THE

RNAseq Analysis via Hisat, StringTie and BAllgown

MANUAL

Timo Lassmann

TODAY

Contents

CHAPTER 1

Introduction

Included are two RNAseq workflows implemented in bpipe[sadedin2012bpipe]. Each produces a Ballgown[frazee2014flexible] R object for differenetial gene expression analysis.

CHAPTER 2

Quick Usage

To run these pipelines simply point them to the input fastq files:

```
bpipe run  -n 4 -r ../pipelines/StringTieDeNovo.groovy  *R*.fq.gz
```

Note: you may have to modify the regular expressions at the start of the pipelines ot recognize your file names.

2.1 Workflow 1: Finding novel isoforms:

This workflow:

1. aligns all reads to the genome using HiSat[kim2015hisat].
2. assembles mapped reads into transcripts using StringTie[pertea2015stringtie]. Known transcripts from gencode v19 are provided to guide the assembly.
3. merges transcripts from all samples using cuffmerge.
4. annotates transcripts usign cuffcompare
5. quantifies transcripts from steop 3 across all samples using StringTie[pertea2015stringtie].

File: StringTieDeNovoHUMAN.groovy

```
##Settings

HISATINDEX=HISATHG19

GENOME=GENOMEHG19

KNOWNTRANSCRIPTS=GENCODEV19_HS
KNOWNSPLICESITES=GENCODEV19_HS_SPLICESITES
```

```
MERGED_TRANSCRIPTS = "merged_asm/merged.gtf"

FINAL_TRANSCRIPTS =  "merged_asm/finalmodel.combined.gtf"

Bpipe.run {"%_R*.fastq.gz"  * [ hisat_align + stringtie]  +
   makeassemblylist + cuffmerge + cuffcompare + "%_R*.fastq.gz"  *
    [  hisat_align + stringtieB] + make_ballgown_obj}
```

File: StringTieDeNovoMOUSE.groovy

```
##Settings

HISATINDEX=HISATMM10
GENOME=GENOMEMM10

KNOWNTRANSCRIPTS=GENCODEVM5_MM
KNOWNSPLICESITES=GENCODEVM5_MM_SPLICESITES

MERGED_TRANSCRIPTS = "merged_asm/merged.gtf"

FINAL_TRANSCRIPTS =  "merged_asm/finalmodel.combined.gtf"

Bpipe.run {"%_R*.fastq.gz"  * [ hisat_align + stringtie]  +
   makeassemblylist + cuffmerge + cuffcompare + "%_R*.fastq.gz"  *
    [  hisat_align + stringtieB] + make_ballgown_obj}
```

2.2 Workflow 2: Quantification of known transcripts.

This workflow:

1. aligns all reads to the genome using HiSat[kim2015hisat].
2. quantifies gencode v19 transcripts from across all samples using StringTie[pertea2015stringtie].

2.3 Settings

```
load "localprograms.groovy"
load "localdatafiles.groovy"
load "../modules/RNAseq.groovy"
```

File: make_ballgown_obj.sh

This is a small script to turn stringtie tables into an R object.

```
#!/bin/bash
```

```
echo "library(ballgown)" > ballgownR_glue.R
echo -ne "sample_directories <- c(" >> ballgownR_glue.R
counter=1
for line in $(find  . -name "e2t.ctab"); do
    a="$(echo $line   | rev | cut -d"/" -f2,3 | rev)"
    if [ $counter = 1 ]; then
        echo -ne "'$a'" >> ballgownR_glue.R
    else
        echo -ne ",'$a'" >> ballgownR_glue.R
    fi
    counter=$(($counter + 1));
done
echo ")" >> ballgownR_glue.R
echo "bg <- ballgown(samples = sample_directories)" >>
   ballgownR_glue.R

echo "save(bg, file = 'bg.rda')" >> ballgownR_glue.R

Rscript  ballgownR_glue.R
```