TD : Metagenomics using Galaxy

1- Traditionnal approach using Megablast

Reads sequenced from environmental samples can be maped against generalist databases such as GenBank (NT) or WGS (Whole Genome Shotgun) in order to be then assigned to specific microbial species (bacteria, virus…). To do so, we will run the Megablast software traditionally used for this task. It uses the greedy algorithm for nucleotide sequence alignment search.

The following exercise is taken from the Galaxy Central « Metagenomics exercise » https://usegalaxy.org/u/james/p/exercise-metagenomics.

Step 1: Import raw reads
Data libraries => Formation => Metagenomics => megablast_ipnut.fna
into your current history.

Step 2: The default read names are unwieldy and not handled well by Megablast. We will rename the reads with a numeric index. Use the "NGS: QC and manipulation > Rename sequences" tool and select "Rename sequences to" "numeric counter".

Step 3: Use "NGS: Mapping > Megablast" to map reads to the NT database. Set the identity threshold to 80% and the E-value cutoff to 0.0001. Inspect the resulting alignments.

Step 4: Filter alignments to include only those with query coverage greater than 0.5. First we will use "FASTA Manipulation > Compute Sequence Length" to find the length of each query sequence (run this on the high quality segments). Next, we want to join the Megablast results with the sequence lengths. Both datasets have sequence identifiers in column 1. Use "Join Subtract and Group > Join two Queries" to put them together. Finally, filter this dataset using "Filter and sort > Filter" to remove lines with low alignment coverage (c5/c15 > 0.5).

Step 5: Now, map sequences back to their taxonomic position. Each megablast result contains the unique identifier GI as the second column. Use the "Metagenomic Analysis > Fetch Taxonomic Representation" tool to map these reads to their taxonomic representation.

Step 6: Use the "Metagenomic Analysis > Find lowest diagnostic ranks" and find reads that are diagnostic for a classification level below Kingdom.

Step 7: Use "Metagenomic Analysis > Summarize Taxonomy" to aggregate over all the diagnostic reads to get a summary of the number of reads diagnostic for each clade.

Step 8: Use "Metagenomic Analysis > Draw Phylogeny" to draw a taxonomic tree with each node annotated with the number of diagnostic hits for that node.

2- Using MetaPhlAn

MetaPhlAn (Metagenomic Phylogenetic Analysis) is a computational tool for profiling the composition of microbial communities from metagenomic shotgun sequencing data. MetaPhlAn relies on unique clade-specific marker genes identified from 3,000 reference genomes, allowing:

- *up to 25,000 reads-per-second (on one CPU) analysis speed (orders of magnitude faster compared to existing methods);*
- *unambiguous taxonomic assignments as the MetaPhlAn markers are clade-specific;*
- *accurate estimation of organismal relative abundance (in terms of number of cells rather than fraction of reads);*
- *species-level resolution for bacterial and archaeal organisms;*
- *extensive validation of the profiling accuracy on several synthetic datasets and on thousands of real metagenomes*.

http://huttenhower.sph.harvard.edu/metaphlan/
(Segata et al, Nat Methods, 2012)

Step 1: Import raw reads
Data libraries => Formation => Metagenomics => metaphlan_input_test.fna
into your current history.

Step 2: Run MetaPhlAn (Metagenomic analyses => MetaPhlAn)

Step 3: Convert the output into a specific input for Krona visualization.
(Metagenomic analyses => MetaPhlAn to Krona)

Step 4: Create the Krona HTML output using « Visualize with Krona ».
Download the output, unzip it locally in your computer and open the HTML file.