Qiime-formatted MaarjAM

Nov 13 build – SSU rDNA

HOW TO...

Data recovery

1. Log in into maarjAM website *#Users must be logged in to download data*
2. Search by fungal taxon, querying AM classes
3. Filter by 18S rDNA marker
4. Export all sequences in set (short form) *#Sequence file → Multi-sequence fasta*
5. Export all biogeodata to Excell *#Biogeodata → ID_to_taxonomy map*
6. Save Type sequences of VT, status 31/03/2013 (fasta) *#Reference Sequence file for chimera detection*

ID-to-taxonomy map

1. Correct biogeodata files *#Currently, there are 46 empty fields after export*
2. Export in CSV format, tab delimited
3. cat *.csv > maarjAM.biogeodata.csv *#Merge CSV files*
4. Open in Excell
5. Delete biogeodata header rows
6. Sort ascending by GenBank Accession Number, save
7. awk script#1 *#Generate ID-2-TX file with 6 levels descriptors*
8. Open in text Editor and delete YYY00000 entries (and duplicates)

Script#1: awk -F"\t" '{if ($8 !~ /^ *$/) {print $2"\tFungi;Glomeromycota;"$3";"$4";"$5";"$6"_"$7"_"$8} else {print $2"\tFungi;Glomeromycota;"$3";"$4";"$5";"$6"_"$7}}' maarjAM.biogeodata.csv > maarjAM.id_to_taxonomy.txt

Multi-sequence fasta file

1. cat *.txt > maarjaAM.fna
2. format_fasta.pl maarjaAM.fna > maarjAM.unsorted.fna
3. fasta_formatter -i maarjAM.unsorted.fna -o maarjAM.unsorted.tab -t
4. Open in Excell
5. Sort ascending by GenBank Accession number. Save as maarjAM.sorted.csv
6. awk script#2 *#Generate Multi-sequence fasta file*

7. Optional ID-2-TX vs FASTA consistency check

Script#2: awk -F "\t" '{if($1 !~ /^gb\|YYY00000/) {split($1,def,"|"); print ">"def[2]"\n"$2}}' maarjAM.sorted.4.csv > maarjAM.4.fasta