AQUA-FAANG: Bioinformatic analysis of regulatory elements

ChIP-seq Practical

Ilias Lavidas, Garth Ilsley, David Urbina

11 May 2021

Objectives

- Familiarize with ChIP-seq data analysis workflow
- Call Peaks from Alignments
- Visualize output

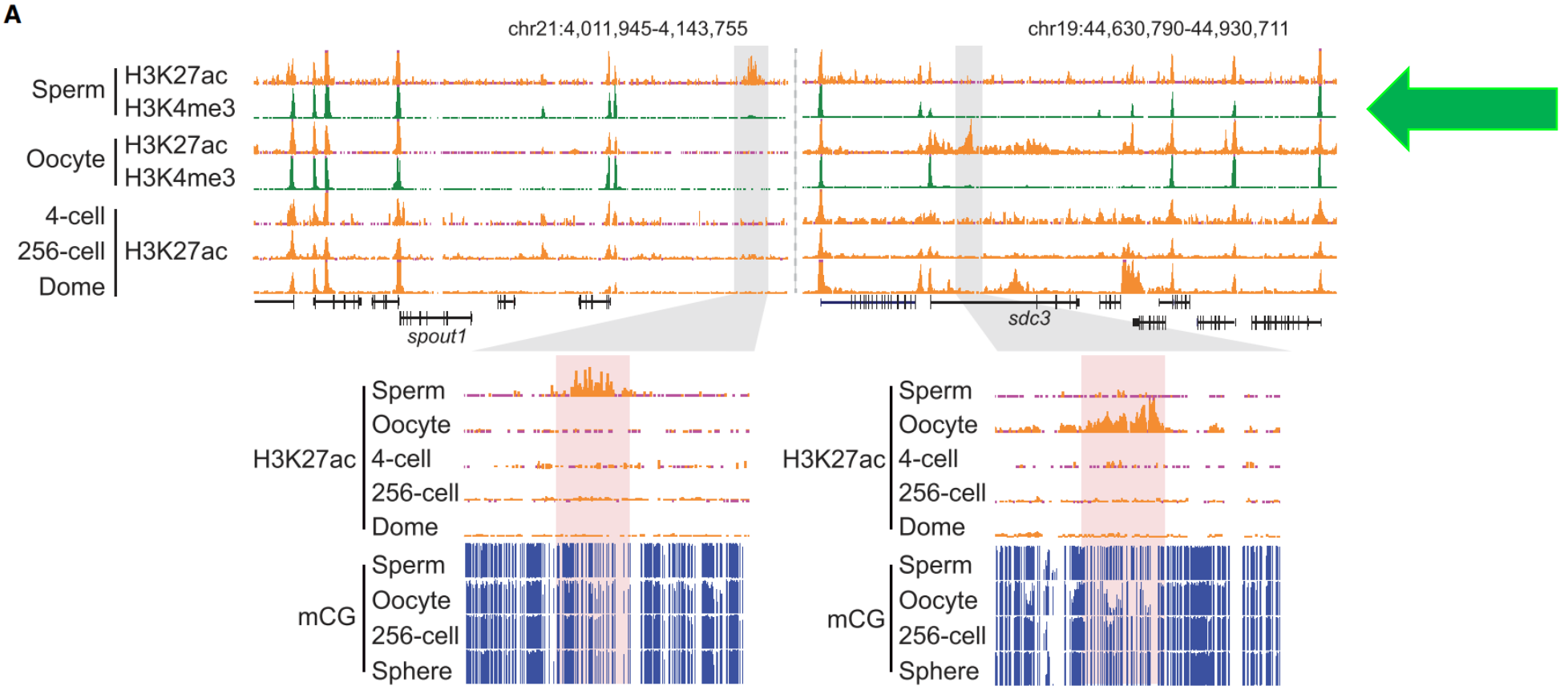
Introduction

`~/train-aquafaang-bioinf/chip-seq/data`

https://hub.docker.com/u/juettemann

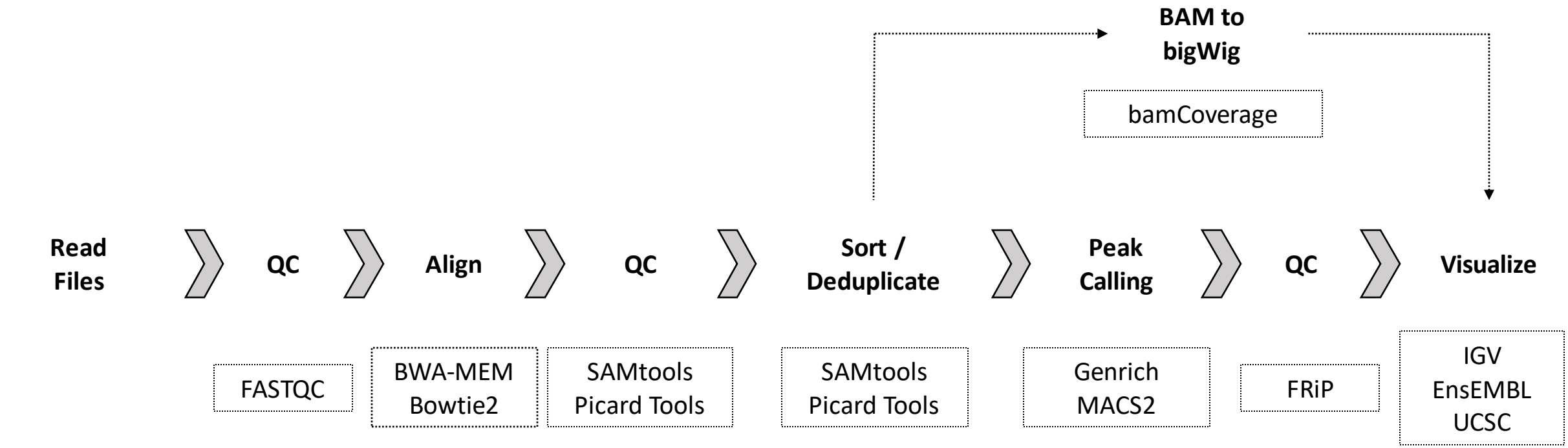
https://github.com/FAANG/train-aquafaang-bioinf.git

Introduction

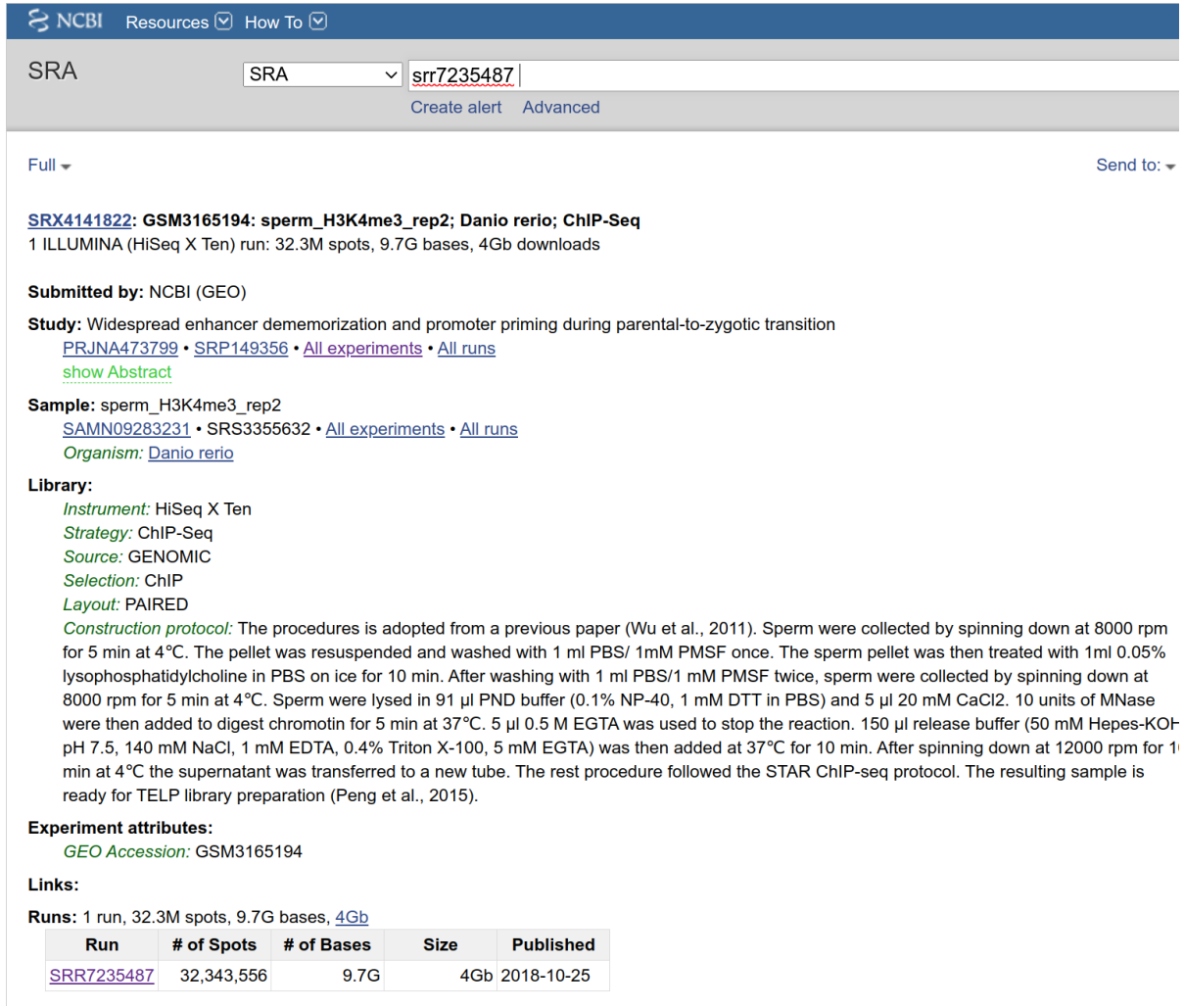


Source: Zhang B et al., Widespread Enhancer Dememorization and Promoter Priming during Parental-to-Zygotic Transition. Mol Cell. 2018 Nov 15;72(4):673-686.e6. doi: 10.1016/j.molcel.2018.10.017. PMID: 30444999.

ChIP-seq Analysis Workflow



Read Files



NCBI Resources How To

SRA SRA srr7235487

Create alert Advanced

Full Send to:

SRX4141822: GSM3165194: sperm_H3K4me3_rep2; Danio rerio; ChIP-Seq
1 ILLUMINA (HiSeq X Ten) run: 32.3M spots, 9.7G bases, 4Gb downloads

Submitted by: NCBI (GEO)

Study: Widespread enhancer dememorization and promoter priming during parental-to-zygotic transition
PRJNA473799 • SRP149356 • All experiments • All runs
show Abstract

Sample: sperm_H3K4me3_rep2
SAMN09283231 • SRS3355632 • All experiments • All runs
Organism: Danio rerio

Library:
Instrument: HiSeq X Ten
Strategy: ChIP-Seq
Source: GENOMIC
Selection: ChIP
Layout: PAIRED
Construction protocol: The procedures is adopted from a previous paper (Wu et al., 2011). Sperm were collected by spinning down at 8000 rpm for 5 min at 4°C. The pellet was resuspended and washed with 1 ml PBS/ 1mM PMSF once. The sperm pellet was then treated with 1ml 0.05% lysophosphatidylcholine in PBS on ice for 10 min. After washing with 1 ml PBS/1 mM PMSF twice, sperm were collected by spinning down at 8000 rpm for 5 min at 4°C. Sperm were lysed in 91 µl PND buffer (0.1% NP-40, 1 mM DTT in PBS) and 5 µl 20 mM CaCl2. 10 units of MNase were then added to digest chromotin for 5 min at 37°C. 5 µl 0.5 M EGTA was used to stop the reaction. 150 µl release buffer (50 mM Hepes-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 0.4% Triton X-100, 5 mM EGTA) was then added at 37°C for 10 min. After spinning down at 12000 rpm for 1 min at 4°C the supernatant was transferred to a new tube. The rest procedure followed the STAR ChIP-seq protocol. The resulting sample is ready for TELP library preparation (Peng et al., 2015).

Experiment attributes:
GEO Accession: GSM3165194

Links:

Runs: 1 run, 32.3M spots, 9.7G bases, 4Gb

Run	# of Spots	# of Bases	Size	Published
SRR7235487	32,343,556	9.7G	4Gb	2018-10-25

Read Files – FASTQ

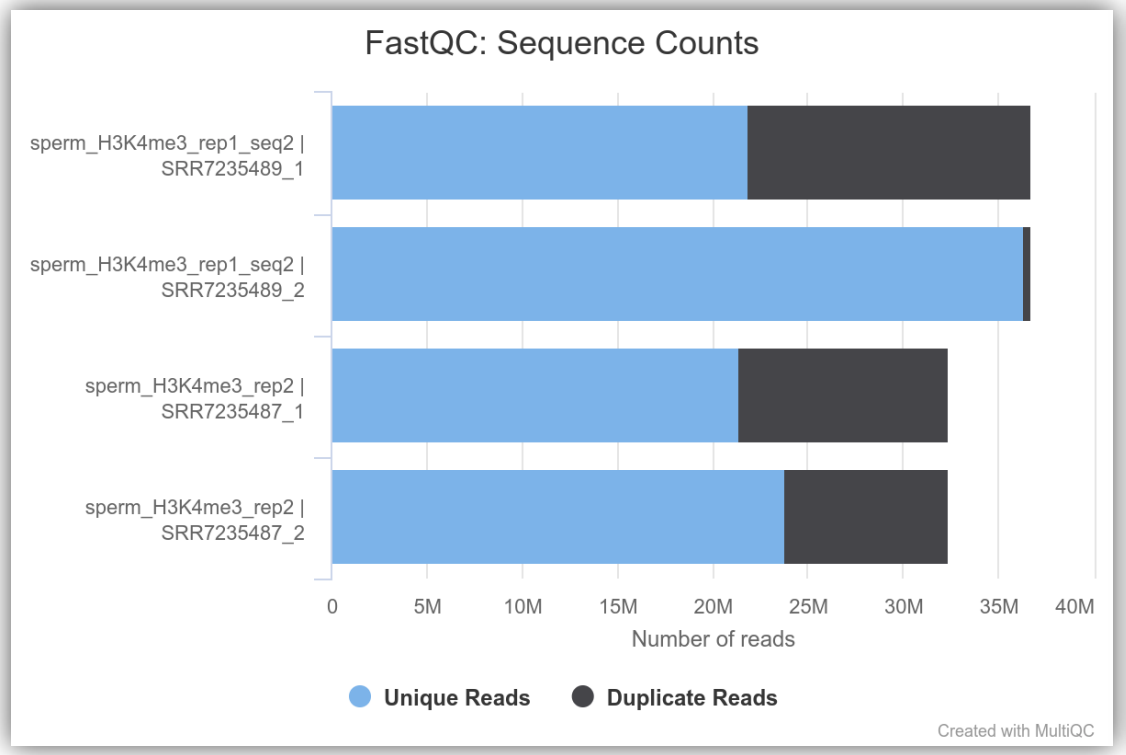
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

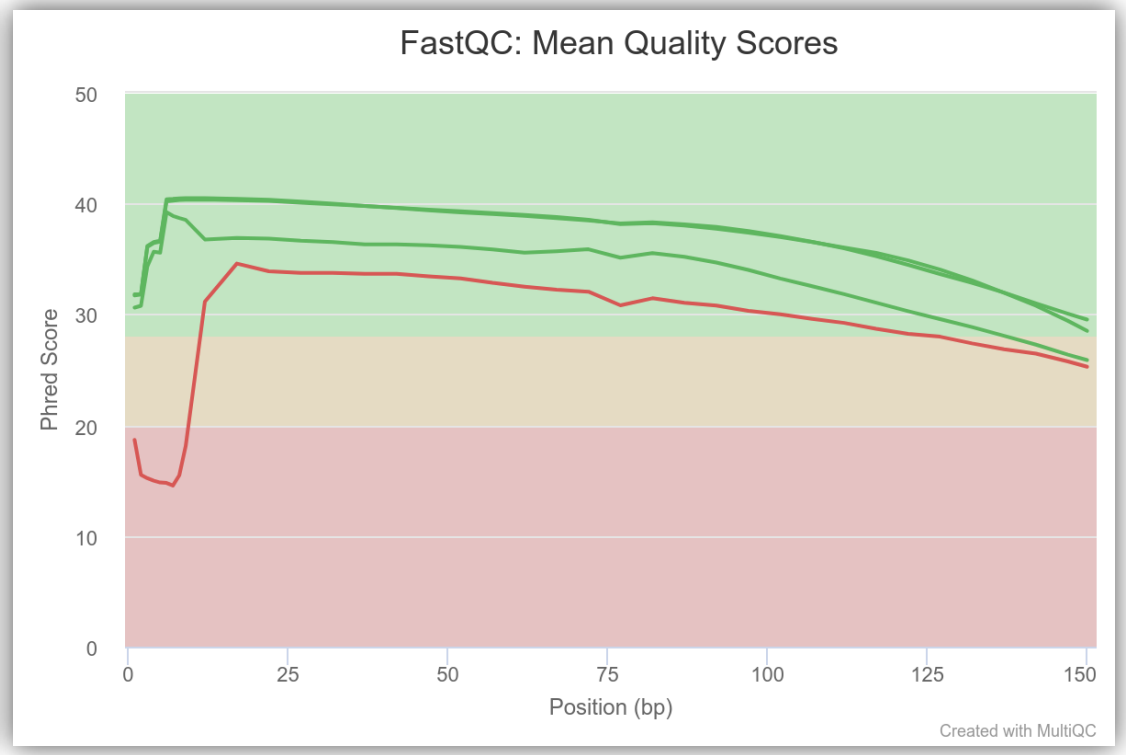
Line 1: sequence identifier, begins with @
Line 2: raw sequence
Line 3: begins with +, optionally followed by seq. id.
Line 4: encoded quality values

```
==> SRR001666_1.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI
```

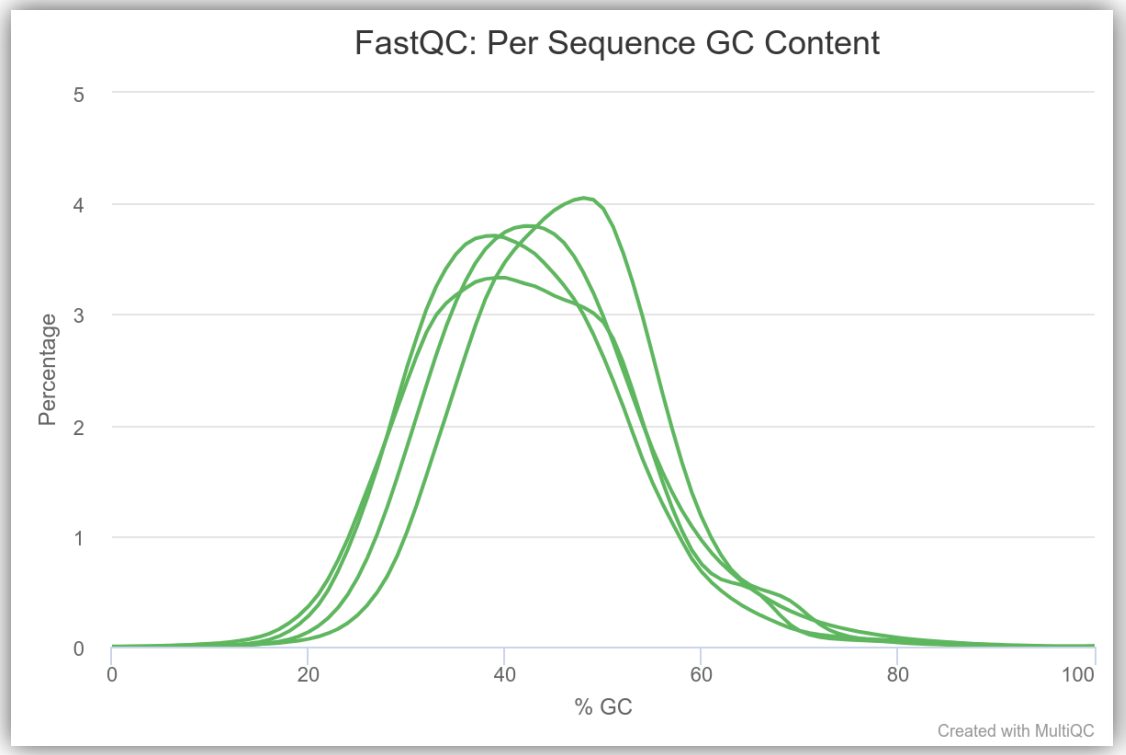
```
==> SRR001666_2.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

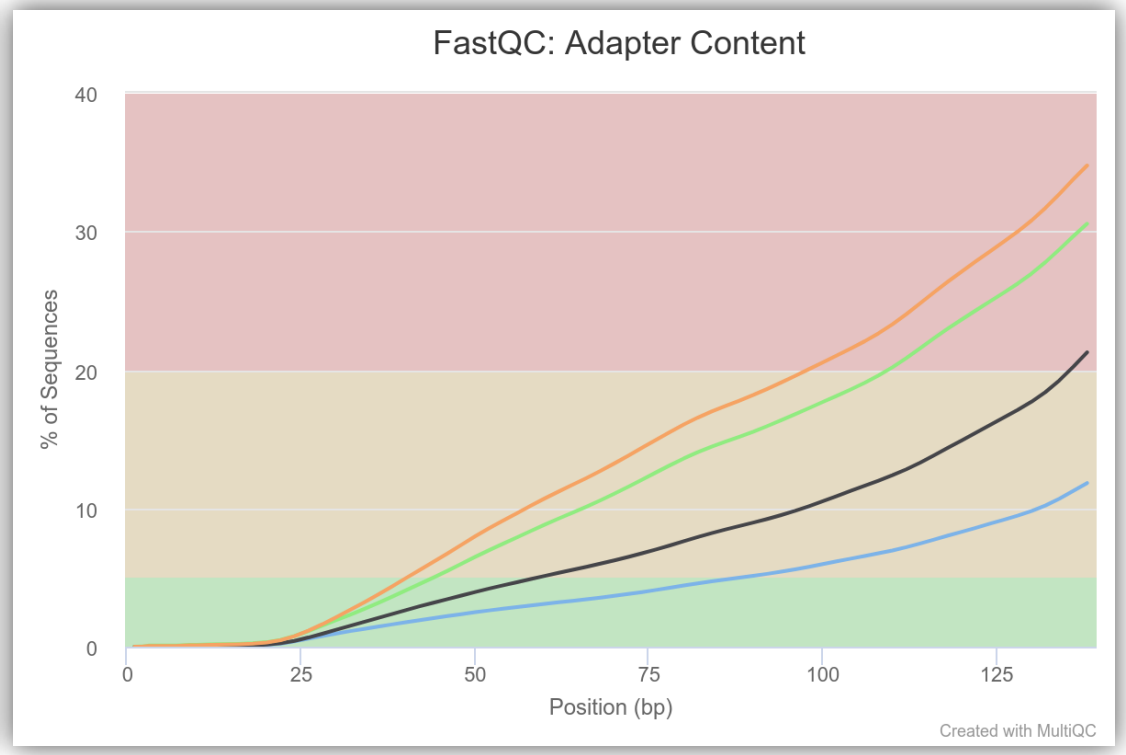
Quality Control – FastQC



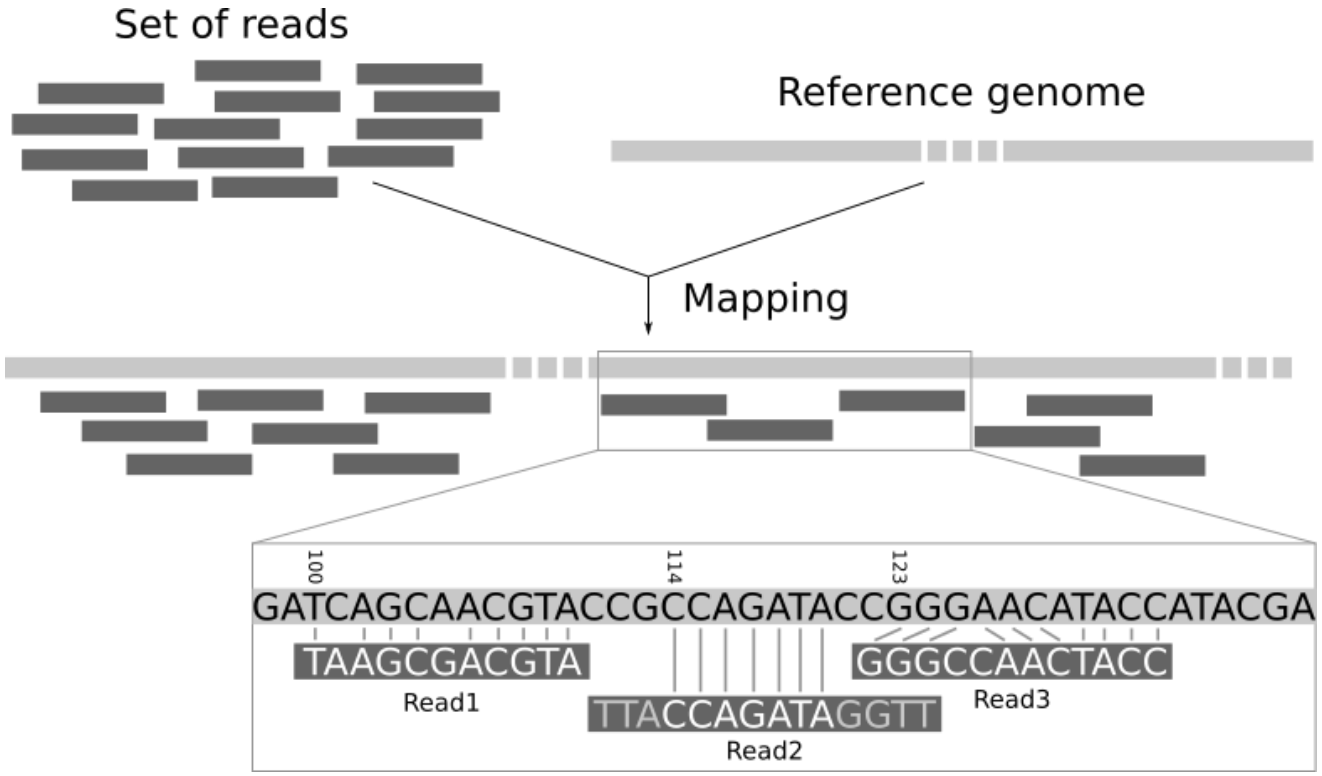


Quality Control – FastQC





Align



Source: https://training.galaxyproject.org/training-material/topics/sequence-analysis/images/mapping/mapping.png

Align - SAM & BAM

SAM: Header section & Alignment section

BAM: binary, compressed SAM

Col	Field	Brief description
1	**QNAME**	**Query template NAME**
2	FLAG	bitwise FLAG
3	**RNAME**	**References sequence NAME**
4	**POS**	**1- based leftmost mapping POSition**
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next read
8	PNEXT	Position of the mate/next read
9	TLEN	observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

More QC – SAMtools

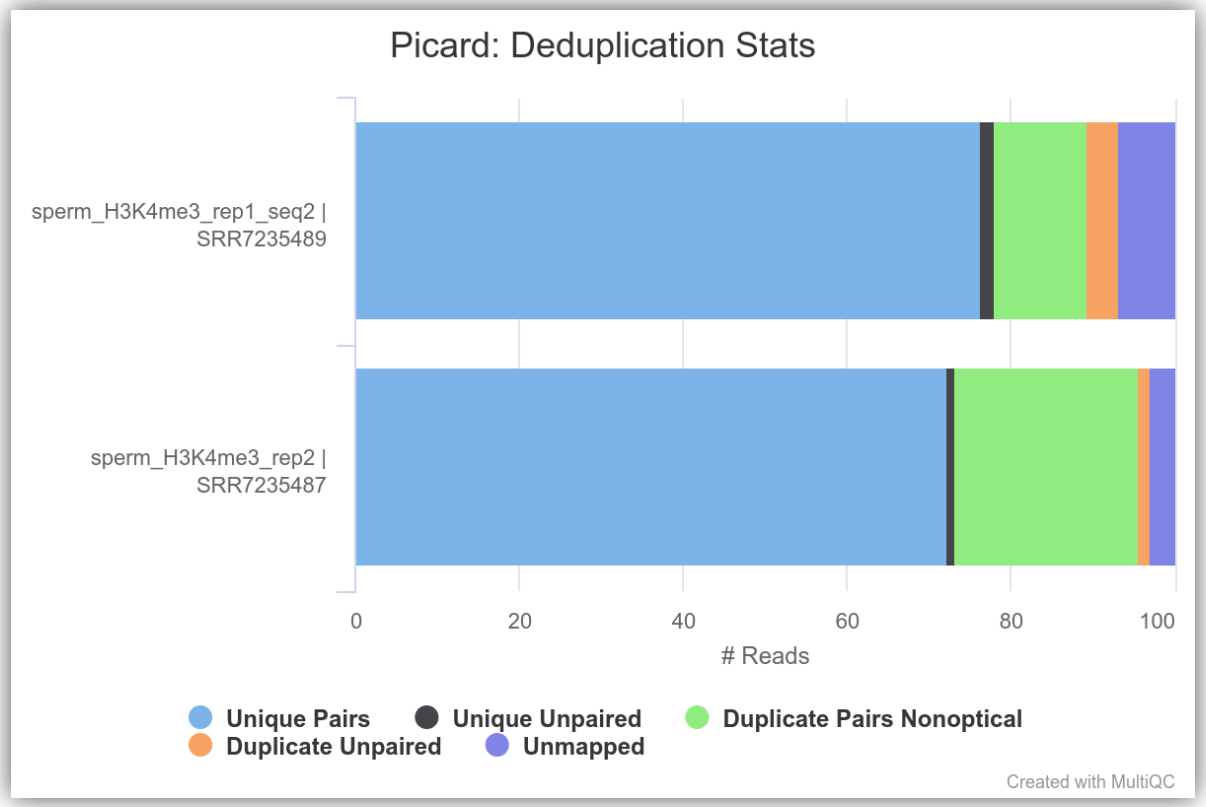
Alignment metrics

This module parses the output from `samtools stats`. All numbers in millions.

Warning: 6 samples hidden. **See toolbox.**

Hover over a data point for more information

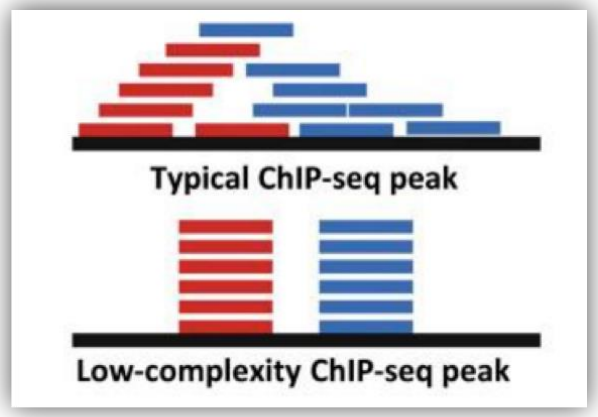
More QC – Picard



Deduplicate

Reads with the same start position are considered duplicates.

- "Good" duplicates: Some level of duplication is expected when sequencing a small part of the genome.
- "Bad" duplicates: Overamplification of low starting material leads to artificially enriched regions



Landt et al, Genome Res. 2012

Good quality peaks: multiple overlapping reads with offsets

Low quality peaks: perfectly stacked reads, PCR artifacts

Genrich can remove duplicates on the fly!

Data

- `/home/training/train-aquafaang-bioinf/chip-seq/data`
  - `sperm_H3K4me3`
    - `roi` → Contains reads that are mapped only on regions of interest (chr 19 & 21)
      - `bam`
      - `bed`
      - `bw`
    - `full` → Contains all reads. Caution: Long runtime
  - `sperm_H3K27ac`
  - `oocyte_H3K27ac`

  For later use
  - `publication` → Data from publication

Sort

Command	`samtools sort`	
Parameters	`-n`	sort by read name
	`-o`	output file
	`-@`	number of threads
	`-m`	required memory per thread

Input Files [BAM]

`sperm_H3K4me3_rep1.roi.bam`

`sperm_H3K4me3_rep2.roi.bam`

Output Files [BAM]

`sperm_H3K4me3_rep1.roi.sorted.bam`

`sperm_H3K4me3_rep2.roi.sorted.bam`

Documentation

```
$ export data_dir="/home/training/train-aquafaang-bioinf/chip-seq/data/sperm_H3K4me3/"

$ export mnt_dir="type=bind,source=$data_dir,target=/mnt"

$ docker run --mount $mnt_dir juettemann/samtools sort -n -@ 7 -m 2G \
-o /mnt/roi/bam/sperm_H3K4me3_rep1.roi.sorted.bam /mnt/roi/bam/sperm_H3K4me3_rep1.roi.bam
```

Sort

Command	**samtools view**

Input Files [BAM]

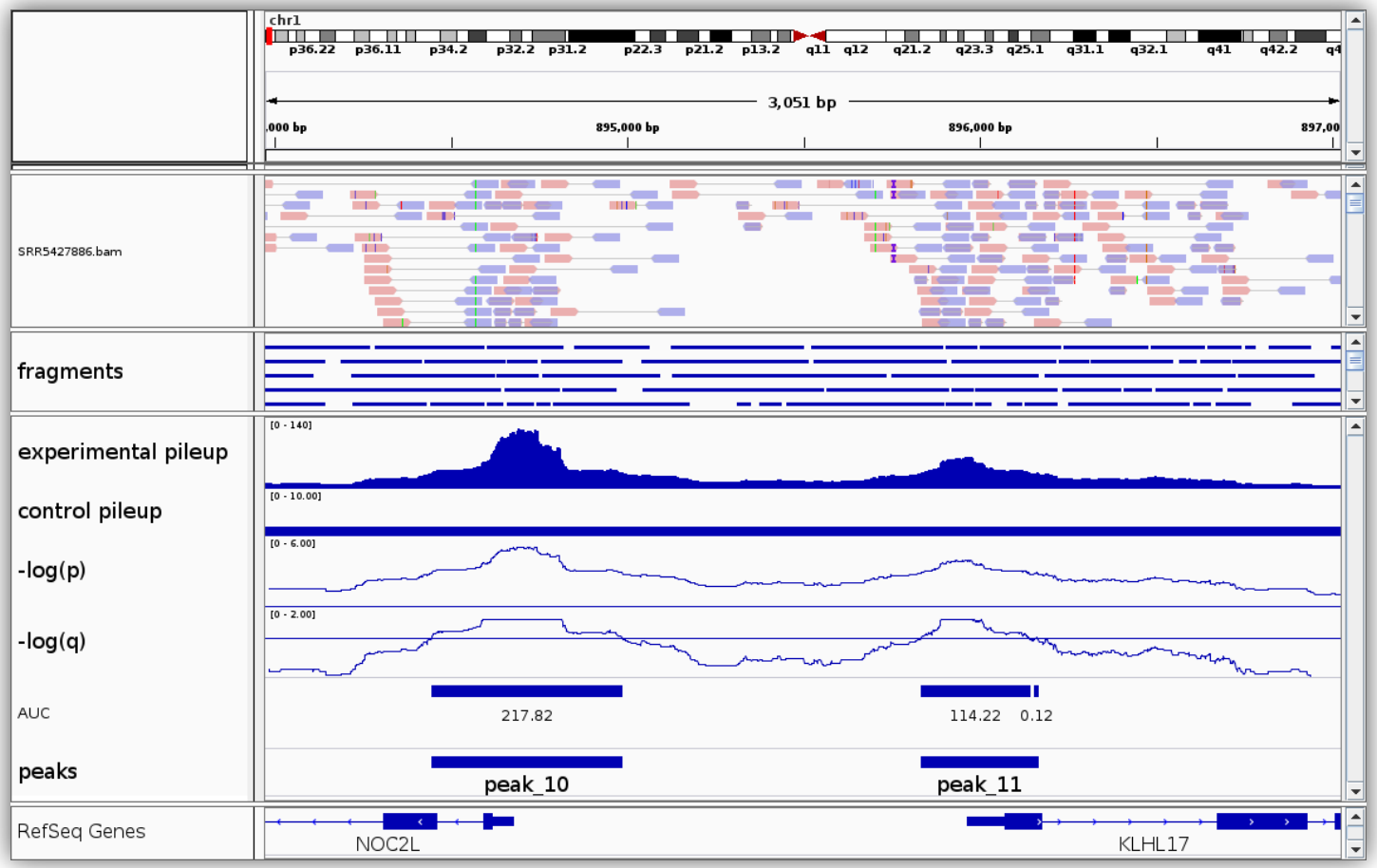
sperm_H3K4me3_rep1.roi.sorted.bam

sperm_H3K4me3_rep2.roi.sorted.bam

Documentation

```
$ docker run --mount $mnt_dir juettemann/samtools view \
/mnt/roi/bam/sperm_H3K4me3_rep1.roi.sorted.bam | head
```

Peak Calling – Genrich method



Source: https://raw.githubusercontent.com/jsh58/Genrich/master/figures/figure1.png

1. Parse alignments for the experimental sample and create an experimental "pileup" by counting the DNA fragments that cover each position of the genome.
2. Create a control pileup using the control sample (if available) and background level.
3. Calculate *p*-values for each genomic position.
4. Calculate the "area under the curve" (AUC) for all regions reaching statistical significance.
5. Combine nearby regions and call peaks whose total AUC is above a threshold.

Peak Calling

Command	**genrich**	
Parameters	**-t**	input experimental files
	-r	remove PCR duplicates
	-o	output narrowPeak file
	-e	list of excluded chromosomes

Input Files [BAM]

sperm_H3K4me3_rep1.roi.sorted.bam

sperm_H3K4me3_rep2.roi.sorted.bam

Output File [narrowPeak]

sperm_H3K4me3.roi.narrowPeak

Documentation

```
$ export excluded_chrs="1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,20,22,23,24,25"

$ docker run --mount $mnt_dir juettemann/genrich -r -v -e $excluded_chrs \
-t /mnt/roi/bam/sperm_H3K4me3_rep1.roi.sorted.bam,/mnt/roi/bam/sperm_H3K4me3_rep2.roi.sorted.bam \
-o /mnt/roi/bed/sperm_H3K4me3.roi.narrowPeak
```

Peak Calling – Genrich Output - narrowPeak

```
$ head ~/train-aquafaang-bioinf/chip-seq/data/sperm_H3K4me3/roi/bed/sperm_H3K4me3.roi.narrowPeak
```

```
19      36444   37763   peak_0  1000    .       1909.176392     5.103715        -1      1009
19      39540   41216   peak_1  1000    .       6766.594727     8.691507        -1      248
19      49840   51068   peak_2  1000    .       8880.161133     15.587051       -1      788
19      58149   59795   peak_3  1000    .       8931.702148     12.211674       -1      758
19      83435   83835   peak_4  958     .       383.198547      3.614118        -1      277
19      131426  132172  peak_5  385     .       287.046326      3.212096        -1      447
19      146393  148042  peak_6  1000    .       10245.839844    12.903536       -1      695
19      161093  162067  peak_7  1000    .       1197.003662     4.899028        -1      565
19      166597  167530  peak_8  1000    .       2016.466919     5.823216        -1      341
19      181269  182503  peak_9  1000    .       6792.544434     11.894626       -1      543
```

1. chrom	Name of the chromosome
2. chromStart	Starting position of the peak (0-based)
3. chromEnd	Ending position of the peak (not inclusive)
4. name	`peak_N`, where `N` is the 0-based count
5. score	Average AUC (total AUC / bp) × 1000, rounded to the nearest int (max. 1000)
6. strand	`.` (no orientation)
7. signalValue	Total area under the curve (AUC)
8. pValue	Summit $-\log_{10}(p\text{-value})$
9. qValue	Summit $-\log_{10}(q\text{-value})$, or `-1` if not available (e.g. without `-q`)
10. peak	Summit position (0-based offset from chromStart): the midpoint of the peak interval with the highest significance (the longest interval in case of ties)

QC – FRiP score

Fraction of Reads in Peaks

$$\text{FRiP score} = \frac{\text{Number of reads that fall inside peaks}}{\text{Total number of reads}}$$

Good quality: FRiP score > 5%

BAM to bigWig

Command	**bamCoverage**	
Parameters	**-b**	input BAM file
	-o	output bigWig file
	-p	number of processors
	-bs	bin size, in bases

Input Files [BAM]

sperm_H3K4me3_rep1.roi.bam

sperm_H3K4me3_rep2.roi.bam

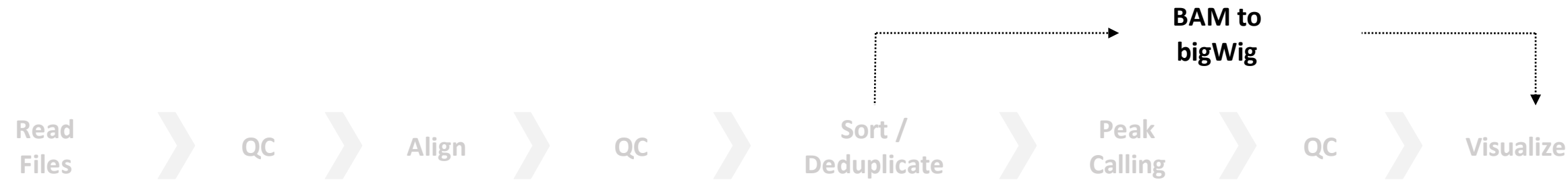
Output Files [bigWig]

sperm_H3K4me3_rep1.roi.bw

sperm_H3K4me3_rep2.roi.bw

Documentation

```
$ docker run --mount $mnt_dir juettemann/deeptools bamCoverage -p 7 -bs 100 \
-b /mnt/roi/bam/sperm_H3K4me3_rep1.roi.bam -o /mnt/roi/bw/sperm_H3K4me3_rep1.roi.bw
```



Visualization options

Ensembl — Documentation

UCSC Genome Browser — Documentation

Documentation

Visualize - IGV

```
$ bash /usr/local/IGV_Linux_2.9.4/igv.sh
```

Regions of interest

- 21:4011945-4143755
- 19:44630790-44930711

Tasks

1. Generate peaks for sperm H3K27ac and oocyte H3K27ac samples. How many did you get for each sample?
2. Visualize your peaks in IGV. Can you spot the differences between the two samples as indicated in the publication figure?
3. Explore Genrich features! Find a way to:
    - Ignore reads that have low mapping quality.
    - Not report peaks that have a very short length.
    - Consider unpaired alignments. (by default, Genrich ignores them)