BIOL 354W - Research Methods in Advance Microbiology

Version 3

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.mvec63e

Published: 29 Jan 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 bioinfomatic, we first need to get confortable using the computational language and basic skills that will allow you to

Open this link in Chrome

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

ANNOTATIONS

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.

The first step to analyzing a sequencing dataset is to assess what is the quality of the sequence information and to edit your data set to retain only the highest quality sequences for all analysis that will follow.

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

ANNOTATIONS

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.

Marcia Smith 31 Jan 2018

After running this scp ~/Desktop/ command, you can actually just navigate to your desktop (as you would normally do while NOT working in terminal) and view the files. They will be available on your computer's desktop!

Assuring DNA sequencing quality using Trimmomatic

Step 4.

Trimmomatic: A flexible read trimming tool for Illumina NGS data (Website - http://www.usadellab.org/cms/?page=trimmomatic)

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 -
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:TruSe
q3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Try to run this command as is with a Q15 and then with Q30. Record the number % of out put sequences per each.

ANNOTATIONS

Marcia Smith 31 Jan 2018

A FEW THINGS TO NOTE

This command is missing a few phrases. It is also impossible to c/p for some reason. This is a copy of my Trimmomatic command for my data:

```
java -jar /opt/BioInfo_tools/Trimmomatic-0.36/trimmomatic-0.36.jar PE -phred33
36_48_CCTTCA_L007_R1_001.fastq 36_48_CCTTCA_L007_R2_001.fastq 001_forward_paired.fq.gz
001_forward_unpaired.fq.gz 001_reverse_paired.fq.gz 001_reverse_unpaired.fq.gz
ILLUMINACLIP:/opt/BioInfo_tools/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Notice that my data files are the 36_48_CCTTCA...... and I am calling my files 001.

So, basically you need to insert into your terminal window:

java -jar /opt/BioInfo_tools/Trimmomatic-0.36/trimmomatic-0.36.jar PE -phred33 [**YOUR FORWARD READ FILE NAME]**.fastq [**YOUR REVERSE READ FILE NAME]**.fastq [**whatever you want to call these files]**_forward_paired.fq.gz [**whatever you want to call these files]**_forward_unpaired.fq.gz [**whatever you want to call these files]**_reverse_paired.fq.gz [**whatever you want to call these files]**_reverse_unpaired.fq.gz ILLUMINACLIP:/opt/BioInfo_tools/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Metagenomic assembly

Step 5.

To assemble our metagenomes we will try two differnet assemblies and compare them. First we will try IDBA_UD and Megahit assemblies.

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/idba_ud -r merged_reads.fa -o output_dir
```

Once the read files are converted into fasta and in consecutive order then the assembly can be run

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
```

QUAST evaluates genome assemblies by computing various metrics.

Binning assembled metagenomes

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.fa" -
reads "concatenated reads fasta" -out "out directory"
```

MaxBin requires the assembled contains file and also the file that contains the sequence reads

Assessing the quality of your bins via CheckM

Step 8.

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