

Genome project tables in the genomes package

Chris Stubben

October 2, 2012

The **genomes** package collects genome project metadata from NCBI using E-utility scripts (equery, esummary, efetch and elink) or from the ENA using the ENA Browser REST URL. The package also includes genome tables from NCBI 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** method 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 14202 rows and 18 columns

pid	name	status
-----	------	--------

```

1      55729      Abiotrophia defectiva ATCC 49176 Assembly
2      174970     Acaricomes phytoseiuli DSM 14247  No data
3      58167      Acaryochloris marina MBIC11017 Complete
4      78283      Acaryochloris sp. CCME 5410 Assembly
5      51533      Acetivibrio cellulolyticus CD2 Assembly
...      ...
14202  68445 Zymomonas mobilis subsp. pomaceae ATCC 29192 Complete
        released ...
1      2009-03-17 ...
2      <NA> ...
3      2007-10-16 ...
4      2011-06-03 ...
5      2010-08-11 ...
...      ...
14202  2011-06-17 ...

```

```
R> summary(proks)
```

```
$`Total genomes`
```

```
[1] 14202 genome projects on Sep 17, 2012
```

```
$`By status`
```

	Total
No data	5765
Assembly	4918
Complete	2179
SRA or Traces	1340

```
$`Recent submissions`
```

	released	name	status
1	2012-09-14	Brachyspira pilosicoli B2904	Complete
2	2012-09-14	Brevibacillus brevis X23	Assembly
3	2012-09-14	Enterococcus sp. GMD1E	Assembly
4	2012-09-14	Enterococcus sp. GMD2E	Assembly
5	2012-09-14	Enterococcus sp. GMD3E	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	1338
Salmonella enterica	750
Staphylococcus aureus	492
Streptococcus agalactiae	315
Helicobacter pylori	286
Enterococcus faecalis	272
Streptococcus pneumoniae	246
Enterococcus faecium	238
Clostridium difficile	228
Acinetobacter baumannii	192

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).

```
R> ## Yersinia pestis
R> yp<-subset(proks, name %like% 'Yersinia pestis*')
R> plotby(yp, labels=TRUE, cex=.5, lty='n')
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.

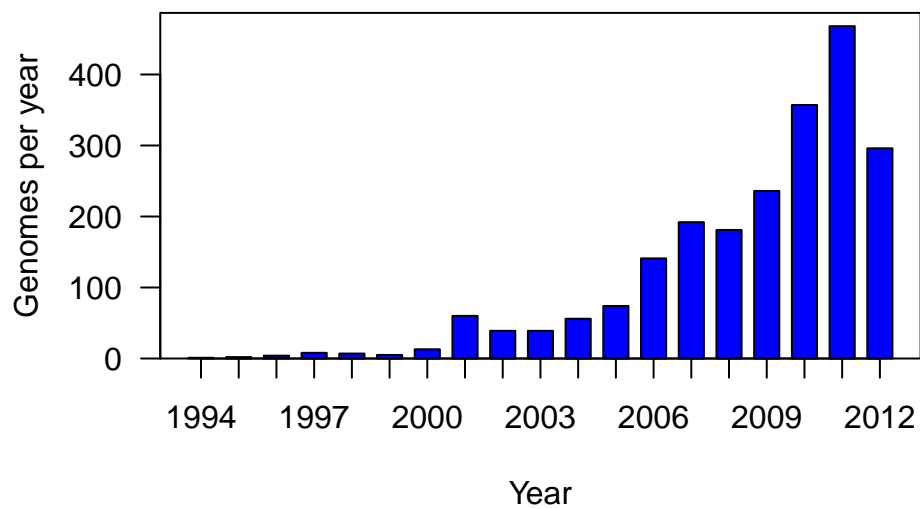


Figure 1: Number of complete microbial genomes released each year at NCBI

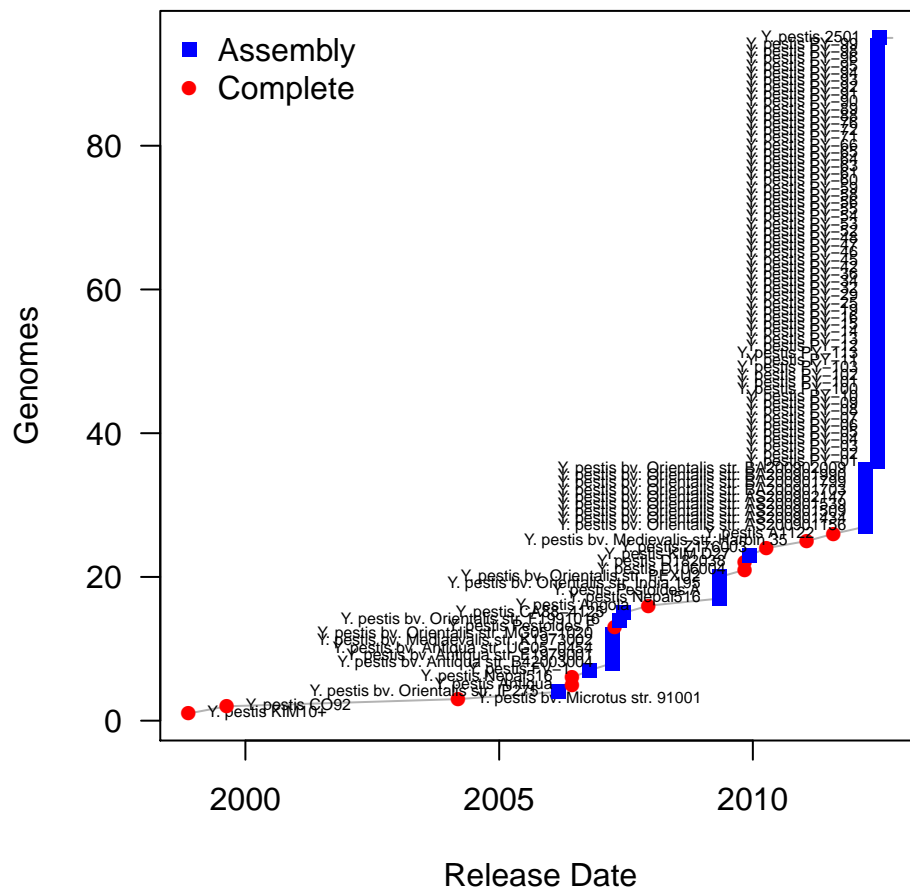


Figure 2: Cumulative plot of *Yersinia pestis* genomes by release date.