Table of Contents:

1. Processing of FastQ Files:

Ion Assist was originally developed to process fasta files that were output from CLC Genomics Workbench. However, we felt the need to process the FastQ files based on their Phred scores to better inform us on the quality of our Ion Torrent data. The processing of these files is not limited to Ion Torrent data, but is intended for the use with Ion Torrent data. To begin the process, you will need a FastQ file and a mapping file detailing the information present in the FastQ files. The process is intended to utilize files with sequences that have been barcoded.

The mapping file should be a tab delimited text file, best made using Microsoft Excel. The format of the files should be that the first line of the file is the header line. The first column should contain "#SampleID", and the second column should contain "BarcodeSequence". You can include up to 98 additional columns with whatever information you like. For example, the 3rd column could be "Age", the 4th could be "Sex", and the 5th could be "BarcodeNumber". There can be up to 1000 rows in each mapping file. We encourage that all the text be in all caps to avoid errors, however, the process may work even when all caps is not used.

Example of a valid mapping file:

#SampleID	BarcodeSequence	LinkerPrimerSequence	BarcodeNumber
Blank1	CTAAGGTAAC	CCTACGGGAGGCAGCAG	1
Blank2	TAAGGAGAAC	CCTACGGGAGGCAGCAG	2
Blank3	AAGAGGATTC	CCTACGGGAGGCAGCAG	3
Blank4	TACCAAGATC	CCTACGGGAGGCAGCAG	4
Blank5	CAGAAGGAAC	CCTACGGGAGGCAGCAG	5
Blank6	CTGCAAGTTC	CCTACGGGAGGCAGCAG	6
Blank7	TTCGTGATTC	CCTACGGGAGGCAGCAG	7
Blank8	TTCCGATAAC	CCTACGGGAGGCAGCAG	8
Blank9	TGAGCGGAAC	CCTACGGGAGGCAGCAG	9
Blank10	CTGACCGAAC	CCTACGGGAGGCAGCAG	10
Blank11	TCCTCGAATC	CCTACGGGAGGCAGCAG	11
Blank12	TAGGTGGTTC	CCTACGGGAGGCAGCAG	12
Blank13	TCTAACGGAC	CCTACGGGAGGCAGCAG	13
Blank14	TTGGAGTGTC	CCTACGGGAGGCAGCAG	14
Blank15	TCTAGAGGTC	CCTACGGGAGGCAGCAG	15
Blank16	TCTGGATGAC	CCTACGGGAGGCAGCAG	16

We recommend using File-->Fastq Processing-->Find and Distribute MIDs with Additional Info to process fastq files. You will be prompted to input quality parameters including Minimum Length, Maximum Length, Maximum Homopolymers, Quality Score Cutoff, Remove Barcode if less than X reads, and to check whether or not to remove primer sequences. Unless your primer sequence has an exact match in the read being examined, it will not be removed. Depending on the speed of your system and the size of the Fastq file, expect the processing to take between 10 minutes and several hours. The output is a log file highlighting the characteristics of the reads processed, individual fasta files for each #sampleID used, and an FNA file containing all the processed reads, which can be read directly into QIIME.

If you choose to use File-->Fastq Processing-->Find and Distribute MIDs, you will not be provided with a detailed log file.

We do not recommend using any analyses under File-->Find and Distribute MIDs or File-->Find and Distribute MIDs Large File, as this was specifically designed for when barcode labeling goes seriously awry.

File-->Fastq Processing-->Find and Distribute MIDs-->Trim Reads Without Primers And Barcodes was designed to process reads that do not have any primers or barcodes. A mapping file is not required for this process.

File-->Fastq Processing-->Find and Distribute MIDs-->Combine Fastq Files does exactly what one would expect.

File-->Fastq Processing-->Find and Distribute MIDs-->Split FNA Files into Fasta files can take a file designed for QIIME and break it up into individual Fasta files for each sample.

File-->Fastq Processing-->Find and Distribute MIDs-->Split Fastq File using mapping file can be used when 2 separate projects have been sequenced in the same run. It will create a separate Fastq file for whatever project you designate in the mapping file. Useful for submissions to the Sequence Read Archive.

2. Creating Random 16S Datasets

This protocol can be used for any datasets and are not really specific to 16S amplicon sequencing. Simply choose a set of Fasta files, enter the number of random reads you want from each dataset and the number of datasets you would like to create. By checking the box for name checking, the protocol will rename any reads with identical names automatically.

3. Merging Files

There are many occasions where files need to be merged, whether they are Fasta or text files. To merge files, go to File-->Merge Files-->Batch. This allows you to merge multiple files into a single file. The original files will be retained, and you can give the merged file a name. The function will error out if you try to merge files that are too large. Generally files >150mb begin to cause crashed. For large files, we recommend you append them rather than merge together. File-->Merge Files-->One at a time, should be rather obvious.

To merge fastq files, see section 9.

4. Appending Files

To append files, go to File-->Append Files-->Normal File. This function is good when trying to merge large files. Choose this function, and always append the smaller file to the larger file. Choose the larger file first, then choose the smaller file to append.

To append larger files, go to File-->Append Files-->Large File is useful when putting 2 large files together. This process will take much longer than when merging smaller files.

5. Splitting Fasta Files

Occasionally you need to split a file into multiple files. Is useful for Fasta files, but theoretically, you could split other types of files as well. We use this function when we have a lot of sequences to blast, and want to split the work across multiple computers.

Go to File-->Split MID File-->. Then choose which option works best. You can choose to keep the names of the sequences intact. If you do not choose to keep them intact, the sequences will be numbered in lieu of names.

6. Read Statistics

To gather statistics on reads in a fasta file, go to File-->Read Statistics. The output generally gives you mean length, mean GC content, and number of reads for each file you choose. Choose the option with Tabs to allow for easier manipulation with Microsoft Excel. You can also do contig statistics. It doesn't really matter whether you use reads or contigs for these analyses, as the software really doesn't care. You can also get the data necessary to make box and whiskers plots with microsoft excel. N50 and Maximum lengths also can be determined for contigs using these functions.

7. Rename Sequences

Will work on both reads and contigs. Is very useful to rename a group of reads or contigs quickly. Go to File-->Rename Sequences. You will be prompted to open up a file or a group of fasta files. You then will be prompted for each individual file to give the reads a name. For example, you can call your reads MouseHeartRead, and the routine will name the reads MouseHeartRead_1, MouseHeartRead_2, etc... We use this routine all the time if we inadvertently have a typo in our readnames, or need to change the names of the reads to run different analyses.

8. Create Fastq files for NCBI Sequence Read Archive

For those who don't submit the Sequence Read Archive very often, submitting can be a tall order, as the interface is not user friendly nor intuitive. When submitting pooled amplicon sequences, it is especially problematic, as they prefer you submit individual fastq files that represent each separate barcoded amplicon. This routine will take your original fastq file along with the mapping file (see section 1 for creating mapping files), and will create fastq files that represent each barcode. These can easily be uploaded to the SRA. Of note, the settings for creating these files is limited to those applicable to Ion Torrent data, so it may or not work on other data types. To create these fastq files go to File-->Fastq processing--> Find And Distribute MIDs to FastQ Files For SRA.

9. Merge Fastq Files

Merging fastq files probably does not come up very often, but in our experience we sometimes sequence different samples and need to go back later and sequence additional reads. These fastq files can be merged later by going to File-->Fastq processing--> Combine FastQ Files.

10. Parsing Blast Files

We do a lot of parsing of the results of blast files for viromes. These routines are specific for parsing viromes, it does so by preferentially identifying virus hits and assigning them to reads/contigs even if there are hits with better E-scores that belong to bacteria. We usually parse at E-scores of 10-5. To perform this analysis, go to Parse-->Parse and Categorize Blast Files Based on Escore Ten-5. The output is a group of files, including a file with all phage hits, all bacteria hits, all human hits, and all other hits. It also puts together a taxonomy file, which probably will not be helpful, and it puts together a parsed file with all hits.

11. Taxonomy from Parsed Blast Files

To get taxonomic compositions from blast files that have been parsed, go to Analysis Phages-->Load Phage File-->To Analyze Taxonomy. This routine will ask for your parsed phage file, and take that file and compile the taxonomy. This routine also can be performed with total numbers or reads corrections or coverage corrections, but requires a CSV file (usually obtained from CLC genomics, but could be made from other programs) to figure out how to perform the corrections. The format of the CSV file is below:

Name	Consensus length	Total read count	Single reads	Reads in pairs	Average coverage
contig_1_mapping	1704	2421	2421	0	253.8732394
contig_2_mapping	15608	32154	32154	0	387.4419528
contig_3_mapping	340	7	7	0	3.702941177
contig_4_mapping	907	20	20	0	3.712238148

The output gives you a Class, Family, Genus, Order, and Phylum File.

12. Compiling Parsed Taxonomy Data

To compile parsed taxonomic data, go to Reformat-->Make Table of Compiled Taxonomic Data. The output is easily usable for figures using Microsoft Excel. The output is essentially a heatmap, and can be used by programs such as Java Treeview or MEV (Multi Experiment Viewer) for visualization.

13. Normalize Heatmap File

Compiled taxonomic data can be normalized, which are helpful for heatmap visualization. Go to Reformat-->Normalize Heatmap File and input the compiled taxonomy data file and it will normalize for better heatmap visualization.

14. Removing Human Contamination

We sometimes get reads in our viromes with hits to human chromosomes. The vast majority of these hits are clonal, and don't likely represent contamination, but we remove them nonetheless. To perform this analysis, we take our reads and do a BLASTN against the human database, and take the results and use in our functions. To remove the reads that had hits to the human database, go to Parse-->Parse

Blasts To Remove Hits From Fasta File. Input the BLASTN results and the Fasta File, and will output a file that has removed the reads with human blast hits.

15. Removing Contaminant Reads

To remove contaminant reads from a sample, place the contaminant read sequences into a separate fasta file. For example, if your contaminant reads are Bradyrhizobia, place bradyrhizobia reads from a database such as greengenes into a separate file. Go to Analysis Bacteria-->Studies-->Campus Antibiotics Study-->Remove Bradyrhizobiaceae Reads. This routine can remove any reads, they don't have to be Bradyrhizobiaceae. You will be prompted to open up the contamination source file, and then to open the file with your reads, and the output file will be a Fasta file with the reads removed.

16. Homologous Virus Diversity Index

The HVDI is somewhat complicated to perform, but here are a set of instructions to allow for anyone to do it. You will need standalone BLAST to create the blast files necessary to perform the analysis. You also will need a CSV file (Usually from CLC genomics Workbench) that provides the contig assembly data, but you can make these files yourself from the output of whatever assembler you choose to use (See Section Taxonomy from Parsed Blast Files). To perform the HVDI, follow these steps:

1. Assemble your reads (We recommend CLC Genomics Workbench)

2. Save your reads as a fasta file and the assembly data as a CSV file (Also recommend CLC Genomics Workbench for the CSV file)

3. Use standalone blast to create a blast database from your file:

```
makeblastdb -dbtype nucl -in MyFile.fa -out MyDatabase
```

4. Use standalone blast to blast the reads against themselves

```
blastn -evalue 0.0000000001 -num_descriptions 10000 -num_alignments 0 -query MyFile.fa -db MyDatabase -out MyFileVsItselfBlastN.txt
```

5. Go to Analysis Phages-->Chemostat Study-->Homologous Virus Diversity Index

You will be prompted to open up the CSV file, then the associated blast file you just created, and then to enter the number of singletons from the assembly process. After you input this information, you will receive several files with the results. The data from the corrected Chao1 Index are currently not correct, and I am working to fix that bug. The output gives you the number of Shannon Index values corrected and uncorrected. The Uncorrected data just represent the Shannon Index, and the other data represent the HVDI. Occasionally, no corrections are needed and both index values are the same.

*Versions of blast output change all the time. The blast files work only with version 2.2.31 of the NCBI standalone blast program.

17. Homologous Virus Diversity Index Rarefactions

You can also perform rarefactions with the HVDI using similar methodology. Create all the same files that were created in the description of the Homologous Virus Diversity Index. Go to Analysis Phages-->Stool Virome Study-->Homologous Virus Diversity Rarefaction. You can do this for a single sample or multiple samples simultaneously. The process of creating these rarefactions is time consuming. Of note, the Chao1 index is working accurately for this routine.

18. Other Routines

Ion Assist is meant to do many other things. Just a quick tour through the menus will indicate that. Most of these functions were never meant for public consumption, so use at your own risk.