16S rRNA sequencing hands-on tutorial using Chipster

Eija Korpelainen, CSC – IT Center for Science, chipster@csc.fi

This example is based on Mothur's MiSeq SOP: http://www.mothur.org/wiki/MiSeq_SOP
The data consist of the following 19 mouse fecal samples collected after weaning from one animal:
-9 early samples, collected on days 0-9 post weaning
-10 late samples, collected on days 141-150 post weaning
All samples were sequenced using Illumina MiSeq. We have overlapping 250 bp paired end reads which come from the 253 bp long V4 region of the 16S rRNA gene.

1.Start Chipster and open a session
Go to **http://chipster.csc.fi/**, and **Launch Chipster**.
Select **Open example session** and **course_16s_mothur_miseq**. This session has 2 x 19 zipped fastq files.

2. Package all the fastq files in a Tar package
Select all the fastq files and run the tool **Utilities / Make a Tar package**. Click **Show parameters** and set **File name for tar package = zippedFastq**. Note that after this point in real life you can delete the individual fastq files, because you can always open the Tar package using the tool **Utilities / Extract .tar.gz file**.

3. Check the quality of reads with FastQC
Select the file **zippedFastq.tar** and the tool **Quality control / Read quality with MultiQC for many FASTQ files**, and click **Run**. Select the result file and the visualization option **Open in external web browser**.
-How long are the reads (in the **General Statistics** section click **Configure columns**, select **Length** and unselect **M sequences**)? Do all samples have the same number of reads? Is the base quality good all along the reads?

4. Combine paired reads into sequence contigs
Select the **zippedFastq.tar** and run the tool **16S rRNA sequencing / Combine paired reads to contigs**.
-Choose **contigs.summary.tsv** file and in the Visualisation method **Spreadsheet**. How many sequences are there in the data? How long are most of the contigs? The longest contig? Are there ambiguous bases in the contigs?
-Choose **contig.numbers.txt** and **View text**. Are there roughly the same number of contigs in each sample?
-Check in the **samples.fastqs.txt** if fastq files were assigned correctly to samples.
-Open **contigs.groups** as text file. What does it contain?

5. Remove suspiciously long sequences and sequences with ambiguous bases
Choose **contigs.fasta.gz** and **contigs.groups** (use ctrl / cmd key to select multiple files). Select tool **Screen sequences for several criteria**, and set **Maximum length of the sequences = 275** and **Maximum number of ambiguous bases = 0**, and run the tool.
-Open **summary.screened.tsv** as spreadsheet. Did we manage to remove all the long contigs and ambiguous bases? How many sequences did we remove in total?

6. Remove identical sequences
Select **screened.fasta.gz** and **screened.groups**, and run the tool **Extract unique sequences**.
-Open **unique.count_table** as spreadsheet. What do the rows and columns represent?
-Open **unique.summary.tsv**. How many unique sequences do we have in the data?

7. Align sequences to reference
Select files **unique.fasta** and **unique.count_table**, and run the tool **Align sequences to reference** so that you select only a small section of the reference alignment by setting **Start** = **13 862** and **End** = **23 444**.
-Open **aligned-summary.tsv** as a spreadsheet. Where in the reference alignment do most of the contigs align? Are there deviants?
-Open also **custom.reference.summary.tsv**. How long is the longest homopolymer in the reference?

8. Remove sequences which align outside the wanted alignment range or contain too long homopolymers

Choose **aligned.fasta.gz** and **unique.count_table**, and tool **Screen sequences for several criteria**. Set **Start position** = **8**, **End position** = **9582**, and **Maximum homopolymer length** = **16**. Run the tool.

-Open **summary.screened.tsv** as spreadsheet. How many unique sequences were removed? How many sequences were removed overall?

9. Remove gaps and overhangs from the alignment. If this creates new identical sequences, remove them

Choose **screened.fasta.gz** and **screened.count_table** and run the tool **Filter sequence alignment**.

-Open **filtered-log.txt** as text file. How many columns we removed? How long was the alignment before and how long is it now?

-Did we generate and then remove any identical sequences (compare **summary.screened.tsv** and **filtered-unique-summary.tsv**)?

10. Precluster the aligned sequences

Select **filtered-unique.fasta** and **filtered-unique.count_table** and the tool **Precluster aligned sequences.** Set the **Number of differences allowed** to **2** and run the tool.

-Open **preclustered-summary.tsv** as spreadsheet. How many unique sequences do we have now?

11 Removing chimeras

Select **preclustered.fasta** and **preclustered.count_table** and the tool **Remove chimeric sequences**. Set **dereplicate** = **true** and make sure that **Silva gold bacteria** is selected as the reference.

-Open **chimeras.removed.summary.tsv** to find out how many chimeras were removed.

12. Classify sequences and remove unwanted "species"

Choose **chimeras.removed.fasta** and **chimeras.removed.count_table**, and run the tool **Classify sequences to taxonomic units** so that you remove lineages Chloroplast-mitochondria-Archaea-Eukaryota-unknown.

-Open **classification-summary.tsv**. What kind of bacteria do we have there?

13. Create species count table and phenodata file

Choose **sequences-taxonomy-assignment.txt** and **picked.count_table**, and run the tool **Produce count table and phenodata** so that you set **Cutting level for taxonomic names** = **4**.

-Open **counttable.tsv** and view it as spreadsheet. Which sample has most Bacteroidales in it? (Hint: click the column title Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales to sort).

-Open the file **countable_transposed.tsv**. How does it differ from the other one?

Repeat the analysis run so that you set **Cutting level for taxonomic names** = **3** and **Rarefy counts** = **yes**. Rename the result file to **counttable-rarefied.tsv**.

14. Fill in the group information in the phenodata files

Open **phenodata.tsv** in **Phenodata editor**. Fill in the **group** column: mark all the early samples (D0-D9) with **1**, and all the late samples (D141-D150) with **2**. Note that you can copy the column contents to the other phenodata file by selecting the cells, right-clicking, and selecting **Copy**.

15. Statistical analysis

Select **countable_transposed.tsv** and run **Quality control / PCA and heatmap of samples with DESeq2**.

-Open **PCA_and_heatmap_deseq2.pdf** and check the PCA plot. Do the groups separate from each other?

Select **countable-rarefied.tsv** and run the tool **Statistical analysis for marker gene studies**. Open the result files **rank-abundance_rarefaction_RDA.pdf** and **stat-results.txt**.

-Does the group variable (early versus late samples) explain, at least partly, the differences in composition between the samples? What is the p-value for this possible effect?

-What are the bacterial groups that differentiate the groups best (use the results from indicator species approach)?