cipher Documentation

Release 1.0.0

Carlos Guzman

Dec 22, 2017

Contents:

CHAPTER 1

Installation

CIPHER requires no manual installation. Just download and unzip the tar file from our GitHub or git clone the repository. The download may be large.

The only required manual software installation on your local computer / cluster are Docker and Nextflow.

By default, CIPHER will run off a Docker container, but this can be optionally turned off by removing the option in the config file. For more information, please read the 'Nextflow Config File' documentation.

CHAPTER 2

Installing Nextflow

For Linux:

1. **We recommend installing Nextflow through the Anaconda package manager.**

```
wget https://repo.continuum.io/archive/Anaconda2-4.3.1-Linux-x86_64.sh
bash Anaconda2-4.3.1-Linux-x86_64.sh
```

2. **Install Nextflow**

```
conda install -c bioconda nextflow
```

For macOS:

1. **We recommend installing Nextflow through the Anaconda package manager.**

```
wget https://repo.continuum.io/archive/Anaconda2-4.3.1-MacOSX-x86_64.sh
bash Anaconda2-4.3.1-Linux-x86_64.sh
```

2. **Install Nextflow**

```
conda install -c bioconda nextflow
```

For Windows:

CIPHER has not been tested on the Windows operating system. However, Windows 10 has introduced the 'Ubuntu Bash Shell' sub-system, which potentially can be used to run CIPHER.

1. Install Ubuntu Bash on your Windows 10 computer. Follow the instructions here: https://www.howtogeek.com/249966/how-to-install-and-use-the-linux-bash-shell-on-windows-10/
2. **Using the Ubuntu Bash terminal we recommend installing nextflow through the Anaconda package manager.**

```
wget https://repo.continuum.io/archive/Anaconda2-4.3.1-Linux-x86_64.sh
bash Anaconda2-4.3.1-Linux-x86_64.sh
```

3. **Install Nextflow**

```
conda install -c bioconda nextflow
```

CHAPTER 3

Installing Docker

Installing Docker can be tricky. Because the instructions for correct installation are constantly changing, we provide links to Docker's official installation process. You'll want to use Linux's installation instructions for Windows via the Ubuntu Bash terminal.

For Linux:

https://docs.docker.com/engine/installation/#supported-platforms

For macS:

https://docs.docker.com/docker-for-mac/

CHAPTER 4

Manual Installation

Some universities may not want to install Docker for security reasons. And others may not want to use Docker at all on their local desktop. This section lists how to manually install all the dependencies for CIPHER.

We recommend using the Anaconda package manager.

For Linux:

1. **We recommend installing Nextflow through the Anaconda package manager.**

```
wget https://repo.continuum.io/archive/Anaconda2-4.3.1-Linux-x86_64.sh
bash Anaconda2-4.3.1-Linux-x86_64.sh
```

2. **Install bioconda packages**

```
conda install -c bioconda nextflow fastqc bbmap star hisat2 bowtie2 bwa␣
↪multiqc macs2 deeptools epic preseq samtools sambamba bedtools bedops␣
↪stringtie subread
```

3. **Install R packages**

```
R

install.packages(c("data.table", "ggplot2", "gplots"))
install.packages("http://hartleys.github.io/QoRTs/QoRTs_LATEST.tar.gz",␣
↪repos=NULL, type="source")
source("https://bioconductor.org/biocLite.R")
biocLite()
biocLite(c("ChIPQC", "RUVSeq", "ChIPseeker"))
```

For macOS:

1. **We recommend installing Nextflow through the Anaconda package manager.**

```
wget https://repo.continuum.io/archive/Anaconda2-4.3.1-MacOSX-x86_64.sh
bash Anaconda2-4.3.1-Linux-x86_64.sh
```

2. **Install bioconda packages**

```
conda install -c bioconda nextflow fastqc bbmap star hisat2 bowtie2 bwa␣
↪multiqc macs2 deeptools epic preseq samtools sambamba bedtools bedops␣
↪stringtie subread
```

3. **Install R packages**

```
R

install.packages(c("data.table", "ggplot2", "gplots"))
install.packages("http://hartleys.github.io/QoRTs/QoRTs_LATEST.tar.gz",␣
↪repos=NULL, type="source")
source("https://bioconductor.org/biocLite.R")
biocLite()
biocLite(c("ChIPQC", "RUVSeq", "ChIPseeker"))
```

CHAPTER 5

Parameters

The following table lists the currently available parameters for CIPHER along with a brief description. Updated: May 2017

REQUIRED PARAMETERS

Parameter	Description
–mode	Choose from available: ‘chip’, ‘rna’, ‘gro’, ‘mnase’, ‘dnase’ , ‘atac’, ‘analysis’.
–config	Tab separated config file with sample information. Check README for more information.
–fasta	Reference genome in FASTA format.
–gtf	Reference genome in GTF format.
–lib	Library information. “s” for single-stranded data, “p” for pair-ended data.
–readLen	The length of your reads.

RNA-seq MODE ONLY

Parameter	Description
--strandInfo	Strandedness information. Choose from “unstranded”, “frFirstStrand”, or “frSecondStrand”.
--expInfo	Tab separated config file for RNA-seq DGE analysis. Check README for more information.

ANALYSIS MODE ONLY

Parameter	Description
--function	Choose from available: “predictEnhancers”, “geneExpressionNearPeaks”. Check README for more information.

OPTIONAL PARAMETERS

Parameter	Description	Default Value
–threads	Number of threads to use.	1
–aligner	The alignment software for mapping reads. Available: bbmap, star, hisat2, bwa, bowtie2.	bbmap
–minid	Minimum alignment identity to look for during BBMap mapping. Higher is faster and less sensitive.	0.76
–qvalue	Minimum FDR cutoff for peak detection in MACS2 ad EPIC.	0.01
–epic_w	Window size ued to scan the genome for peak detection in EPIC.	200
–epic_g	A multiple of epic_w used to determine the gap size in EPIC.	3
–maxindel	Maximum indel length searched during mapping. 200k recommended for vertebrate genomes.	200k
–intronlen	Maximum intron length during mapping. 20 recommended for vertebrate genomes.	20
–egs	The effective genome size of your species. Is automatically calculated by default. 80GB of RAM.	Auto
–egs_ratio	Effective genome as fraction of genome size. Must be between 0 and 1. Automatically calculated.	Auto
–outdir	Name of ourput directory.	results

ALIGNER-SPECIFIC OPTIONAL PARAMETERS

BWA

Parameter	Description	Default Value
–bwa_k	See -k option in BWA user manual for more information.	19
–bwa_w	See -w option in BWA user manual for more information.	100
–bwa_d	See -d option in BWA user manual for more information.	19
–bwa_r	See -r option in BWA user manual for more information.	1.5
–bwa_c	See -c option in BWA user manual for more information.	10000
–bwa_A	See -A option in BWA user manual for more information.	1
–bwa_B	See -B option in BWA user manual for more information.	4
–bwa_O	See -O option in BWA user manual for more information.	6
–bwa_E	See -E option in BWA user manual for more information.	1
–bwa_L	See -L option in BWA user manual for more information.	5
–bwa_U	See -U option in BWA user manual for more information.	9
–bwa_T	See -T option in BWA user manual for more information.	30

BOWTIE2

Parameter	Description	Default Value
–bt2_D	See -D option in Bowtie2 user manual for more information.	20
–bt2_R	See -R option in Bowtie2 user manual for more information.	3
–bt2_N	See -N option in Bowtie2 user manual for more information.	0
–bt2_L	See -L option in Bowtie2 user manual for more information.	20
–bt2_i	See -i option in Bowtie2 user manual for more information.	S,5,1,0.50
–bt2_trim5	See –trim5 option in Bowtie2 user manual for more information.	0
–bt2_trim3	See –trim3 option in Bowtie2 user manual for more information.	0
–local	Set this parameter to map alignments in local mode.	false

HISAT2

Parameter	Description	Default Value
–hs_k	See -k option in HISAT2 user manual for more information.	5
–hs_trim5	See –trim5 option in HISAT2 user manual for more information.	0
–hs_trim3	See –trim3 option in HISAT2 user manual for more information.	0
–hs_mp	See –mp option in HISAT2 user manual for more information.	6,2
–hs_sp	See –sp option in HISAT2 user manual for more information.	2,1
–hs_np	See –np option in HISAT2 user manual for more information.	1
–hs_rdg	See –rdg option in HISAT2 user manual for more information.	5,3
–hs_rfg	See –rfg option in HISAT2 user manual for more information.	5,3
–hs_pen_cansplice	See –pen-cansplice option in HISAT2 user manual for more information.	0
–hs_pen_noncansplice	See –pen-noncansplice option in HISAT2 user manual for more information.	12
–hs_min_intronlen	See –min-intronlen option in HISAT2 user manual for more information.	20
–hs_max_intronlen	See –max-intronlen option in HISAT2 user manual for more information.	500000
–hs_max_seeds	See –max-seeds option in HISAT2 user manual for more information.	5

STAR

Parameter	Description	Default Value
–star_clip3pNbases	See –clip3pNbases option in STAR user manual for more information.	0
–star_clip5pNbases	See –clip5pNbases option in STAR user manual for more information.	0
–star_outFilterMultimapScoreRange	See –outFilterMultimapScoreRange option in STAR user manual for more information.	1
–star_outFilterMultimapNmax	See –outFilterMultimapNmax option in STAR user manual for more information.	10
–star_outFilterMismatchNmax	See –outFilterMismatchNmax option in STAR user manual for more information.	10
–star_outFilterScoreMin	See –outFilterScoreMin option in STAR user manual for more information.	0
–star_alignEndsType	See –alignEndsType option in STAR user manual for more information.	Local
–star_winAnchorMultimapNmax	See –winAnchorMultimapNmax option in STAR user manual for more information.	50
–star_quantMode	See –quantMode option in STAR user manual for more information.	•
–star_twopassMode	See –twopassMode option in STAR user manual for more information.	None

FOR OPTIMIZING AND TESTING

Parameter	Description	Default Value
–subsample	Set this flag to subsample reads for teting.	false

CHAPTER 6

ChIP-seq Workflow

All workflows require the following files:

1. A config file (described below)
2. Reference genome FASTA file
3. Reference genome GTF file

Config File A config file is a tab separated text file that includes information regarding the name, location, and input of your experiment.

Single-End Config This is the config file format for single-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1.fastq.gz  control1    ␣
↪   sample1
sample1         sample1_rep2    /path/to/sample1_rep2.fastq.gz  control1    ␣
↪   sample1
sample2         sample2_rep1    /path/to/sample2_rep1.fastq.gz  control2    ␣
↪   sample2
sample2         sample2_rep2    /path/to/sample2_rep2.fastq.gz  control2    ␣
↪   sample2
control1        control1_rep1   /path/to/control1_rep1.fastq.gz control1    ␣
↪   input
control2        control2_rep1   /path/to/control2_rep1.fastq.gz control2    ␣
↪   input
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path: The path to the fastq file to be processed. Can be gzipped or not.
4. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.

5. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Pair-End Config This is the config file format for pair-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1_R1.fastq.gz /path/to/
↪sample1_rep1_R2.fastq.gz     control1        sample1
sample1         sample1_rep2    /path/to/sample1_rep2_R1.fastq.gz /path/to/
↪sample1_rep2_R2.fastq.gz     control1        sample1
sample2         sample2_rep1    /path/to/sample2_rep1_R1.fastq.gz /path/to/
↪sample2_rep1_R2.fastq.gz     control2        sample2
sample2         sample2_rep2    /path/to/sample2_rep2_R1.fastq.gz /path/to/
↪sample2_R2_rep1.fastq.gz     control2        sample2
control1        control1_rep1   /path/to/control1_rep1_R1.fastq.gz /path/to/
↪control1_rep1_R2.fastq.gz   control1        input
control2        control2_rep1   /path/to/control2_rep1_R1.fastq.gz /path/to/
↪control2_rep1_R2.fastq.gz   control2        input
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path1: The path to the first fastq file to be processed. Can be gzipped or not.
4. Path2: The path to the second fastq file to be processed. Can be gzipped or not.
5. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
6. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Simple ChIP-seq Tutorial (single-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode chip --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75
```

Simple ChIP-seq Tutorial (pair-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode chip --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib p --readLen 75
```

Simple ChIP-seq Tutorial (single-end, 75 length reads, use bowtie2 aligner instead of default bbmap, use 5 threads)

```
nextflow run /path/to/main.nf --mode chip --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75 --aligner bowtie2 -
↪-threads 5
```

CHAPTER 7

RNA-seq Workflow

RNA-seq workflows require the following files:

1. A config file (described below)
2. Reference genome FASTA file
3. Reference genome GTF file
4. A experimental config file (described below)

Config File A config file is a tab separated text file that includes information regarding the name, location, and input of your experiment.

Single-End Config This is the config file format for single-ended data.

```
sample1           sample1_rep1    /path/to/sample1_rep1.fastq.gz
sample1           sample1_rep2    /path/to/sample1_rep2.fastq.gz
sample1     sample1_rep3    /path/to/sample1_rep2.fastq.gz
sample2           sample2_rep1    /path/to/sample2_rep1.fastq.gz
sample2           sample2_rep2    /path/to/sample2_rep2.fastq.gz
sample2     sample2_rep3    /path/to/sample1_rep2.fastq.gz
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path: The path to the fastq file to be processed. Can be gzipped or not.

Pair-End Config This is the config file format for pair-ended data.

```
sample1           sample1_rep1    /path/to/sample1_rep1_R1.fastq.gz /path/to/
↪sample1_rep1_R2.fastq.gz
sample1           sample1_rep2    /path/to/sample1_rep2_R1.fastq.gz /path/to/
↪sample1_rep2_R2.fastq.gz
```

```
sample1     sample1_rep3    /path/to/sample1_rep3_R1.fastq.gz /path/to/
↪sample1_rep3_R2.fastq.gz
sample2           sample2_rep1    /path/to/sample2_rep1_R1.fastq.gz /path/to/
↪sample2_rep1_R2.fastq.gz
sample2           sample2_rep2    /path/to/sample2_rep2_R1.fastq.gz /path/to/
↪sample2_rep2_R2.fastq.gz
sample2     sample2_rep3    /path/to/sample2_rep1_R1.fastq.gz /path/to/
↪sample2_rep3_R2.fastq.gz
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path1: The path to the first fastq file to be processed. Can be gzipped or not.
4. Path2: The path to the second fastq file to be processed. Can be gzipped or not.

Experimental Config File A experimental config file is a tab separated text file that has sample and condition information. Your experimental config file must be ordered in the same way that your config file is. (sample1 in config file == sample1 in expInfo file). Additionally you must include the headers in the expInfo config file.

```
sample  condition
ctrl1    WT
ctrl2    WT
ctrl3    WT
ko1      KO
ko2      KO
ko3      KO
```

The headers "sample" and "condition" must be included.

The columns represent:

1. SampleID: Refers to the ID in your config file. Used to differentiate between different replicates.
2. Condition: Refers to the condition or experimental variable of your sample. CIPHER currently only supports two condition DGE analysis.

Simple RNA-seq Tutorial (single-end, 75 length reads, frFirstStrand)

```
nextflow run /path/to/main.nf --mode rna --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75 --strandInfo␣
↪frFirstStrand --expInfo exp_config.txt
```

Simple RNA-seq Tutorial (pair-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode rna --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib p --readLen 75 --strandInfo␣
↪frFirstStrand --expInfo exp_config.txt
```

Simple RNA-seq Tutorial (single-end, 75 length reads, use star aligner instead of default bbmap, use 5 threads)

```
nextflow run /path/to/main.nf --mode rna --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75 --strandInfo␣
↪frFirstStrand --expInfo exp_config.txt --aligner star --threads 5
```

CHAPTER 8

DNase-seq Workflow

All workflows require the following files:

1. A config file (described below)
2. Reference genome FASTA file
3. Reference genome GTF file

Config File A config file is a tab separated text file that includes information regarding the name, location, and input of your experiment.

Single-End Config This is the config file format for single-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1.fastq.gz  -       ␣
↪sample1
sample1         sample1_rep2    /path/to/sample1_rep2.fastq.gz  -       ␣
↪sample1
sample2         sample2_rep1    /path/to/sample2_rep1.fastq.gz  -       ␣
↪sample2
sample2         sample2_rep2    /path/to/sample2_rep2.fastq.gz  -       ␣
↪sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path: The path to the fastq file to be processed. Can be gzipped or not.
4. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
5. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Pair-End Config This is the config file format for pair-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1_R1.fastq.gz /path/to/
↪sample1_rep1_R2.fastq.gz     -       sample1
sample1         sample1_rep2    /path/to/sample1_rep2_R1.fastq.gz /path/to/
↪sample1_rep2_R2.fastq.gz     -       sample1
sample2         sample2_rep1    /path/to/sample2_rep1_R1.fastq.gz /path/to/
↪sample2_rep1_R2.fastq.gz     -       sample2
sample2         sample2_rep2    /path/to/sample2_rep2_R1.fastq.gz /path/to/
↪sample2_R2_rep1.fastq.gz     -       sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path1: The path to the first fastq file to be processed. Can be gzipped or not.
4. Path2: The path to the second fastq file to be processed. Can be gzipped or not.
5. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
6. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Simple MNase-seq Tutorial (single-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode dnase --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75
```

Simple MNase-seq Tutorial (pair-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode dnase --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib p --readLen 75
```

Simple MNase-seq Tutorial (single-end, 75 length reads, use bowtie2 aligner instead of default bbmap, use 5 threads)

```
nextflow run /path/to/main.nf --mode dnase --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75 --aligner bowtie2 -
↪-threads 5
```

CHAPTER 9

MNase-seq Workflow

All workflows require the following files:

1. A config file (described below)
2. Reference genome FASTA file
3. Reference genome GTF file

Config File A config file is a tab separated text file that includes information regarding the name, location, and input of your experiment.

Single-End Config This is the config file format for single-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1.fastq.gz  -       ␣
↪sample1
sample1         sample1_rep2    /path/to/sample1_rep2.fastq.gz  -       ␣
↪sample1
sample2         sample2_rep1    /path/to/sample2_rep1.fastq.gz  -       ␣
↪sample2
sample2         sample2_rep2    /path/to/sample2_rep2.fastq.gz  -       ␣
↪sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path: The path to the fastq file to be processed. Can be gzipped or not.
4. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
5. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Pair-End Config This is the config file format for pair-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1_R1.fastq.gz /path/to/
↪sample1_rep1_R2.fastq.gz      -       sample1
sample1         sample1_rep2    /path/to/sample1_rep2_R1.fastq.gz /path/to/
↪sample1_rep2_R2.fastq.gz      -       sample1
sample2         sample2_rep1    /path/to/sample2_rep1_R1.fastq.gz /path/to/
↪sample2_rep1_R2.fastq.gz      -       sample2
sample2         sample2_rep2    /path/to/sample2_rep2_R1.fastq.gz /path/to/
↪sample2_R2_rep1.fastq.gz      -       sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path1: The path to the first fastq file to be processed. Can be gzipped or not.
4. Path2: The path to the second fastq file to be processed. Can be gzipped or not.
5. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
6. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Simple MNase-seq Tutorial (single-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode mnase --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75
```

Simple MNase-seq Tutorial (pair-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode mnase --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib p --readLen 75
```

Simple MNase-seq Tutorial (single-end, 75 length reads, use bowtie2 aligner instead of default bbmap, use 5 threads)

```
nextflow run /path/to/main.nf --mode mnase --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75 --aligner bowtie2 -
↪-threads 5
```

CHAPTER 10

GRO-seq Workflow

All workflows require the following files:

1. A config file (described below)
2. Reference genome FASTA file
3. Reference genome GTF file

Config File A config file is a tab separated text file that includes information regarding the name, location, and input of your experiment.

Single-End Config This is the config file format for single-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1.fastq.gz  -       ␣
↪sample1
sample1         sample1_rep2    /path/to/sample1_rep2.fastq.gz  -       ␣
↪sample1
sample2         sample2_rep1    /path/to/sample2_rep1.fastq.gz  -       ␣
↪sample2
sample2         sample2_rep2    /path/to/sample2_rep2.fastq.gz  -       ␣
↪sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path: The path to the fastq file to be processed. Can be gzipped or not.
4. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
5. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Pair-End Config This is the config file format for pair-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1_R1.fastq.gz /path/to/
↪sample1_rep1_R2.fastq.gz      -       sample1
sample1         sample1_rep2    /path/to/sample1_rep2_R1.fastq.gz /path/to/
↪sample1_rep2_R2.fastq.gz      -       sample1
sample2         sample2_rep1    /path/to/sample2_rep1_R1.fastq.gz /path/to/
↪sample2_rep1_R2.fastq.gz      -       sample2
sample2         sample2_rep2    /path/to/sample2_rep2_R1.fastq.gz /path/to/
↪sample2_R2_rep1.fastq.gz      -       sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path1: The path to the first fastq file to be processed. Can be gzipped or not.
4. Path2: The path to the second fastq file to be processed. Can be gzipped or not.
5. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
6. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Simple MNase-seq Tutorial (single-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode gro --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75
```

Simple MNase-seq Tutorial (pair-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode gro --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib p --readLen 75
```

Simple MNase-seq Tutorial (single-end, 75 length reads, use bowtie2 aligner instead of default bbmap, use 5 threads)

```
nextflow run /path/to/main.nf --mode gro --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75 --aligner bowtie2 -
↪-threads 5
```

CHAPTER 11

ATAC-seq Workflow

All workflows require the following files:

1. A config file (described below)
2. Reference genome FASTA file
3. Reference genome GTF file

Config File A config file is a tab separated text file that includes information regarding the name, location, and input of your experiment.

Single-End Config This is the config file format for single-ended data.

```
sample1     sample1_rep1    /path/to/sample1_rep1.fastq.gz  -    sample1
sample1     sample1_rep2    /path/to/sample1_rep2.fastq.gz  -    sample1
sample2     sample2_rep1    /path/to/sample2_rep1.fastq.gz  control2   ␣
↪sample2
sample2     sample2_rep2    /path/to/sample2_rep2.fastq.gz  control2   ␣
↪sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path: The path to the fastq file to be processed. Can be gzipped or not.
4. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
5. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Pair-End Config This is the config file format for pair-ended data.

```
sample1         sample1_rep1    /path/to/sample1_rep1_R1.fastq.gz /path/to/
↪sample1_rep1_R2.fastq.gz      -       sample1
sample1         sample1_rep2    /path/to/sample1_rep2_R1.fastq.gz /path/to/
↪sample1_rep2_R2.fastq.gz      -       sample1
sample2         sample2_rep1    /path/to/sample2_rep1_R1.fastq.gz /path/to/
↪sample2_rep1_R2.fastq.gz      -       sample2
sample2         sample2_rep2    /path/to/sample2_rep2_R1.fastq.gz /path/to/
↪sample2_R2_rep1.fastq.gz      -       sample2
```

The columns represent:

1. MergeID: The merge ID that will be used should your files be merged together. Should be the same for all replicates.
2. ID: The ID that will be used to name the majority of your files that are not merged. Recommended to be used to differentiate between different technical replicates.
3. Path1: The path to the first fastq file to be processed. Can be gzipped or not.
4. Path2: The path to the second fastq file to be processed. Can be gzipped or not.
5. ControlID: The ID indicating what control file to be used for peak calling and other downstream analysis. Use "-" (without quotes) if there is no control for a particular sample.
6. Mark: The ID that signifies the type of mark or histone being processed. Use "input" if the line refers to a control. If the line is NOT a control, then use the MergeID name.

Simple MNase-seq Tutorial (single-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode atac --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75
```

Simple MNase-seq Tutorial (pair-end, 75 length reads)

```
nextflow run /path/to/main.nf --mode atac --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib p --readLen 75
```

Simple MNase-seq Tutorial (single-end, 75 length reads, use bowtie2 aligner instead of default bbmap, use 5 threads)

```
nextflow run /path/to/main.nf --mode atac --config /path/to/config.txt --fasta /
↪path/to/fasta.fa --gtf /path/to/gtf.gtf --lib s --readLen 75 --aligner bowtie2 -
↪-threads 5
```

CHAPTER 12

Indices and tables

- genindex
- modindex
- search

CIPHER is a workflow platform written in the Nextflow DSL that was developed to enhance reproducibility among research, and to simplify data processing for non-computational scientists. CIPHER can be used for the efficient pre-processing and analysis of high-throughput sequencing data including: ChIP-seq, RNA-seq, DNase-seq, MNase-seq, GRO-seq, and ATAC-seq.

For support, questions, or feature requests contact: cag104@ucsd.edu or submit an issue at our github.

CIPHER is run from the command line, can be run on a local desktop or HPC and only requires the installation of Nextflow and Docker (optionally). Please visit the 'Installation' page for more details.