Bioinformatics, 33(8), 2017, 1216–1217
doi: 10.1093/bioinformatics/btw821
Advance Access Publication Date: 20 January 2017
Applications Note

OXFORD

Genome analysis

Biomartr: genomic data retrieval with R

Hajk-Georg Drost* and Jerzy Paszkowski

The Sainsbury Laboratory, University of Cambridge, Cambridge, United Kingdom

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 20, 2016; revised on December 16, 2016; editorial decision on December 20, 2016; accepted on December 22, 2016

Abstract

Motivation: Retrieval and reproducible functional annotation of genomic data are crucial in biology. However, the current poor usability and transparency of retrieval methods hinders reproducibility. Here we present an open source R package, *biomartr*, which provides a comprehensive easy-to-use framework for automating data retrieval and functional annotation for meta-genomic approaches. The functions of *biomartr* achieve a high degree of clarity, transparency and reproducibility of analyses.

Results: The *biomartr* package implements straightforward functions for bulk retrieval of all genomic data or data for selected genomes, proteomes, coding sequences and annotation files present in databases hosted by the National Center for Biotechnology Information (NCBI) and European Bioinformatics Institute (EMBL-EBI). In addition, *biomartr* communicates with the BioMart database for functional annotation of retrieved sequences. Comprehensive documentation of *biomartr* functions and five tutorial vignettes provide step-by-step instructions on how to use the package in a reproducible manner.

Availability and Implementation: The open source *biomartr* package is available at https://github.com/HajkD/biomartr and https://cran.r-project.org/web/packages/biomartr/index.html.

Contact: hgd23@cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Modern genome studies are no longer limited to single genome analyses or pairwise genomic comparisons but increasingly involve meta-genomic approaches. For this purpose, the NCBI and EMBL-EBI organize and maintain specialized sequence databases that fulfil various scientific requirements. Among the most important and best-curated databases are Genbank, RefSeq and ENSEMBL. Genbank is an annotated collection of all publicly available DNA sequences (Benson *et al.*, 2013). The RefSeq collection offers a comprehensive, integrated, non-redundant and well-annotated set of sequences, including genomic DNA, transcripts and proteins (Pruitt *et al.*, 2007). The ENSEMBL database provides DNA sequence assemblies and curated Ensembl gene builds from various projects (Yates *et al.*, 2016).

Current meta-genomic pipelines consist of custom-prepared scripts that automatically retrieve selected genomes from these resources. Post-processing, handling and analysis usually uses the Perl, Python, or R programming languages. Although powerful sequence retrieval frameworks have been implemented in Perl (BioPerl), Python (BioPython) and R, their use requires appropriate programming expertise. Furthermore, none of these frameworks combines meta-genomic scale sequence retrieval with functional annotation. These deficiencies also apply to the currently available R packages *seqinr* and *biomaRt*. The *seqinr* package aims to automate sequence retrieval in R but is not designed for meta-genomic approaches and does not include functional annotation. The *biomaRt* package aims to provide functional annotation methods but these are also not designed for meta-genomic approaches and are not easy to use for non-programming experts. To provide a fast, transparent and easy-to-use framework for combined genomic data retrieval and efficient functional annotation of genetic features in meta-genomic approaches, we have designed the R package *biomartr* for use with the NCBI, ENSEMBL, ESNEMBLGENOMES and BioMart infrastructures (Smedley *et al.*, 2009). The major advantage of *biomartr*

is that it does not require profound programming expertise. It is optimized to handle multiple genomes simultaneously and allows, for example, assignment of Gene Ontology (GO) information and sequence homology relationships between different organisms by communicating with the BioMart database. The interface functions communicating with the BioMart database use a novel organism centered notation for information retrieval. Instead of learning the underlying database and dataset linking convention of BioMart, users can type the scientific name of an organism of interest (e.g. 'Homo sapiens') to retrieve a list of all available information provided by BioMart for this particular organism of interest. In summary, the *biomartr* package provides researchers with a powerful tool for efficient, straightforward and reproducible handling of large-scale meta-genomic data and intuitive organism centered interface functions to retrieve functional annotation information from the BioMart database.

2 Implementation

The *biomartr* package is released under the GNU General Public License within the CRAN project (R Core Team). The package can be downloaded from https://cran.r-project.org/web/packages/biomartr/index.html. The source code is publically available at https://github.com/HajkD/biomartr. The *biomartr* package depends on the R packages *Biostrings*, *data.table*, *dplyr*, *readr*, *downloader*, *RCurl*, *XML*, *biomaRt* (Durinck *et al.*, 2005), *httr* and *stringr*. The functionality of packages such as *biomaRt* (Durinck *et al.*, 2005) and *seqinr* (Charif and Lobry, 2007) are included in *biomartr* and significantly extended. This is achieved by additional data retrieval functions and the direct combination of *biomaRt* and *seqinr* functionality with improved retrieval capability.

3 Functions and examples

Thirty-seven functions are provided by the *biomartr* package. For genome and database retrieval, the functions *listDatabases()*, *listKingdoms()*, *listGroups()*, *listSubgroups()*, *listGenomes()* and *is.genome.available()* enable the listing of all databases and genomic sequences that are available for automated retrieval.

For example, the entire NCBI nr database can then be downloaded easily using just one command:

```
download.database.all(db = 'nr', path = 'nr')
```

Analogous to the retrieval of databases as described above, selected genomes can also be retrieved using the following function.

As exemplified below by the human genome, download can be triggered by typing:

```
getGenome (db = 'refseq', organism = 'Homo sapiens')
```

The command *getGenome()* also documents the source and version of the downloaded files. Corresponding download of proteomes, coding sequences and annotation files can be obtained by applying

getProteome(), *getCDS()* and *getGFF()*, respectively. The *db* argument can be specified to retrieve genomes from other NCBI or ENSEMBL databases. For meta-genome approaches, *biomartr* includes the *meta.retrieval()* function to download the genomes of entire kingdoms:

```
# Download all vertebrate genomes
meta.retrieval(kingdom = 'vertebrate_mammalian', type = 'genome')
```

Hence, for example, all mammalian genomes can be downloaded with just one command. The *type* argument can also be specified for proteomes, coding sequences and annotation files. For functional annotation, available datasets and BioMart connections for a specific organism of interest can be obtained by typing:

```
organismBM(organism = 'Homo sapiens')
```

For example, available sequence homology relationships to other organisms can be retrieved by running the command:

```
organismAttributes(organism = 'Homo sapiens', topic = 'homolog')
```

Finally, users can retrieve GO information for a particular gene or gene set, e.g. human gene *GUCA2A* by running the command:

```
getGO(organism = 'Homo sapiens',
genes = 'GUCA2A',
filters = 'hgnc_symbol')
```

Tutorials are available at https://github.com/HajkD/biomartr#tutorials and also in the Supplementary Tutorial.

4 Conclusions

The functions provided by *biomartr* enable fast data retrieval and functional annotation queries to prominent sequence and annotation databases such as NCBI, ENSEMBL, ENSEMBLGENOMES and BioMart. In addition, all data retrieval functions implemented in *biomartr* automatically archive and log the source, date, version, taxid and type of data retrieved. Thus, *biomartr* improves reproducibility and transparency in genomic data handling. It can be integrated easily into meta-genomic analyses.

Funding

This work was supported by an European Research Council grant named EVOBREED [grant number 322621] (to JP) and a Gatsby Fellowship [grant number AT3273/GLE] (to JP).

Conflict of Interest: none declared.

References

Benson,D.A. *et al.* (2013) Genbank. *Nucleic Acids Res.*, **41**, D36–D42.

Charif,D. and Lobry,J.R. (2007) SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution*. Springer Berlin Heidelberg, pp. 207–232.

Durinck,S. *et al.* (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Smedley,D. *et al.* (2009) BioMart – biological queries made easy. *BMC Genomics*, **10**, 22.

Yates *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.