Week 3

- Unix Problem Set Q&A
- Review ChIP-Seq and RNA-Seq pipelines

- Loading files to the cluster using FileZilla and rsync

- On the ghpcc06 cluster, read criteria for running bowtie2, picard, samtools, macs2, etc.
- See sample bsub commands for running each module
- Look at bowtie2 and macs2 output files
- Use IGV to view .bam and .bed files

- On the ghpcc06 cluster, read criteria for running rsem-calculate-expression on the cluster
  https://bioinfo.umassmed.edu/index.php?p=33
- RNA-Seq – Look at RSEM output files
- See sample bsub commands for rsem-calculate-expression

Tasks:
Create lab schedule for using the ghpcc06 cluster
Explore the cluster and examine modules and tools for data analysis that are available
Upload Fastq files to ghpcc06 cluster

Week 4

- Introducing NGSplot: Quick mining and visualization of NGS data by integrating genomic databases
  https://github.com/shenlab-sinai/ngsplot
- Introducing HOMER: Software for motif discovery and next-gen sequencing analysis
  http://homer.ucsd.edu/homer/ngs/

- Running NGSplot on the ghpcc06 cluster for making ChIP-Seq and RNA-Seq heatmaps and extracting tag density files
- Running HOMER on the ghpcc06 cluster for annotating bed files and performing motif analysis
- Running bedtools on the ghpcc06 cluster for comparing .bed files of different experiments

Week 3

- Unix Problem Set Q&A
- Review ChIP-Seq and RNA-Seq pipelines

- Loading files to the cluster using FileZilla and rsync

- On the ghpcc06 cluster, read criteria for running bowtie2, picard, samtools, macs2, etc.
- See sample bsub commands for running each module
- Look at bowtie2 and macs2 output files
- Use IGV to view .bam and .bed files

- On the ghpcc06 cluster, read criteria for running rsem-calculate-expression on the cluster
  https://bioinfo.umassmed.edu/index.php?p=33
- RNA-Seq – Look at RSEM output files
- See sample bsub commands for rsem-calculate-expression

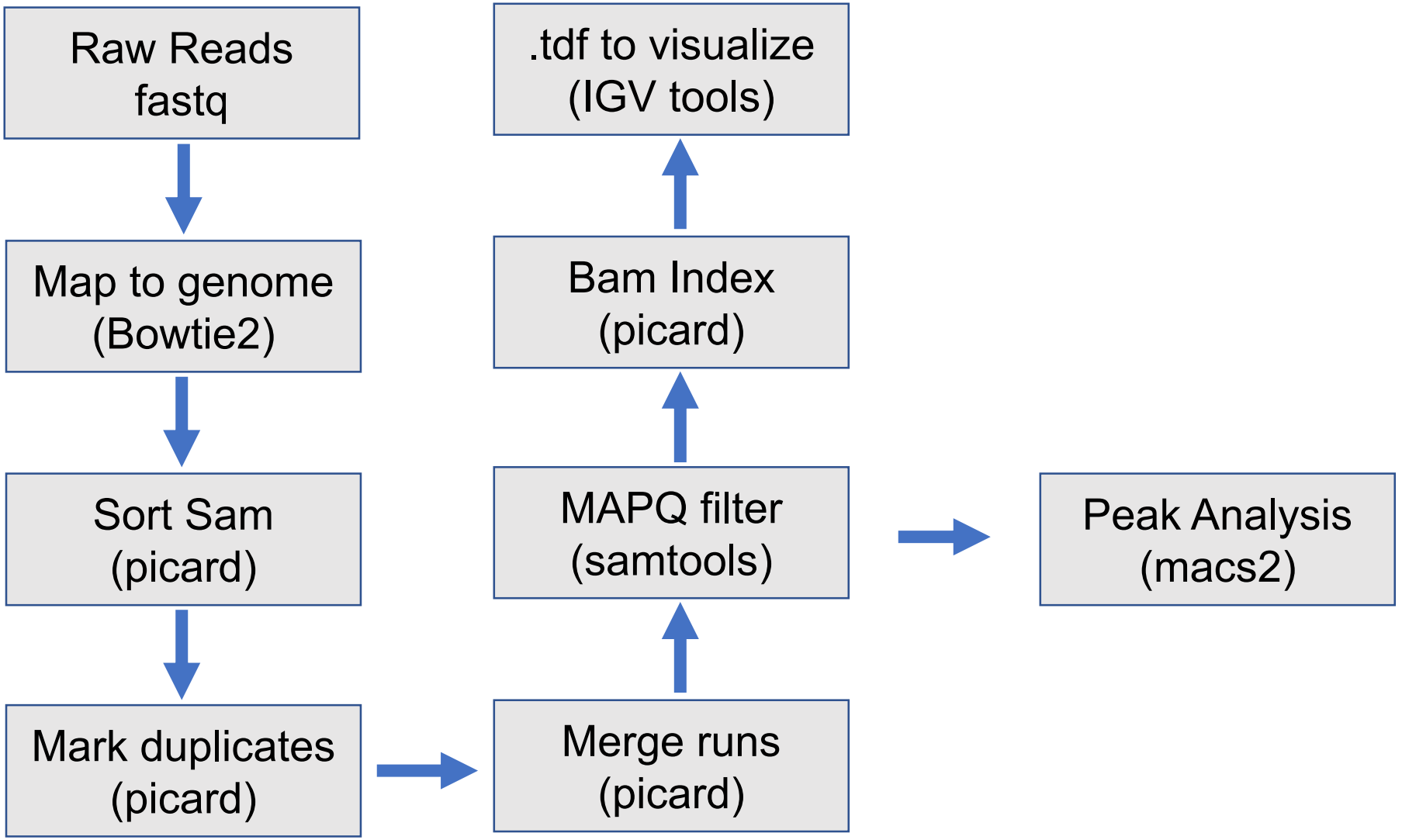
Tasks:
Create lab schedule for using the ghpcc06 cluster
Explore the cluster and examine modules and tools for data analysis that are available
Upload Fastq files to ghpcc06 cluster

Week 4

- Introducing NGSplot: Quick mining and visualization of NGS data by integrating genomic databases
  https://github.com/shenlab-sinai/ngsplot
- Introducing HOMER: Software for motif discovery and next-gen sequencing analysis
  http://homer.ucsd.edu/homer/ngs/

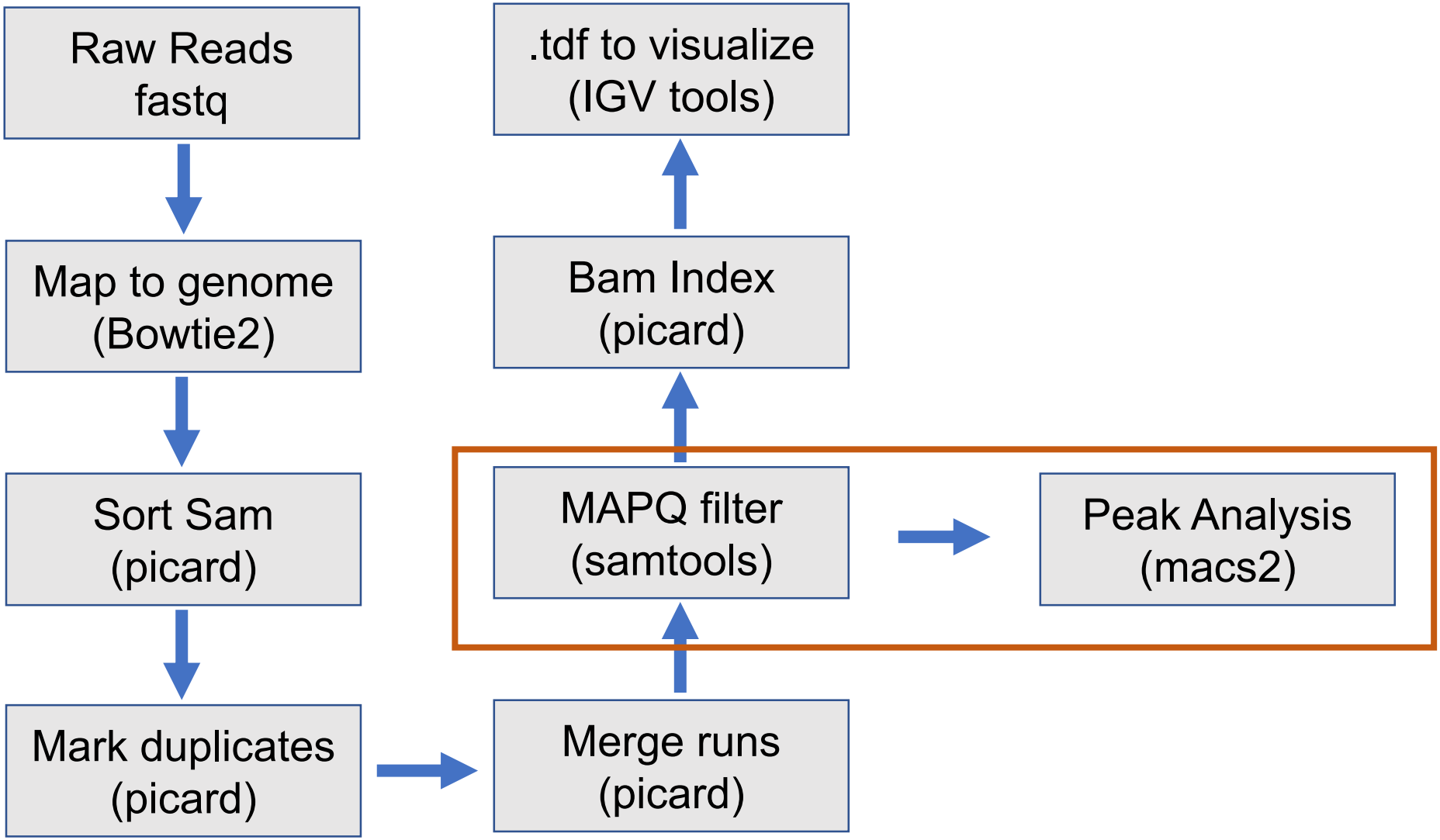
- Running NGSplot on the ghpcc06 cluster for making ChIP-Seq and RNA-Seq heatmaps and extracting tag density files
- Running HOMER on the ghpcc06 cluster for annotating bed files and performing motif analysis
- Running bedtools on the ghpcc06 cluster for comparing .bed files of different experiments

My ChIP-Sequencing Workflow

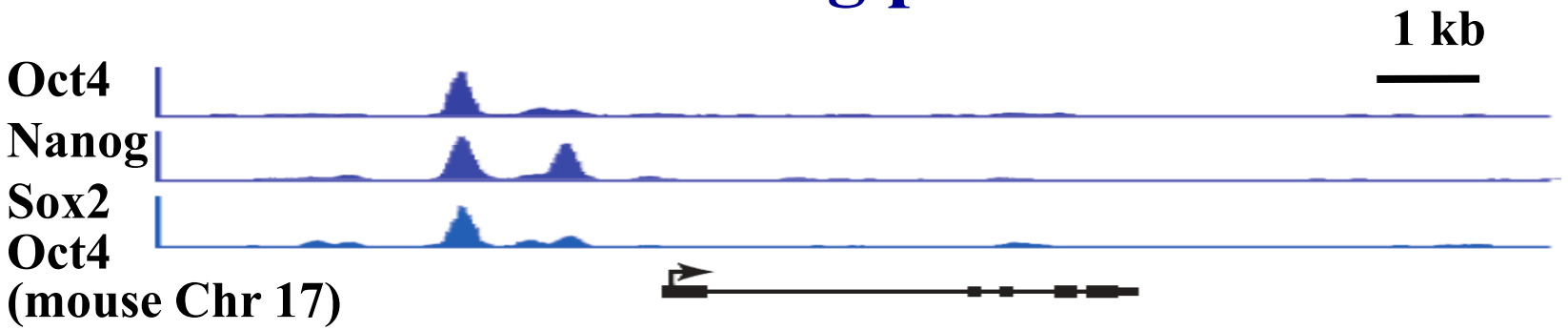


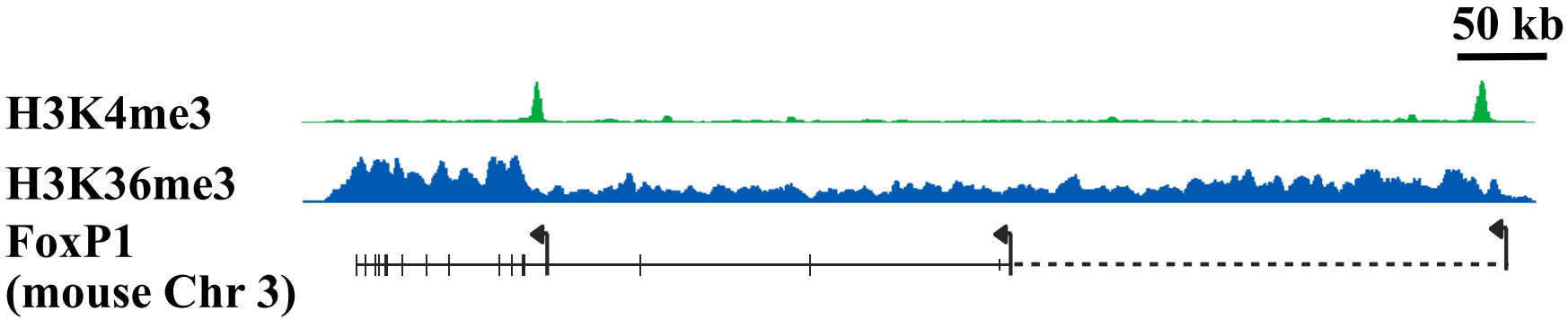
```
[ss45w@ghpcc06 mergedhr4D1_Prmt5ChIP]$ cd /nl/umw_anthony_imbalzano/Sabriya/TereChIP/
[ss45w@ghpcc06 TereChIP]$
[ss45w@ghpcc06 TereChIP]$
[ss45w@ghpcc06 TereChIP]$ ls -lh
total 8.8G
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano  63M Jun 22 12:00 log
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 2.7K Jun 22 15:04 log3
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 3.9K Jun 22 09:59 logPicard
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 4.2K Jun 22 11:55 logS
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano  20K Jun 22 14:37 logW
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 9.0K Jun 22 12:16 logX
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 6.4K Jun 22 12:06 logZ
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 1.4K Jun 22 14:18 markdup_metrics.txt
-rwx------ 1 ss45w umw_anthony_imbalzano 574M Jun 21 23:38 ProlnoCu_IN_rep1.fq.gz
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 4.5G Jun 22 08:44 ProlnoCu_IN_rep1.sam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 722M Jun 22 12:06 ProlnoCu_IN_rep1.sorted.bam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 659M Jun 22 13:08 ProlnoCu_IN_rep1_sorted.dedup.bam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 5.3M Jun 22 15:04 ProlnoCu_IN_rep1_sorted.dedup.q20.bai
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 398M Jun 22 14:37 ProlnoCu_IN_rep1_sorted.dedup.q20.bam
```

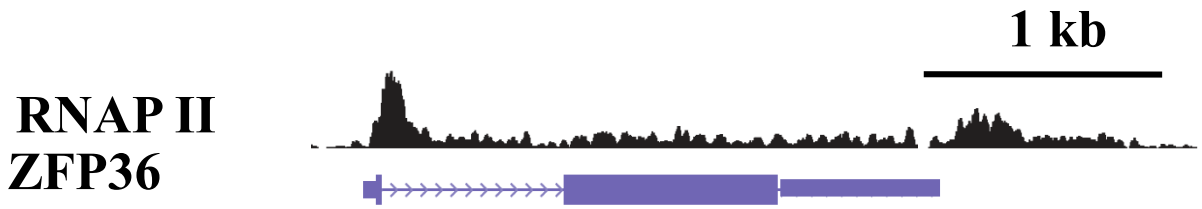
My ChIP-Sequencing Workflow



Binding profile







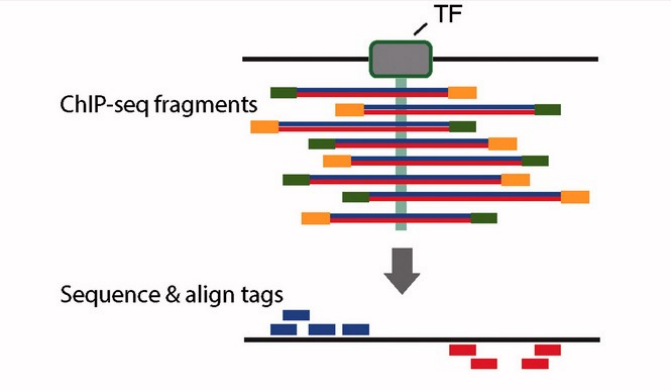
Kagey et al. (2010). Nature 467: 430
Mikkelsen et al. (2007). Nature 448: 553
Pepke et al. (2009). Nat. Methods 6: S22

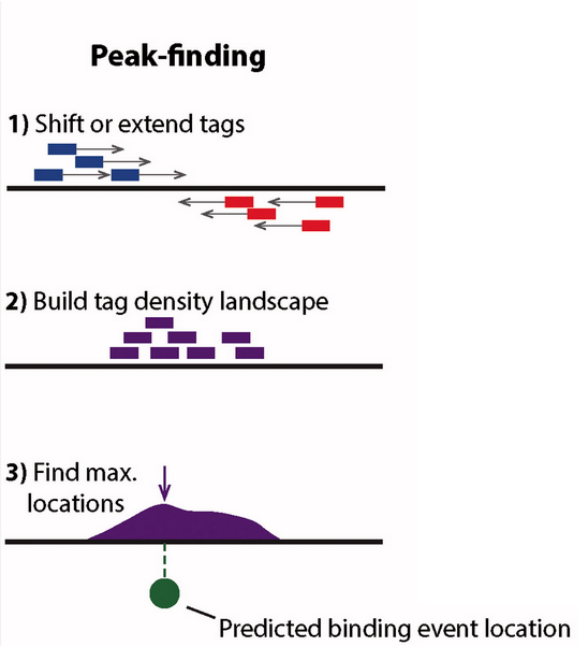
MACS2

A commonly used tool for identifying transcription factor binding sites is named Model-based Analysis of ChIP-seq (MACS). The MACS algorithm captures the influence of genome complexity to evaluate the significance of enriched ChIP regions. Although it was developed for the detection of transcription factor binding sites it is also suited for larger regions.

MACS improves the spatial resolution of binding sites through **combining the information of both sequencing tag position and orientation.** MACS can be easily used either for the ChIP sample alone, or along with a control sample which increases specificity of the peak calls. The MACS workflow

https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html





Mahony and Pugh, 2015, Critical Reviews in Biochemistry and Molecular Biology, Volume 50, 2015 - Issue 4

```
[ss45w@ghpcc06 mergedhr4D1_Prmt5ChIP]$ module load macs/1.4.2
 openssl 1.0.1q is located under /share/pkg/openssl/1.0.1q
 python 2.7.5 is located under /share/pkg/python/2.7.5
[ss45w@ghpcc06 mergedhr4D1_Prmt5ChIP]$ macs2 callpeak -h
usage: macs2 callpeak [-h] -t TFILE [TFILE ...] [-c [CFILE [CFILE ...]]]
                      [-f {AUTO,BAM,SAM,BED,ELAND,ELANDMULTI,ELANDEXPORT,BOWTIE,BAMPE,BEDPE}]
                      [-g GSIZE] [--keep-dup KEEPDUPLICATES]
                      [--buffer-size BUFFER_SIZE] [--outdir OUTDIR] [-n NAME]
                      [-B] [--verbose VERBOSE] [--trackline] [--SPMR]
                      [-s TSIZE] [--bw BW] [-m MFOLD MFOLD] [--fix-bimodal]
                      [--nomodel] [--shift SHIFT] [--extsize EXTSIZE]
                      [-q QVALUE] [-p PVALUE] [--to-large] [--ratio RATIO]
                      [--down-sample] [--seed SEED] [--tempdir TEMPDIR]
                      [--nolambda] [--slocal SMALLLOCAL] [--llocal LARGELOCAL]
                      [--broad] [--broad-cutoff BROADCUTOFF]
                      [--cutoff-analysis] [--call-summits]
                      [--fe-cutoff FECUTOFF]

optional arguments:
  -h, --help            show this help message and exit

Input files arguments:
  -t TFILE [TFILE ...], --treatment TFILE [TFILE ...]
                        ChIP-seq treatment file. If multiple files are given
                        as '-t A B C', then they will all be read and pooled
                        together. REQUIRED.
  -c [CFILE [CFILE ...]], --control [CFILE [CFILE ...]]
                        Control file. If multiple files are given as '-c A B
                        C', they will be pooled to estimate ChIP-seq
                        background noise.
  -f {AUTO,BAM,SAM,BED,ELAND,ELANDMULTI,ELANDEXPORT,BOWTIE,BAMPE,BEDPE}, --format {AUTO,BAM,SAM,BED,ELAND,ELANDMULTI,ELANDEXP
ORT,BOWTIE,BAMPE,BEDPE}
                        Format of tag file, "AUTO", "BED" or "ELAND" or
                        "ELANDMULTI" or "ELANDEXPORT" or "SAM" or "BAM" or
                        "BOWTIE" or "BAMPE" or "BEDPE". The default AUTO
                        option will let MACS decide which format (except for
                        BAMPE and BEDPE which should be implicitly set) the
                        file is. Please check the definition in README. Please
                        note that if the format is set as BAMPE or BEDPE,
                        MACS2 will call its special Paired-end mode to call
                        peaks by piling up the actual ChIPed fragments defined
                        by both aligned ends, instead of predicting the
                        fragment size first and extending reads. Also please
                        note that the BEDPE only contains three columns, and
                        is NOT the same BEDPE format used by BEDTOOLS.
                        DEFAULT: "AUTO"
  -g GSIZE, --gsize GSIZE
                        Effective genome size. It can be 1.0e+9 or 1000000000,
                        or shortcuts:'hs' for human (2.7e9), 'mm' for mouse
                        (1.87e9), 'ce' for C. elegans (9e7) and 'dm' for
```

```
                        fruitfly (1.2e8), Default:hs
  --keep-dup KEEPDUPLICATES
                        It controls the MACS behavior towards duplicate tags
                        at the exact same location -- the same coordination
                        and the same strand. The 'auto' option makes MACS
                        calculate the maximum tags at the exact same location
                        based on binomal distribution using 1e-5 as pvalue
                        cutoff; and the 'all' option keeps every tags. If an
                        integer is given, at most this number of tags will be
                        kept at the same location. The default is to keep one
                        tag at the same location. Default: 1
  --buffer-size BUFFER_SIZE
                        Buffer size for incrementally increasing internal
                        array size to store reads alignment information. In
                        most cases, you don't have to change this parameter.
                        However, if there are large number of
                        chromosomes/contigs/scaffolds in your alignment, it's
                        recommended to specify a smaller buffer size in order
                        to decrease memory usage (but it will take longer time
                        to read alignment files). Minimum memory requested for
                        reading an alignment file is about # of CHROMOSOME *
                        BUFFER_SIZE * 2 Bytes. DEFAULT: 100000

Output arguments:
  --outdir OUTDIR       If specified all output files will be written to that
                        directory. Default: the current working directory
  -n NAME, --name NAME  Experiment name, which will be used to generate output
                        file names. DEFAULT: "NA"
  -B, --bdg             Whether or not to save extended fragment pileup, and
                        local lambda tracks (two files) at every bp into a
                        bedGraph file. DEFAULT: False
  --verbose VERBOSE     Set verbose level of runtime message. 0: only show
                        critical message, 1: show additional warning message,
                        2: show process information, 3: show debug messages.
                        DEFAULT:2
  --trackline           Tells MACS to include trackline with bedGraph files.
                        To include this trackline while displaying bedGraph at
                        UCSC genome browser, can show name and description of
                        the file as well. However my suggestion is to convert
                        bedGraph to bigWig, then show the smaller and faster
                        binary bigWig file at UCSC genome browser, as well as
                        downstream analysis. Require -B to be set. Default:
                        Not include trackline.
  --SPMR                If True, MACS will save signal per million reads for
                        fragment pileup profiles. Require -B to be set.
                        Default: False
```

```
Shifting model arguments:
  -s TSIZE, --tsize TSIZE
                        Tag size. This will overide the auto detected tag
                        size. DEFAULT: Not set
  --bw BW               Band width for picking regions to compute fragment
                        size. This value is only used while building the
                        shifting model. DEFAULT: 300
  -m MFOLD MFOLD, --mfold MFOLD MFOLD
                        Select the regions within MFOLD range of high-
                        confidence enrichment ratio against background to
                        build model. Fold-enrichment in regions must be lower
                        than upper limit, and higher than the lower limit. Use
                        as "-m 10 30". DEFAULT:5 50
  --fix-bimodal         Whether turn on the auto pair model process. If set,
                        when MACS failed to build paired model, it will use
                        the nomodel settings, the --exsize parameter to extend
                        each tags towards 3' direction. Not to use this
                        automate fixation is a default behavior now. DEFAULT:
                        False
  --nomodel             Whether or not to build the shifting model. If True,
                        MACS will not build model. by default it means
                        shifting size = 100, try to set extsize to change it.
                        DEFAULT: False
  --shift SHIFT         (NOT the legacy --shiftsize option!) The arbitrary
                        shift in bp. Use discretion while setting it other
                        than default value. When NOMODEL is set, MACS will use
                        this value to move cutting ends (5') towards 5'->3'
                        direction then apply EXTSIZE to extend them to
                        fragments. When this value is negative, ends will be
                        moved toward 3'->5' direction. Recommended to keep it
                        as default 0 for ChIP-Seq datasets, or -1 * half of
                        EXTSIZE together with EXTSIZE option for detecting
                        enriched cutting loci such as certain DNAseI-Seq
                        datasets. Note, you can't set values other than 0 if
                        format is BAMPE or BEDPE for paired-end data. DEFAULT:
                        0.
  --extsize EXTSIZE     The arbitrary extension size in bp. When nomodel is
                        true, MACS will use this value as fragment size to
                        extend each read towards 3' end, then pile them up.
                        It's exactly twice the number of obsolete SHIFTSIZE.
                        In previous language, each read is moved 5'->3'
                        direction to middle of fragment by 1/2 d, then
                        extended to both direction with 1/2 d. This is
                        equivalent to say each read is extended towards 5'->3'
                        into a d size fragment. DEFAULT: 200. EXTSIZE and
                        SHIFT can be combined when necessary. Check SHIFT
```

```
[ss45w@ghpcc06 TereChIP]$ less logW
[ss45w@ghpcc06 TereChIP]$ ls -lh
total 15G
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano  63M Jun 23 17:03 log
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 2.7K Jun 22 15:04 log3
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 6.8K Jun 23 18:35 log4
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 3.9K Jun 22 09:59 logPicard
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 4.2K Jun 22 11:55 logS
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano  23K Jun 24 09:50 logW
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano  18K Jun 23 18:57 logX
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 6.4K Jun 22 12:06 logZ
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 1.4K Jun 23 18:57 markdup_metrics2.txt
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 1.4K Jun 22 14:18 markdup_metrics.txt
-rwx------ 1 ss45w umw_anthony_imbalzano 574M Jun 21 23:38 ProlnoCu_IN_rep1.fq.gz
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 4.5G Jun 22 08:44 ProlnoCu_IN_rep1.sam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 722M Jun 22 12:06 ProlnoCu_IN_rep1.sorted.bam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 659M Jun 22 13:08 ProlnoCu_IN_rep1_sorted.dedup.bam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 5.3M Jun 22 15:04 ProlnoCu_IN_rep1_sorted.dedup.q20.bai
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 398M Jun 22 14:37 ProlnoCu_IN_rep1_sorted.dedup.q20.bam
-rwx------ 1 ss45w umw_anthony_imbalzano 456M Jun 21 23:20 ProlnoCu_IP_rep1.fq.gz
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 3.4G Jun 23 17:32 ProlnoCu_IP_rep1.sam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 617M Jun 23 18:35 ProlnoCu_IP_rep1.sorted.bam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 586M Jun 23 18:57 ProlnoCu_IP_rep1.sorted.dedup.bam
-rw-rw-r-- 1 ss45w umw_anthony_imbalzano 391M Jun 24 09:50 ProlnoCu_IP_rep1.sorted.dedup.q20.bam
```

#Peak analysis
IP.sorted.dedup.q20.bam and Input.sorted.dedup.q20.bam --> .peak analysis using macs2

#example

module load macs/1.4.2

bsub -q short -o log5 -n 2 -R rusage[mem=20000] -W120 -R span[hosts=1] "macs2 callpeak -t ProlnoCu_IP_rep1.sorted.dedup.q20.bam -c ProlnoCu_IN_rep1_sorted.dedup.q20.bam -f BAM -g mm -q 0.05 -n ProlnoCu _rep1_q0.05"

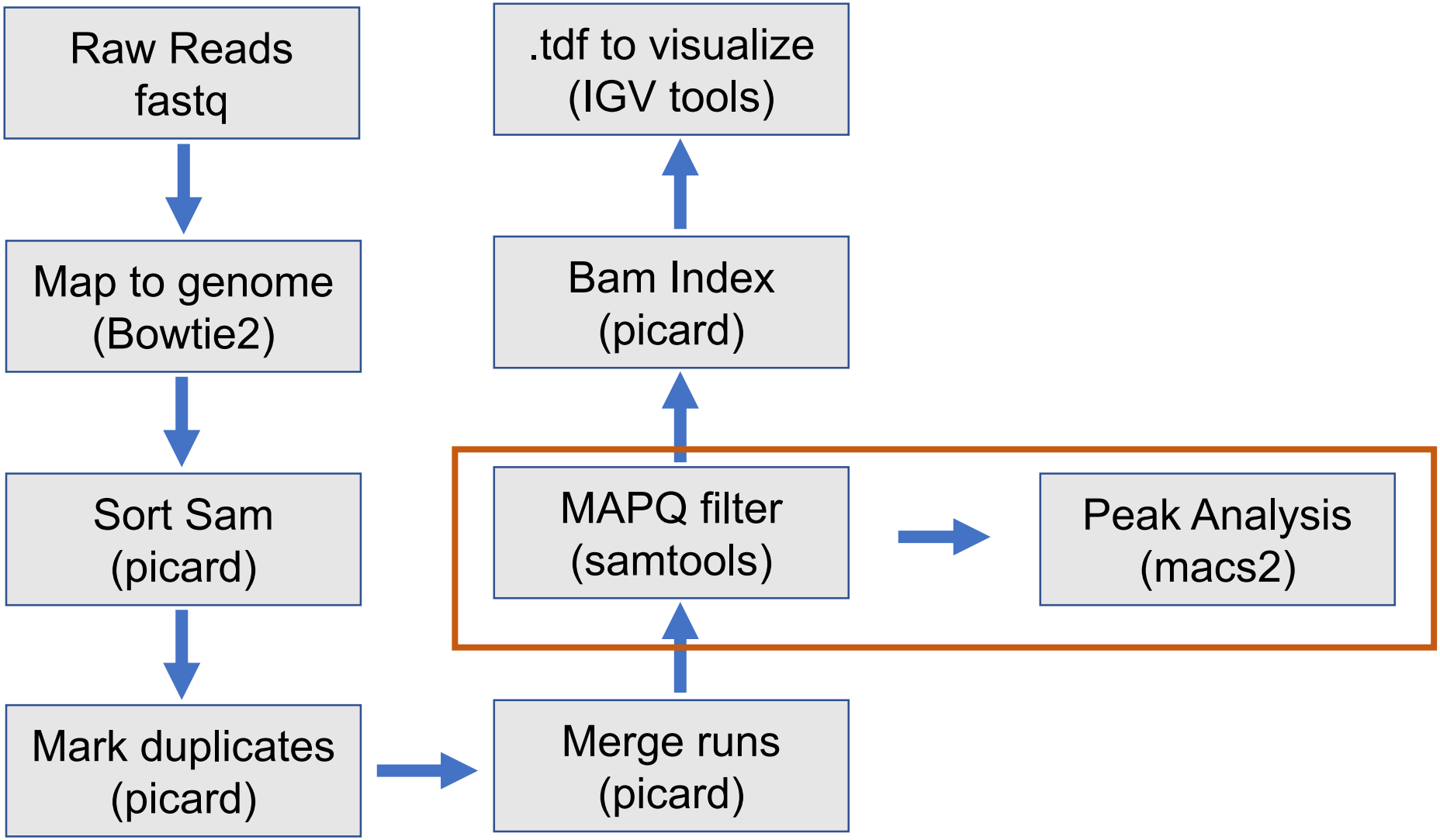
If ChIP-Seq is paired-end:

bsub -q short -n 1 -R rusage[mem=20000] -W120 "macs2 callpeak -t ProlnoCu_IP_rep1_sorted.dedup.q20.bam -c ProlnoCu_IN_rep1_sorted.dedup.q20.bam -f BAMPE -g mm -q 0.05 -n D1Prmt5_n1n2_q0.05_PE"

For broader peaks, another option (but better broad peak software exists):

bsub -q short -o log -n 1 -R rusage[mem=20000] -W120 "macs2 callpeak -t ProlnoCu_IP_rep1_sorted.dedup.q20.bam -c ProlnoCu_IN_rep1_sorted.dedup.q20.bam -f BAM -g mm --broad --broad-cutoff 0.05 -n D2_H3k27me3_q0.1"

My ChIP-Sequencing Workflow



Run Report

```
INFO  @ Wed, 24 Jun 2020 09:59:06: #2 finished!
INFO  @ Wed, 24 Jun 2020 09:59:06: #2 predicted fragment length is 100 bps
INFO  @ Wed, 24 Jun 2020 09:59:06: #2 alternative fragment length(s) may be 100 bps
INFO  @ Wed, 24 Jun 2020 09:59:06: #2.2 Generate R script for model : ProlnoCu_rep1_q0.05_model.r
INFO  @ Wed, 24 Jun 2020 09:59:06: #3 Call peaks...
INFO  @ Wed, 24 Jun 2020 09:59:06: #3 Pre-compute pvalue-qvalue table...
INFO  @ Wed, 24 Jun 2020 09:59:38: #3 Call peaks for each chromosome...
INFO  @ Wed, 24 Jun 2020 09:59:56: #4 Write output xls file... ProlnoCu_rep1_q0.05_peaks.xls
INFO  @ Wed, 24 Jun 2020 09:59:56: #4 Write peak in narrowPeak format file... ProlnoCu_rep1_q0.05_peaks.narrowPeak
INFO  @ Wed, 24 Jun 2020 09:59:56: #4 Write summits bed file... ProlnoCu_rep1_q0.05_summits.bed
INFO  @ Wed, 24 Jun 2020 09:59:56: Done!

------------------------------------------------------------
Sender: LSF System <lsfadmin@c38b03>
Subject: Job 7737666: <macs2 callpeak -t ProlnoCu_IP_rep1.sorted.dedup.q20.bam -c ProlnoCu_IN_rep1_sorted.dedup.q20.bam -f BAM -g mm -q 0.05 -n Pro
lnoCu_rep1_q0.05> in cluster <umghpcc> Done

Job <macs2 callpeak -t ProlnoCu_IP_rep1.sorted.dedup.q20.bam -c ProlnoCu_IN_rep1_sorted.dedup.q20.bam -f BAM -g mm -q 0.05 -n ProlnoCu_rep1_q0.05>
was submitted from host <ghpcc06> by user <ss45w> in cluster <umghpcc> at Wed Jun 24 09:57:56 2020
Job was executed on host(s) <2*c38b03>, in queue <short>, as user <ss45w> in cluster <umghpcc> at Wed Jun 24 09:57:57 2020
</home/ss45w> was used as the home directory.
</nl/umw_anthony_imbalzano/Sabriya/TereChIP> was used as the working directory.
Started at Wed Jun 24 09:57:57 2020
Terminated at Wed Jun 24 09:59:56 2020
Results reported at Wed Jun 24 09:59:56 2020

Your job looked like:

------------------------------------------------------------
# LSBATCH: User input
macs2 callpeak -t ProlnoCu_IP_rep1.sorted.dedup.q20.bam -c ProlnoCu_IN_rep1_sorted.dedup.q20.bam -f BAM -g mm -q 0.05 -n ProlnoCu_rep1_q0.05
------------------------------------------------------------

Successfully completed.

Resource usage summary:

    CPU time :                                   113.48 sec.
    Max Memory :                                 177 MB
    Average Memory :                             115.50 MB
    Total Requested Memory :                     40000.00 MB
    Delta Memory :                               39823.00 MB
    Max Swap :                                   -
    Max Processes :                              3
    Max Threads :                                4
    Run time :                                   121 sec.
    Turnaround time :                            120 sec.
```

MACS output

```
# This file is generated by MACS version 2.1.1.20160226
# Command line: callpeak -t D0_Prmt5_n1n2merged_newdedup.bam -c D0_Input_n1n2merged_newdedup.bam -f BAMPE -g mm -q 0.05 -n D0Prmt5_n1_q0.0.5_PE_new
dedup
# ARGUMENTS LIST:
# name = D0Prmt5_n1_q0.0.5_PE_newdedup
# format = BAMPE
# ChIP-seq file = ['D0_Prmt5_n1n2merged_newdedup.bam']
# control file = ['D0_Input_n1n2merged_newdedup.bam']
# effective genome size = 1.87e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is off
# Paired-End mode is on

# fragment size is determined as 112 bps
# total fragments in treatment: 46902866
# fragments after filtering in treatment: 46902865
# maximum duplicate fragments in treatment = 1
# Redundant rate in treatment: 0.00
# total fragments in control: 52039179
# fragments after filtering in control: 52039179
# maximum duplicate fragments in control = 1
# Redundant rate in control: 0.00
# d = 112
chr     start   end     length  abs_summit      pileup  -log10(pvalue)  fold_enrichment -log10(qvalue)  name
chr1    5019461 5019631 171     5019575 16.00   5.88270 3.37027 3.00123 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_1
chr1    5022840 5023126 287     5022957 20.00   7.83652 3.81368 4.73355 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_2
chr1    5082895 5083069 175     5082917 19.00   6.53375 3.36342 3.56727 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_3
chr1    5128645 5129110 466     5128871 21.00   8.52978 3.99528 5.35599 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_4
chr1    6214521 6214835 315     6214572 26.00   9.06821 3.69403 5.85294 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_5
chr1    7088504 7089069 566     7088922 37.00   23.36820        7.53355 19.52630        D0Prmt5_n1_q0.0.5_PE_newdedup_peak_6
chr1    7131196 7131376 181     7131216 21.00   6.57816 3.20772 3.60012 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_7
chr1    7397602 7398361 760     7397980 723.00  1251.02661      119.72334       1245.72278      D0Prmt5_n1_q0.0.5_PE_newdedup_peak_8
chr1    9545118 9545496 379     9545388 61.00   40.39735        9.03995 36.33358        D0Prmt5_n1_q0.0.5_PE_newdedup_peak_9
chr1    9548327 9548493 167     9548365 15.00   5.25200 3.17202 2.46890 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_10
chr1    9578765 9578966 202     9578839 19.00   5.94720 3.12120 3.05358 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_11
chr1    9748313 9748615 303     9748461 28.00   10.38592        3.96766 7.05751 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_12
chr1    10037675        10038208        534     10038049        25.00   10.15810        4.19443 6.84449 D0Prmt5_n1_q0.0.5_PE_newdedup_peak_13
```

THE FASEB JOURNAL • RESEARCH • www.fasebj.org

The classic metal-sensing transcription factor MTF1 promotes myogenesis in response to copper

Cristina Tavera-Montañez,[*,1] Sarah J. Hainer,[†,1,2] Daniella Cangussu,[*] Shellaina J. V. Gordon,[*] Yao Xiao,[‡] Pablo Reyes-Gutierrez,[*] Anthony N. Imbalzano,[*] Juan G. Navea,[‡] Thomas G. Fazzio,[†] and Teresita Padilla-Benavides[*,3]

[*]Department of Biochemistry and Molecular Pharmacology and [†]Department of Molecular, Cell, and Cancer Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA; and [‡]Department of Chemistry, Skidmore College, Saratoga Springs, New York, USA

Data analysis

Single-end Fastq reads were split by barcode adapter sequences and adapter sequences were removed using the Fastx toolkit. Reads were mapped to the mm10 genome using bowtie, allowing up to 3 mismatches. Aligned reads were processed using Hypergeometric Optimization of Motif EnRichment (HOMER) (95). University of California Santa Cruz (UCSC) UCSC genome browser tracks were generated using the "makeUCSCfile" command. Mapped reads were aligned over all annotated mm10 transcriptional start sites (TSSs) using the "annotatePeaks" command, generating 20 bp bins and summing the reads within each window. After anchoring mapped reads over reference TSSs, aggregation plots were generated by averaging data obtained from 2 biologic replicates. Peaks were called individually from replicate data sets using the "findPeaks" command and then overlapping peaks were identified using the "mergePeaks" command. For peak calling, a false discovery rate of 0.001 was used as a threshold. Motifs were identified using the "findMotifs" command. Analysis of data from GSE24852 (96) was performed similarly. Data were downloaded from GSE24852 and converted to Fastq files using SRAtoolkit Fastq-dump and mapped reads were converted to mm10. Aligned reads were processed in HOMER (95), as previously described.

Week 3

- Unix Problem Set Q&A
- Review ChIP-Seq and RNA-Seq pipelines

- Loading files to the cluster using FileZilla and rsync

- On the ghpcc06 cluster, read criteria for running bowtie2, picard, samtools, macs2, etc.
- See sample bsub commands for running each module
- Look at bowtie2 and macs2 output files
- Use IGV to view .bam and .bed files

- On the ghpcc06 cluster, read criteria for running rsem-calculate-expression on the cluster
  https://bioinfo.umassmed.edu/index.php?p=33
- RNA-Seq – Look at RSEM output files
- See sample bsub commands for rsem-calculate-expression

Tasks:
Create lab schedule for using the ghpcc06 cluster
Explore the cluster and examine modules and tools for data analysis that are available
Upload Fastq files to ghpcc06 cluster

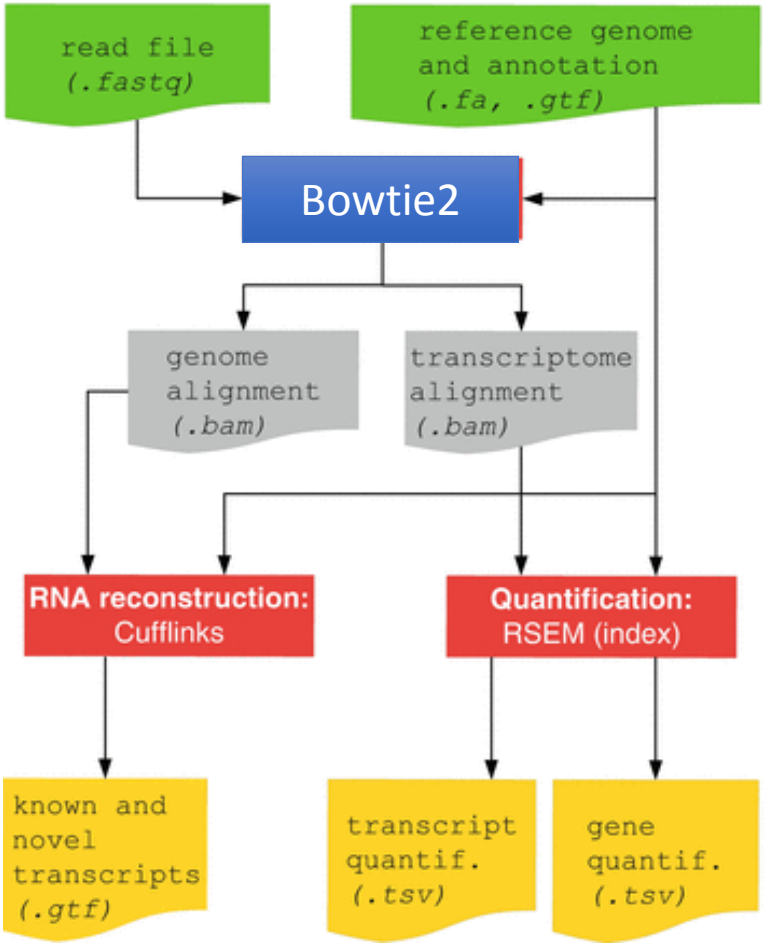
Week 4

- Introducing NGSplot: Quick mining and visualization of NGS data by integrating genomic databases
  https://github.com/shenlab-sinai/ngsplot
- Introducing HOMER: Software for motif discovery and next-gen sequencing analysis
  http://homer.ucsd.edu/homer/ngs/

- Running NGSplot on the ghpcc06 cluster for making ChIP-Seq and RNA-Seq heatmaps and extracting tag density files
- Running HOMER on the ghpcc06 cluster for annotating bed files and performing motif analysis
- Running bedtools on the ghpcc06 cluster for comparing .bed files of different experiments

RNA-Seq Pipline

Methods and Tools

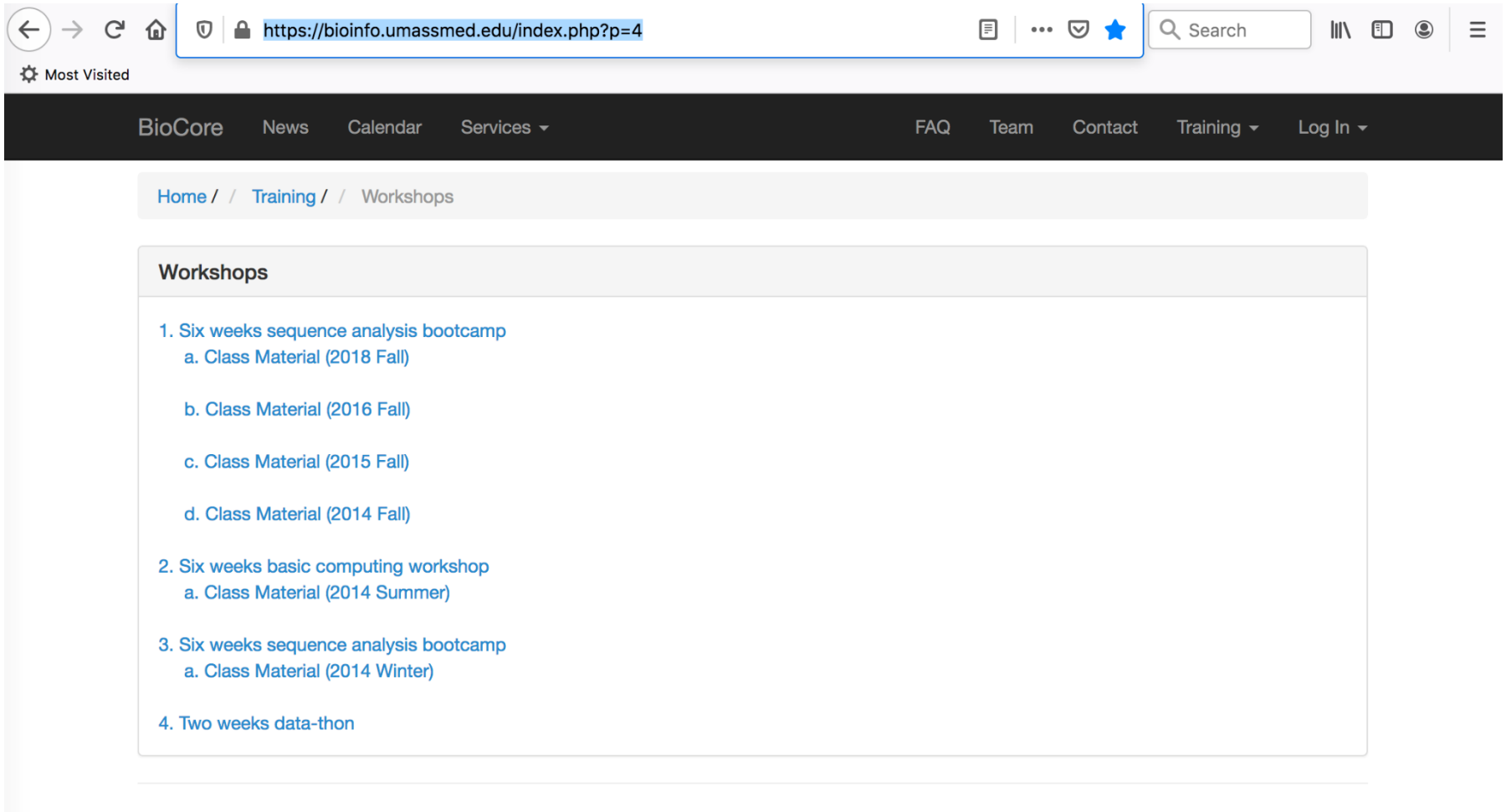


Bowtie and Bowtie2 use Burrows-Wheeler indexing for aligning reads. With bowtie2 there is no upper limit on the read length

Bowtie2 to align reads in a splice-aware manner and aids the discovery of new splice junctions

RSEM is a software package for estimating gene and isoform expression levels from RNA-Seq data. The RSEM package provides an user-friendly interface, supports threads for parallel computation of the EM algorithm, single-end and paired-end read data, quality scores, variable-length reads and RSPD estimation.

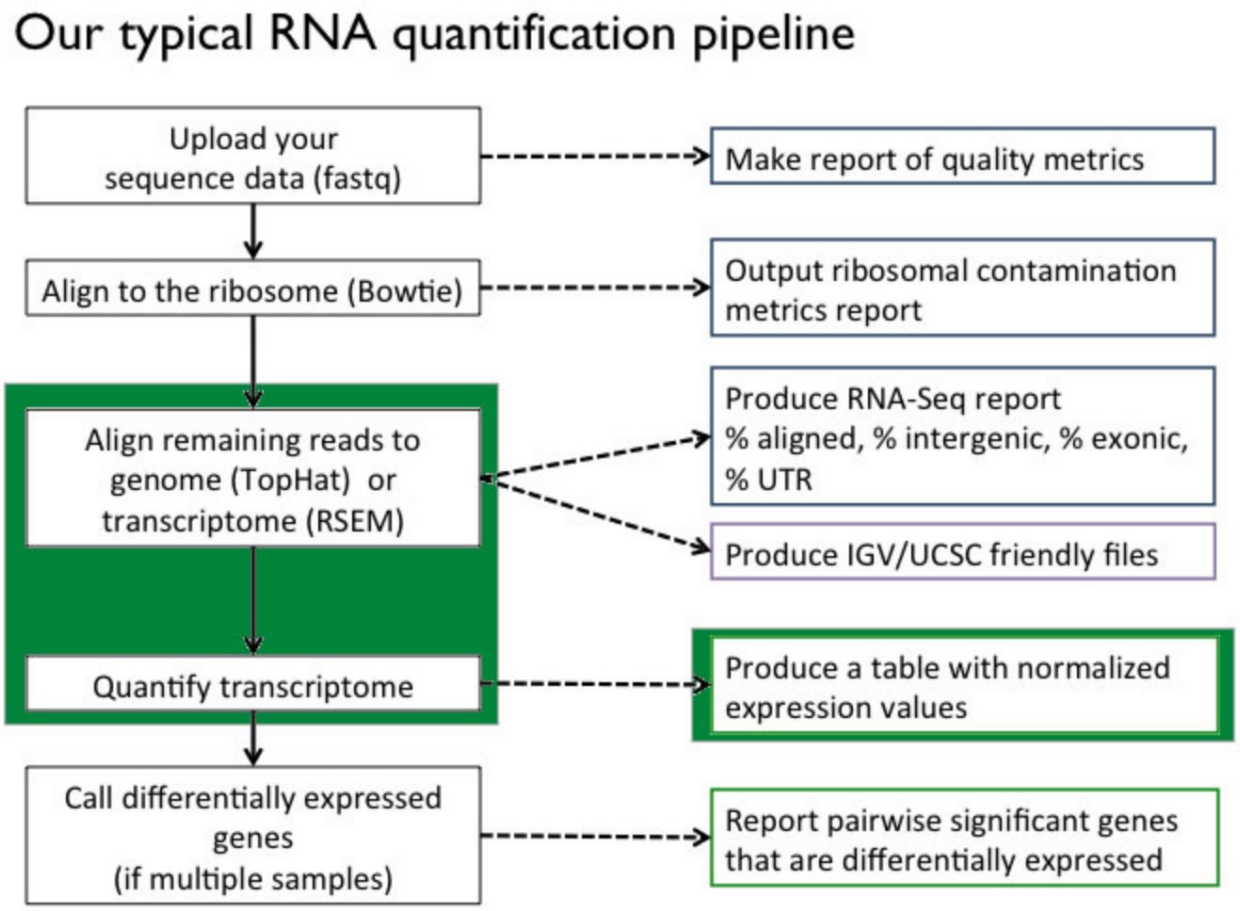
https://bioinfo.umassmed.edu/index.php?p=4



https://bioinfo.umassmed.edu/index.php?p=33

expected learning outcome

To understand the basics of RNA-Seq data, how to use RNA-Seq for different objectives and to familiarize yourself with some standard software packages for such analysis.



https://bioinfo.umassmed.edu/index.php?p=33

Exercise 1: prepare for genomic alignment

Both TopHat and RSEM rely on bowtie to perform read alignment (similar to the BWA aligner you used in genome assembly tutorial). Bowtie like BWA uses very efficient genome compression algorithm (Burrows-Wheeler transform) that allows for quick matching of sequences with less than 3 missmatches. To use these alignments it is necessary to create the BW transform of our genome before mapping reads. The bowtie-build2 program in the Bowtie distribution creates the BW index. Change your directory to genome.quantification. Invoke the BW transform on the mm10.fa file found in the directory genome.quantification:

1. First load necessary modules

```
module load RSEM/1.2.11
```

```
module load tophat/2.0.9
```

```
module load bowtie2/2.3.2
```

```
module load samtools/0.0.19
```

```
module load java/1.7.0_25
```

2. Build bowtie2 index files

```
cd ~/RNASeqWS/transcriptomics/genome.quantification
```

```
bowtie2-build -f mm10.fa mm10
```

We named it mm10 (following the UCSC genome browser naming convention. Although we named it similarly to the full genome, this sequence file only contains a very small region of the mouse genome. Our alignment database will be called mm10, and will include partial sequences for chromosomes 7 and 16.

https://bioinfo.umassmed.edu/index.php?p=33

3. In addition we also will prepare the transcriptome for RSEM alignment. RSEM will align directly to the set of transcripts included (*ucsc.gtf* file). The transcript file was downloaded directly from the UCSC table browser. The file does not contain the information necessary to map isoforms to genes, we therefore compiled a table, *ucsc_into_genesymbol.rsem* that contains this information. To generate the necessary index files use the command below. Please not that the \ is used to span multiple lines:

```
rsem-prepare-reference \
--gtf ucsc.gtf --transcript-to-gene-map ucsc_into_genesymbol.rsem \
mm10.fa mm10.rsem
```

Expected result: Your *genome.quantification* directory now should contain the following files (Tip: use ls -1):

Bowtie2 indeces:

```
mm10.1.bt2
mm10.2.bt2
mm10.3.bt2
mm10.4.bt2
mm10.rev.1.bt2
mm10.rev.2.bt2
```

Bowtie indeces:

```
mm10.rsem.ti
mm10.rsem.grp
mm10.rsem.chrlist
mm10.rsem.transcripts.fa
mm10.rsem.seq
mm10.rsem.idx.fa
mm10.rsem.1.ebwt
mm10.rsem.2.ebwt
mm10.rsem.3.ebwt
mm10.rsem.4.ebwt
mm10.rsem.rev.1.ebwt
mm10.rsem.rev.2.ebwt
```

rsem-prepare-reference --gtf /path/to/trasncriptome/data/annotation.gtf --bowtie2 /path/to/trasncriptome/data/genome.fasta /nl/umw_anthony_imbalzano/Sabriya/RNASeq/output

```
[ss45w@ghpcc06 ~]$ module load RSEM/1.2.11
R 3.0.1 is located under /share/pkg/R/3.0.1
If you are going to install R modules locally, you likely need to load a gcc module first.  If you don't it may give er
rors regarding lquadmath.
perl 5.18.1 is located under /share/pkg/perl/5.18.1
RSEM 1.2.11 is located under /share/pkg/RSEM/1.2.11
[ss45w@ghpcc06 ~]$ rsem-prepare-reference
Invalid number of arguments!
NAME
    rsem-prepare-reference

SYNOPSIS
    rsem-prepare-reference [options] reference_fasta_file(s) reference_name

ARGUMENTS
    reference_fasta_file(s)
        Either a comma-separated list of FASTA formatted files OR a
        directory name. If a directory name is specified, RSEM will read all
        files with suffix ".fa" or ".fasta" in this directory. The files
        should contain either the sequences of transcripts or an entire
        genome, depending on whether the --gtf option is used.

    reference name
        The name of the reference used. RSEM will generate several
        reference-related files that are prefixed by this name. This name
        can contain path information (e.g. /ref/mm9).

OPTIONS
    --gtf <file>
        If this option is on, RSEM assumes that 'reference_fasta_file(s)'
        contains the sequence of a genome, and will extract transcript
        reference sequences using the gene annotations specified in <file>,
        which should be in GTF format.

        If this option is off, RSEM will assume 'reference_fasta_file(s)'
        contains the reference transcripts. In this case, RSEM assumes that
        name of each sequence in the FASTA files is its transcript_id.

        (Default: off)

    --transcript-to-gene-map <file>
        Use information from <file> to map from transcript (isoform) ids to
        gene ids. Each line of <file> should be of the form:
```

```
    gene_id transcript_id

    with the two fields separated by a tab character.

    If you are using a GTF file for the "UCSC Genes" gene set from the
    UCSC Genome Browser, then the "knownIsoforms.txt" file (obtained
    from the "Downloads" section of the UCSC Genome Browser site) is of
    this format.

    If this option is off, then the mapping of isoforms to genes depends
    on whether the --gtf option is specified. If --gtf is specified,
    then RSEM uses the "gene_id" and "transcript_id" attributes in the
    GTF file. Otherwise, RSEM assumes that each sequence in the
    reference sequence files is a separate gene.

    (Default: off)

--no-polyA
    Do not add poly(A) tails to the end of reference isoforms. (Default:
    adding poly(A) tails to all transcripts)

--no-polyA-subset <file>
    Add poly(A) tails to all transcripts except those listed in <file>.
    <file> is a file containing a list of transcript_ids. (Default: add
    poly(A) tails to all transcripts. This option cannot be used with
    '--no-polyA'.)

--polyA-length <int>
    The length of the poly(A) tails to be added. (Default: 125)

--bowtie-path <path>
    The path to the Bowtie executables. (Default: the path to Bowtie
    executables is assumed to be in the user's PATH environment
    variable)

--no-bowtie
    Do not build Bowtie indices. Specify this option if you wish to use
    an alternative aligner for mapping reads to transcripts. You should
    align against the sequences generated in the output file
    'reference_name.idx.fa'. (Default: off)

--no-ntog
    Disable the conversion of 'N' characters to 'G' characters in the
    reference sequences prepared for aligners. This conversion is in
    particular desired for Bowtie aligner to align against all positions
    in the reference. (Default: off)
```

```
    --bowtie2
        Build Bowtie 2 indices instead of Bowtie indices. Turn on this
        option will automatically turn on '--no-bowtie' and '--no-ntog'
        options. (Default: off)

    --bowtie2-path
        The path to the Bowtie 2 executables. (Default: the path to Bowtie 2
        executables is assumed to be in the user's PATH environment
        variable)

    -q/--quiet
        Suppress the output of logging information. (Default: off)

    -h/--help
        Show help information.

DESCRIPTION
    This program extracts/preprocesses the reference sequences and builds
    Bowtie indices using default parameters. If an alternative aligner is to
    be used, then the '--no-bowtie' option should be specified to disable
    building Bowtie indices. This program is used in conjunction with the
    'rsem-calculate-expression' program.

OUTPUT
    This program will generate 'reference_name.grp', 'reference_name.ti',
    'reference_name.transcripts.fa', 'reference_name.seq',
    'reference_name.chrlist' (if '--gtf' is on), 'reference_name.idx.fa',
    and corresponding Bowtie index files (unless '--no-bowtie' is
    specified).

    'reference_name.grp', 'reference_name.ti', 'reference_name.seq', and
    'reference_name.chrlist' are used by RSEM internally.

    'reference_name.transcripts.fa' contains the extracted reference
    transcripts in FASTA format. Poly(A) tails are added unless '--no-polyA'
    is set.

    'reference_name.idx.fa' is used by aligners to build their own indices.
    If '--no-ntog' is set, this file should be identical to
    'reference_name.transcripts.fa'.
```

EXAMPLES

1) Suppose we have mouse RNA-Seq data and want to use the UCSC mm9 version of the mouse genome. We have downloaded the UCSC Genes transcript annotations in GTF format (as mm9.gtf) using the Table Browser and the knownIsoforms.txt file for mm9 from the UCSC Downloads. We also have all chromosome files for mm9 in the directory '/data/mm9'. We want to put the generated reference files under '/ref' with name 'mouse_125'. We'll add poly(A) tails with length 125. Please note that GTF files generated from UCSC's Table Browser do not contain isoform-gene relationship information. For the UCSC Genes annotation, this information can be obtained from the knownIsoforms.txt file. Suppose we want to build Bowtie indices and Bowtie executables are found in '/sw/bowtie'.

There are two ways to write the command:

```
 rsem-prepare-reference --gtf mm9.gtf \
                        --transcript-to-gene-map knownIsoforms.txt \
                        --bowtie-path /sw/bowtie \
                        /data/mm9/chr1.fa,/data/mm9/chr2.fa,...,/data/mm9/chrM.fa \
                        /ref/mouse_125
```

OR

```
 rsem-prepare-reference --gtf mm9.gtf \
                        --transcript-to-gene-map knownIsoforms.txt \
                        --bowtie-path /sw/bowtie \
                        /data/mm9 \
                        /ref/mouse_125
```

2) Suppose we only have transcripts from EST tags in 'mm9.fasta'. In addition, we also have isoform-gene information in 'mapping.txt'. We do not want to add any poly(A) tails. The reference_name will be set to 'mouse_0'. In addition, we do not want to build Bowtie indices, and will use an alternative aligner to align reads against the 'mouse_0.idx.fa' output file:

```
 rsem-prepare-reference --transcript-to-gene-map mapping.txt \
                        --no-polyA \
                        --no-bowtie \
                        mm9.fasta \
                        mouse_0
```

exercise 2 Quantify with the RSEM program

RSEM depends on an existing annotation and will only scores transcripts that are present in the given annotation file. We will compare the alignments produced by RSEM and tophat and this will become clear.

The first step is to prepare the transcript set that we will quantify. We selected the UCSC genes which is a very comprehensive, albeit a bit noisy dataset. As with all the data in this activity we will only use the subset of the genes that map to the genome regions we are using.

2.1 Calculate expression

RSEM now is ready to align and then attempt to perform read assignment and counting for each isoform in the file provided above. You must process each one of the 6 libraries:

```
cd ~/RNASeqWS/transcriptomics
```

```
mkdir rsem

rsem-calculate-expression --paired-end -p 2 \
 --output-genome-bam fastq.quantification/control_rep1.1.fq \
fastq.quantification/control_rep1.2.fq genome.quantification/mm10.rsem rsem/ctrl1.rsem
```

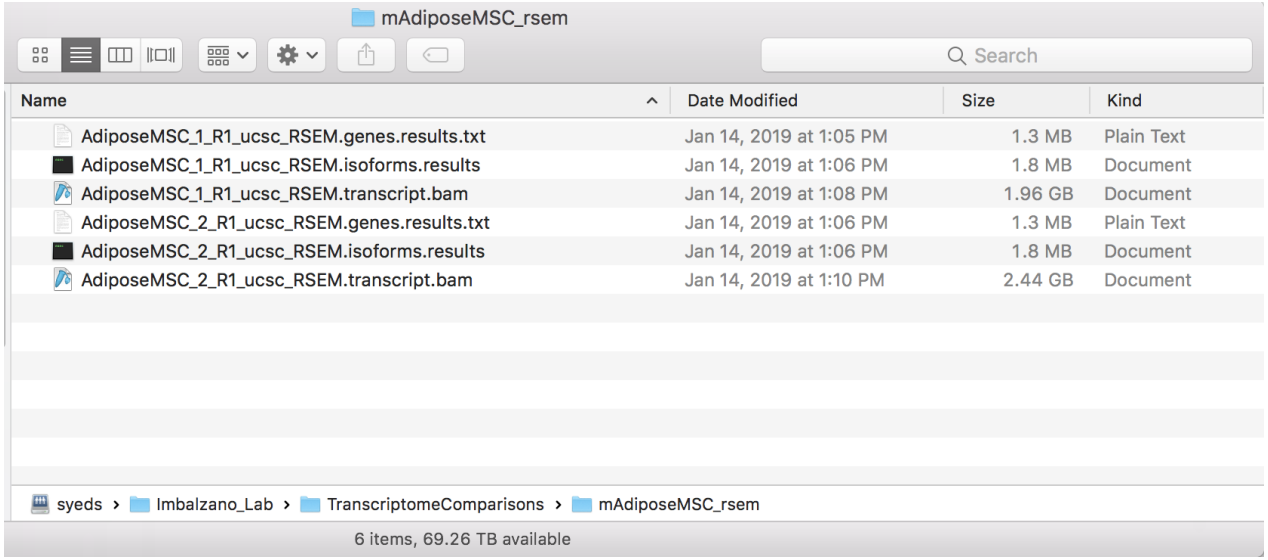
Single-end RNA-Seq analysis

bsub -n 1 -o logS -q long -R rusage[mem=80000] -W999 "rsem-calculate-expression --bowtie2 --num-threads 12 RNASeq/msAdiposeMSC_GSE116558_1_R1.fastq.gz RNASeq/mm10 RNASeq/AdiposeMSC_1_R1_ucsc_RSEM”

```
[ss45w@ghpcc06 ~]$ rsem-calculate-expression
Invalid number of arguments!
NAME
    rsem-calculate-expression

SYNOPSIS
     rsem-calculate-expression [options] upstream_read_file(s) reference_name sample_name
     rsem-calculate-expression [options] --paired-end upstream_read_file(s) downstream_read_file(s) reference_name sample_name
     rsem-calculate-expression [options] --sam/--bam [--paired-end] input reference_name sample_name

ARGUMENTS
    upstream_read_files(s)
        Comma-separated list of files containing single-end reads or
        upstream reads for paired-end data. By default, these files are
        assumed to be in FASTQ format. If the --no-qualities option is
        specified, then FASTA format is expected.

    downstream_read_file(s)
        Comma-separated list of files containing downstream reads which are
        paired with the upstream reads. By default, these files are assumed
        to be in FASTQ format. If the --no-qualities option is specified,
        then FASTA format is expected.

    input
        SAM/BAM formatted input file. If "-" is specified for the filename,
        SAM/BAM input is instead assumed to come from standard input. RSEM
        requires all alignments of the same read group together. For
        paired-end reads, RSEM also requires the two mates of any alignment
        be adjacent. See Description section for how to make input file obey
        RSEM's requirements.

    reference_name
        The name of the reference used. The user must have run
        'rsem-prepare-reference' with this reference_name before running
        this program.

    sample_name
        The name of the sample analyzed. All output files are prefixed by
        this name (e.g., sample_name.genes.results)

OPTIONS
    --paired-end
        Input reads are paired-end reads. (Default: off)

    --no-qualities
        Input reads do not contain quality scores. (Default: off)

    --strand-specific
        The RNA-Seq protocol used to generate the reads is strand specific,
```

```
--sam
    Input file is in SAM format. (Default: off)

--bam
    Input file is in BAM format. (Default: off)

--sam-header-info <file>
    RSEM reads header information from input by default. If this option
    is on, header information is read from the specified file. For the
    format of the file, please see SAM official website. (Default: "")

-p/--num-threads <int>
    Number of threads to use. Both Bowtie/Bowtie2 and expression
    estimation will use this many threads. (Default: 1)

--no-bam-output
    Do not output any BAM file. (Default: off)

--output-genome-bam
    Generate a BAM file, 'sample_name.genome.bam', with alignments
    mapped to genomic coordinates and annotated with their posterior
    probabilities. In addition, RSEM will call samtools (included in
    RSEM package) to sort and index the bam file.
    'sample_name.genome.sorted.bam' and
    'sample_name.genome.sorted.bam.bai' will be generated. (Default:
    off)

--sampling-for-bam
    When RSEM generates a BAM file, instead of outputing all alignments
    a read has with their posterior probabilities, one alignment is
    sampled according to the posterior probabilities. The sampling
    procedure includes the alignment to the "noise" transcript, which
    does not appear in the BAM file. Only the sampled alignment has a
    weight of 1. All other alignments have weight 0. If the "noise"
    transcript is sampled, all alignments appeared in the BAM file
    should have weight 0. (Default: off)

--calc-ci
    Calculate 95% credibility intervals and posterior mean estimates.
    (Default: off)

--seed-length <int>
    Seed length used by the read aligner. Providing the correct value is
    important for RSEM. If RSEM runs Bowtie, it uses this value for
    Bowtie's seed length parameter. Any read with its or at least one of
    its mates' (for paired-end reads) length less than this value will
    be ignored. If the references are not added poly(A) tails, the
    minimum allowed value is 5, otherwise, the minimum allowed value is
    25. Note that this script will only check if the value >= 5 and give
```

RSEM output



Gene results

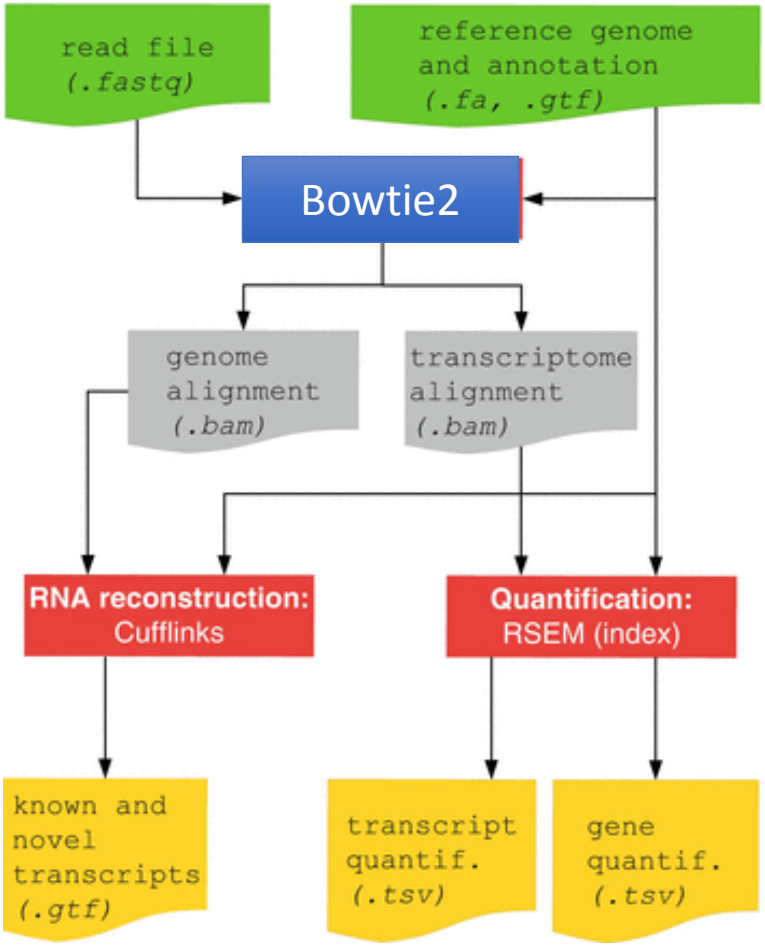
A	B	C	D	E	F	G
gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
0610005C13Rik	NR_038165,NR_038166	1018.5	983.5	0	0	0
0610007P14Rik	NM_021446	1185	1150	1296	58.24	31.74
0610009B22Rik	NM_025319	795	760	826.71	56.21	30.63
0610009L18Rik	NR_038126	619	584	87	7.7	4.2
0610009O20Rik	NM_024179	2404	2369	391.08	8.53	4.65
0610010B08Rik	NM_001177543	4539	4504	195.99	2.25	1.23
0610010F05Rik	NM_027860	4140	4105	1023	12.88	7.02
0610010K14Rik	NM_001177601,NM_001177603,NM_00117	806.29	771.29	710	47.57	25.92
0610011F06Rik	NM_026686	811	776	318	21.18	11.54
0610012G03Rik	NR_027897	1471	1436	1547.11	55.68	30.34
0610030E20Rik	NM_026696	4634	4599	172.12	1.93	1.05
0610031O16Rik	NR_045760	824	789	0	0	0
0610037L13Rik	NM_028754	1551	1516	981	33.44	18.22
0610038B21Rik	NR_028125	1523	1488	0	0	0
0610039K10Rik	NR_028113	951	916	0	0	0
0610040B10Rik	NR_027874	440	405	14	1.79	0.97
0610040F04Rik	NR_040757,NR_104577	784	749	4	0.28	0.15

Transcript variant results

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct
NR_038165	0610005C13Rik	915	880	0	0	0	0
NR_038166	0610005C13Rik	1122	1087	0	0	0	0
NM_021446	0610007P14Rik	1185	1150	1296	58.24	31.74	100
NM_025319	0610009B22Rik	795	760	826.71	56.21	30.63	100
NR_038126	0610009L18Rik	619	584	87	7.7	4.2	100
NM_024179	0610009O20Rik	2404	2369	391.08	8.53	4.65	100
NM_001177543	0610010B08Rik	4539	4504	195.99	2.25	1.23	100
NM_027860	0610010F05Rik	4140	4105	1023	12.88	7.02	100
NM_001177601	0610010K14Rik	886	851	83.3	5.06	2.76	10.63
NM_001177603	0610010K14Rik	781	746	310.33	21.5	11.71	45.19
NM_001177606	0610010K14Rik	679	644	0	0	0	0
NM_001177607	0610010K14Rik	706	671	107.02	8.24	4.49	17.33
NM_026757	0610010K14Rik	780	745	0	0	0	0
NM_145758	0610010K14Rik	882	847	209.36	12.77	6.96	26.85
NM_026686	0610011F06Rik	811	776	318	21.18	11.54	100
NR_027897	0610012G03Rik	1471	1436	1547.11	55.68	30.34	100
NM_026696	0610030E20Rik	4634	4599	172.12	1.93	1.05	100
NR_045760	0610031O16Rik	824	789	0	0	0	0
NM_028754	0610037L13Rik	1551	1516	981	33.44	18.22	100

RNA-Seq Pipline

Methods and Tools



Bowtie and Bowtie2 use Burrows-Wheeler indexing for aligning reads. With bowtie2 there is no upper limit on the read length

Bowtie2 to align reads in a splice-aware manner and aids the discovery of new splice junctions

RSEM is a software package for estimating gene and isoform expression levels from RNA-Seq data. The RSEM package provides an user-friendly interface, supports threads for parallel computation of the EM algorithm, single-end and paired-end read data, quality scores, variable-length reads and RSPD estimation.

Differential Gene Expression analysis remains (week 6)

….then figures like this are possible

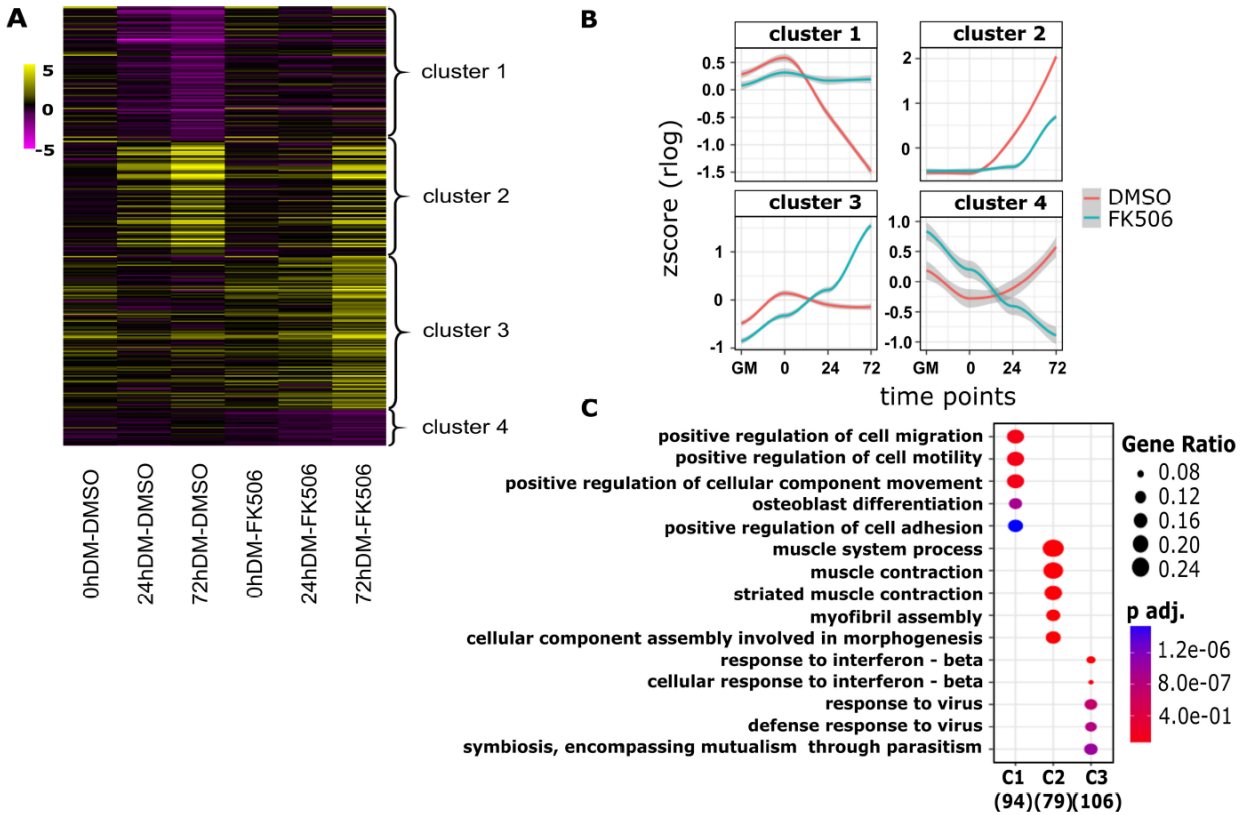


FIG 3 Cluster analysis of differentially expressed genes at three differentiation time points in myoblasts treated with Cn inhibitor. (A) Heat map comparing differential expression levels of 308 FK506 treatment-specific genes, categorized in four different clusters. Each column represents an experimental sample (times 0, 24, and 72 h in differentiation medium [DM]) compared to the proliferating myoblast sample cultured in growth medium (GM). Each row represents a specific gene. The colors range from yellow (high expression) to magenta (low expression) and represent the relative $\log_2$ ratios of expression levels. (B) Kinetic expression patterns of the four clusters of genes. Lines represent the LOESS (locally estimated scatterplot smoothing) fitting for the relative expression levels of all genes in the cluster. The gray shading represents 95% confidence intervals. (C) Gene ontology analysis of differentially expressed genes within cluster 1 (C1), C2, and C3 identified the top enriched GO terms with the corresponding enrichment *P* values and gene ratios.

Week 3

- Unix Problem Set Q&A
- Review ChIP-Seq and RNA-Seq pipelines

- Loading files to the cluster using FileZilla and rsync

- On the ghpcc06 cluster, read criteria for running bowtie2, picard, samtools, macs2, etc.
- See sample bsub commands for running each module
- Look at bowtie2 and macs2 output files
- Use IGV to view .bam and .bed files

- On the ghpcc06 cluster, read criteria for running rsem-calculate-expression on the cluster
  https://bioinfo.umassmed.edu/index.php?p=33
- RNA-Seq – Look at RSEM output files
- See sample bsub commands for rsem-calculate-expression

Tasks:
Create lab schedule for using the ghpcc06 cluster
Explore the cluster and examine modules and tools for data analysis that are available
Upload Fastq files to ghpcc06 cluster

Week 4

- Introducing NGSplot: Quick mining and visualization of NGS data by integrating genomic databases
  https://github.com/shenlab-sinai/ngsplot
- Introducing HOMER: Software for motif discovery and next-gen sequencing analysis
  http://homer.ucsd.edu/homer/ngs/

- Running NGSplot on the ghpcc06 cluster for making ChIP-Seq and RNA-Seq heatmaps and extracting tag density files
- Running HOMER on the ghpcc06 cluster for annotating bed files and performing motif analysis
- Running bedtools on the ghpcc06 cluster for comparing .bed files of different experiments