HW2: RNA-seq Ver2

BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

Getting started with HISAT, StringTie, and Ballgown

https://www.nature.com/articles/nprot.2016.095#an1

Getting started with HISAT, StringTie, and Ballgown

$ wget ftp://ftp.ccb.jhu.edu/pub/RNAseq_protocol/chrX_data.tar.gz

You can copy the data from /public/ahnt/courses/bcb5250/rna_seq_lab/chrX_data.tar.gz

Required tools

EQUIPMENT

- Data (example RNA-seq reads, indexes and gene annotations for use in this protocol are available at ftp://ftp.ccb.jhu.edu/pub/RNAseq_protocol; see Equipment Setup for details)
- HISAT2 software (http://ccb.jhu.edu/software/hisat2 or http://github.com/infphilo/hisat2, version 2.0.1 or later)
- StringTie software (http://ccb.jhu.edu/software/stringtie or https://github.com/gpertea/stringtie , version 1.2.2 or later)
- SAMtools (http://samtools.sourceforge.net, version 0.1.19 or later)
- R (https://www.r-project.org, version 3.2.2 or later)
- Hardware (64-bit computer running either Linux or Mac OS X (10.7 Lion or later); 4 GB of RAM (8 GB preferred); see Equipment Setup)

/public/ahnt/courses/bcb5250/rna_seq_lab/software/

Add each program path into your bashrc

Ballgown is a Bioconductor package, so we need to install that using R. While we are at it, we will install various dependencies too.

```
install.packages("devtools")
install.packages("dplyr")

source("https://www.bioconductor.org/biocLite.R")
biocLite(c("alyssafrazee/RSkittleBrewer", "ballgown", "genefilter"))
```

Mapping

- Mapping is performed using HISAT2 and usually the first step, prior to mapping, is to create an index of the reference genome. The indices are provided in the data folder but let's create them again.
  - $ cd chrX_data
  - $ mkdir my_index
  - $ cd my_index
  - $ /public/ahnt/courses/bcb5250/rna_seq_lab/software/hisat2-2.0.0-beta/extract_splice_sites.py ../genes/chrX.gtf > chrX.ss
  - $ /public/ahnt/courses/bcb5250/rna_seq_lab/software/hisat2-2.0.0-beta/extract_exons.py ../genes/chrX.gtf > chrX.exon
  - $ head -3 chrX.ss
  - $ head -3 chrX.exon
  - $ hisat2-build -p 8 --ss chrX.ss --exon chrX.exon ../genome/chrX.fa chrX_tran

Mapping

- $ mkdir map
- $ cat map.sh

- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188044_chrX_1.fastq.gz -2 chrX_data/samples/ERR188044_chrX_2.fastq.gz -S map/ERR188044_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188104_chrX_1.fastq.gz -2 chrX_data/samples/ERR188104_chrX_2.fastq.gz -S map/ERR188104_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188234_chrX_1.fastq.gz -2 chrX_data/samples/ERR188234_chrX_2.fastq.gz -S map/ERR188234_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188245_chrX_1.fastq.gz -2 chrX_data/samples/ERR188245_chrX_2.fastq.gz -S map/ERR188245_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188257_chrX_1.fastq.gz -2 chrX_data/samples/ERR188257_chrX_2.fastq.gz -S map/ERR188257_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188273_chrX_1.fastq.gz -2 chrX_data/samples/ERR188273_chrX_2.fastq.gz -S map/ERR188273_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188337_chrX_1.fastq.gz -2 chrX_data/samples/ERR188337_chrX_2.fastq.gz -S map/ERR188337_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188383_chrX_1.fastq.gz -2 chrX_data/samples/ERR188383_chrX_2.fastq.gz -S map/ERR188383_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188401_chrX_1.fastq.gz -2 chrX_data/samples/ERR188401_chrX_2.fastq.gz -S map/ERR188401_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188428_chrX_1.fastq.gz -2 chrX_data/samples/ERR188428_chrX_2.fastq.gz -S map/ERR188428_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR188454_chrX_1.fastq.gz -2 chrX_data/samples/ERR188454_chrX_2.fastq.gz -S map/ERR188454_chrX.sam
- hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1 chrX_data/samples/ERR204916_chrX_1.fastq.gz -2 chrX_data/samples/ERR204916_chrX_2.fastq.gz -S map/ERR204916_chrX.sam

Mapping

- Sort
  - $ cat sort.sh
  - samtools view -@ 8 -S -b map/ERR188044_chrX.sam | samtools sort -@ 8 - map/ERR188044_chrX
  - samtools view -@ 8 -S -b map/ERR188104_chrX.sam | samtools sort -@ 8 - map/ERR188104_chrX
  - samtools view -@ 8 -S -b map/ERR188234_chrX.sam | samtools sort -@ 8 - map/ERR188234_chrX
  - samtools view -@ 8 -S -b map/ERR188245_chrX.sam | samtools sort -@ 8 - map/ERR188245_chrX
  - samtools view -@ 8 -S -b map/ERR188257_chrX.sam | samtools sort -@ 8 - map/ERR188257_chrX
  - samtools view -@ 8 -S -b map/ERR188273_chrX.sam | samtools sort -@ 8 - map/ERR188273_chrX
  - samtools view -@ 8 -S -b map/ERR188337_chrX.sam | samtools sort -@ 8 - map/ERR188337_chrX
  - samtools view -@ 8 -S -b map/ERR188383_chrX.sam | samtools sort -@ 8 - map/ERR188383_chrX
  - samtools view -@ 8 -S -b map/ERR188401_chrX.sam | samtools sort -@ 8 - map/ERR188401_chrX
  - samtools view -@ 8 -S -b map/ERR188428_chrX.sam | samtools sort -@ 8 - map/ERR188428_chrX
  - samtools view -@ 8 -S -b map/ERR188454_chrX.sam | samtools sort -@ 8 - map/ERR188454_chrX
  - samtools view -@ 8 -S -b map/ERR204916_chrX.sam | samtools sort -@ 8 - map/ERR204916_chrX

Assembly

- Now we need to assemble the mapped reads into transcripts. StringTie can assemble transcripts with or without annotation; as noted in the protocol, annotation can be helpful when the number of reads for a transcript is too low for an accurate assembly.
  - $ mkdir assembly
  - $ cat stringtie.sh
  - stringtie map/ERR188044_chrX.bam -l ERR188044 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188044_chrX.gtf
  - stringtie map/ERR188104_chrX.bam -l ERR188104 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188104_chrX.gtf
  - stringtie map/ERR188234_chrX.bam -l ERR188234 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188234_chrX.gtf
  - stringtie map/ERR188245_chrX.bam -l ERR188245 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188245_chrX.gtf
  - stringtie map/ERR188257_chrX.bam -l ERR188257 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188257_chrX.gtf
  - stringtie map/ERR188273_chrX.bam -l ERR188273 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188273_chrX.gtf
  - stringtie map/ERR188337_chrX.bam -l ERR188337 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188337_chrX.gtf
  - stringtie map/ERR188383_chrX.bam -l ERR188383 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188383_chrX.gtf
  - stringtie map/ERR188401_chrX.bam -l ERR188401 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188401_chrX.gtf
  - stringtie map/ERR188428_chrX.bam -l ERR188428 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188428_chrX.gtf
  - stringtie map/ERR188454_chrX.bam -l ERR188454 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR188454_chrX.gtf
  - stringtie map/ERR204916_chrX.bam -l ERR204916 -p 8 -G chrX_data/genes/chrX.gtf -o assembly/ERR204916_chrX.gtf

Assembly

- Before merging we need to modify mergelist.txt. The modified mergelist.txt should look like this:
  - $ cat chrX_data/mergelist.txt
  - assembly/ERR188044_chrX.gtf
  - assembly/ERR188104_chrX.gtf
  - assembly/ERR188234_chrX.gtf
  - assembly/ERR188245_chrX.gtf
  - assembly/ERR188257_chrX.gtf
  - assembly/ERR188273_chrX.gtf
  - assembly/ERR188337_chrX.gtf
  - assembly/ERR188383_chrX.gtf
  - assembly/ERR188401_chrX.gtf
  - assembly/ERR188428_chrX.gtf
  - assembly/ERR188454_chrX.gtf
  - assembly/ERR204916_chrX.gtf

Assembly

- Merge
  - $ stringtie --merge -p 8 -G chrX_data/genes/chrX.gtf -o stringtie_merged.gtf chrX_data/mergelist.txt
  - $ cat stringtie_merged.gtf | head

- How many transcripts?
  - cat stringtie_merged.gtf | grep -v "^#" | awk '$3=="transcript" {print}' | wc –l

- compare the assembled transcripts to known transcripts
  - $ gffcompare -r chrX_data/genes/chrX.gtf -G -o merged stringtie_merged.gtf
  - 2343 reference transcripts loaded.
  - 241 duplicate reference transcripts discarded.
  - 3547 query transfrags loaded.

Estimate their abundances

- Now that we have our assembled transcripts, we can estimate their abundances.
  - $ cat estimate.sh
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188044/ERR188044_chrX.gtf map/ERR188044_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188104/ERR188104_chrX.gtf map/ERR188104_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188234/ERR188234_chrX.gtf map/ERR188234_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188245/ERR188245_chrX.gtf map/ERR188245_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188257/ERR188257_chrX.gtf map/ERR188257_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188273/ERR188273_chrX.gtf map/ERR188273_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188337/ERR188337_chrX.gtf map/ERR188337_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188383/ERR188383_chrX.gtf map/ERR188383_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188401/ERR188401_chrX.gtf map/ERR188401_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188428/ERR188428_chrX.gtf map/ERR188428_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR188454/ERR188454_chrX.gtf map/ERR188454_chrX.bam
  - stringtie -e -B -p 8 -G stringtie_merged.gtf -o ballgown/ERR204916/ERR204916_chrX.gtf map/ERR204916_chrX.bam

Differential expression

- Ballgown is a Bioconductor package, so we need to install that using R. While we are at it, we will install various dependencies too.
  - install.packages("devtools")
  - install.packages("dplyr")
  - install.packages("ggplot2")
  - install.packages("cowplot")

  - source("https://www.bioconductor.org/biocLite.R")
  - biocLite(c("alyssafrazee/RSkittleBrewer", "ballgown", "genefilter"))

Differential expression

- Let me provide the R script separately.