Supplemental Information 3: Genome Size

Simpson, Bettauer et al.

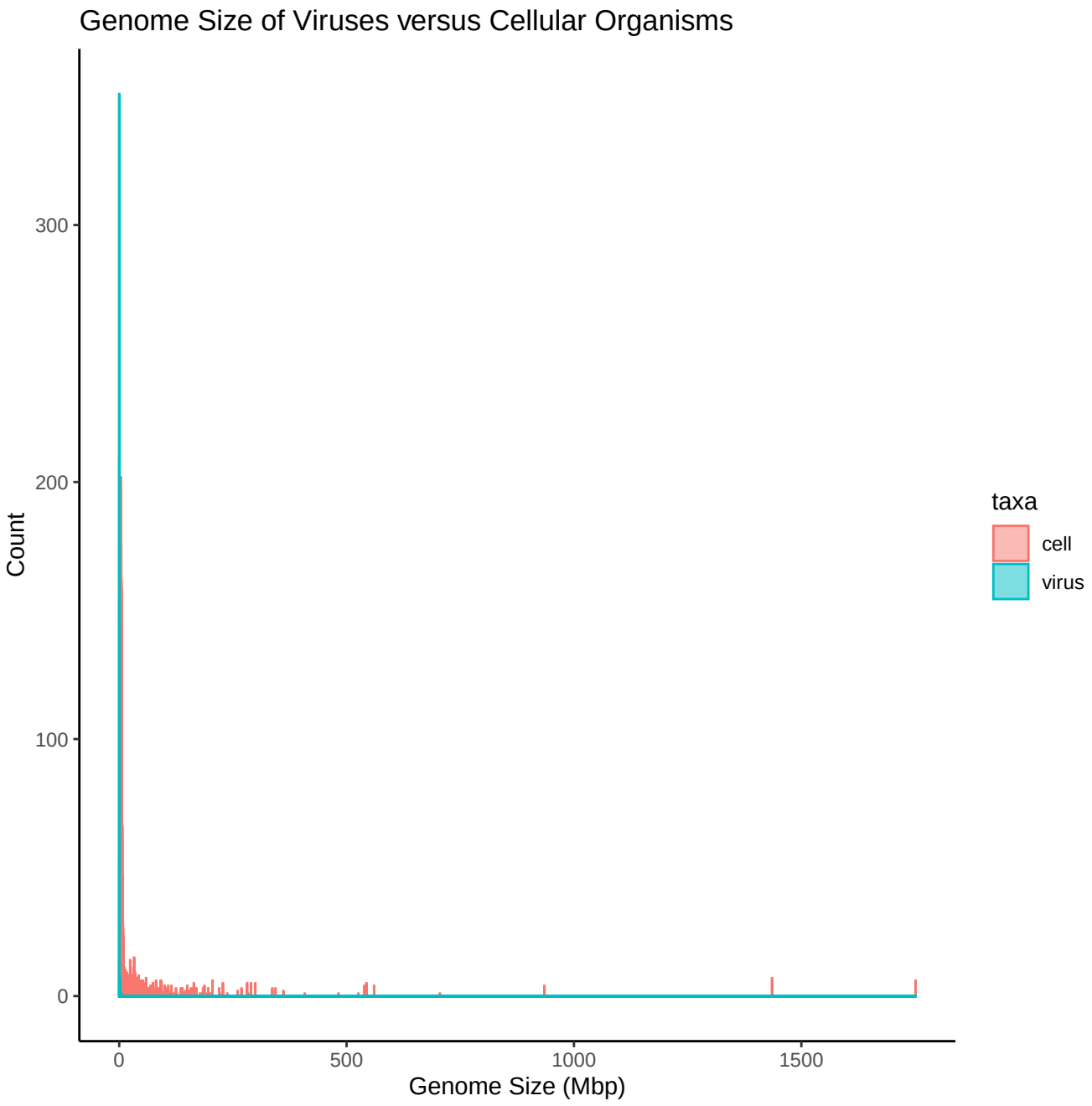
April 2020

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.

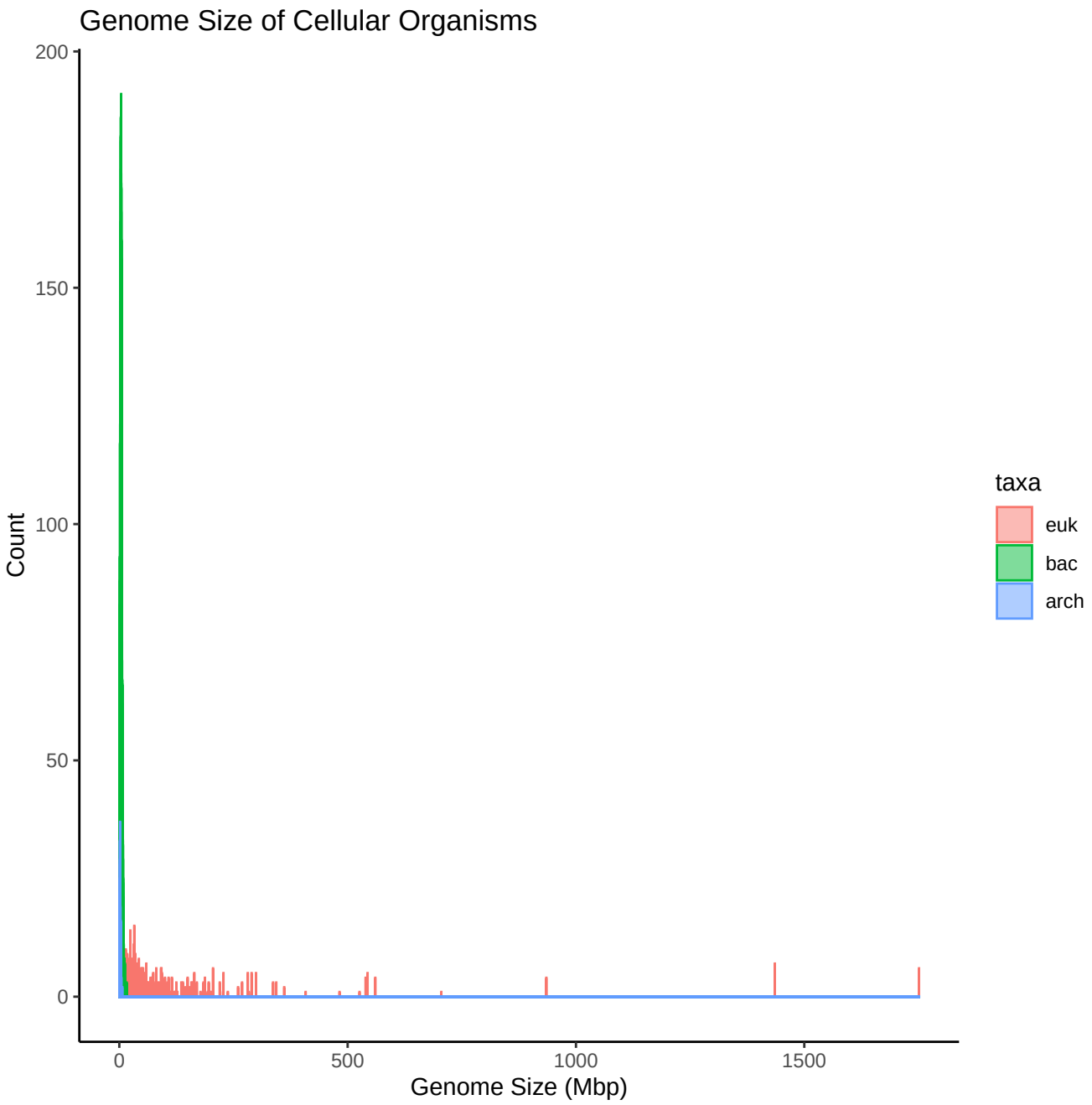
```
options(warn = -1)
library(xtable); library(ggplot2); library(vcd); library(MASS); library(FNN); library(rlang)
source("~/repo/reefmicrobiome/src/functions.R")

# Load the tree data.frame with Bracken counts etc.
REEF_DIR <- "/home/data/refined/reef/R/"
load( paste0(REEF_DIR, "pure.tree.april.15.RData" ) )  # loads tree data.frame
original <- tree # for safe keeping
date <- "april.19"
```

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



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_s
cell.hist <- rbind( cell.hist, data.frame( taxa = "arch", genome_size = arch$genome_size[!is.na(arch$genom
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
```



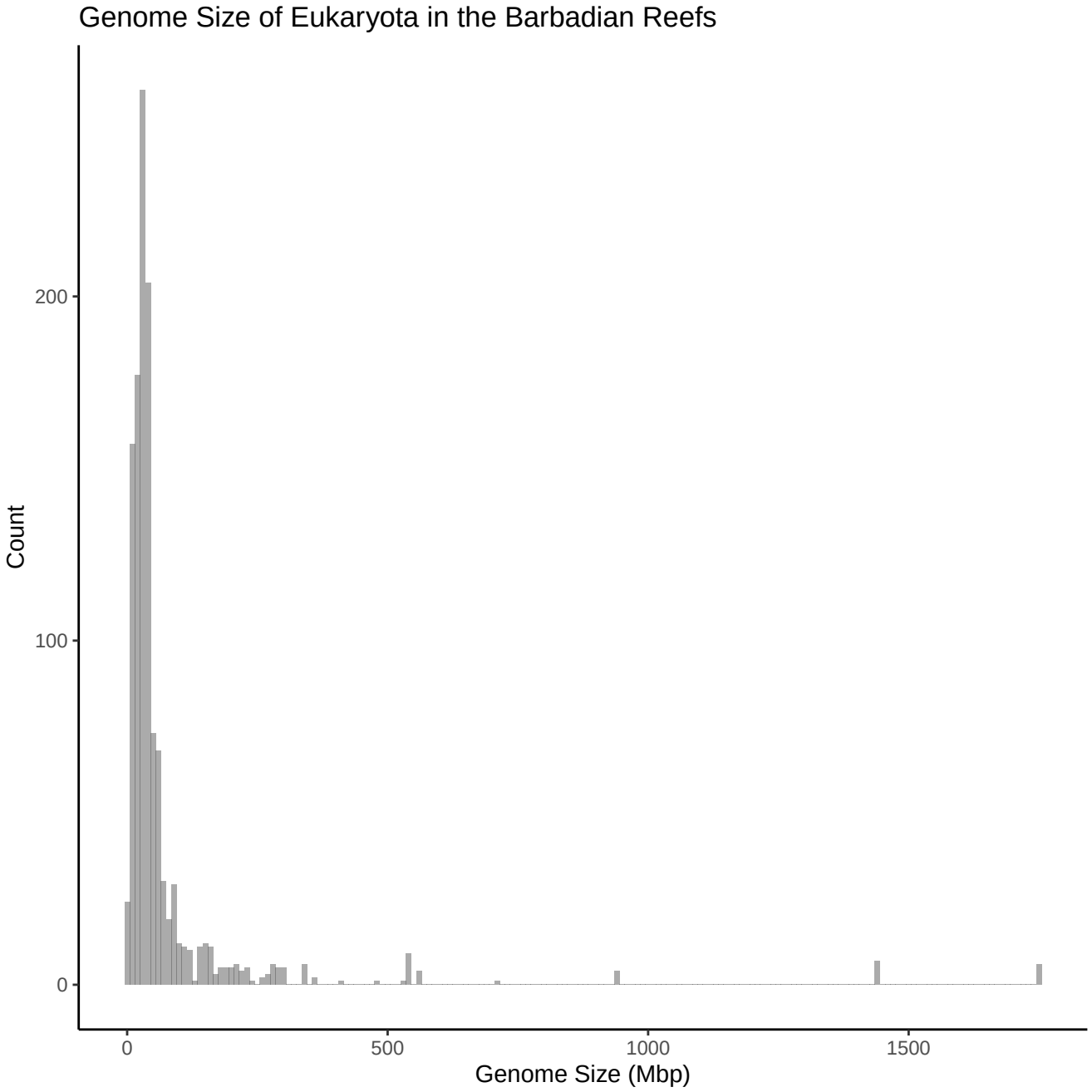
The largest bacterial genome in our data is Minicystis rosea at 16 Mbp and the largest Archaea genome is 6 Mbp.

```
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, "genome
 }
```

```
##
##  Chara braunii   69332 1751.21   root.cellular organisms.Eukaryota.Viridiplantae.Streptophyta.Streptoph
##  Euglena gracilis   3039 1435.5   root.cellular organisms.Eukaryota.NA.Euglenozoa.Euglenida.NA.Euglenop
##  Symbiodinium kawagutii   104179 935.067   root.cellular organisms.Eukaryota.Sar.Alveolata.Dinophyceae.S
##  Nephromyces sp. ex Molgula occidentalis   2544991 560.451   root.cellular organisms.Eukaryota.Sar.Alve
##  Hemileia vastatrix   203904 543.605   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Bas
##  Saccharina japonica   88149 539.4045   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Ochrophyta.
##  Dunaliella salina   3046 343.704   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core ch
##  Kappaphycus alvarezii   38544 336.721   root.cellular organisms.Eukaryota.Rhodophyta.Florideophyceae.NA
##  Digenea simplex   945030 299.321   root.cellular organisms.Eukaryota.Rhodophyta.Florideophyceae.NA.Cera
##  Mesostigma viride   41882 289.67   root.cellular organisms.Eukaryota.Viridiplantae.Streptophyta.Mesosti
##  Cymbomonas tetramitiformis   36881 281.27   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta
##  Haematococcus lacustris   44745 268.7845   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta
##  Tetraselmis striata   3165 227.954   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core
##  Oxytricha trifallax   1172189 219.9967   root.cellular organisms.Eukaryota.Sar.Alveolata.Ciliophora.Int
##  Physarum polycephalum   5791 205.176   root.cellular organisms.Eukaryota.NA.NA.Eumycetozoa.Myxogastria
##  Ectocarpus siliculosus   2880 195.811   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Ochrophyta
##  Chromera velia   505693 187.455   root.cellular organisms.Eukaryota.Sar.Alveolata.Colpodellida.NA.Chro
##  Botryococcus braunii   38881 184.382   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.cor
##  Phytophthora infestans   4787 177.7035   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.
##  Cladosiphon okamuranus   309737 169.731   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Ochrophy
##  Trichomonas vaginalis   5722 164.072   root.cellular organisms.Eukaryota.NA.Parabasalia.Trichomonadida
##  Cantharellus lutescens   104198 160.367   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Tetradesmus obliquus   3088 157.946   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core
##  NA   658196 153.876   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Fungi incertae sedis.Mucoro
##  Gonium pectorale   33097 148.806   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core ch
##  NA   588596 145.3173   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Fungi incertae sedis.Mucor
##  Yamagishiella unicocca   51707 137.536   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.c
##  Tetrabaena socialis   47790 135.78   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core
##  Plasmopara halstedii   4781 127.0636   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Pe
##  Tuber melanosporum   39416 124.946   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Chlamydomonas reinhardtii   3055 120.405   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta
##  Sphaeroforma arctica   72019 115.142   root.cellular organisms.Eukaryota.Opisthokonta.Ichthyosporea.Ich
##  Diplonema papillatum   91374 107.915   root.cellular organisms.Eukaryota.NA.Euglenozoa.Diplonemea.NA.D
##  Pyropia yezoensis   2788 107.591   root.cellular organisms.Eukaryota.Rhodophyta.Bangiophyceae.Bangiales
##  Ophiocordyceps sinensis   72228 106.603   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Chondrus crispus   2769 104.98   root.cellular organisms.Eukaryota.Rhodophyta.Florideophyceae.NA.Gigart
##  Cyanophora paradoxa   2762 99.9404   root.cellular organisms.Eukaryota.Glaucocystophyceae.Cyanophoracea
##  Ulva mutabilis   498180 98.4847   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.Ulvophyce
##  Chlorella sp. ArM0029B   1415603 92.9613   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta
##  Plasmopara viticola   143451 92.59207   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Pe
##  Thalassiosira oceanica   159749 92.1856   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Ochrophy
##  Psammoneis japonica   517775 91.4306   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Ochrophyta.
##  Ulva prolifera   3117 87.8893   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.Ulvophycea
##  Porphyra umbilicalis   2786 87.889   root.cellular organisms.Eukaryota.Rhodophyta.Bangiophyceae.Bangial
##  Phytophthora sojae   67593 84.23815   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Pero
##  Acanthamoeba castellanii   5755 80.8823   root.cellular organisms.Eukaryota.NA.NA.Longamoebia.NA.Acantl
##  Chlamydomonas applanata   35704 78.5042   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.
##  Amanita bisporigera   87325 75.3463   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Bas
##  Chlorokybus atmophyticus   3144 74.3303   root.cellular organisms.Eukaryota.Viridiplantae.Streptophyta
##  Hyphochytrium catenoides   42384 73.08465   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Hyphoch
##  Hyaloperonospora arabidopsidis   272952 72.35685   root.cellular organisms.Eukaryota.Sar.Stramenopiles
##  Paramecium tetraurelia   5888 72.0945   root.cellular organisms.Eukaryota.Sar.Alveolata.Ciliophora.Intr
##  Ganoderma boninense   34458 69.75715   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Bas
##  Monoraphidium neglectum   145388 69.7118   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta
##  Asterionella formosa   210441 68.4198   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Ochrophyta
```

```
##  Sterkiella histriomuscorum   94289 66.3686   root.cellular organisms.Eukaryota.Sar.Alveolata.Ciliophora
##  Diaporthe helianthi   158607 63.672   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Rostrostelium ellipticum   361140 62.1602   root.cellular organisms.Eukaryota.NA.NA.Eumycetozoa.Dictyo
##  Aphanomyces invadans   157072 61.7834   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Sa
##  Toxoplasma gondii   5811 61.5763   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa.Conoida
##  NA   554055 61.0189   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core chlorophytes.Tr
##  Eimeria mitis   44415 60.4151   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa.Conoidasid
##  Armillaria ostoyae   47428 60.1068   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Basi
##  Colletotrichum graminicola   31870 59.9141   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dika
##  Aphanomyces astaci   112090 59.69445   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Sa
##  Phytophthora cinnamomi   4785 59.63075   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.
##  Parachlorella kessleri   3074 59.1878   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.co
##  Chrysochromulina tobinii   1460289 59.0731   root.cellular organisms.Eukaryota.NA.Haptophyta.NA.Prymne
##  Besnoitia besnoiti   94643 58.8459   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa.Conoi
##  Moneuplotes crassus   5936 58.5666   root.cellular organisms.Eukaryota.Sar.Alveolata.Ciliophora.Intram
##  Chlorella sorokiniana   3076 58.33862   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.co
##  NA   690256 57.55417   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.sacchar
##  Mastigamoeba balamuthi   108607 57.2666   root.cellular organisms.Eukaryota.NA.NA.NA.NA.Mastigamoebida
##  Colletotrichum gloeosporioides   474922 57.2517   root.cellular organisms.Eukaryota.Opisthokonta.Fungi
##  Auxenochlorella pyrenoidosa   3078 56.993   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyt
##  Aphanomyces euteiches   100861 56.9046   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.
##  Aureococcus anophagefferens   44056 56.6606   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Pela
##  Rhizoctonia solani   456999 56.0285   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Bas
##  Salpingoeca rosetta   946362 55.4403   root.cellular organisms.Eukaryota.Opisthokonta.Choanoflagellata
##  Clonostachys rosea   29856 55.30035   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Eimeria necatrix   51315 55.0079   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa.Conoida
##  Amorphotheca resinae   5101 54.4745   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Phytophthora capsici   4784 53.41358   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Pe
##  Porodaedalea pini   108901 53.3469   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Basi
##  Pyropia haitanensis   1262161 53.2546   root.cellular organisms.Eukaryota.Rhodophyta.Bangiophyceae.Ban
##  Fusarium oxysporum   5507 52.02895   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Stichococcus bacillaris   37433 51.9927   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.
##  Raphidocelis subcapitata   307507 51.1627   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyt
##  Stylonychia lemnae   5949 50.1645   root.cellular organisms.Eukaryota.Sar.Alveolata.Ciliophora.Intrama
##  Morchella importuna   1174673 49.9687   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.A
##  Fistulifera solaris   1519565 49.7366   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Ochrophyta
##  Colletotrichum higginsianum   80884 49.4357   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dik
##  Phialocephala scopiformis   149040 48.8763   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dika
##  Ichthyophthirius multifiliis   5932 48.8   root.cellular organisms.Eukaryota.Sar.Alveolata.Ciliophora.
##  Coremiostelium polycephalum   142831 48.621   root.cellular organisms.Eukaryota.NA.NA.Eumycetozoa.Dict
##  Colletotrichum orchidophilum   1209926 48.5565   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.
##  NA   1578925 48.3379   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.sacchar
##  Phytophthora parasitica   4792 47.7626   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.
##  NA   554065 46.1595   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core chlorophytes.Tr
##  Leptosphaeria maculans   5022 45.9865   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.A
##  Eimeria maxima   5804 45.9751   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa.Conoidasid
##  Eimeria acervulina   5801 45.8306   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa.Conoid
##  Trichoderma virens   29875 45.8201   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Fusarium solani   169388 45.8133   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomy
##  Fusarium proliferatum   948311 45.78918   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Aspergillus mulundensis   1810919 45.3419   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikar
##  Rhizopus oryzae   64495 45.04514   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Fungi incertae
##  Fusarium fujikuroi   5127 44.92106   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Agrocybe aegerita   5400 44.7908   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Basidi
##  Chrysoporthe austroafricana   354353 44.6689   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Di
##  Pyricularia grisea   148305 44.31875   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
```

```
##  Pseudocercospora fijiensis   1873960 44.1245   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Di
##  Pythium insidiosum   114742 44.08819   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Py
##  Cyclospora cayetanensis   88456 43.98724   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa
##  NA   563466 43.82915   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.sacchar
##  Eimeria falciformis   84963 43.6713   root.cellular organisms.Eukaryota.Sar.Alveolata.Apicomplexa.Cono
##  Achlya hypogyna   1202772 43.3985   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.Saprol
##  Venturia effusa   50376 42.9507   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomyc
##  Lobosporangium transversale   64571 42.7689   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Fung
##  [Nectria] haematococca   140110 42.65957   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Fusarium verticillioides   117187 42.40943   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dika
##  Trichosporon coremiiforme   82509 42.3533   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikar
##  NA   569365 42.30655   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.sacchar
##  Lentinula edodes   5353 41.46406   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Basidi
##  Crithidia fasciculata   5656 41.2974   root.cellular organisms.Eukaryota.NA.Euglenozoa.Kinetoplastea.M
##  Trypanosoma congolense   5692 41.2334   root.cellular organisms.Eukaryota.NA.Euglenozoa.Kinetoplastea.
##  Annulohypoxylon stygium   326628 41.1319   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Chlorella vulgaris   3077 41.10782   root.cellular organisms.Eukaryota.Viridiplantae.Chlorophyta.core
##  Neurospora crassa   5141 40.98185   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascom
##  Naegleria gruberi   5762 40.9641   root.cellular organisms.Eukaryota.NA.Heterolobosea.NA.NA.Vahlkampfi
##  Trichosporon ovoides   82524 40.9337   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ba
##  Phialemoniopsis curvata   1093900 40.3666   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikar
##  NA   568076 40.3173   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.saccharo
##  Trichoderma harzianum   5544 40.25984   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.A
##  Aspergillus alliaceus   209559 40.15425   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Aspergillus sojae   41058 40.054   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomy
##  Fusarium culmorum   5516 40.0196   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomy
##  Aspergillus caelatus   61420 40.0164   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Arthrobotrys oligospora   13349 40.00234   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Bodo saltans   75058 39.8644   root.cellular organisms.Eukaryota.NA.Euglenozoa.Kinetoplastea.Metakinet
##  Schizophyllum commune   5334 39.7789   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ba
##  Hypoxylon pulicicidum   1243767 39.6241   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Pyricularia oryzae   318829 39.4074   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Cryphonectria parasitica   5116 39.2611   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Phanerochaete chrysosporium   5306 39.2051   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dika
##  Thraustotheca clavata   74557 39.1007   root.cellular organisms.Eukaryota.Sar.Stramenopiles.Oomycota.S
##  Sparassis crispa   139825 39.0203   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Basid
##  Sordaria macrospora   5147 38.8634   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Fusarium venenatum   56646 38.6602   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Dichomitus squalens   114155 38.59503   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ba
##  NA   1725355 38.5165   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.sacchar
##  Paraphaeosphaeria sporulosa   1460663 38.464   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Di
##  Aspergillus pseudotamarii   132259 38.2439   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dika
##  Exophiala oligosperma   215243 38.2245   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.
##  Trichoderma gamsii   398673 38.1993   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Zymoseptoria tritici   1047171 38.1077   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.
##  Trichoderma asperellum   101201 38.0794   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Parastagonospora nodorum   13684 38.03676   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikar
##  Aspergillus welwitschiae   1341132 37.84645   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dik
##  Neurospora discreta   29879 37.76317   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Aspergillus oryzae   5062 37.63278   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Aspergillus bombycis   109264 37.4746   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.A
##  Trametes hirsuta   5327 37.434   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Basidiom
##  Purpureocillium lilacinum   33203 37.4171   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikar
##  Fusarium coffeatum   231269 37.40235   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Aspergillus pseudonomius   1506151 37.24685   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dik
##  Endocarpon pusillum   364733 37.1732   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
```

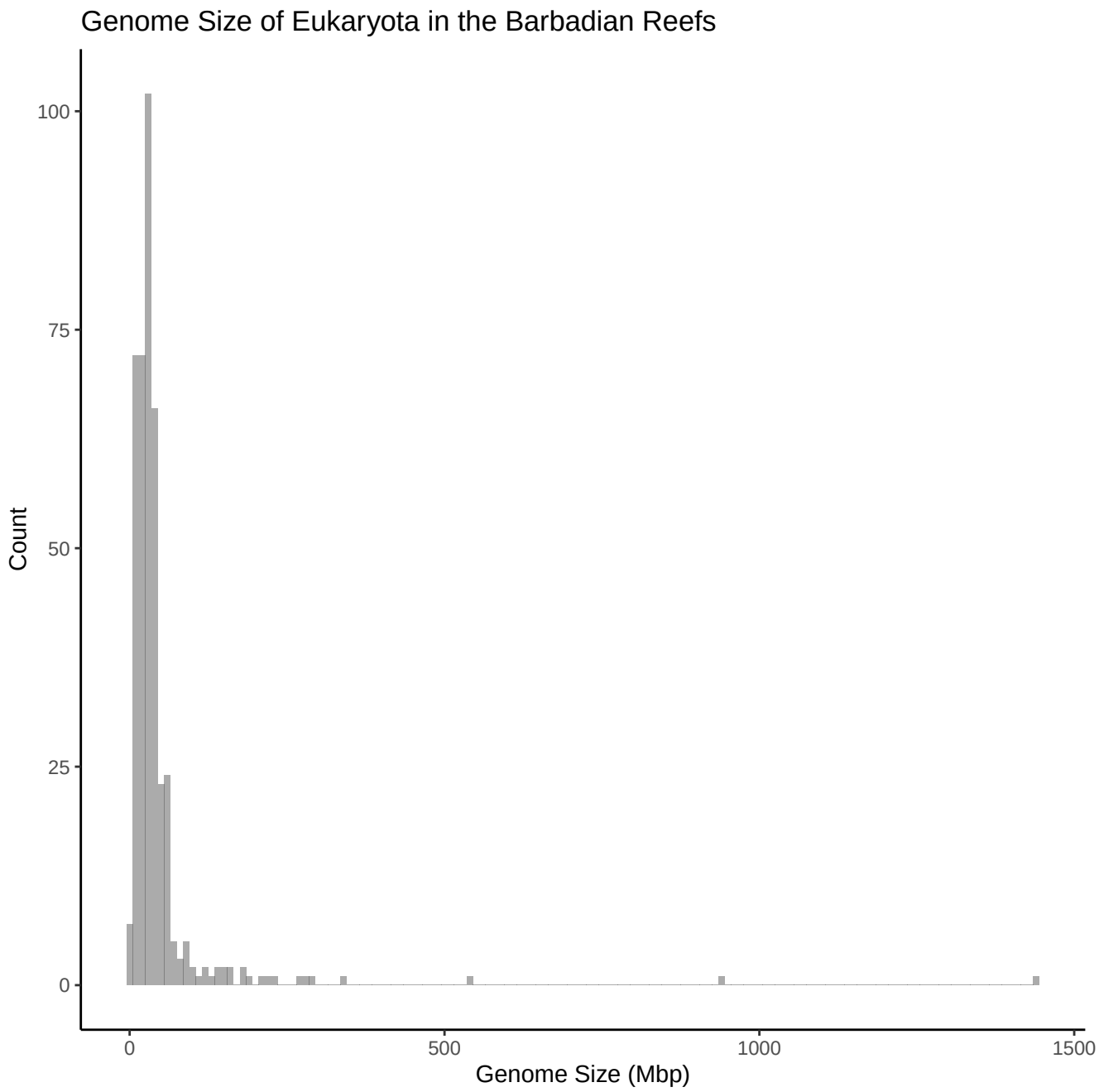
```
##  Cladophialophora bantiana   89940 37.0879   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikar
##  Aspergillus flavus   5059 37.00717   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Fusarium graminearum   5518 36.80952   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Fusarium pseudograminearum   101028 36.70635   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Di
##  NA   2587410 36.5797   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.sacchar
##  Trichoderma atroviride   63577 36.53947   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Bipolaris sorokiniana   45130 36.5329   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Flammulina velutipes   38945 36.4597   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Bas
##  Cercospora beticola   122368 36.28903   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Mucor circinelloides   36080 36.2576   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Fungi incer
##  Talaromyces pinophilus   128442 35.8832   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  NA   655981 35.8182   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.saccharo
##  Trypanosoma cruzi   5693 35.76182   root.cellular organisms.Eukaryota.NA.Euglenozoa.Kinetoplastea.Metal
##  Beauveria bassiana   176275 35.63905   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Pyrenophora tritici-repentis   45151 35.51074   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Di
##  Drechslerella brochopaga   47238 35.4316   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Aspergillus niger   5061 35.42943   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascom
##  Aspergillus nomius   41061 35.2805   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asco
##  Fonsecaea monophora   254056 35.2298   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Cafeteria roenbergensis   33653 35.04308   root.cellular organisms.Eukaryota.Sar.Stramenopiles.NA.NA.B
##  Diplodia corticola   236234 34.9861   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Fonsecaea erecta   1367422 34.748   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascom
##  Pleurotus ostreatus   5322 34.72905   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Bas
##  NA   1659845 34.5224   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Basidiomycota.Agar
##  NA   2587412 34.5062   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomycota.sacchar
##  Exophiala lecanii-corni   91925 34.4207   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Paramoeba pemaquidensis   180228 34.387   root.cellular organisms.Eukaryota.NA.NA.Flabellinia.NA.Param
##  Podospora comata   48703 34.3855   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascomy
##  Verticillium dahliae   27337 34.03032   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Alternaria alternata   5599 33.95789   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Asc
##  Alternaria arborescens   156630 33.83602   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya
##  Lachnellula hyalina   1316788 33.8283   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.As
##  Fonsecaea nubica   856822 33.7874   root.cellular organisms.Eukaryota.Opisthokonta.Fungi.Dikarya.Ascom
```

```
# to_kill <- c(69332, 2544991, 88149, 38544, 945030, 2880, 309737, 104198, 658196, 39416, 2788, 72228, 276
#         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.

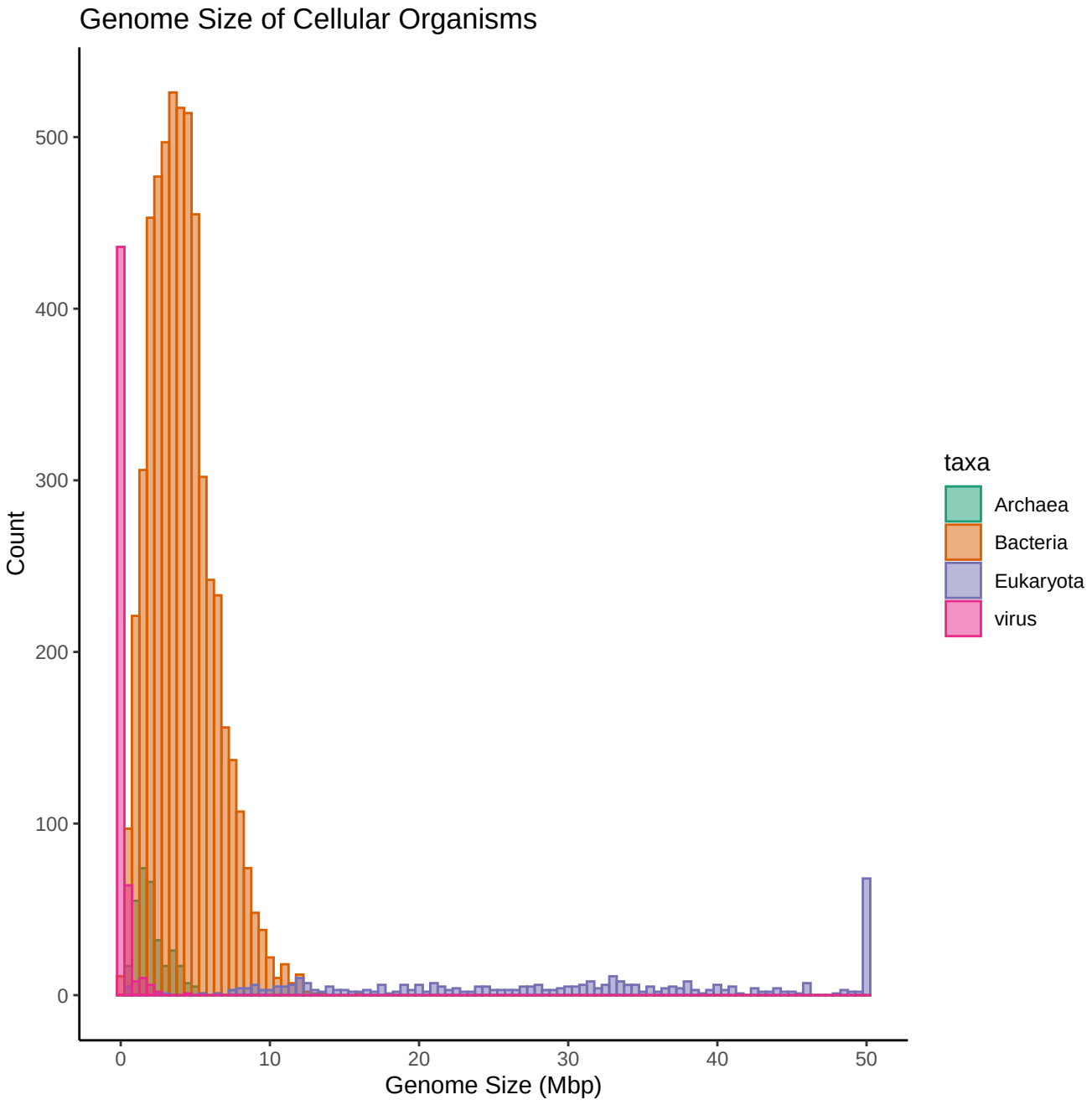
```
load(file = paste0(paste0("/home/data/refined/reef/R/ultra.pure.tree.", date), ".RData"))

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))
```

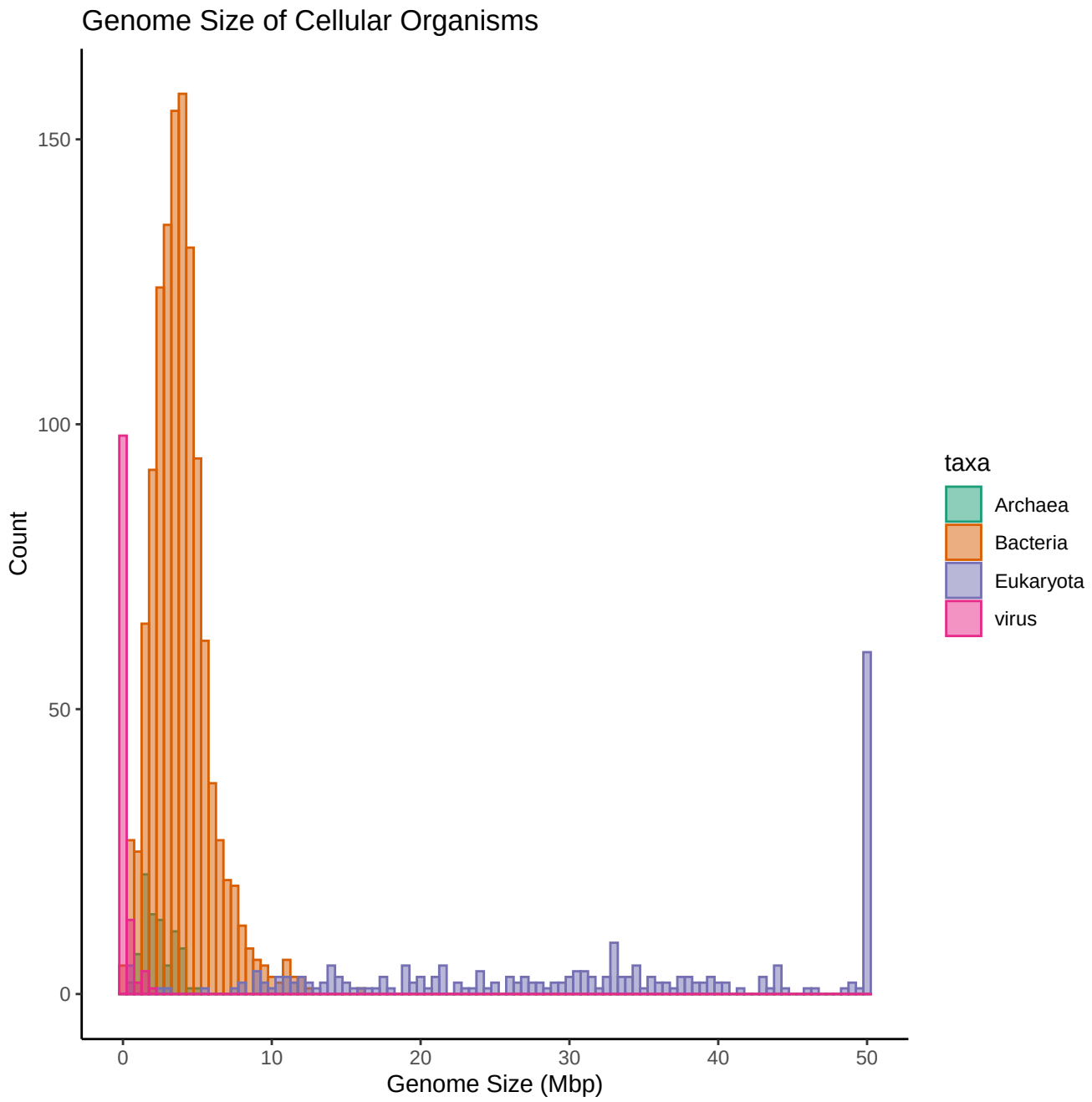
```
largest_species <- largest[ which(largest$rank == "species"), ]

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

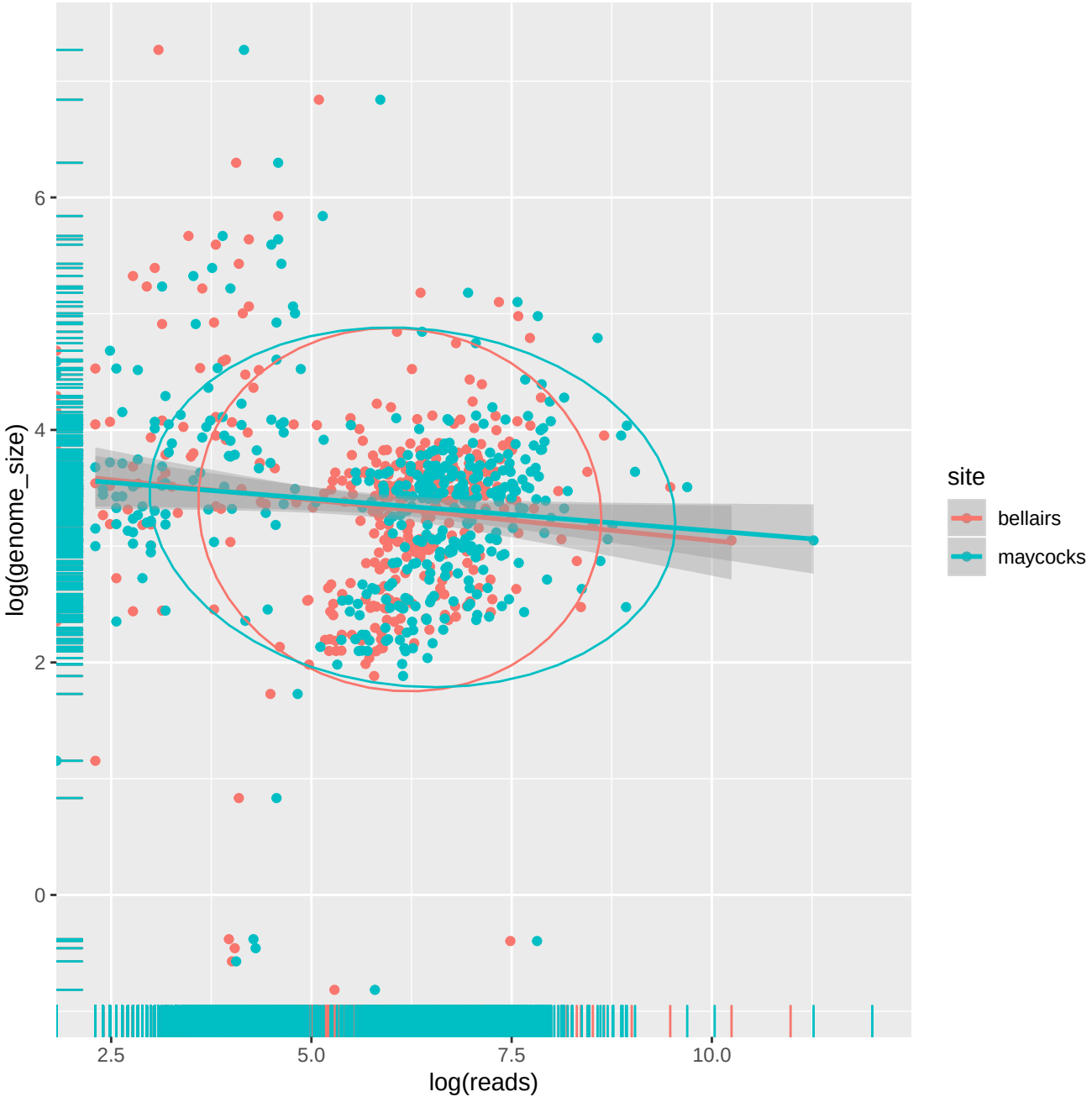
```
largest_species <- largest[ which(largest$rank == "genus"), ]
largest_species$genome_size <- Winsorize(largest_species$genome_size, maxval = 50, na.rm=TRUE)

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
```



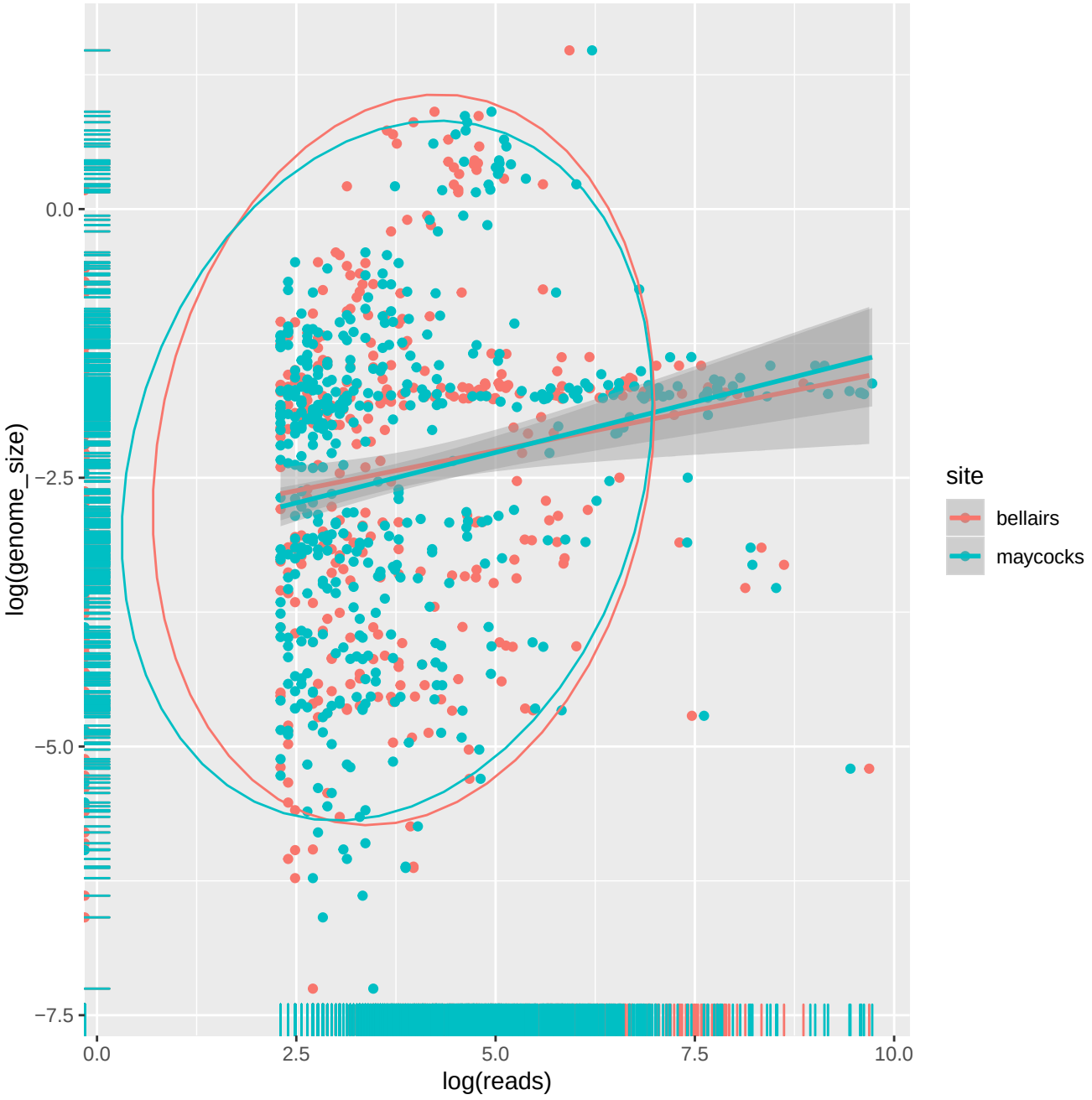
1 Correlations between genome size and read count

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.

```
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$br_bel
euk_species_b <- euk_species; euk_species_b$site <- "maycocks"; euk_species_b$reads <- euk_species$br_may
euk_tmp <- rbind(euk_species_a, euk_species_b)
ggplot(euk_tmp, aes(x=log(reads), y=log(genome_size), color = site)) +
  geom_point() + geom_rug()+  geom_smooth(method=lm) +
  stat_ellipse()
```

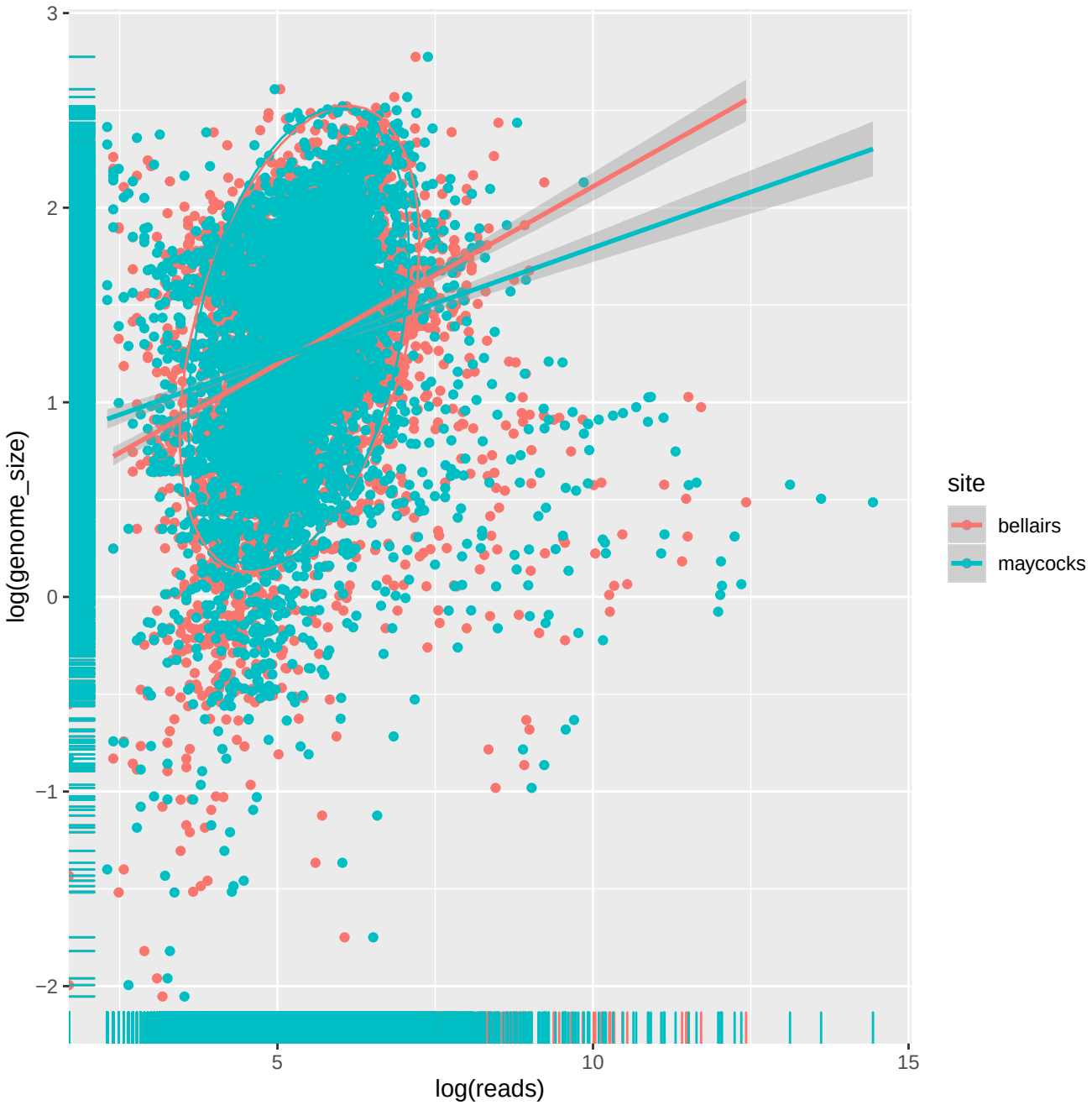


```
virus <- induce_tree(10239); virus$taxa <- "virus"
virus_species <- virus[ which(virus$rank == "species"), ]
virus_species_a <- virus_species; virus_species_a$site <- "bellairs"; virus_species_a$reads <- virus_speci
virus_species_b <- virus_species; virus_species_b$site <- "maycocks"; virus_species_b$reads <- virus_speci
virus_tmp <- rbind(virus_species_a, virus_species_b)
ggplot(virus_tmp, aes(x=log(reads), y=log(genome_size), color = site)) +
  geom_point() + geom_rug()+ geom_smooth(method=lm) +
  stat_ellipse()
```

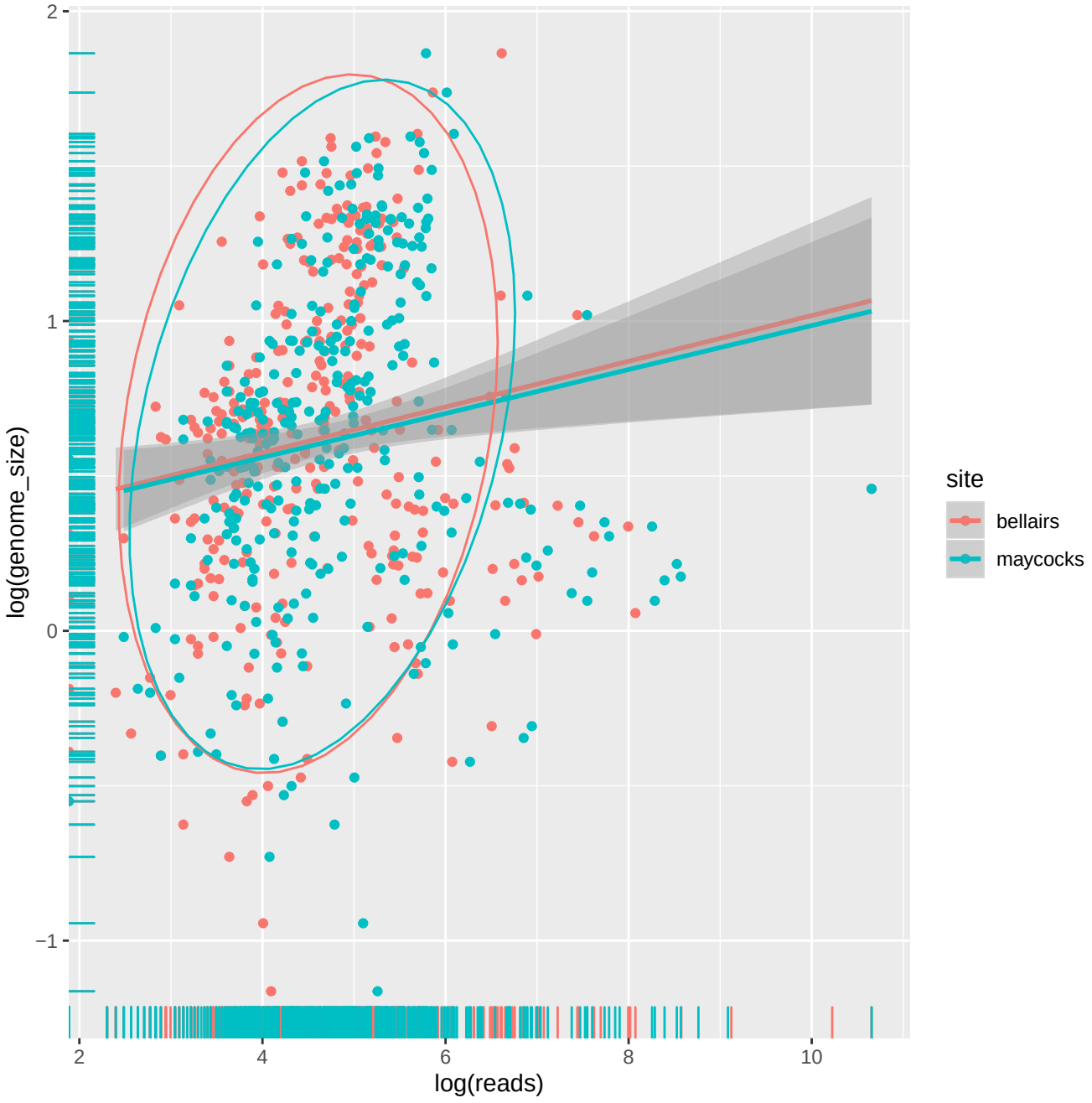



```
bac <- induce_tree(2) ; bac$taxa <- "Bacteria"
bac_species <- bac[ which(bac$rank == "species"), ]
bac_species_a <- bac_species; bac_species_a$site <- "bellairs"; bac_species_a$reads <- bac_species$br_bel
bac_species_b <- bac_species; bac_species_b$site <- "maycocks"; bac_species_b$reads <- bac_species$br_may
bac_tmp <- rbind(bac_species_a, bac_species_b)
```

```
ggplot(bac_tmp, aes(x=log(reads), y=log(genome_size), color = site)) +
  geom_point() + geom_rug()+ geom_smooth(method=lm) +
  stat_ellipse()
```




```
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$br
arch_species_b <- arch_species; arch_species_b$site <- "maycocks"; arch_species_b$reads <- arch_species$br
arch_tmp <- rbind(arch_species_a, arch_species_b)
ggplot(arch_tmp, aes(x=log(reads), y=log(genome_size), color = site)) +
  geom_point() + geom_rug()+ geom_smooth(method=lm) +
  stat_ellipse()
```



2 Correcting for genome size

To renormalize the counts what if we create a ratio based on the totals number of counts sequenced at specific location and divide it by the length of a genome size. A given organism's counts by that ratio. - Shawn

2.1 A

ttempting method mentioned above.

```
bel_tot <- 4708322 #Tot # of BellairsCounts
may_tot <- 10181105 #Tot # of MaycocksCounts
colnames(gnempool)[6] <-c("gn_size")
```