Comparison of the frequency vectors between Bellairs and Cement.

Hallett group

July 2019

We perform some cursory exploration of the frequency tables obtained from Bracken (REF) and Kraken for the Barbados and Cement reef marine samples. The code to perform this analysis is primarily in `src/percolate.R`, as some of the routines are too time consuming to be performed within the R markdown setting here. Column names from Bracken have been modified for readability below. Some of the code is reproduced here to explain the series of steps. We begin by reading in the two Braken files from the `data` directory.

```
> options(warn = -1)
> root <- rprojroot::find_root(".git/index"); setwd(root);
> setwd(file.path(root, "experiments/exp3-bracken2data.frame"))
> library(xtable)
> source("../../src/functions.R")
```

The final tree will always be kept in
`repo/reefmicrobiome/data/tree.latest`. This data.frame contains all the information we have about our samples. It was generated using the code in `ad_hoc_scripts.R` and the functions in `functions.R`

```
> # load the finalized tree  computed in percolate.R with functions from abundance_comparison.R
> REEF_DIR <- "/home/data/refined/reef/R/"
> load( paste0(REEF_DIR, "raw.tree.april.15.RData" ) )
> tree[1:10, 1:7 ]
                       name tax_id  parent         rank embl_code division_id
1                      root      1      NA      no rank                     8
2                  Bacteria      2  131567 superkingdom                     0
3              Azorhizobium      6  335928        genus                     0
4   Azorhizobium caulinodans      7       6      species        AC           0
5        Buchnera aphidicola      9   32199      species        BA           0
6                Cellvibrio     10 1706371        genus                     0
7        Cellulomonas gilvus     11    1707      species        CG           0
8              Dictyoglomus     13  203488        genus                     0
9  Dictyoglomus thermophilum     14      13      species        DT           0
10             Methylophilus     16   32011        genus                     0
    br_bel
1  7374451
2  3498456
3      542
4      542
5     4202
6     1413
7     1570
8      276
9      150
10     178
```

The `make_table()` function takes as input a `tax_id` and pretty prints the information regarding the children of that node.

```
> root <- make_table( 1 );   # 1 is the root
> make_table( 131567 ) # superkingdoms 131567

       Name Tax. Id. Parent         Rank Local.Freq.Bel Local.Freq.May
1  Bacteria        2 131567 superkingdom          0.497          0.608
2 Eukaryota     2759 131567 superkingdom          0.480          0.379
3   Archaea     2157 131567 superkingdom          0.023          0.014
  log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1    -0.201         0.474         0.587    -0.113
2     0.237         0.458         0.366     0.092
3     0.535         0.022         0.013     0.009

> make_table( 2, precision = 4)  # 2 is tax_id for bacteria

                             Name Tax. Id. Parent    Rank Local.Freq.Bel
1                  Proteobacteria     1224      2  phylum         0.4805
2                            <NA>  1783272      2 no rank         0.3714
3                            <NA>  1783270      2 no rank         0.1066
4                            <NA>  1783257      2 no rank         0.0158
5           environmental samples    48479      2 no rank         0.0117
6           unclassified Bacteria     2323      2 no rank         0.0053
7                    Spirochaetes   203691      2  phylum         0.0024
8                    Fusobacteria    32066      2  phylum         0.0019
9                   Acidobacteria    57723      2  phylum         0.0016
10                    Thermotogae   200918      2  phylum         0.0007
11                    Nitrospirae    40117      2  phylum         0.0004
12                      Aquificae   200783      2  phylum         0.0004
13          Thermodesulfobacteria   200940      2  phylum         0.0003
14                  Synergistetes   508458      2  phylum         0.0002
15                Deferribacteres   200930      2  phylum         0.0002
16 Caldiserica/Cryosericota group  2498710      2 no rank         0.0001
17                  Elusimicrobia    74152      2  phylum         0.0001
18                Calditrichaeota  1930617      2  phylum         0.0001
19                    Dictyoglomi    68297      2  phylum         0.0001
20                 Chrysiogenetes   200938      2  phylum         0.0000
21           Coprothermobacterota  2138240      2  phylum         0.0000
   Local.Freq.May log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1          0.2505    0.6513        0.2280        0.1471    0.0808
2          0.6929   -0.6235        0.1762        0.4069   -0.2307
3          0.0349    1.1172        0.0506        0.0205    0.0301
4          0.0057    1.0261        0.0075        0.0033    0.0042
5          0.0086    0.3158        0.0056        0.0050    0.0005
6          0.0026    0.7036        0.0025        0.0015    0.0010
7          0.0014    0.5244        0.0011        0.0008    0.0003
8          0.0012    0.4157        0.0009        0.0007    0.0002
9          0.0006    0.9801        0.0008        0.0004    0.0004
10         0.0005    0.4145        0.0003        0.0003    0.0001
11         0.0002    0.8165        0.0002        0.0001    0.0001
12         0.0003    0.3855        0.0002        0.0001    0.0000
13         0.0002    0.3672        0.0001        0.0001    0.0000
14         0.0001    0.7282        0.0001        0.0001    0.0000
15         0.0001    0.5596        0.0001        0.0001    0.0000
16         0.0001    0.3932        0.0001        0.0000    0.0000
17         0.0001    0.6247        0.0001        0.0000    0.0000
18         0.0000    0.8899        0.0000        0.0000    0.0000
19         0.0001    0.3585        0.0000        0.0000    0.0000
20         0.0000    0.9872        0.0000        0.0000    0.0000
21         0.0000    0.2249        0.0000        0.0000    0.0000
```

```
> make_table( 1224, precision = 4) # proteobacteris is 1224
                           Name Tax. Id. Parent      Rank Local.Freq.Bel
1           Alphaproteobacteria    28211   1224     class         0.5184
2           Gammaproteobacteria     1236   1224     class         0.2568
3            Betaproteobacteria    28216   1224     class         0.0885
4   unclassified Proteobacteria    32045   1224   no rank         0.0726
5    delta/epsilon subdivisions    68525   1224 subphylum         0.0605
6                          <NA>  1553900   1224     class         0.0020
7                          <NA>   580370   1224     class         0.0004
8             Acidithiobacillia  1807140   1224     class         0.0004
9         environmental samples    47936   1224   no rank         0.0003
10                         <NA>  2008785   1224     class         0.0001
   Local.Freq.May log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1          0.5641   -0.0845        0.1182        0.0830    0.0352
2          0.2393    0.0705        0.0585        0.0352    0.0233
3          0.0837    0.0559        0.0202        0.0123    0.0079
4          0.0226    1.1671        0.0166        0.0033    0.0132
5          0.0868   -0.3599        0.0138        0.0128    0.0010
6          0.0018    0.0789        0.0004        0.0003    0.0002
7          0.0004    0.1835        0.0001        0.0001    0.0000
8          0.0004   -0.0585        0.0001        0.0001    0.0000
9          0.0009   -1.0823        0.0001        0.0001   -0.0001
10         0.0000    0.3212        0.0000        0.0000    0.0000
```

```
> make_table( 1218 )   # 1218 is the tax_id of Prochlorococcus
                     Name Tax. Id. Parent    Rank Local.Freq.Bel Local.Freq.May
1                    <NA>  2627481   1218 no rank          0.538          0.524
2 Prochlorococcus marinus     1219   1218 species          0.458          0.473
3   environmental samples    98167   1218 no rank          0.004          0.003
  log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1     0.026         0.040         0.157    -0.117
2    -0.032         0.034         0.142    -0.108
3     0.294         0.000         0.001    -0.001
```

```
> make_table( 10239 )  # Viruses is tax_id 10239
                    Name Tax. Id. Parent    Rank Local.Freq.Bel Local.Freq.May
1   unclassified viruses    12429  10239 no rank          0.624          0.274
2           Caudovirales    28883  10239   order          0.258          0.616
3  environmental samples   186616  10239 no rank          0.077          0.055
4                   <NA>   549779  10239  family          0.012          0.016
5                   <NA>  2559587  10239 no rank          0.011          0.013
6           Ortervirales  2169561  10239   order          0.007          0.007
7        Phycodnaviridae    10501  10239  family          0.004          0.008
8                   <NA>   548681  10239   order          0.002          0.003
9             Poxviridae    10240  10239  family          0.001          0.002
10          Iridoviridae    10486  10239  family          0.001          0.001
11         Baculoviridae    10442  10239  family          0.001          0.001
12      Marseilleviridae   944644  10239  family          0.000          0.001
13          Microviridae    10841  10239  family          0.000          0.000
14           Nudiviridae  1511852  10239  family          0.000          0.000
15          Asfarviridae   137992  10239  family          0.000          0.000
16        Polydnaviridae    10482  10239  family          0.000          0.000
17      Papillomaviridae   151340  10239  family          0.000          0.000
18        Hepadnaviridae    10404  10239  family          0.000          0.000
19          Adenoviridae    10508  10239  family          0.000          0.000
20                  <NA>   687329  10239  family          0.000          0.000
```

```
21            Inoviridae    10860  10239  family          0.000          0.000
22           Nimaviridae   196937  10239  family          0.000          0.000
23        Hytrosaviridae  1285590  10239  family          0.000          0.000
24          Circoviridae    39724  10239  family          0.000          0.000
25          Parvoviridae    10780  10239  family          0.000          0.000
26        Ligamenvirales  1511857  10239   order          0.000          0.000
27         Lavidaviridae  1914302  10239  family          0.000          0.000
28        Polyomaviridae   151341  10239  family          0.000          0.000
29     Alphasatellitidae  1458186  10239  family          0.000          0.000
30         Genomoviridae  1910928  10239  family          0.000          0.000
31         Geminiviridae    10811  10239  family          0.000          0.000
32           Ascoviridae    43682  10239  family          0.000          0.000
   log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1      0.824         0.022         0.007     0.015
2     -0.871         0.009         0.016    -0.007
3      0.334         0.003         0.001     0.001
4     -0.284         0.000         0.000     0.000
5     -0.188         0.000         0.000     0.000
6      0.081         0.000         0.000     0.000
7     -0.674         0.000         0.000    -0.000
8     -0.688         0.000         0.000    -0.000
9     -0.446         0.000         0.000    -0.000
10    -0.431         0.000         0.000    -0.000
11    -0.633         0.000         0.000    -0.000
12    -0.347         0.000         0.000    -0.000
13    -0.084         0.000         0.000     0.000
14     0.192         0.000         0.000     0.000
15     0.140         0.000         0.000     0.000
16    -0.519         0.000         0.000    -0.000
17    -0.630         0.000         0.000    -0.000
18     0.154         0.000         0.000     0.000
19    -0.674         0.000         0.000    -0.000
20     0.058         0.000         0.000     0.000
21     0.114         0.000         0.000     0.000
22       Inf         0.000         0.000     0.000
23       Inf         0.000         0.000     0.000
24     0.024         0.000         0.000     0.000
25       Inf         0.000         0.000     0.000
26       Inf         0.000         0.000     0.000
27      -Inf         0.000         0.000    -0.000
28      -Inf         0.000         0.000    -0.000
29      -Inf         0.000         0.000    -0.000
30      -Inf         0.000         0.000    -0.000
31      -Inf         0.000         0.000    -0.000
32      -Inf         0.000         0.000    -0.000
```

```
> make_table( 2157 ) # Archaea
```

```
                            Name Tax. Id. Parent    Rank Local.Freq.Bel
1                  Euryarchaeota    28890   2157  phylum          0.628
2                           <NA>  1783275   2157 no rank          0.324
3          environmental samples    48510   2157 no rank          0.019
4                           <NA>  1783276   2157 no rank          0.014
5                   Asgard group  1935183   2157 no rank          0.011
6           unclassified Archaea    29294   2157 no rank          0.004
7 Candidatus Hydrothermarchaeota  1935019   2157  phylum          0.000
  Local.Freq.May log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
```

```
1          0.815    -0.260         0.014         0.011     0.003
2          0.127     0.939         0.007         0.002     0.006
3          0.023    -0.191         0.000         0.000     0.000
4          0.017    -0.167         0.000         0.000     0.000
5          0.014    -0.228         0.000         0.000     0.000
6          0.005    -0.231         0.000         0.000     0.000
7          0.001    -0.150         0.000         0.000     0.000
```

```
> make_table( 2759 ) # Eukaryota
                      Name Tax. Id. Parent    Rank Local.Freq.Bel
1             Opisthokonta    33154   2759 no rank          0.587
2            Viridiplantae    33090   2759 kingdom          0.369
3                      Sar  2698737   2759 no rank          0.027
4                     <NA>  2611352   2759 no rank          0.004
5                     <NA>  2608109   2759  phylum          0.004
6               Rhodophyta     2763   2759  phylum          0.003
7                     <NA>   554915   2759 no rank          0.003
8            Cryptophyceae     3027   2759   class          0.002
9    environmental samples    61964   2759 no rank          0.001
10                    <NA>  2611341   2759 no rank          0.001
11                    <NA>   554296   2759 no rank          0.000
12      Glaucocystophyceae    38254   2759   class          0.000
13         Malawimonadidae   136087   2759  family          0.000
14                    <NA>  2683617   2759 no rank          0.000
15                    <NA>  2608240   2759 no rank          0.000
16 unclassified eukaryotes    42452   2759 no rank          0.000
   Local.Freq.May log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1           0.571     0.029         0.269         0.209     0.060
2           0.386    -0.046         0.169         0.141     0.028
3           0.024     0.130         0.012         0.009     0.004
4           0.004    -0.183         0.002         0.002     0.000
5           0.006    -0.587         0.002         0.002    -0.001
6           0.002     0.449         0.001         0.001     0.001
7           0.003     0.003         0.001         0.001     0.000
8           0.001     0.322         0.001         0.000     0.000
9           0.002    -0.199         0.001         0.001     0.000
10          0.000     0.108         0.000         0.000     0.000
11          0.001    -0.431         0.000         0.000    -0.000
12          0.000     0.148         0.000         0.000     0.000
13          0.000    -0.208         0.000         0.000     0.000
14          0.000    -2.213         0.000         0.000    -0.000
15          0.000    -0.459         0.000         0.000    -0.000
16          0.000     0.170         0.000         0.000     0.000
```

```
> make_table( 33154 ) # parent of fungi
              Name Tax. Id. Parent    Rank Local.Freq.Bel Local.Freq.May
1          Metazoa    33208  33154 kingdom          0.859          0.847
2            Fungi     4751  33154 kingdom          0.139          0.150
3 Choanoflagellata    28009  33154   class          0.001          0.001
4    Rotosphaerida  2686024  33154   order          0.000          0.001
5    Ichthyosporea   127916  33154   class          0.000          0.000
6       Filasterea  2687318  33154   class          0.000          0.001
  log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1     0.014         0.231         0.177     0.054
2    -0.077         0.037         0.031     0.006
3    -0.441         0.000         0.000    -0.000
```

```
4    -0.159         0.000         0.000     0.000
5     0.077         0.000         0.000     0.000
6    -0.384         0.000         0.000    -0.000
```

```
> make_table( 4751 ) # fungi

                 Name Tax. Id. Parent       Rank Local.Freq.Bel
1              Dikarya   451864   4751 subkingdom          0.961
2  Fungi incertae sedis   112252   4751    no rank          0.035
3 environmental samples    57731   4751    no rank          0.004
4    unclassified Fungi    89443   4751    no rank          0.000
  Local.Freq.May log(BvsM) Glob.Freq.Bel Glob.Freq.May DeltaFreq
1          0.967    -0.006          0.036          0.030     0.006
2          0.030     0.164          0.001          0.001     0.000
3          0.003     0.110          0.000          0.000     0.000
4          0.000     0.252          0.000          0.000     0.000
```

You can easily then send these tables to file using the `write.csv()` function.

1 Background: how the tree was constructed

We now describe how we built the data.frame `tree` and describe the purpose of each of its columns.

1.1 NCBI Taxonomy

```
> tree[1,]

  name tax_id parent    rank embl_code division_id  br_bel   br_may
1 root      1     NA no rank                     8 7374451 13127272
  bell_orig_est_reads bell_orig_fraction br_bel_frac br_may_frac
1                  0                  0           0           0
  may_orig_est_reads may_orig_fraction Local.Freq.Bel Local.Freq.May
1                 0                 0             NA             NA
  Glob.Freq.Bel Glob.Freq.May DeltaFreq Multinom Polarity Polarity.Adj path
1             1             1         0        0        1            1 root
```

```
> tree[1, 1:6]

  name tax_id parent    rank embl_code division_id
1 root      1     NA no rank                     8
```

All of these fields originate from the NCBI Taxonomy download. File `ad_hoc_scripts.R` contains code that addedthe names of each taxa after trying to find the *scientific name* amongst synonyms. The column `tax_id` is used throughout the code to find taxa of interest. The `parent` column defines the structure of the tree (each node points to its unique parent).

1.2 Importing the Bracken counts for our two sites

I have never observed a big difference in the results if I look at Kraken versus Bracken. The script `script_bracken.bash` contains the shell commands used to generate these files. Therefore, I suggest we use only the Bracken mappings of reads to nodes in the tree from here on in.

```
> tree[1,]

  name tax_id parent    rank embl_code division_id  br_bel   br_may
1 root      1     NA no rank                     8 7374451 13127272
  bell_orig_est_reads bell_orig_fraction br_bel_frac br_may_frac
1                  0                  0           0           0
```

```
  may_orig_est_reads may_orig_fraction Local.Freq.Bel Local.Freq.May
1                  0                 0             NA             NA
  Glob.Freq.Bel Glob.Freq.May DeltaFreq Multinom Polarity Polarity.Adj path
1             1             1         0        0        1            1 root
```

```
> tree[1, 7:8]
```

```
   br_bel   br_may
1 7374451 13127272
```

These two fields represent the Braken counts from Bellairs and Maycocks respectively. The first step was to assign the `est_reads` from the `bellairs.bracken` and `cement.bracken` files (cement was the earlier name for Maycocks). This information was renamed `br_bel` and `br_may`. At the same time the `fraction` fieds of the Bracken files were assigned to variables `br_bel_frac` and `br_may_frac`. These reads were assigned to the leaves of `tree`. The code is located in `ad_hoc_scripts.R`.

The next step was to percolate these reads "up" the tree of life to the root. More precisely, consider a node t with children $c_1, \ldots, c_k$. At the Bellairs site child c_i has with read count $\texttt{br\_bel}_i$. Then, `br_bel` for node t is $\Sigma_{1 \leq i \leq k} \texttt{br\_bel}_i$. This is defined analogously for Maycocks.

Before doing this, the root of the tree of life was manually set to `NA` (see `ad_hoc_scripts.R`), as there was a mistake in the NCBI download (the root pointed to itself which causes a problem for recursion).

```
> tree[1,]
```

```
  name tax_id parent    rank embl_code division_id  br_bel   br_may
1 root      1     NA no rank                     8 7374451 13127272
  bell_orig_est_reads bell_orig_fraction br_bel_frac br_may_frac
1                   0                  0           0           0
  may_orig_est_reads may_orig_fraction Local.Freq.Bel Local.Freq.May
1                  0                 0             NA             NA
  Glob.Freq.Bel Glob.Freq.May DeltaFreq Multinom Polarity Polarity.Adj path
1             1             1         0        0        1            1 root
```

```
> #percolate(1)  # note that this function takes about 1 day to run.
> tree[1,]
```

```
  name tax_id parent    rank embl_code division_id  br_bel   br_may
1 root      1     NA no rank                     8 7374451 13127272
  bell_orig_est_reads bell_orig_fraction br_bel_frac br_may_frac
1                   0                  0           0           0
  may_orig_est_reads may_orig_fraction Local.Freq.Bel Local.Freq.May
1                  0                 0             NA             NA
  Glob.Freq.Bel Glob.Freq.May DeltaFreq Multinom Polarity Polarity.Adj path
1             1             1         0        0        1            1 root
```

Next nodes with 0 counts for both Bellairs and Maycocks were removed (see `ad_hoc_scripts.R`). Columns were rearranged and the scientific name from NCBI Taxonomy was assigned, if it existed (see comment above and `ad_hoc_scripts.R`).

1.3 Global versus Local Frequencies

There are two distinct concepts of frequencies that each have advantages and disadvantages. Consider a node t with total reads r and children $c_1, \ldots, c_k$ with total reads $r_1, \ldots, r_k$ respectively. The *local frequency* f_i for child c_i is equal to $\frac{r_i}{r}$. The *global frequency* f_i for child c_i is equal to $\frac{r_i}{R}$ where R is the total number of reads at the root of the tree of life.

The local frequencies were assigned to the `tree` as follows.

```
> #void <- local_frequencies(1)
> tree[1, ]
```

```
  name tax_id parent    rank embl_code division_id  br_bel   br_may
1 root      1     NA no rank                     8 7374451 13127272
  bell_orig_est_reads bell_orig_fraction br_bel_frac br_may_frac
1                   0                  0           0           0
  may_orig_est_reads may_orig_fraction Local.Freq.Bel Local.Freq.May
1                  0                 0             NA             NA
  Glob.Freq.Bel Glob.Freq.May DeltaFreq Multinom Polarity Polarity.Adj path
1             1             1         0        0        1            1 root
```

```
> # Local.Freq.Bel and Local.Freq.May
```

The global frequencies were assigned as follows.

```
> #void <- global_frequencies(1)
> tree[1, ]
```

```
  name tax_id parent    rank embl_code division_id  br_bel   br_may
1 root      1     NA no rank                     8 7374451 13127272
  bell_orig_est_reads bell_orig_fraction br_bel_frac br_may_frac
1                   0                  0           0           0
  may_orig_est_reads may_orig_fraction Local.Freq.Bel Local.Freq.May
1                  0                 0             NA             NA
  Glob.Freq.Bel Glob.Freq.May DeltaFreq Multinom Polarity Polarity.Adj path
1             1             1         0        0        1            1 root
```

```
> # Glob.Freq.Bel and Glob.Freq.May, Delta.Freq
```

For convenience, I added a column to the `tree` data.frame to record the difference in global frequencies between the two sites.

```
> # tree$DeltaFreq <- tree$Glob.Freq.Bel - tree$Glob.Freq.May
> top <- tree[ order( -abs(tree$DeltaFreq) ),   ]
> top[1:max(which(top$DeltaFreq > 0.01)), c(1,2,3, 13:15) ]
```

```
                                               name  tax_id  parent
1546689                             Synechococcales 1890424    1117
867                                   Cyanobacteria    1117 1798711
1463898                                        <NA> 1798711 1783272
1452374                                        <NA> 1783272       2
937                                 Prochlorococcus    1218    1213
933                                   Prochloraceae    1213 1890424
2172523                                        <NA> 2627481    1218
2                                          Bacteria       2  131567
938                         Prochlorococcus marinus    1219    1218
2189                                      Eukaryota    2759  131567
942                                  Proteobacteria    1224       2
15460                                  Opisthokonta   33154    2759
15499                                       Metazoa   33208   33154
4892                                      Eumetazoa    6072   33208
15501                                     Bilateria   33213    6072
1576581                    Prochlorococcus sp. RS50 1924285 2627481
15680                                 Deuterostomia   33511   33213
6204                                       Chordata    7711   33511
6233                                     Vertebrata    7742   89593
64245                                      Craniata   89593    7711
6253                                  Gnathostomata    7776    7742
88654                                    Teleostomi  117570    7776
88655                                  Euteleostomi  117571  117570
12009                           Alphaproteobacteria   28211    1224
17250                                  Streptophyta   35493   33090
```

```
100718                               Streptophytina  131221   35493
2530                                    Embryophyta    3193  131221
36872                                  Tracheophyta   58023    3193
36873                                 Spermatophyta   58024   78536
54490                                 Euphyllophyta   78536   58023
2705                                  Magnoliopsida    3398   58024
1147225                             Mesangiospermae 1437183    3398
1546690                            Synechococcaceae 1890426 1890424
878                                   Synechococcus    1129 1890426
2171151                                        <NA> 2626047    1129
1452372                                        <NA> 1783270       2
45537                  Bacteroidetes/Chlorobi group   68336 1783270
752                                   Bacteroidetes     976   68336
1204513                Prochlorococcus sp. MIT 0604 1501268 2627481
47921                                eudicotyledons   71240 1437183
66286                                    Gunneridae   91827   71240
1147241                                Pentapetalae 1437201   91827
15396                                 Viridiplantae   33090    2759
6345                                 Actinopterygii    7898  117571
22572                                   Neopterygii   41665  186623
150323                                  Actinopteri  186623    7898
1193998                         Osteoglossocephalai 1489341   32443
15163                                     Teleostei   32443   41665
949                             Gammaproteobacteria    1236    1224
150325                                Clupeocephala  186625 1489341
166460                              Rhodobacterales  204455   28211
14797                              Rhodobacteraceae   31989  204455
47951                                        rosids   71275 1437201
6660                                  Sarcopterygii    8287  117571
1059359                        Dipnotetrapodomorpha 1338369    8287
15224                                     Tetrapoda   32523 1338369
15225                                       Amniota   32524   32523
88821                                Flavobacteriia  117743     976
162959                             Flavobacteriales  200644  117743
1194009                           Euteleosteomorpha 1489388  186625
9817                           unclassified viruses   12429   10239
2222380                           Marine virus AFVG 2693321   12429
29430                             Flavobacteriaceae   49546  200644
93670                                 Ctenosquamata  123367  123366
93671                               Acanthomorphata  123368  123367
93668                                  Neoteleostei  123365 1489388
93669                                  Eurypterygia  123366  123365
93672                            Euacanthomorphacea  123369  123368
14837                   unclassified Proteobacteria   32045    1224
57171   unclassified Proteobacteria (miscellaneous)   81684   32045
1622355                                        <NA> 1977087   81684
163469                               Actinobacteria  201174 1783272
1475737           Prochlorococcus sp. REDSEA-S17_B1 1811562 2627481
1194359                             Percomorphaceae 1489872  123369
274                                     Rhizobiales     356   28211
1380                                 Actinobacteria    1760  201174
101016                           cellular organisms  131567       1
66294                                        fabids   91835   71275
21668                                      Mammalia   40674   32524
15226                                        Theria   32525   40674
47950                                      asterids   71274 1437201
```

```
7546                                       Eutheria    9347   32525
15520                                   Protostomia   33317   33213
1147057                               Boreoeutheria 1437010    9347
        may_orig_est_reads may_orig_fraction Local.Freq.Bel
1546689                  0           0.00000     0.91752304
867                      0           0.00000     0.99978363
1463898                  0           0.00000     0.57262139
1452374                  0           0.00000     0.37143386
937                      0           0.00000     0.99868166
933                      0           0.00000     0.79789824
2172523                  0           0.00000     0.53819325
2                        0           0.00000     0.49685028
938                1860875           0.14170     0.45823811
2189                     0           0.00000     0.47970110
942                      0           0.00000     0.48052598
15460                    0           0.00000     0.58748724
15499                    0           0.00000     0.85854216
4892                     0           0.00000     0.99742612
15501                    0           0.00000     0.98663073
1576581             818259           0.06231     0.32916447
15680                    0           0.00000     0.78793977
6204                     0           0.00000     0.98721974
6233                     0           0.00000     1.00000000
64245                    0           0.00000     0.99685080
6253                     0           0.00000     0.99975539
88654                    0           0.00000     0.99783106
88655                    0           0.00000     1.00000000
12009                    0           0.00000     0.51842218
17250                    0           0.00000     0.90834601
100718                   0           0.00000     0.99987713
2530                     0           0.00000     0.99977633
36872                    0           0.00000     0.98821816
36873                    0           0.00000     0.99974740
54490                    0           0.00000     0.99893161
2705                     0           0.00000     0.99838723
1147225                  0           0.00000     0.99065955
1546690                  0           0.00000     0.18837866
878                      0           0.00000     0.98426685
2171151                  0           0.00000     0.98367560
1452372                  0           0.00000     0.10656301
45537                    0           0.00000     0.98394071
752                      0           0.00000     0.98744612
1204513             499786           0.03806     0.23260319
47921                    0           0.00000     0.82448896
66286                    0           0.00000     1.00000000
1147241                  0           0.00000     1.00000000
15396                    0           0.00000     0.36872207
6345                     0           0.00000     0.54579900
22572                    0           0.00000     1.00000000
150323                   0           0.00000     0.99395912
1193998                  0           0.00000     0.99996436
15163                    0           0.00000     0.99705928
949                      0           0.00000     0.25677072
150325                   0           0.00000     0.96563094
166460                   0           0.00000     0.24514899
14797                    0           0.00000     0.94832251
```

```
47951                    0          0.00000     0.61123936
6660                     0          0.00000     0.45420100
1059359                  0          0.00000     0.99686615
15224                    0          0.00000     0.99997616
15225                    0          0.00000     0.94437244
88821                    0          0.00000     0.51128891
162959                   0          0.00000     0.99616082
1194009                  0          0.00000     0.71459996
9817                     0          0.00000     0.62438492
2222380                  0          0.00000     0.96295083
29430                    0          0.00000     0.90372659
93670                    0          0.00000     1.00000000
93671                    0          0.00000     1.00000000
93668                    0          0.00000     0.89188195
93669                    0          0.00000     1.00000000
93672                    0          0.00000     0.96327867
14837                    0          0.00000     0.07262214
57171                    0          0.00000     1.00000000
1622355              43659          0.00332     1.00000000
163469                   0          0.00000     0.25957697
1475737             231112          0.01760     0.12944870
1194359                  0          0.00000     0.93004824
274                      0          0.00000     0.17079834
1380                     0          0.00000     0.95834643
101016                   0          0.00000     0.95481928
66294                    0          0.00000     0.63868067
21668                    0          0.00000     0.68277108
15226                    0          0.00000     0.99413652
47950                    0          0.00000     0.36949440
7546                     0          0.00000     0.97269624
15520                    0          0.00000     0.21205247
1147057                  0          0.00000     0.98023588
```

1.4 The Multinomial Statistic

I implemented the multinomial test and applied it to each node in the `tree` datastructure. The relevant function is in `functions.R`.

```
> #multinomial_tree_test(1)
> # Fraction significant
> length(which(tree$Multinom < 0.01))  / nrow(tree)
```

```
[1] 0.08315555
```

1.5 Taxa Path to Root

Finally, I added a character to the `tree` data.frame that describes the phylogenetic path to the root.

```
> # tree$path <- unlist(lapply( tree$tax_id, FUN = function(x) { return(paste(path2root(x)$name, collapse=
```