Sequencing Analysis Pipeline

(SAP)

Instructions

Nadège Pulgar-Vidal

Description and Purpose

- Reads we get back from sequencing have to go through several bioinformatics tools to yield useful and interpretable data
- These tools (Trimmomatic, Tophat2 with Bowtie2-build, Cuffquant, Cuffnorm, and Cuffdiff) are run from the terminal/command line and require lengthy commands with several settings as well as input files in a specific order and format

- This can be complicated for someone with no previous experience using the terminal/command line and is tedious when dealing with a great quantity of reads
- The goal of the Sequencing Analysis Pipeline (SAP) is to provide a friendlier, interactive prompt system that will guide the user through the necessary inputting process and then build the commands and/or run the reads through all the aforementioned programs

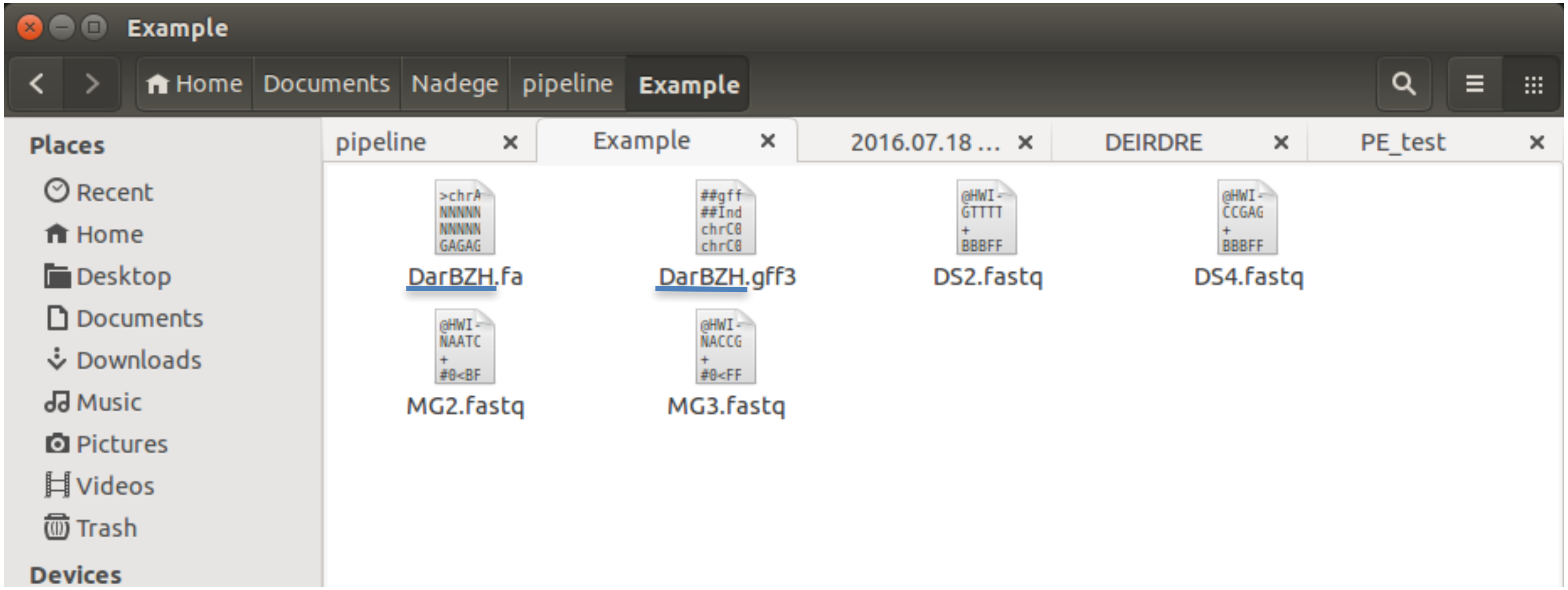
Additional features

- The prompt system includes several verification steps and has multiple error checks in place (input format, existing files, reasonable values)
- The output of the different programs is neatly organized into folders
- A “commands.txt” file is built containing all the commands required to run the user’s reads through the pipeline’s programs
- A “trim_log.txt” file saves the Trimmomatic console output, including the number of input and surviving reads
- The alignment summaries for individual reads provided by Tophat2 are concatenated into “all_align_summaries.txt” for easier access

Instructions

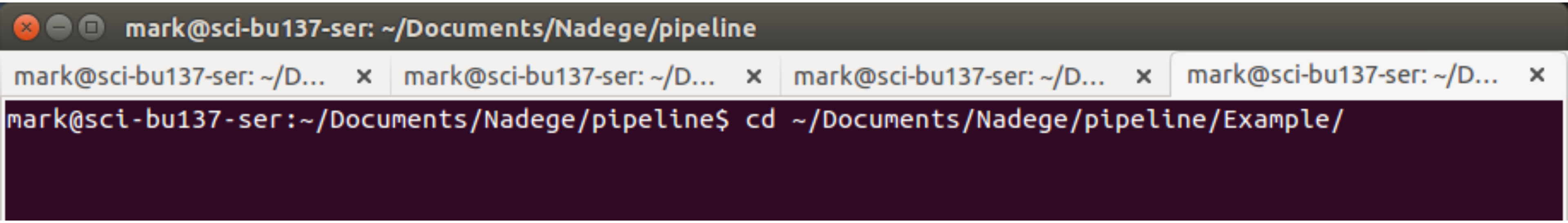
Step 1: Prep

Group all your reads files (.fastq), your genome file (.gff/.gff3), and your fasta file (.fa) into one folder. (In this example we will use a small sample of reads) Make sure the genome file and the fasta file have the same name before the extension.

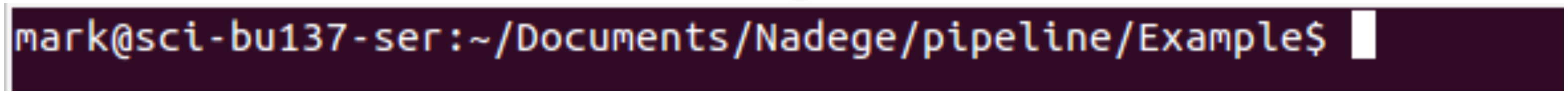


Step 2: Navigate to directory

Open terminal/command line (example uses terminal in Ubuntu) and navigate to the directory you placed all your files in (Example in this case).


mark@sci-bu137-ser: ~/Documents/Nadege/pipeline
mark@sci-bu137-ser: ~/D...
mark@sci-bu137-ser: ~/D...
mark@sci-bu137-ser: ~/D...
mark@sci-bu137-ser: ~/D...

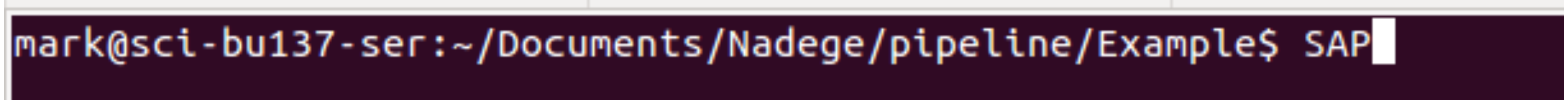
```
mark@sci-bu137-ser:~/Documents/Nadege/pipeline$ cd ~/Documents/Nadege/pipeline/Example/
```



```
mark@sci-bu137-ser:~/Documents/Nadege/pipeline/Example$ 
```

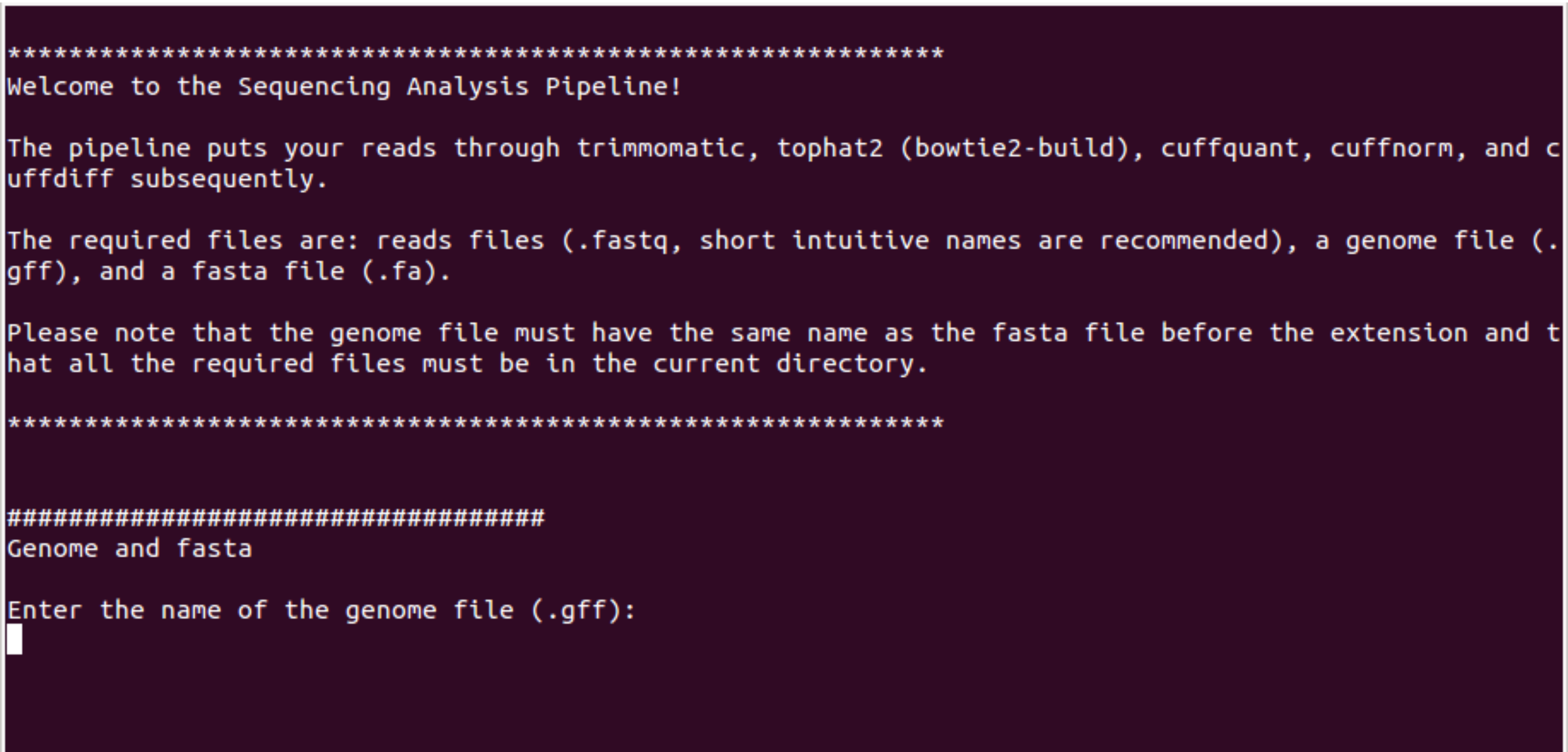
Step 3: Launch

Launch the program by typing the bin shortcut “SAP” then press enter



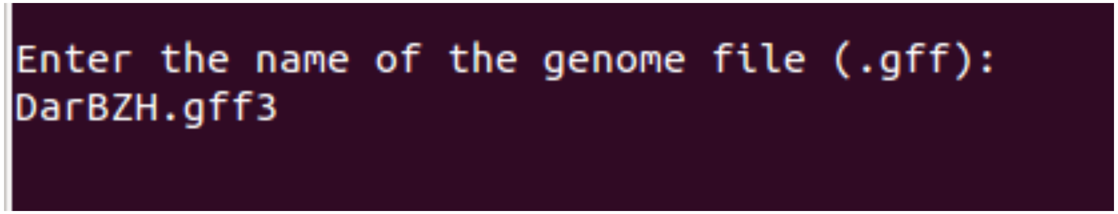
```
mark@sci-bu137-ser:~/Documents/Nadege/pipeline/Example$ SAP
```

You will see the program opening banner and the first prompt:



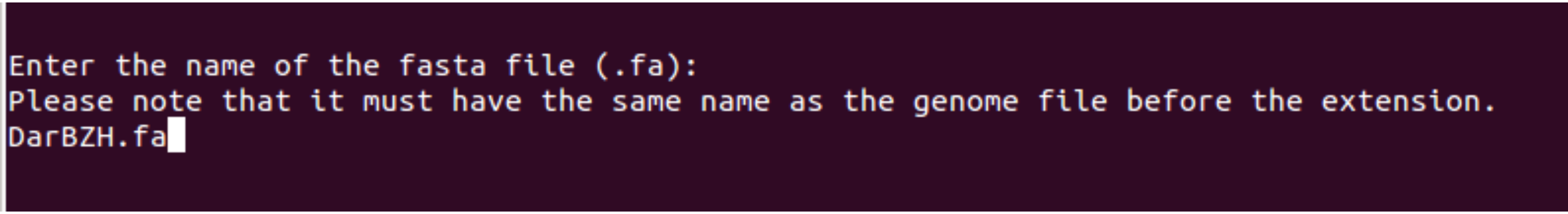
```
*************************************************************
Welcome to the Sequencing Analysis Pipeline!

The pipeline puts your reads through trimmomatic, tophat2 (bowtie2-build), cuffquant, cuffnorm, and c
uffdiff subsequently.

The required files are: reads files (.fastq, short intuitive names are recommended), a genome file (.
gff), and a fasta file (.fa).

Please note that the genome file must have the same name as the fasta file before the extension and t
hat all the required files must be in the current directory.

*************************************************************



###################################
Genome and fasta

Enter the name of the genome file (.gff):
```

Step 4: Genome and fasta files



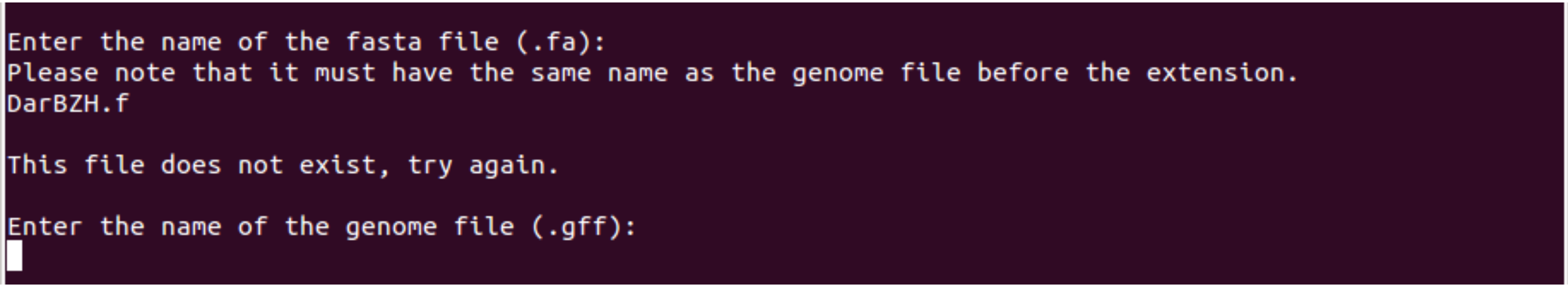
```
Enter the name of the genome file (.gff):
DarBZH.gff3
```

Type the name of the genome file including the extension (.gff/.gff3) then press enter. Do the same for the fasta file (.fa) at the next prompt.

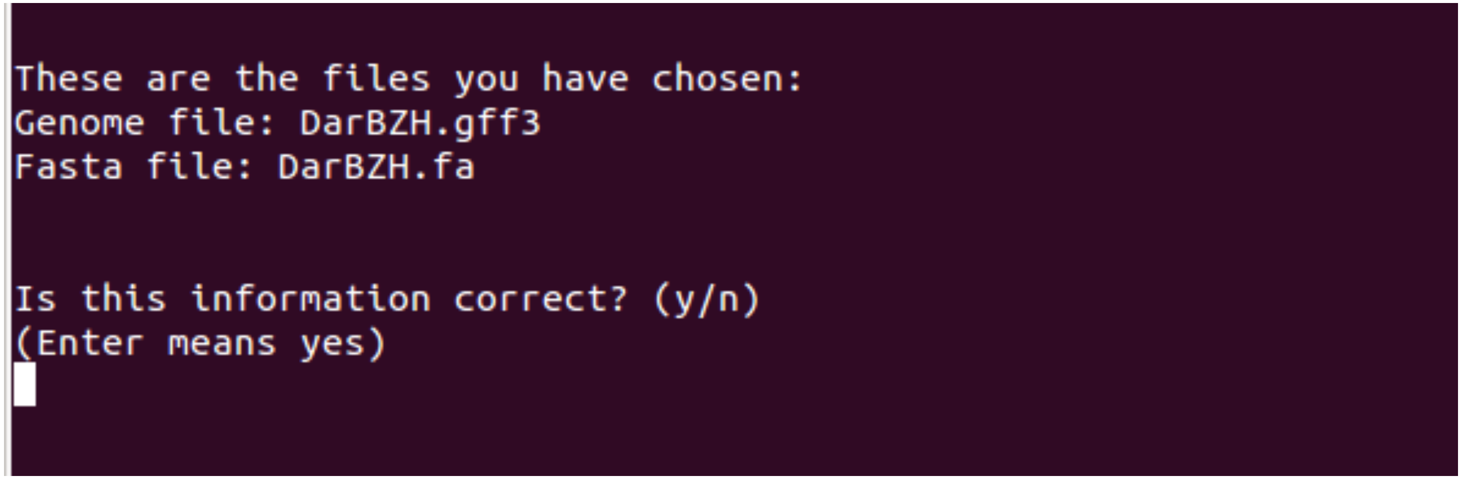


```
Enter the name of the fasta file (.fa):
Please note that it must have the same name as the genome file before the extension.
DarBZH.fa
```

If you mistype or the file isn’t in the folder, the program will print out a simple error message then prompt you to try again starting at the genome file input.



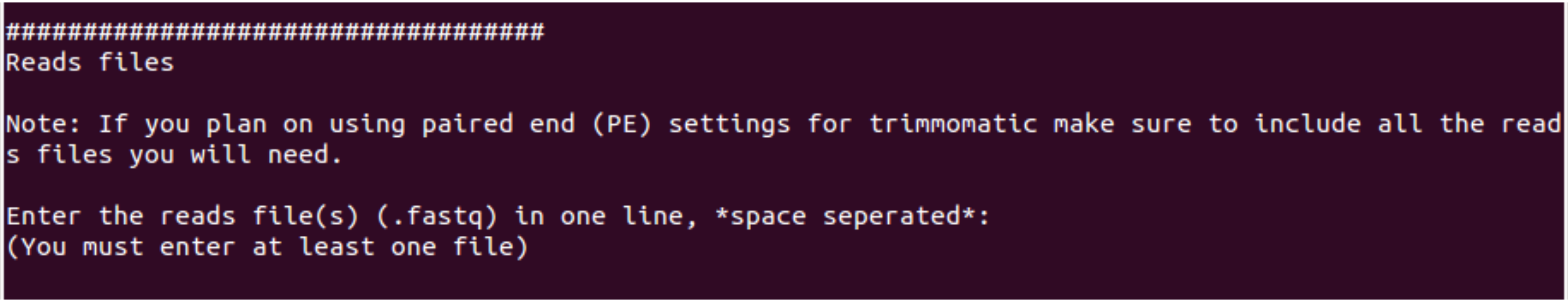
```
Enter the name of the fasta file (.fa):
Please note that it must have the same name as the genome file before the extension.
DarBZH.f

This file does not exist, try again.

Enter the name of the genome file (.gff):
```



```
These are the files you have chosen:
Genome file: DarBZH.gff3
Fasta file: DarBZH.fa


Is this information correct? (y/n)
(Enter means yes)
█
```

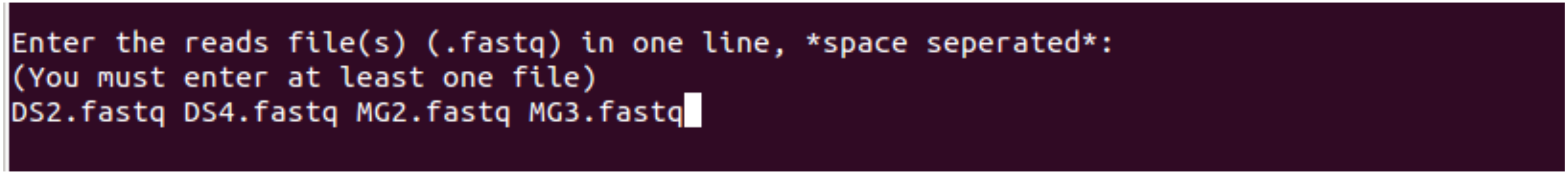
Once you have entered existing file names for genome and fasta files and these two files have the same name before the extension, the program will ask you to verify the information you entered. If the information is incorrect type “n” or “no” then press enter and the program will take you back to the beginning of the genome and fasta file prompt. If you enter nothing or anything other than “n” (case insensitive) the program will assume the information is correct and move on. All verification steps work the same way.

Step 5: Reads files



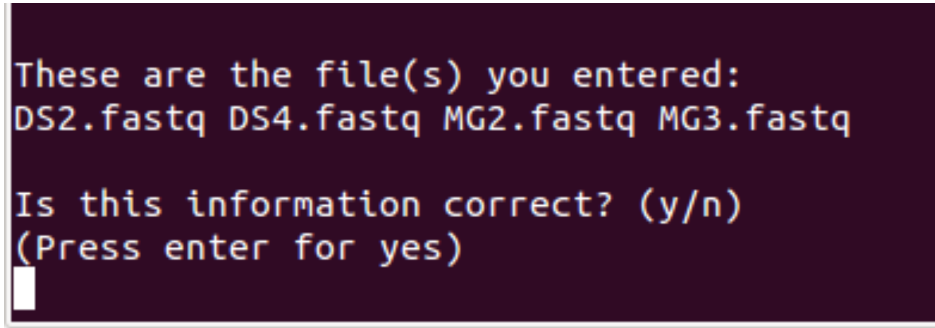
```
###################################
Reads files

Note: If you plan on using paired end (PE) settings for trimmomatic make sure to include all the read
s files you will need.

Enter the reads file(s) (.fastq) in one line, *space seperated*:
(You must enter at least one file)
```

Once you reach this prompt type out the names of your reads files with extensions in a space separated list as shown in the example below:

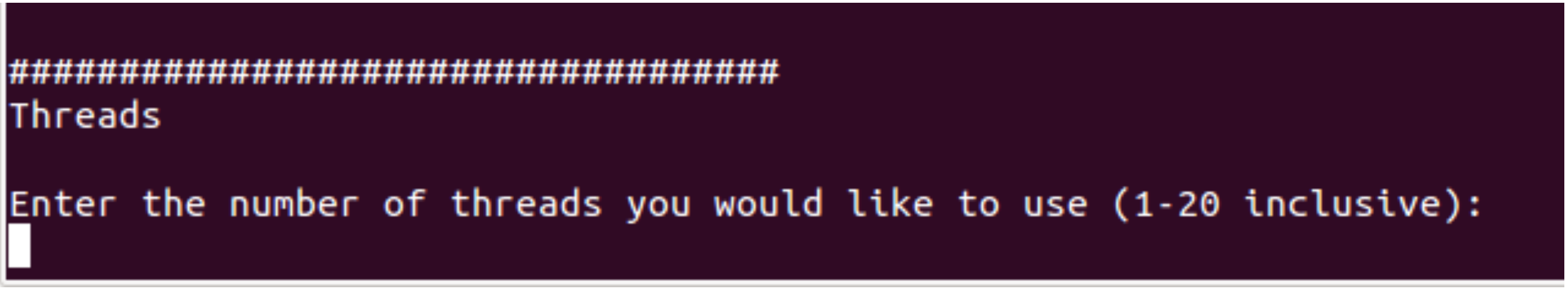


```
Enter the reads file(s) (.fastq) in one line, *space seperated*:
(You must enter at least one file)
DS2.fastq DS4.fastq MG2.fastq MG3.fastq
```

Once you have entered a list of files that are all valid, you will be prompted for verification. Once more, everything other than “n” or “no” will be interpreted as yes.



```
These are the file(s) you entered:
DS2.fastq DS4.fastq MG2.fastq MG3.fastq

Is this information correct? (y/n)
(Press enter for yes)
```

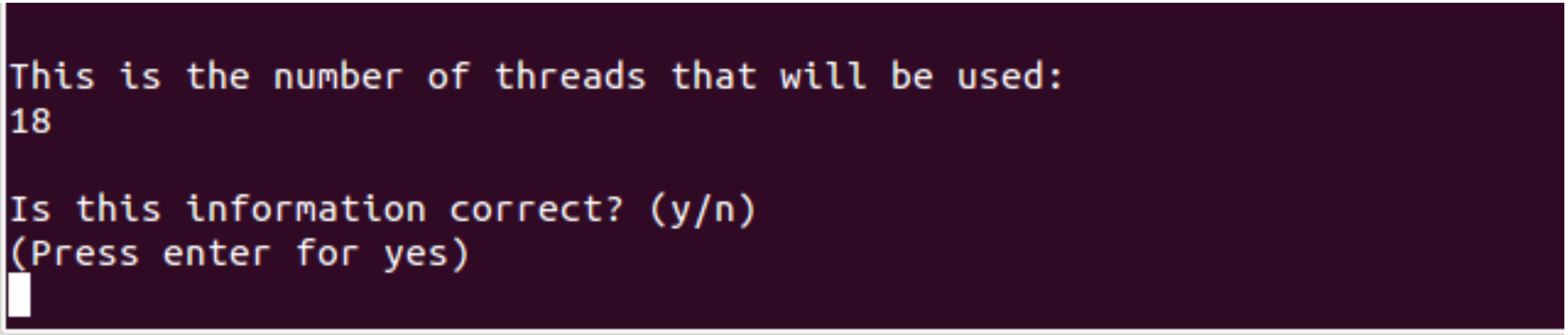


```
####################################
Threads

Enter the number of threads you would like to use (1-20 inclusive):
```

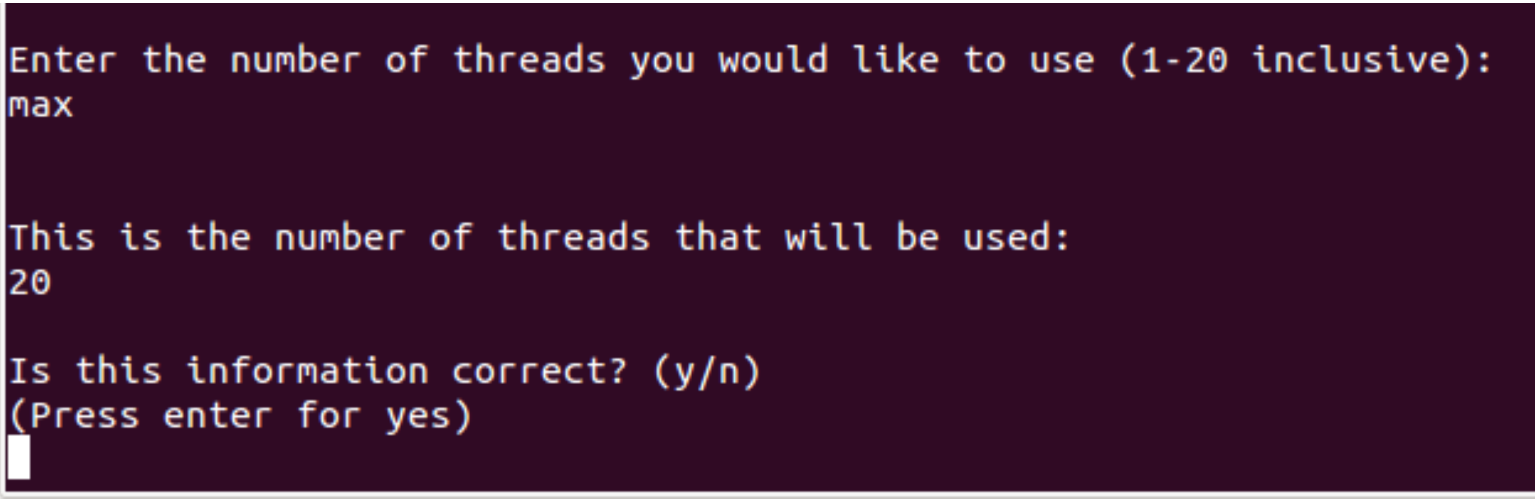
Step 6: Threads

Here the program prompts you for how many threads you would like to use. Under normal conditions this is restricted from 1 to 20 inclusive. Type valid a number then press enter. If you enter an invalid value the program will display a simple error message and tell you to try again. If you enter nothing the current value will be kept (initially 15 threads). If you enter a number with a decimal place it will be truncated to an integer, not rounded.

Note: All programs except bowtie will use this thread number

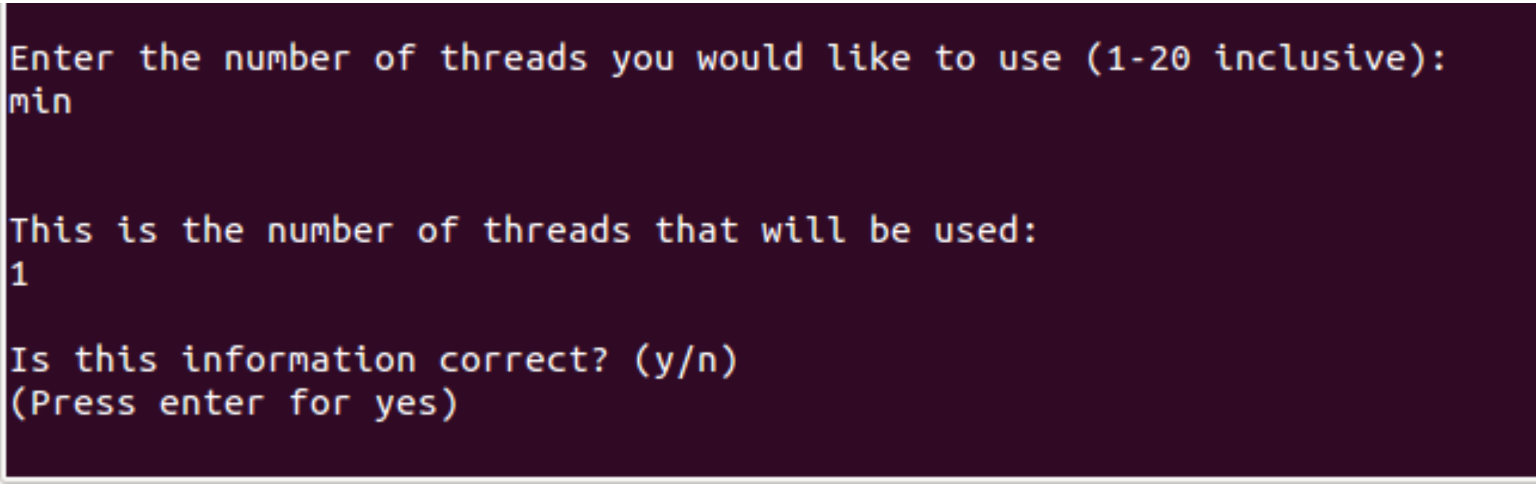


```
This is the number of threads that will be used:
18

Is this information correct? (y/n)
(Press enter for yes)
```



```
Enter the number of threads you would like to use (1-20 inclusive):
max


This is the number of threads that will be used:
20

Is this information correct? (y/n)
(Press enter for yes)
```

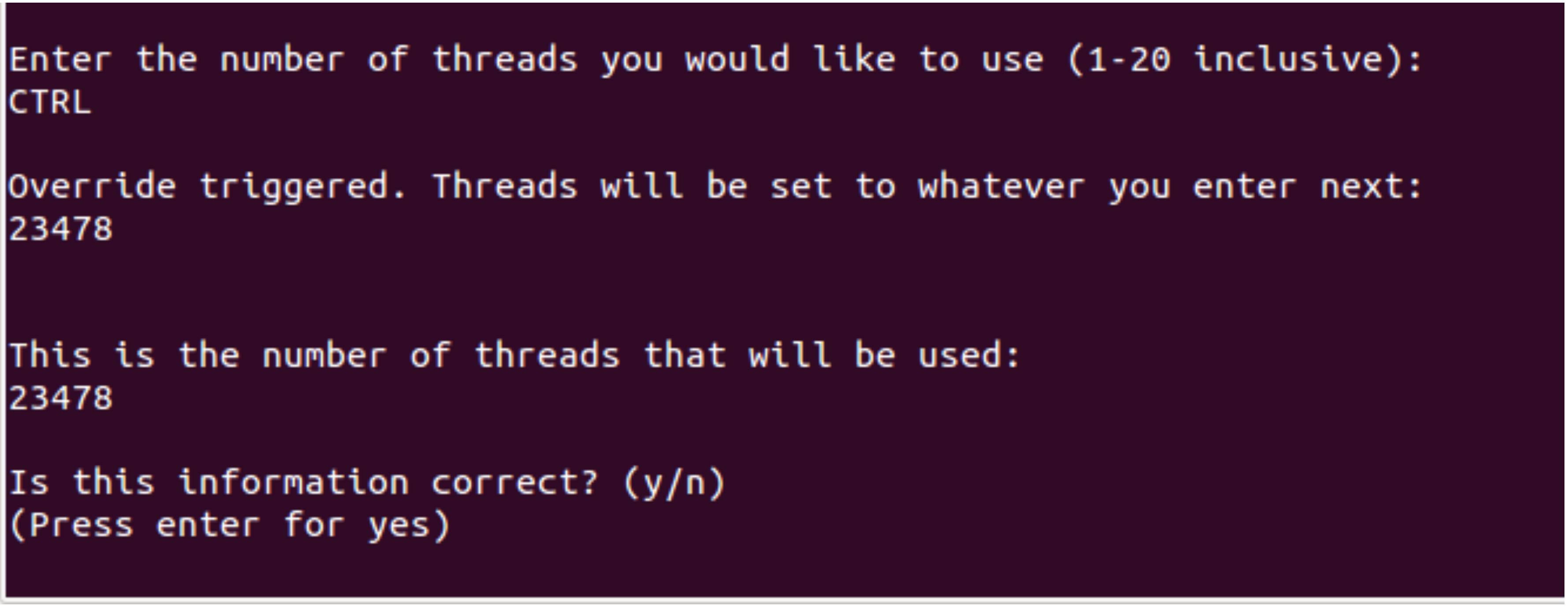
There are some special inputs for the threads prompt:
-“max” will set the threads to whatever the maximum is (currently 20 in the code)
-“min” will set the threads to whatever the minimum is (currently 1 in the code)



```
Enter the number of threads you would like to use (1-20 inclusive):
min


This is the number of threads that will be used:
1

Is this information correct? (y/n)
(Press enter for yes)
```

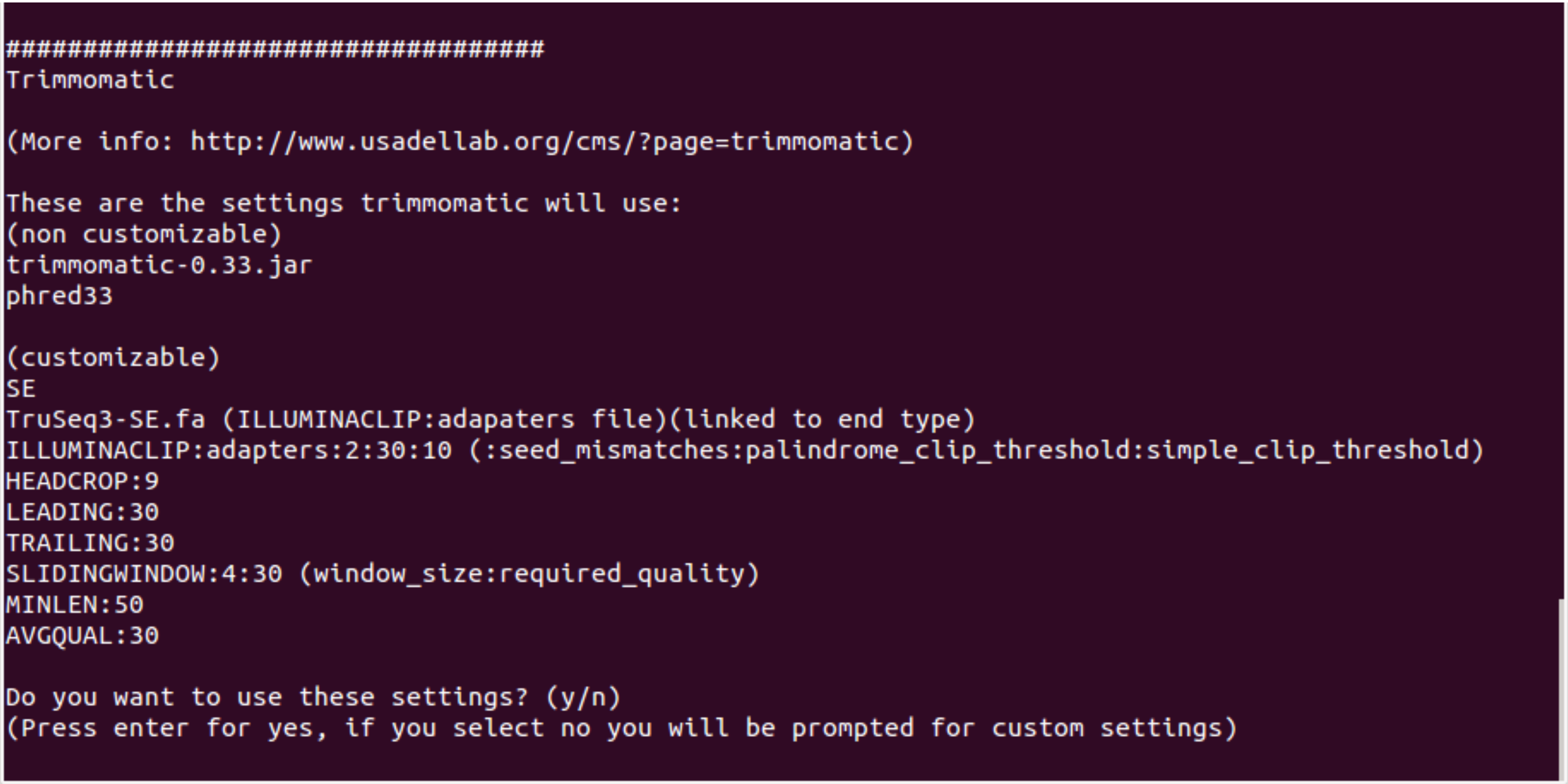
There is also a special code that allows the user to bypass the normal error checks for thread number input: “CTRL”

This will allow you to assign any value they want (including non numerical values) to the threads variable that will be used to build all commands.
If you enter an invalid value your commands will not run properly so be careful.

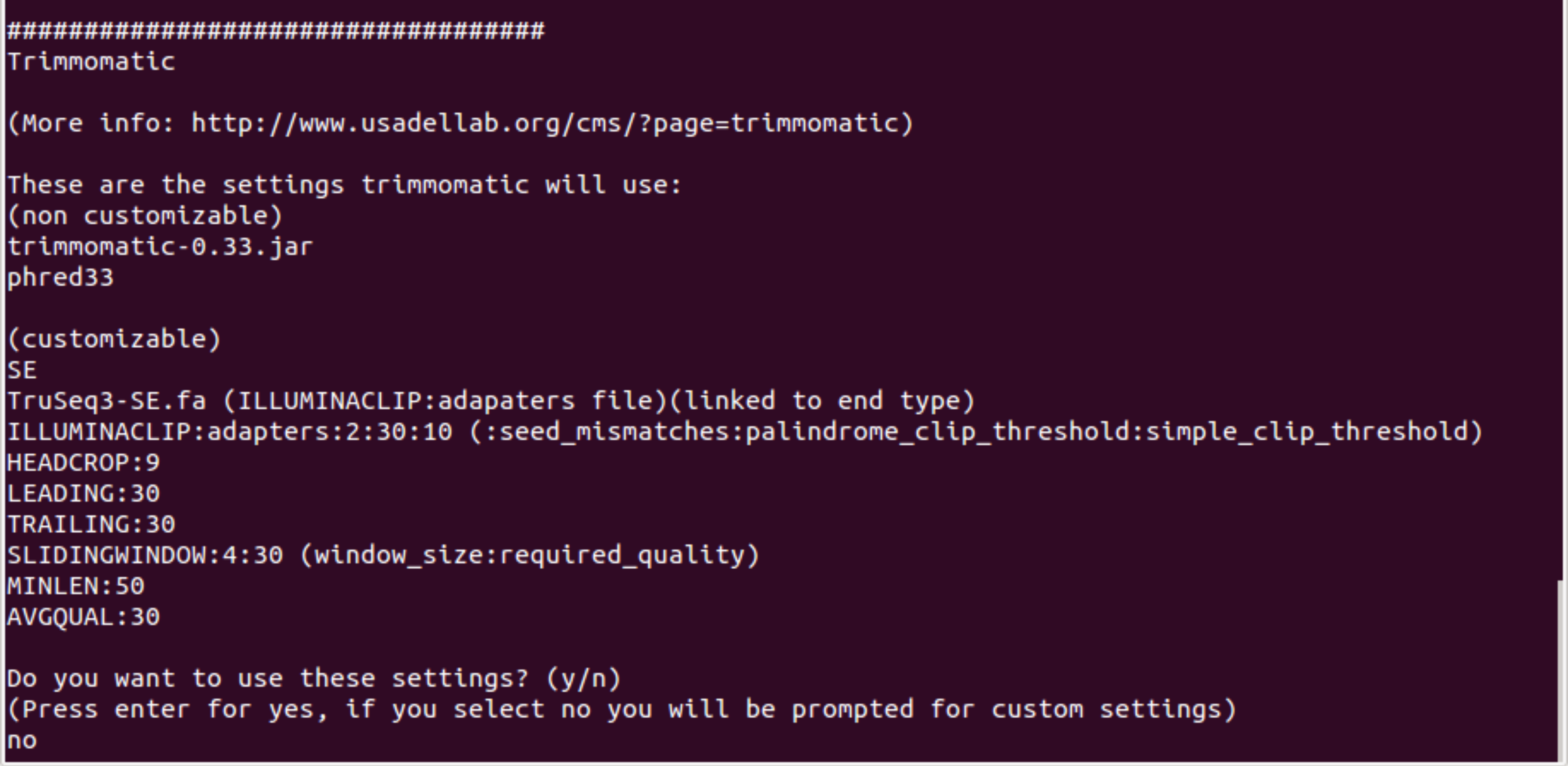


```
Enter the number of threads you would like to use (1-20 inclusive):
CTRL

Override triggered. Threads will be set to whatever you enter next:
23478


This is the number of threads that will be used:
23478

Is this information correct? (y/n)
(Press enter for yes)
```

Step 7: Trimmomatic settings



```
####################################
Trimmomatic

(More info: http://www.usadellab.org/cms/?page=trimmomatic)

These are the settings trimmomatic will use:
(non customizable)
trimmomatic-0.33.jar
phred33

(customizable)
SE
TruSeq3-SE.fa (ILLUMINACLIP:adapaters file)(linked to end type)
ILLUMINACLIP:adapters:2:30:10 (:seed_mismatches:palindrome_clip_threshold:simple_clip_threshold)
HEADCROP:9
LEADING:30
TRAILING:30
SLIDINGWINDOW:4:30 (window_size:required_quality)
MINLEN:50
AVGQUAL:30

Do you want to use these settings? (y/n)
(Press enter for yes, if you select no you will be prompted for custom settings)
```

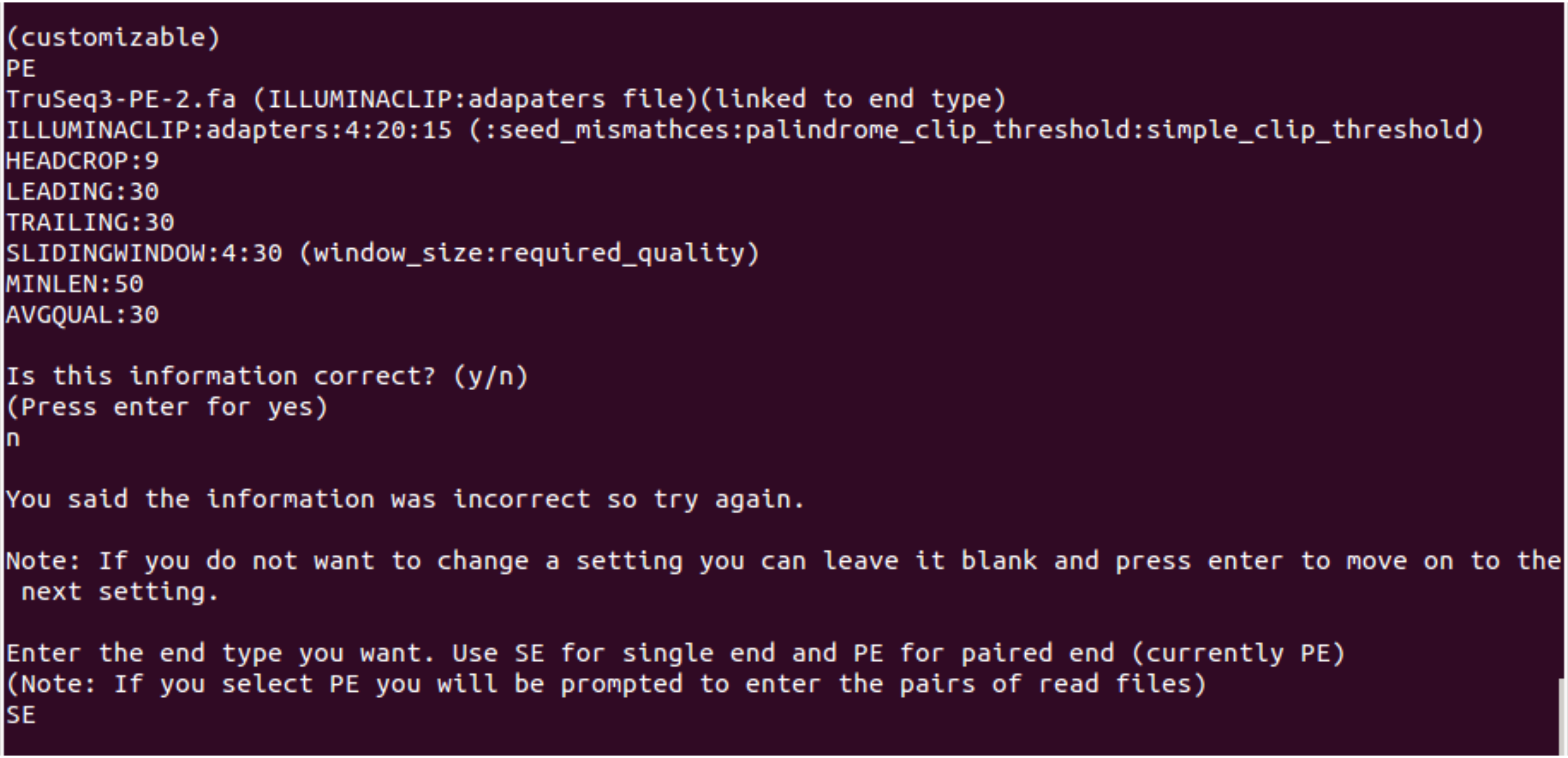


```
###################################
Trimmomatic

(More info: http://www.usadellab.org/cms/?page=trimmomatic)

These are the settings trimmomatic will use:
(non customizable)
trimmomatic-0.33.jar
phred33

(customizable)
SE
TruSeq3-SE.fa (ILLUMINACLIP:adapaters file)(linked to end type)
ILLUMINACLIP:adapters:2:30:10 (:seed_mismatches:palindrome_clip_threshold:simple_clip_threshold)
HEADCROP:9
LEADING:30
TRAILING:30
SLIDINGWINDOW:4:30 (window_size:required_quality)
MINLEN:50
AVGQUAL:30

Do you want to use these settings? (y/n)
(Press enter for yes, if you select no you will be prompted for custom settings)
no
```

The program will print the default settings for Trimmomatic and then prompt you to ask whether you want to use these defaults or change them.

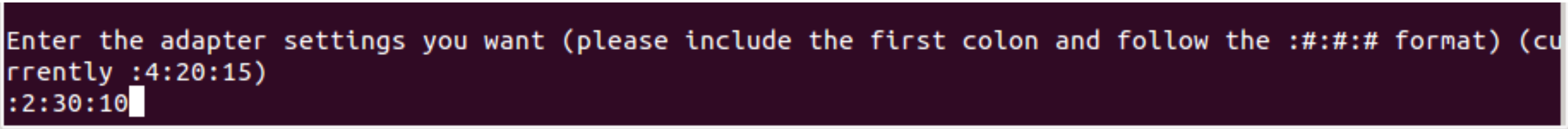
If you want to change the settings enter “n” or “no”. If you enter anything else or nothing the program will assume you want to use these default settings and move on.

For all of the settings prompts, if you enter nothing or an invalid value (out of bounds or wrong format), the program will use the current value for that setting.



```
(customizable)
PE
TruSeq3-PE-2.fa (ILLUMINACLIP:adapaters file)(linked to end type)
ILLUMINACLIP:adapters:4:20:15 (:seed_mismathces:palindrome_clip_threshold:simple_clip_threshold)
HEADCROP:9
LEADING:30
TRAILING:30
SLIDINGWINDOW:4:30 (window_size:required_quality)
MINLEN:50
AVGQUAL:30

Is this information correct? (y/n)
(Press enter for yes)
n

You said the information was incorrect so try again.

Note: If you do not want to change a setting you can leave it blank and press enter to move on to the
 next setting.

Enter the end type you want. Use SE for single end and PE for paired end (currently PE)
(Note: If you select PE you will be prompted to enter the pairs of read files)
SE
```

The first setting you are prompted for is the end type. SE/se (or S/s) for single end reads (default) and PE/pe (or P/p) for paired end reads. The adapters file will be adjusted accordingly.

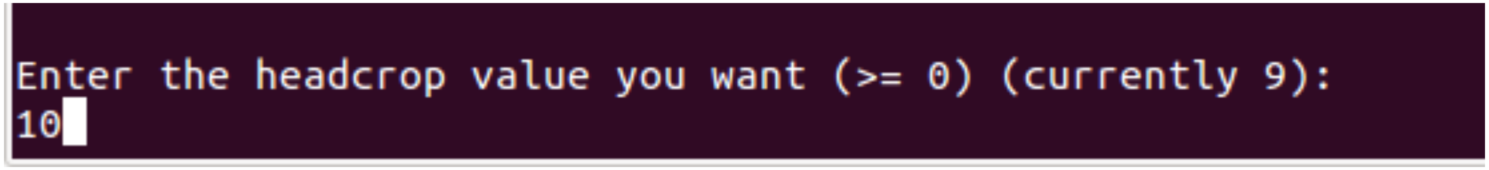


```
Enter the adapter settings you want (please include the first colon and follow the :#:#:# format) (cu
rrently :4:20:15)
:2:30:10
```

Then you will be prompted for the adapter settings. The numbers correspond to these settings:

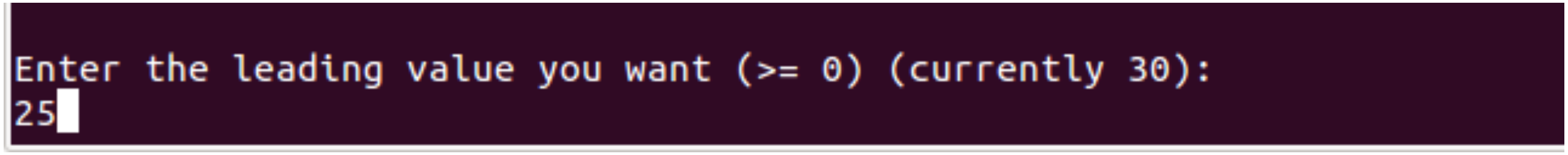
: seed_mismatches : palindrome_clip_threshold : simple_clip_threshold

Type out the settings you want in the correct format (:#:#:#) then press enter.



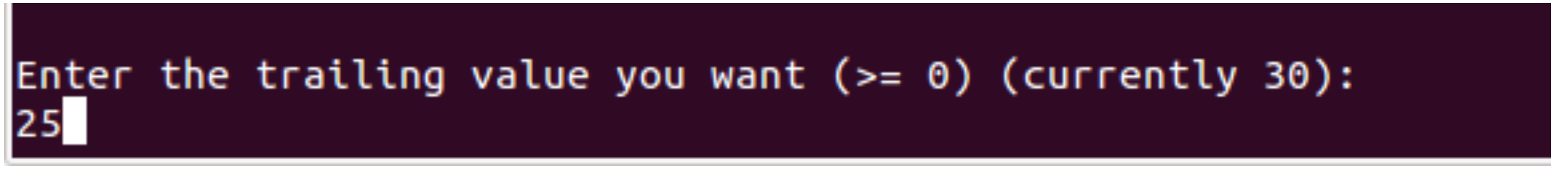
```
Enter the headcrop value you want (>= 0) (currently 9):
10
```

Then you will be prompted for the headcrop value (how many base pairs to trim off the beginning of the read). Type a non-negative integer then press enter.



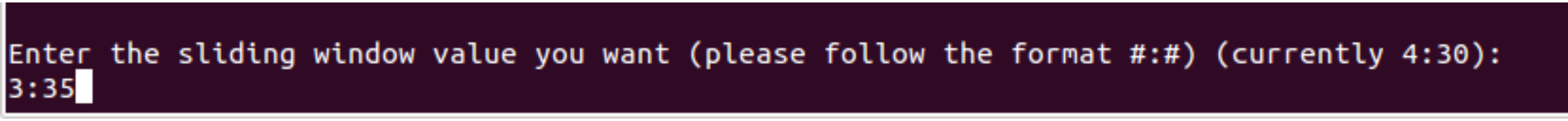
```
Enter the leading value you want (>= 0) (currently 30):
25
```

Then you will be prompted for the leading value (minimum quality at the start of the read). Type a non-negative integer then press enter.



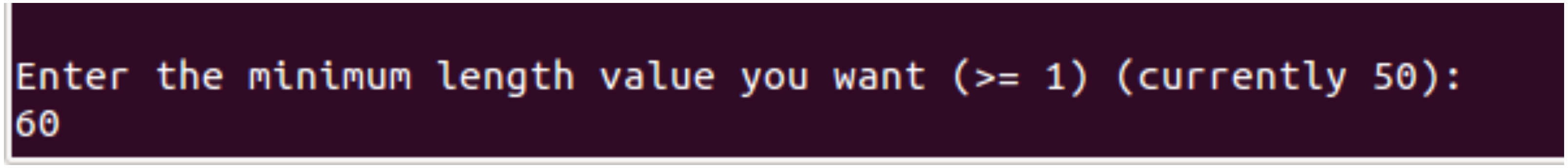
```
Enter the trailing value you want (>= 0) (currently 30):
25
```

Then you will be prompted for trailing value (minimum quality at the end of the read). Type a non-negative integer then press enter.



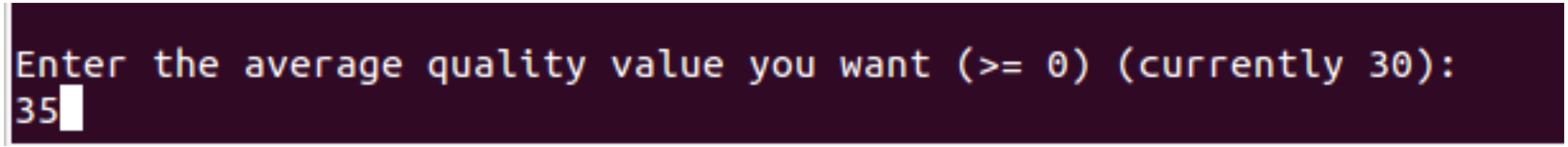
```
Enter the sliding window value you want (please follow the format #:#) (currently 4:30):
3:35
```

Then you will be prompted for sliding window settings (cut if the average quality falls below threshold within the window size). The numbers correspond to:
window_size : required_quality
Type out the settings you want in the correct format (#:#) then press enter.



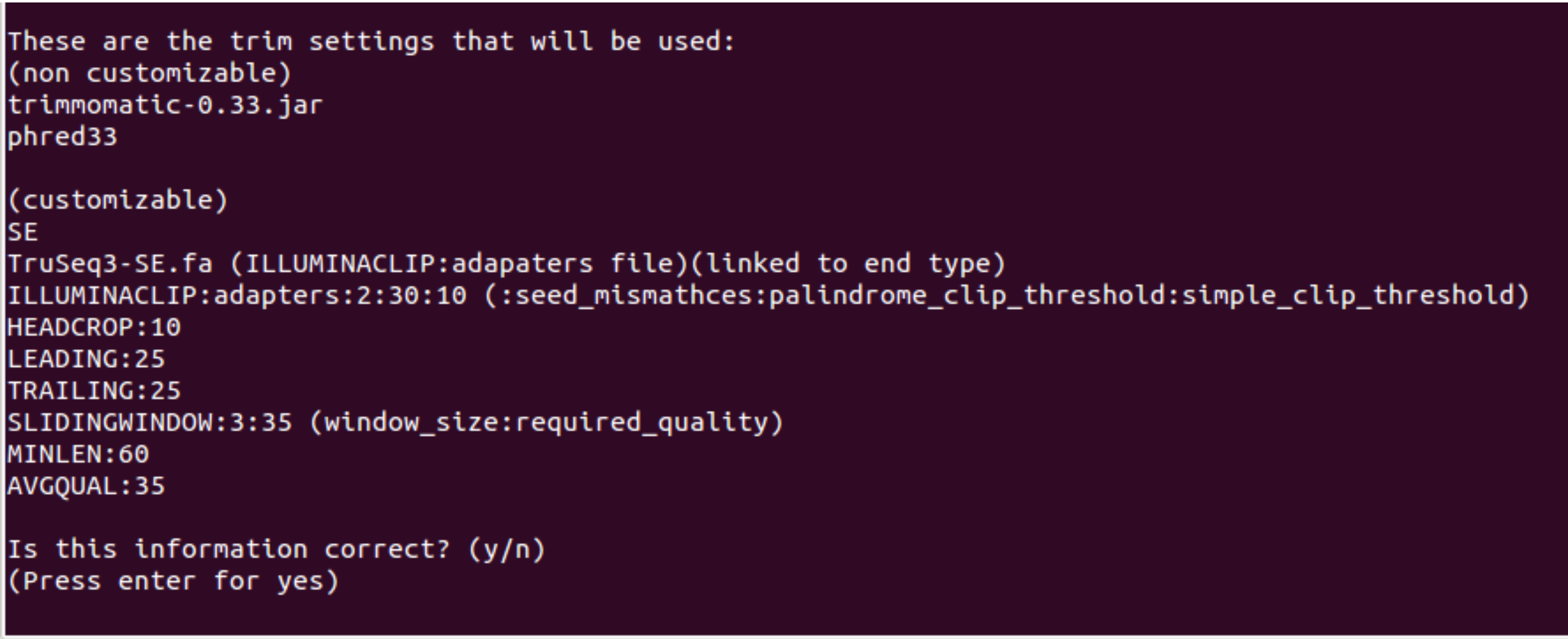
```
Enter the minimum length value you want (>= 1) (currently 50):
60
```

Then you will be prompted for the minimum length value (shorter reads will be discarded). Type a positive integer then press enter.



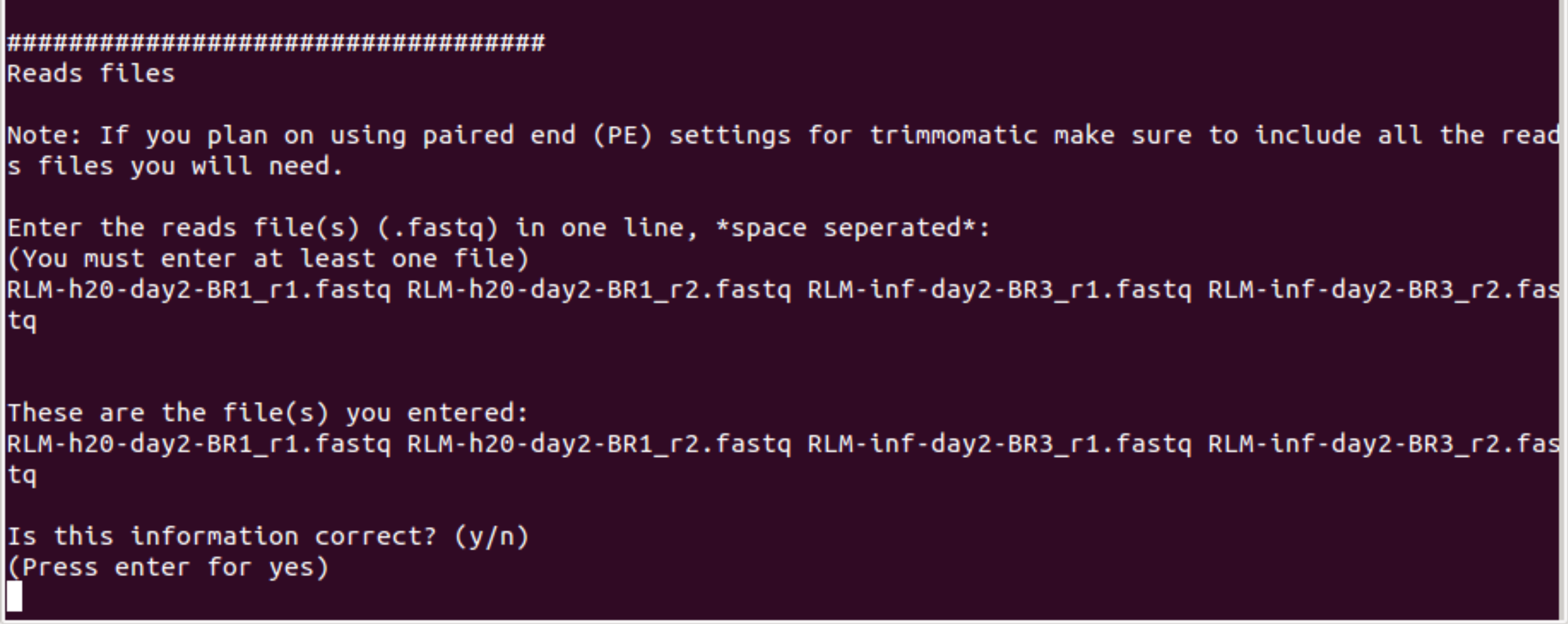
```
Enter the average quality value you want (>= 0) (currently 30):
35
```

Then you will be prompted for the average quality value (reads with lower average quality will be discarded). Type a non-negative integer then press enter.



```
These are the trim settings that will be used:
(non customizable)
trimmomatic-0.33.jar
phred33

(customizable)
SE
TruSeq3-SE.fa (ILLUMINACLIP:adapaters file)(linked to end type)
ILLUMINACLIP:adapters:2:30:10 (:seed_mismathces:palindrome_clip_threshold:simple_clip_threshold)
HEADCROP:10
LEADING:25
TRAILING:25
SLIDINGWINDOW:3:35 (window_size:required_quality)
MINLEN:60
AVGQUAL:35

Is this information correct? (y/n)
(Press enter for yes)
```

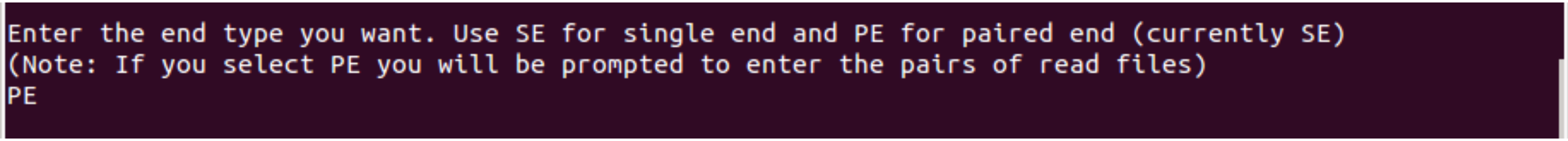
Once you’ve gone through all the settings the program will print all the current settings and there will be a verification step. Enter “n” to re-edit setting or press enter (or enter anything other than “n”) to proceed.



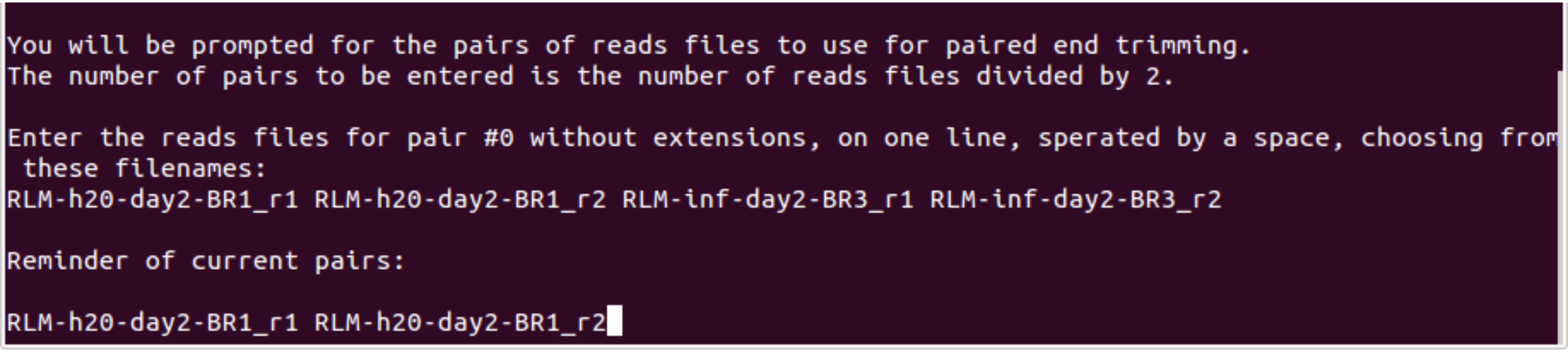
```
##################################
Reads files

Note: If you plan on using paired end (PE) settings for trimmomatic make sure to include all the read
s files you will need.

Enter the reads file(s) (.fastq) in one line, *space seperated*:
(You must enter at least one file)
RLM-h20-day2-BR1_r1.fastq RLM-h20-day2-BR1_r2.fastq RLM-inf-day2-BR3_r1.fastq RLM-inf-day2-BR3_r2.fas
tq


These are the file(s) you entered:
RLM-h20-day2-BR1_r1.fastq RLM-h20-day2-BR1_r2.fastq RLM-inf-day2-BR3_r1.fastq RLM-inf-day2-BR3_r2.fas
tq

Is this information correct? (y/n)
(Press enter for yes)
```

Step 7.5: Paired end settings

If you selected PE in the trim settings you will be prompted to identify reads files pairs from the list of reads files you entered. Watch out for bugs in the case of an uneven number of files. The number of pairs you will be prompted for is the result of doing an integer division by two of the number of reads files you entered.



```
Enter the end type you want. Use SE for single end and PE for paired end (currently SE)
(Note: If you select PE you will be prompted to enter the pairs of read files)
PE
```



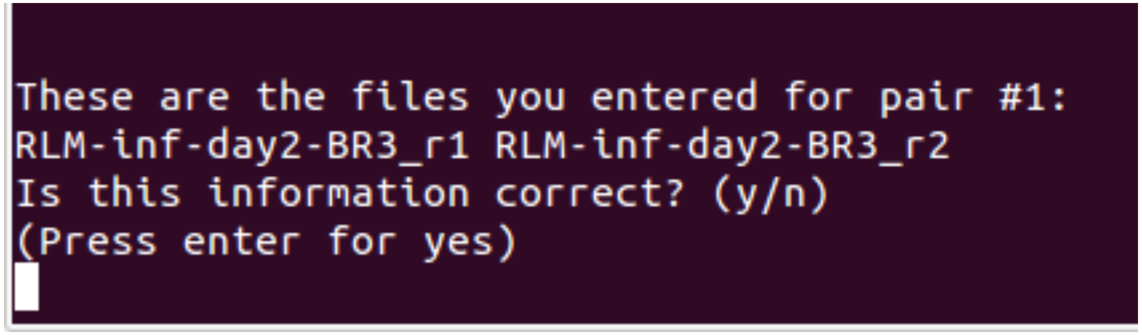
```
You will be prompted for the pairs of reads files to use for paired end trimming.
The number of pairs to be entered is the number of reads files divided by 2.

Enter the reads files for pair #0 without extensions, on one line, sperated by a space, choosing from
 these filenames:
RLM-h20-day2-BR1_r1 RLM-h20-day2-BR1_r2 RLM-inf-day2-BR3_r1 RLM-inf-day2-BR3_r2

Reminder of current pairs:

RLM-h20-day2-BR1_r1 RLM-h20-day2-BR1_r2
```

Once this prompt appears you should type the names of the two files belonging to a pair from the list printed, as shown above, then press enter.

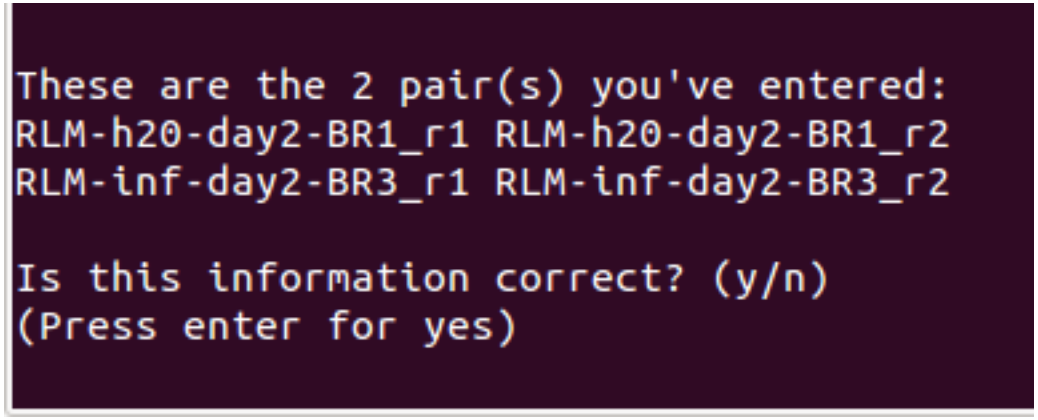
Under “Reminder of current pairs” the program will print previously entered pairs so you know which ones you’ve already input.

Each time you successfully enter a pair (both files are in the list of reads files you entered), there will be a verification step that works the same way as the previous ones (enter “n” for no enter anything else or nothing for yes).

Note that the numbering of pairs starts at #0, so if you had 14 reads files the last pair number you will be prompted for will be pair #6.



```
These are the files you entered for pair #1:
RLM-inf-day2-BR3_r1 RLM-inf-day2-BR3_r2
Is this information correct? (y/n)
(Press enter for yes)
```

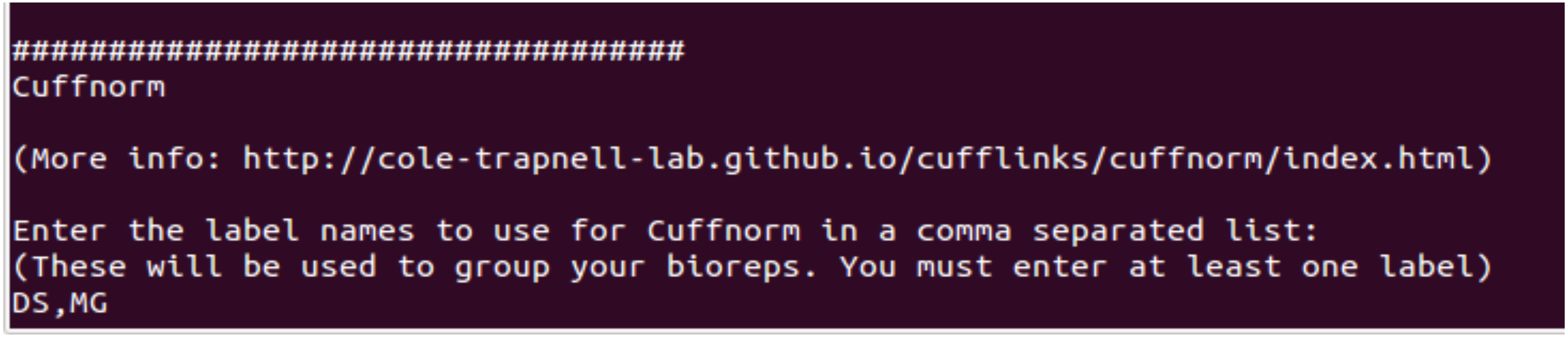


```
These are the 2 pair(s) you've entered:
RLM-h20-day2-BR1_r1 RLM-h20-day2-BR1_r2
RLM-inf-day2-BR3_r1 RLM-inf-day2-BR3_r2

Is this information correct? (y/n)
(Press enter for yes)
```

Once you have entered something valid for each pair and verified each, there will be a final verification where the program will display all the pairs you entered and ask you to check them as shown above. Once more, enter “n” to go back and change the pairs you entered or enter anything else to move on.

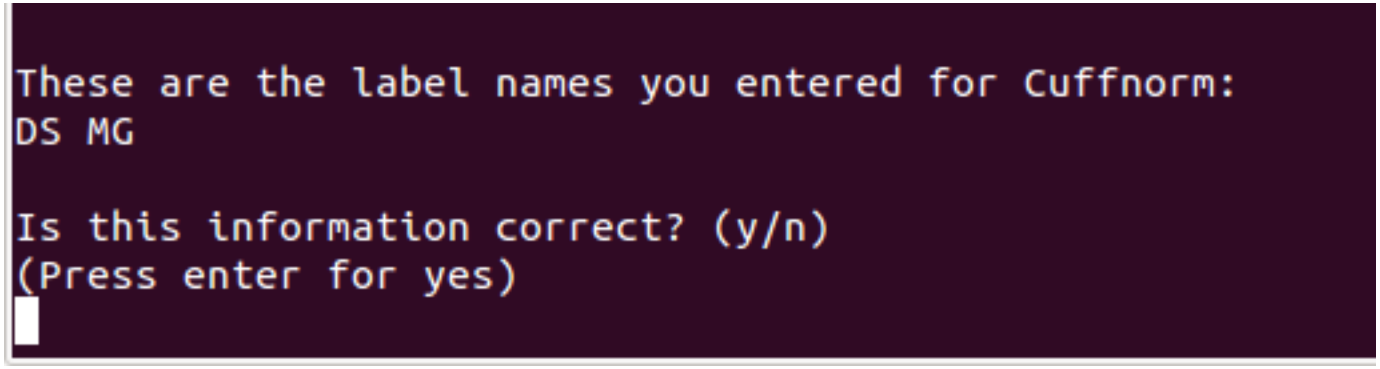
I recommend having the pair filenames written out in a text file and copy-paste them to avoid typos since this final verification would bring you back to the beginning and make you enter all the pairs over again.

Step 8: Cuffnorm labels and files

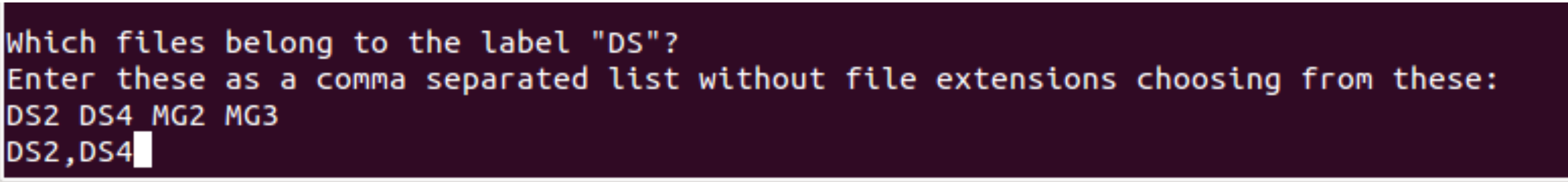


```
###################################
Cuffnorm

(More info: http://cole-trapnell-lab.github.io/cufflinks/cuffnorm/index.html)

Enter the label names to use for Cuffnorm in a comma separated list:
(These will be used to group your bioreps. You must enter at least one label)
DS,MG
```

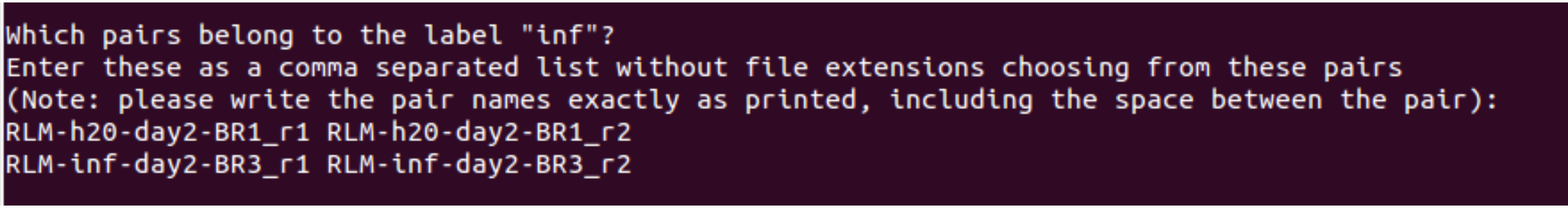
Cuffnorm labels are used to group bioreps/treatments together when normalizing your data. Type out the labels you want to use in a comma separated list then press enter. Once you do there will be a verification step (image below). The labels printed in the verification step will be separated by a space. If you did not type out your labels in a comma separated list, the formatting will not be as expected at this step. Note that you must enter at least one label.



```
These are the label names you entered for Cuffnorm:
DS MG

Is this information correct? (y/n)
(Press enter for yes)
```

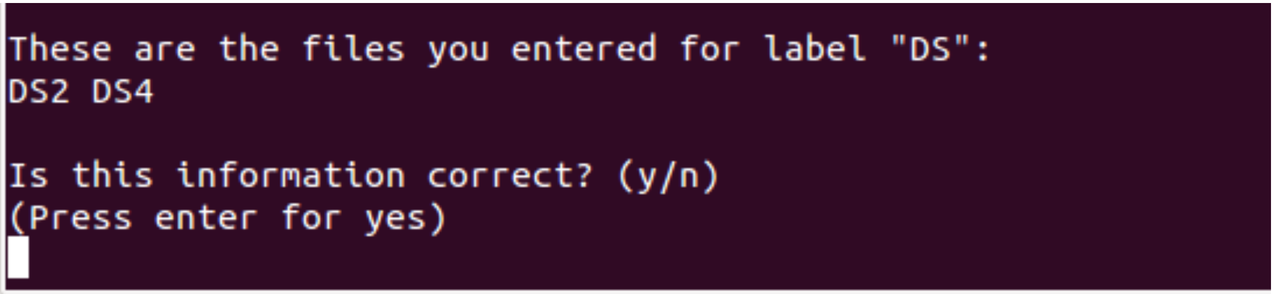


```
Which files belong to the label "DS"?
Enter these as a comma separated list without file extensions choosing from these:
DS2 DS4 MG2 MG3
DS2,DS4
```



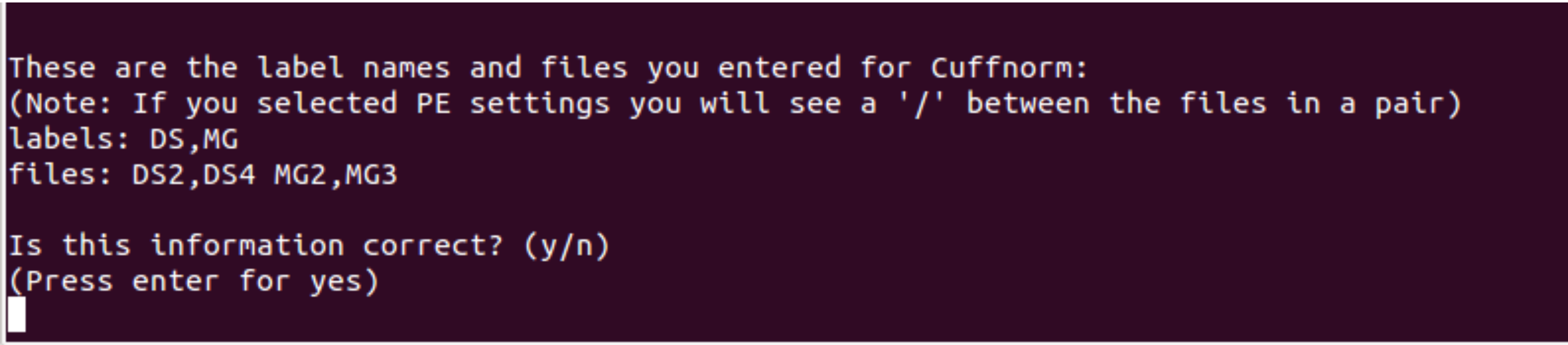
```
Which pairs belong to the label "inf"?
Enter these as a comma separated list without file extensions choosing from these pairs
(Note: please write the pair names exactly as printed, including the space between the pair):
RLM-h20-day2-BR1_r1 RLM-h20-day2-BR1_r2
RLM-inf-day2-BR3_r1 RLM-inf-day2-BR3_r2
```

Once you have verified your labels, you will be prompted to assign files to each label. For each label the program will print a list of the reads files you entered, or the pairs you inputted if you selected PE settings, for you to choose from. Type out the names of the files /pairs belonging to the given label in a comma separated list then press enter. For pairs, include the space between the 2 files in the pair. The files/pairs you choose must be in the printed list. After you enter valid files for a label there will be a verification step, as shown in the image below.



```
These are the files you entered for label "DS":
DS2 DS4

Is this information correct? (y/n)
(Press enter for yes)
```

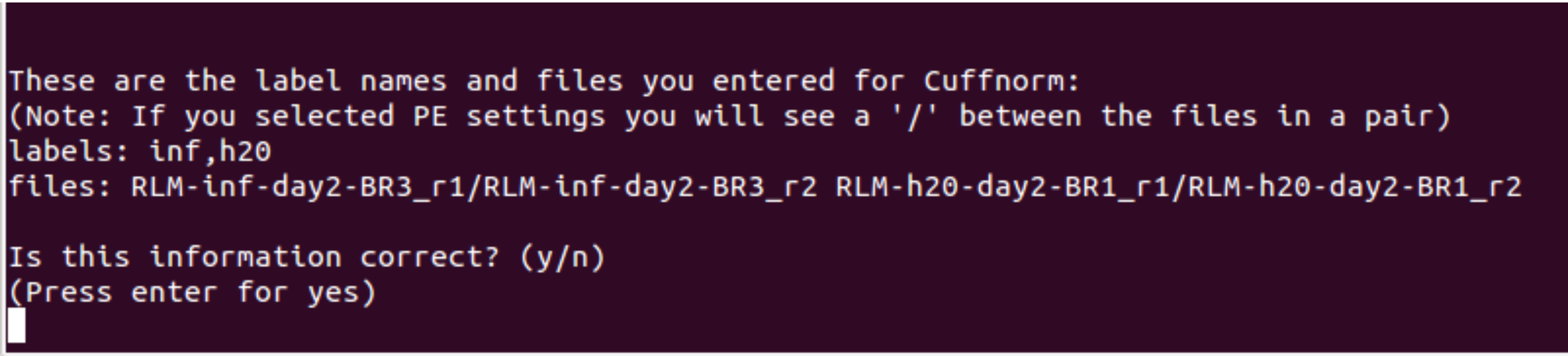
Note that the program does not check for duplicates: if you include the same file/pair in more than one label or more than once in the same label, the program will not catch it.



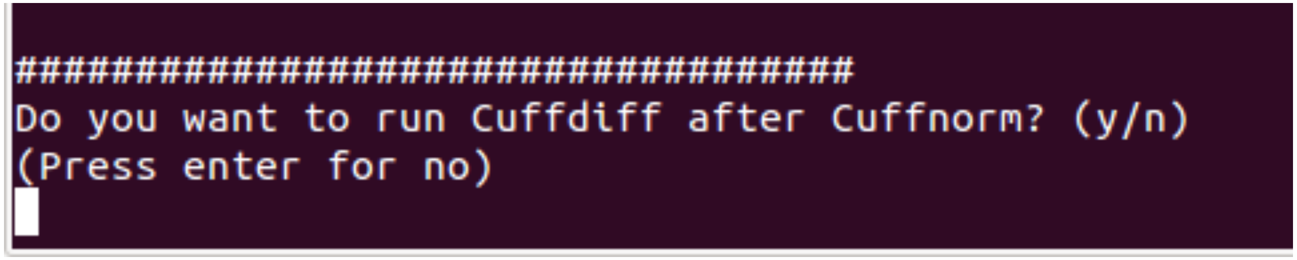
```
These are the label names and files you entered for Cuffnorm:
(Note: If you selected PE settings you will see a '/' between the files in a pair)
labels: DS,MG
files: DS2,DS4 MG2,MG3

Is this information correct? (y/n)
(Press enter for yes)
```

Once you have assigned at least one file/pair to each label and your input was valid, there will be a final verification for labels and files. An SE example is shown above and a PE example is shown below. Notice that the label groups are space separated in both cases, but In SE settings the files are comma separated, while in PE settings, files within pairs are slash separated and pairs within a labels are comma separated. This verification step works the same as the previous ones.

The program does not check to see if you have included all your files/pairs within your labels, so the verification step is important.



```
These are the label names and files you entered for Cuffnorm:
(Note: If you selected PE settings you will see a '/' between the files in a pair)
labels: inf,h20
files: RLM-inf-day2-BR3_r1/RLM-inf-day2-BR3_r2 RLM-h20-day2-BR1_r1/RLM-h20-day2-BR1_r2

Is this information correct? (y/n)
(Press enter for yes)
```

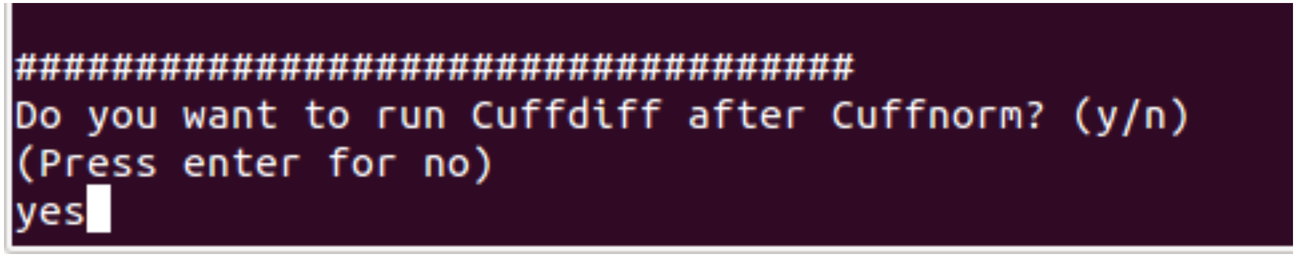


```
###################################
Do you want to run Cuffdiff after Cuffnorm? (y/n)
(Press enter for no)
```

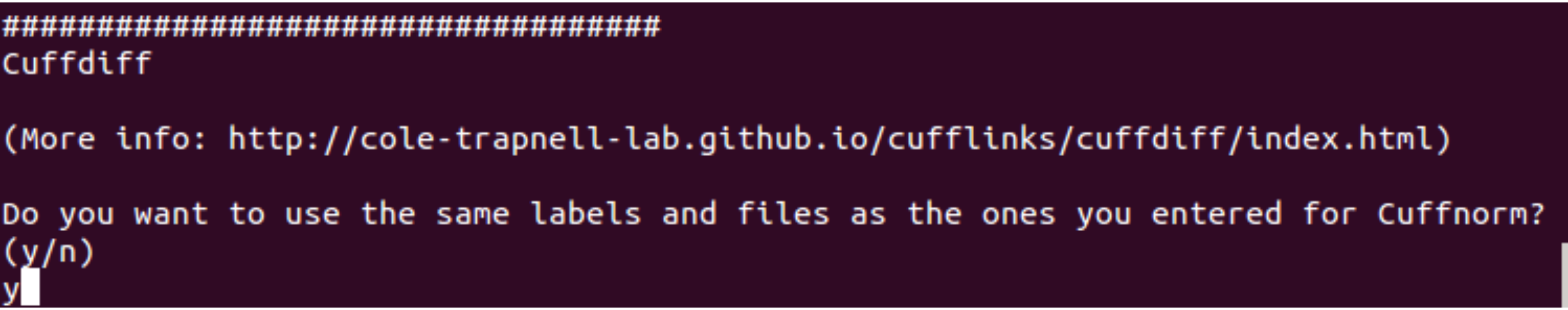
Step 9: Cuffdiff labels and files (optional)

Running Cuffdiff once at the end of the pipeline is optional. If you know which comparison you would like to run or plan on running all the comparisons at once, I recommend getting this program to run/build it for you. The Cuffdiff format is almost identical to the Cuffnorm format and they use the same input files.

If you do want to use Cuffdiff with the pipeline type “yes” or “y” at this prompt to enter the subset of Cuffdiff prompts. Enter anything else or nothing and the program will move on to building and/or running your commands.

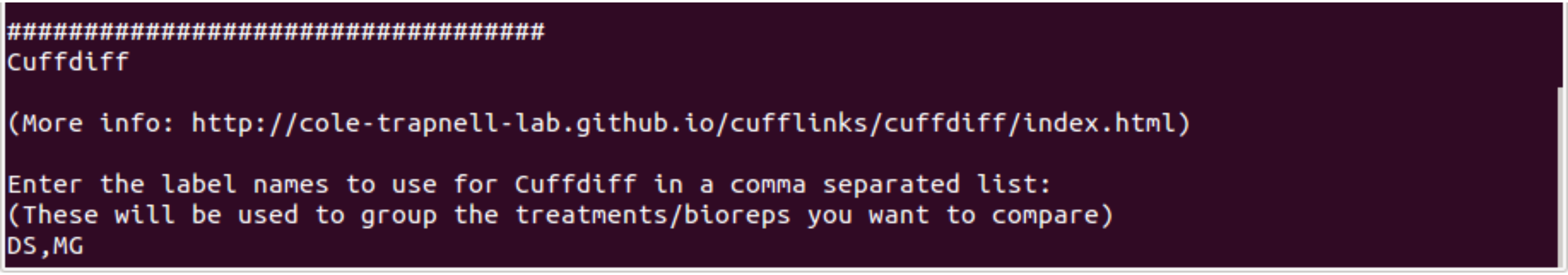


```
###################################
Do you want to run Cuffdiff after Cuffnorm? (y/n)
(Press enter for no)
yes
```



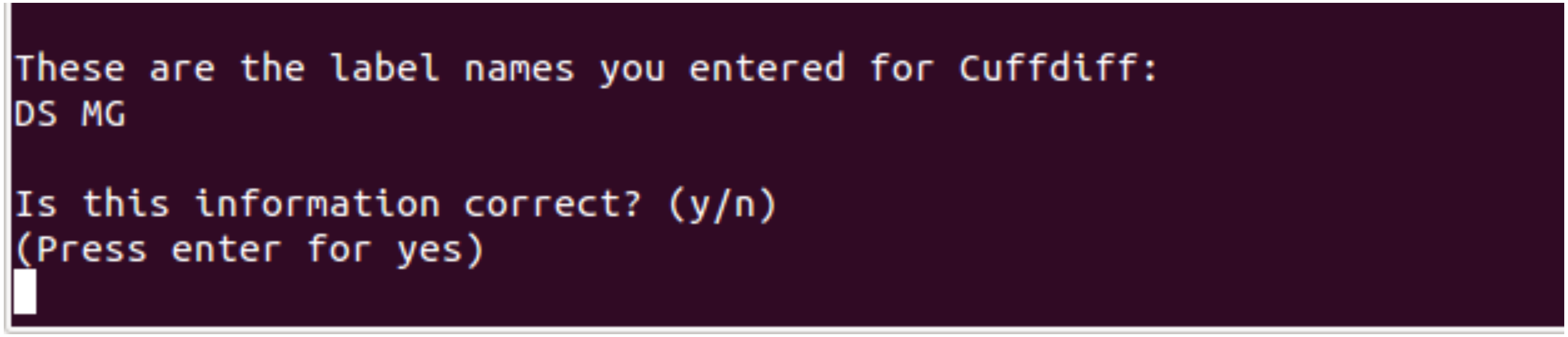
```
###################################
Cuffdiff

(More info: http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html)

Do you want to use the same labels and files as the ones you entered for Cuffnorm?
(y/n)
y
```

If you do choose to run Cuffdiff you will be asked if you want to use the same labels and files you entered for Cuffnorm. This will be particularly useful if you plan on using the cummeRbund R script since it needs CuffDiff output will all your samples (Note: some genome annotations give blank genes files but the isoforms are normal. This is fine for other analyses but cummeRbund will not be agreeable).
This prompt is answered the same as all others and anything other than a ’Y’ or ‘y’ will be interpreted as a ‘no’.

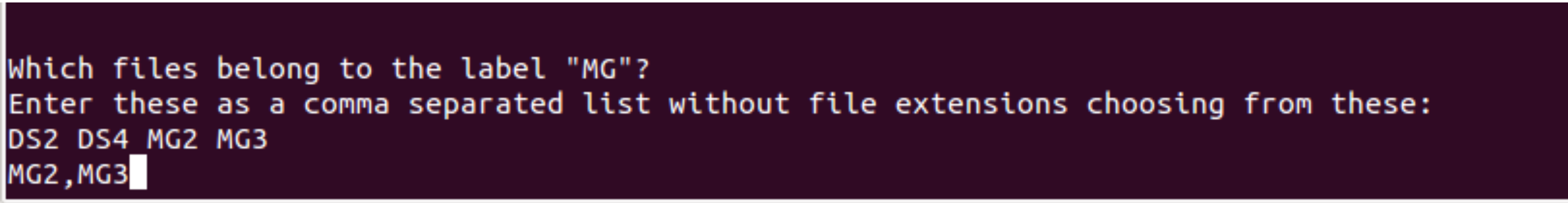


```
###################################
Cuffdiff

(More info: http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html)

Enter the label names to use for Cuffdiff in a comma separated list:
(These will be used to group the treatments/bioreps you want to compare)
DS,MG
```

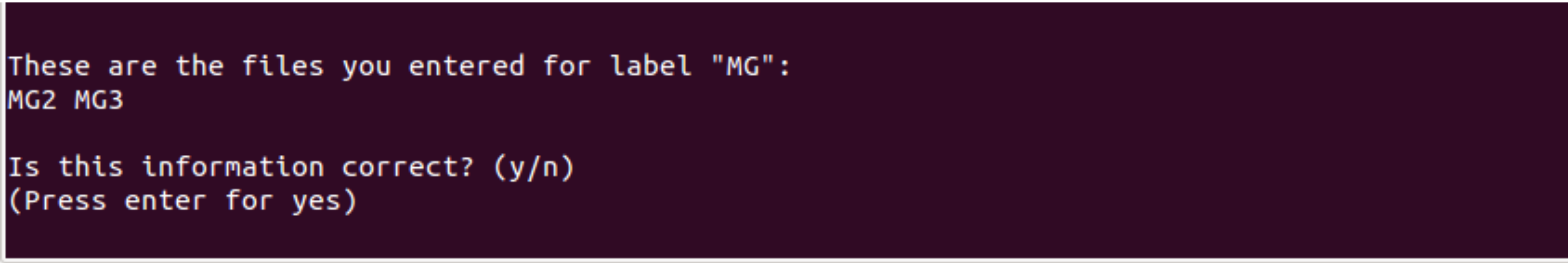
Entering Cuffdiff labels and files works exactly the same way as the entry of Cuffnorm labels and files, so please refer to that section or the images on the next slide if you need more detail. The same input format is expected and there will be verification steps for labels, files belonging to a label, and the final check. Once more, you must enter at least one label, you must enter at least one file/pair per label, the program will not check for duplicates, and it will not check that all the files/pairs have been included.



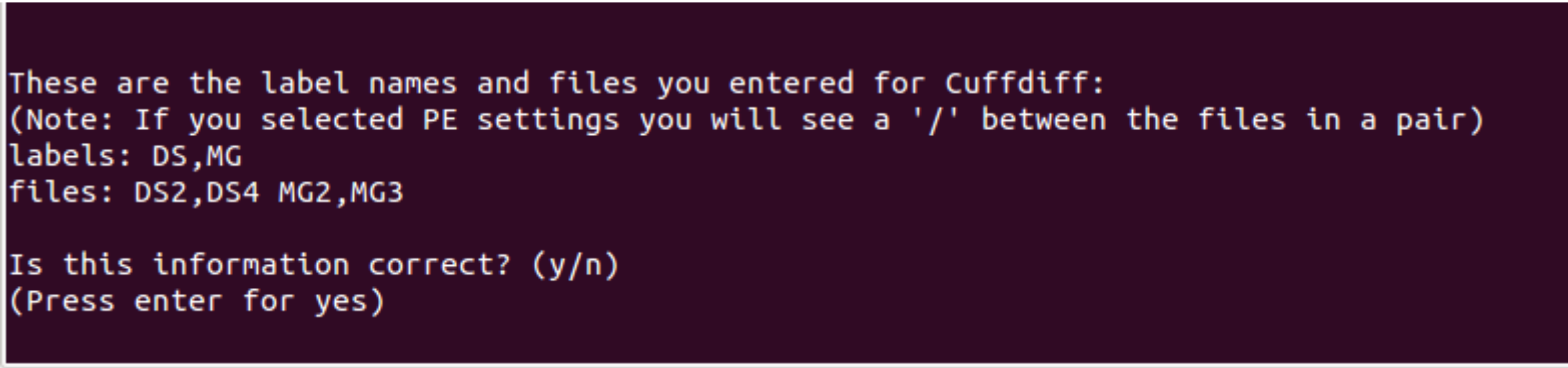
```
These are the label names you entered for Cuffdiff:
DS MG

Is this information correct? (y/n)
(Press enter for yes)
```



```
Which files belong to the label "MG"?
Enter these as a comma separated list without file extensions choosing from these:
DS2 DS4 MG2 MG3
MG2,MG3
```

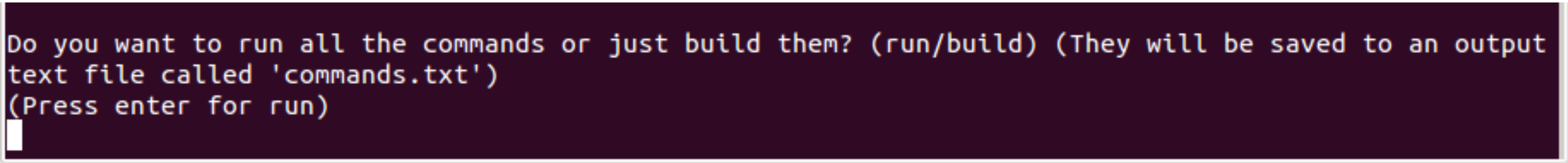


```
These are the files you entered for label "MG":
MG2 MG3

Is this information correct? (y/n)
(Press enter for yes)
```



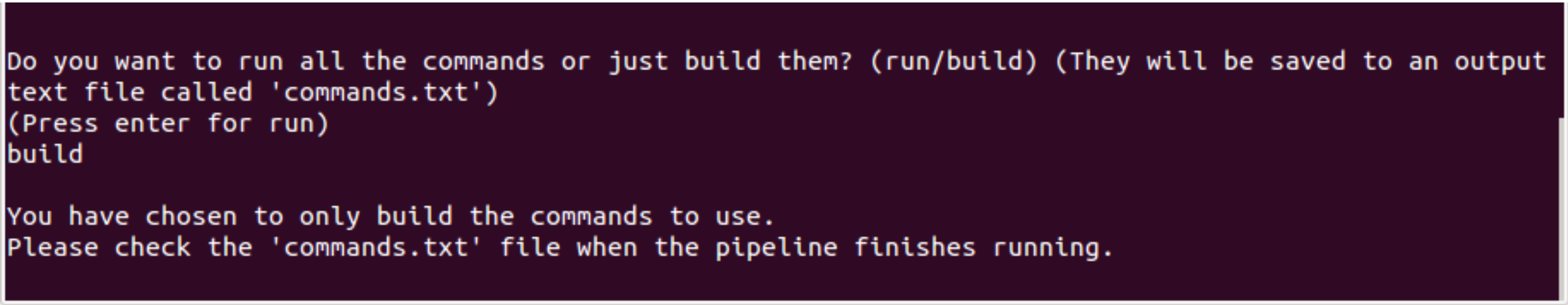
```
These are the label names and files you entered for Cuffdiff:
(Note: If you selected PE settings you will see a '/' between the files in a pair)
labels: DS,MG
files: DS2,DS4 MG2,MG3

Is this information correct? (y/n)
(Press enter for yes)
```

Step 10: Running or building



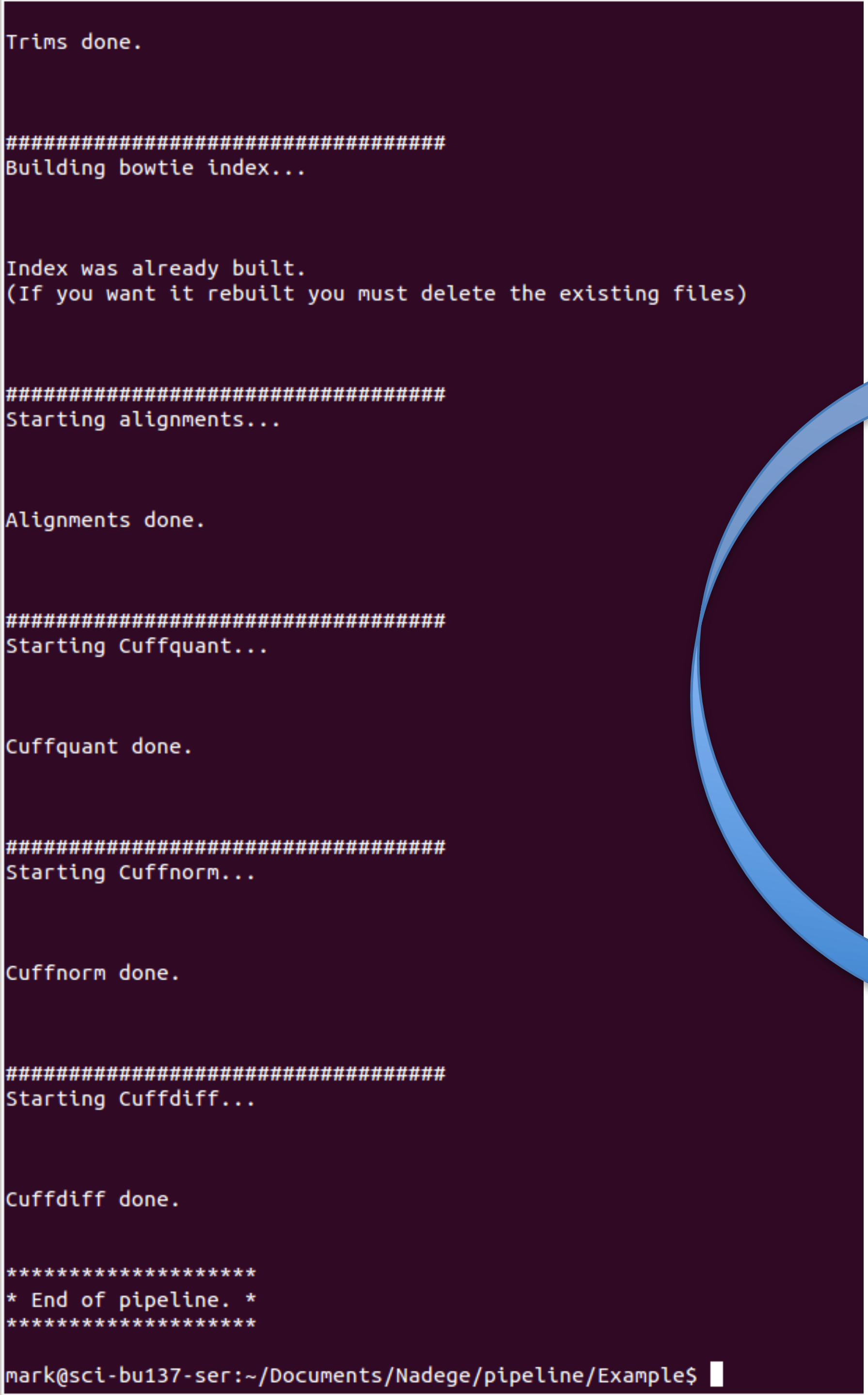
```
Do you want to run all the commands or just build them? (run/build) (They will be saved to an output
text file called 'commands.txt')
(Press enter for run)
```

Once you’ve finished entering all the files, settings, and labels, the program will prompt you to ask if the commands should be run or built. If you choose “run” the program will still build and save the commands in “commands.txt” but these will be run as they’re being built. If you choose “build” the program will only build “commands.txt”.



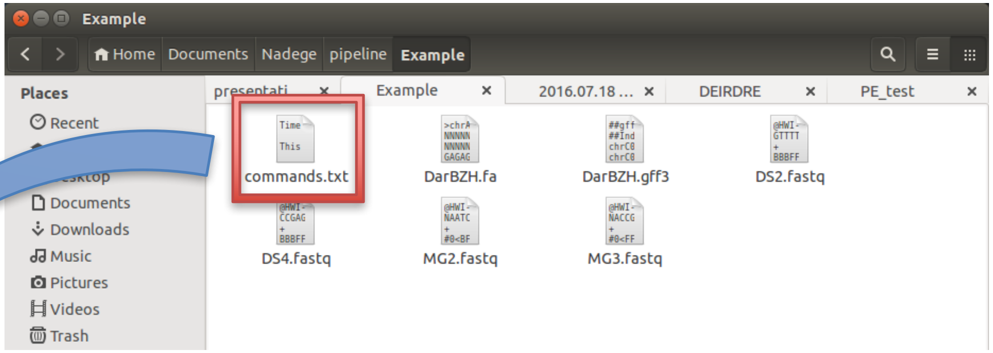
```
Do you want to run all the commands or just build them? (run/build) (They will be saved to an output
text file called 'commands.txt')
(Press enter for run)
build

You have chosen to only build the commands to use.
Please check the 'commands.txt' file when the pipeline finishes running.
```

If you want to only build the commands type “b” or “build” then press enter. If you enter anything else or nothing the program will assume you want to run the commands. If you choose build, there will be a message acknowledging this and a delay so you can read this message.

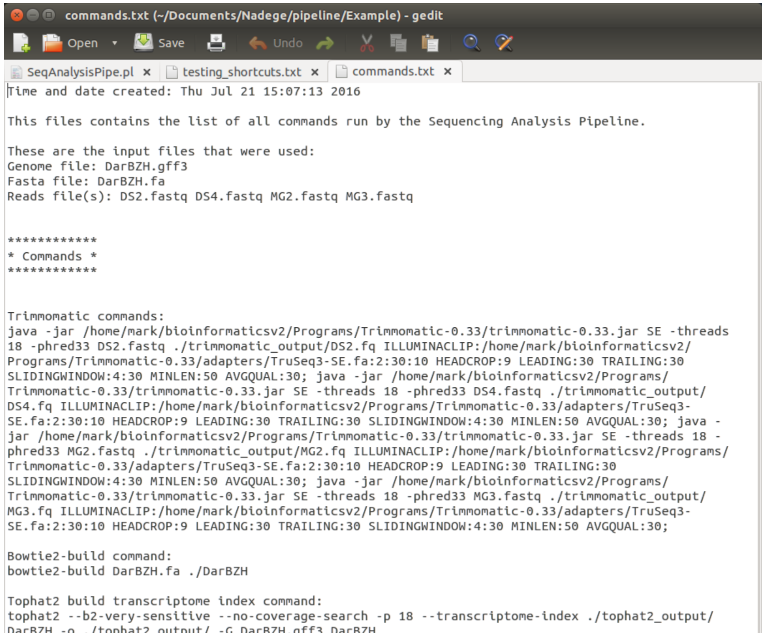


```
Trims done.


###################################
Building bowtie index...


Index was already built.
(If you want it rebuilt you must delete the existing files)


###################################
Starting alignments...


Alignments done.


###################################
Starting Cuffquant...


Cuffquant done.


###################################
Starting Cuffnorm...


Cuffnorm done.


###################################
Starting Cuffdiff...


Cuffdiff done.


********************
* End of pipeline. *
********************

mark@sci-bu137-ser:~/Documents/Nadege/pipeline/Example$ 
```

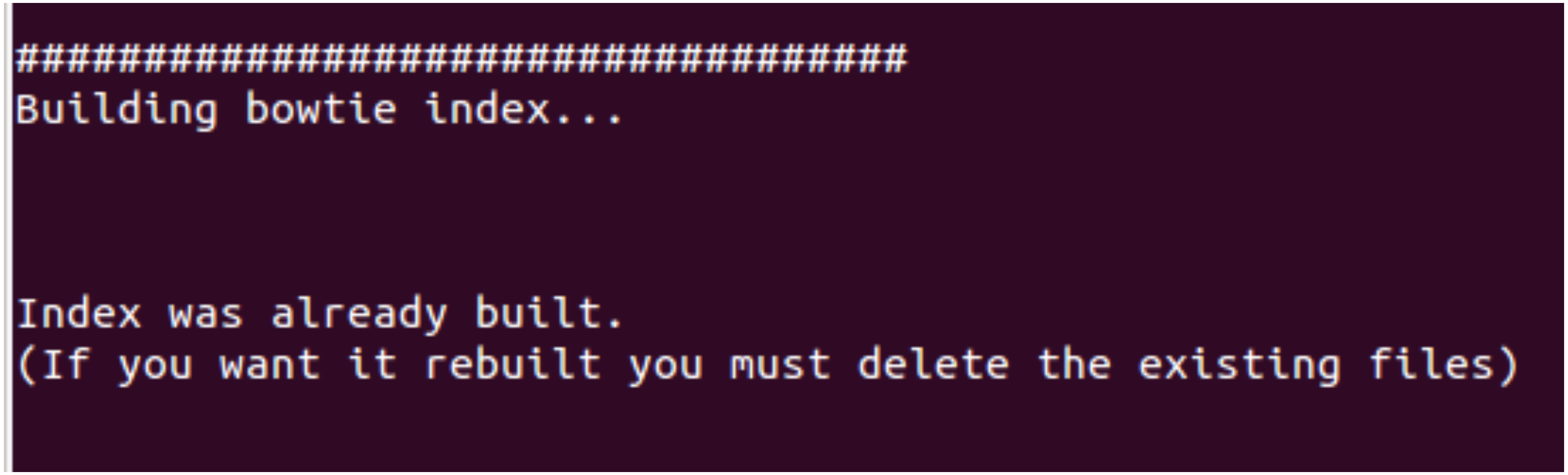
This is what choosing build looks like after the delay:



Notice the “commands.txt” file that was created in the folder containing your reads, genome, and fasta files.

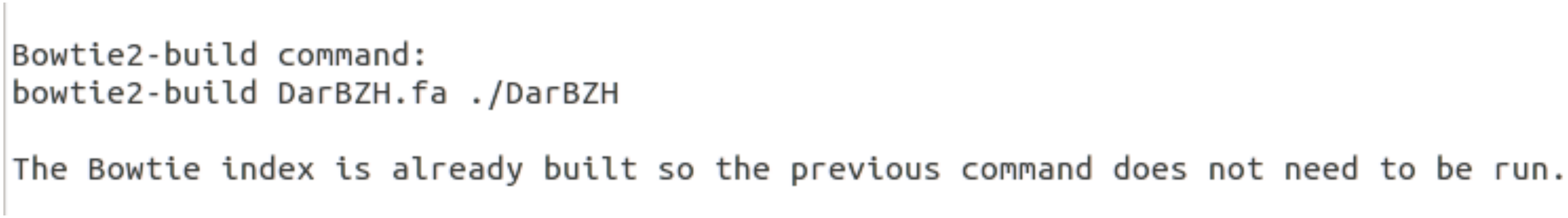


```
Time and date created: Thu Jul 21 15:07:13 2016

This files contains the list of all commands run by the Sequencing Analysis Pipeline.

These are the input files that were used:
Genome file: DarBZH.gff3
Fasta file: DarBZH.fa
Reads file(s): DS2.fastq DS4.fastq MG2.fastq MG3.fastq


************
* Commands *
************


Trimmomatic commands:
java -jar /home/mark/bioinformaticsv2/Programs/Trimmomatic-0.33/trimmomatic-0.33.jar SE -threads
18 -phred33 DS2.fastq ./trimmomatic_output/DS2.fq ILLUMINACLIP:/home/mark/bioinformaticsv2/
Programs/Trimmomatic-0.33/adapters/TruSeq3-SE.fa:2:30:10 HEADCROP:9 LEADING:30 TRAILING:30
SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:30; java -jar /home/mark/bioinformaticsv2/Programs/
Trimmomatic-0.33/trimmomatic-0.33.jar SE -threads 18 -phred33 DS4.fastq ./trimmomatic_output/
DS4.fq ILLUMINACLIP:/home/mark/bioinformaticsv2/Programs/Trimmomatic-0.33/adapters/TruSeq3-
SE.fa:2:30:10 HEADCROP:9 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:30; java -
jar /home/mark/bioinformaticsv2/Programs/Trimmomatic-0.33/trimmomatic-0.33.jar SE -threads 18 -
phred33 MG2.fastq ./trimmomatic_output/MG2.fq ILLUMINACLIP:/home/mark/bioinformaticsv2/Programs/
Trimmomatic-0.33/adapters/TruSeq3-SE.fa:2:30:10 HEADCROP:9 LEADING:30 TRAILING:30
SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:30; java -jar /home/mark/bioinformaticsv2/Programs/
Trimmomatic-0.33/trimmomatic-0.33.jar SE -threads 18 -phred33 MG3.fastq ./trimmomatic_output/
MG3.fq ILLUMINACLIP:/home/mark/bioinformaticsv2/Programs/Trimmomatic-0.33/adapters/TruSeq3-
SE.fa:2:30:10 HEADCROP:9 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:30;

Bowtie2-build command:
bowtie2-build DarBZH.fa ./DarBZH

Tophat2 build transcriptome index command:
tophat2 --b2-very-sensitive --no-coverage-search -p 18 --transcriptome-index ./tophat2_output/
DarBZH -o ./tophat2_output/ -G DarBZH.gff3 DarBZH
```

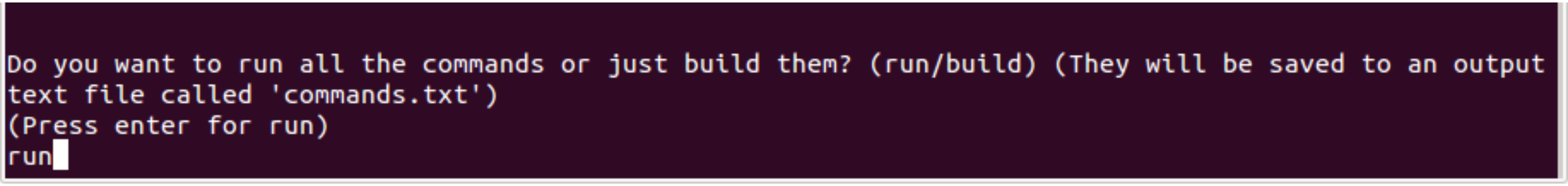


```
###################################
Building bowtie index...



Index was already built.
(If you want it rebuilt you must delete the existing files)
```

If the bowtie index is already built the program will inform you in the terminal/command line (see above). The command to build the index will still be built and saved to “commands.txt” but a note will be added saying it does not need to be run (see below).



```
Bowtie2-build command:
bowtie2-build DarBZH.fa ./DarBZH

The Bowtie index is already built so the previous command does not need to be run.
```

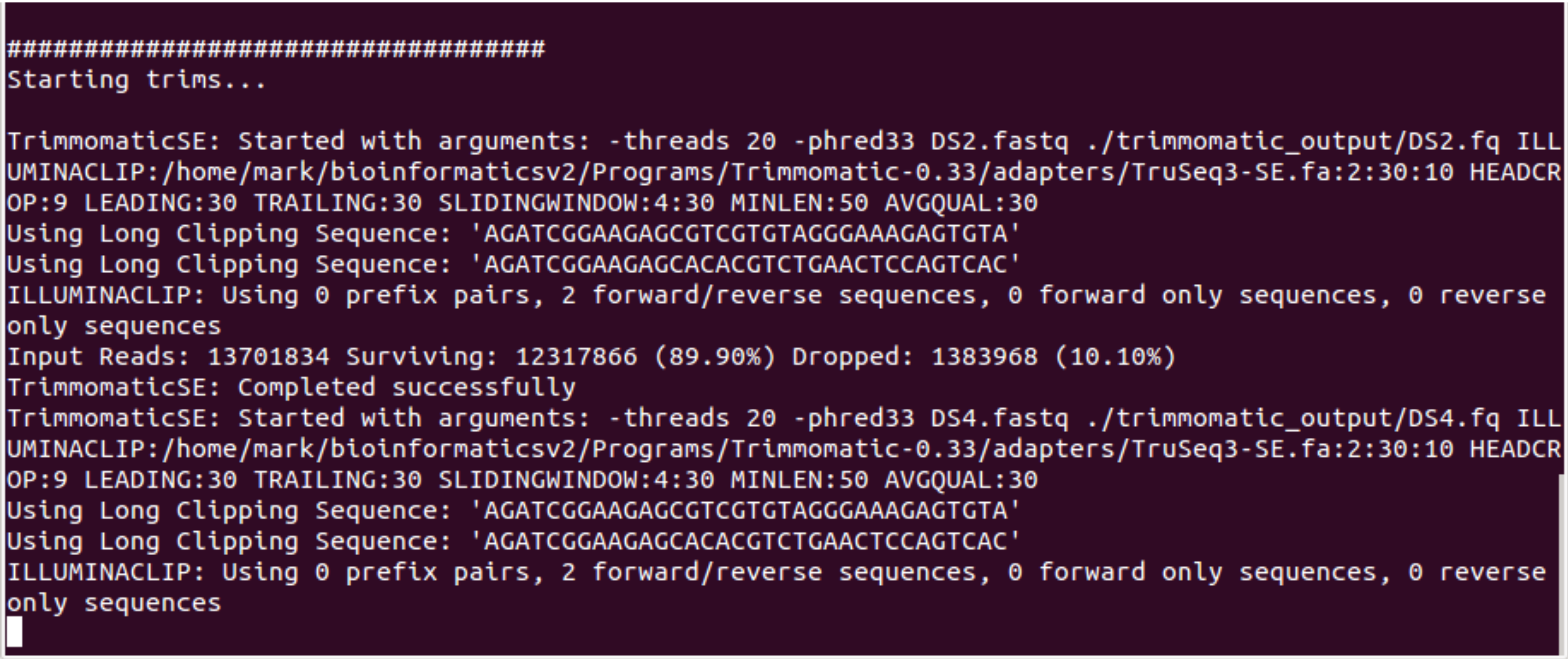


```
Do you want to run all the commands or just build them? (run/build) (They will be saved to an output
text file called 'commands.txt')
(Press enter for run)
run
```

If you chose to run the commands the program will print each program’s normal progress output as shown in the next few slides. Running all the commands will take a while, even more so with large and/or many reads files, so just leave it running until it’s done. You will known the program has finished running when the “End of pipeline” message has been printed.

Important: Do not rename or modify any of the files and folders until the “End of pipeline” message has been printed. If you do, the programs in the pipeline might not be able to find their input/output files or folders which would likely cause an error and interrupt the rest of the pipeline. If you need to look at the output of a specific step in the pipeline, I recommend waiting until the end of that step and then, only viewing the file contents.

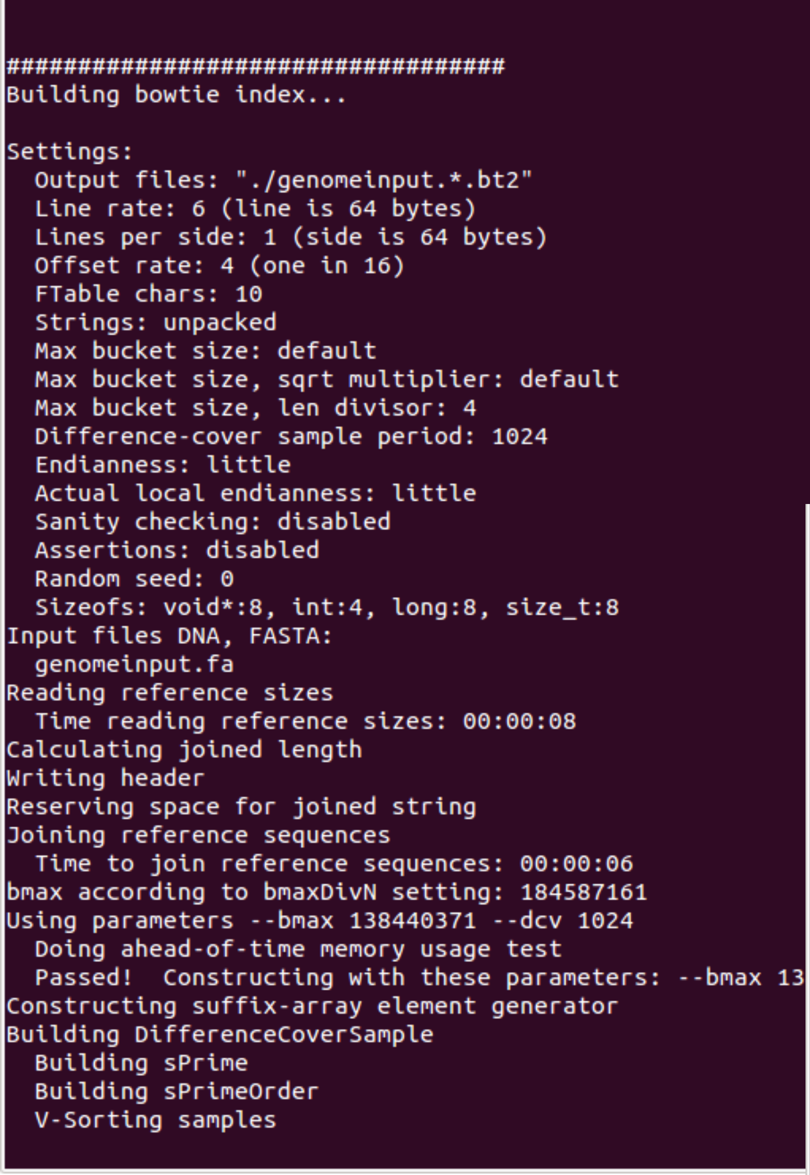
This is what running Trimmomatic from the pipeline looks like:



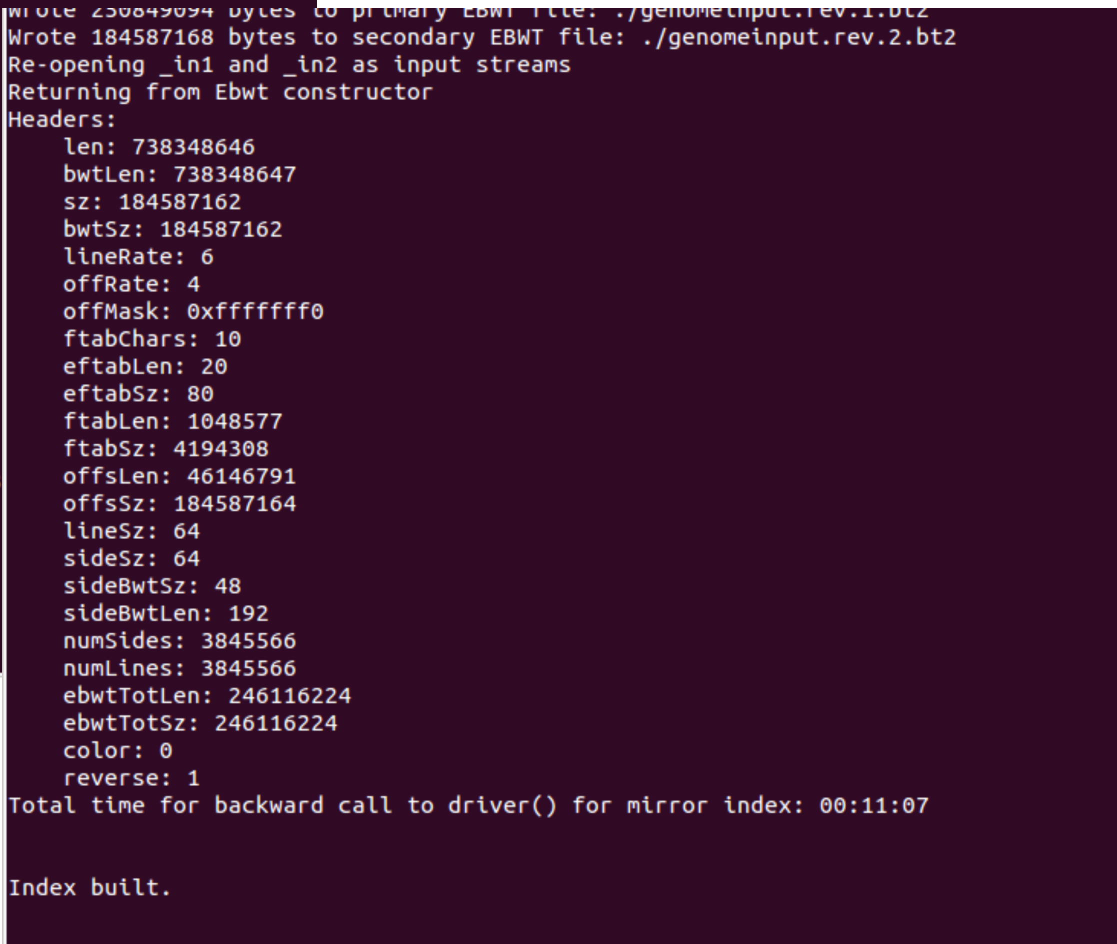
```
###################################
Starting trims...

TrimmomaticSE: Started with arguments: -threads 20 -phred33 DS2.fastq ./trimmomatic_output/DS2.fq ILL
UMINACLIP:/home/mark/bioinformaticsv2/Programs/Trimmomatic-0.33/adapters/TruSeq3-SE.fa:2:30:10 HEADCR
OP:9 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:30
Using Long Clipping Sequence: 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA'
Using Long Clipping Sequence: 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC'
ILLUMINACLIP: Using 0 prefix pairs, 2 forward/reverse sequences, 0 forward only sequences, 0 reverse
only sequences
Input Reads: 13701834 Surviving: 12317866 (89.90%) Dropped: 1383968 (10.10%)
TrimmomaticSE: Completed successfully
TrimmomaticSE: Started with arguments: -threads 20 -phred33 DS4.fastq ./trimmomatic_output/DS4.fq ILL
UMINACLIP:/home/mark/bioinformaticsv2/Programs/Trimmomatic-0.33/adapters/TruSeq3-SE.fa:2:30:10 HEADCR
OP:9 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:30
Using Long Clipping Sequence: 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA'
Using Long Clipping Sequence: 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC'
ILLUMINACLIP: Using 0 prefix pairs, 2 forward/reverse sequences, 0 forward only sequences, 0 reverse
only sequences
```

This is what building the bowtie index looks like:

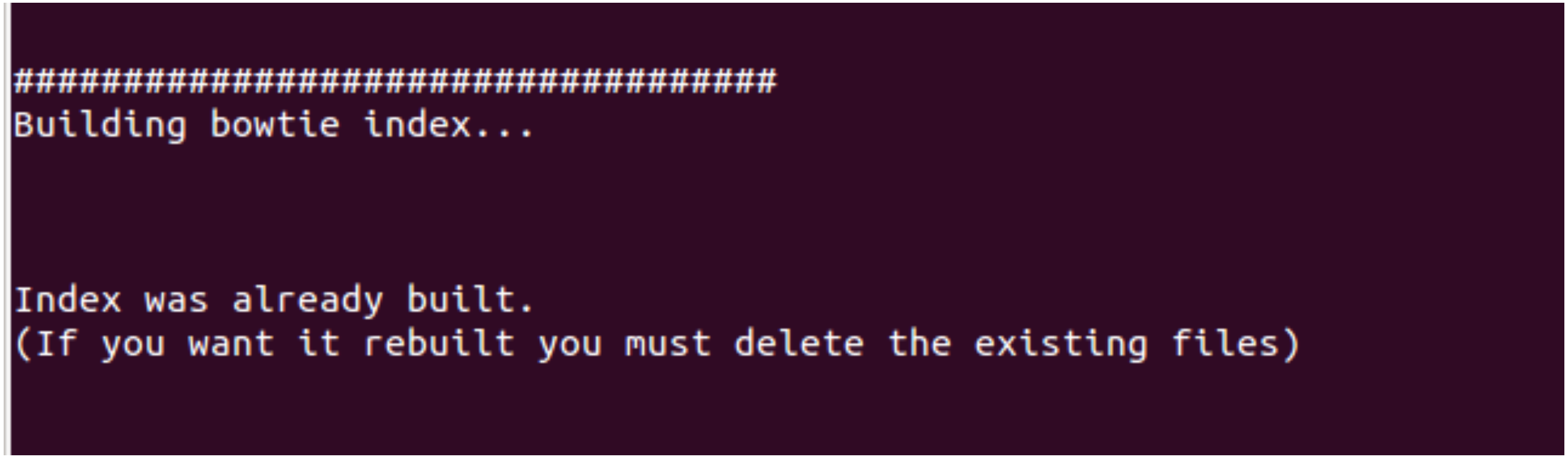
This is a long process which is why there is a check in place to skip this step if it is already built.



```
###################################
Building bowtie index...

Settings:
  Output files: "./genomeinput.*.bt2"
  Line rate: 6 (line is 64 bytes)
  Lines per side: 1 (side is 64 bytes)
  Offset rate: 4 (one in 16)
  FTable chars: 10
  Strings: unpacked
  Max bucket size: default
  Max bucket size, sqrt multiplier: default
  Max bucket size, len divisor: 4
  Difference-cover sample period: 1024
  Endianness: little
  Actual local endianness: little
  Sanity checking: disabled
  Assertions: disabled
  Random seed: 0
  Sizeofs: void*:8, int:4, long:8, size_t:8
Input files DNA, FASTA:
  genomeinput.fa
Reading reference sizes
  Time reading reference sizes: 00:00:08
Calculating joined length
Writing header
Reserving space for joined string
Joining reference sequences
  Time to join reference sequences: 00:00:06
bmax according to bmaxDivN setting: 184587161
Using parameters --bmax 138440371 --dcv 1024
  Doing ahead-of-time memory usage test
  Passed!  Constructing with these parameters: --bmax 13
Constructing suffix-array element generator
Building DifferenceCoverSample
  Building sPrime
  Building sPrimeOrder
  V-Sorting samples
```



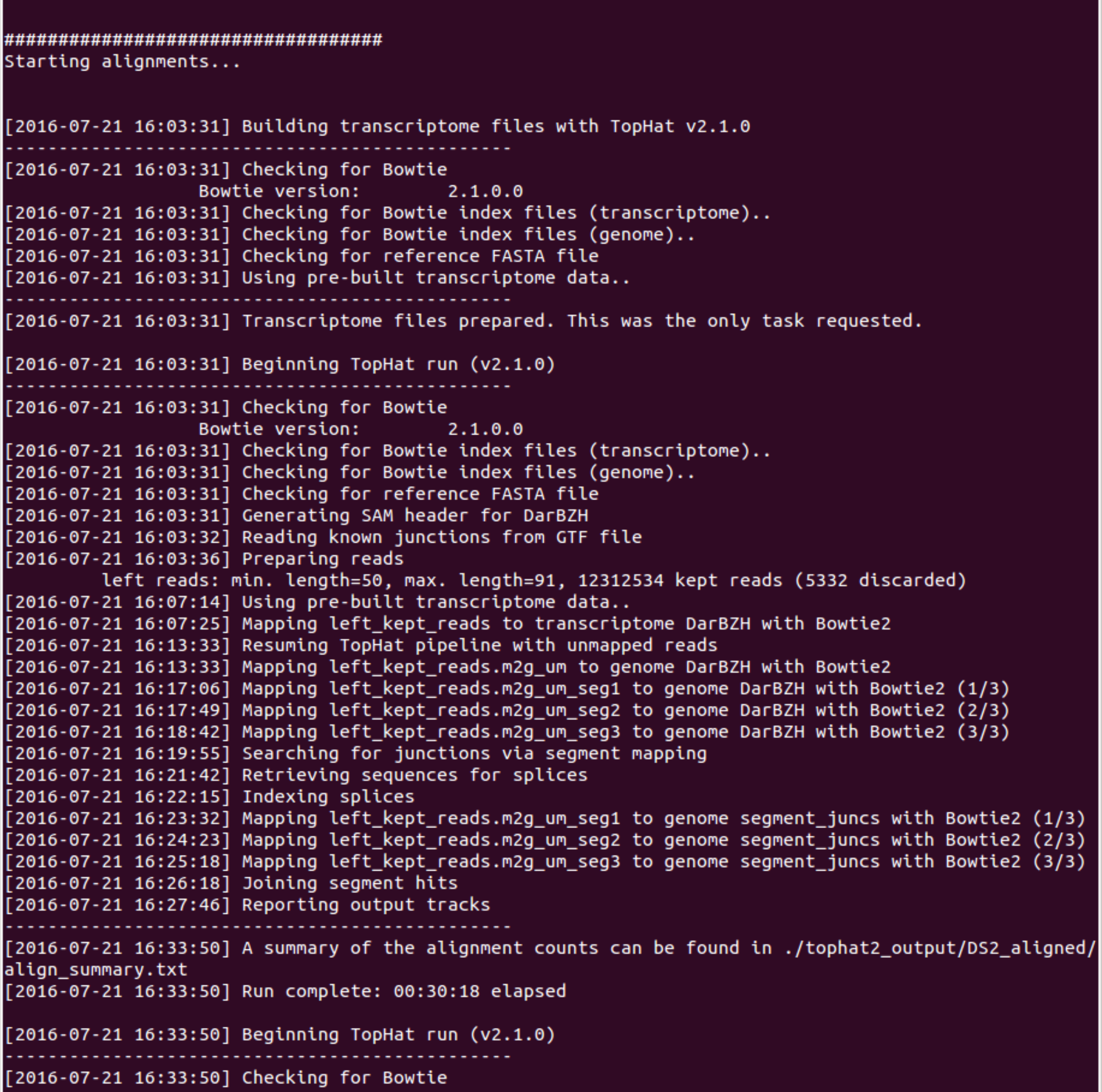
```
Wrote 236849094 bytes to primary EBWT file: ./genomeinput.rev.1.bt2
Wrote 184587168 bytes to secondary EBWT file: ./genomeinput.rev.2.bt2
Re-opening _in1 and _in2 as input streams
Returning from Ebwt constructor
Headers:
    len: 738348646
    bwtLen: 738348647
    sz: 184587162
    bwtSz: 184587162
    lineRate: 6
    offRate: 4
    offMask: 0xfffffff0
    ftabChars: 10
    eftabLen: 20
    eftabSz: 80
    ftabLen: 1048577
    ftabSz: 4194308
    offsLen: 46146791
    offsSz: 184587164
    lineSz: 64
    sideSz: 64
    sideBwtSz: 48
    sideBwtLen: 192
    numSides: 3845566
    numLines: 3845566
    ebwtTotLen: 246116224
    ebwtTotSz: 246116224
    color: 0
    reverse: 1
Total time for backward call to driver() for mirror index: 00:11:07


Index built.
```



```
###################################
Building bowtie index...



Index was already built.
(If you want it rebuilt you must delete the existing files)
```

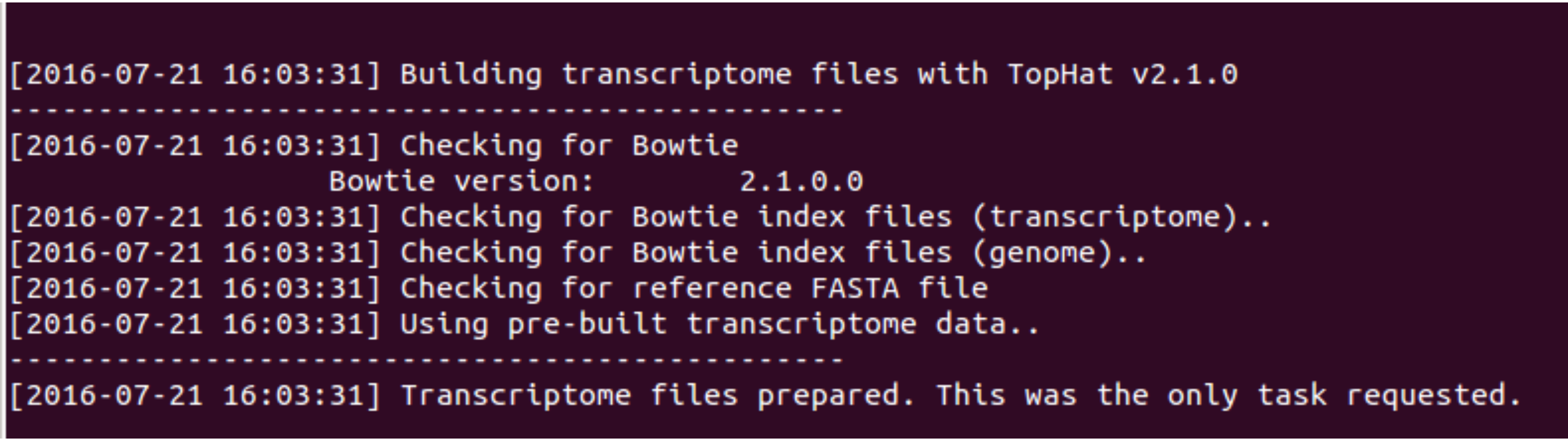
If the pipeline finds index files with the same names as the ones it would have built, it will skip building the bowtie index to save time and the message above will be displayed. As printed above, if you want the index to be built again, you will have to move or delete the existing files so that they are not in the directory where your reads, genome, and fasta files are located. The bowtie index is composed of 6 files prefixed with the same name as your genome and fasta files and ending in .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, and .rev.2.bt2. If your fasta file was exceptionally large these extensions will end in bt2l (lowercase ‘L’).

This is what running Tophat2 from the pipeline looks like:



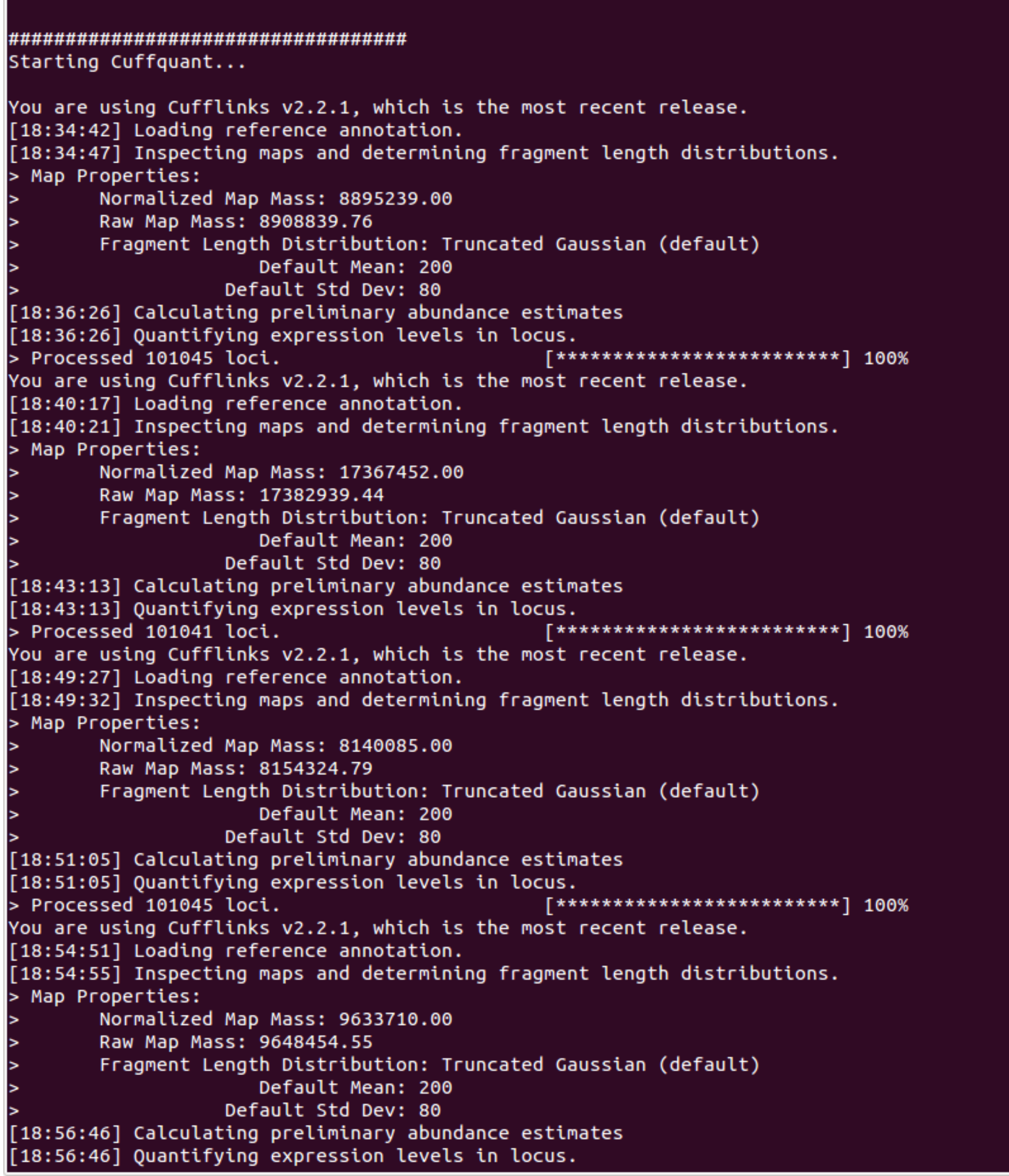
```
###################################
Starting alignments...


[2016-07-21 16:03:31] Building transcriptome files with TopHat v2.1.0
-----------------------------------------------
[2016-07-21 16:03:31] Checking for Bowtie
                  Bowtie version:        2.1.0.0
[2016-07-21 16:03:31] Checking for Bowtie index files (transcriptome)..
[2016-07-21 16:03:31] Checking for Bowtie index files (genome)..
[2016-07-21 16:03:31] Checking for reference FASTA file
[2016-07-21 16:03:31] Using pre-built transcriptome data..
-----------------------------------------------
[2016-07-21 16:03:31] Transcriptome files prepared. This was the only task requested.

[2016-07-21 16:03:31] Beginning TopHat run (v2.1.0)
-----------------------------------------------
[2016-07-21 16:03:31] Checking for Bowtie
                  Bowtie version:        2.1.0.0
[2016-07-21 16:03:31] Checking for Bowtie index files (transcriptome)..
[2016-07-21 16:03:31] Checking for Bowtie index files (genome)..
[2016-07-21 16:03:31] Checking for reference FASTA file
[2016-07-21 16:03:31] Generating SAM header for DarBZH
[2016-07-21 16:03:32] Reading known junctions from GTF file
[2016-07-21 16:03:36] Preparing reads
         left reads: min. length=50, max. length=91, 12312534 kept reads (5332 discarded)
[2016-07-21 16:07:14] Using pre-built transcriptome data..
[2016-07-21 16:07:25] Mapping left_kept_reads to transcriptome DarBZH with Bowtie2
[2016-07-21 16:13:33] Resuming TopHat pipeline with unmapped reads
[2016-07-21 16:13:33] Mapping left_kept_reads.m2g_um to genome DarBZH with Bowtie2
[2016-07-21 16:17:06] Mapping left_kept_reads.m2g_um_seg1 to genome DarBZH with Bowtie2 (1/3)
[2016-07-21 16:17:49] Mapping left_kept_reads.m2g_um_seg2 to genome DarBZH with Bowtie2 (2/3)
[2016-07-21 16:18:42] Mapping left_kept_reads.m2g_um_seg3 to genome DarBZH with Bowtie2 (3/3)
[2016-07-21 16:19:55] Searching for junctions via segment mapping
[2016-07-21 16:21:42] Retrieving sequences for splices
[2016-07-21 16:22:15] Indexing splices
[2016-07-21 16:23:32] Mapping left_kept_reads.m2g_um_seg1 to genome segment_juncs with Bowtie2 (1/3)
[2016-07-21 16:24:23] Mapping left_kept_reads.m2g_um_seg2 to genome segment_juncs with Bowtie2 (2/3)
[2016-07-21 16:25:18] Mapping left_kept_reads.m2g_um_seg3 to genome segment_juncs with Bowtie2 (3/3)
[2016-07-21 16:26:18] Joining segment hits
[2016-07-21 16:27:46] Reporting output tracks
-----------------------------------------------
[2016-07-21 16:33:50] A summary of the alignment counts can be found in ./tophat2_output/DS2_aligned/
align_summary.txt
[2016-07-21 16:33:50] Run complete: 00:30:18 elapsed

[2016-07-21 16:33:50] Beginning TopHat run (v2.1.0)
-----------------------------------------------
[2016-07-21 16:33:50] Checking for Bowtie
```

The first run of Tophat2 within the alignment section will look different from the others because it is being run without trimmed reads to build a transcriptome index. All the other Tophat2 runs will use this same transcriptome index without having to prepare it.



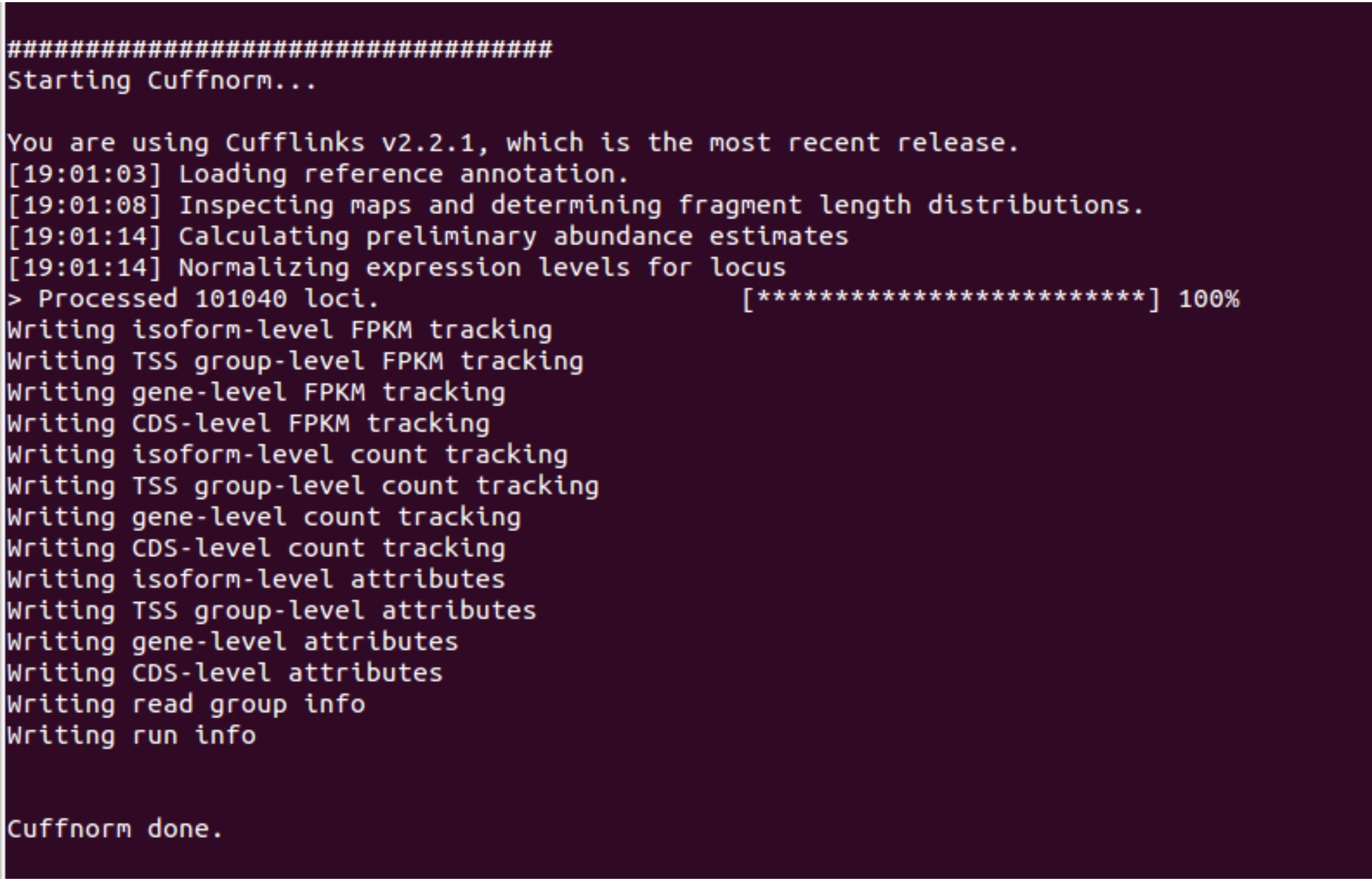
```
[2016-07-21 16:03:31] Building transcriptome files with TopHat v2.1.0
-----------------------------------------
[2016-07-21 16:03:31] Checking for Bowtie
                  Bowtie version:        2.1.0.0
[2016-07-21 16:03:31] Checking for Bowtie index files (transcriptome)..
[2016-07-21 16:03:31] Checking for Bowtie index files (genome)..
[2016-07-21 16:03:31] Checking for reference FASTA file
[2016-07-21 16:03:31] Using pre-built transcriptome data..
-----------------------------------------
[2016-07-21 16:03:31] Transcriptome files prepared. This was the only task requested.
```

This is what running Cuffquant from the pipeline looks like:



```
###################################
Starting Cuffquant...

You are using Cufflinks v2.2.1, which is the most recent release.
[18:34:42] Loading reference annotation.
[18:34:47] Inspecting maps and determining fragment length distributions.
> Map Properties:
>       Normalized Map Mass: 8895239.00
>       Raw Map Mass: 8908839.76
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[18:36:26] Calculating preliminary abundance estimates
[18:36:26] Quantifying expression levels in locus.
> Processed 101045 loci.                       [*************************] 100%
You are using Cufflinks v2.2.1, which is the most recent release.
[18:40:17] Loading reference annotation.
[18:40:21] Inspecting maps and determining fragment length distributions.
> Map Properties:
>       Normalized Map Mass: 17367452.00
>       Raw Map Mass: 17382939.44
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[18:43:13] Calculating preliminary abundance estimates
[18:43:13] Quantifying expression levels in locus.
> Processed 101041 loci.                       [*************************] 100%
You are using Cufflinks v2.2.1, which is the most recent release.
[18:49:27] Loading reference annotation.
[18:49:32] Inspecting maps and determining fragment length distributions.
> Map Properties:
>       Normalized Map Mass: 8140085.00
>       Raw Map Mass: 8154324.79
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[18:51:05] Calculating preliminary abundance estimates
[18:51:05] Quantifying expression levels in locus.
> Processed 101045 loci.                       [*************************] 100%
You are using Cufflinks v2.2.1, which is the most recent release.
[18:54:51] Loading reference annotation.
[18:54:55] Inspecting maps and determining fragment length distributions.
> Map Properties:
>       Normalized Map Mass: 9633710.00
>       Raw Map Mass: 9648454.55
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[18:56:46] Calculating preliminary abundance estimates
[18:56:46] Quantifying expression levels in locus.
```

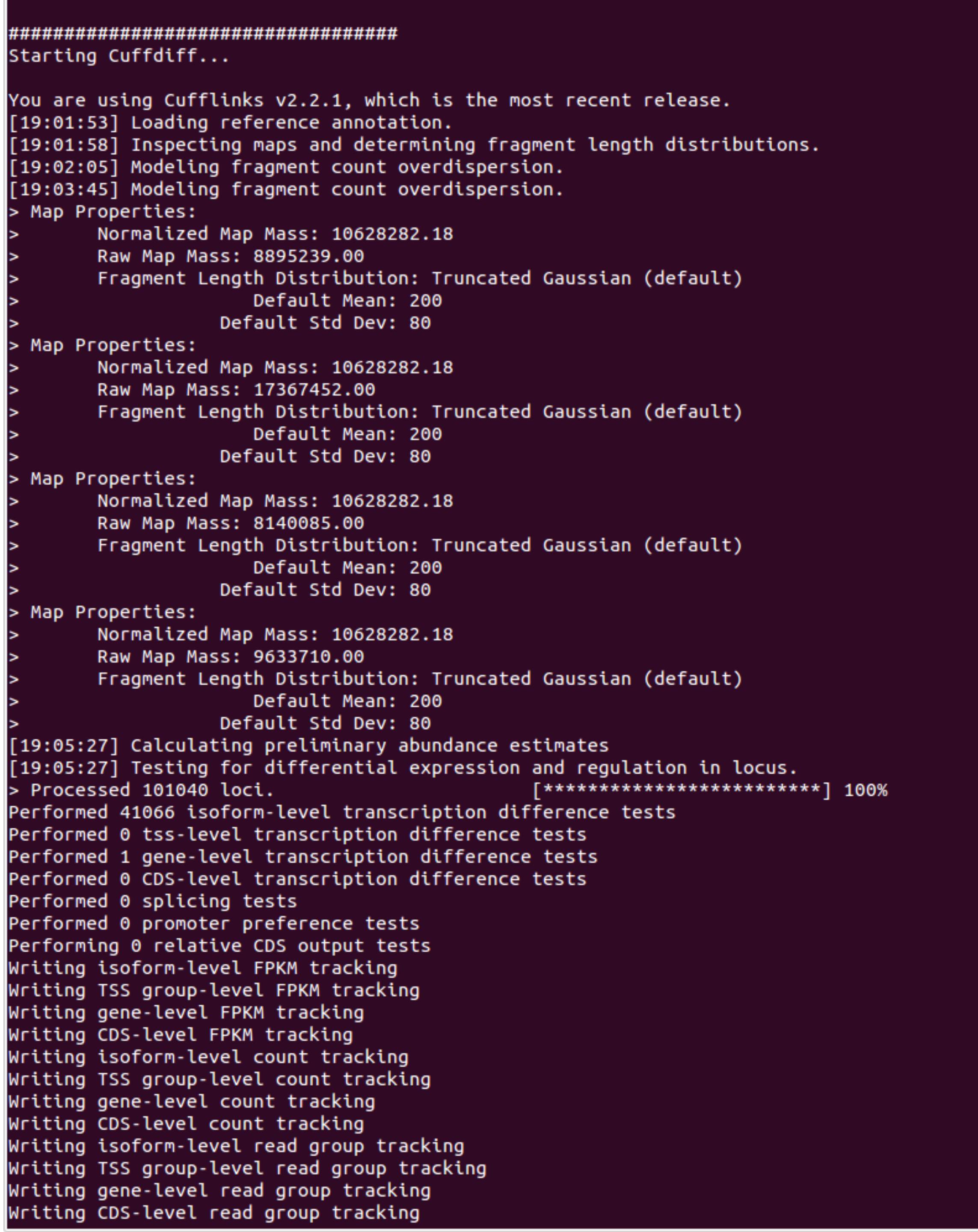
This is what running Cuffnorm from the pipeline looks like:



```
###################################
Starting Cuffnorm...

You are using Cufflinks v2.2.1, which is the most recent release.
[19:01:03] Loading reference annotation.
[19:01:08] Inspecting maps and determining fragment length distributions.
[19:01:14] Calculating preliminary abundance estimates
[19:01:14] Normalizing expression levels for locus
> Processed 101040 loci.                       [*************************] 100%
Writing isoform-level FPKM tracking
Writing TSS group-level FPKM tracking
Writing gene-level FPKM tracking
Writing CDS-level FPKM tracking
Writing isoform-level count tracking
Writing TSS group-level count tracking
Writing gene-level count tracking
Writing CDS-level count tracking
Writing isoform-level attributes
Writing TSS group-level attributes
Writing gene-level attributes
Writing CDS-level attributes
Writing read group info
Writing run info


Cuffnorm done.
```

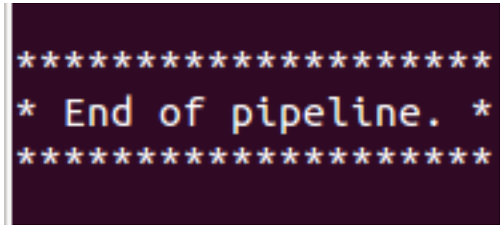
This is what running Cuffdiff from the pipeline looks like:

Note: If you chose not to run Cuffdiff you will not see this section



```
###################################
Starting Cuffdiff...

You are using Cufflinks v2.2.1, which is the most recent release.
[19:01:53] Loading reference annotation.
[19:01:58] Inspecting maps and determining fragment length distributions.
[19:02:05] Modeling fragment count overdispersion.
[19:03:45] Modeling fragment count overdispersion.
> Map Properties:
>       Normalized Map Mass: 10628282.18
>       Raw Map Mass: 8895239.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
> Map Properties:
>       Normalized Map Mass: 10628282.18
>       Raw Map Mass: 17367452.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
> Map Properties:
>       Normalized Map Mass: 10628282.18
>       Raw Map Mass: 8140085.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
> Map Properties:
>       Normalized Map Mass: 10628282.18
>       Raw Map Mass: 9633710.00
>       Fragment Length Distribution: Truncated Gaussian (default)
>                     Default Mean: 200
>                  Default Std Dev: 80
[19:05:27] Calculating preliminary abundance estimates
[19:05:27] Testing for differential expression and regulation in locus.
> Processed 101040 loci.                       [*************************] 100%
Performed 41066 isoform-level transcription difference tests
Performed 0 tss-level transcription difference tests
Performed 1 gene-level transcription difference tests
Performed 0 CDS-level transcription difference tests
Performed 0 splicing tests
Performed 0 promoter preference tests
Performing 0 relative CDS output tests
Writing isoform-level FPKM tracking
Writing TSS group-level FPKM tracking
Writing gene-level FPKM tracking
Writing CDS-level FPKM tracking
Writing isoform-level count tracking
Writing TSS group-level count tracking
Writing gene-level count tracking
Writing CDS-level count tracking
Writing isoform-level read group tracking
Writing TSS group-level read group tracking
Writing gene-level read group tracking
Writing CDS-level read group tracking
```

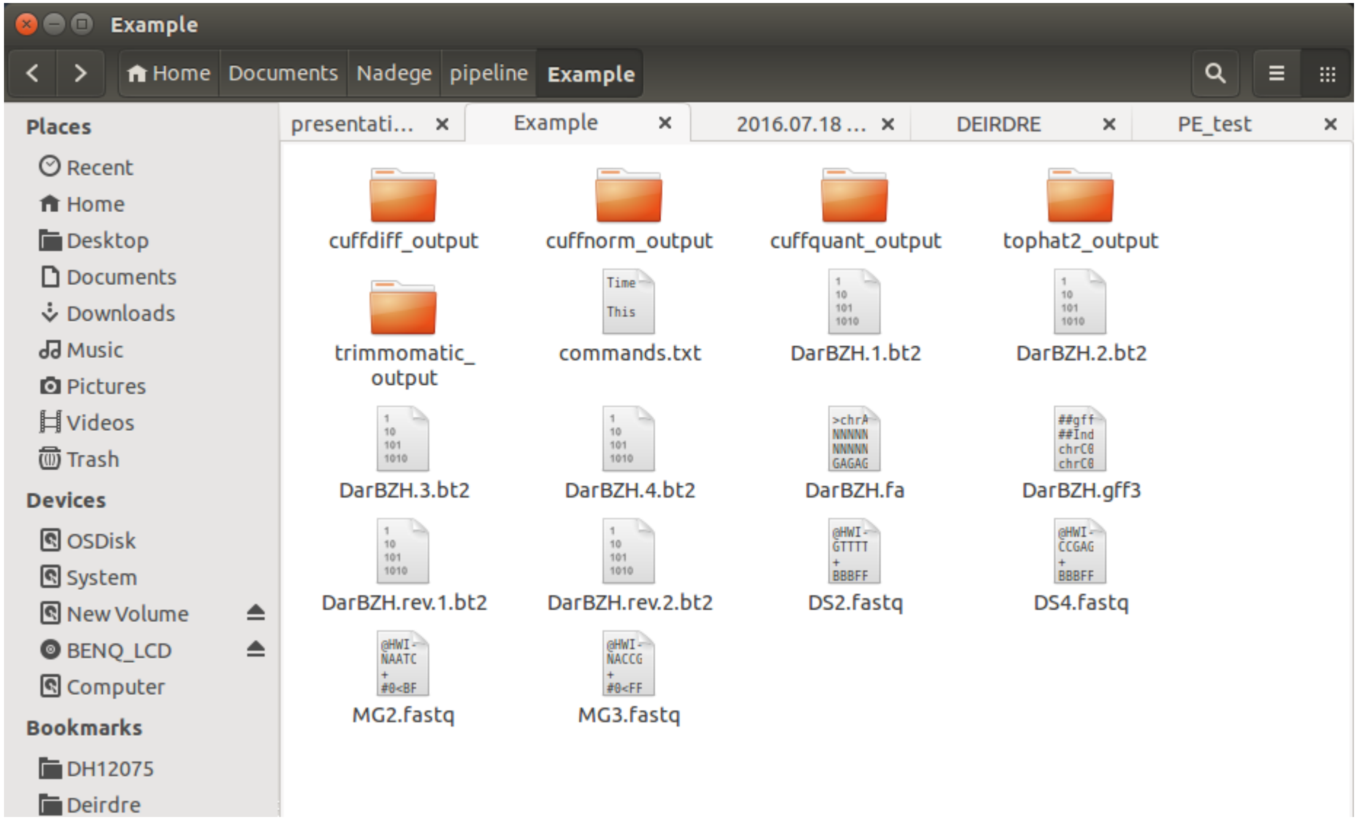
This is the “End of pipeline” message that will print once all the commands have been built/run and the pipeline’s execution is done.



```
********************
* End of pipeline. *
********************
```

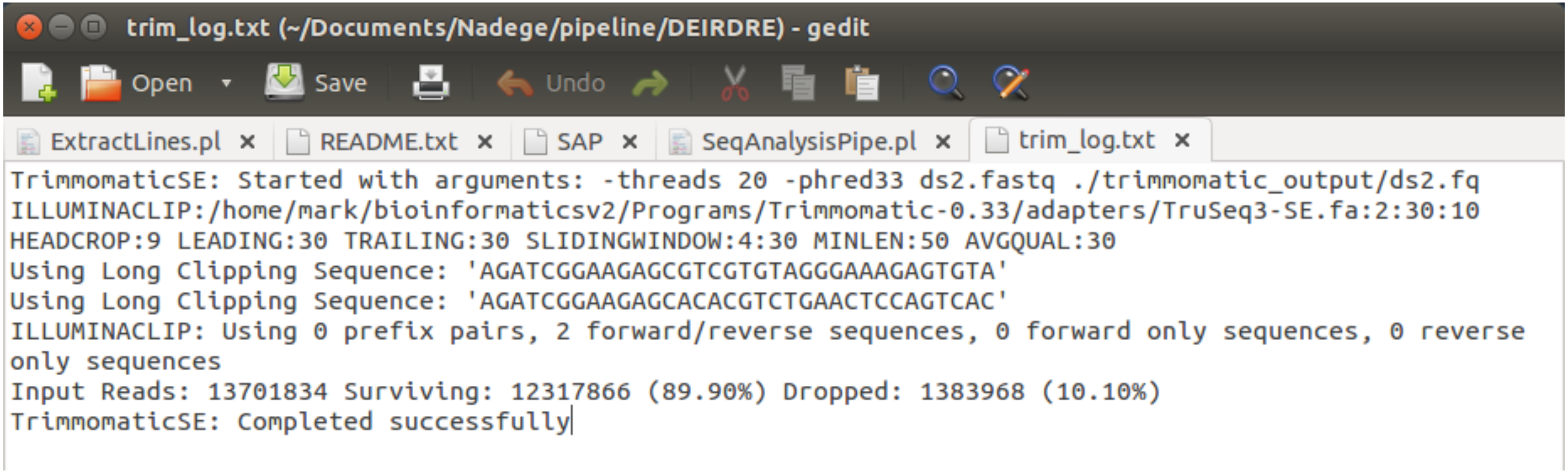
Pipeline Output

This is the original folder you placed your reads, genome, and fasta files in. At the end of the pipeline you should see a “commands.txt” file, the bowtie index files, and output directories for each step of the pipeline in addition to the files you placed there before using the pipeline. If you chose not to run Cuffdiff there will not be an output directory for it.



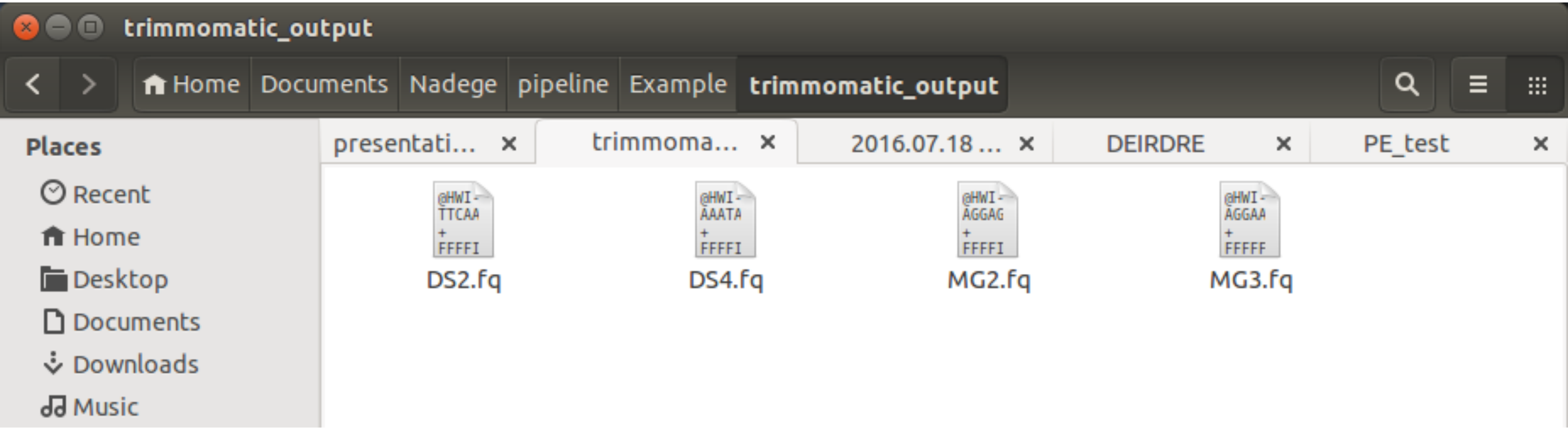
You will also find in that same original folder a “trim_log.txt” file. This contains the console output of Trimmomatic, meaning it will have the numbers of input reads and number of dropped reads among the other things Trimmomatic prints out. An short example of the contents of the trim_log file is shown below.

trim_log.txt

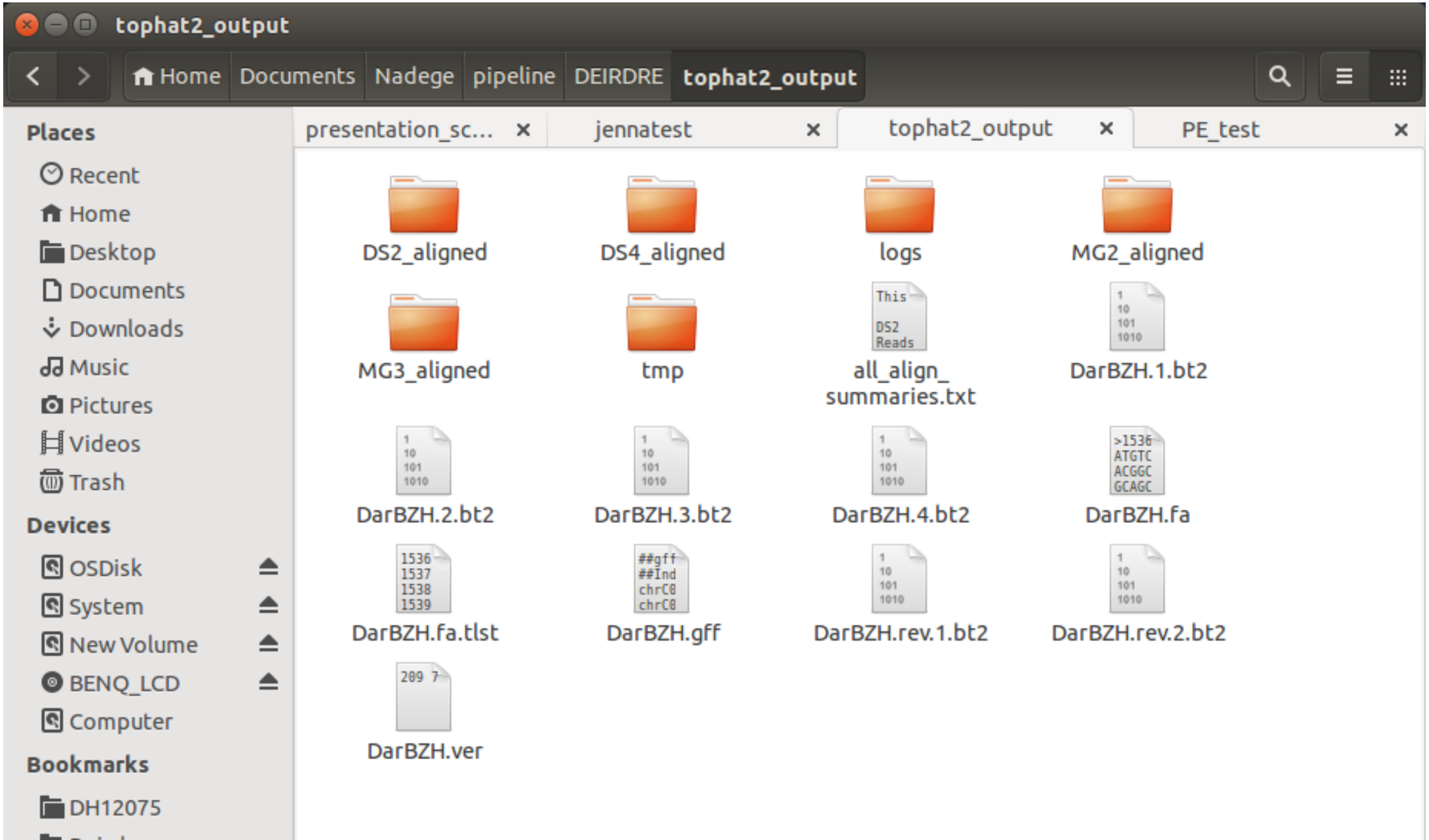


```
TrimmomaticSE: Started with arguments: -threads 20 -phred33 ds2.fastq ./trimmomatic_output/ds2.fq
ILLUMINACLIP:/home/mark/bioinformaticsv2/Programs/Trimmomatic-0.33/adapters/TruSeq3-SE.fa:2:30:10
HEADCROP:9 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50 AVGQUAL:30
Using Long Clipping Sequence: 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA'
Using Long Clipping Sequence: 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC'
ILLUMINACLIP: Using 0 prefix pairs, 2 forward/reverse sequences, 0 forward only sequences, 0 reverse
only sequences
Input Reads: 13701834 Surviving: 12317866 (89.90%) Dropped: 1383968 (10.10%)
TrimmomaticSE: Completed successfully
```

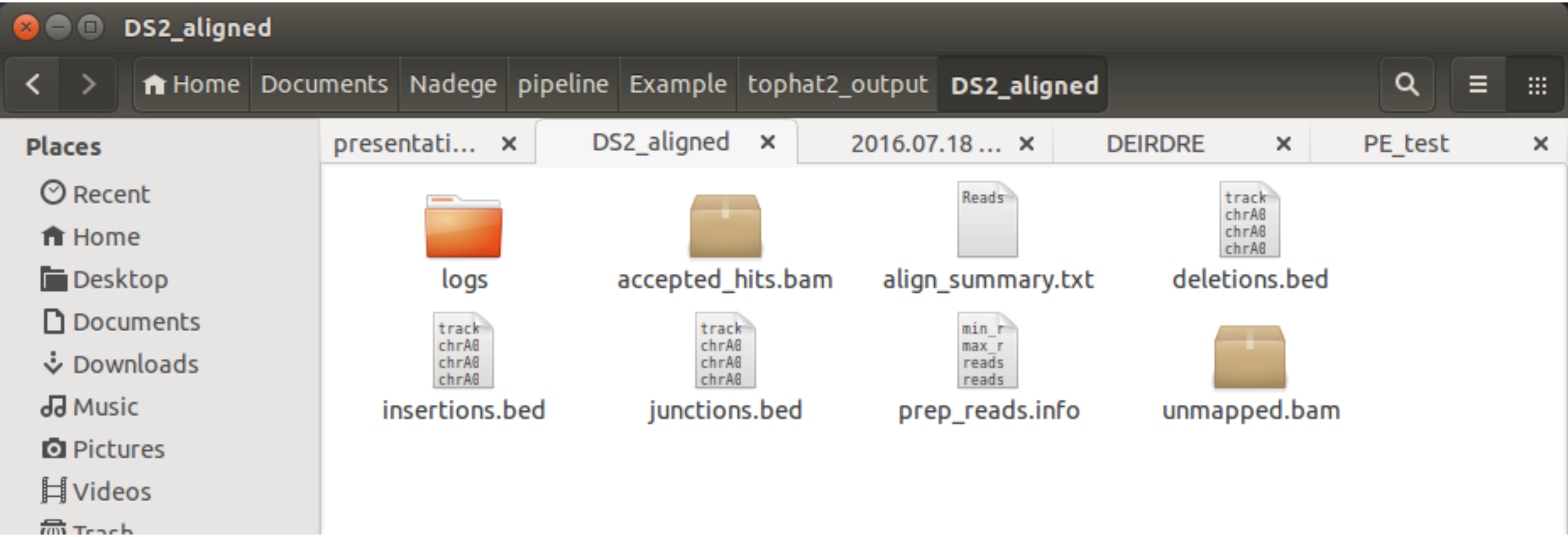
These are the contents of the Trimmomatic output directory. You will have your trimmed reads files (.fq) here:



These are the contents of the Tophat2 output directory: it will contain a folder for each reads file, the files that make up the transcriptome index, other folders created by Tophat2 during its execution, and an “all_align_summaries.txt” file that contains the “align_summary.txt” files of each individual reads file.

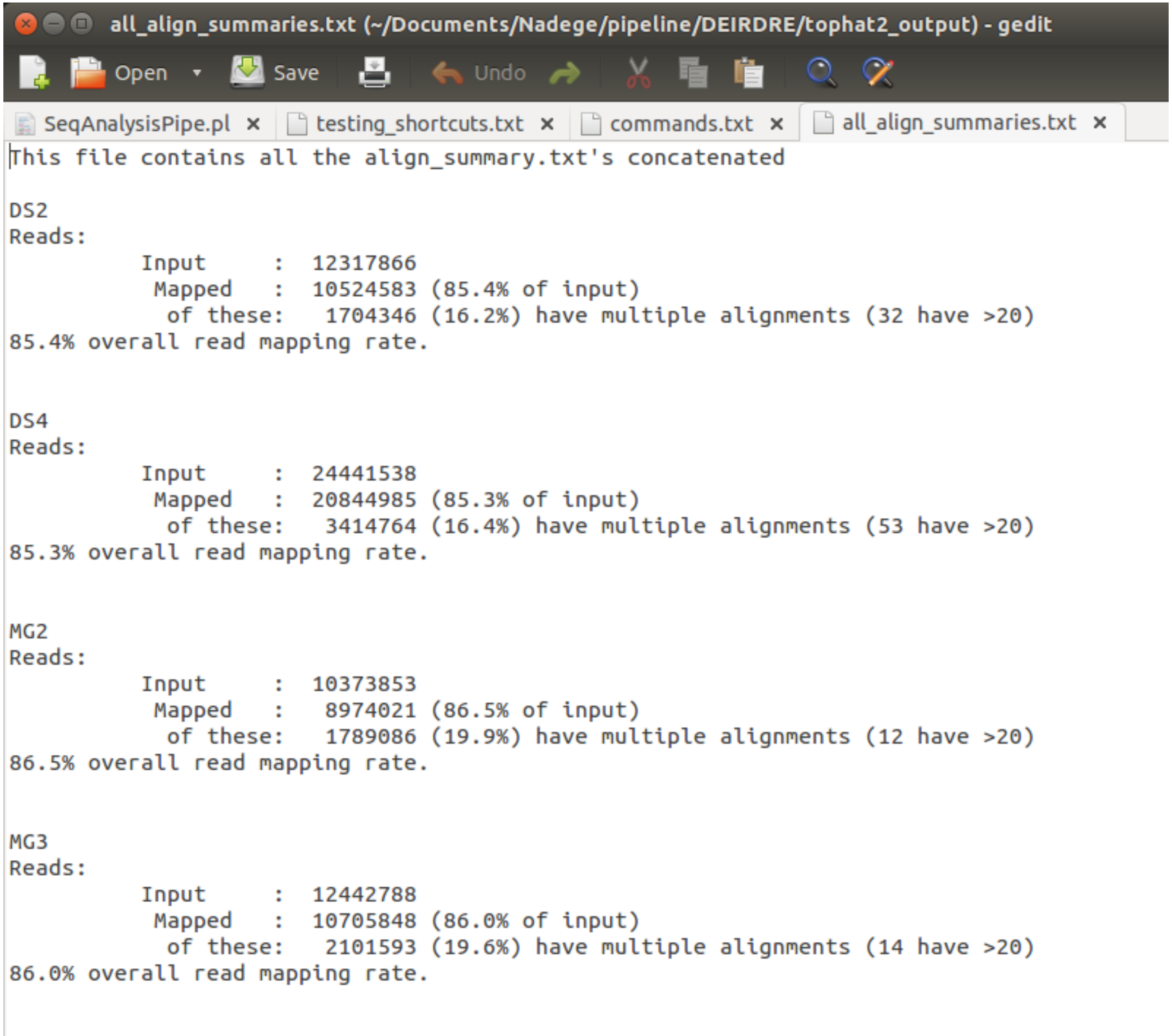


This is what a single reads file directory contains:



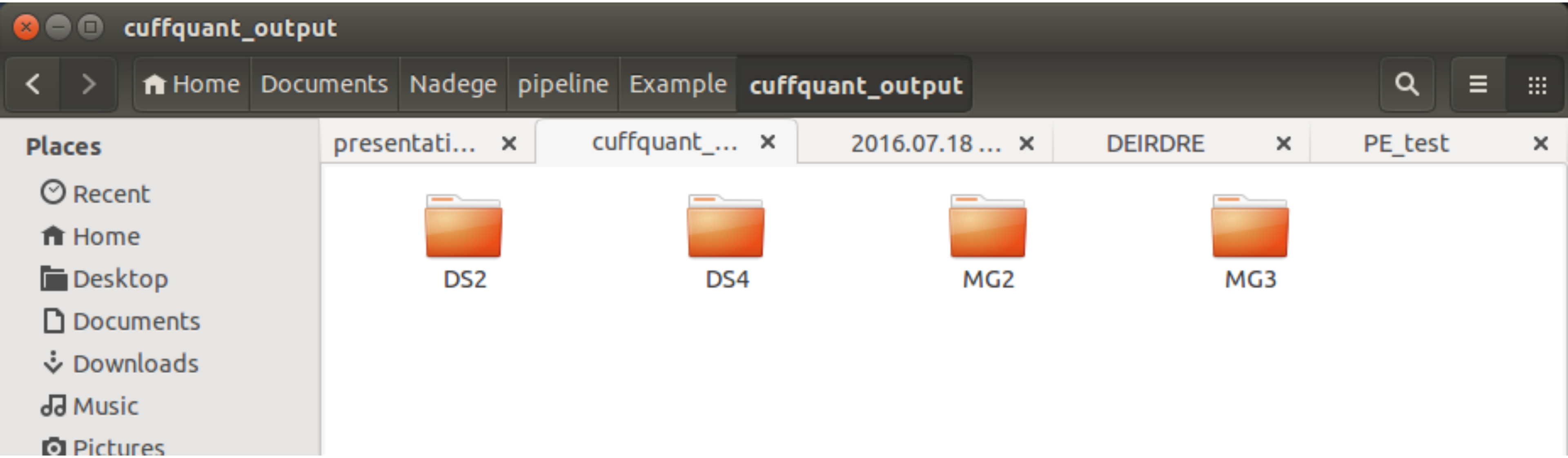
If you selected SE settings each reads file will have its own "filename_aligned" folder. If you selected PE settings each pair of reads file will have its own "filename1_and_filename2_aligned" folder.

This is what the “all_align_summary.txt” file looks like:

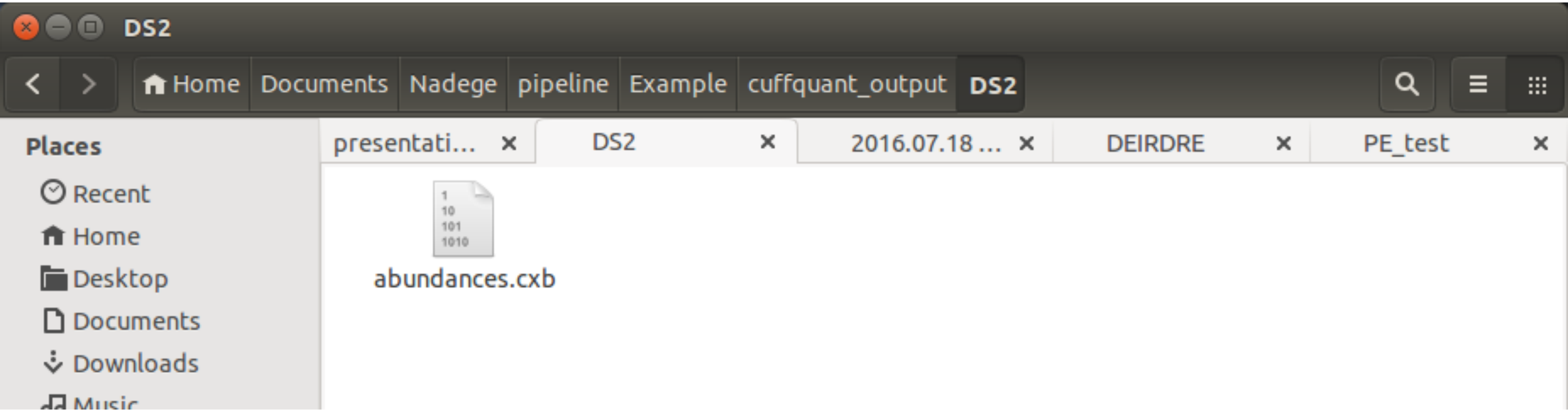


```
This file contains all the align_summary.txt's concatenated

DS2
Reads:
          Input     :  12317866
           Mapped   :  10524583 (85.4% of input)
            of these:   1704346 (16.2%) have multiple alignments (32 have >20)
85.4% overall read mapping rate.


DS4
Reads:
          Input     :  24441538
           Mapped   :  20844985 (85.3% of input)
            of these:   3414764 (16.4%) have multiple alignments (53 have >20)
85.3% overall read mapping rate.


MG2
Reads:
          Input     :  10373853
           Mapped   :   8974021 (86.5% of input)
            of these:   1789086 (19.9%) have multiple alignments (12 have >20)
86.5% overall read mapping rate.


MG3
Reads:
          Input     :  12442788
           Mapped   :  10705848 (86.0% of input)
            of these:   2101593 (19.6%) have multiple alignments (14 have >20)
86.0% overall read mapping rate.
```

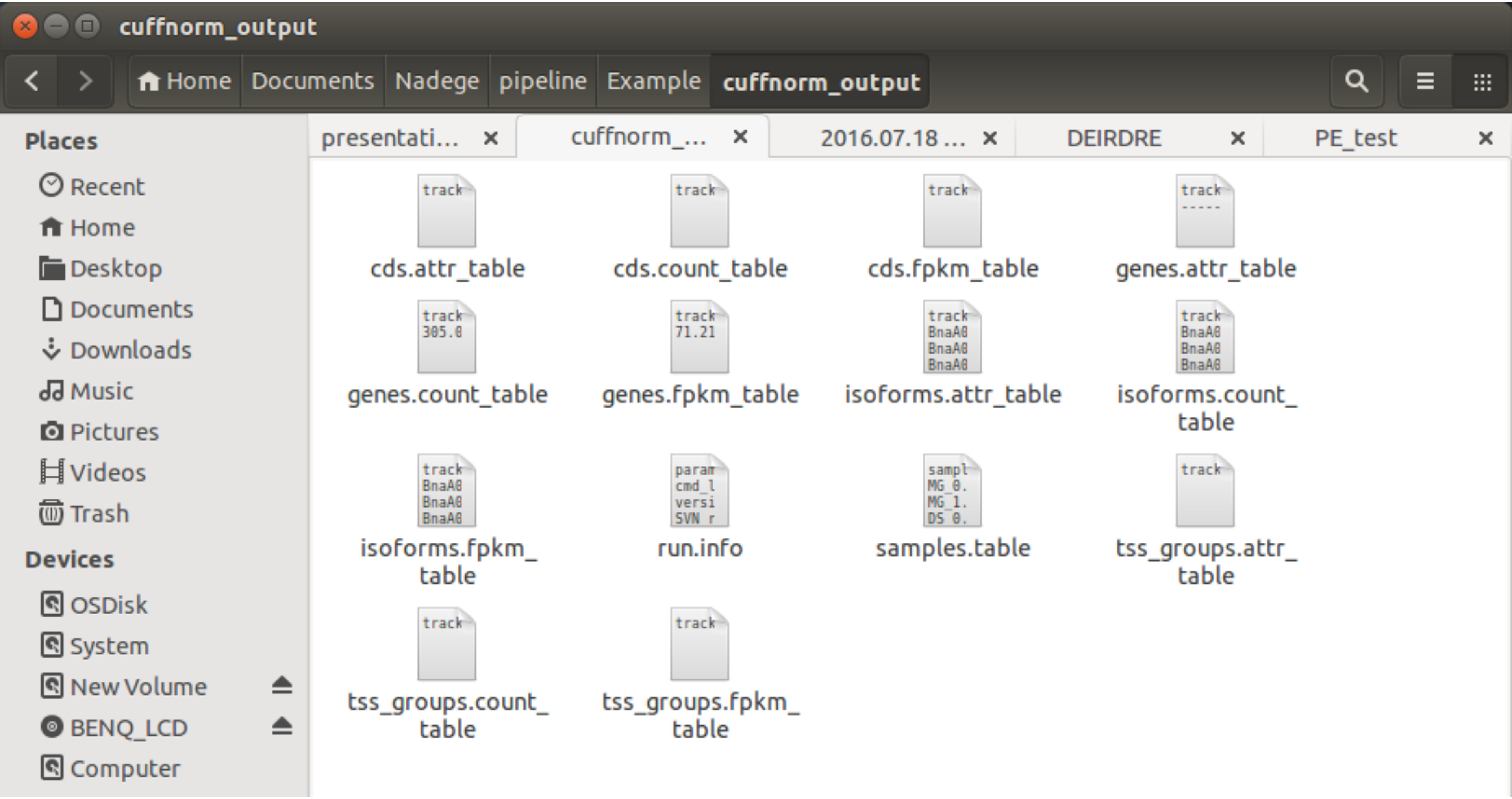
This is what the Cuffquant output directory will look like:



Each folder (one for each reads file/pair) will contain an "abundances.cxb" file. These are the input files for Cuffnorm and Cuffdiff.

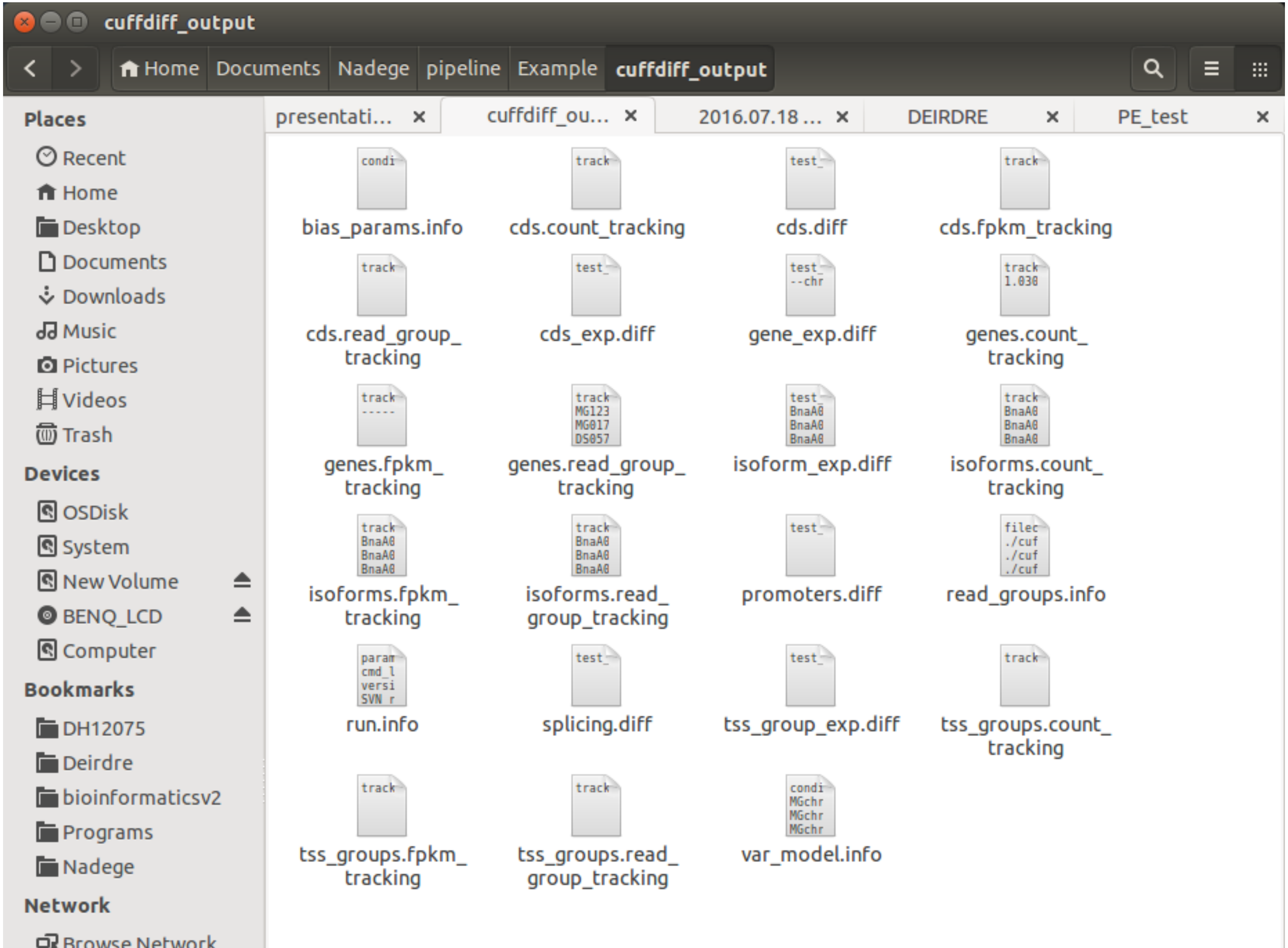


These are the files the Cuffnorm output directory will contain:



cds.attr_table
cds.count_table
cds.fpkm_table
genes.attr_table
genes.count_table
genes.fpkm_table
isoforms.attr_table
isoforms.count_table
isoforms.fpkm_table
run.info
samples.table
tss_groups.attr_table
tss_groups.count_table
tss_groups.fpkm_table

These are the files the Cuffdiff output directory will contain (if run):



bias_params.info
cds.count_tracking
cds.diff
cds.fpkm_tracking
cds.read_group_tracking
cds_exp.diff
gene_exp.diff
genes.count_tracking
genes.fpkm_tracking
genes.read_group_tracking
isoform_exp.diff
isoforms.count_tracking
isoforms.fpkm_tracking
isoforms.read_group_tracking
promoters.diff
read_groups.info
run.info
splicing.diff
tss_group_exp.diff
tss_groups.count_tracking
tss_groups.fpkm_tracking
tss_groups.read_group_tracking
var_model.info

Links

- Trimmomatic: http://www.usadellab.org/cms/?page=trimmomatic and http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf
- Bowtie2(-build): http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#the-bowtie2-build-indexer
- Tophat2: https://ccb.jhu.edu/software/tophat/manual.shtml
- Cuffquant: http://cole-trapnell-lab.github.io/cufflinks/cuffquant/
- Cuffnorm: http://cole-trapnell-lab.github.io/cufflinks/cuffnorm/index.html
- Cuffdiff: http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html
- Cufflinks: http://cole-trapnell-lab.github.io/cufflinks/manual/