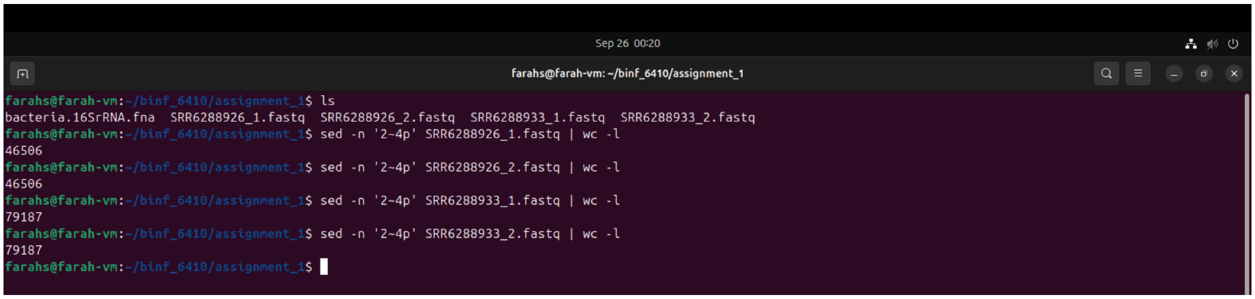
Assignment #1: Exploring spatial variation in gut microbiota composition along the intestinal tract of healthy individuals

All code and files produced for the assignment can be found here:
https://github.com/farah-y-sadoon/binf_6410/tree/main/assignment_1

Question 1: How many sequences are present in each of the four FASTQ files before merging?

Answer:
SRR6288926_1.fastq, SRR6288926_2.fastq = 46506 (DL)
SRR6288933_1.fastq, SRR6288933_2.fastq = 79187 (PM)

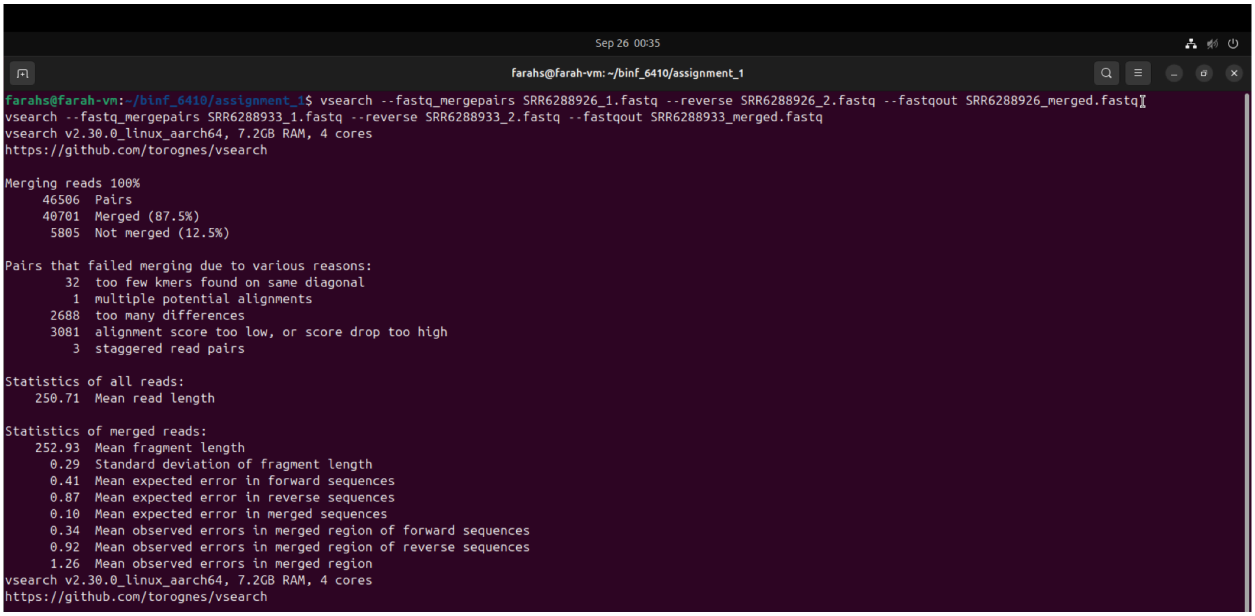


```
farahs@farah-vm:~/binf_6410/assignment_1$ ls
bacteria.16SrRNA.fna  SRR6288926_1.fastq  SRR6288926_2.fastq  SRR6288933_1.fastq  SRR6288933_2.fastq
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288926_1.fastq | wc -l
46506
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288926_2.fastq | wc -l
46506
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288933_1.fastq | wc -l
79187
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288933_2.fastq | wc -l
79187
farahs@farah-vm:~/binf_6410/assignment_1$ 
```

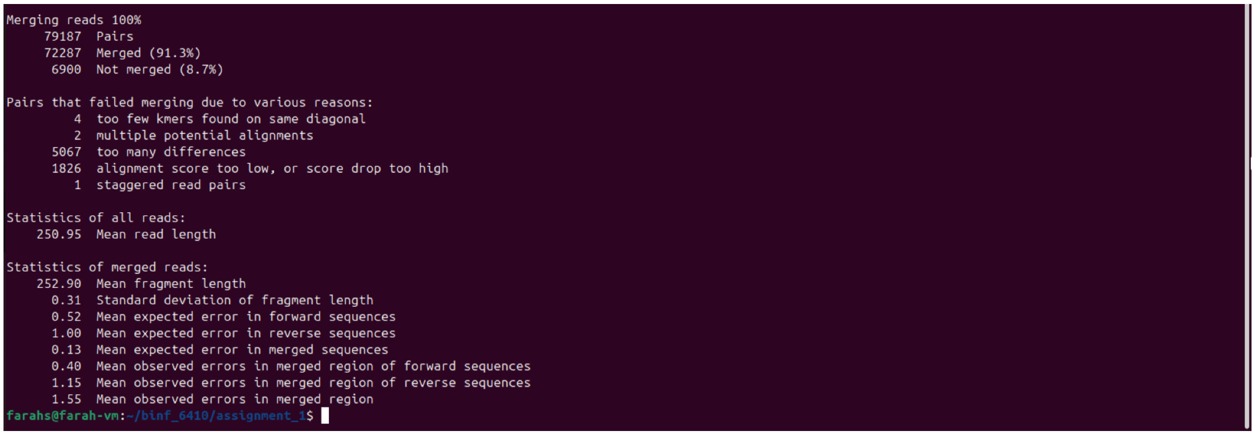
Question 2. After merging the forward and reverse reads for each FASTQ pair, how many sequences are there in each of the resulting two merged FASTQ files?

Answer:
SRR6288926_merged.fastq = 40701 (DL)
SRR6288933_merged.fastq = 72287 (PM)

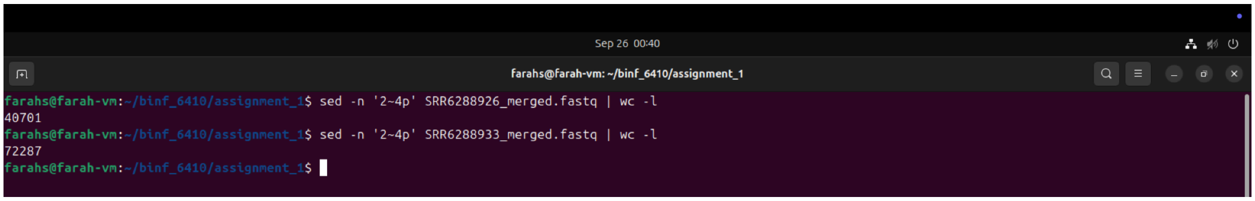
Process for merging reads and outputting results into new fastq files called *_merged.fastq



```
farahs@farah-vm:~/binf_6410/assignment_1$ vsearch --fastq_mergepairs SRR6288926_1.fastq --reverse SRR6288926_2.fastq --fastqout SRR6288926_merged.fastq
vsearch --fastq_mergepairs SRR6288933_1.fastq --reverse SRR6288933_2.fastq --fastqout SRR6288933_merged.fastq
vsearch v2.30.0_linux_aarch64, 7.2GB RAM, 4 cores
https://github.com/torognes/vsearch

Merging reads 100%
     46506  Pairs
     40701  Merged (87.5%)
      5805  Not merged (12.5%)

Pairs that failed merging due to various reasons:
        32  too few kmers found on same diagonal
         1  multiple potential alignments
      2688  too many differences
      3081  alignment score too low, or score drop too high
         3  staggered read pairs

Statistics of all reads:
    250.71  Mean read length

Statistics of merged reads:
    252.93  Mean fragment length
      0.29  Standard deviation of fragment length
      0.41  Mean expected error in forward sequences
      0.87  Mean expected error in reverse sequences
      0.10  Mean expected error in merged sequences
      0.34  Mean observed errors in merged region of forward sequences
      0.92  Mean observed errors in merged region of reverse sequences
      1.26  Mean observed errors in merged region
vsearch v2.30.0_linux_aarch64, 7.2GB RAM, 4 cores
https://github.com/torognes/vsearch
```



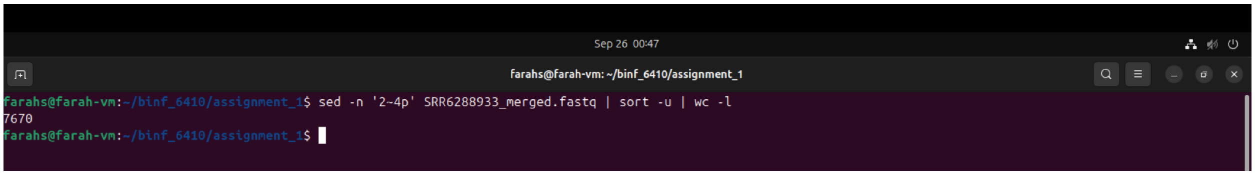
```
Merging reads 100%
     79187  Pairs
     72287  Merged (91.3%)
      6900  Not merged (8.7%)

Pairs that failed merging due to various reasons:
         4  too few kmers found on same diagonal
         2  multiple potential alignments
      5067  too many differences
      1826  alignment score too low, or score drop too high
         1  staggered read pairs

Statistics of all reads:
    250.95  Mean read length

Statistics of merged reads:
    252.90  Mean fragment length
      0.31  Standard deviation of fragment length
      0.52  Mean expected error in forward sequences
      1.00  Mean expected error in reverse sequences
      0.13  Mean expected error in merged sequences
      0.40  Mean observed errors in merged region of forward sequences
      1.15  Mean observed errors in merged region of reverse sequences
      1.55  Mean observed errors in merged region
farahs@farah-vm:~/binf_6410/assignment_1$ 
```



```
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288926_merged.fastq | wc -l
40701
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288933_merged.fastq | wc -l
72287
farahs@farah-vm:~/binf_6410/assignment_1$ 
```

Question 3. How many unique sequences are present in the PM?

Answer: 7670

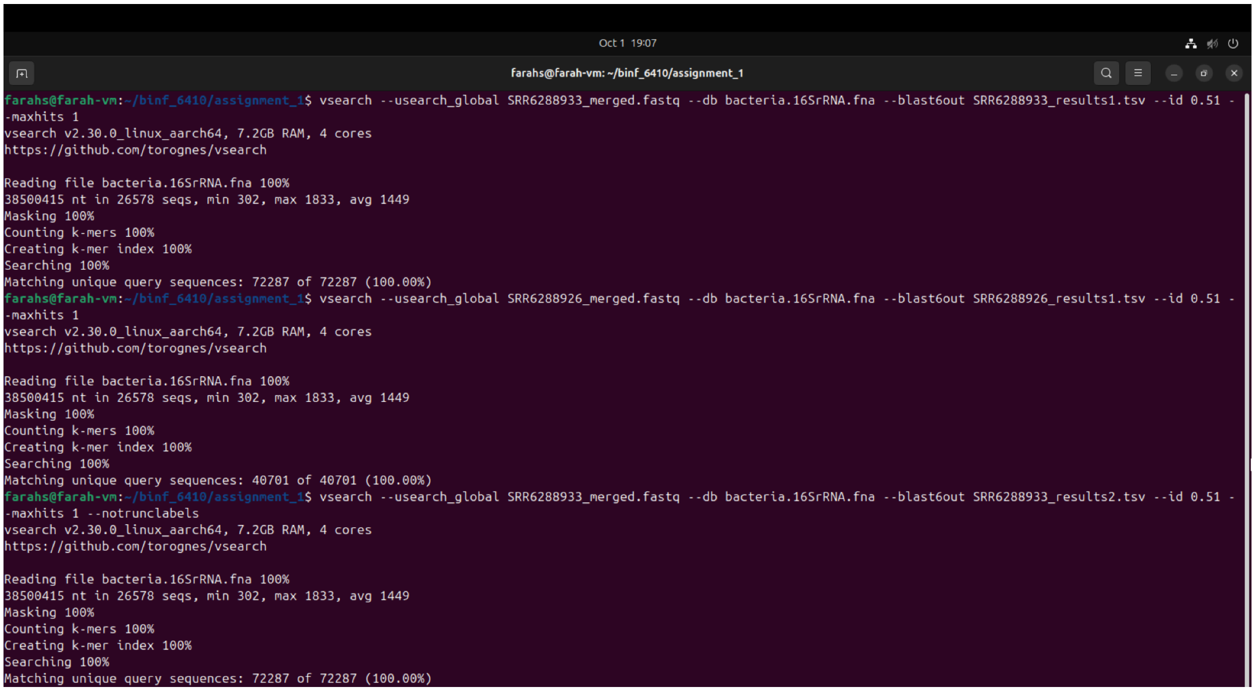


```
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288933_merged.fastq | sort -u | wc -l
7670
farahs@farah-vm:~/binf_6410/assignment_1$ 
```

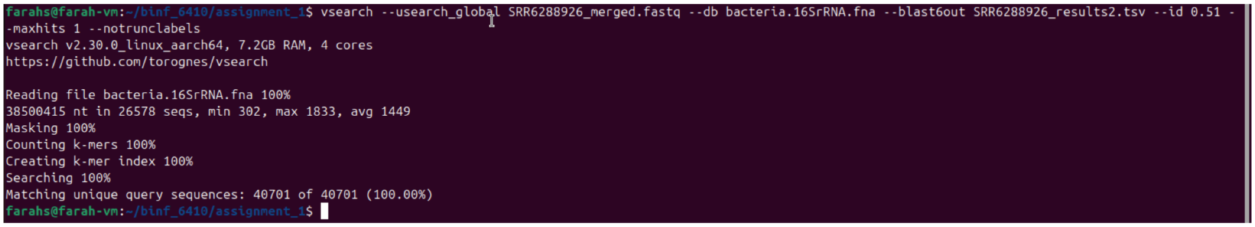
Question 4. How many unique species are present in the DL?

Answer: 347

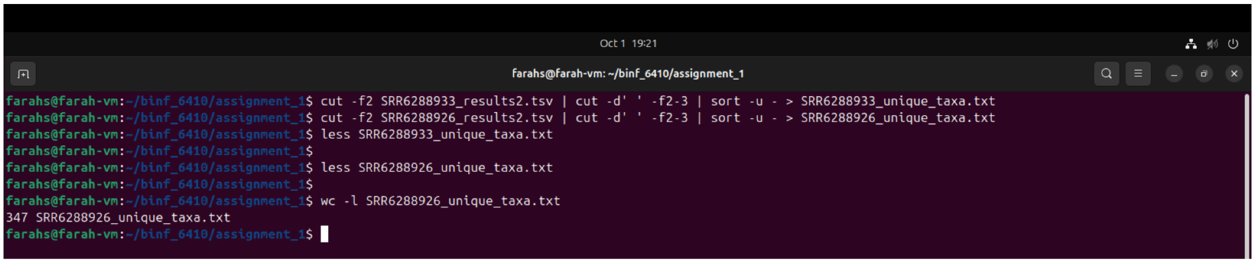
Process for matching sequences in the *merged.fastq files and outputting them to *results1.tsv and *results2.tsv (truncated and non-truncated labels respectively), then creating new files called *unique_taxa.txt for only unique taxa for each sample (DL and PM)



```
farahs@farah-vm:~/binf_6410/assignment_1$ vsearch --usearch_global SRR6288933_merged.fastq --db bacteria.16SrRNA.fna --blast6out SRR6288933_results1.tsv --id 0.51 -
-maxhits 1
vsearch v2.30.0_linux_aarch64, 7.2GB RAM, 4 cores
https://github.com/torognes/vsearch

Reading file bacteria.16SrRNA.fna 100%
38500415 nt in 26578 seqs, min 302, max 1833, avg 1449
Masking 100%
Counting k-mers 100%
Creating k-mer index 100%
Searching 100%
Matching unique query sequences: 72287 of 72287 (100.00%)
farahs@farah-vm:~/binf_6410/assignment_1$ vsearch --usearch_global SRR6288926_merged.fastq --db bacteria.16SrRNA.fna --blast6out SRR6288926_results1.tsv --id 0.51 -
-maxhits 1
vsearch v2.30.0_linux_aarch64, 7.2GB RAM, 4 cores
https://github.com/torognes/vsearch

Reading file bacteria.16SrRNA.fna 100%
38500415 nt in 26578 seqs, min 302, max 1833, avg 1449
Masking 100%
Counting k-mers 100%
Creating k-mer index 100%
Searching 100%
Matching unique query sequences: 40701 of 40701 (100.00%)
farahs@farah-vm:~/binf_6410/assignment_1$ vsearch --usearch_global SRR6288933_merged.fastq --db bacteria.16SrRNA.fna --blast6out SRR6288933_results2.tsv --id 0.51 -
-maxhits 1 --notrunclabels
vsearch v2.30.0_linux_aarch64, 7.2GB RAM, 4 cores
https://github.com/torognes/vsearch

Reading file bacteria.16SrRNA.fna 100%
38500415 nt in 26578 seqs, min 302, max 1833, avg 1449
Masking 100%
Counting k-mers 100%
Creating k-mer index 100%
Searching 100%
Matching unique query sequences: 72287 of 72287 (100.00%)
```



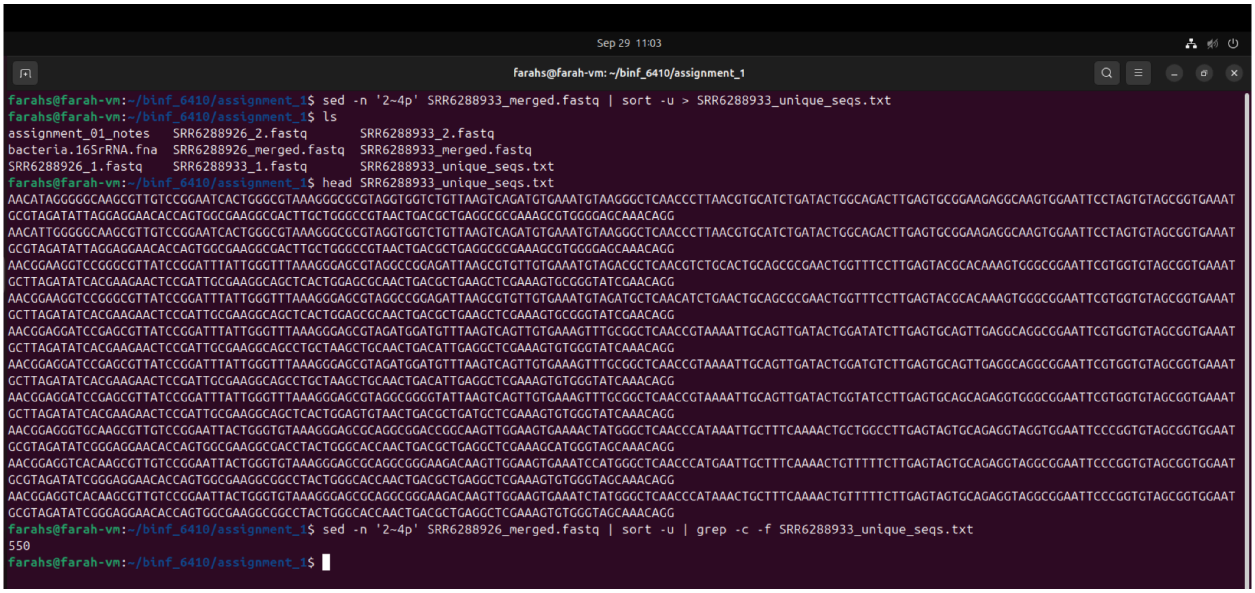
```
farahs@farah-vm:~/binf_6410/assignment_1$ vsearch --usearch_global SRR6288926_merged.fastq --db bacteria.16SrRNA.fna --blast6out SRR6288926_results2.tsv --id 0.51 -
-maxhits 1 --notrunclabels
vsearch v2.30.0_linux_aarch64, 7.2GB RAM, 4 cores
https://github.com/torognes/vsearch

Reading file bacteria.16SrRNA.fna 100%
38500415 nt in 26578 seqs, min 302, max 1833, avg 1449
Masking 100%
Counting k-mers 100%
Creating k-mer index 100%
Searching 100%
Matching unique query sequences: 40701 of 40701 (100.00%)
farahs@farah-vm:~/binf_6410/assignment_1$ 
```



```
farahs@farah-vm:~/binf_6410/assignment_1$ cut -f2 SRR6288933_results2.tsv | cut -d' ' -f2-3 | sort -u - > SRR6288933_unique_taxa.txt
farahs@farah-vm:~/binf_6410/assignment_1$ cut -f2 SRR6288926_results2.tsv | cut -d' ' -f2-3 | sort -u - > SRR6288926_unique_taxa.txt
farahs@farah-vm:~/binf_6410/assignment_1$ less SRR6288933_unique_taxa.txt
farahs@farah-vm:~/binf_6410/assignment_1$
farahs@farah-vm:~/binf_6410/assignment_1$ less SRR6288926_unique_taxa.txt
farahs@farah-vm:~/binf_6410/assignment_1$
farahs@farah-vm:~/binf_6410/assignment_1$ wc -l SRR6288926_unique_taxa.txt
347 SRR6288926_unique_taxa.txt
farahs@farah-vm:~/binf_6410/assignment_1$ 
```

Question 5. How many unique sequences are shared between DL and PM sites?

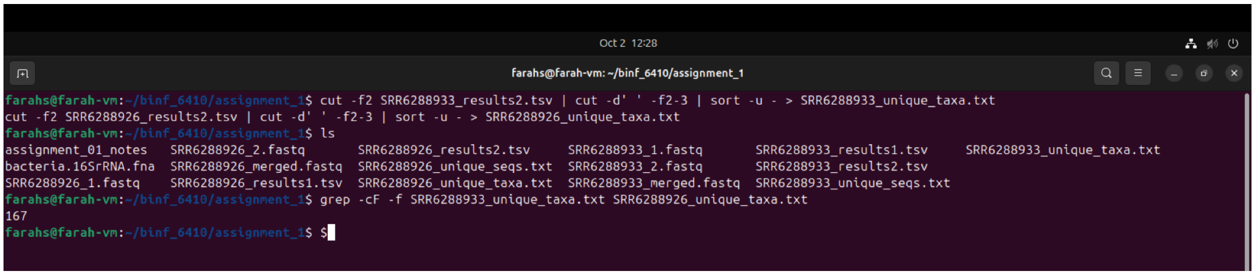
Answer: 550



```
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288933_merged.fastq | sort -u > SRR6288933_unique_seqs.txt
farahs@farah-vm:~/binf_6410/assignment_1$ ls
assignment_01_notes   SRR6288926_2.fastq       SRR6288933_2.fastq
bacteria.16SrRNA.fna  SRR6288926_merged.fastq  SRR6288933_merged.fastq
SRR6288926_1.fastq    SRR6288933_1.fastq       SRR6288933_unique_seqs.txt
farahs@farah-vm:~/binf_6410/assignment_1$ head SRR6288933_unique_seqs.txt
AACATAGGGGGCAAGCGTTGTCCGGAATCACTGGGCGTAAAGGGCGCGTAGGTGGTCTGTTAAGTCAGATGTGAAATGTAAGGGCTCAACCCTTAACGTGCATCTGATACTGGCAGACTTGAGTGCGGAAGAGGCAAGTGGAATTCCTAGTGTAGCGGTGAAAT
GCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCGACTTGCTGGGCCGTAACTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACAGG
AACATTGGGGGCAAGCGTTGTCCGGAATCACTGGGCGTAAAGGGCGCGTAGGTGGTCTGTTAAGTCAGATGTGAAATGTAAGGGCTCAACCCTTAACGTGCATCTGATACTGGCAGACTTGAGTGCGGAAGAGGCAAGTGGAATTCCTAGTGTAGCGGTGAAAT
GCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCGACTTGCTGGGCCGTAACTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACAGG
AACGGAAGGTCCGGGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCCGGAGATTAAGCGTGTTGTGAAATGTAGACGCTCAACGTCTGCACTGCAGCGCGAACTGGTTTCCTTGAGTACGCACAAAGTGGGCGGAATTCGTGGTGTAGCGGTGAAAT
GCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCTCACTGGAGCGCAACTGACGCTGAAGCTCGAAAGTGCGGGTATCGAACAGG
AACGGAAGGTCCGGGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCCGGAGATTAAGCGTGTTGTGAAATGTAGATGCTCAACATCTGAACTGCAGCGCGAACTGGTTTCCTTGAGTACGCACAAAGTGGGCGGAATTCGTGGTGTAGCGGTGAAAT
GCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCTCACTGGAGCGCAACTGACGCTGAAGCTCGAAAGTGCGGGTATCGAACAGG
AACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAAT
GCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGTGGGTATCAAACAGG
AACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATGTCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAAT
GCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGTGGGTATCAAACAGG
AACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGGGTATTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGTATCCTTGAGTGCAGCAGAGGTGGGCGGAATTCGTGGTGTAGCGGTGAAAT
GCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCTCACTGGAGTGTAACTGACGCTGATGCTCGAAAGTGTGGGTATCAAACAGG
AACGGAGGGTGCAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGACCGGCAAGTTGGAAGTGAAAACTATGGGCTCAACCCATAAATTGCTTTCAAAACTGCTGGCCTTGAGTAGTGCAGAGGTAGGTGGAATTCCCGGTGTAGCGGTGGAAT
GCGTAGATATCGGGAGGAACACCAGTGGCGAAGGCGACCTACTGGGCACCAACTGACGCTGAGGCTCGAAAGCATGGGTAGCAAACAGG
AACGGAGGTCACAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGAAGACAAGTTGGAAGTGAAATCCATGGGCTCAACCCATGAATTGCTTTCAAAACTGTTTTTCTTGAGTAGTGCAGAGGTAGGCGGAATTCCCGGTGTAGCGGTGGAAT
GCGTAGATATCGGGAGGAACACCAGTGGCGAAGGCGGCCTACTGGGCACCAACTGACGCTGAGGCTCGAAAGTGTGGGTAGCAAACAGG
AACGGAGGTCACAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGAAGACAAGTTGGAAGTGAAATCTATGGGCTCAACCCATAAACTGCTTTCAAAACTGTTTTTCTTGAGTAGTGCAGAGGTAGGCGGAATTCCCGGTGTAGCGGTGGAAT
GCGTAGATATCGGGAGGAACACCAGTGGCGAAGGCGGCCTACTGGGCACCAACTGACGCTGAGGCTCGAAAGTGTGGGTAGCAAACAGG
farahs@farah-vm:~/binf_6410/assignment_1$ sed -n '2~4p' SRR6288926_merged.fastq | sort -u | grep -c -f SRR6288933_unique_seqs.txt
550
farahs@farah-vm:~/binf_6410/assignment_1$ 
```

Question 6. How many unique species are shared between PM and DL sites?

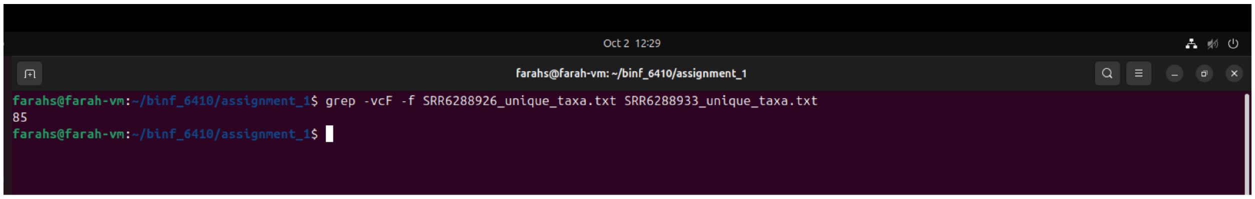
Answer: 167



```
farahs@farah-vm:~/binf_6410/assignment_1$ cut -f2 SRR6288933_results2.tsv | cut -d' ' -f2-3 | sort -u - > SRR6288933_unique_taxa.txt
cut -f2 SRR6288926_results2.tsv | cut -d' ' -f2-3 | sort -u - > SRR6288926_unique_taxa.txt
farahs@farah-vm:~/binf_6410/assignment_1$ ls
assignment_01_notes   SRR6288926_2.fastq       SRR6288926_results2.tsv     SRR6288933_1.fastq       SRR6288933_results1.tsv     SRR6288933_unique_taxa.txt
bacteria.16SrRNA.fna  SRR6288926_merged.fastq  SRR6288926_unique_seqs.txt  SRR6288933_2.fastq       SRR6288933_results2.tsv
SRR6288926_1.fastq    SRR6288926_results1.tsv  SRR6288926_unique_taxa.txt  SRR6288933_merged.fastq  SRR6288933_unique_seqs.txt
farahs@farah-vm:~/binf_6410/assignment_1$ grep -cF -f SRR6288933_unique_taxa.txt SRR6288926_unique_taxa.txt
167
farahs@farah-vm:~/binf_6410/assignment_1$ $
```

Question 7. How many unique species are exclusively present in PM but not in DL?

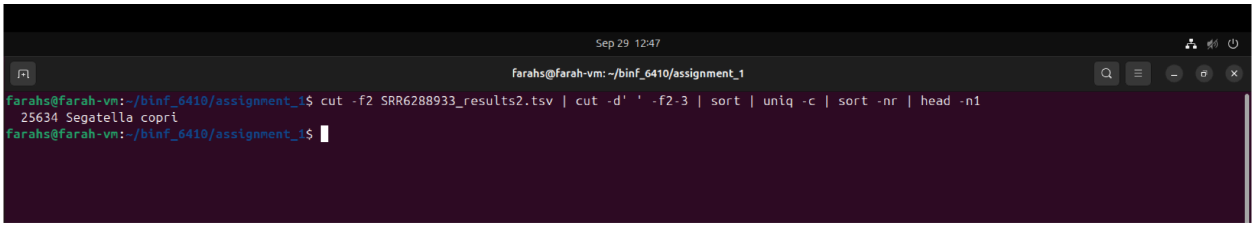
Answer: 85



```
farahs@farah-vm:~/binf_6410/assignment_1$ grep -vcF -f SRR6288926_unique_taxa.txt SRR6288933_unique_taxa.txt
85
farahs@farah-vm:~/binf_6410/assignment_1$ 
```

Question 8. What is the most abundant species in PM and how many sequence reads of it were detected?

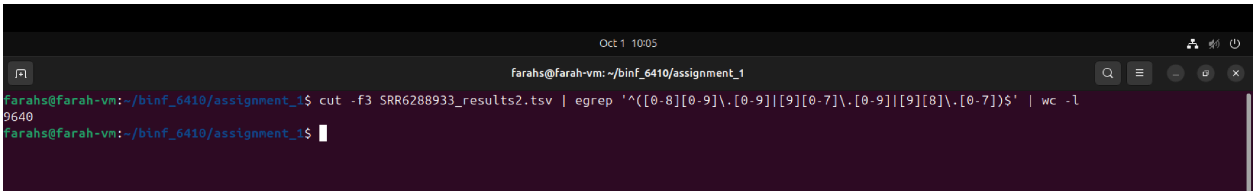
Answer: Most abundant species in PM = Segatella copri, 25, 634 sequence reads



```
farahs@farah-vm:~/binf_6410/assignment_1$ cut -f2 SRR6288933_results2.tsv | cut -d' ' -f2-3 | sort | uniq -c | sort -nr | head -n1
  25634 Segatella copri
farahs@farah-vm:~/binf_6410/assignment_1$
```

Question 9. How many query sequences in PM matched to a database reference at exactly, or less than 98.7% identity? Reminder: the % identity column is the 3rd column in a typical blast-formatted results table.

Answer: 9640



```
farahs@farah-vm:~/binf_6410/assignment_1$ cut -f3 SRR6288933_results2.tsv | egrep '^([0-8][0-9]\.[0-9]|[9][0-7]\.[0-9]|[9][8]\.[0-7])$' | wc -l
9640
farahs@farah-vm:~/binf_6410/assignment_1$
```

Question 10. Consider a scenario where you're comparing "Species A" abundance across two samples. The FASTQ of "Sample 1" contains 1000 reads, while the FASTQ of "Sample 2" contains 2000 reads. Why would it be incorrect to assume that Sample 2 has a higher abundance of "Species A" than Sample 1?

It would be incorrect to assume that "Sample 2" has a higher abundance of "Species A" because number of reads does not necessarily correspond to species abundance. Many factors could influence the number of reads generated by sequencing, including the quality of the samples and the effectiveness of library preparation. In a FASTQ file, sequence quality scores are recorded; therefore, when counting raw reads in a FASTQ file, one may be including low-quality sequences and inflating the abundance of "Species A" in "Sample 2" if not filtering for reliable sequences first.

Additionally, some level of normalization is required because counting the raw reads from Sample 2 introduces a sampling bias. For example:

- Sample 2 has 1000/2000 reads from Species A (50% abundance)
- Sample 1 has 800/1000 reads from Species A (80% abundance)

Normalization is needed here to define the relative abundance of transcripts, which would involve looking at a defined number of reads between two samples and then defining abundance by the proportion of Species A reads in each.

Question 11. Here, we assigned taxonomy based on the query sequence's top matching % sequence identity compared to known reference sequences in the NCBI RefSeq 16S rRNA Targeted Loci Database. Why might this lead to inaccurate taxonomic classification

in cases where query sequences are less than 98.7% similar to known database references?

Inaccurate taxonomic classification may occur because the accepted threshold for species-level identification based on the 16S rRNA gene is 98.7% (Stackebrandt & Ebers, 2006, Yarza et al., 2014). Threshold values represent the amount of individual variation in a sequence that is acceptable within a given taxa. This means that sequences with less than 98.7% similarity could represent different species and instead belong to higher taxonomic groups depending on their similarity. Work by Yarza et al. (2014) defines what threshold values are for genera and above based on sequences from the 16S rRNA gene. According to findings, the threshold values are as follows: 95.4% or lower for genera, 86.5% or lower for families, 82.0% or lower for classes, 78.5% or lower for orders, and 75.0% or lower for phyla.

Conversely, if novel species are sequenced and matched to known database references at less than the accepted threshold value, they may be inaccurate because it is not represented in the database. If the database does not include many reference sequences, it would lead to incorrect classification due to lack of data. Moreover, it would be misleading to assign species-level classification based on the highest existing match because it inhibits the ability to identify distinct and/or novel species.

Therefore, for the current analysis, we cannot be confident that any sequence classified as a given bacterial species based on a sequence identity of less than 98.7% similarity is accurate.

References

Stackebrandt, E. & Ebers, J. Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today 33, 152-155 (2006). https://microbiologysociety.org/static/uploaded/a8800d1f-de21-432d-a5399c13c9ad1643.pdf

Yarza, P., Yilmaz, P., Pruesse, E. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12, 635–645 (2014). https://doi.org/10.1038/nrmicro3330