Genome project tables in the genomes package

Chris Stubben

October 8, 2013

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 24105 rows and 20 columns

       pid                                                name        status
```

```
1      55729                Abiotrophia defectiva ATCC 49176 SRA or Traces
2     199097                Acaricomes phytoseiuli DSM 14247      Assembly
3      58167                  Acaryochloris marina MBIC11017      Complete
4      78283                    Acaryochloris sp. CCMEE 5410      Assembly
5     197021 Acetanaerobacterium sp. hmp_mda_pilot_jcvi_0106 SRA or Traces
...      ...                                                ...           ...
24105 200373           Zymophilus raffinosivorans DSM 20765      Assembly
        released ...
1           <NA> ...
2     2013-04-20 ...
3     2007-10-16 ...
4     2011-06-03 ...
5           <NA> ...
...          ... ...
24105 2013-04-23 ...
```

```
R> summary(proks)

$`Total genomes`
[1] 24105 genome projects on Oct 08, 2013

$`By status`
              Total
Assembly      11497
No data        8229
Complete       2652
SRA or Traces  1727

$`Recent submissions`
  released   name                                          status
1 2013-09-16 Campylobacter jejuni subsp. jejuni 00-2425 Complete
2 2013-09-16 Cronobacter sakazakii 8399                  Assembly
3 2013-09-16 Enterococcus faecalis SDVK1A                Assembly
4 2013-09-16 Morganella morganii subsp. morganii GM1DA1 Assembly
5 2013-09-16 Pantoea sp. AS-PWVM4                        Assembly

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
Escherichia coli           1958
Staphylococcus aureus      1529
Salmonella enterica        1446
Acinetobacter baumannii    1033
Mycobacterium tuberculosis  608
Enterococcus faecalis       377
Streptococcus agalactiae    365
Helicobacter pylori         347
Streptococcus pneumoniae    298
Enterococcus faecium        294
```

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.

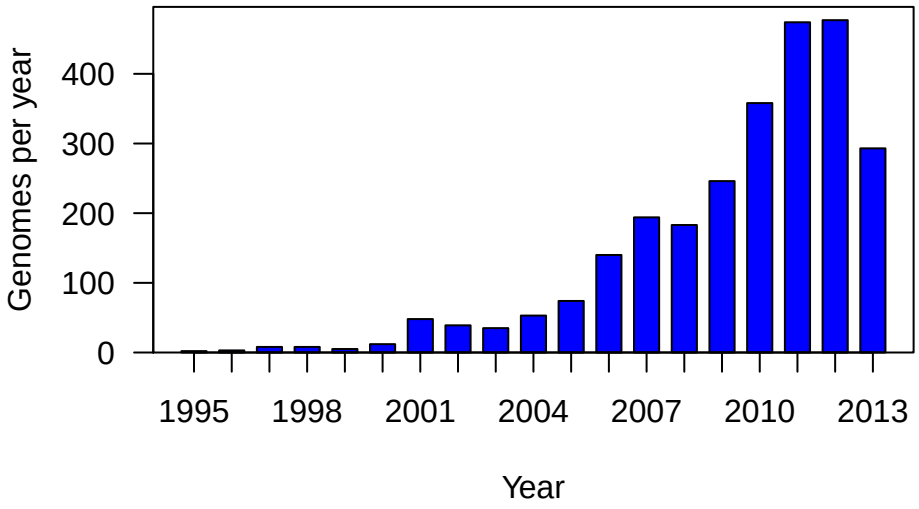

Figure 1: Number of complete microbial genomes released each year at NCBI

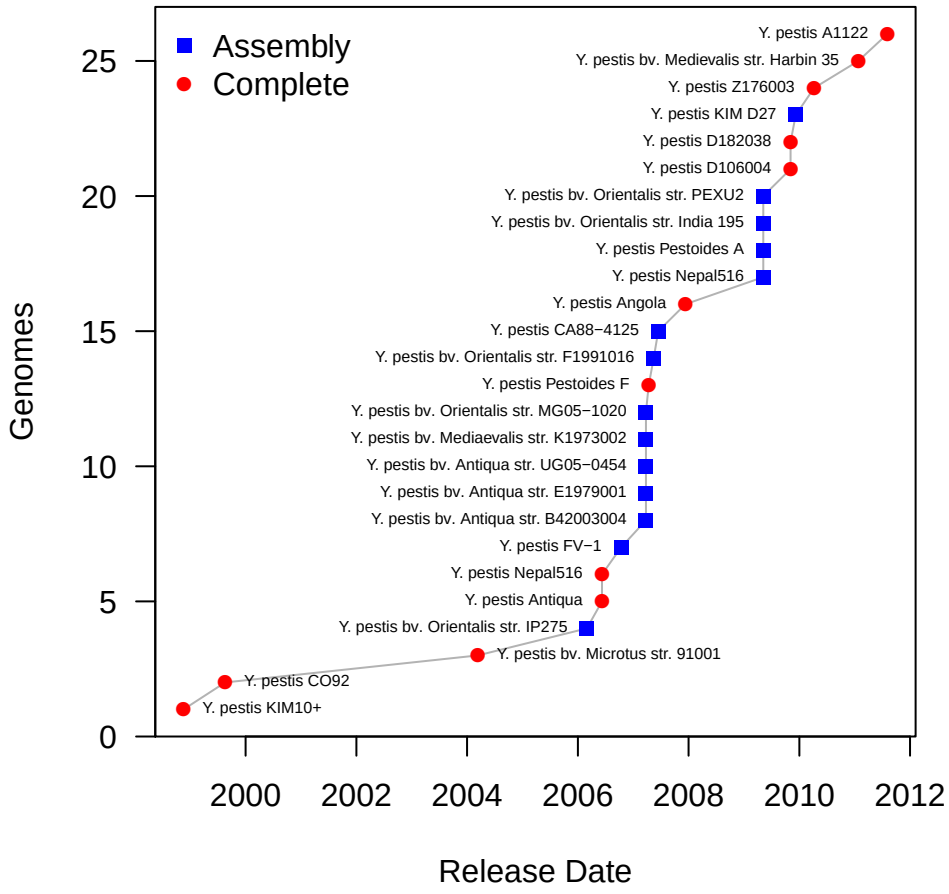

Figure 2: Cumulative plot of *Yersinia pestis* genomes by release date.