Jonathan Wang

Lab 3 Report 1

Answers

Part 1

(1a) How many transcripts did you find in total, in the merged file?

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings> python3 part1ab.py < stringtie_merged.gtf  > part1ab.txt
```

Counted the number of “feature” where it equal “transcript”

- 9947 transcripts in total

(1b) How many distinct genes did you find in total? (Note that each gene has its own gene_id.)
Counted distinct gene_id in stringtie_merged.gtf

- 2926 distinct genes found

(1c) How many transcripts are in the Chr17 annotation file?

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reference_ch17> python3 part1cde.py < CHESS_v3.0_chr17.gtf > part1cde.txt
```

Counted the number of “feature” where it equal “transcript”

- 8324 transcripts in total

(1d) In the Chr17 annotation, how many transcripts are protein coding? (Note that each transcript record has a gene_type label.)
Counted the number of lines where “feature” == “transcript” and “gene_type” == “protein_coding”

- 5770 transcripts are protein coding

(1e) In the Chr17 annotation, how many distinct genes are protein coding?
Counted the number of distinct gene_id with “gene_type” == “protein_coding”

- 1323 distinct genes are protein coding

Part 2

(2a) How many of your transcripts exactly match all the introns of a known gene from the CHESS annotation?

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings> python part2abc.py < merged.annotated.gtf > part2abc.txt
```

Counted the number of transcripts with class code “=”

- 8149 transcripts directly match all the introns of a known gene from the CHESS annotation.

(2b) How many novel transcripts (i.e., they match a protein-coding gene, but they do not match any of the intron chains in the annotated transcripts) did you find in protein-coding gene loci?
Counted the number of transcripts that don't have class_code "u" and "="

- 1755 novel transcripts in protein-coding gene loci.

(2c) How many of your novel transcripts occur at entirely novel locations (code "u" from gffcompare)?
Counted the number of transcripts that have class_code "u"

- 43 of novel transcripts occur at entirely novel locations.

Part 3

(3a) Among all the transcripts you assembled, and among all 11 samples, which one has the highest TPM? Report the transcript record (just the 'transcript' line) for this one as well as the sample in which you found it.

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> python part3abcScripts.py > part3abc.txt
```

- File with Highest TPM: SRR479070_aligned_reestimate.gtf
- Highest TPM: 64376.265625
- Line with highest TPM: ['chr17', 'StringTie', 'transcript', '81509971', '81512851', '1000', '-', '.', 'gene_id "MSTRG.1643"; transcript_id "CHS.23541.3"; ref_gene_name "ACTG1"; cov "6350.230957"; FPKM "19994.509766"; TPM "64376.265625";']
- Sample: SRR479070

(3b) Looking across all 11 samples, how many distinct transcripts have a TPM above 0?

Generated set across all 11 samples with unique transcript_ids that have TPM > 0 → counted number of items in set

- 6605 distinct transcripts

(3c) How many distinct genes have a TPM above 0?
Generated set across all 11 samples with unique gene_ids that have TPM > 0 → counted number of items in set

- 1728 distinct genes

(3d) For every transcript, find its maximum TPM in all 11 samples. Report how many distinct transcripts have a maximum TPM greater than 50.

bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> python part3dScripts.py > part3d.txt

- 3439 distinct transcripts with maximum TPM greater than 50

You can find the maximum TPM for every distinct transcript in all 11 samples in part3d.txt

(3e) This one takes a bit more work. Sample SRR47952 is a control sample, and SRR47954 is a sample that was treated with a cancer drug, diarylpropionitrile (DPN). What you are doing in this exercise is just the beginning of an analysis to determine what genes were affected by the drug treatment. For these two samples, SRR47952 and SRR47954, compute the total expression in TPM for each gene. This requires you to sum up all of the transcript TPM values for each gene. There will be nearly 3000 genes in your output, but we only want you to report the top 10 most-highly expressed genes, along with their total TPM values, for each sample. You will notice that the lists for SRR47952 and SRR47954 are different–think about whether you can attribute those differences to the drug treatment.

bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> python part3eScripts.py < SRR479052_aligned_reestimate.gtf > SRR479052_part3e.txt

bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> python part3eScripts.py < SRR479054_aligned_reestimate.gtf > SRR479054_part3e.txt

```
The top 10 most-highly expressed genes for SRR479052
Gene ID: MSTRG.1643, TPM Count: 59688.204346
Gene ID: MSTRG.397, TPM Count: 50563.534505
Gene ID: MSTRG.894, TPM Count: 25460.202881
Gene ID: MSTRG.766, TPM Count: 23531.269165
Gene ID: MSTRG.136, TPM Count: 22613.946777
Gene ID: MSTRG.267, TPM Count: 18773.950684
Gene ID: MSTRG.1144, TPM Count: 15469.583821
Gene ID: MSTRG.1408, TPM Count: 12852.768072
Gene ID: MSTRG.27, TPM Count: 12799.637696
Gene ID: MSTRG.548, TPM Count: 10884.794376
```

```
The top 10 most-highly expressed genes for SRR479054
Gene ID: MSTRG.1643, TPM Count: 55634.097519
Gene ID: MSTRG.397, TPM Count: 47694.92952
Gene ID: MSTRG.894, TPM Count: 25525.385204
Gene ID: MSTRG.766, TPM Count: 22697.981886
Gene ID: MSTRG.136, TPM Count: 20070.359863
Gene ID: MSTRG.267, TPM Count: 19625.865669
Gene ID: MSTRG.1144, TPM Count: 14512.044401
Gene ID: MSTRG.1408, TPM Count: 13507.776377
Gene ID: MSTRG.27, TPM Count: 13386.172852
Gene ID: MSTRG.548, TPM Count: 11991.068077
```

It appears the top 10 most-highly expressed genes are common among the two samples, but the total TPM seemed to decrease for half of the 10 expressed genes SRR479054 (the sample treated with the cancer drug): the other five gene ids to not decrease in total TPM from SRR479052 to SRR479054 were MSTRG.894, MSTRG.267, MSTRG.1408, MSTRG.27, and MSTRG.548.

I decided to look at the genes that had a decrease in TPM and ultimately found that all of the genes that decreased in expression were clinically related and proven to have ties to progression of cancer and for some genes, even more specifically parathyroid adenoma.

MSTRG.1643 → ACTG1 gene (Actin gamma 1) → 10.1210/jendso/bvac096
MSTRG.397 → Ubiquitin B gene → 10.3390/jcm8030297
MSTRG.766 → Ribosomal Protein L19 → https://doi.org/10.1158/1078-0432.CCR-05-2445
MSTRG.136 → Profilin 1 → 10.1186/1477-5956-9-29
MSTRG.1144 → NME/NM23 nucleoside diphosphate kinase 1 → https://doi.org/10.1002/humu.23337

Protocol

1) Run StringTie2 to assemble all of the alignments you created for Project 3, part 1. First you'll run StringTie, guided by the example in step 3 of the protocol in the Nature Protocols paper by Pertea et al. Next you'll merge these 11 files, following the example in step 4 of the protocol, by using stringtie with the --merge option.

1st Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479052_aligned.gtf -l SRR479052_aligned ../aligned_readings/SRR479052_aligned.bam
```

2nd Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479054_aligned.gtf -l SRR479054_aligned ../aligned_readings/SRR479054_aligned.bam
```

3rd Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings>  stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479056_aligned.gtf -l SRR479056_aligned ../aligned_readings/SRR479056_aligned.bam
```

4th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479058_aligned.gtf -l SRR479058_aligned ../aligned_readings/SRR479058_aligned.bam
```

5th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479061_aligned.gtf -l SRR479061_aligned ../aligned_readings/SRR479061_aligned.bam
```

6th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479064_aligned.gtf -l SRR479064_aligned ../aligned_readings/SRR479064_aligned.bam
```

7th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479066_aligned.gtf -l SRR479066_aligned ../aligned_readings/SRR479066_aligned.bam
```

8th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479068_aligned.gtf -l SRR479068_aligned ../aligned_readings/SRR479068_aligned.bam
```

9th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479070_aligned.gtf -l SRR479070_aligned ../aligned_readings/SRR479070_aligned.bam
```

10th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479073_aligned.gtf -l SRR479073_aligned ../aligned_readings/SRR479073_aligned.bam
```

11th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o SRR479076_aligned.gtf -l SRR479076_aligned ../aligned_readings/SRR479076_aligned.bam
```

2) Next you'll merge these 11 files, following the example in step 4 of the protocol, by using stringtie with the --merge option. After you do this, answer the following:

Create mergelist.txt

```
SRR479052_aligned.gtf
SRR479058_aligned.gtf
SRR479066_aligned.gtf
SRR479073_aligned.gtf
SRR479054_aligned.gtf
SRR479061_aligned.gtf
SRR479068_aligned.gtf
SRR479076_aligned.gtf
SRR479056_aligned.gtf
SRR479064_aligned.gtf
SRR479070_aligned.gtf
```

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/assembled_readings> stringtie --merge -p 8 -G /opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -o ../merged_readings/stringtie_merged.gtf mergelist.txt
```

2) Use gffcompare to compare your file to the guide annotation from CHESS 3.0.

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings> gffcompare -r
/opt/ccb/data/grch38/CHESS_v3.0_chr17.gtf -G -o merged stringtie_merged.gtf
```

3) Now use StringTie2 to re-estimate the expression levels for all the transcripts in each of your 11 samples, using your merged file as the reference (provided to StringTie2 with the "-G" option). These levels will be expressed in TPM, or transcripts per million. Note that Step 6 of the protocol shows how to do this with Ballgown, but you won't be using Ballgown, so you can name your output files whatever you like here. We suggest that if you originally named a file something like "SRR1234.gtf", you might use "SRR1234_reestimate.gtf" here.

1st Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e
-p 8 -G
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_
merged.gtf -o SRR479052_aligned_reestimate.gtf
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479052
_aligned.bam
```

2nd Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e
-p 8 -G
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_
merged.gtf -o SRR479054_aligned_reestimate.gtf
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479054
_aligned.bam
```

3rd Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e
-p 8 -G
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_
merged.gtf -o SRR479054_aligned_reestimate.gtf
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479054
_aligned.bam
```

4th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e
-p 8 -G
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_
merged.gtf -o SRR479056_aligned_reestimate.gtf
/home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479056
_aligned.bam
```

5th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479058_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479058_aligned.bam
```

6th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479061_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479061_aligned.bam
```

7th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479064_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479064_aligned.bam
```

8th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479066_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479066_aligned.bam
```

9th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479068_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479068_aligned.bam
```

10th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479070_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479070_aligned.bam
```

10th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479073_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479073_aligned.bam
```

11th Sample

```
bme-manatee:~/lab3/RNAseq_project/parathyroid_tumor_samples/reestimate_readings> stringtie -e -p 8 -G /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/merged_readings/stringtie_merged.gtf -o SRR479076_aligned_reestimate.gtf /home/WIN/jwang428/lab3/RNAseq_project/parathyroid_tumor_samples/aligned_readings/SRR479076_aligned.bam
```