

# Genome project tables in the genomes package

Chris Stubben

October 18, 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`.

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")
```

You can also install the package on earlier versions of R using `install.packages`, but this has not been tested completely.

```
R> install.packages("genomes",
  repos="http://www.bioconductor.org/packages/release/bioc")
```

Finally, 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. On some systems (Mac 10.4), you may need to add the `type='source'` option to install the package source. In addition, the `genomes` package depends on some functions in the `lattice` and `XML` packages, so these two should be installed first (`lattice` is usually installed by default, but not `XML`).

```
R> install.packages("genomes",
  repos="http://www.bioconductor.org/packages/devel/bioc")
```

## 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 4951 rows and 32 columns
```

```
      pid                name      status
1  30807      'Nostoc azollae' 0708 Complete
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 Assembly
...     ...                ...      ...
4951 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  2010-08-11 ...
...           ... ...
4951           <NA> ...
```

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

```
$`Total genomes`
```

```
[1] 4951 genome projects on Oct 06, 2010
```

```
$`By status`
```

```
      Total
In Progress 2198
Assembly    1509
Complete    1244
```

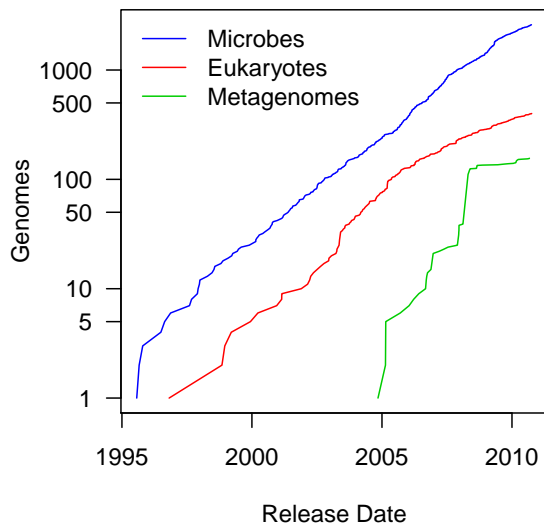


Figure 1: Cumulative plot of genome projects by release date at NCBI.

```
$`Recent submissions`
```

	RELEASED	NAME	STATUS
1	2010-10-01	Helicobacter pylori Cuz20	Complete
2	2010-10-01	Helicobacter pylori PeCan4	Complete
3	2010-10-01	Helicobacter pylori Sat464	Complete
4	2010-10-01	Helicobacter pylori SJM180	Complete
5	2010-10-01	Lactobacillus plantarum subsp. plantarum ST-III	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).

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

```
days
683
```

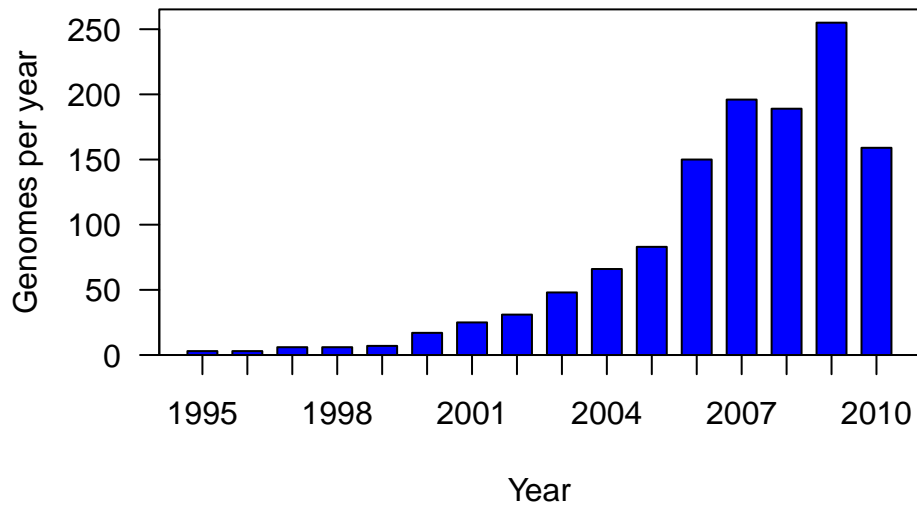


Figure 2: Number of complete microbial genomes released each year at NCBI

```
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	392
Streptococcus pneumoniae	212
Staphylococcus aureus	91
Mycobacterium tuberculosis	81
Salmonella enterica	79
Acinetobacter baumannii	74
Propionibacterium acnes	74
Enterococcus faecalis	72
Streptococcus mutans	60
Bacillus cereus	53

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

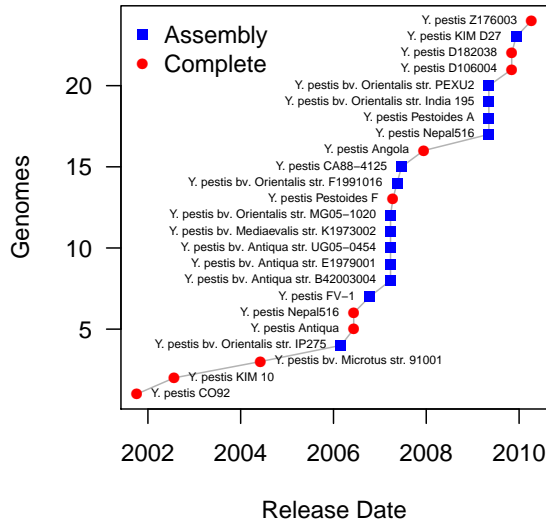


Figure 3: Cumulative plot of *Yersinia pestis* genomes by release date.

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, lpty = "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 currently 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 illustrated 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.

```
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	50	50
Rickettsia	2	19	21
Rhizobium	18	0	18
Wolbachia	18	0	18
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	2391
Non-Pathogen	414
Intracellular pathogen	115
Parasite	92
Acidophile	72
Probiotic	60
Meticillin resistant	48
Radiation resistant	39
Catalase positive	34
Symbiont	33

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

	goldstamp	name	phenotype
153	Gc00422	Alkalilimnicola ehrlichei MLHE-1	Arsenic metabolizer
156	Gc00666	Alkaliphilus orelandii OhILAs	Arsenic metabolizer
336	Gi00970	Bacillus selenitireducens MLMS-1	Arsenic metabolizer
338	Gc01337	Bacillus selenitireducens MLS-10	Arsenic metabolizer
2008	Gc00526	Herminiimonas arsenicoxydans ULPAs1	Arsenic metabolizer
3711	Gi00788	Thiomonas sp. 3As	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.