Genome project tables in the genomes package

Chris Stubben

April 4, 2012

The genomes package collects genome project metadata from NCBI (http://www.ncbi.nlm.nih.gov) and the ENA (http://www.ebi.ac.uk/ena) 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. At a minimum, the table should have a column listing the project name, status, and release date. 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 11105 rows and 17 columns

             acc                                          name   status
```

```
1     PRJNA55729             Abiotrophia defectiva ATCC 49176 Assembly
2     PRJNA58167               Acaryochloris marina MBIC11017 Complete
3     PRJNA78283                 Acaryochloris sp. CCMEE 5410 Assembly
4     PRJNA51533               Acetivibrio cellulolyticus CD2 Assembly
5     PRJNA80697           Acetobacteraceae bacterium AT-5844 Assembly
...          ...                                          ...      ...
11105 PRJNA68445 Zymomonas mobilis subsp. pomaceae ATCC 29192 Complete
        released ...
1     2009-03-17 ...
2     2007-10-16 ...
3     2011-06-03 ...
4     2010-08-11 ...
5     2011-12-16 ...
...          ... ...
11105 2011-06-17 ...

R> summary(proks)

$`Total genomes`
[1] 11105 genome projects on Apr 03, 2012

$`By status`
              Total
No data        4419
Assembly       3275
Complete       2132
SRA or Traces  1279

$`Recent submissions`
  RELEASED
1 2012-03-27
2 2012-03-27
3 2012-03-27
4 2012-03-27
5 2012-03-26
  NAME
1 Salmonella enterica subsp. enterica serovar Heidelberg str. 41579
2 Streptococcus pneumoniae 459-5
3 Streptococcus pneumoniae SV35
4 Streptococcus pneumoniae SV36
5 Helicobacter pylori NAB47
  STATUS
1 Assembly
```

```
2 Assembly
3 Assembly
4 Assembly
5 Assembly
```

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

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          1100
Staphylococcus aureus      416
Salmonella enterica        392
Streptococcus agalactiae   312
Helicobacter pylori        275
Streptococcus pneumoniae   245
Enterococcus faecium       233
Clostridium difficile      224
Enterococcus faecalis      224
Vibrio cholerae            174
```

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* (Figure 2).

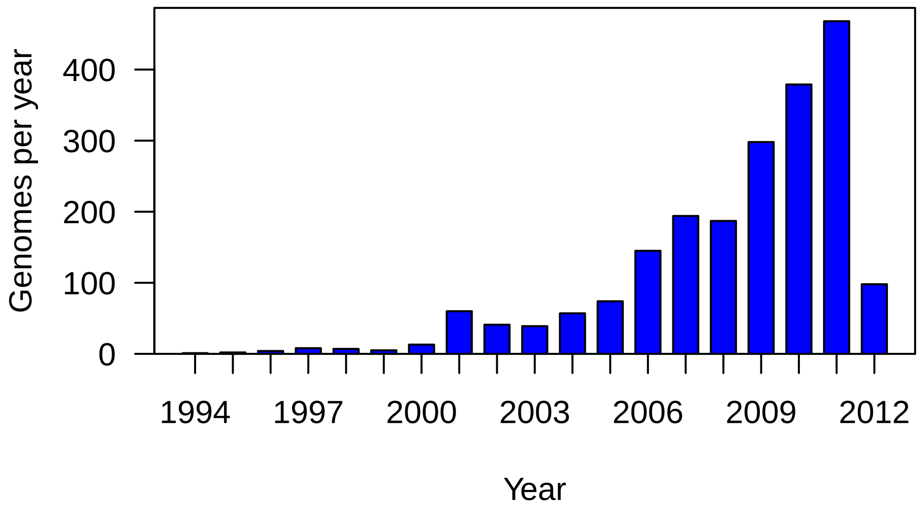

Figure 1: Number of complete microbial genomes released each year at NCBI

```
R> yp <- subset(proks, name %like% "Yersinia pestis*")
R> plotby(yp, labels = TRUE, cex = 0.5, lbty = "n")
```

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.

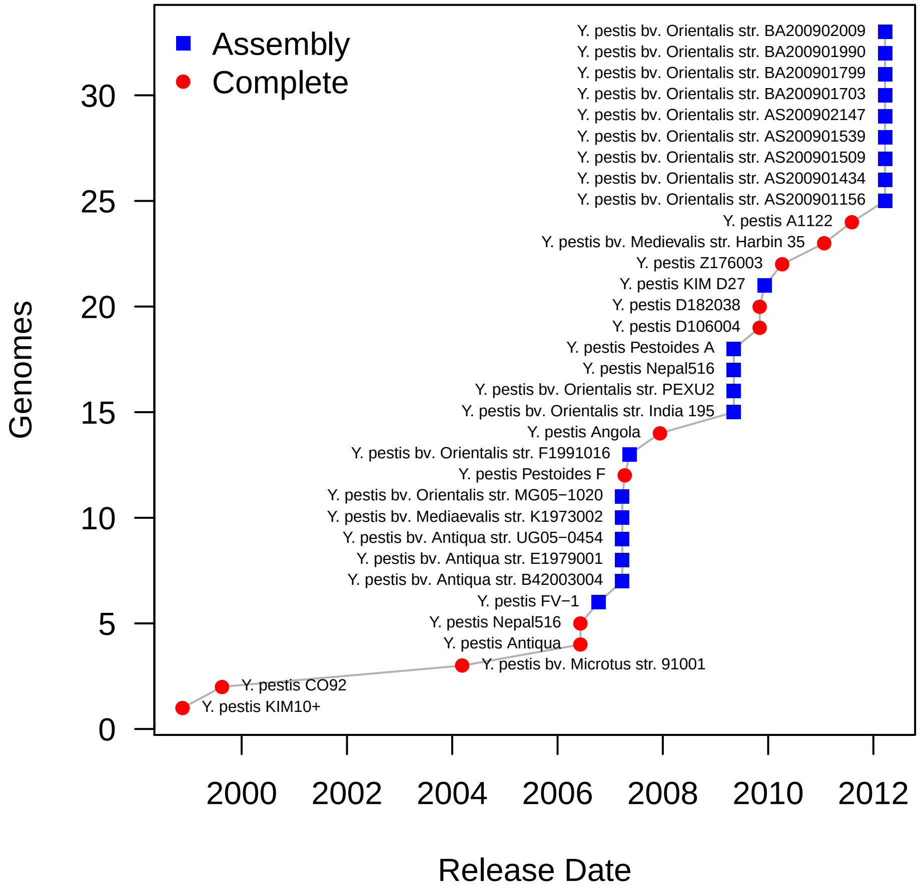

Figure 2: Cumulative plot of *Yersinia pestis* genomes by release date.