Scientific Report – Short Term Scientific Mission (STSM)

Manoj Kumar Gopala Krishnan

COST Action FA1103

Host institution: Austrian Institute for Technology
Period: 28/04/2014 to 04/05/2014
Reference code: COST-STSM-ECOST-STSM-FA1103-280414-042309

STSM Topic:

Bioinformatics tools for analysis of endophytic bacterial communities.

(i) Abstract

(ii) Purpose of the STSM

The main goal of the STSM was to learn bioinformatics tools for analysis of the bacterial and fungal community sequence data from arcto-alpine plants. In addition, the goal was to learn to apply R statistics for downstream statistical analysis of these communities. In particular, the purpose was to learn bioinformatics tools applicable to sequence data generated by IonTorrent semiconductor technology and proper statistical methods to identify linkages between complex and divergent bacterial and fungal communities.

The group of Professor Angela Sessitsch at the Austrian Institute of Technology Austria has strong experience on NGS sequence analysis of plant microbiomes and bioinformatics knowledge on NGS data processing and downstream community analysis, as well as necessary infrastructure for these.

(iii) Description of the work carried out during the STSM

Samples used for analysis:

Soil samples collected from three geographical regions: Kilpisjarvi, Finland (low-arctic), Ny-Alesund, Svalbard (high-arctic) and Innsbruck, Austria (Alps). Top (1-5 cm depth), bulk (10-15 cm depth) and rhizosphere samples (*Oxyria digyna* and *Saxifraga oppositifolia*) were used for the analysis. DNA extracted from the soil was further amplified and sequenced for fungal phylogenetic diversity analysis based on ITS region using ITS-4 and ITS-7 primers (1). About 700,000 reads from a total of 96 samples from bulk and rhizosphere soils were used for analysis. Then the obtained sequences were used for downstream bioinformatics analysis. Since we are still optimizing the primer set for the endophytic bacteria we used the fungal communities here to set the pipeline. The same pipeline will be further used to analyse the bacterial communities too.

Bioinformatics analysis:

Mothur and QIIME are two open source software packages available to analyse the diversity of microbial communities primarily based on high-throughout sequence data (2, 3). Data generated from different sequencing techniques like Sanger, 454, Ion-torrent and Illumina can be analysed using these software. In our analysis we have chosen QIIME to analyse our fungal community data obtained through ion-torrent sequencing. Even though, ion-torrent sequencing technique is widely used there is no proper pipeline set up to analyze the generated data. Hence, during the STSM training we created a new pipeline to analyse the ion-torrent data.

The X.sff file was obtained from ion-torrent server (http://130.234.117.6/login/?next=/data/#table) and quality checked in PRINSEQ (http://prinseq.sourceforge.net/). The flogram file was then converted into X.fasta file using mother command "mothur > sffinfo (sff =X.sff)". Data from read sequences, quality, flows and ancillary metadata were analysed using Quantitative Insights Into Microbial Ecology (QIIME) pipeline. A meta data file was prepared with information such as sample name, barcodes, forward primer, reverse primer, sampling region, sampling site, soil type and plant species of the rhizosphere soil used for analysis (4). Then metadata file in tab delimited text format (X.txt) was validated using the command "validate_mapping_file.py -m X.txt -o mapping_output". Any errors or warnings in the metadata file can be checked and corrected by referring the output X.html file.

The mapping file (X.txt) along with fasta (X.fasta) file was used for trimming the sequence data. Quality filtering consisted of discarding reads <200 nt and >1000 nt, excluding homopolymer runs >8 nt and ambiguous bases >6, accepting 2 barcode correction and turning the primer check off. A value of 25 has been considered as a minimum average Phred quality score allowed in reads.

The following command was used "split_libraries.py -b variable_length -H 8 -p -e 2 -m X.txt -f X.fasta -q X.qual -l 200 -L 1000 -z truncate_only -s 25 -w 50 -o split_library_output/"

-H = length of homopolymers (set to 8)

-e = errors in barcode (set to 2)

-l = minimum sequence length

-L = maximum sequence length

-s = quality score (Set to 25)

Sequences were de-replicated, sorted, noisy filtered and chimera checked in UPARSE, considering a pairwise identity percentage of 0.97. The chimeras were removed from the sequences using UNITE fungal database which consists a total of 352,622 fungal ITS sequences. Sequences from the same database were used to assign OTU's to the respective taxonomy by blasting with rdp database.

The commands used are listed below with respective functions.

-derep_fulllength = remove duplicated sequences

-sortbysize = sort sequences in ascending order

-cluster_otus = cluster markers gene sequences into OTU's

-db/macqiime/unite97.fasta = removes chimeric sequences

fasta_number.py = replaces fasta labels with OTU labels

assign_taxonomy.py = assign OTU's with taxonomy sequences

-m rdp = blasting with rdp database

Finally, the taxonomy assigned OTU table can be converted into BIOM table (X.biom) and furthermore converted in to text (X.txt) and used for further downstream analysis. To correct for sampling effort, a randomly selected subset based on the number of sequences in the poorest sample was calculated in QIIME and used for further analyses.

R statistics was used to determine alpha diversity, beta diversity and to plot ordination graphs.

(iv) Description of the main results obtained

Therefore using the pipeline we were able to compare the fungal communities across the bio-geographic regions using PRIMER 6 and R statistics. The preliminary results indicated that climatic regions primarily influence the fungal community structure. In addition, we also observed the effect of vegetation and selection of rhizosphere fungal communities by individual plant species. Moreover, we also identified OTU's associated with climatic zones, sampling site and plant species.

We are planning to publish the results obtained from this STSM program. I will send a copy of the publication with detailed data after the acceptance. We are planning to complete and submit the article by December 2014. The resulting pipeline we established during this STSM is given below.

#**Check mapping file first:**

```
validate_mapping_file.py -m fun_comb.txt -o mapping_output
```

#**Trimming Barcodes**

```
split_libraries.py -b variable_length -H 8 -p -e 2 -m map.txt -f seq.fasta -q  seq.qual -l 200 -L 1000
-z truncate_only -s 25 -w 50 -o split_library_output/
```

#**UPARSE:**

```
/macqiime/usearch7 -derep_fulllength split_library_output/seqs.fna -output seqs_der.fna
-sizeout -threads 4
```

```
/macqiime/usearch7 -sortbysize seqs_der.fna -minsize 2 -output seqs_der_nosingle.fna
```

```
/macqiime/usearch7 -cluster_otus seqs_der_nosingle.fna -otus seqs_der_nosingle_rep.fasta
```

```
/macqiime/usearch7 -uchime_ref seqs_der_nosingle_rep.fasta -db /macqiime/unite97.fasta
-strand plus -nonchimeras seqs_der_nosingle_rep_nonchimeras.fasta -chimeras chimeras.fasta
-threads 4
```

```
/macqiime/fasta_number.py seqs_der_nosingle_rep_nonchimeras.fasta OTU_ >
seqs_der_nosingle_rep_nonchimeras_OTU.fasta
```

```
/macqiime/usearch7 -usearch_global split_library_output/seqs.fna -db
seqs_der_nosingle_rep_nonchimeras_OTU.fasta -strand plus -id 0.97 -uc otu_map.uc -threads 4
```

```
python /macqiime/uc2otutab_mod.py otu_map.uc > uparse_table.txt
```

#Back to QIIME-ing

```
biom convert --table-type="otu table" -i uparse_table.txt -o uparse_table.biom
```

#Taxonomy assignment

```
assign_taxonomy.py -i seqs_der_nosingle_rep_nonchimeras_OTU.fasta -r
/macqiime/unite97.fasta  -t /macqiime/unite97.txt -m rdp  -o taxonomy_rdp --rdp_max_memory
8000
```

#Build the BIOM table

```
biom add-metadata --sc-separated taxonomy --observation-header OTUID,taxonomy
--observation-metadata-fp
taxonomy_rdp/seqs_der_nosingle_rep_nonchimeras_OTU_tax_assignments.txt -i
uparse_table.biom -o otu_table.biom
```

```
biom convert -i otu_table.biom  -o otu_table.txt -b --header-key taxonomy
```

```
biom convert -i table_table.txt -o otu_table.biom --table-type "otu table" --process-obs-metadata
taxonomy
```

```
biom summarize-table -i otu_table.biom -o summary_otu_table.txt (--qualitative)
```

#Summarizing taxonomy

```
summarize_taxa.py -i otu_table.biom -o taxonomy_summaries/
```

(v) Future collaboration with host institution

As I will set up the pipeline developed in home institution, the interaction with the host institution will continue with optimization and further development of the pipeline, further collaborative data analysis and preparation of 1-2 manuscripts on the data obtained.

(vi) Foreseen publications/articles resulting or to result from the STSM (if applicable)

I expect to publish one or two paper(s) with the collaboration of the host institute. Furthermore, the skills I learned during this STSM will definitely help in analyzing my sequence data in future and result in at least 2 to 3 more publications during my doctoral studies.

(viii) References

1. Saanakajsa M. Velmala, Tiina Rajala, Jussi Heinonsalo, Andy F.S. Taylor and Taina Pennanen; Profiling functions of ectomycorrhizal diversity and root structuring in seedlings of Norway spruce (Picea abies) with fast- and slow-growing phenotypes; New Phytologist (2014) 201: 610-622.

2. J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld and Rob Knight; QIIME allows analysis of high-throughput community sequencing data: Nature Methods (2010) 7; 335-336.

3. Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, Brian B. Oakley, Donovan H. Parks, Courtney J. Robinson, Jason W. Sahl, Blaz Stres, Gerhard G. Thallinger, David J. Van Horn and Carolyn F. Weber; Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities: Applied and Environmental Microbiology (2009) 75(23);7537-7541.

4. http://qiime.org/documentation/file_formats.html