Homework 1

Name: Aris Podotas

University: National and Kapodistrian University of Athens

Program: Data Science and Informaion Technologies

Specialization: Bioinformatics - Biomedical Data

Lesson: Introduction To Bioinformatics

User: user03

Date: November 2024

Contents

Note: Files will be within the user03 folder in the virtual machine. **Note:** Commands will not be explicitly written here.

1 Analysis

Note: a intron has two splice sites, the 5' and 3' splice site.

1.1 find a read with unmapped pair

genome:173,708-173,775

1.2 find a read that has 2 mismatches

genome:136,210-136,277

1.3 find a read-pair with pair orientation: F2R1

genome:313,112-313,179

1.4 find a gene that agrees in 6 splice-sites with its annotation

genome:329662-331216

1.5 find a location where less than 96at least 2 other nucleotides occur.

genome:270,404

1.6 find a gene where the peak with the highest number of reads has at least 3 times the reads of the highest peak in a different sample (with non-zero reads for that gene)

genome:2670-3431

2

After running the command (the command itself is contained in the file) and looking in the "cuffcmp.stats" file we have:

```
user03@UoA-Intro2Bio24: ~/
# Cuffcompare v2.2.1 | Command line was:
#cuffcompare ./merged_asm/merged.gtf -r genes.gff3
#

#= Summary for dataset: ./merged_asm/merged.gtf :
#     Query mRNAs :     209 in     207 loci  (59 multi-exon transcripts)
#            (2 multi-transcript loci, ~1.0 transcripts per locus)
# Reference mRNAs :     200 in     200 loci  (77 multi-exon)
# Super-loci w/ reference transcripts:      180
#-------------------|    Sn    |  Sp    |  fSn |  fSp
        Base level:      76.3    85.5     -       -
        Exon level:      26.7    32.7    29.2    35.7
      Intron level:      52.1    96.6    52.1    96.6
Intron chain level:      49.4    64.4    50.6    66.1
  Transcript level:       0.0     0.0     0.0     0.0
       Locus level:      19.0    18.4    19.5    18.8

     Matching intron chains:      38
              Matching loci:      38

          Missed exons:      91/363     ( 25.1%)
           Novel exons:      15/297     (  5.1%)
        Missed introns:      77/163     ( 47.2%)
         Novel introns:       2/88      (  2.3%)
           Missed loci:      20/200     ( 10.0%)
            Novel loci:      13/207     (  6.3%)

 Total union super-loci across all input datasets: 193
```

Figure 1: Cuffcompare: cuffcmp.stats file contents

3

Showing the contents of all the files under the align_summary.txt label:

```
Left reads:
          Input     :    101577
           Mapped   :     92230 (90.8% of input)
            of these:        14 ( 0.0%) have multiple alignments (0 have >20)
Right reads:
          Input     :    101577
           Mapped   :     95250 (93.8% of input)
            of these:        13 ( 0.0%) have multiple alignments (0 have >20)
92.3% overall read mapping rate.

Aligned pairs:     88881
     of these:        12 ( 0.0%) have multiple alignments
                      31 ( 0.0%) are discordant alignments
87.5% concordant pair alignment rate.
```

Figure 2: Tophat: df summary statistics

```
Left reads:
          Input     :    107676
           Mapped   :     99116 (92.1% of input)
            of these:        19 ( 0.0%) have multiple alignments (0 have >20)
Right reads:
          Input     :    107676
           Mapped   :    100451 (93.3% of input)
            of these:        18 ( 0.0%) have multiple alignments (0 have >20)
92.7% overall read mapping rate.

Aligned pairs:     95068
     of these:        18 ( 0.0%) have multiple alignments
                      32 ( 0.0%) are discordant alignments
88.3% concordant pair alignment rate.
```

Figure 3: Tophat: hs summary statistics

```
Left reads:
          Input     :     37915
           Mapped   :     34949 (92.2% of input)
            of these:        14 ( 0.0%) have multiple alignments (0 have >20)
Right reads:
          Input     :     37915
           Mapped   :     35392 (93.3% of input)
            of these:        14 ( 0.0%) have multiple alignments (0 have >20)
92.8% overall read mapping rate.

Aligned pairs:     33565
     of these:        14 ( 0.0%) have multiple alignments
                      19 ( 0.1%) are discordant alignments
88.5% concordant pair alignment rate.
```

Figure 4: Tophat: log summary statistics

```
Left reads:
          Input     :     99328
           Mapped   :     91593 (92.2% of input)
            of these:        19 ( 0.0%) have multiple alignments (0 have >20)
Right reads:
          Input     :     99328
           Mapped   :     94339 (95.0% of input)
            of these:        19 ( 0.0%) have multiple alignments (0 have >20)
93.6% overall read mapping rate.

Aligned pairs:     89032
     of these:        17 ( 0.0%) have multiple alignments
                      14 ( 0.0%) are discordant alignments
89.6% concordant pair alignment rate.
```

Figure 5: Tophat: plat summary statistics

3.1 Right reads

The final right reads average is:

$$\frac{93.8 + 93.3 + 93.3 + 95.0}{4} = 93.85\%$$

3.2 Left reads

The final left reads average is:

$$\frac{90.8 + 92.1 + 92.2 + 92.2}{4} = 91.825\%$$

4

Note: The files mentioned will be within user03 in the virtual machine
Note: Input functionality at each step will be
Note: Because all answers will be yes for this exercise as was said in the requirements.
Note: The files that call each other need *chmod* to be run on them before we can use them since they dont have permission to run other files yet.

Multiple attempts were made and documented (kept within the vm). Each one had a layer of progression, for example the first attempt had one file that ran on one thread not even qualifying for the solution. These are the files:

4.1 Attempt 1:

File: script.sh

```
#!/bin/bash

main () {
        bowtie-build genome.fa genome
        tophat -I 1000 -i 20 --bowtie1 --library-type fr-firststrand -o tophat.Sp_ds.dir genome Sp_ds.left.fq Sp_ds.right.fq
        mv tophat.Sp_ds.dir/accepted_hits.bam tophat.Sp_ds.dir/Sp_ds.bam
        samtools index tophat.Sp_ds.dir/Sp_ds.bam
        cufflinks --overlap-radius 1 --library-type fr-firststrand -o cufflinks.Sp_ds.dir tophat.Sp_ds.dir/Sp_ds.bam
  mv cufflinks.Sp_ds.dir/transcripts.gtf cufflinks.Sp_ds.dir/Sp_ds.transcripts.gtf
        tophat -I 1000 -i 20 --bowtie1 --library-type fr-firststrand -o tophat.Sp_hs.dir genome Sp_hs.left.fq Sp_hs.right.fq
        mv tophat.Sp_hs.dir/accepted_hits.bam tophat.Sp_hs.dir/Sp_hs.bam
        samtools index tophat.Sp_hs.dir/Sp_hs.bam
        cufflinks --overlap-radius 1 --library-type fr-firststrand -o cufflinks.Sp_hs.dir tophat.Sp_hs.dir/Sp_hs.bam
        mv cufflinks.Sp_hs.dir/transcripts.gtf cufflinks.Sp_hs.dir/Sp_hs.transcripts.gtf
        tophat -I 1000 -i 20 --bowtie1 --library-type fr-firststrand -o tophat.Sp_log.dir genome Sp_log.left.fq Sp_log.right.fq
```

```
16          mv      tophat.Sp_log.dir/accepted_hits.bam     tophat.
    Sp_log.dir/Sp_log.bam
17          samtools        index   tophat.Sp_log.dir/Sp_log.bam
18          cufflinks       --overlap-radius 1              --library-
    type  fr-firststrand  -o      cufflinks.Sp_log.dir    tophat.Sp_log.
    dir/Sp_log.bam
19          mv      cufflinks.Sp_log.dir/transcripts.gtf    cufflinks.
    Sp_log.dir/Sp_log.transcripts.gtf
20          tophat  -I      1000    -i      20      --bowtie1       --
    library-type  fr-firststrand  -o      tophat.Sp_plat.dir
21          mv      tophat.Sp_plat.dir/accepted_hits.bam    tophat.
    Sp_plat.dir/Sp_plat.bam
22          samtools        index   tophat.Sp_plat.dir/Sp_plat.bam
23          cufflinks --overlap-radius 1     --library-type fr-
    firststrand  -o      cufflinks.Sp_plat.dir   tophat.Sp_plat.dir/
    Sp_plat.bam
24          mv      cufflinks.Sp_plat.dir/transcripts.gtf   cufflinks.
    Sp_plat.dir/Sp_plat.transcripts.gtf
25          echo    cufflinks.Sp_ds.dir/Sp_ds.transcripts.gtf       >>
        assemblies.txt
26          echo    cufflinks.Sp_hs.dir/Sp_hs.transcripts.gtf       >>
        assemblies.txt
27          echo    cufflinks.Sp_log.dir/Sp_log.transcripts.gtf     >>
        assemblies.txt
28          echo    cufflinks.Sp_plat.dir/Sp_plat.transcripts.gtf   >>
        assemblies.txt
29          cat     assemblies.txt
30          cuffmerge       -s      genome.fa       assemblies.txt
31          java    -Xmx2G  -jar    /Users/bhaas/IGV/current//igv.jar
        -g      `pwd`/genome.fa `pwd`/merged_asm/merged.gtf,`pwd`/genes
    .bed,`pwd`/tophat.Sp_ds.dir/Sp_ds.bam,`pwd`/tophat.Sp_hs.dir/Sp_hs.
    bam,`pwd`/tophat.Sp_log.dir/Sp_log.bam,`pwd`/tophat.Sp_plat.dir/
    Sp_plat.bam
32          cuffdiff        --library-type  fr-firststrand          -o
        diff_out        -b      genome.fa       -L      Sp_ds,Sp_hs,
    Sp_log,Sp_plat      -u   merged_asm/merged.gtf    tophat.Sp_ds.dir/
    Sp_ds.bam      tophat.Sp_hs.dir/Sp_hs.bam      tophat.Sp_log.dir/
    Sp_log.bam    tophat.Sp_plat.dir/Sp_plat.bam
33          head    diff_out/gene_exp.diff
34          return 0
35      }
36
37  main
```

This is a single file non working version that is literally the commands from the pdf file transferred to a bash file. This does not constitute a solution to the problem.

4.2 Attempt 2:

File: mthscript.sh

```
1   #!/bin/bash
2
3   main () {
```

```
4           bowtie-build genome.fa genome -p 4
5           tophat -p 4 -I  1000    -i      20      --bowtie1       --
    library-type  fr-firststrand  -o      tophat.Sp_ds.dir        genome
      Sp_ds.left.fq   Sp_ds.right.fq
6           mv      tophat.Sp_ds.dir/accepted_hits.bam      tophat.
    Sp_ds.dir/Sp_ds.bam
7           samtools        -@ 4 index      tophat.Sp_ds.dir/Sp_ds.bam
8           cufflinks       -p 4 --overlap-radius 1 --library-type  fr-
    firststrand  -o      cufflinks.Sp_ds.dir     tophat.Sp_ds.dir/Sp_ds.
    bam
9           mv cufflinks.Sp_ds.dir/transcripts.gtf  cufflinks.Sp_ds.dir
    /Sp_ds.transcripts.gtf
10          tophat -p 4 -I  1000    -i      20      --bowtie1       --
    library-type  fr-firststrand  -o      tophat.Sp_hs.dir        genome
      Sp_hs.left.fq   Sp_hs.right.fq
11          mv      tophat.Sp_hs.dir/accepted_hits.bam      tophat.
    Sp_hs.dir/Sp_hs.bam
12          samtools        -@ 4 index      tophat.Sp_hs.dir/Sp_hs.bam
13          cufflinks       -p 4 --overlap-radius 1         --library-
    type  fr-firststrand  -o      cufflinks.Sp_hs.dir     tophat.Sp_hs.
    dir/Sp_hs.bam
14          mv      cufflinks.Sp_hs.dir/transcripts.gtf     cufflinks.
    Sp_hs.dir/Sp_hs.transcripts.gtf
15          tophat -p 4 -I  1000    -i      20      --bowtie1       --
    library-type  fr-firststrand  -o      tophat.Sp_log.dir       genome
      Sp_log.left.fq  Sp_log.right.fq
16          mv      tophat.Sp_log.dir/accepted_hits.bam     tophat.
    Sp_log.dir/Sp_log.bam
17          samtools        -@ 4 index      tophat.Sp_log.dir/Sp_log.
    bam
18          cufflinks       -p 4 --overlap-radius 1         --library-
    type  fr-firststrand  -o      cufflinks.Sp_log.dir    tophat.Sp_log.
    dir/Sp_log.bam
19          mv      cufflinks.Sp_log.dir/transcripts.gtf    cufflinks.
    Sp_log.dir/Sp_log.transcripts.gtf
20          tophat -p 4 -I  1000    -i      20      --bowtie1       --
    library-type  fr-firststrand  -o      tophat.Sp_plat.dir
21          mv      tophat.Sp_plat.dir/accepted_hits.bam    tophat.
    Sp_plat.dir/Sp_plat.bam
22          samtools        -@ 4 index      tophat.Sp_plat.dir/Sp_plat.
    bam
23          cufflinks -p 4 --overlap-radius 1       --library-type fr-
    firststrand  -o      cufflinks.Sp_plat.dir   tophat.Sp_plat.dir/
    Sp_plat.bam
24          mv      cufflinks.Sp_plat.dir/transcripts.gtf   cufflinks.
    Sp_plat.dir/Sp_plat.transcripts.gtf
25          echo    cufflinks.Sp_ds.dir/Sp_ds.transcripts.gtf       >>
        assemblies.txt
26          echo    cufflinks.Sp_hs.dir/Sp_hs.transcripts.gtf       >>
        assemblies.txt
27          echo    cufflinks.Sp_log.dir/Sp_log.transcripts.gtf     >>
        assemblies.txt
28          echo    cufflinks.Sp_plat.dir/Sp_plat.transcripts.gtf   >>
        assemblies.txt
29          cat     assemblies.txt
30          cuffmerge       -p 4 -s genome.fa       assemblies.txt
```

```
31          java    -Xmx2G  -jar    /Users/bhaas/IGV/current//igv.jar
        -g      `pwd`/genome.fa `pwd`/merged_asm/merged.gtf,`pwd`/genes
    .bed,`pwd`/tophat.Sp_ds.dir/Sp_ds.bam,`pwd`/tophat.Sp_hs.dir/Sp_hs.
    bam,`pwd`/tophat.Sp_log.dir/Sp_log.bam,`pwd`/tophat.Sp_plat.dir/
    Sp_plat.bam
32          cuffdiff        --library-type  fr-firststrand          -o
        diff_out        -b      genome.fa       -L      Sp_ds,Sp_hs,
    Sp_log,Sp_plat      -u   merged_asm/merged.gtf    tophat.Sp_ds.dir/
    Sp_ds.bam      tophat.Sp_hs.dir/Sp_hs.bam      tophat.Sp_log.dir/
    Sp_log.bam    tophat.Sp_plat.dir/Sp_plat.bam
33          head    diff_out/gene_exp.diff
34          return 0
35
36      }
37
38      main
```

This is a modified version of the script that runs on 4 threads, there is still some amount of repetition in the code that should be abstracted away. It is a single file.

4.3 Attempt 3:

Files: executer.sh, abstract.sh (in that order)

```
1       #!/bin/bash
2
3       main() {
4           runs=('ds' 'hs' 'log' 'plat')
5           bowtie-build genome.fa genome -p 4
6           for data in "${runs[@]}"
7           do
8               echo "Running
    ----------------------------------------------------------------------------
    $data"
9               ./abstract.sh $data
10          done
11          java    -Xmx2G  -jar    /Users/bhaas/IGV/current//igv.jar
        -g      `pwd`/genome.fa `pwd`/merged_asm/merged.gtf,`pwd`/genes
    .bed,`pwd`/tophat.Sp_ds.dir/Sp_ds.bam,`pwd`/tophat.Sp_hs.dir/Sp_hs.
    bam,`pwd`/tophat.Sp_log.dir/Sp_log.bam,`pwd`/tophat.Sp_plat.dir/
    Sp_plat.bam
12          cuffdiff        --library-type  fr-firststrand          -o
        diff_out        -b      genome.fa       -L      Sp_ds,Sp_hs,
    Sp_log,Sp_plat      -u   merged_asm/merged.gtf    tophat.Sp_ds.dir/
    Sp_ds.bam      tophat.Sp_hs.dir/Sp_hs.bam      tophat.Sp_log.dir/
    Sp_log.bam    tophat.Sp_plat.dir/Sp_plat.bam
13          head    diff_out/gene_exp.diff
14      }
15
16      main
```

```
1   #!/bin/bash
2
```

```
tophat -p 4 -I 1000 -i 20 --bowtie1 --library-type fr-firststrand -o tophat.Sp_$1.dir genome Sp_ds.left.fq Sp_ds.right.fq
mv tophat.Sp_$1.dir/accepted_hits.bam tophat.Sp_$1.dir/Sp_$1.bam
samtools index tophat.Sp_$1.dir/Sp_$1.bam -@ 4
cufflinks -p 4 --overlap-radius 1 --library-type fr-firststrand -o cufflinks.Sp_$1.dir tophat.Sp_$1.dir/Sp_$1.bam
mv cufflinks.Sp_$1.dir/transcripts.gtf cufflinks.Sp_$1.dir/Sp_$1.transcripts.gtf
echo cufflinks.Sp_$1.dir/Sp_$1.transcripts.gtf >> assemblies.txt
cat assemblies.txt
cuffmerge -p 4 -s genome.fa assemblies.txt
```

These are two files that interplay running the commands with 4 threads and using a more efficient method for calling similar commands with slight name changes such that the lines of code actually written are minimized.

5 Optional task

Files: hsat1.sh, hsat2.sh

```
#!/bin/bash

main() {
        runs=('ds' 'hs' 'log' 'plat')
        hisat2-build genome.fa genome -p 4
        echo > assemblies.txt
        for data in "${runs[@]}"
        do
                echo "Running ---------- $data"
                ./hsat2.sh $data
        done
        cuffdiff --library-type fr-firststrand -o hsat -b genome.fa -L Sp_ds,Sp_hs,Sp_log,Sp_plat -u merged_asm/merged.gtf Sp_ds.sorted.bam Sp_hs.sorted.bam Sp_log.sorted.bam Sp_plat.sorted.bam
}

main
```

```
#!/bin/bash

hisat2 -p 4 --max-intronlen 1000 --fr --min-intronlen 20 -x genome -1 Sp_$1.left.fq -2 Sp_$1.right.fq -S Sp_$1.sam
samtools view -bS Sp_$1.sam > Sp_$1.bam
samtools sort -o Sp_$1.sorted.bam Sp_$1.bam
samtools index Sp_$1.sorted.bam
cufflinks -p 4 --overlap-radius 1 --library-type fr-firststrand -o hisat.cufflinks.Sp_$1.dir Sp_$1.sorted.bam
mv hisat.cufflinks.Sp_$1.dir/transcripts.gtf hisat.cufflinks.Sp_$1.dir/Sp_$1.transcripts.gtf
```

```
echo     hisat.cufflinks.Sp_$1.dir/Sp_$1.transcripts.gtf >>
   assemblies.txt
cat      assemblies.txt
cuffmerge        -p 4 -s genome.fa        assemblies.txt
```

These resemble the executer.sh and abstract.sh because they are modified copies. the alterations to the flags and the commands came from reading the documentation of each of the tools and finding common flags with -h.

For the difference between the hisat2 and tophat.

```
user03@UoA-Intro2Bio24:~/rnaseq_workshop$ awk '$14 == "yes" && $10 <
   -2.0' hsat/gene_exp.diff | wc -l
awk '$14 == "yes" && $10 > 2.0' hsat/gene_exp.diff | wc -l
4
2
user03@UoA-Intro2Bio24:~/rnaseq_workshop$ awk '$14 == "yes" && $10 <
   -2.0' diff_out/gene_exp.diff | wc -l
0
user03@UoA-Intro2Bio24:~/rnaseq_workshop$ awk '$14 == "yes" && $10 >
   2.0' diff_out/gene_exp.diff | wc -l
0
user03@UoA-Intro2Bio24:~/rnaseq_workshop$
```

Which essentially means that hisat 2 produces more statistically significant results, we can expect that this is due to the different alignment method it uses. At face value it means that hisat2 identifies up regulated and down regulated genes but tophat does not.

Thus concludes this ITBI exercise