Microbial eukaryotes in high-mountain lakes

Table of Contents

1 Dereplication of reads

```
grep -v ">^" amplicons_linearized.fasta | \
grep -v [^ACGTacgt] | sort -d | uniq -c | \
while read abundance sequence ; do
    hash=$(printf ${sequence} | sha1sum)
    hash=<${hash:0:40}
    printf "%s_%d_%s\n" "${hash}" "${abundance}" "${sequence}"
done | sort -t "_" -k2,2nr -k1.2,1d | \
sed -e 's/\_/\n/2' > amplicons_linearized_dereplicated.fasta
```

2 OTU contingency table

```
STATS="samples.stats"
SWARMS="samples.swarms"
AMPLICON_TABLE="amplicon_contingency_table.csv"
OTU_TABLE="OTU_contigency_table.csv"

# Header
echo -e "OTU\t$(head -n 1 "${AMPLICON_TABLE}")" > "${OTU_TABLE}"

# Compute "per sample abundance" for each OTU
awk -v SWARM="${SWARMS}" \
    -v TABLE="${AMPLICON_TABLE}" \
    'BEGIN {FS = " "
            while ((getline < SWARM) > 0) {
                swarms[$1] = $0
            }
            FS = "\t"
            while ((getline < TABLE) > 0) {
                table[$1] = $0
            }
            while ((getline < CHIMERA) > 0) {
                split($2, a, "_")
                chimera[a[1]] = $18
            }
           }

     {# Parse the stat file (OTUs sorted by decreasing abundance)
      seed = $3 "_" $4
      n = split(swarms[seed], OTU, "[ _]")
      for (i = 1; i < n; i = i + 2) {
```

```
        s = split(table[OTU[i]], abundances, "\t")
        for (j = 1; j < s; j++) {
            samples[j] += abundances[j+1]
        }
    }
    printf "%s\t%s", NR, $3
    for (j = 1; j < s; j++) {
        printf "\t%s", samples[j]
    }
printf "\n"
delete samples
}' "${STATS}" >> "${OTU_TABLE}"
```

HTML generated by org-mode 6.33x in emacs 23