### MetaDNA\_RNA\_analysis workflow

### 2019. Nov.4th

### Part i:

### Instructions on how to get access to Genomics Compute Cluster:

<https://www.it.northwestern.edu/research/user-services/quest/genomics.html>

**Part ii:**

**A summary on the softwares used in the workflow:**

**Keep a record of your sequencing strategy, Sequencing strategy:**

DNA-seq insert size (not counting adapter) ?bp; PE length ?bp,; sequening depth: ? reads or ? total bp reads

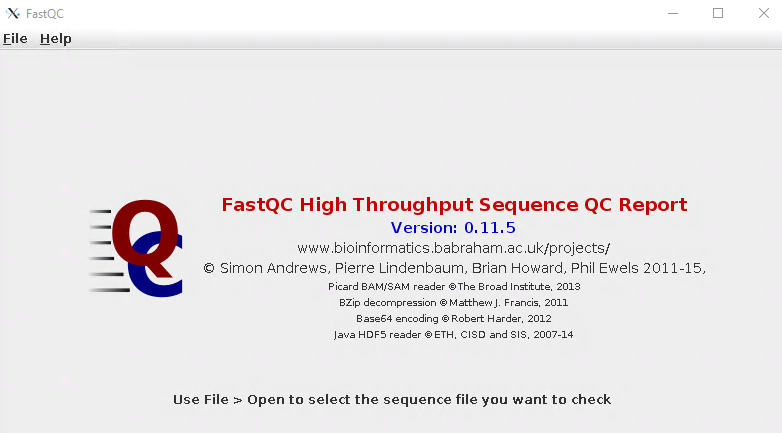
1. ***Fastqc---- to check the quality of the metagenome sequence data***

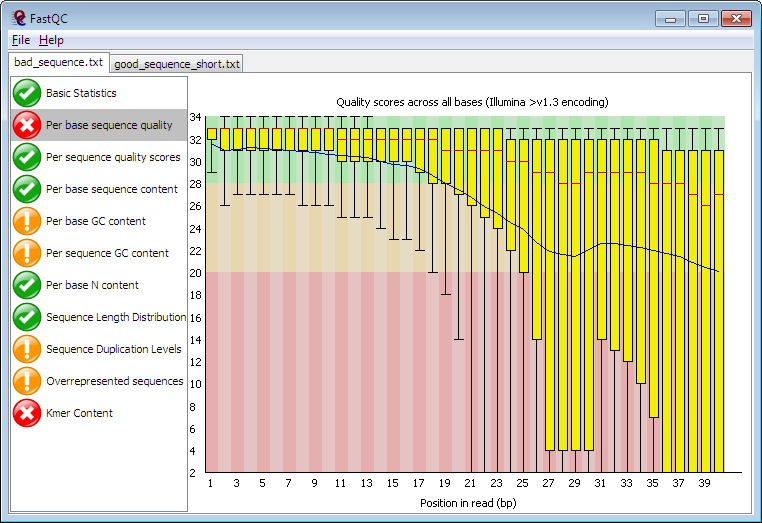
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

fastqc is already installed in quest, to use it, type “module load fastqc”, and then type “fastqc”, to open the graphical (interactive) interface



Below is the fastqc graphical interface that would appear, refer to the bottom line, open the fq file and visualize the quality of the sequencing data returned





1. ***Fastx ---- there are a number of tools in fastx, they include but are not limited to: converting fq file to fa file, trim the fq sequences at certain length according to your observations on the quality of your sequences (e.g.*** Qc > 28)

<http://hannonlab.cshl.edu/fastx_toolkit/>

I have installed fastx in my home folder on quest: /home/ywx1845/software/fastx/

You may not have the permission to get access to the files in my home folder, and may need to install fastx to your own home folder following the installation instructions in the link above; and modify the path to the fastx software in the submission .sh files below

The two submission scripts I used in this step have been uploaded in the box folder, and they are:

*2.1\_fastx\_trim\_sequence\_at\_certain\_length.sh*

*2.2\_fastx\_convert\_fq\_to\_fa.sh*

1. ***Assembly ---- assemble metagenome DNA raw reads to longer contigs***

There are several assembly softwares, among which include IDBA, megahit, metaspades;

I would recommend IDBA and metaspades;

The benefit of IDBA over metaspades is that it support co-assembly and requires less memory;

the benefit of metaspades is that its performance is better than IDBA when assembly DNA reads from one single metagenome dataset (not co-assembly);

metaspades require huge memory and the assembly cannot finish in 48 hours in our genomics node, to use metaspades, email NU computer center, specify your needs; in my case, they have helped assigned to me the access to node b1054 which has 1T memory and does not have time limitation for one single job submitted.

Megahit: <http://www.metagenomics.wiki/tools/assembly/megahit>

IDBA: <https://github.com/loneknightpy/idba>

Metaspades: <http://bioinf.spbau.ru/en/spades3.7>

The submission scripts I used in this assembly step have been uploaded in the box folder, and they are:

3.1\_IDBA\_single\_assembly.sh

3.1\_IDBA\_coassembly.sh

3.2\_metaspades\_assembly.sh

3.3\_megahit\_assembly.sh

1. ***Evaluate the assembly results using quast***

<http://bioinf.spbau.ru/quast>

The submission scripts I used in this assembly step have been uploaded in the box folder

4\_QUAST\_assembly\_evaluation.sh

1. ***Reads mapping***

Bowtie mapping:

<http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml>

1. to generate the coverage and length info of each contig assembled;
2. the coverage info generated could be used to calculate a) the abundance of each genome bin obtained; and b) the ratio of reads used to generate the contigs assembled; and c) the DNA/RBA RPKM values of each gene or genome (refer to below in green on the principles in the calculation of the RPKM values of genomes and genes, so far, there is no software available for the RPKM calculation, so we need to write script wither in python or R to do such calculation)

The submission scripts I used in this reads mapping step have been uploaded in the box folder:

5.1\_DNA\_bowtie\_mapping.sh

5.2\_RNA\_bowtie\_mapping.sh

Map the DNA reads and RNA reads to the contigs, respectively\

# Why RPKM? Normalize the length of the contigs or the length of the genes, for example, gene A and gene B have the same copy number in a community or in a genome, if gene A is longer than gene B, a larger number of reads would be mapped to gene A, while if the length of the two genes are considered as in the PRKM calculation, the RPKM value of the two genes would be the same.

# <https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/>

# <https://www.biostars.org/p/68126/>

Machine generated alternative text:
RPI<M/FPKM 
•RPKM, Reads Per Kilobase of exon model per Million mapped reads (single 
reads) 
•FPKM, Fragments Per Kilobase of exon model per Million mapped reads 
(paired-end sequencing) 
Number of reads (pairs) aligned 
RPKM(FPKM) = 
length of transcript 
Toal reads (pairs) number 
a 
Low 
Short transcript 
Ikb 
High 
Long transcript 
1 
2 
3 
10 
4 

**Reference for the proposal of RPKM quantification**

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 5, 621–628. doi:10.1038/nmeth.1226

For Metatranscriptome, if you want to also consider the abundance of a gene within the community.

Machine generated alternative text:
RPKM-DNA = 
Equation S2 
RPKM-RNA 
RPKM - RNA 
RPKM - DNA 
number of mapped cDNA reads 
o transcri t 
otanu ero 
1 kb 
from metagenome 
in meta enome 
c 
re 
1 
106 
number of mapped mRNA reads from metatranscriptome 
— lengt o transcri t 
Ota number o m NA reads in metatranscriptome 
1 kb 
Equation S3 
MRPKM= 
x 
106 

DNA-RPKM: represents the relative abundance of a gene (ORF) within the community

RNA-RPKM: stands for the total transcription of that gene (ORF) within the community

MRPKM: represents the absolute transcriptional activity of a gene (ORF) within the community

Reference:

Xia, Y., Wang, Y., Fang, H.H.P., Jin, T., Zhong, H., Zhang, T., 2014. Thermophilic microbial cellulose decomposition and methanogenesis pathways recharacterized by metatranscriptomic and metagenomic analysis. Sci. Rep. 4, 6708. doi:10.1038/srep06708

And a lot of other papers applied similar ideas but used different terms

The R script named as “clean\_RNA\_RPKM\_values\_of\_gens\_in\_genomes\_htseq\_processing.R” was the script used in RNA-RPKM calculation based on the htseq results generated from the 6 “run.htseq….sh”

1. ***Genome binning***

The basic idea underlying the genome binning is differential coverage of contigs in different samples (that is why you need to map your contigs to raw reads of at least two different samples to get the corresponding \*sorted.bam files with differential coverage info);

the contigs from one genome are supposed to be of similar coverage info and similar GC content and will be clustered together

below is one visualization generated in mmgenome binning in R, you may refer to this visualization to have a general idea on the underlying principle

A close up of a map

Description automatically generated

The tools I have used from genome binning: metabat and mmgenome

metabat : <https://bitbucket.org/berkeleylab/metabat/src/master/>

mmgenome: <http://madsalbertsen.github.io/mmgenome/>

my evaluation is that metabat is more user friendly (automatic)and much faster than mmgenome and metabat could generate a larger number of draft genome bins than mmgenome;

you can also try mmgenome if you are familiar with how R works, you have more control on the quality of your genome bins with mmgenome, but it is more time-consuming;

The submission scripts I used in genome assembly using matabat have been uploaded in the box folder;

#step1. Generate depth file from the two .sorted.bam files

6.1\_genome\_binning\_step1\_metabat.sh

#step2. Recover genome bins from the depth file generated in the 1st step

6.2\_genome\_binning\_step2\_metabat.sh

If you want to try mmgenome, you need to work with R as summarized in the link above, there is no submission scripts needed

1. ***Check the quality of the draft genome bins you have obtained***

**CheckM:** <https://ecogenomics.github.io/CheckM/>

Check the completeness, contamination and strain heterogeneity of the draft genome bins you recovered from the metagenome datasets;

The submission script I used in genome assembly using matabat have been uploaded in the box folder

7\_checkM\_genome\_quality\_check.sh

When I get the checkm results, my habit is that I will discard these draft genome bins with completeness <80% or <85%, or the contamination is > 10% or >5%; the cutoff varied in papers

1. ***Genome annotation platforms***

Annotation of the genes in each genome for their potential functional roles, the platform I have used are kegg (online), kbase (online) and prokka, I would suggest you try all these three platforms and compare the results

Kegg: <https://www.kegg.jp/ghostkoala/>

Input file: faa files of each genome (faa file will be returned in the checkm result folder), upload the faa files online and check your email for the annotation results

Kbase: <https://kbase.us/>

Upload the original genome bin files online and there are a number of tools for the users to choose in the annotation

Prokka: <https://github.com/tseemann/prokka>

Input file: original genome bin files

The submission script I used in the prokka annotation:

8\_genome\_annotation\_prokka.sh

1. ***Phylogeny annotation of the draft genome bins recovered***

***GTDB:***

***Option1:***

<https://github.com/Ecogenomics/GTDBTk>

The submission script I used in the prokka annotation:

9\_phylogeny\_genome\_GTDB.sh

Please check the modules needed to be loaded in the above submission file, it may take a while for you to get all of them installed, in case you encounter any problem in the installation, either check the link above or consult the NU computer center for assistance, the staff are very helpful and efficient; or in some cases, you may need to consult the developers with a new ‘issue‘ in the above github link

***Option2:***

the GTDB have also been included as a plugin in kbase, while I have not tried it yet

<https://kbase.us/applist/apps/kb_gtdbtk/run_kb_gtdbtk/release?gclid=Cj0KCQiAtf_tBRDtARIsAIbAKe06MuRsXUJubxsYRoX1q88nN6GXfPSc-gBxkmxNnM3XYCGcL7KvOBQaAvz4EALw_wcB>

1. ***Other tools (optional)***

***10.1 FASTANI :*** <https://github.com/ParBLiSS/FastANI>

whole-genome Average Nucleotide Identity (ANI). ANI is defined as mean nucleotide identity of orthologous gene pairs shared between two microbial genomes; This is useful for you to compare the similarity between the genome obtained and the reference genomes;

The submission script I used in the fastani comparison that have been uploaded in box:

10.1\_fastani.sh

In cases you need to download a large number of reference genomes from NCBI and you have already know the accession number of the genomes, refer to the instructions here to download: <https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/>

* 1. **sequence alignment:**

Tools I have used:

**muscle:** <http://www.drive5.com/muscle/muscle.html>

**mafft:** <https://mafft.cbrc.jp/alignment/software/>

**mega:** <https://www.megasoftware.net/>

**RAxML:** <http://www.exelixis-lab.org/software.html>

I only installed muscle on quest, and I installed mafft, mega and raxml locally on my own computer, below is the script I used for muscle alignment jobs on quest:

10.2\_muscle\_alignment.sh

Note: double check the sequences if you may get quite unexpected results, the lessons I have learnt in my analysis included: 1) sometimes, clone sequences should be reverse complemented before alignment; and 2) it is not safe to translate amino acid sequences to nucleotide sequences and do the alignment, better to use the originally nucleotide sequences from the assembly to do the alignment. These two aspects may mislead you that your sequences are of higher novelty than they actually are

How to check whether the sequences should be reversed complemented: get several high quality reference sequences of the same gene (for example, if you are working with amplicon sequences of amoA gene, download from NCBI several high quality reference sequences of amoA genes), do alignment and visualize the alignment through MEGA, if the sequences you have are poorly aligned to the reference sequences of the same gene, something might be wrong, either the sequencing quality is not good, or some assembly error, or try reverse complement (there is such RC option in MEGA) and then redo the alignment to see whether a much better alignment could be achieved.

* 1. **phylogenetic tree generation**

Tools I have used:

Fasttree: <http://www.microbesonline.org/fasttree/>

Mega: <https://www.megasoftware.net/>

I only installed fasttree on quest, and I installed mega locally on my own computer, below is the script I used for fasttree generation on quest:

10.3\_fasttree\_nt.sh

Visualization of the phylogeny tree through ITOL: <https://itol.embl.de/>

* 1. **prodigal : ORF annotation from contigs**

<https://github.com/hyattpd/Prodigal>

Fast, reliable protein-coding gene prediction from contigs assembled,

1. the prokka mentioned above could also predict protein-coding genes from contigs in genomes; but when I need to annotated genes in all contigs obtained through assembly, I will use prodigal, as I find prokka is more suitable for genomes.
2. Unlike prokka that would return the function annotation of each gene, the prodigal only find the regions of genes but there is no function assignment to each gene (e.g, tell us which gene is the amoA gene, or ARG gene, ect.); to get function annotation of each gene returned by prodigal, we need to blast the gene sequences against databased (nr database, or functional gene database curated by ourselves)

the script I used for finding genes through prodigal on quest have been uploaded to box and its name is:

10.4\_prodigal\_ORF\_finder.sh

**10.5 blast**

Blast is quit commonly used in sequence annotation, if you have one or two sequences, blast in NCBI online should be a good choice: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

However, if you have a large number of genes (tens of thousands) as annotated from contigs, online blast is not a good option.

Download nr database from ncbi:

Introduction on the nr database: <https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/>

How to download nr database: <https://www.ncbi.nlm.nih.gov/books/NBK537770/>

Nr database is a fasta file that is a collection of genes with function; fasta file cannot be used as a database directly, we need to format it before it could be blast against

How to create a blast database from the nr database: <https://ncbi.github.io/magicblast/cook/blastdb.html>

the script I used for make blast database of nxrB genes and the blast command line has been uploaded to box and its name is :

10.5\_blast\_nxrb.sh

* 1. **split fasta file**

in case you are working with a very large fasta file (>1G), it is better to split this one large fasta file to several small files and then do the blast; such split is necessary because blast may not finish in 48 hours that is the max duration for single job submitted to quest; once we split the fasta file into several smaller ones, we can submit several jobs separately for each smaller fasta file, and ensure each of the job could finish in 48 hours.

The perl script I used to split fasta into several smaller ones has been uploaded to quest, and its name is : 10.6\_fasta-splitter.pl

How to use this perl script：

For example, you want to split the assembly.orfs.faa file to 6 smaller files, and put all these 6 files into a folder named “assembly.orfs.faa\_split/:

 mkdir ./ assembly.orfs.faa\_split

perl fasta-splitter.pl --n-parts 6 --out-dir ./assembly.orfs.faa\_split/ ./assembly.orfs.faa

* 1. **sortmerna**

This sortmerna could be used to remove the ribosome RNA sequences from the RNA datasets and we are supposed to map only the mRNA reads to the contigs and genes

<https://github.com/biocore/sortmerna>

one alternative approach that could be used to pick out the ribosome RNA sequences is through blast the RNA reads to silva database and delete the RNA reads with positive matches

### Part iii: summary

Summarized above are all the software and submission script used in the metagenome and metatranscriptome data analysis;

Following the above step, you are supposed to have obtained data or complete info on :

1. check the quality of the sequences;
2. trim the sequences if necessary;
3. assemble the raw read to contigs,
4. map the DNA/RNA raw reads to the contigs, and calculate the ratio of DNA/RNA reads used in contig assembly;
5. obtain a number of draft genomes, calculate the abundance and expression of each genome;
6. check the quality of the draft genomes obtained and discard those genomes of low quality;
7. phylogeny and taxonomy annotation of the draft genome obtained;
8. find all genes in the contigs assembled
9. blast nr database or local databases to annotate the functions of the genes
10. fastani of the genomes obtained and the reference genomes
11. functional gene alignment and phylogeny tree construction