Supplemental Information 6: Genome Size

Simpson, Bettauer et al.

April 2020

This supplemental methods describes our investigations of the relative sizes of genomes. We examine the 50 most abundant genres at each site spread across the all the kingdoms and domains in our data by considering the size of the genomes. If the genomes of species different greatly, we will need to correct for genome size in the calculation of frequency.

We begin by loading the refined data after cleaning. At the end of this file, a new version of the tree is saved and the linear models for correcting for genome size.

We begin by reading in the NCBI's summary file of all genes obtained from `ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPO`

```
> cell <- induce_tree(131567)
> virus <- induce_tree(10239)
> unclassified <- induce_tree(12908)
> root.hist <- data.frame( taxa = "cell", genome_size = cell$genome_size[!is.na(cell$genome_size)])
> root.hist <- rbind( root.hist, data.frame( taxa = "virus", genome_size = virus$genome_size[!is.na(virus$g
> # root.hist <-  rbind( root.hist, data.frame( taxa = "unclass", genome_size = unclassified$genome_size[!i
> # removed becasue it is empty
>
> p<-ggplot(root.hist, aes(x=genome_size, fill=taxa, color=taxa)) +
+   geom_histogram(position="identity", alpha=0.5, binwidth = 0.1) +
+   labs(title="Genome Size of Viruses versus Cellular Organisms" ,x="Genome Size (Mbp)", y = "Count")+
+   theme_classic()
> p
```

Notice that there remain some very large genomes at approxiamtely 1.5 billion bp. We investigate these large (non-viral) genomes next.

```
> euk <- induce_tree(2759);
> bac <- induce_tree(2)
> arch <- induce_tree(2157)
> cell.hist <- data.frame( taxa = "euk", genome_size = euk$genome_size[!is.na(euk$genome_size)])
> cell.hist <- rbind( cell.hist, data.frame( taxa = "bac", genome_size = bac$genome_size[!is.na(bac$genome_
> cell.hist <- rbind( cell.hist, data.frame( taxa = "arch", genome_size = arch$genome_size[!is.na(arch$geno
> p<-ggplot(cell.hist, aes(x=genome_size, fill=taxa, color=taxa)) +
+   geom_histogram(position="identity", alpha=0.5, binwidth = 0.1) +
+   labs(title="Genome Size of Cellular Organisms" ,x="Genome Size (Mbp)", y = "Count")+
+   theme_classic()
> p
```

The largest bacterial genome in our data is *Minicystis rosea* at 16 Mbp and the largest Archaea genome is 6 Mbp.

```
> largest <- arrange(bac,desc(genome_size))
> largest_species <- largest[ which(largest$rank == "species"), ]
> largest <- arrange(arch,desc(genome_size))
> largest_species <- largest[ which(largest$rank == "species"), ]
>
```

Therefore, we focus our attention on Eukaryota only of which there are many large genomes.

```
> largest <- arrange(euk,desc(genome_size))
> largest_species <- largest[ which(largest$rank == "species"), ]
> p<-ggplot(largest, aes(x=genome_size)) +
+   geom_histogram(position="identity", alpha=0.5, binwidth=10) +
+   labs(title="Genome Size of Eukaryota in the Barbadian Reefs" ,x="Genome Size (Mbp)", y = "Count")+
+   theme_classic()
> p
```

Many, but certainly not all, of the large genomes correspond to multicellular organisms. We remove the largest (> 30 Mbp) from further analysis. We adjusted the number of counts for each of the remaining genomes that are below this cut off below.

The following taxa were removed.

```
> for (i in 1:200) {
+    cat("\n", largest_species[i, "name"], "\t\t",  largest_species[i, "tax_id"], largest_species[i, "genom
+  }
```

```
 Euglena gracilis                  3039 1435.5          root.cellular organisms.Eukaryota.NA.Euglenozoa.Eug
 Symbiodinium kawagutii                  104179 935.067          root.cellular organisms.Eukaryota.Sar.Alve
 Hemileia vastatrix                  203904 543.605          root.cellular organisms.Eukaryota.Opisthokonta
 Dunaliella salina                  3046 343.704          root.cellular organisms.Eukaryota.Viridiplantae.C
 Mesostigma viride                  41882 289.67          root.cellular organisms.Eukaryota.Viridiplantae.S
 Cymbomonas tetramitiformis                  36881 281.27          root.cellular organisms.Eukaryota.Viridi
 Haematococcus lacustris                  44745 268.7845          root.cellular organisms.Eukaryota.Viridip
 Tetraselmis striata                  3165 227.954          root.cellular organisms.Eukaryota.Viridiplantae
 Oxytricha trifallax                  1172189 219.9967          root.cellular organisms.Eukaryota.Sar.Alveo
 Physarum polycephalum                  5791 205.176          root.cellular organisms.Eukaryota.NA.NA.Eumyc
 Chromera velia                  505693 187.455          root.cellular organisms.Eukaryota.Sar.Alveolata.Co
 Botryococcus braunii                  38881 184.382          root.cellular organisms.Eukaryota.Viridiplant
 Phytophthora infestans                  4787 177.7035          root.cellular organisms.Eukaryota.Sar.Stram
 Trichomonas vaginalis                  5722 164.072          root.cellular organisms.Eukaryota.NA.Parabasa
 Tetradesmus obliquus                  3088 157.946          root.cellular organisms.Eukaryota.Viridiplanta
 Gonium pectorale                  33097 148.806          root.cellular organisms.Eukaryota.Viridiplantae.C
 NA                  588596 145.3173          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Fungi in
 Yamagishiella unicocca                  51707 137.536          root.cellular organisms.Eukaryota.Viridipla
 Tetrabaena socialis                  47790 135.78          root.cellular organisms.Eukaryota.Viridiplantae
 Plasmopara halstedii                  4781 127.0636          root.cellular organisms.Eukaryota.Sar.Stramen
 Chlamydomonas reinhardtii                  3055 120.405          root.cellular organisms.Eukaryota.Viridip
 Sphaeroforma arctica                  72019 115.142          root.cellular organisms.Eukaryota.Opisthokont
 Diplonema papillatum                  91374 107.915          root.cellular organisms.Eukaryota.NA.Euglenoz
 Cyanophora paradoxa                  2762 99.9404          root.cellular organisms.Eukaryota.Glaucocystoph
 Ulva mutabilis                  498180 98.4847          root.cellular organisms.Eukaryota.Viridiplantae.Ch
 Chlorella sp. ArM0029B                  1415603 92.9613          root.cellular organisms.Eukaryota.Viridip
 Plasmopara viticola                  143451 92.59207          root.cellular organisms.Eukaryota.Sar.Strame
 Thalassiosira oceanica                  159749 92.1856          root.cellular organisms.Eukaryota.Sar.Stra
 Psammoneis japonica                  517775 91.4306          root.cellular organisms.Eukaryota.Sar.Stramen
 Ulva prolifera                  3117 87.8893          root.cellular organisms.Eukaryota.Viridiplantae.Chlo
 Phytophthora sojae                  67593 84.23815          root.cellular organisms.Eukaryota.Sar.Strameno
 Acanthamoeba castellanii                  5755 80.8823          root.cellular organisms.Eukaryota.NA.NA.Lo
 Chlamydomonas applanata                  35704 78.5042          root.cellular organisms.Eukaryota.Viridipl
 Hyphochytrium catenoides                  42384 73.08465          root.cellular organisms.Eukaryota.Sar.St
 Paramecium tetraurelia                  5888 72.0945          root.cellular organisms.Eukaryota.Sar.Alveol
 Monoraphidium neglectum                  145388 69.7118          root.cellular organisms.Eukaryota.Viridip
 Asterionella formosa                  210441 68.4198          root.cellular organisms.Eukaryota.Sar.Strame
 Sterkiella histriomuscorum                  94289 66.3686          root.cellular organisms.Eukaryota.Sar.A
 Diaporthe helianthi                  158607 63.672          root.cellular organisms.Eukaryota.Opisthokonta
 Rostrostelium ellipticum                  361140 62.1602          root.cellular organisms.Eukaryota.NA.NA
```

```
Aphanomyces invadans                  157072 61.7834          root.cellular organisms.Eukaryota.Sar.Strame
Toxoplasma gondii                  5811 61.5763          root.cellular organisms.Eukaryota.Sar.Alveolata.A
NA                  554055 61.0189          root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.co
Eimeria mitis                  44415 60.4151          root.cellular organisms.Eukaryota.Sar.Alveolata.Apic
Colletotrichum graminicola                  31870 59.9141          root.cellular organisms.Eukaryota.Opist
Aphanomyces astaci                  112090 59.69445          root.cellular organisms.Eukaryota.Sar.Stramen
Phytophthora cinnamomi                  4785 59.63075          root.cellular organisms.Eukaryota.Sar.Stram
Parachlorella kessleri                  3074 59.1878          root.cellular organisms.Eukaryota.Viridiplan
Chrysochromulina tobinii                  1460289 59.0731          root.cellular organisms.Eukaryota.NA.Ha
Besnoitia besnoiti                  94643 58.8459          root.cellular organisms.Eukaryota.Sar.Alveolata
Moneuplotes crassus                  5936 58.5666          root.cellular organisms.Eukaryota.Sar.Alveolata
Chlorella sorokiniana                  3076 58.33862          root.cellular organisms.Eukaryota.Viridiplan
NA                  690256 57.55417          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
Mastigamoeba balamuthi                  108607 57.2666          root.cellular organisms.Eukaryota.NA.NA.NA
Colletotrichum gloeosporioides                  474922 57.2517          root.cellular organisms.Eukaryota
Auxenochlorella pyrenoidosa                  3078 56.993          root.cellular organisms.Eukaryota.Viridi
Aphanomyces euteiches                  100861 56.9046          root.cellular organisms.Eukaryota.Sar.Stram
Aureococcus anophagefferens                  44056 56.6606          root.cellular organisms.Eukaryota.Sar
Rhizoctonia solani                  456999 56.0285          root.cellular organisms.Eukaryota.Opisthokonta
Salpingoeca rosetta                  946362 55.4403          root.cellular organisms.Eukaryota.Opisthokont
Clonostachys rosea                  29856 55.30035          root.cellular organisms.Eukaryota.Opisthokonta
Eimeria necatrix                  51315 55.0079          root.cellular organisms.Eukaryota.Sar.Alveolata.A
Amorphotheca resinae                  5101 54.4745          root.cellular organisms.Eukaryota.Opisthokonta
Phytophthora capsici                  4784 53.41358          root.cellular organisms.Eukaryota.Sar.Stramen
Fusarium oxysporum                  5507 52.02895          root.cellular organisms.Eukaryota.Opisthokonta
Stichococcus bacillaris                  37433 51.9927          root.cellular organisms.Eukaryota.Viridipl
Raphidocelis subcapitata                  307507 51.1627          root.cellular organisms.Eukaryota.Viridi
Stylonychia lemnae                  5949 50.1645          root.cellular organisms.Eukaryota.Sar.Alveolata
Fistulifera solaris                  1519565 49.7366          root.cellular organisms.Eukaryota.Sar.Strame
Colletotrichum higginsianum                  80884 49.4357          root.cellular organisms.Eukaryota.Opis
Phialocephala scopiformis                  149040 48.8763          root.cellular organisms.Eukaryota.Opist
Ichthyophthirius multifiliis                  5932 48.8          root.cellular organisms.Eukaryota.Sar.Alv
Coremiostelium polycephalum                  142831 48.621          root.cellular organisms.Eukaryota.NA.N
Colletotrichum orchidophilum                  1209926 48.5565          root.cellular organisms.Eukaryota.
NA                  1578925 48.3379          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
Phytophthora parasitica                  4792 47.7626          root.cellular organisms.Eukaryota.Sar.Stram
NA                  554065 46.1595          root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.co
Leptosphaeria maculans                  5022 45.9865          root.cellular organisms.Eukaryota.Opisthokon
Eimeria maxima                  5804 45.9751          root.cellular organisms.Eukaryota.Sar.Alveolata.Apic
Eimeria acervulina                  5801 45.8306          root.cellular organisms.Eukaryota.Sar.Alveolata
Trichoderma virens                  29875 45.8201          root.cellular organisms.Eukaryota.Opisthokonta
Fusarium solani                  169388 45.8133          root.cellular organisms.Eukaryota.Opisthokonta.Fu
Fusarium proliferatum                  948311 45.78918          root.cellular organisms.Eukaryota.Opisthok
Aspergillus mulundensis                  1810919 45.3419          root.cellular organisms.Eukaryota.Opisth
Rhizopus oryzae                  64495 45.04514          root.cellular organisms.Eukaryota.Opisthokonta.Fu
Fusarium fujikuroi                  5127 44.92106          root.cellular organisms.Eukaryota.Opisthokonta
Chrysoporthe austroafricana                  354353 44.6689          root.cellular organisms.Eukaryota.Opi
Pyricularia grisea                  148305 44.31875          root.cellular organisms.Eukaryota.Opisthokont
Pseudocercospora fijiensis                  1873960 44.1245          root.cellular organisms.Eukaryota.Opi
Pythium insidiosum                  114742 44.08819          root.cellular organisms.Eukaryota.Sar.Stramen
Cyclospora cayetanensis                  88456 43.98724          root.cellular organisms.Eukaryota.Sar.Alv
NA                  563466 43.82915          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
Eimeria falciformis                  84963 43.6713          root.cellular organisms.Eukaryota.Sar.Alveolat
Achlya hypogyna                  1202772 43.3985          root.cellular organisms.Eukaryota.Sar.Stramenop
Venturia effusa                  50376 42.9507          root.cellular organisms.Eukaryota.Opisthokonta.Fun
Lobosporangium transversale                  64571 42.7689          root.cellular organisms.Eukaryota.Opis
```

```
[Nectria] haematococca                  140110 42.65957          root.cellular organisms.Eukaryota.Opistho
Fusarium verticillioides                  117187 42.40943          root.cellular organisms.Eukaryota.Opist
Trichosporon coremiiforme                  82509 42.3533          root.cellular organisms.Eukaryota.Opisth
NA                  569365 42.30655          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
Crithidia fasciculata                  5656 41.2974          root.cellular organisms.Eukaryota.NA.Euglenoz
Trypanosoma congolense                  5692 41.2334          root.cellular organisms.Eukaryota.NA.Eugleno
Chlorella vulgaris                  3077 41.10782          root.cellular organisms.Eukaryota.Viridiplantae
Neurospora crassa                  5141 40.98185          root.cellular organisms.Eukaryota.Opisthokonta.F
Naegleria gruberi                  5762 40.9641          root.cellular organisms.Eukaryota.NA.Heterolobose
Trichosporon ovoides                  82524 40.9337          root.cellular organisms.Eukaryota.Opisthokont
Phialemoniopsis curvata                  1093900 40.3666          root.cellular organisms.Eukaryota.Opisth
NA                  568076 40.3173          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.A
Trichoderma harzianum                  5544 40.25984          root.cellular organisms.Eukaryota.Opisthokon
Aspergillus alliaceus                  209559 40.15425          root.cellular organisms.Eukaryota.Opisthok
Aspergillus sojae                  41058 40.054          root.cellular organisms.Eukaryota.Opisthokonta.Fu
Fusarium culmorum                  5516 40.0196          root.cellular organisms.Eukaryota.Opisthokonta.Fu
Aspergillus caelatus                  61420 40.0164          root.cellular organisms.Eukaryota.Opisthokont
Arthrobotrys oligospora                  13349 40.00234          root.cellular organisms.Eukaryota.Opistho
Bodo saltans                  75058 39.8644          root.cellular organisms.Eukaryota.NA.Euglenozoa.Kinet
Hypoxylon pulicicidum                  1243767 39.6241          root.cellular organisms.Eukaryota.Opisthok
Pyricularia oryzae                  318829 39.4074          root.cellular organisms.Eukaryota.Opisthokonta
Cryphonectria parasitica                  5116 39.2611          root.cellular organisms.Eukaryota.Opisthok
Sordaria macrospora                  5147 38.8634          root.cellular organisms.Eukaryota.Opisthokonta
Fusarium venenatum                  56646 38.6602          root.cellular organisms.Eukaryota.Opisthokonta
NA                  1725355 38.5165          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
Paraphaeosphaeria sporulosa                  1460663 38.464          root.cellular organisms.Eukaryota.Opi
Aspergillus pseudotamarii                  132259 38.2439          root.cellular organisms.Eukaryota.Opist
Exophiala oligosperma                  215243 38.2245          root.cellular organisms.Eukaryota.Opisthoko
Trichoderma gamsii                  398673 38.1993          root.cellular organisms.Eukaryota.Opisthokonta
Zymoseptoria tritici                  1047171 38.1077          root.cellular organisms.Eukaryota.Opisthoko
Trichoderma asperellum                  101201 38.0794          root.cellular organisms.Eukaryota.Opisthok
Parastagonospora nodorum                  13684 38.03676          root.cellular organisms.Eukaryota.Opisth
Aspergillus welwitschiae                  1341132 37.84645          root.cellular organisms.Eukaryota.Opis
Neurospora discreta                  29879 37.76317          root.cellular organisms.Eukaryota.Opisthokont
Aspergillus oryzae                  5062 37.63278          root.cellular organisms.Eukaryota.Opisthokonta
Aspergillus bombycis                  109264 37.4746          root.cellular organisms.Eukaryota.Opisthokon
Purpureocillium lilacinum                  33203 37.4171          root.cellular organisms.Eukaryota.Opisth
Fusarium coffeatum                  231269 37.40235          root.cellular organisms.Eukaryota.Opisthokont
Aspergillus pseudonomius                  1506151 37.24685          root.cellular organisms.Eukaryota.Opis
Endocarpon pusillum                  364733 37.1732          root.cellular organisms.Eukaryota.Opisthokont
Cladophialophora bantiana                  89940 37.0879          root.cellular organisms.Eukaryota.Opisth
Aspergillus flavus                  5059 37.00717          root.cellular organisms.Eukaryota.Opisthokonta
Fusarium graminearum                  5518 36.80952          root.cellular organisms.Eukaryota.Opisthokont
Fusarium pseudograminearum                  101028 36.70635          root.cellular organisms.Eukaryota.Opi
NA                  2587410 36.5797          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
Trichoderma atroviride                  63577 36.53947          root.cellular organisms.Eukaryota.Opisthok
Bipolaris sorokiniana                  45130 36.5329          root.cellular organisms.Eukaryota.Opisthokon
Talaromyces pinophilus                  128442 35.8832          root.cellular organisms.Eukaryota.Opisthok
NA                  655981 35.8182          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.A
Beauveria bassiana                  176275 35.63905          root.cellular organisms.Eukaryota.Opisthokont
Pyrenophora tritici-repentis                  45151 35.51074          root.cellular organisms.Eukaryota.Op
Drechslerella brochopaga                  47238 35.4316          root.cellular organisms.Eukaryota.Opistho
Aspergillus niger                  5061 35.42943          root.cellular organisms.Eukaryota.Opisthokonta.F
Aspergillus nomius                  41061 35.2805          root.cellular organisms.Eukaryota.Opisthokonta
Fonsecaea monophora                  254056 35.2298          root.cellular organisms.Eukaryota.Opisthokont
Diplodia corticola                  236234 34.9861          root.cellular organisms.Eukaryota.Opisthokonta
```

```
 Fonsecaea erecta                  1367422 34.748          root.cellular organisms.Eukaryota.Opisthokonta.
 NA                  1659845 34.5224          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
 NA                  2587412 34.5062          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
 Exophiala lecanii-corni                  91925 34.4207          root.cellular organisms.Eukaryota.Opistho
 Paramoeba pemaquidensis                  180228 34.387          root.cellular organisms.Eukaryota.NA.NA.F
 Podospora comata                  48703 34.3855          root.cellular organisms.Eukaryota.Opisthokonta.F
 Verticillium dahliae                  27337 34.03032          root.cellular organisms.Eukaryota.Opisthokon
 Alternaria alternata                  5599 33.95789          root.cellular organisms.Eukaryota.Opisthokont
 Alternaria arborescens                  156630 33.83602          root.cellular organisms.Eukaryota.Opisth
 Lachnellula hyalina                  1316788 33.8283          root.cellular organisms.Eukaryota.Opisthokon
 Fonsecaea nubica                  856822 33.7874          root.cellular organisms.Eukaryota.Opisthokonta.
 Crithidia acanthocephali                  59798 33.7832          root.cellular organisms.Eukaryota.NA.Eug
 Penicillium arizonense                  1835702 33.7291          root.cellular organisms.Eukaryota.Opisth
 NA                  2656787 33.6398          root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
 Curvularia lunata                  5503 33.5359          root.cellular organisms.Eukaryota.Opisthokonta.F
 Nannochloropsis limnetica                  120807 33.5089          root.cellular organisms.Eukaryota.Sar.
 Plasmodium ovale                  36330 33.50455          root.cellular organisms.Eukaryota.Sar.Alveolata
 Leishmania peruviana                  5681 33.399          root.cellular organisms.Eukaryota.NA.Euglenozo
 Fonsecaea multimorphosa                  979981 33.3916          root.cellular organisms.Eukaryota.Opisth
 Penicillium rubens                  1108849 33.3861          root.cellular organisms.Eukaryota.Opisthokont
 Trichoderma citrinoviride                  58853 33.2152          root.cellular organisms.Eukaryota.Opist
 Cordyceps militaris                  73501 33.0958          root.cellular organisms.Eukaryota.Opisthokonta
 Plasmodium gonderi                  77519 33.0063          root.cellular organisms.Eukaryota.Sar.Alveolat
 Rhinocladiella mackenziei                  86056 33.00585          root.cellular organisms.Eukaryota.Opis
 Exophiala spinifera                  91928 32.89853          root.cellular organisms.Eukaryota.Opisthokont
 Leishmania donovani                  5661 32.89695          root.cellular organisms.Eukaryota.NA.Euglenoz
 Cladonia uncialis                  174080 32.8515          root.cellular organisms.Eukaryota.Opisthokonta
 Cladosporium phlei                  1116209 32.816          root.cellular organisms.Eukaryota.Opisthokonta
 Alternaria solani                  48100 32.8036          root.cellular organisms.Eukaryota.Opisthokonta.
 Sporothrix schenckii                  29908 32.77065          root.cellular organisms.Eukaryota.Opisthokon
 NA                 653948 32.7668           root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.
 Dictyostelium purpureum                  5786 32.64365          root.cellular organisms.Eukaryota.NA.NA.E
 Verticillium alfalfae                  1051613 32.6412          root.cellular organisms.Eukaryota.Opistho
 Penicillium expansum                  27334 32.47772          root.cellular organisms.Eukaryota.Opisthokon
 Epichloe festucae                  35717 32.47605          root.cellular organisms.Eukaryota.Opisthokonta
 Trichoderma reesei                  51453 32.32405          root.cellular organisms.Eukaryota.Opisthokonta
 Rhizopus microsporus                  58291 32.2729          root.cellular organisms.Eukaryota.Opisthokont
 Leishmania mexicana                  5665 32.0572          root.cellular organisms.Eukaryota.NA.Euglenozo
 Cercospora sojina                  438356 31.8845          root.cellular organisms.Eukaryota.Opisthokonta
 Exophiala xenobiotica                  348802 31.84035          root.cellular organisms.Eukaryota.Opistho
 Verruconis gallopava                  253628 31.7805          root.cellular organisms.Eukaryota.Opisthokon
 Bipolaris oryzae                  101162 31.674          root.cellular organisms.Eukaryota.Opisthokonta.F
 Verticillium nonalfalfae                  1051616 31.5967          root.cellular organisms.Eukaryota.Opis
 Leishmania amazonensis                  5659 31.5636          root.cellular organisms.Eukaryota.NA.Euglen
 Aspergillus fischeri                  36630 31.38315          root.cellular organisms.Eukaryota.Opisthokon
 Aspergillus nidulans                  162425 31.2995          root.cellular organisms.Eukaryota.Opisthokon
 NA                 691883 31.2965           root.cellular organisms.Eukaryota.Opisthokonta.Rotosphaerida.
 Ramularia collo-cygni                  112498 31.27745          root.cellular organisms.Eukaryota.Opistho

>
>
> # to_kill <- c(69332, 2544991, 88149, 38544, 945030, 2880, 309737, 104198, 658196, 39416, 2788, 72228, 2
> #        33653, 5322)
> #
> # #pre.modified <- tree
> # for (i in 1:length(to_kill)) {
> #   void <- remove_update_tree( to_kill[i] )
> #
> #   to_remove <- intersect( which(tree$br_bel==0), which(tree$br_may==0) )
> #   if (length(to_remove)>0) tree <- tree[ -to_remove, ]
> # }
>
> #save(tree, file = paste0(paste0("/home/data/refined/reef/R/ultra.pure.tree.", date), ".RData"))
> #write.csv(tree, file = paste0(paste0("/home/data/refined/reef/R/ultra.pure.tree.", date), ".csv"))
>
```

Let's revisit briefly after these deletions.

```
> load(file = paste0(paste0("/home/data/refined/reef/R/ultra.pure.tree.", date), ".RData"))
> euk <- induce_tree(2759);
> largest <- arrange(euk,desc(genome_size))
> largest_species <- largest[ which(largest$rank == "species"), ]
> p<-ggplot(largest_species, aes(x=genome_size)) +
+   geom_histogram(position="identity", alpha=0.5, binwidth=10) +
+   labs(title="Genome Size of Eukaryota in the Barbadian Reefs" ,x="Genome Size (Mbp)", y = "Count")+
+   theme_classic()
> p
```

```
> euk <- induce_tree(2759); euk$taxa <- "Eukaryota"
> virus <- induce_tree(10239); virus$taxa <- "virus"
> bac <- induce_tree(2) ; bac$taxa <- "Bacteria"
> arch <- induce_tree(2157); arch$taxa <- "Archaea"
> everyone <- do.call("rbind", list(euk, virus, bac, arch))
> largest <- arrange(everyone,desc(genome_size))
> largest_species <- largest[ which(largest$rank == "species"), ]
> library("DescTools")
> largest_species$genome_size <- Winsorize(largest_species$genome_size, maxval = 50, na.rm=TRUE)
> p<-ggplot(largest_species, aes(x=genome_size, fill=taxa, color=taxa)) +
+   geom_histogram(position="identity", alpha=0.5, binwidth = 0.5) +
+   labs(title="Genome Size of Cellular Organisms" ,x="Genome Size (Mbp)", y = "Count")+
+   scale_color_brewer(palette="Dark2")+
+   scale_fill_brewer(palette="Dark2")+
+   theme_classic()
> p
>
```

Now we repeat the above plot at the genus level rather than at the species level. Although the two plots are very similar, we note that our calculation at each internal nodes *t* in the tree of life should be fixed. Currently, we simply compute the average across all the children of *t* but we should rather compute a weighted average. As it is, the averaage genome size at or near the root is disportionality high because it subject to a few large Eukaryota genomes

```
> largest_species <- largest[ which(largest$rank == "genus"), ]
> largest_species$genome_size <- Winsorize(largest_species$genome_size, maxval = 50, na.rm=TRUE)
> p<-ggplot(largest_species, aes(x=genome_size, fill=taxa, color=taxa)) +
+   geom_histogram(position="identity", alpha=0.5, binwidth = 0.5) +
+   labs(title="Genome Size of Cellular Organisms" ,x="Genome Size (Mbp)", y = "Count")+
+   scale_color_brewer(palette="Dark2")+
+   scale_fill_brewer(palette="Dark2")+
+   theme_classic()
> p
```

Finally, save this all to version 1.0.

1 Correlations between genome size and read count

In this section, we look to see if there is a relationship between the number of reads that are mapped to an organism and the size of the genome. For this analysis, we will treat each superkingdom separately. First, we load a version of the final `tree`.

```
> root <- rprojroot::find_root(".git/index");  source(file.path(root, "src/init.R"))
> figurefile <- "/home/data/refined/reef/R/figures/genome_size"
>  clrs <- c(glasbey(), glasbey());
>   # make the colors a bit easier to read
>  clrs[3] <- clrs[19]
> #  tmp <- clrs[2]; clrs[2] <- clrs[6]; clrs[6] <- tmp
> #  tmp <- clrs[1]; clrs[1] <- clrs[12]; clrs[12] <- tmp
> #  tmp <- clrs[3]; clrs[3] <- clrs[12]; clrs[12] <- tmp
>
```

We begin with Eukaryota. Note that we have to also adjust for the observation that the number of reads mapped to Bellairs and Maycocks is significantly different (approximately $4.5M$ versus $9.2M$ reads respectively).

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

```
  name tax_id parent    rank embl_code division_id  br_bel  br_may
1 root      1     NA no rank                     8 4486003 9194536
  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           NA root
  genome_size isLeaf
1    21.40583  FALSE
```

```
> bel <- tree[1, "br_bel"]; may <- tree[1, "br_may"]
> genome_size_adjustments <- list()
> lqtile <- 0.1; rqtile <- 0.9
> tree <- original
> euk <- induce_tree(2759); euk$taxa <- "Eukaryota"
> euk_species <- euk[ which(euk$rank == "species"), ]
> euk_species_a <- euk_species; euk_species_a$site <- "bellairs"; euk_species_a$reads <- (euk_species$br_be
> euk_species_b <- euk_species; euk_species_b$site <- "maycocks"; euk_species_b$reads <- (euk_species$br_ma
> # added pseudocount above NO
> euk_tmp <- rbind(euk_species_a, euk_species_b)
> tmp <- arrange(euk_tmp, genome_size)
> tmp <- tmp[ which(!is.na(tmp$genome_size)), ]
> euk_tmp <- tmp[floor(nrow(tmp)*lqtile):floor(nrow(tmp)*(rqtile)),]
> # now we remove species with low number of reads
> euk_tmp <- euk_tmp[which(log(euk_tmp$reads) > -12), ]
> f <- lm( formula = log(reads) ~ log(genome_size), data = euk_tmp)
> p <- ggplot(euk_tmp, aes(x=log(genome_size), y=log(reads), color = site)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+          axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(shape=20, alpha=0.8)+
+   labs(title="Eukaryota",
+       x="Log Genome Size (Mbps)", y="Log fraction of reads (versus total reads for site)") +
+   geom_smooth(method=lm)  +
+   geom_abline(intercept = f$coefficients[1], slope = f$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> #ggsave( filename = "eukaryota.species.png", path = figurefile, device = "png", dpi = 300)
> f$coefficients

     (Intercept) log(genome_size)
      -9.2217193        0.0141026

> euk_tmp$log_reads_adj <- euk_tmp$reads
> genome_size_adjustments["eukaryota"] <- f
```

There is no evidence that the number of reads increases with genome size at least when analysis is performed in a manner where attention is restricted to only Eukaryota. Note that the slope of this fit is in fact negative if we do not remove the low count sites (log fraction reads is below -12).

```
> lqtile <- 0.0; rqtile <- 1
> tree <- original
> bel <- tree[1, "br_bel"]; may <- tree[1, "br_may"]
> virus <- induce_tree(10239); virus$taxa <- "virus"
> virus_species <- virus[ which(virus$rank == "species"), ]
> virus_species_a <- virus_species; virus_species_a$site <- "bellairs"; virus_species_a$reads <- virus_spec
> virus_species_b <- virus_species; virus_species_b$site <- "maycocks"; virus_species_b$reads <- virus_spec
> virus_tmp <- rbind(virus_species_a, virus_species_b)
> tmp <- arrange(virus_tmp, genome_size)
> tmp <- tmp[ which(!is.na(tmp$genome_size)), ]
> virus_tmp <- tmp[(floor(nrow(tmp)*lqtile))+1:floor(nrow(tmp)*(rqtile)),]
> virus_tmp <- virus_tmp[ which(log(virus_tmp$reads) > -14), ]
> f <- lm( formula = log(reads) ~ log(genome_size), data = virus_tmp)
> p <- ggplot(virus_tmp, aes(x=log(genome_size), y=log(reads), color = site)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+            axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(shape=20, alpha=0.8)+
+   labs(title="Viruses",
+       x="Log Genome Size (Mbps)", y="Log fraction of  reads (versus total reads for site)") +
+   geom_smooth(method=lm)  +
+   geom_abline(intercept = f$coefficients[1], slope = f$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> #ggsave( filename = "virus.species.png", path = figurefile, device = "png", dpi = 300)
> f$coefficients

     (Intercept) log(genome_size)
     -11.2573953        0.2050719

> virus_tmp$log_reads_adj <- log(virus_tmp$reads) -
+               (log(virus_tmp$genome_size)*f$coefficients[2])
> f_after <- lm( formula = log_reads_adj ~ log(genome_size), data = virus_tmp)
> p <- ggplot(virus_tmp, aes(x=log(genome_size), y=log_reads_adj, color = site)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+            axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(shape=20, alpha=0.8)+
+   labs(title="Viruses",
+        x="Log Genome Size (Mbps)", y="Log fraction of  reads (versus total reads for site)") +
+   geom_smooth(method=lm)  +
+   geom_abline(intercept = f_after$coefficients[1], slope = f_after$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> genome_size_adjustments["viruses"] <- f
```

```
> lqtile <- 0.0; rqtile <- 1
> tree <- original
> bel <- tree[1, "br_bel"]; may <- tree[1, "br_may"]
> arch <- induce_tree(2157); arch$taxa <- "Archaea"
> arch_species <- arch[ which(arch$rank == "species"), ]
> arch_species_a <- arch_species; arch_species_a$site <- "bellairs"; arch_species_a$reads <- arch_species$b
> arch_species_b <- arch_species; arch_species_b$site <- "maycocks"; arch_species_b$reads <- arch_species$b
> arch_tmp <- rbind(arch_species_a, arch_species_b)
> tmp <- arrange(arch_tmp, genome_size)
> tmp <- tmp[ which(!is.na(tmp$genome_size)), ]
> arch_tmp <- tmp[(floor(nrow(tmp)*lqtile))+1:floor(nrow(tmp)*(rqtile)),]
> arch_tmp <- arch_tmp[ which(log(arch_tmp$reads) > -14), ]
> f <- lm( formula = log(reads) ~ log(genome_size), data = arch_tmp)
> p <- ggplot(arch_tmp, aes(x=log(genome_size), y=log(reads), color = site)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+            axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(shape=20, alpha=0.8)+
+   labs(title="Archaea",
+        x="Log Genome Size (Mbps)", y="Log number of reads") +
+   geom_rug(outside = TRUE, color="slategray2")+
+   geom_smooth(method=lm)  +
+   geom_abline(intercept = f$coefficients[1], slope = f$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> #ggsave( filename = "archaea.species.png", path = figurefile, device = "png", dpi = 300)
> f$coefficients
```

```
(Intercept) log(genome_size)
-11.1575223        0.3201326
```

```
> arch_tmp$log_reads_adj <- log(arch_tmp$reads) -
+              (log(arch_tmp$genome_size)*f$coefficients[2])
> f_after <- lm( formula = log_reads_adj ~ log(genome_size), data = arch_tmp)
> p <- ggplot(arch_tmp, aes(x=log(genome_size), y=log_reads_adj, color = site)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+            axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(shape=20, alpha=0.8)+
+   labs(title="Archaea",
+        x="Log Genome Size (Mbps)", y="Log fraction of reads") +
+   geom_rug(outside = TRUE, color="slategray2")+
+   geom_smooth(method=lm)  +
+   geom_abline(intercept = f_after$coefficients[1], slope = f_after$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> genome_size_adjustments["archaea"] <- f
```

```
> lqtile <- 0.0; rqtile <- 1
> tree <- original
> bel <- tree[1, "br_bel"]; may <- tree[1, "br_may"]
> bac <- induce_tree(2) ; bac$taxa <- "Bacteria"
> bac_species <- bac[ which(bac$rank == "species"), ]
> bac_species_a <- bac_species; bac_species_a$site <- "bellairs"; bac_species_a$reads <- bac_species$br_bel
> bac_species_b <- bac_species; bac_species_b$site <- "maycocks"; bac_species_b$reads <- bac_species$br_may
> bac_tmp <- rbind(bac_species_a, bac_species_b)
> tmp <- arrange(bac_tmp, genome_size)
> tmp <- tmp[ which(!is.na(tmp$genome_size)), ]
> bac_tmp <- tmp[(floor(nrow(tmp)*lqtile))+1:floor(nrow(tmp)*(rqtile)),]
> bac_tmp <- bac_tmp[ which(log(bac_tmp$reads) > -14), ]
> f <- lm( formula = log(reads) ~ log(genome_size), data = bac_tmp)
> p <- ggplot(bac_tmp, aes(x=log(genome_size), y=log(reads), color = site)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+          axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(shape=20, alpha=0.8)+
+   labs(title="Bacteria",
+       x="Log Genome Size (Mbps)", y="Log fraction of reads") +
+   geom_rug(outside = TRUE, color="slategray2")+
+   geom_smooth(method=lm)  +
+   geom_abline(intercept = f$coefficients[1], slope = f$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> #ggsave( filename = "bacteria.species.png", path = figurefile, device = "png", dpi = 300)
> f$coefficients
```

```
    (Intercept) log(genome_size)
     -10.843223         0.420101
```

```
> bac_tmp$log_reads_adj <- log(bac_tmp$reads) -
+               (log(bac_tmp$genome_size)*f$coefficients[2])
> f_after <- lm( formula = log_reads_adj ~ log(genome_size), data = bac_tmp)
> p <- ggplot(bac_tmp, aes(x=log(genome_size), y=log_reads_adj, color = site)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+          axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(shape=20, alpha=0.8)+
+   labs(title="Bacteria",
+       x="Log Genome Size (Mbps)", y="Log fraction of reads") +
+   geom_rug(outside = TRUE, color="slategray2")+
+   geom_smooth(method=lm)  +
+   geom_abline(intercept = f_after$coefficients[1], slope = f_after$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> genome_size_adjustments["bacteria"] <- f
> lqtile <- 0.0; rqtile <- 1
> tree <- original
> tot_tmp <- do.call("rbind", list(euk_tmp, virus_tmp, bac_tmp, arch_tmp))
> tmp <- arrange(tot_tmp, genome_size)
> tmp <- tmp[ which(!is.na(tmp$genome_size)), ]
> tot_tmp <- tmp[(floor(nrow(tmp)*lqtile))+1:floor(nrow(tmp)*(rqtile)),]
> tot_tmp <- tot_tmp[ which(log(tot_tmp$reads) > -14), ]
> f <- lm( formula = log(reads) ~ log(genome_size), data = tot_tmp)
> p <- ggplot(tot_tmp, aes(x=log(genome_size), y=log(reads), color = taxa)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+            axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(aes(shape=factor(site)), alpha=0.8)+
+   labs(title="All",
+       x="Log Genome Size (Mbps)", y="Log fraction of reads") +
+   geom_rug(outside = TRUE, color="slategray2")+
+ #  geom_smooth(method=lm)  +
+   geom_abline(intercept = f$coefficients[1], slope = f$coefficients[2], color="purple",
+                   size=0.5, alpha = 0.75)
> p
> #ggsave( filename = "all.png", path = figurefile, device = "png", dpi = 300)
> f$coefficients

    (Intercept) log(genome_size)
    -10.8238770        0.4024304

> tot_tmp$log_reads_adj <- log(tot_tmp$reads) -
+              (log(tot_tmp$genome_size)*f$coefficients[2])
> f_after <- lm( formula = log_reads_adj ~ log(genome_size), data = tot_tmp)
> p <- ggplot(tot_tmp, aes(x=log(genome_size), y=log_reads_adj, color = taxa)) +
+   scale_color_manual(values=clrs) +
+   theme( axis.title.x = element_text(size = 11), axis.text.x = element_text(size = 8),
+            axis.title.y = element_text(size = 1))+
+   theme_tufte()+
+   geom_rug(outside = TRUE, color="slategray2")+
+   coord_cartesian(clip = "off") +
+   geom_point(aes(shape = factor(site)),  alpha=0.8)+
+   labs(title="All",
+       x="Log Genome Size (Mbps)", y="Log fraction of reads") +
+   geom_rug(outside = TRUE, color="slategray2")+
+ #  geom_smooth(method=lm)  +
+   geom_abline(intercept = f_after$coefficients[1], slope = f_after$coefficients[2], color="darkgreen",
+                   size=0.5, alpha = 0.75) +
+   geom_text(aes(x= -7, label="\nLinear Fit", y=f_after$coefficients[1]), colour="black", angle=0, text=el
+
+     geom_hline(yintercept=-5.8,  color = "black")+
+   geom_text(aes(x=-6.5, label="\nViruses", y=-5.8), colour="black", angle=0, text=element_text(size=3)) +
+   geom_text_repel(
+       max.iter=100000,
+       aes(label=subset(tot_tmp, ((taxa == "virus") & (log_reads_adj > -5.8 )))$name),
+       size = 2,
+       data = subset(tot_tmp, ((taxa == "virus") & (log_reads_adj > -5.8 ))),
+        segment.size  = 0.1,
+       nudge_x = -5,
+       nudge_y = +1
+      # nudge_x       = top_left,   segment.size  = 0.1,    direction     = "y",    hjust         = 0.5,
+     ) +
+
+   geom_hline(yintercept=-8,  color = "blue")+
+   geom_text(aes(x=-6.5, label="\nArchaea", y=-8), colour="blue", angle=0, text=element_text(size=3)) +
+   geom_text_repel(
+       max.iter=100000,
+       aes(label=subset(tot_tmp, ((taxa == "Archaea") & (log_reads_adj > -8 )))$name),
+       size = 2,
+       data = subset(tot_tmp, ((taxa == "Archaea") & (log_reads_adj > -8 ))),
+       segment.size  = 0.1,
+       nudge_x = -5,
+       nudge_y = +5
+      # nudge_x       = top_left,   segment.size  = 0.1,    direction     = "y",    hjust         = 0.5,
+     ) +
+
+   geom_hline(yintercept=-4.5,  color = "red")+
+   geom_text(aes(x=2.0, label="\nBacteria", y=-4.5), colour="red", angle=0, text=element_text(size=3)) +
+   geom_text_repel(
+       max.iter=100000,
+       aes(label=subset(tot_tmp, ((taxa == "Bacteria") & (log_reads_adj > -4.5 )))$name),
+       size = 2,
+       data = subset(tot_tmp, ((taxa == "Bacteria") & (log_reads_adj > -4.5 ))),
+       segment.size  = 0.1,
+       nudge_x = +1,
+       nudge_y = +1
+      # nudge_x       = top_left,   segment.size  = 0.1,    direction     = "y",    hjust         = 0.5,
+     ) +
+
+   geom_hline(yintercept=-8.5,  color = "purple") +
+   geom_text(aes(x=3.5, label="\nEukaryota", y=-8.5), colour="purple", angle=0, text=element_text(size=3)
+   geom_text_repel(
+       max.iter=100000,
+       aes(label=subset(tot_tmp, ((taxa == "Eukaryota") & (log_reads_adj > -8.5 )))$name),
+       size = 2,
+       data = subset(tot_tmp, ((taxa == "Eukaryota") & (log_reads_adj > -8.5 ))),
+       segment.size  = 0.1,
+       nudge_x = +1,
+       nudge_y = +1
+      # nudge_x       = top_left,   segment.size  = 0.1,    direction     = "y",    hjust         = 0.5,
+     )
> p
> #ggsave( filename = "all.adjusted.png", path = figurefile, device = "png", dpi = 300)
>
> genome_size_adjustments["all"] <- f
```

The R object that stores the linear fit f is stored in the `reef` folder. Finally we add to our `tree` data structure an attribute corresponding to the corrected read count corrected by f.

```
> # Write the linear models to file
> #save( file = file.path(REEF_DIR, "genome_size_adjustment_1.0.RData"), genome_size_adjustments)
```