Genome project tables in the genomes package

Chris Stubben

April 11, 2014

The `genomes` package collects genome project metadata from NCBI using E-utility scripts (esearch, esummary, efetch and elink) or from the ENA using the ENA Browser REST URL. The packages also includes genome tables from NCBI and and provides tools to summarize, compare and plot the data in the R programming environment. Genome tables are a defined class (*genomes*) and each table is a data frame where rows are genome projects and columns are the fields describing the associated metadata. A number of methods are available that operate on genome tables including `print`, `summary`, `plot` and `update`.

There are a number of ways to install this package. If you are running the most recent R version, you can use the `biocLite` command.

```
R> source("http://bioconductor.org/biocLite.R")
R> biocLite("genomes")
```

Since the format of online genome tables may change (and then `update` commands may fail), I would recommend downloading the development version for fixes in between the six month release cycle.

```
R> install.packages("genomes",
    repos="http://www.bioconductor.org/packages/devel/bioC", type="source")
```

Genome tables from the Genome database at NCBI include prokaryotic (`proks`), eukaryotic (`euks`) and virus genomes (`virus`). The `print` methods displays the first few rows and columns of the table (either select less than seven rows or convert the object to a `data.frame` to print all columns). The `summary` function displays the download date, a count of projects by status, and a list of recent submissions. The `plot` method displays a cumulative plot of genomes by release date.

```
R> data(proks)
R> proks

  A genomes data.frame with 20064 rows and 23 columns

        pid                                                 name
```

```
1      33011                Abiotrophia defectiva ATCC 49176
2     174970                Acaricomes phytoseiuli DSM 14247
3      12997                  Acaryochloris marina MBIC11017
4      16707                    Acaryochloris sp. CCMEE 5410
5     197021 Acetanaerobacterium sp. hmp_mda_pilot_jcvi_0106
...      ...                                              ...
20064 182445           Zymophilus raffinosivorans DSM 20765
                   status   released ...
1     Scaffolds or contigs 2009-03-17 ...
2     Scaffolds or contigs 2013-04-20 ...
3                 Complete 2007-10-16 ...
4     Scaffolds or contigs 2011-06-03 ...
5            SRA or Traces       <NA> ...
...                    ...        ... ...
20064 Scaffolds or contigs 2013-04-23 ...
```

```
R> summary(proks)

$`Total genomes`
[1] 20064 genome projects on Jan 07, 2014

$`By status`
                     Total
Scaffolds or contigs 13126
SRA or Traces         4147
Complete              2791

$`Recent submissions`
  released   name                               status
1 2014-01-03 Bifidobacterium sp. MSTE12         Scaffolds or contigs
2 2014-01-03 Lachnospiraceae bacterium MSX33    Scaffolds or contigs
3 2013-12-31 Enterobacter cloacae P101          Complete
4 2013-12-31 Lactobacillus florum 8D            Scaffolds or contigs
5 2013-12-31 Morganella sp. EGD-HP17            Scaffolds or contigs
```

```
R> plot(proks, log='y', las=1)
R>
```

Most importantly, the `update` method downloads the latest version of the table from NCBI and displays a message listing the number of project IDs added and removed (not run).

```
R> update(proks)
```

A number of additional functions assist in selecting, sorting and grouping genomes. The `species` and `genus` functions can be used to extract the species or genus from a scientific name. The `table2` function formats and sorts a contingency table by counts.

```
R> spp<-species(proks$name)
R> table2(spp)
```

```
                          Total
Staphylococcus aureus      2382
Escherichia coli           1764
Salmonella enterica         989
Mycobacterium tuberculosis  700
Acinetobacter baumannii     466
Helicobacter pylori         381
Enterococcus faecalis       349
Streptococcus agalactiae    302
Streptococcus pneumoniae    295
Enterococcus faecium        262
```

The `month` and `year` functions can be used to extract the month or year from the release date (Figure 1).

```
R> complete <- subset(proks, status == "Complete")
R> x <- table(year(complete$released))
R> barplot(x, col="blue", ylim=c(0,max(x)*1.04), space=0.5, las=1,
 axis.lty=1, xlab="Year", ylab="Genomes per year")
R> box()
```

Because subsets of tables are often needed, the binary operator `like` allows pattern matching using wildcards. The `plotby` function can then be used to plot the release dates by status using labeled points, in this case to identify complete and draft sequences of *Yersinia pestis* released before 2012 (Figure 2).

```
R> ## Yersinia pestis
R> yp<-subset(proks, name %like% 'Yersinia pestis*' & year(released)<2012 )
R> plotby(yp, labels=TRUE, cex=.5, lbty='n', curdate=FALSE)
R>
```

A number of recent functions have been added that allow R users to query NCBI databases or the European Nucleotide Archive. These functions will be described in a separate vignette.

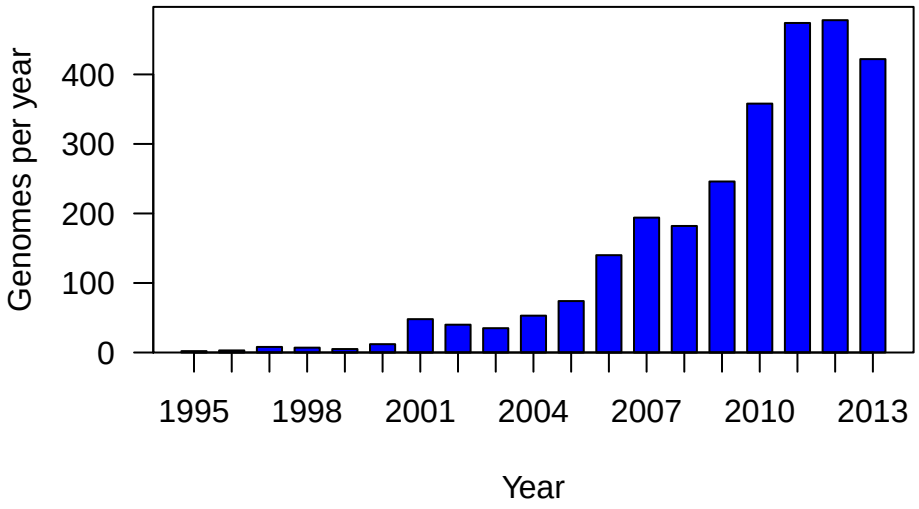

Figure 1: Number of complete microbial genomes released each year at NCBI

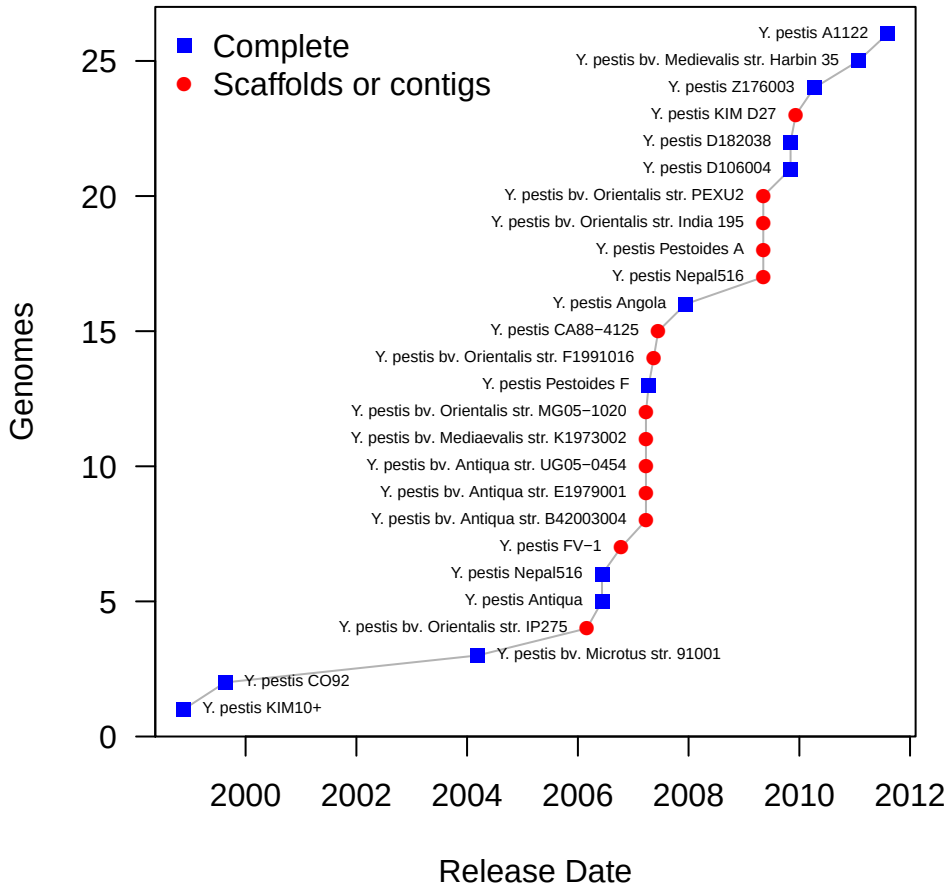

Figure 2: Cumulative plot of *Yersinia pestis* genomes by release date.