Metagenomic Activity Workbook

By Kate Waring
December, 2020

In fulfillment of the Knowledge Mobilization component of MICB 505

Introduction

This workbook is intended to serve as a resource for the introductory level metagenomics learner or teacher. For those that are new to the field of metagenomics, these activities convey the basics in a very low tech, beginner-friendly style. The metagenomics novice will need nothing more than a printer, pencil, scissors, and some basic microbiological knowledge from a high school or undergraduate biology course to complete the activities in this workbook. By removing the pressure of coding and software usage, the broad ideas behind metagenomic workflows can be more easily conveyed, which makes room for the learner to seek the accompanying software tools from a position of knowledge power rather than getting overwhelmed with too many new things. Additionally, this format has the learner actively decision making while performing steps of the process in a more substantial way than simply clicking on something, using 'Ctrl + F', or copying and pasting. By forcing the learner to do some very basic calculations, it provides them with a more intuitive understanding of the functions that bioinformatic software is performing. For the instructor, this format would allow for a quick and easy didactic exercise during class to break up the monotony of a lecture while also varying the instruction format by allowing for active participation.

Future improvements currently identified for this project include:

- A "Follow-up Questions" section for each activity to encourage extrapolation from the activity to the real process,
- Improvements to the short background text included in each of the activities,
- Addition of materials that span the entire range of metagenomics topics (i.e. from initial DNA extraction through taxonomy assignment and gene annotation),
- Additional complexity levels for the materials for each topic,
- Inclusion of a narrative element to the workbook activities that ties them all together (for example, a story line that puts performance of the activities into a greater context).

Your ideas for additional activities (or fully formed ones) are welcome! Any submissions used would include full credit to the submitter. The workbook is hosted in a GitHub repository (https://github.com/krwaring/metagenomic-activity-book) to allow for easy distribution of materials, but also for community-sourced improvements via the GitHub pull requests feature.

Table of Contents

Data Cleanup: Contaminant Removal

Learning objectives:

1. Identify the need for contaminant sequence removal step in raw sequence files, and what general types of contaminants need to be removed.
2. Recognize the principles underlying software performance of the contaminant removal step by performing a manual contaminant removal on a simplified .fasta data set.

Level of Difficulty: 2/5

When you receive your raw data from the sequencer, it contains a few types of contaminating sequences that should be removed before proceeding to data analysis. The preparation procedure for DNA sequencing with an Illumina sequencer involves adding adaptor nucleotide sequences to the pieces of DNA being sequenced. These will cause problems for your data analysis, so they should be removed. Additionally, a library of calibration sequences called “phiX” (aka Coliphage phi-X174) are often added to check that the sequencer was working correctly, you’ll have to take those out as well. The contaminant removal step is essentially a search and find mission for types of data that don’t belong in your data files.

To do: Locate the Illumina adaptor and Phix contaminant sequences in the below set of 16 reads and cross them out to remove them from your data set. How many reads remain after filtering? What percentage of the original read set does this amount to?

Illumina Adaptor Sequence

CTGTCT

Phix Sequences

CAAACATTGGGCCAAATGA

GGGCGTTGTATGGTTGCCA

AATACCCCCAGACGTCGGT

TACCGAGTTTCCGATTCGC

READ ID	SEQUENCE
1	CTGTCTATGATAAAAGTTCAATTGA
2	CTGTCTTGCAATGCGCTTATTGAAA
3	CTGTCTGGGCTGTTATGGGTTCA
4	CTGTCTAACGGCTCGGATCCCCA
5	CTGTCTTACCGAGTTTCCGATTCGC
6	CTGTCTATCGGTCAATCTTGGGGGG
7	CTGTCTTTCGGGTCGTATTGTGA
8	CTGTCTTGGTCATGGTTTTGGTGA
9	CTGTCTGGGAGTTCCTCCGATAC
10	CTGTCTTGTAAGGTAAAGAATGGTA
11	CTGTCTCAAACATTGGGCCAAATGA
12	CTGTCTAGTCTTTCTTCTTGGAATG
13	CTGTCTGTAGGTTAGTAAAGCATG
14	CTGTCTACTTGACCCGGGGCAAAGT
15	CTGTCTAATACCCCCAGACGTCGGT
16	CTGTCATGATAGGGTTGCCAAGCAA

# of sequences remaining post-contaminant removal: _____

$$\frac{\text{\# of sequences remaining}}{\text{\# of original sequences}} \times 100\% = \_\_\_\_\_ \text{ \% original read set remaining}$$

Data Cleanup: Contaminant Removal - **Answer Key**

Locate the Illumina adaptor and Phix contaminant sequences in the below set of 16 reads, and cross them out to remove them from your data set.

Illumina adaptor sequences are filled in with light grey below, Phix sequences are filled in with dark grey.

Illumina Adaptor Sequence

CTGTCT

Phix Sequences

CAAACATTGGGCCAAATGA

GGGCGTTGTATGGTTGCCA

AATACCCCCAGACGTCGGT

TACCGAGTTTCCGATTCGC

READ ID	SEQUENCE
1	CTGTCTATGATAAAAGTCAATTGA
2	CTGTCTTGCAATGCGCTTATTGAAA
3	CTGTCTGGGCGTTGTATGGTTGCA
4	CTGTCTAACGGCTCGGATCCCCGT
5	CTGTCTTACCGAGTTTCCGATTCG
6	CTGTCTATCGGTCAATCTTGGGGG
7	CTGTCTTTCGGGGTCTTATAATTG
8	CTGTCTTGGTCATGGTTTTGTTGA
9	CTGTCTCAAAATCTCCTCCGATCT
10	CTGTCTTATAAGGTAAAGAATGGTA
11	CTGTCTCAAACATTGGGCCAAATGA
12	CTGTCTAGTCTTTCTTGGTTCAATT
13	CTGTCTGTAGGTTTAGTAAAGCATG
14	CTGTCTACTTGACCCGGGGCAAAGT
15	CTGTCTAATACCCCCAGACGTCGGT
16	CTGTCTACGATATGGGAGCACAAGCAA

# of sequences remaining post-contaminant removal: **12**

$$\frac{\text{\# of sequences remaining}}{\text{\# of original sequences}} \text{ x } 100\% = \frac{\mathbf{12}}{\mathbf{16}} = \mathbf{75}\% \text{ original read set remaining}$$

Data Cleanup: Quality Filtering

Learning objectives:

1. Identify the need for quality filtering in raw .fastq files.
2. Become familiar with the Phred Quality scoring system used to assign base call accuracies in .fastq files.
3. Recognize the principles underlying software performance of the quality filtering step by performing a manual filter on a simplified .fastq data set.

Level of Difficulty: 3/5

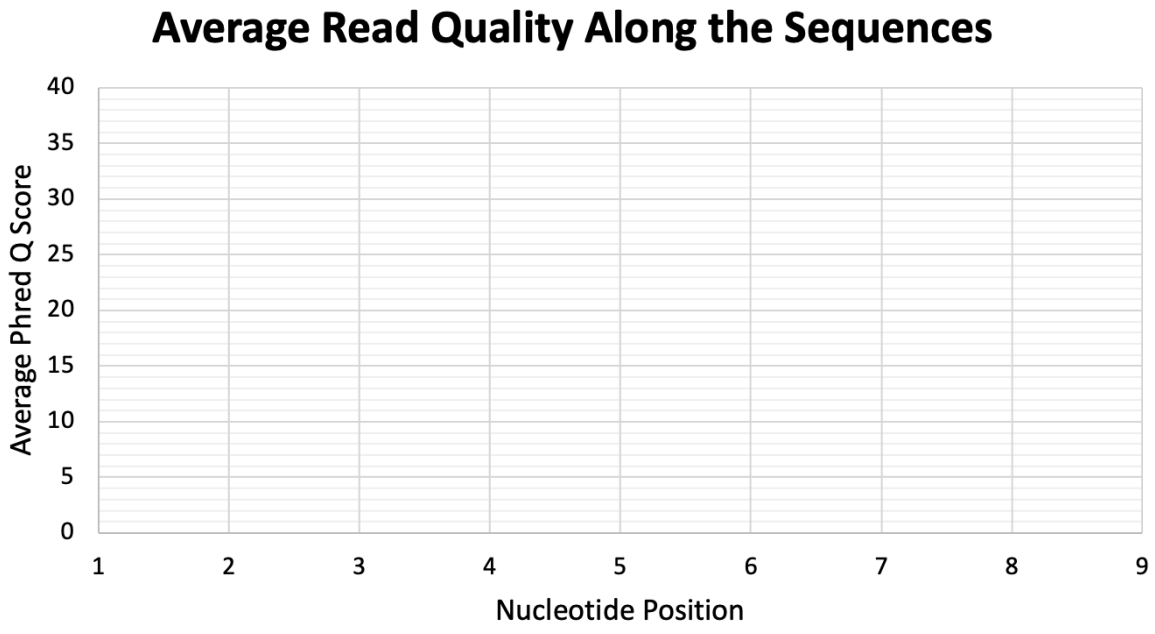
After removing contaminants, you'll need to get rid of the data that Illumina wasn't very sure about in a step called "quality trimming". When an Illumina sequencer makes a base call (decides what nucleotide is present at a certain location in a read sequence), it also assigns a probability of error to the decision (the inverse of which would be the level of confidence). This can be thought of as sort of like Illumina's report card grade for that particular base call. When this confidence level is recorded, instead of writing a number Illumina uses an ASCII (pronounced "ask-ee") character, based on the conversion key shown below. The data selected for removal from the read set are determined based on the data's Phred Quality score (aka Q Score) associated with the ASCII character. Low Phred scores correspond to low confidence base calls by Illumina. The minimum allowable quality score may vary based on the intended application and other context factors. Note: Quality scores are only present alongside the base calls in .fastq files, while .fasta files only contain the base calls.

ASCII character	Phred Q-Score	Probability of correct base assignment	Illumina Report Card Grade	ASCII character	Phred Q-Score	Probability of correct base assignment	Illumina Report Card Grade
"	1	21%	F	6	21	99.2%	A
#	2	37%	F	7	22	99.4%	A
$	3	50%	F	8	23	99.5%	A
%	4	60%	F	9	24	99.6%	A
&	5	68%	F	:	25	99.7%	A
'	6	75%	D	;	26	99.75%	A
(	7	80%	D	<	27	99.80%	A
)	8	84%	D	=	28	99.84%	A
*	9	87%	D	>	28	99.84%	A
+	10	90%	D	?	30	99.90%	A
,	11	92%	C	@	31	99.92%	A
-	12	94%	C	A	32	99.94%	A
.	13	95%	C	B	33	99.95%	A
/	14	96%	C	C	34	99.96%	A
0	15	96.8%	B	D	35	99.97%	A
1	16	97.5%	B	E	36	99.975%	A
2	17	98.0%	B	F	37	99.980%	A
3	18	98.4%	B	G	38	99.984%	A
4	19	98.7%	B	H	39	99.987%	A
5	20	99.0%	A	I	40	99.990%	A

To do:
1) Using the conversion key provided, convert the ASCII characters for sequences 1-4 below into Phred quality score values in the grey boxes. Then plot the average Phred quality score for each base pair position on the provided plot.

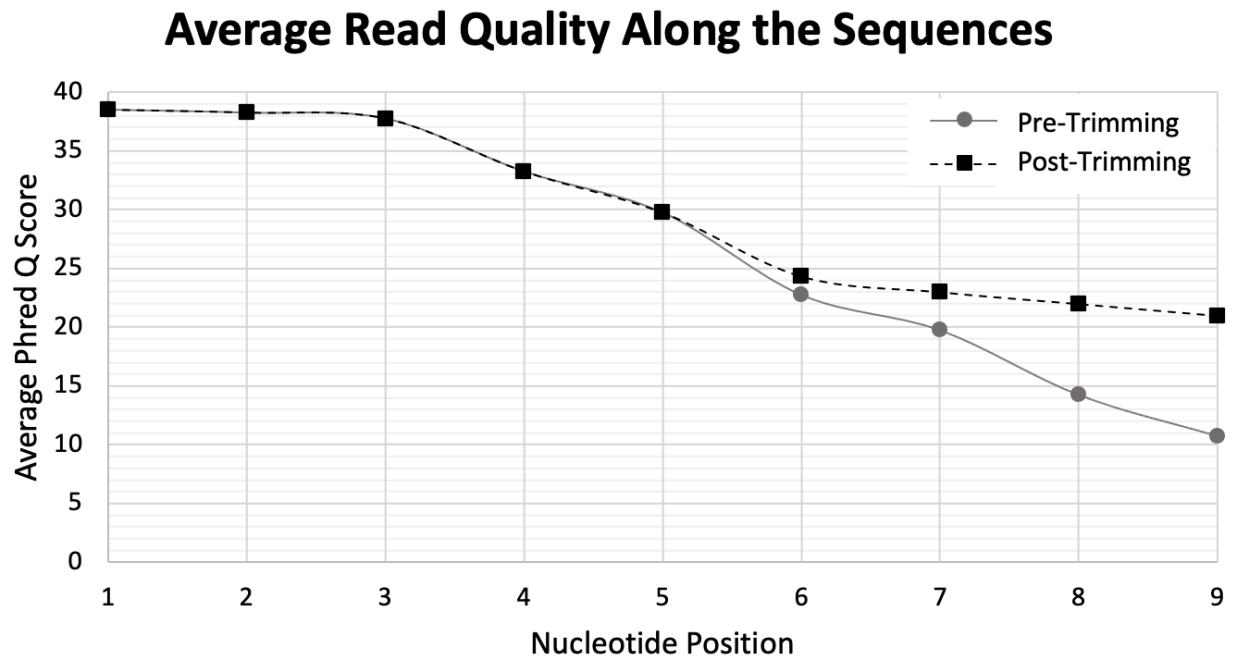
2) Now "trim" any quality scores below 20 by crossing out that portion of the sequence to ensure only high quality data remains. Recalculate and replot the average Q-score values for each position.

	1	2	3	4	5	6	7	8	9	
	T	A	T	T	A	A	T	G	T	Sequence 1
	H	H	F	C	A	<	9	7	6	ASCII characters 1
Quality Score 1										
	A	A	G	C	A	A	G	C	G	Sequence 2
	H	H	I	A	@	:	7	-	,	ASCII characters 2
Quality Score 2										
	A	C	G	T	G	G	C	T	A	Sequence 3
	G	G	F	E	=	6	4	0	+	ASCII characters 3
Quality Score 3										
	A	T	C	A	T	G	C	A	G	Sequence 4
	G	F	F	@	>	3	/	)	"	ASCII characters 4
Quality Score 4										
1) Average Quality Score pre-trimming										
2) Average Quality Score post-trimming										
	1	2	3	4	5	6	7	8	9	Basepair Position



Data Cleanup: Quality Filtering- **Answer Key**

	T	A	T	T	A	A	T	G	T	Sequence 1
	H	H	F	C	A	<	9	7	6	ASCII characters 1
Quality Score 1	39	39	37	34	32	27	24	22	21	
	A	A	G	C	A	A	G	C	G	Sequence 2
	H	H	I	A	@	:	7	-	,	ASCII characters 2
Quality Score 2	39	39	40	32	31	25	22	~~12~~	~~11~~	
	A	C	G	T	G	G	C	T	A	Sequence 3
	G	G	F	E	=	6	4	0	+	ASCII characters 3
Quality Score 3	38	38	37	36	28	21	~~19~~	~~15~~	~~10~~	
	A	T	C	A	T	G	C	A	G	Sequence 4
	G	F	F	@	>	3	/	)	"	ASCII characters 4
Quality Score 4	38	37	37	31	28	~~18~~	~~14~~	~~8~~	~~1~~	
1) Average Quality Score pre-trimming	39	38	38	33	30	23	20	14	11	
2) Average Quality Score post-trimming	39	38	38	33	30	24	23	22	21	
	1	2	3	4	5	6	7	8	9	Basepair Position



Assembly

Learning objectives:

1. Identify the most basic elements underlying the *de novo* assembly process.
2. List potential strategies and complications involved in the software performance of the *de novo* assembly process.

Level of Difficulty: 3/5

Once you've got a filtered set of reads, you may decide you want to piece them together to try to reassemble the genome(s) they came from. This process is kind of like if you were to piece back together a newspaper that had been run through a shredder. What strategies might you use while reassembling? Say you've taped back together as much as you can of the original newspaper, but because the pieces were so small and the shredder ate some of them, the best you can do is several patchy reassembled sections that you can't connect together, and you're left with a bunch of repetitive words and characters like "a", "the", and commas. The sections of assembled pieces are like metagenomic "contigs", and those repetitive leftovers may still be useful down the road in read mapping.

Follow up questions:

- How are nucleotides different from words in a newspaper or song? What context clues in DNA might be used instead of common words, sentence structure, punctuation, and font?
- What other clues in the DNA might guide a metagenomic assembly?
- Why do you suppose some pieces might be missing in metagenomic assembly?
- How might repeated genes affect the assembly process?

To do:
A few copies of two common nursery rhyme songs representing different microbial genomes have been cut up into pieces, and some of the pieces may be missing. Try to reassemble as much of the rhymes as you can from the following list of “reads”. Note that a few areas of the rhymes have repeated words, and both rhymes loop back on themselves at the ends, kind of like a circular genome.

-vous? dorm
? Frere J
am row r
atines son
cques Fre
dream r
e is but
eam ro
eam row r
errily merrily
es dorm
es matin
g dong d
g dong Fr
ines ding d
ing din
ing don
m merrily merr
merrily mer
merrily mer
ng ding
ng ding d
nnez les
onnez le
ous? Dor
ous? Fre
ous? sonn
r boat ge
rere Jacq
rilly merrily
rmez-vo
rmez-vou
row ro
row row
row your
rrily lif
t gently
ut a dre
w row r
w your bo
wn the s
y down t
y life is
z les mat