# BIOL 354W - Research Methods in Advance Microbiology

**Version 12**

**Rosa Leon**

## Abstract

This protocol series will guide students through the experience of analyzing metagenomic data.

**Citation:** Rosa Leon BIOL 354W - Research Methods in Advance Microbiology. **protocols.io** dx.doi.org/10.17504/protocols.io.ngfdbtn
**Published:** 01 Mar 2018

## Protocol

### Introduction to BIOL 354W, sequencing data and bioinformatics

**Step 1.**

BIOL 354W Jan 16th

BIOL 354W Jan 18th

### Command line tutorial

**Step 2.**

In order to do bioinformatics, we first need to become comfortable using the computational language and basic skills that will allow you to analyze data.

Open this link in Chrome

LINK:
http://rik.smith-unna.com/command_line_bootcamp/?

NOTES
**Marcia Smith** 29 Jan 2018

change to:

In order to do bioinformatics, we first need to become comfortable using the computational language and basic skills that will allow you to analyze data.

### DNA quality assessment and assurance

**Step 3.**

he first step in analyzing the sequencing data set is to asses the quality of the sequence, and then to edit the dataset in order to retain only the highest quality sequences for the following analysis.

To this end we will use: FastQC - A high throughput sequence QC analysis tool

Familiarize your self with the software by looking at their web page - check out the video tutorial!

cmd COMMAND

```
scp -r username@bio-
server-2.willamette.edu:/home/username/folder_with_fastqc_file ~/Desktop/
```

Now that the software has run and you have folders and files with date, you should look at the data to assess the quality and make decision about the quality control step that we will work on next. For this you can unzip you folder where there will be detail information about the results, as well as a summary of the run. You can also download the .html file to look at the graphic representation of the run, the same format you experienced on the fasqc web and tutorial

NOTES

**Rosa Leon** 14 Jan 2018

You can perform the fastqc file on .fastq files and also in .fastq.gz files or compressed files

**Rosa Leon** 30 Jan 2018

This step most be done from a Terminal window that is looking at your own computer and not conected to the sever

**Marcia Smith** 29 Jan 2018

Change to:

The first step in analyzing the sequencing data set is to asses the quality of the sequence, and then to edit the dataset in order to retain only the highest quality sequences for the following analysis.

## Assuring DNA sequencing quality using Trimmomatic

**Step 4.**

Trimmomatic: A flexible read trimming tool for Illumina NGS data (Website )

Description

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended

data.The selection of trimming steps and their associated parameters are supplied on the command line.

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

cmd COMMAND

```
 java -jar /opt/BioInfo_tools/Trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 5 -
phred33 input_forward.fq.gz input_reverse.fq.gz output_forward_paired.fq.gz output_forward_
unpaired.fq.gz output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:/opt/
BioInfo_tools/Trimmomatic-0.36/adapters/TruSeq3-
PE.fa:2:30:10 LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:36
```

input_forward.fq.gz = " the exact name of your forward or R1 sequence file" input_reverse.fq.gz = " the exact name of your forward or R2 sequence file" output_forward_paired.fq.gz = write in what you would like the output to be called Eg. 3A_trimmed_R1_paired.fastq.gz" output_forward_unpaired.fq.gz = write in what you would like the output to be called Eg. 3A_trimmed_R1_unpaired.fastq.gz" output_reverse_paired.fq.gz = write in what you would like the output to be called Eg. 3A_trimmed_R2_paired.fastq.gz" output_reverse_unpaired.fq.gz = write in what you would like the output to be called Eg. 3A_trimmed_R2_unpaired.fastq.gz" Try to run this command as it is with quality of Q15 (SLIDINGWINDOW:4:15) as currently stated in the command and then with Q30 (SLIDINGWINDOW:4:30). Record the number % of out put sequences per each.

## Metagenomic assembly

**Step 5.**

To assemble our metagenomes we will try two differnet assemblies and compare them.  We will try IDBA_UD and Megahit assemblies. These is going to be one of the most time intensive process that we will do in the class.

Megahit github - https://github.com/voutcn/megahit/

Megahit article - https://academic.oup.com/bioinformatics/article/31/10/1674/177884

IDBA_UD - https://github.com/loneknightpy/idba

IDBA_UD article - https://academic.oup.com/bioinformatics/article/28/11/1420/266973

cmd COMMAND

```
/opt/BioInfo_tools/idba/bin/idba_ud -r merged_reads.fa -o output_dir --num_threads 5
```

Once the read files are converted into fasta and in consecutive order then the assembly can be run
merged_reads.fa = your new generated merged fasta sequences files exactly as you called them
output_dir = a folder to store the assembly output, you choose the folder name

## Assessing the quality of the assemblies

**Step 6.**

We can investigate assembly statistics to compare which assembly is best between the two assemblies utilized. For this we can use a software called Quast.

Metrics based only on contigs:

- Number of large contigs (i.e., longer than 500 bp) and total length of them.
- Length of the largest contig.
- N50 (length of a contig, such that all the contigs of at least the same length together cover at least 50% of the assembly).
- Number of predicted genes, discovered either by GeneMark.hmm (for prokaryotes), GeneMark-ES or GlimmerHMM (for eukaryotes), or MetaGeneMark (for metagenomes).

cmd COMMAND

```
/opt/BioInfo_tools/quast/metaquast.py contig.fa --gene-finding -t  5
```

QUAST evaluates genome assemblies by computing various metrics.

## Binning assembled metagenomes with MaxBin

**Step 7.**

MaxBin is a software for binning assembled metagenomic sequences based on an Expectation-Maximization algorithm.

Users provide the assembled metagenomic sequences and the reads coverage information or sequencing reads. MaxBin will report genome-related statistics, including estimated completeness, GC content and genome size in the binning summary page.

MaxBin article - https://academic.oup.com/bioinformatics/article/32/4/605/1744462

cmd COMMAND

```
perl /opt/BioInfo_tools/MaxBin-2.2.4/run_MaxBin.pl -contig "assembled_contigs.fa" -
reads "interleaved reads fasta" -out "out directory" -thread 5
```

MaxBin requires the assembled contains file and also the file that contains the sequence reads
assembled_contigs.fa = your contigs file (remember to add the full path if you are in a different directory) concatenated reads fasta = the path to your reads, these reads most all be in one file

and concatenate ( or paired R1 followed by R2 reads). This you can get from your IDBA fq2fa run out directory = a directory that you create to save your bins

## Assessing the quality of your bins via CheckM

**Step 8.**

Checkm article - http://genome.cshlp.org/content/25/7/1043

Also check out the websit for information on CheckM - CheckM website

Befor running Checkm the software pplacer must be included in the PATH by addind

export PATH="/opt/anaconda3/bin:$PATH" to the .bashrc file in your home directory under the

# User specific aliases and functions section.

cmd COMMAND

```
/usr/bin/checkm qa lineage.ms . -o 2
```

This command will help you generate an expanded information table about each of your bins. Run this command from within the directory where your checkm data is located copy the table that this command generated onto an excel sheet and analyze to then run VizBin

## Use VizBin to further curate your bins

**Step 9.**

VizBin is a java software that calculates kmer composition and creats a pictographical output that shows the similatity between contigs realted to how close they are postition to each other. We will use VizBin to help us de-contaminate out bins

VizBin will generate a visualization window. Each point represents a genomic fragment (by default of length >= 1,000nt). VizBin is designed with the user in mind. All that is needed is a fasta file containing the sequences of interest. A step-by-step guide on using VizBin - including a description of loading the data, selecting points, and exporting the sequences represented by the selected points - is provided on the tutorial page of VizBin's github wiki

In order to run VizBin with you data you must download you bins fasta files onto your desktop.

To download go to the VizBin page

## Perform taxonomic identification using Phylosift

**Step 10.**

Phylosift software searches for single copy marker genes and finds thier taxonomic classification

Before running this command take a moment to learn about the sorftware at the Phylosift webpage

cmd COMMAND

```
/usr/local/phylosift_v1.0.1/bin/phylosift all your_bin.fasta --threads 3
```

To run Phylosfit you only need to have change your_bin.fasta for the files (and path if required) for each of your individual bins

## Prokka - software for annotations

**Step 11.**

Learn about how to set up a prokka run and what the outputs are by looking at the git hub prokka webpage

cmd COMMAND

```
/opt/BioInfo_tools/prokka-1.11/bin/prokka contigs.fasta
```

We will annotated our curated bins using PROKKA

## Compare genomes to various databases

**Step 12.**

In order to assess the metabolic potential of you Metagenome Assembled Genomes (MAGs) we will compare their predicted proteins against a few different databases. These databases will provide information about what pathways or protein groups your annotated proteins belong to. This will help you assess what kind of metabolic potential your organsims possess.

We will start by taking our annotated proteins and running it in the BlastKoala web platform. http://www.kegg.jp/blastkoala/

Use your PROKKA.faa file to copy the protein annotations and past on the box label Enter FASTA sequences or upload the PROKKA.faa file. Add you email so they can keep you update on the progress of your analysis.

## Compare genomes to various databases

**Step 13.**

In order to run the next few steps we need to add another set of software to our path

cmd COMMAND

```
nano .bashrc

##copy and paste

User specific aliases and functions
export PATH=$PATH:/opt/ncbi-blast-2.7.1+/bin/

## Save file changes by "control + O" and then "control + X", then close the window and log
```

```
in again to the server
```

This step is crucial to successfully run the next few steps

## Compare genomes to various databases

**Step 14.**

Compare annotated proteins to the Cluster of Orthologous Genes (COG)

cmd COMMAND

```
perl /opt/BioInfo_tools/cdd2cog.pl -r output_file.out -c /opt/BioInfo_tools/COG/cddid.tbl -
f /opt/BioInfo_tools/COG/fun.txt -w /opt/BioInfo_tools/COG/whog -a
```

Once we have generated a blast output, which provides a comparison of our annotated bins to the COG database we can use the cdd2cog perl script to count and parse that information for us