Code Document:

#After we download the reference.ft file we use the below code on our personal terminal account so it's uploaded.

Scp [redacted]

Scp: Secure Copy Protocol- Used to transfer files between local and remote systems over SSH (Secure Shell)
#To load SRR files after downloading them
Scp [redacted]

Scp [redacted]

#To download blast
module load blast+/2.14.0
then
module spider blast+/2.14.0

#To create BLAST-compatible data frame from input sequence file
makeblastdb -dbtype nucl -in reference.ft

makeblastdb: formats a file of nucleotide or protein sequences, so that it can be queried using blastn, blastx, etc. Output files for nucleotides will include .nhr, .nin, .nsq extensions.

dbtype nucl: Specifies the type of sequence (nucl/prot), in this case nucleotide.

-in: what comes after it is the input file name.
#To match sequences from SRR files to the reference file
blastn -db reference.ft -query SRR6288926.fasta -max_target_seqs 1 -outfmt "6 qseqid sseqid bitscore" -out DL_match.txt

blastn -db reference.ft -query SRR6288933.fasta -max_target_seqs 1 -outfmt "6 qseqid sseqid bitscore" -out PM_match.txt

blastn: Matches a query sequence of interest against a reference sequence.
-db: What comes after it is the database that would be reference for matching.
-query: What comes after it is the sequence of interest that will be matched against the reference.
-max_target_seqs 1: Only one (the top match) will be outputted. One match per sequence.
-outfmt "6 qseqid sseqid bitscore": Output format Specifications. Format 6 with 3 columns will be outputted. The Query sequence Id, Database sequence Id, bit score alignment respectively. Bit score shows the quality of the match.
-out: What comes after it will be the output file name.

To Extract Genera from the PM and DL match files
cut -f2 PM_match.txt | sed -E 's/_/ /g' | sed -E 's/ [0-9]$//g' | sort | uniq > ./PM_genera_list

cut -f2 DL_match.txt | sed -E 's/_/ /g' | sed -E 's/ [0-9]$//g' | sort | uniq > ./DL_genera_list

cut -f2: Extracts field/column 2 from each line in the match.txt files.
Sed -E: Extended regular expressions to perform substitutions on the extracted data.
's/_/ /g': Replaces all underscores with spaces on all occurrences in each line. (on column 2)
's/ [0-9]$//g': all the spaces followed by a single digit at the end of the line will be removed.
Sort: sorts the resulting lines alphabetically.
Uniq: removes duplicate lines from the sorted output.
> ./: redirects the output file to the filename specified within the current directory.

To create common genera file
cat DL_genera_list PM_genera_list | sort | uniq -d > ./DL_PM_common.txt

To create genera unique to PM file
cat DL_PM_common.txt PM_genera_list | sort | uniq -u > ./PM_unique.txt

To create genera unique to DL file
cat DL_PM_common.txt DL_genera_list | sort | uniq -u > ./DL_unique.txt

cat: reads the contents of the file or files.
Sort: sorts the combined outputs alphabetically.
Uniq-d: Outputs only the lines that are duplicated/common.
Uniq-u: Outputs only the lines that are unique to either PM/DL genera list. (Those that are not in common)

To copy the .txt files from graham to local PC we put the following code in our personal Terminal

scp [redacted]
[redacted]
scp [redacted]
[redacted]
scp [redacted]
[redacted]
scp [redacted]
[redacted]
scp [redacted]
[redacted]
scp [redacted]
[redacted]
scp [redacted]
[redacted]

Responses Document:

- How many sequences are in each of the SRR files?
  grep -c ">" SRR6288933.fasta 158374
  grep -c ">" SRR6288926.fasta 93012

  grep -c: Counts the number of matching lines.
- What is the most common bacterial genus in each of the two samples?
  cut -f2 PM_match.txt | sort | uniq -c | sort -nr | head -1 54612 Prevotella_4
  cut -f2 DL_match.txt | sort | uniq -c | sort -nr | head -1 10196 Clostridium

  Extracts the second column in .txt files, sorts them alphabetically. Counts the occurrences of each unique line.
  Sort -nr: Sorts the counted occurrences based on numerical and reverse order. (The Bigger number at the top)
  Head-1: retrieves the first line of output.
- How many genera are unique to the distal lumen, unique to proximal mucosa, and common between them?
  wc -l DL_unique.txt 35
  wc -l PM_unique.txt 19
  wc -l DL_PM_common.txt 117

  wc -l: Counts the lines
- Your aim is to understand bacteria distributions in human colons. What could be some limitations of this experiment and analysis? Provide a brief explanation.

  1. Sampling Bias: Colon bacteria are different from one region of the colon to the other and getting samples (biopsies) from each section is challenging. Stool samples might not give comprehensive results.
  2. Sequencing Bias: While sequencing some Bactria species might be favored over the others.
  3. Gut microbiome differs between individuals based on their diet, genetics, etc.
  4. Finding a link between the taxonomy identification of the Bacteria and how they affect the host is challenging.

- We are assigning sequences to genera based on their best BLAST score. What are the sequences with the lowest BLAST score matches? Explain why or why not you would include or exclude these sequences.
  sort -nk3 PM_match.txt | head

  SRR6288933.1721 Desulfovibrio_2 60.2
  SRR6288933.23857 Fusicatenibacter_1 65.8
  SRR6288933.68163 Anaerovorax_1 71.3

```
SRR6288933.66890  Bradyrhizobium_1     73.1
SRR6288933.68029  Bilophila_1     75.0
SRR6288933.19845  Intestinibacter_1       76.8
SRR6288933.24113  Parabacteroides_4       76.8
SRR6288933.77425  Allobaculum_1       76.8
SRR6288933.13945  Clostridium     82.4
SRR6288933.73219  Blautia_4       82.4
```

```
sort -nk3 DL_match.txt | head
```

```
SRR6288926.8271   Lachnospiracea_incertae_sedis_2 56.5
SRR6288926.26751  Anaerococcus_5      67.6
SRR6288926.9175   Anaerococcus_5      67.6
SRR6288926.22142  Paenibacillus_1     73.1
SRR6288926.25206  Eubacterium_2       73.1
SRR6288926.41219  Lachnospiracea_incertae_sedis_2 73.1
SRR6288926.5525   Ethanoligenens_1    73.1
SRR6288926.35825  Anaerococcus_5      75.0
SRR6288926.715    Eubacterium_1       75.0
SRR6288926.12430  Roseburia_5     76.8
```

sort -nk3: Sorted numerically(low to high) by the values of the third column.

Sequences with low scores, usually indicate weak similarity to the reference sequence. Sometimes, the low score might be due to contamination from the sequencing process. On the other hand, in some cases, low scoring sequences might be indication of novel sequences/ genera that might be found useful in the future. Including these data, can also help illustrate the range of bacteria or allow other researchers to do further research on them.