Genome project tables in the genomes package

Chris Stubben

April 22, 2010

The `genomes` package collects genome project metadata and provides tools to track, sort, group, summarize and plot the data. The genome project tables from the National Center for Biotechnology Information (NCBI) and the Genomes On Line Database (GOLD) are the primary sources of data and include a rapidly growing collection of organisms from all domains of life (viruses, archaea, bacteria, protists, fungi, plants, and animals) plus metagenomic sequences.

Genome tables are a defined class (*genomes*) in the package 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`.

NCBI tables

Genome tables at NCBI are downloaded from the Genome Project database. The primary tables include a list of prokaryotic projects (`lproks`), eukaryotic projects (`leuks`), and metagenomic projects (`lenvs`). 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 in Figure 1 (use `lines` to add additional tables). The `update` method is not illustrated below, but can be used to download the latest version of the table from NCBI.

```
R> data(lproks)
R> lproks
```

```
  A genomes data.frame with 3861 rows and 31 columns

      pid                                          name      status
1   30807                       'Nostoc azollae' 0708    Assembly
2   33011            Abiotrophia defectiva ATCC 49176    Assembly
```

```
3    12997               Acaryochloris marina MBIC11017    Complete
4    16707                 Acaryochloris sp. CCMEE 5410 In Progress
5    45843               Acetivibrio cellulolyticus CD2 In Progress
...    ...                                                ...         ...
3861 34927 Zymomonas mobilis subsp. pomaceae ATCC 29192 In Progress
       released ...
1    2009-03-06 ...
2    2009-03-17 ...
3    2007-10-16 ...
4          <NA> ...
5          <NA> ...
...         ... ...
3861       <NA> ...
```

```
R> summary(lproks)
```

```
$`Total genomes`
[1] 3861 genome projects on Apr 12, 2010

$`By status`
            Total
In Progress  1585
Assembly     1159
Complete     1117

$`Recent submissions`
  RELEASED   NAME                                    STATUS
1 2010-04-09 Prevotella ruminicola 23                Complete
2 2010-04-09 Rhodobacter capsulatus SB1003           Complete
3 2010-04-08 Bacillus megaterium QM B1551            Complete
4 2010-04-08 Yersinia pestis Z176003                 Complete
5 2010-04-05 Aminobacterium colombiense DSM 12261 Complete
```

```
R> plot(lproks, log = "y", las = 1)
R> data(leuks)
R> data(lenvs)
R> lines(leuks, col = "red")
R> lines(lenvs, col = "green3")
R> legend("topleft", c("Microbes", "Eukaryotes", "Metagenomes"),
     lty = 1, bty = "n", col = c("blue", "red", "green3"))
```

For microbial genome projects, the number of complete genomes doubles every 22 months and a new microbial genome is released about every other day. At least in 2008, fewer complete genomes were released than the previous year (Figure 2).

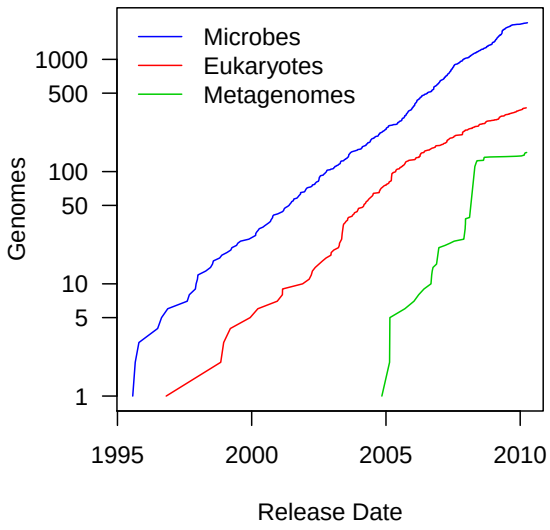

Figure 1: Cumulative plot of genome projects by release date at NCBI.

```
R> complete <- subset(lproks, status == "Complete")
R> doublingTime(complete)

days
 670

R> x <- table(format(complete$released, "%Y"))
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()
```

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

```
R> table2(species(lproks$name))

                          Total
Escherichia coli            263
Salmonella enterica         127
Staphylococcus aureus        76
Mycobacterium tuberculosis   70
Enterococcus faecalis        57
Bacillus cereus              53
Vibrio cholerae              48
```

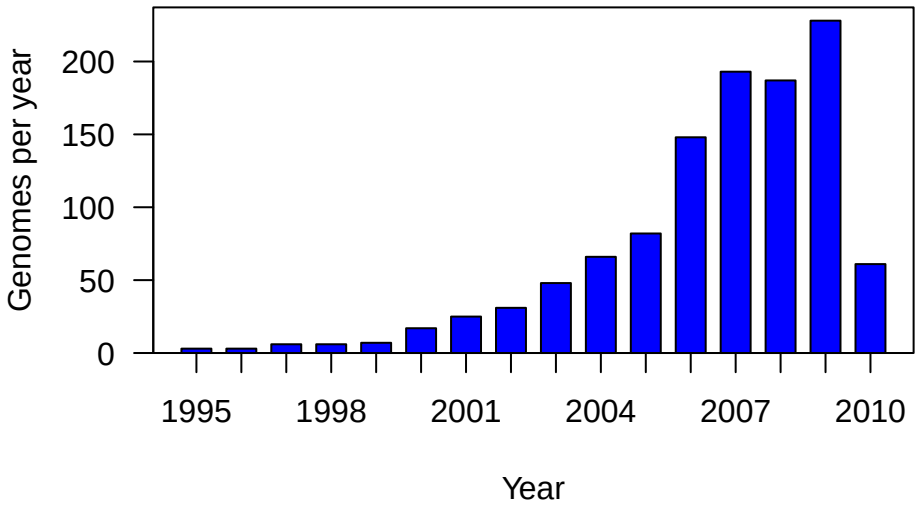

Figure 2: Number of complete microbial genomes released each year at NCBI

```
Streptococcus pneumoniae      47
Helicobacter pylori           39
Pseudomonas syringae          36
```

Because subsets of tables are often needed, the binary operator `like` allows pattern matching using wildcards. The `plotby` function below expands on the default plot method and adds the ability to plot by groups (default is status) using either labeled points or multiple lines like Figure 1. For example, the release dates of complete and draft sequences of *Yersinia pestis* are displayed in Figure 3.

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

GOLD and other tables

The Genomes Online Database (GOLD) is a comprehensive resource that collects detailed project metadata from over 7,000 genomes. There are currenlty over 100 columns in this large table with specific fields relating to the organism, host, environment, and sequencing methods. Just two of the hundreds of possible queries are illustated below. In first example, a list of endosymbiotic intracellular organisms is divided into pathogens and commensal bacteria. In the second example, the comma-separated list of phenotypes is split and a new table is created listing the GOLD identifier, name, and a single phenotype. Then genomes matching "Arsenic metabolizer" are displayed.

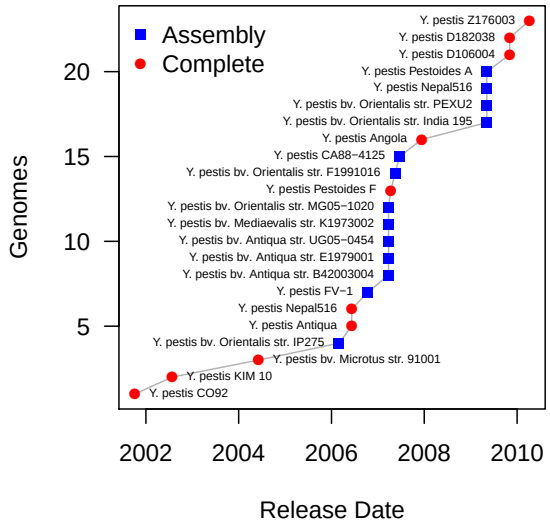

Figure 3: Cumulative plot of *Yersinia pestis* genomes by release date.

```
R> data(gold)
R> obligate <- subset(gold, symbiotic.interaction == "Endosymbiotic intracellular",
     c(goldstamp, name, phenotype))
R> obligate$pathogen <- "Pathogen"
R> obligate$pathogen[obligate$phenotype %like% "Non-*|Symb*|Carb"] <- "Commensal"
R> obligate$pathogen[obligate$phenotype == ""] <- "Commensal"
R> table2(genus(obligate$name), obligate$pathogen)
```

	Commensal	Pathogen	Total
Chlamydia	0	45	45
Rickettsia	2	18	20
Rhizobium	18	0	18
Wolbachia	17	0	17
Chlamydophila	0	12	12
Buchnera	11	0	11
Coxiella	0	7	7
Ehrlichia	0	7	7
Anaplasma	0	6	6
Mesorhizobium	5	0	5

```
R> x <- subset(gold, phenotype != "")
R> x2 <- strsplit(x$phenotype, ", ")
R> gold2 <- as.data.frame(cbind(goldstamp = rep(x$goldstamp,
     sapply(x2, length)), name = rep(x$name, sapply(x2, length)),
     phenotype = unlist(x2)))
R> table2(gold2$phenotype)
```

```
                      Total
Pathogen               1894
Non-Pathogen            249
Intracellular pathogen  114
Acidophile               71
Parasite                 58
Probiotic                50
Meticillin resistant     44
Radiation resistant      37
Catalase positive        34
Symbiont                 32
```

```
R> subset(gold2, phenotype %like% "Arsenic metabol*")
```

```
     goldstamp                                name           phenotype
107    Gc00422    Alkalilimnicola ehrlichei MLHE-1 Arsenic metabolizer
110    Gc00666      Alkaliphilus oremlandii OhILAs Arsenic metabolizer
280    Gi00970    Bacillus selenitireducens MLMS-1 Arsenic metabolizer
281    Gi00921    Bacillus selenitireducens MLS-10 Arsenic metabolizer
1732   Gc00526 Herminiimonas arsenicoxydans ULPAs1 Arsenic metabolizer
2929   Gi00788                      Thiomonas sp. Arsenic metabolizer
```

Finally, genome data from the Human Microbiome Project is stored in the `hmp` dataset and includes additional information such as the primary body site occupied by a sequenced organism.