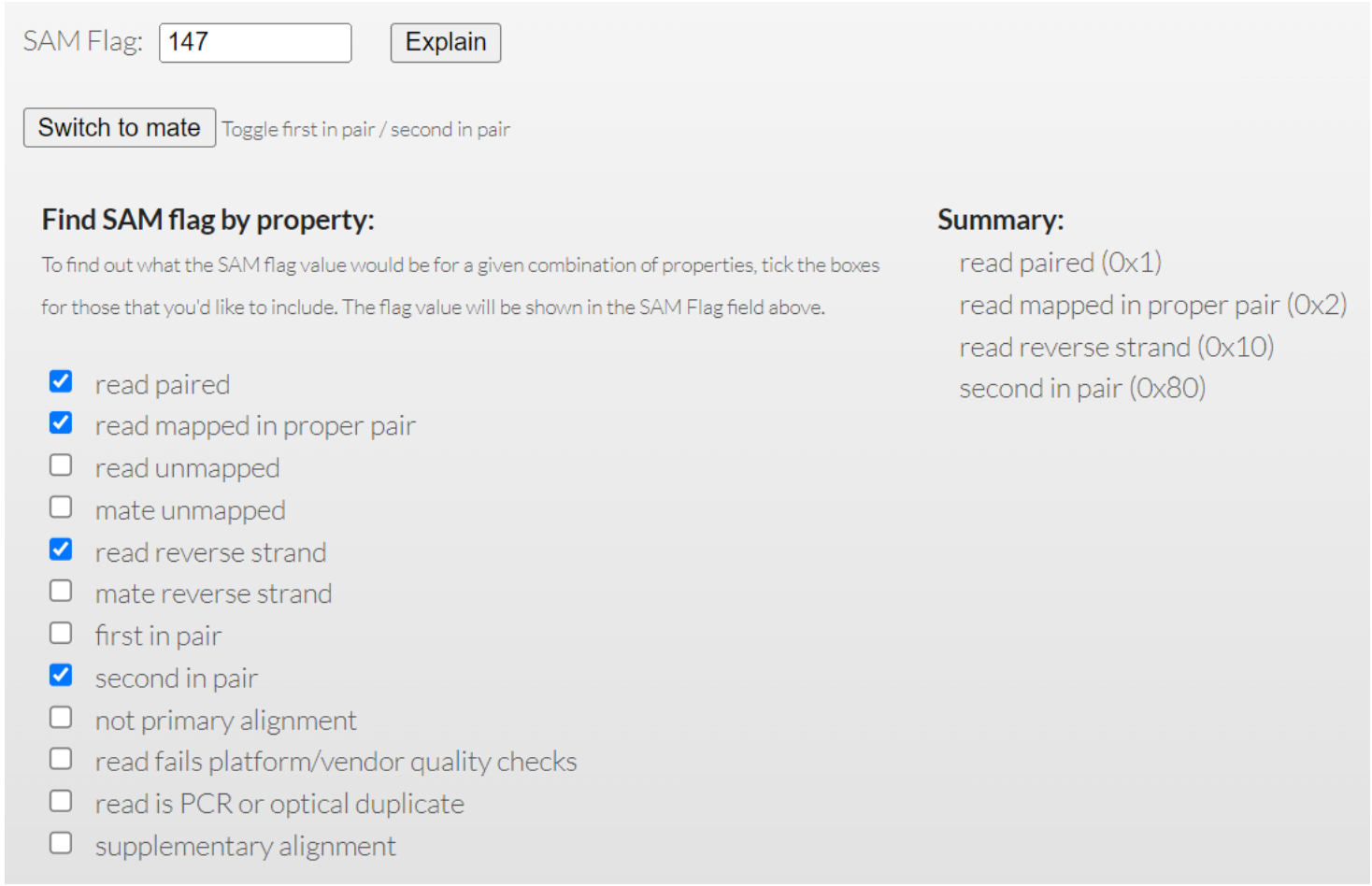
FLAGS

https://broadinstitute.github.io/picard/explain-flags.html



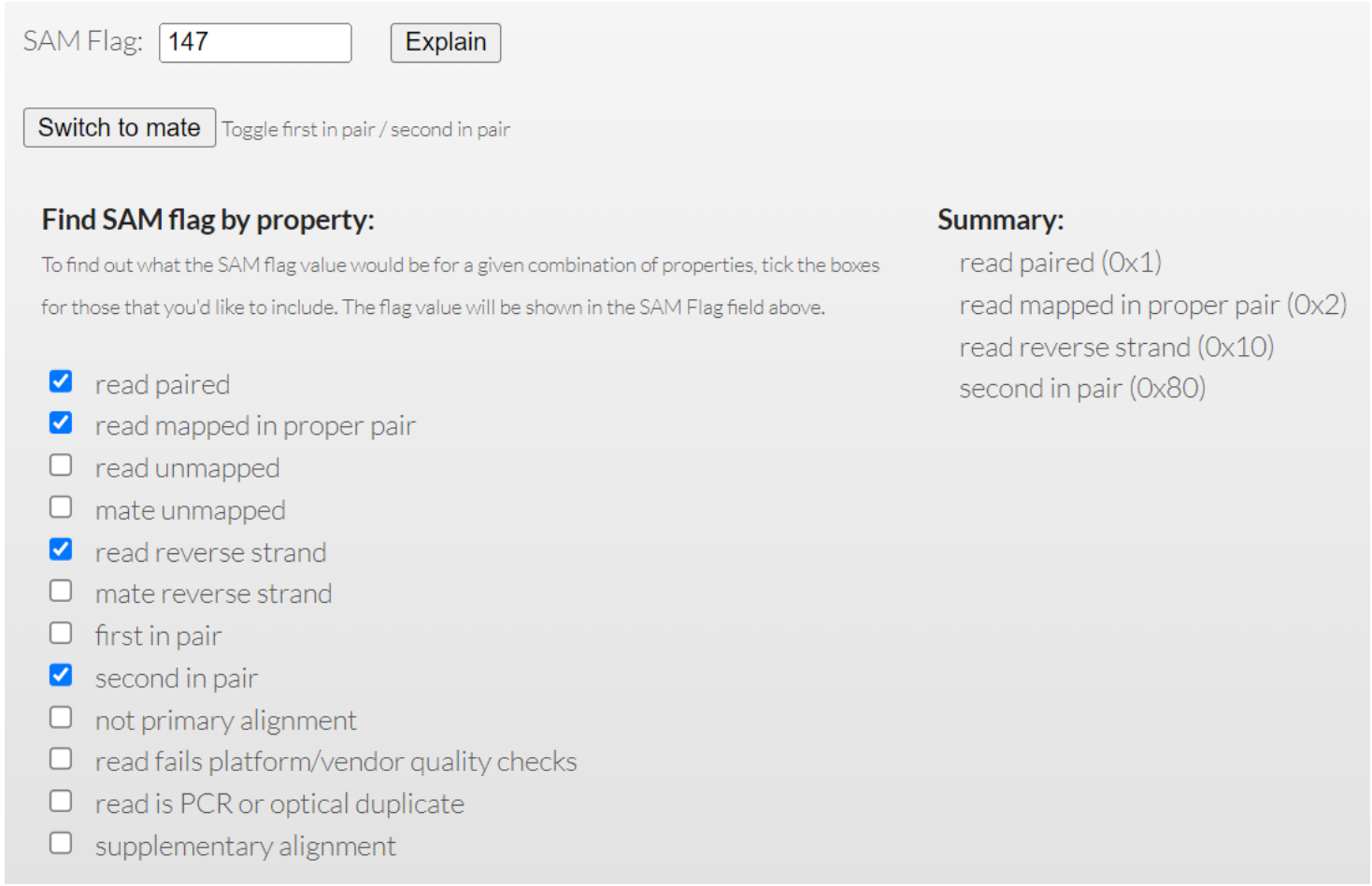
FLAGS

https://broadinstitute.github.io/picard/explain-flags.html



FLAGS

https://broadinstitute.github.io/picard/explain-flags.html



Hands on (5 min)
Converting between formats

Convert the bam file to a sam file. Compare the sizes

```
module load StdEnv/2020 samtools/1.16.1
1. samtools view –h –o sam_file bam_file
2. ls -lh
```

Optional: subset the sam/bam file to contain only region chr11:5240000-5260000

```
samtools index bam_file
samtools view [options] file region
```

Hands on
Converting between formats

ANSWER

```
samtools view -h -o
left_ventricle_34m_chr11.sam
left_ventricle_34m_chr11.bam

samtools index left_ventricle_34m_chr11.bam

samtools view -h -o
left_ventricle_34m_chr11_subset.sam
left_ventricle_34m_chr11.bam chr11:5240000-
5260000
```

Files for genomic regions analysis

Wig and bigWig

wig (wiggle format)

.wig

- Plot quantitative data along the genome
- Fixed or variable step
- Variable format (header specifies variableStep/fixedStep, chrom, start, step)*

```
fixedStep  chrom=chrN
start=position  step=stepInterval
[span=windowSize]
  dataValue1
  dataValue2
  ... etc ...
```

```
variableStep  chrom=chrN
[span=windowSize]
  chromStartA  dataValueA
  chromStartB  dataValueB
  ... etc ...  ... etc ...
```

*when converting bedGraph -> bigWig -> wig,

it has the same format as a bedGraph

bigWig

.bigwig
.bw

- Binary file

bedGraph

bedGraph

- Plot quantitative data along the genome
- Fixed or variable step
- Fixed format (chrom start end value)

.bedGraph

```
[aubag1@workshop2021a Data]$ head left_ventricle_34m_minus.bedGraph
chr1    13129   13229   0.00092
chr1    13244   13344   0.0046
chr1    13344   13444   0.00092
chr1    13463   13476   0.0046
chr1    13476   13479   0.0092
chr1    13479   13529   0.01073
chr1    13529   13531   0.01533
chr1    13531   13563   0.01686
chr1    13563   13575   0.01226
chr1    13575   13579   0.00766
```

Bed and bigBed

.bed

bed

- Represents genomic regions
- Minimum 3 columns (chrom start end)
- BED6: (BED3 name score strand)
- BED12: (BED6 thickStart thickEnd itemRgb

Start codon

End codon

blockCount blockSizes blockStarts)

exons

Sizes of blocks (;)

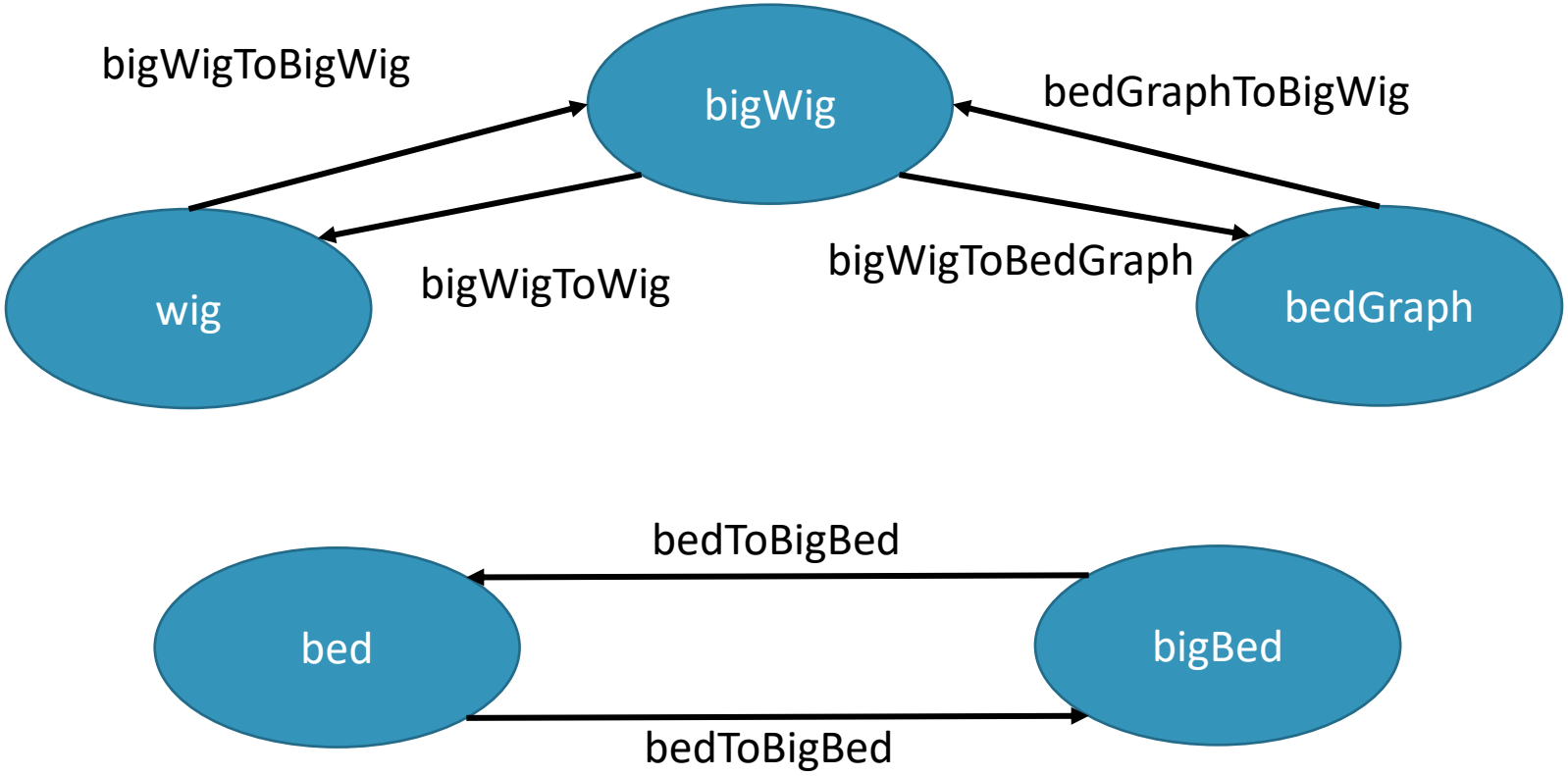
Starts of blocks (;)

.bigBed
.bb

bigBed

- Binary file

Converting between formats



Gtf

1. Seqname (chromosome)
2. Source
3. Feature
4. Start
5. End
6. Score
7. Strand
8. Frame (0: first base is start of codon, 1: second base is start of codon, 2: third base is start of codon)
9. Attribute

BED vs GTF

BED6

```
chr11   75779   76143   ENST00000519787.1       7       +
chr11   86648   87586   ENST00000424047.1       7       -
chr11   112966  125927  ENST00000622626.1       3       -
```

GTF

```
chr11   knownGene       transcript      112967  125927  .       -
chr11   knownGene       exon    112967  113111  .       -       .
chr11   knownGene       exon    113116  113174  .       -       .
```

Formats along the genome

fastq

fasta

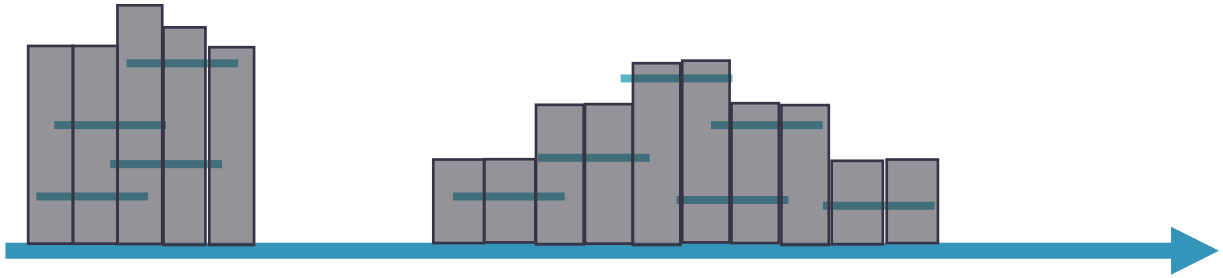
Formats along the genome

sam/bam

bed/bigBed

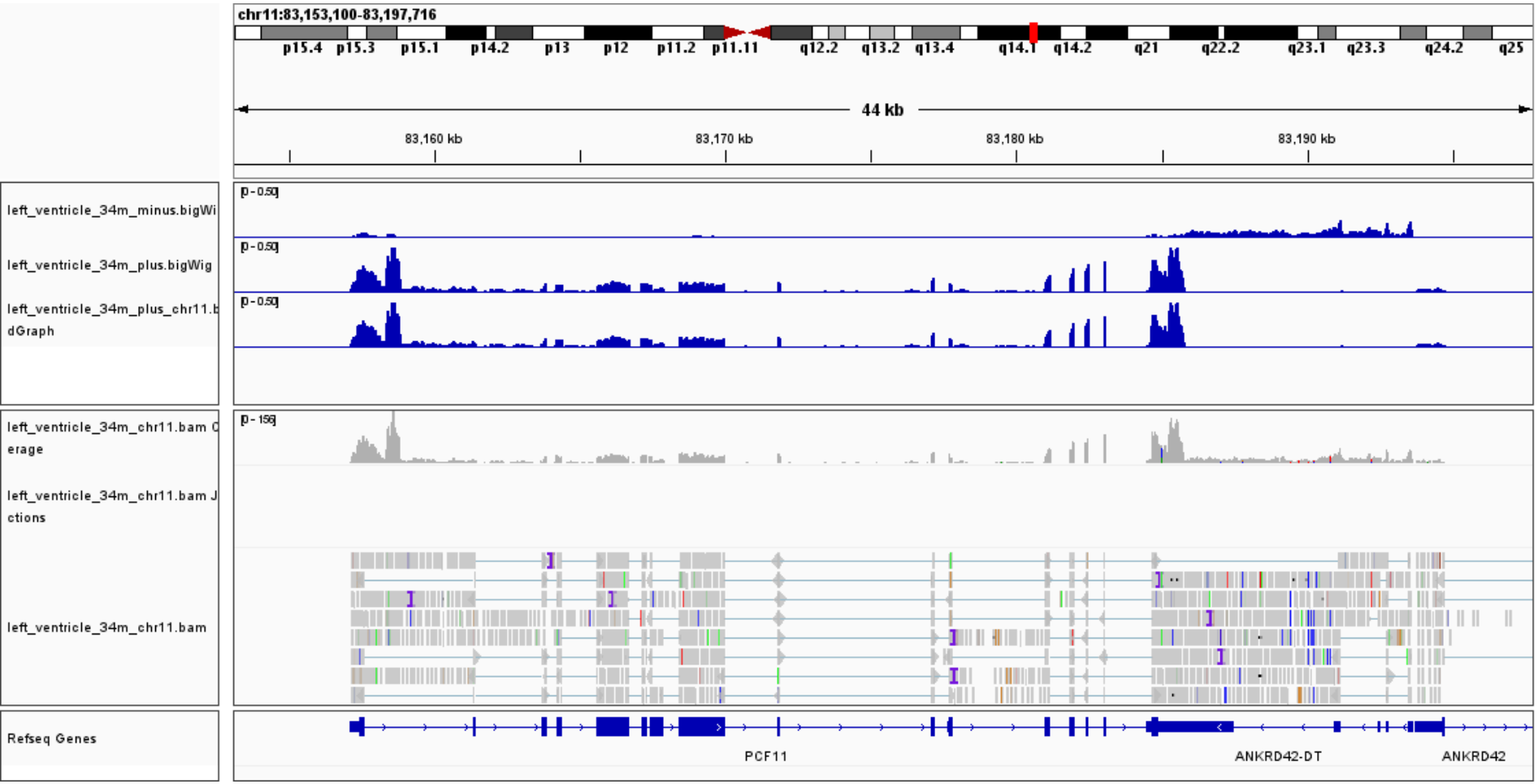
Formats along the genome

bedGraph
wig/bigWig

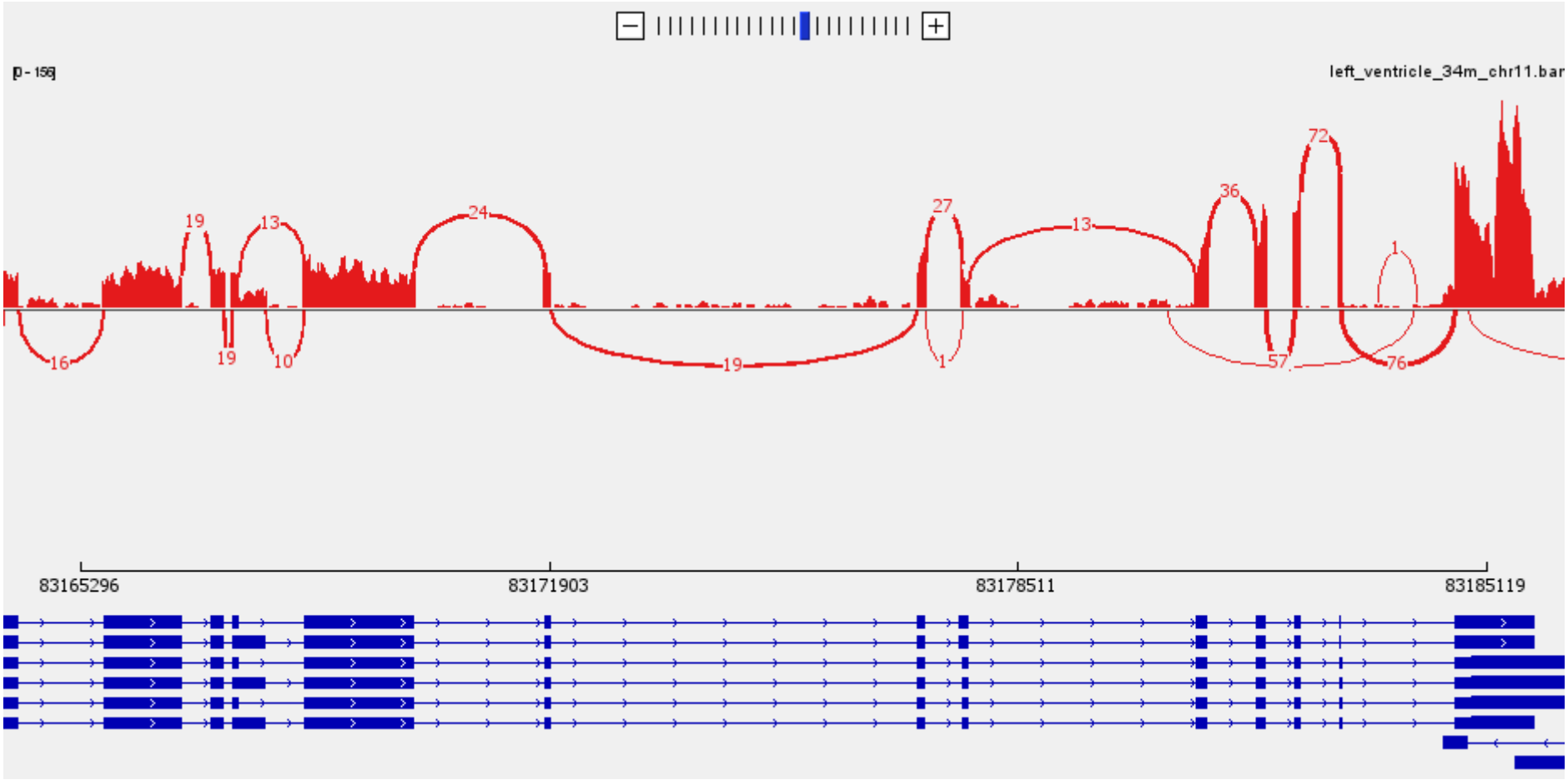


bed/bigBed

Formats along the genome - IGV



Sashimi plots - IGV



Liftover to change reference

- Changes the genomic coordinates between assemblies
- Across version or across species
- Alternative to reprocessing

Liftover to change reference

Liftover tool

- ✓ Quick and easy
- ✓ Good for well-characterized, conserved regions
- X Imperfect, less precise
- X Some regions have conflicts (split)
- X Dependent on format
- • RNA-seq, ChIP-seq

Reprocessing

- X Can be long
- ✓ Works every time
- ✓ Harmonizes processing
- • SNPs, Hi-C

Hands on (5 min)
Lifting genes with the liftover tool

- Lift the positions of (some) chr11 genes over to another assembly/organism
- What are the results? How many are lost

Subset the first columns of the bed file

```
cut –f1-3 genes_hg38_chr11.bed > out.bed
```

Copy the first few lines of the file OR download it

https://genome.ucsc.edu/cgi-bin/hgLiftOver

Hands on:
Lifting genes with the liftover tool

ANSWER

Taking the first 10 genes…

-> hg19: all genes are transposed

-> T2T: all genes are transposed

-> mm10: one gene cannot be transposed (sequence does not exist)

-> susScr11 (pig): one gene cannot be transposed

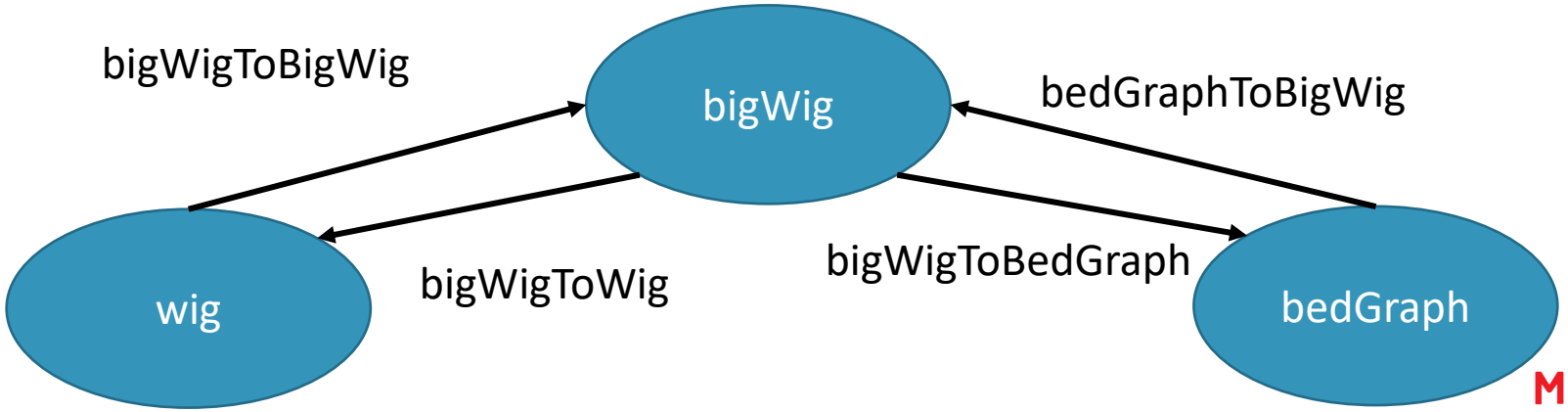
Bonus exercise

Hands on:
Subset the bigwig file

The bigwig file cannot be directly subsetted. We must go through the wig or bedGraph format.

Subset left_ventricle_34m_plus.bigWig, to keep chr11 only, then re-convert to bigWig

```
module load mugqic/kentUtils/302.1.0
grep chr11 file
fetchChromSizes hg38 > hg38.chromsizes
```



Hands on:
Subset the bigwig file

ANSWER

```
bigWigToBedGraph left_ventricle_34m_plus.bigWig
left_ventricle_34m_plus.bedGraph

grep chr11 left_ventricle_34m_plus.bedGraph
> left_ventricle_34m_plus_chr11.bedGraph

fetchChromSizes hg38 > hg38.chromsizes

bedGraphToBigWig
left_ventricle_34m_plus_chr11.bedGraph
hg38.chrom.sizes
left_ventricle_34m_plus_chr11.bigWig
```