

Genome project tables in the genomes package

Chris Stubben

October 13, 2014

The `genomes` package collects genome project metadata from NCBI using E-utility scripts (`esearch`, `esummary`, `efetch` and `elink`) or the NCBI genomes FTP. The package also includes 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`.

Genome tables from the Genomes FTP 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 27570 rows and 25 columns
```

```
      pid          name          status
1    33011 Abiotrophia defectiva ATCC 49176 Scaffold
2   174970 Acaricomes phytoseiuli DSM 14247  Contig
3    12997   Acaryochloris marina MBIC11017 Gapless Chromosome
4    16707   Acaryochloris sp. CCMEE 5410  Contig
5    45843   Acetivibrio cellulolyticus CD2 Scaffold
...      ...          ...          ...
27570 182445 Zymophilus raffinovorans DSM 20765 Scaffold
      released ...
1    2009-03-17 ...
2    2013-04-20 ...
3    2007-10-16 ...
4    2011-06-03 ...
5    2010-08-11 ...
```

```
...
27570 2013-04-23 ...
```

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

```
$`Total genomes`
[1] 27570 genome projects on Sep 04, 2014
```

```
$`By status`
```

	Total
Contig	13074
Scaffold	10718
Gapless Chromosome	3053
Chromosome	373
Chromosome with gaps	343
Complete	9

```
$`Recent submissions`
```

released	name	status
1 2014-09-02	Altuibacter lentus	Scaffold
2 2014-09-02	Bacillus cereus ATCC 4342	Scaffold
3 2014-09-02	Bacillus licheniformis	Scaffold
4 2014-09-02	Bacillus megaterium	Scaffold
5 2014-09-02	Paenibacillus macerans	Scaffold

```
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 `month` and `year` functions can be used to extract the month or year from the release date. The `table2` function formats and sorts a contingency table by counts.

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

	Total
Staphylococcus aureus	4178
Escherichia coli	2292
Mycobacterium tuberculosis	1765
Salmonella enterica	907
Acinetobacter baumannii	816
Helicobacter pylori	432
Klebsiella pneumoniae	386
Enterococcus faecalis	352
Streptococcus agalactiae	308
Streptococcus pneumoniae	297

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 1).

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

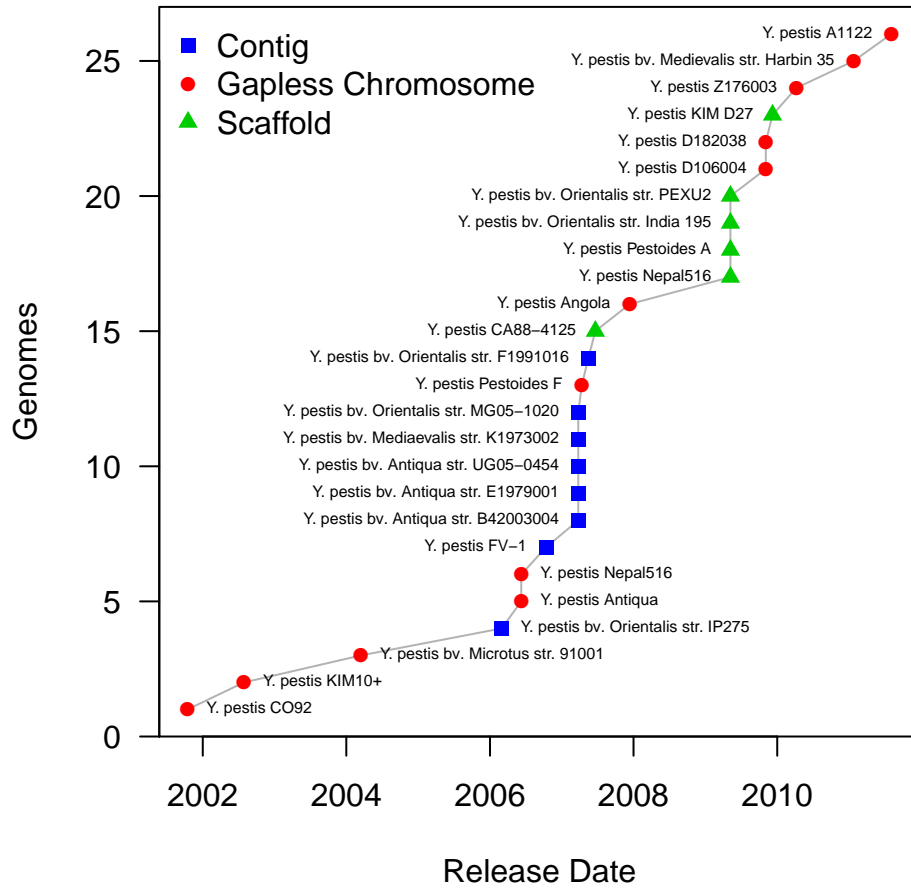


Figure 1: Cumulative plot of *Yersinia pestis* genomes released before 2012.