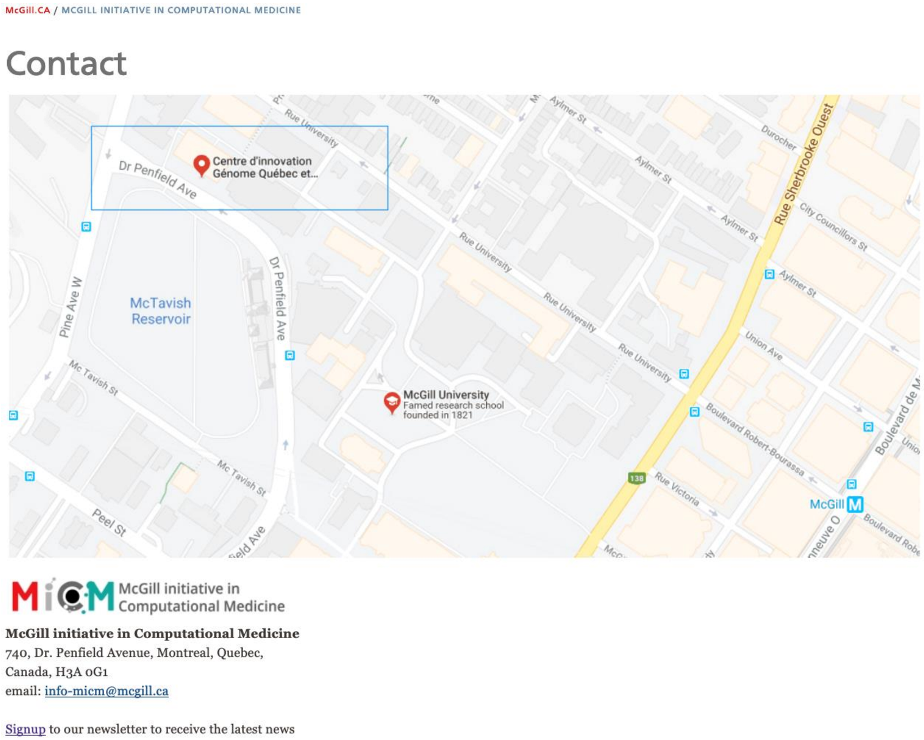
Introduction to RNA-seq formats

December 1st, 2022
Instructor: Audrey Baguette
TA: Rached Alkallas

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



McGill.CA / MCGILL INITIATIVE IN COMPUTATIONAL MEDICINE

Contact

McGill initiative in
Computational Medicine

McGill initiative in Computational Medicine
740, Dr. Penfield Avenue, Montreal, Quebec,
Canada, H3A 0G1
email: info-micm@mcgill.ca

Signup to our newsletter to receive the latest news

https://www.mcgill.ca/micm

Overview

- Raw sequence files: fasta and fastq (25 min)
  - Fasta vs fastq: what is the difference?
  - Decoding fastq quality scores
  - Hands on: Cutting a read at Q30 (5 min)
- Aligning reads (20 min)
  - How to choose the reference?
  - 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
  - Liftover to change reference
  - Hands on: Lifting genes with the liftover tool (10 min)

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
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
```

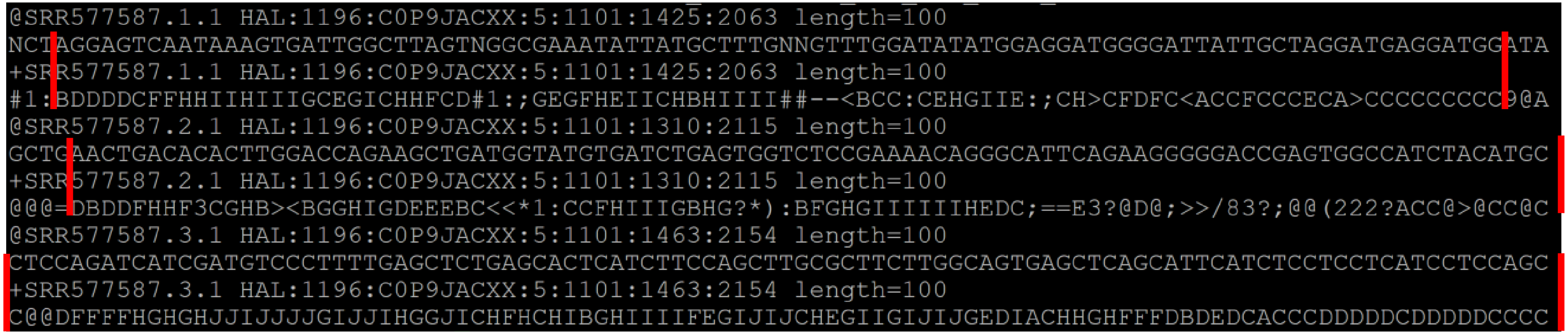
Find the ASCII scores at

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)



```
@SRR577587.1.1 HAL:1196:C0P9JACXX:5:1101:1425:2063 length=100
NCTAGGAGTCAATAAAGTGATTGGCTTAGTNGGCGAAATATTATGCTTTGNNGTTTGGATATATGGAGGATGGGGATTATTGCTAGGATGAGGATGGATA
+SRR577587.1.1 HAL:1196:C0P9JACXX:5:1101:1425:2063 length=100
#1:BDDDDCFFHHIIHIIIGCEGICHHFCD#1:;GEGFHEIICHBHIIII##--<BCC:CEHGIIE:;CH>CFDFC<ACCFCCCECA>CCCCCCCCC9@A
@SRR577587.2.1 HAL:1196:C0P9JACXX:5:1101:1310:2115 length=100
GCTGAACTGACACACTTGGACCAGAAGCTGATGGTATGTGATCTGAGTGGTCTCCGAAAACAGGGCATTCAGAAGGGGGACCGAGTGGCCATCTACATGC
+SRR577587.2.1 HAL:1196:C0P9JACXX:5:1101:1310:2115 length=100
@@@=DBDDFHHF3CGHB><BGGHIGDEEEBC<<*1:CCFHIIIGBHG?*):BFGHGIIIIIIHEDC;==E3?@D@;>>/83?;@@(222?ACC@>@CC@C
@SRR577587.3.1 HAL:1196:C0P9JACXX:5:1101:1463:2154 length=100
CTCCAGATCATCGATGTCCCTTTTGAGCTCTGAGCACTCATCTTCCAGCTTGCGCTTCTTGGCAGTGAGCTCAGCATTCATCTCCTCCTCATCCTCCAGC
+SRR577587.3.1 HAL:1196:C0P9JACXX:5:1101:1463:2154 length=100
C@@DFFFFHGHGHJJIJJJJGIJJIHGGJICHFHCHIBGHIIIIFEGIJIJCHEGIIGIJIJGEDIACHHGHFFFDBDEDCACCCDDDDDCDDDDDCCCC
```

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

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
```

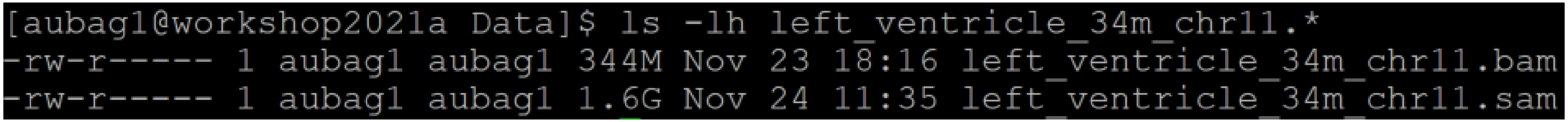
SAM vs BAM format

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
```

Hands on
Converting between formats

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

```
1. samtools index bam_file*
2. samtools view –h –o sam_file
   bam_file
3. ls -lh
```

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

```
Samtools view [options] file region
```

* Was done to subset the file already, not need to do it again

Hands on
Converting between formats

ANSWER

```
#samtools index left_ventricle_34m_chr11.bam

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

 samtools view -bam -o
 left_ventricle_34m_subset.bam
 left_ventricle_34m_chr11.bam chr11:5240000-
 5260000
```

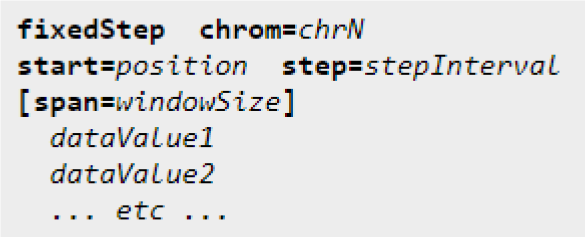
Files for genomic regions analysis

Wig and bigWig

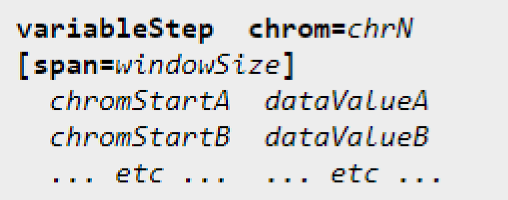
.wig

wig (wiggle format)

- 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

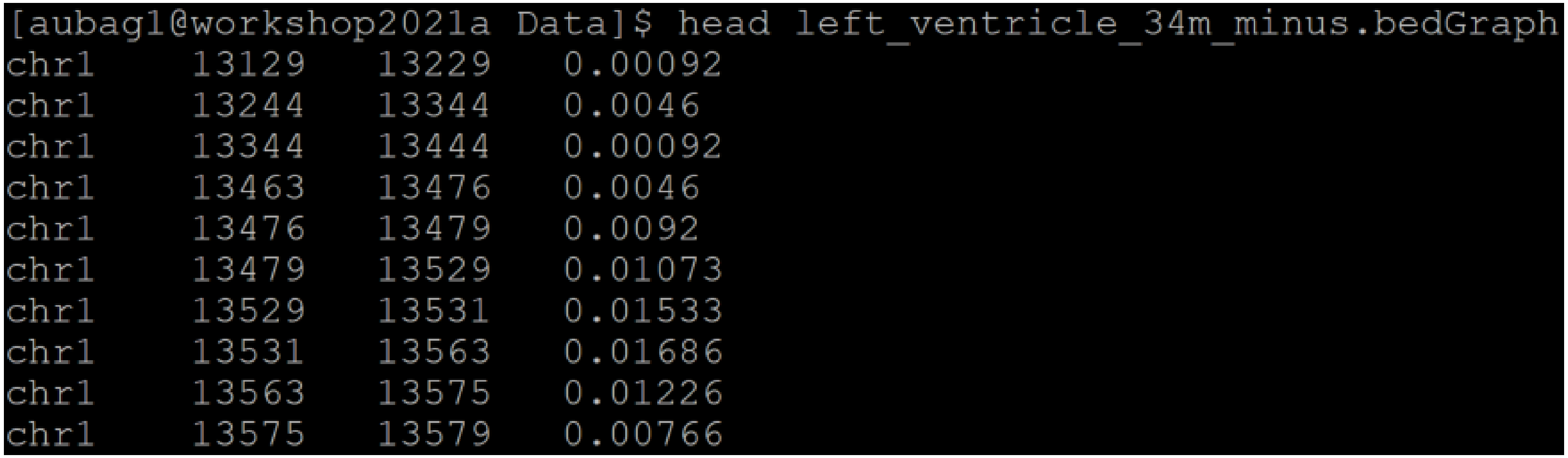
- Binary file

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

.bigBed
.bb

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

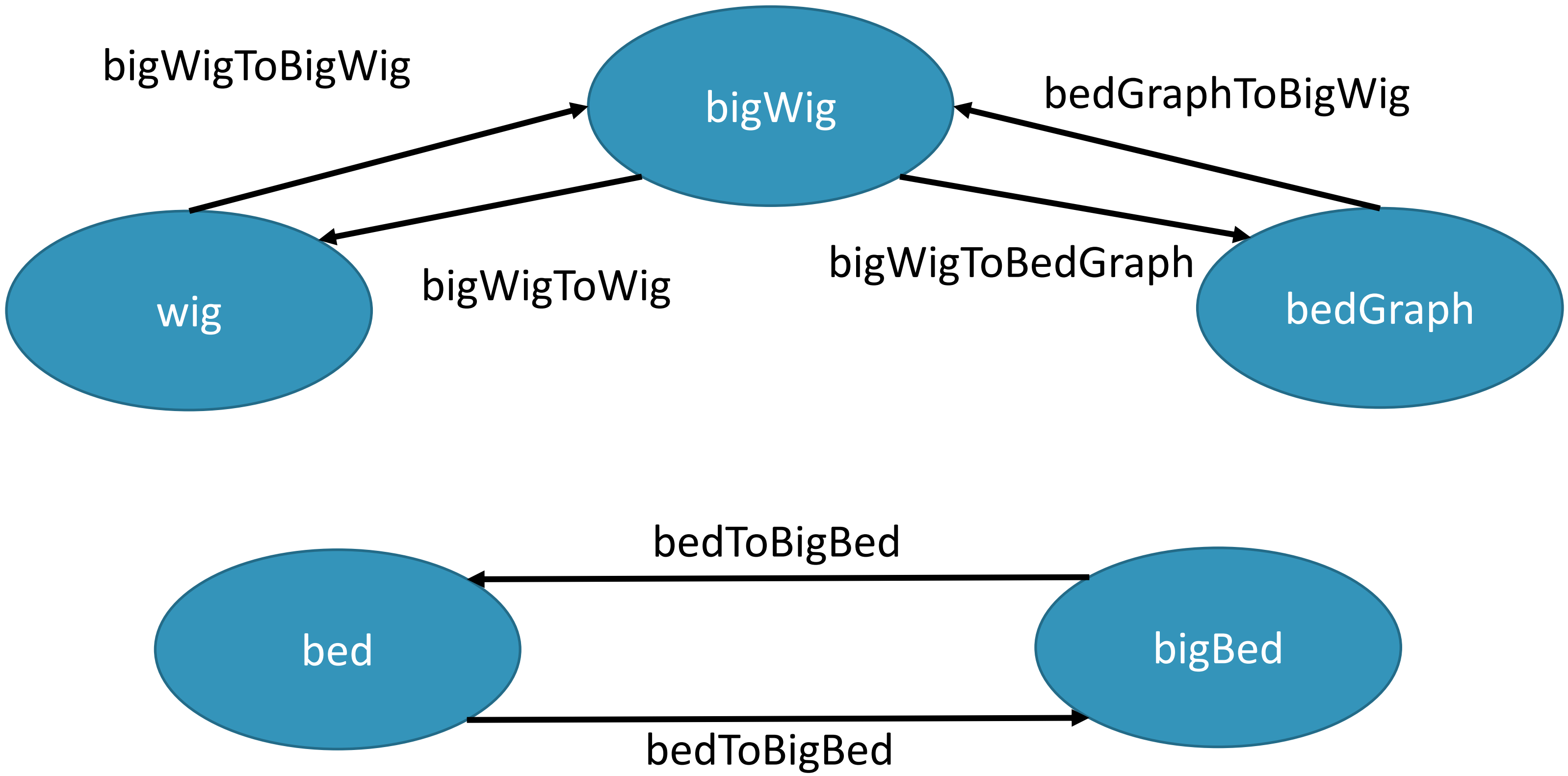
Sizes of blocks (;)

Starts of blocks (;)

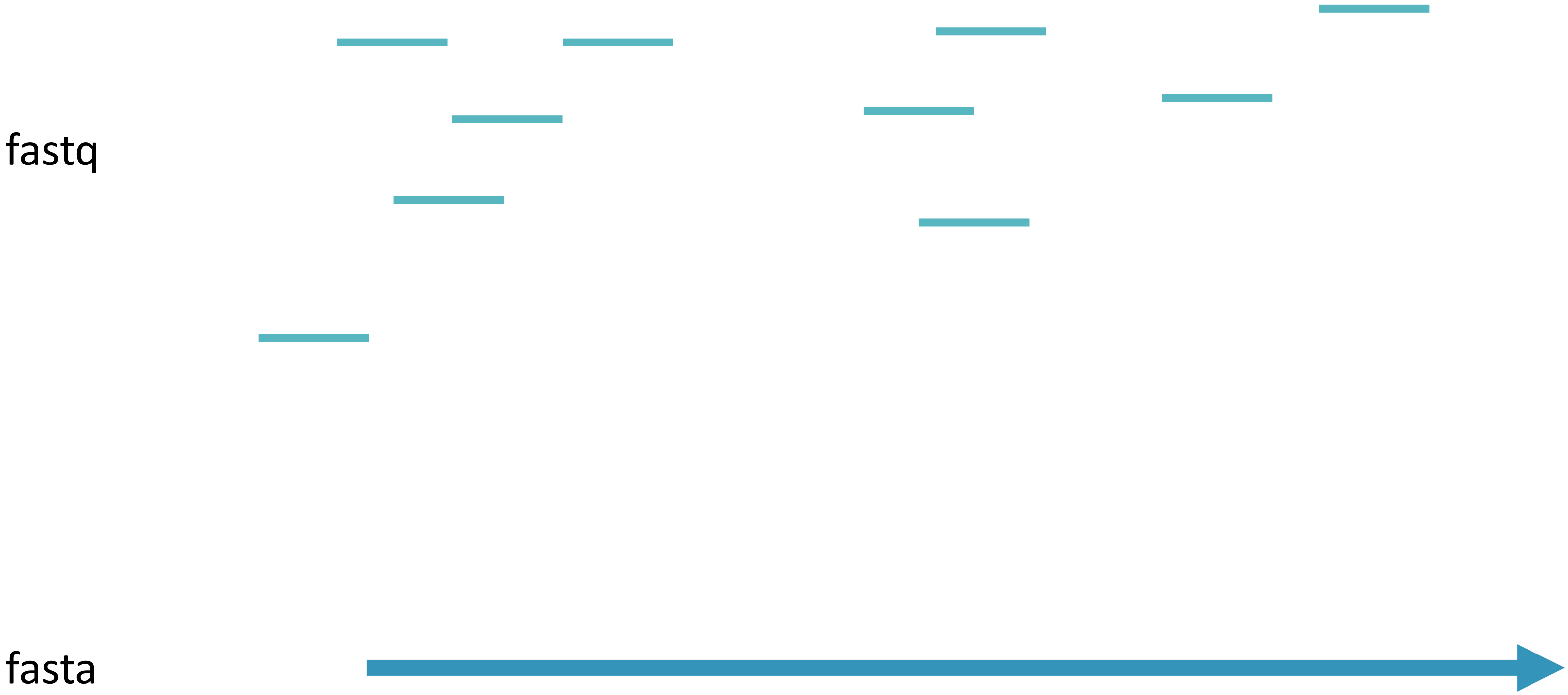
bigBed

- Binary file

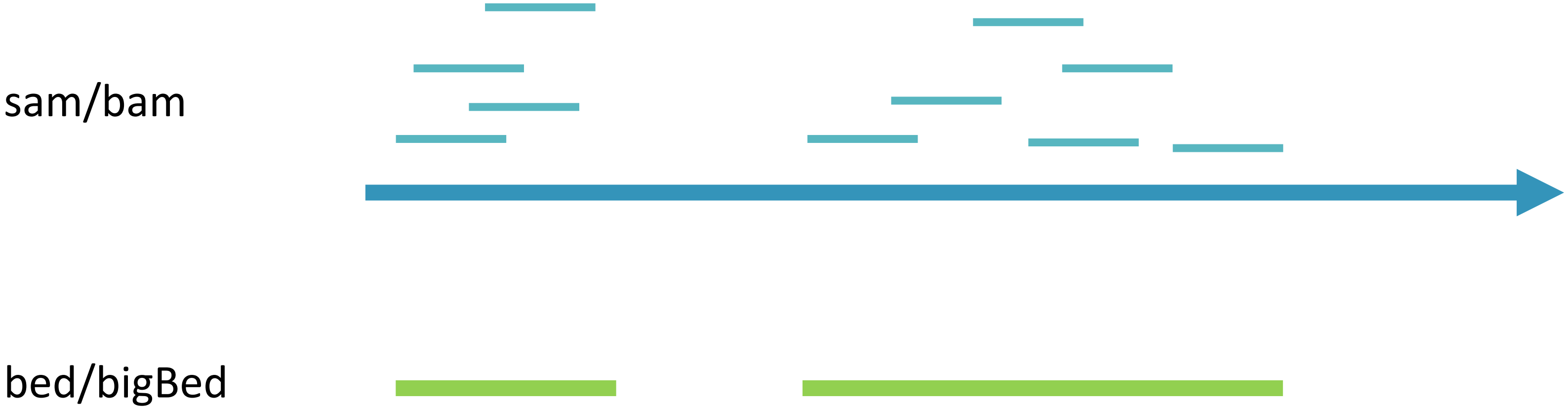
Converting between formats



Formats along the genome



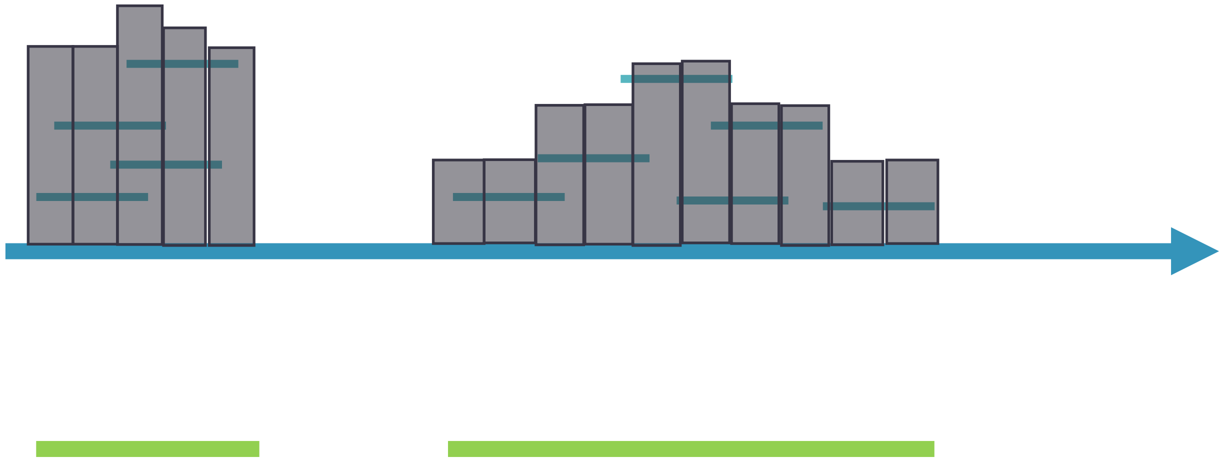
Formats along the genome



sam/bam

bed/bigBed

Formats along the genome



bedGraph
wig/bigWig

bed/bigBed

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:
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

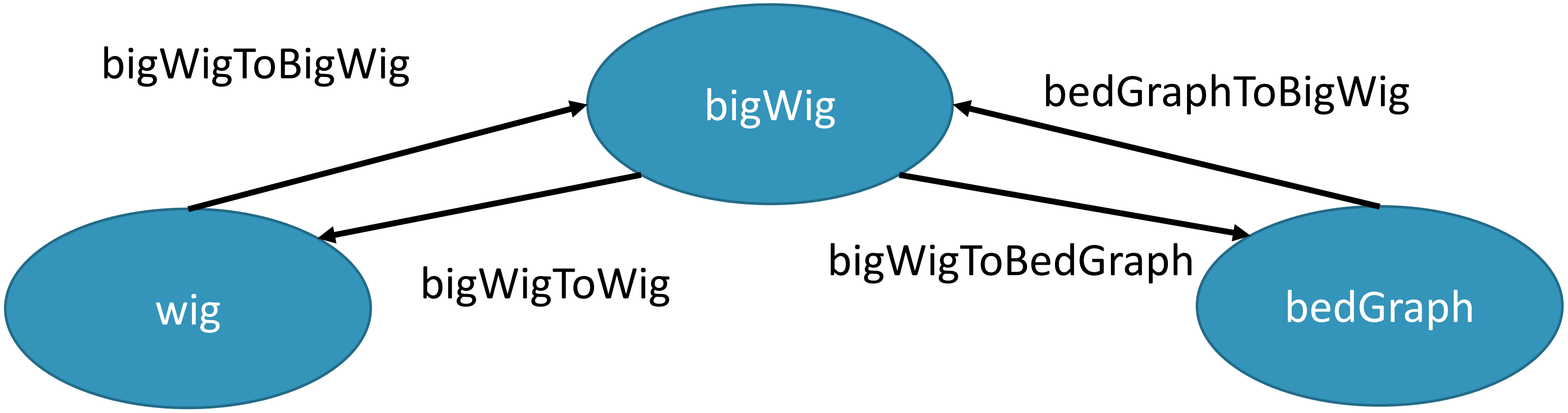
Bonus exercise

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

```
grep chr11 file
```



Hands on:
Subset the bigwig file

ANSWER

```
bigWigToBedGraph
left_ventricle_34m_plus.bigWig
left_ventricle_34m_plus.bedGraph

grep chr11 left_ventricle_34m_plus.bedGrap
> left_ventricle_34m_plus_chr11.bedGraph

bedGraphToBigWig
left_ventricle_34m_plus_chr11.bedGraph
left_ventricle_34m_plus_chr11.bigWig
```