GSBS Bootstrappers:
Bedtools Workshop

The GSBS Bootstrappers

Part 0 - UNIX Review & Class Setup

Part I – General Overview of NGS Analysis

Create Library

Create Library

Sequence Library

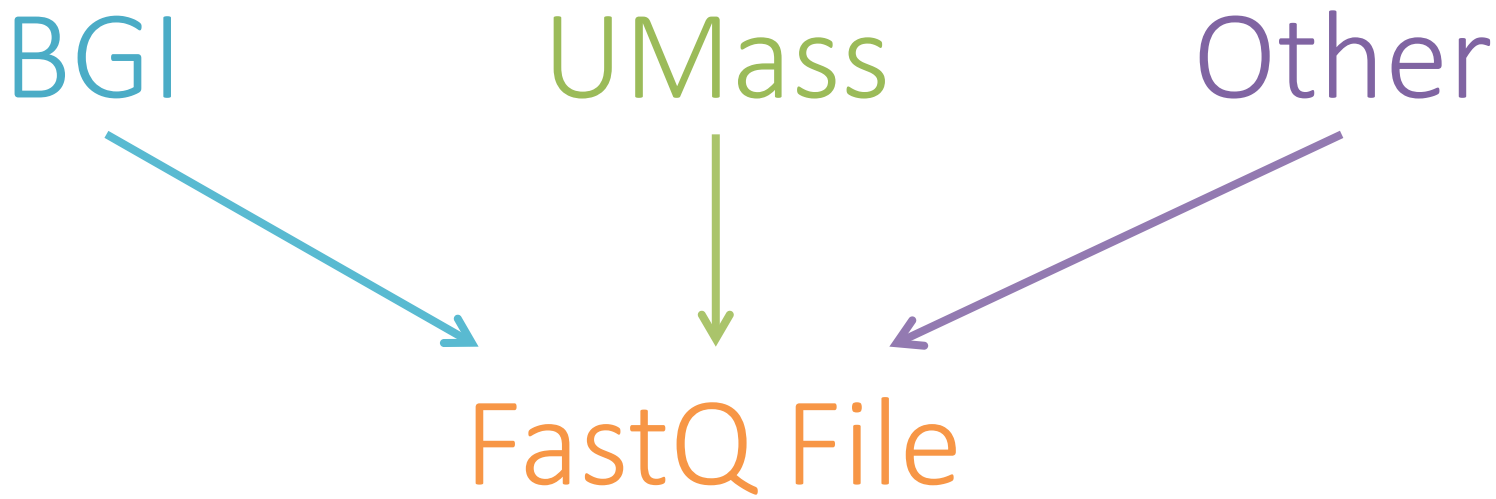
Create Library

Sequence Library

BGI UMass Other

Create Library

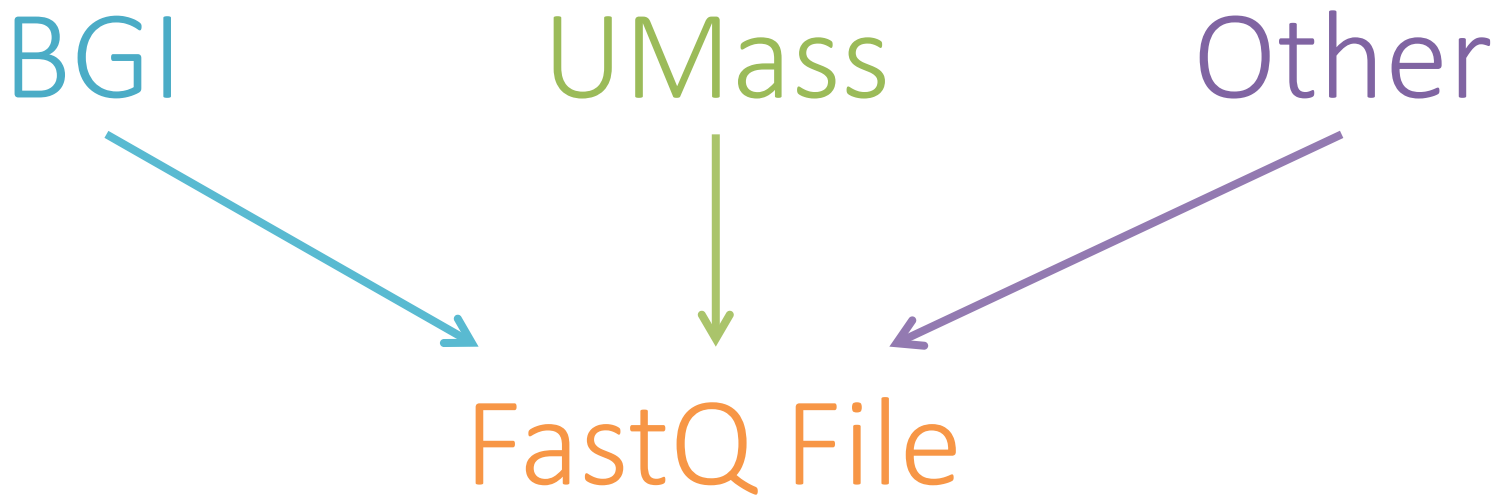
Sequence Library



The GSBS Bootstrappers

Create Library

Sequence Library




```
@GM19092.Ribo.1 7LYMFP1_0320:5:1101:3750:1882 length=29
GTACTGCGCGACAATATCCAGGGCATCAC
+GM19092.Ribo.1 7LYMFP1_0320:5:1101:3750:1882 length=29
S\cceecgggcgfhihghiiiiihfiifg
```

Create Library

Sequence Library

Quality Control/Clean Reads

FastX FastQC Cutadapt

The GSBS Bootstrappers

Create Library

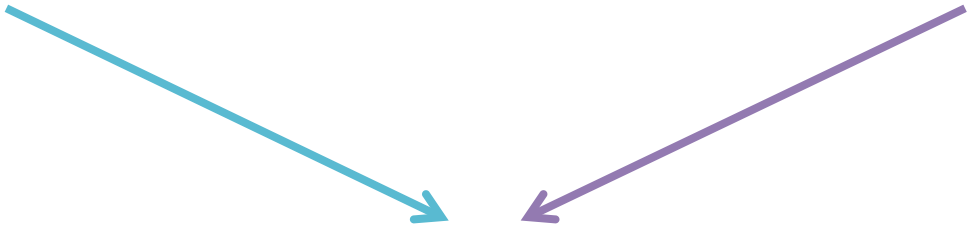
Sequence Library

Quality Control/Clean Reads

Align Reads to Genome

RNA
(STAR, TopHat)

Genome
(BWA, Bowtie2)



BAM/SAM File

The GSBS Bootstrappers

SAM Format

- SAM = sequence alignment/map
- File containing containing information about read alignment such as position, quality, and indels
- Binary version of file is called BAM
- Use samtools to explore and manipulate files

SAM Format

Col	Field	Description
1	QNAME	Query (pair) NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	ISIZE	Inferred insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Create Library

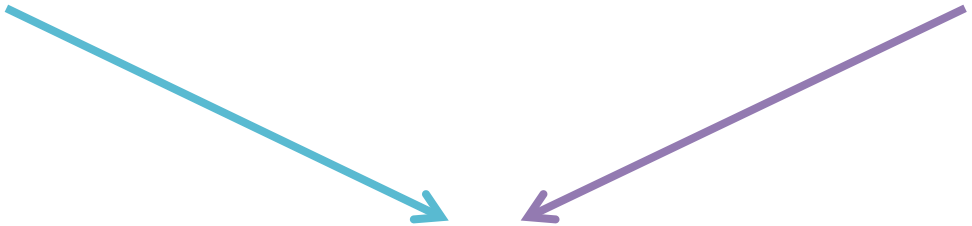
Sequence Library

Quality Control/Clean Reads

Align Reads to Genome

RNA
(STAR, TopHat)

Genome
(BWA, Bowtie2)



BAM/SAM File

The GSBS Bootstrappers

Create Library

Sequence Library

Quality Control/Clean Reads

Align Reads to Genome

Additional Analysis

Gene Expression

Peak Calling

miR Abudance

Create Library

Sequence Library

Quality Control/Clean Reads

Align Reads to Genome

Additional Analysis

Gene Expression

Peak Calling

miR Abudance

BED Format

- BED = Browser Extensible Data format
- In its simplest form, a BED file contains three columns:

Chromosome	Start	End
chr6	20256	21945

- Additional columns can be used to designate more information about each interval

BED6 Format

- Contains six columns

Chromosome	Start	End	Name	Score	Strand

- Not all columns need to have distinct values, "." can be used to denote an empty value

BED6 Format – Genomic Features

Chromosome	Start	End	Name	Score	Strand
chr1	2000	3000	Gene-A	.	+
Chr2	550	600	Gene-B	.	-

BED6 Format – Genomic Features

Chromosome	Start	End	Name	Score	Strand
chr1	2000	3000	Gene-A	.	+
Chr2	550	600	Gene-B	.	-

```
chr15   102382322   102390527   OR4F13P       .   +
chr15   102416167   102417104   OR4F28P       .   +
chr15   102427557   102427970   WBP1LP5       .   -
chr15   102443357   102443407   AC140725.7    .   +
chr15   102462245   102463298   OR4F4         .   -
chr15   102467008   102467910   OR4G2P        .   -
chr15   102495088   102496615   FAM138E       .   +
chr15   102500051   102501611   MIR1302-10    .   -
chr15   102501356   102516768   WASH3P        .   +
chr15   102516758   102519298   DDX11L9       .   -
```

ENCODE NarrowPeak Format

- Format used by ENCODE for transcription factor and histone modification ChIP-seq peaks

Chromosome	Start	Stop	Name	Score	Strand

SignalValue	pvalue	qvalue	peak

- Similar to Bed6 format but with four additional columns

ENCODE NarrowPeak Format

Chromosome	Start	Stop	Name	Score	Strand
chr1	935658	935738	.	0	.

SignalValue	pvalue	qvalue	peak
182	5.09	-1	50

Part II – Running Bedtools

Step 1: Start an Interactive Job

```
bsub -R rusage[mem=1000] -W 2:00 -q interactive -Is bash
```

The GSBS Bootstrappers

Step 1: Start an Interactive Job

Amount of time (in hours)

```
bsub -R rusage[mem=1000] -W 2:00 -q interactive -Is bash
```

Amount of memory (in bytes)

The GSBS Bootstrappers

Step 2: Load Bedtools Module

```
module load bedtools/2.25.0
```

Step 3: Run Bedtools Command

```
bash-4.1$ bedtools
bedtools: flexible tools for genome arithmetic and DNA
sequence analysis.
usage:    bedtools <subcommand> [options]

The bedtools sub-commands include:

...
...
```

The GSBS Bootstrappers

Part III – Bamtobed

bamtobed

```
bedtools bamtobed –i file.bam
```

Example:

```
bedtools bamtobed –I RNAseq_K562_chr15.bam >
RNAseq_K562_chr15.bed
```

bamtobed Options

-bedpe = write alignments in paired-end format (requires bam file to be sorted by read name)

-split = report each read split as a unique bed entry

-cigar = add cigar string as 7th column

What we will cover:

Class 2 -

- Intersect
- Jaccard
- Merge
- Complement
- Subtract

Class 3 -

- Coverage/Multicov
- Genomecvg
- Shuffle

Class 4 -

- Map

The GSBS Bootstrappers