# R 프로그래밍 #10

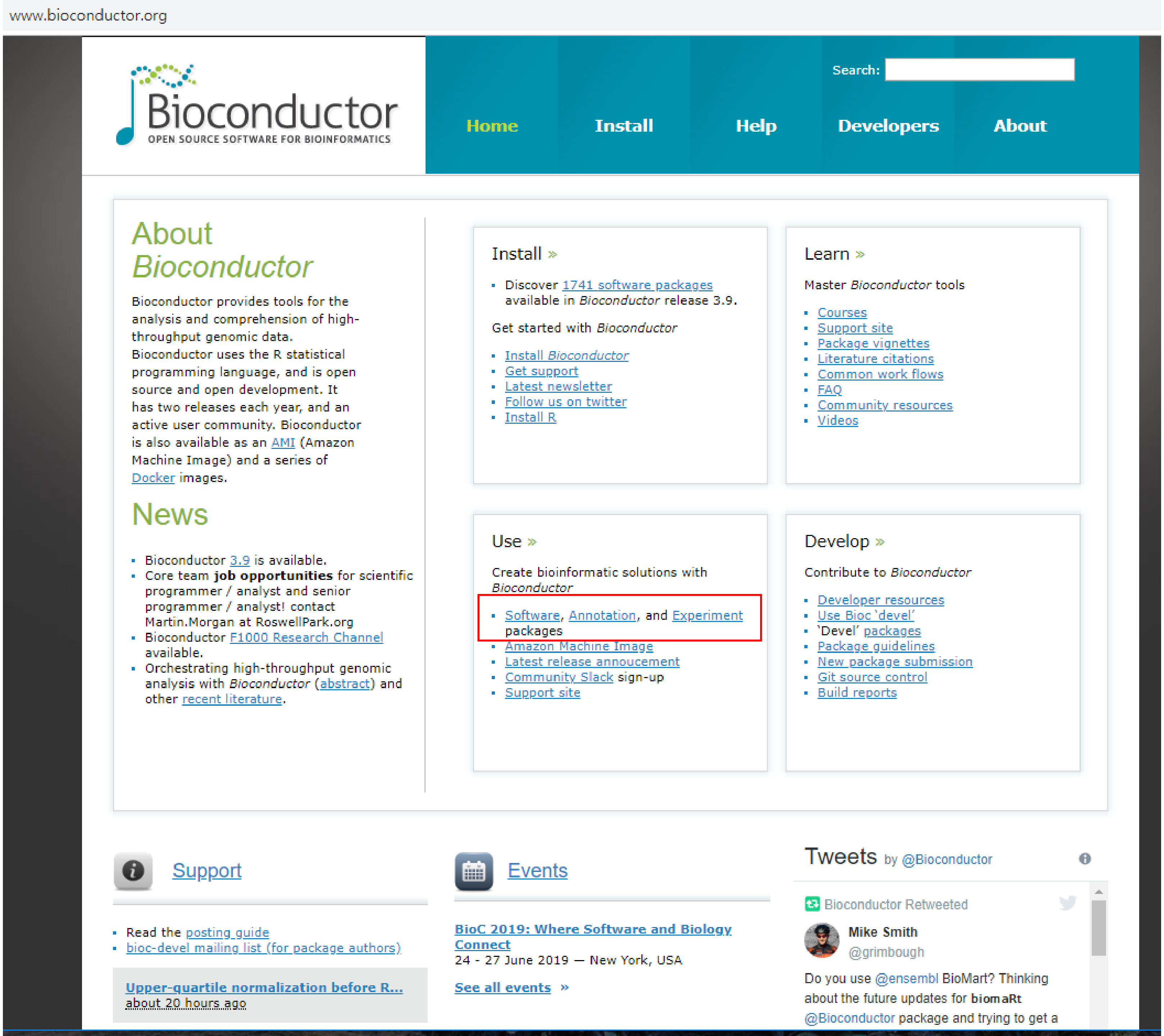
2019. 05. 16

한국생명공학연구원
김하성

Required packages: reutils, seqinr

# Sequence analysis

- A Little Book of R For Bioinformatics, Avril Coghlan (https://media.readthedocs.org/pdf/a-little-book-of-r-for-bioinformatics/latest/a-little-book-of-r-for-bioinformatics.pdf)
- https://web.stanford.edu/class/bios221/labs/biostrings/lab_1_biostrings.html
- https://bioconductor.org/packages/release/bioc/vignettes/Biostrings



## About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

## News

- Bioconductor 3.9 is available.
- Core team **job opportunities** for scientific programmer / analyst and senior programmer / analyst! contact Martin.Morgan at RoswellPark.org
- Bioconductor F1000 Research Channel available.
- Orchestrating high-throughput genomic analysis with *Bioconductor* (abstract) and other recent literature.

### Install »

- Discover 1741 software packages available in *Bioconductor* release 3.9.

Get started with *Bioconductor*

- Install *Bioconductor*
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

### Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

### Use »

Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and Experiment packages
- Amazon Machine Image
- Latest release annoucement
- Community Slack sign-up
- Support site

### Develop »

Contribute to *Bioconductor*

- Developer resources
- Use Bioc 'devel'
- 'Devel' packages
- Package guidelines
- New package submission
- Git source control
- Build reports

### Support

- Read the posting guide
- bioc-devel mailing list (for package authors)

**Upper-quartile normalization before R...**
about 20 hours ago

### Events

**BioC 2019: Where Software and Biology Connect**
24 - 27 June 2019 — New York, USA

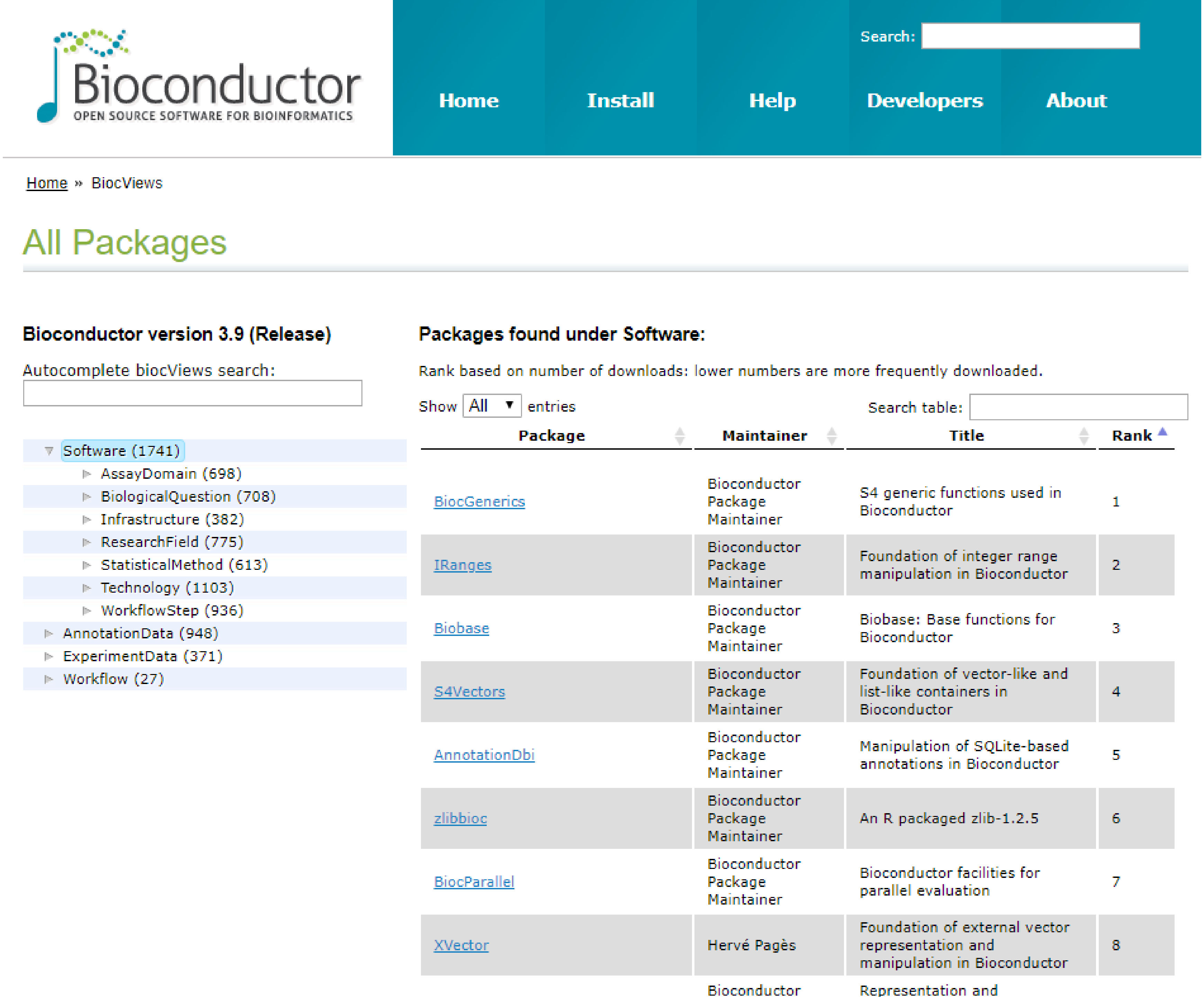
**See all events** »

### Tweets by @Bioconductor

Bioconductor Retweeted

**Mike Smith**
@grimbough

Do you use @ensembl BioMart? Thinking about the future updates for **biomaRt** @Bioconductor package and trying to get a



Home » BiocViews

# All Packages

### Bioconductor version 3.9 (Release)

Autocomplete biocViews search:

- Software (1741)
  - AssayDomain (698)
  - BiologicalQuestion (708)
  - Infrastructure (382)
  - ResearchField (775)
  - StatisticalMethod (613)
  - Technology (1103)
  - WorkflowStep (936)
- AnnotationData (948)
- ExperimentData (371)
- Workflow (27)

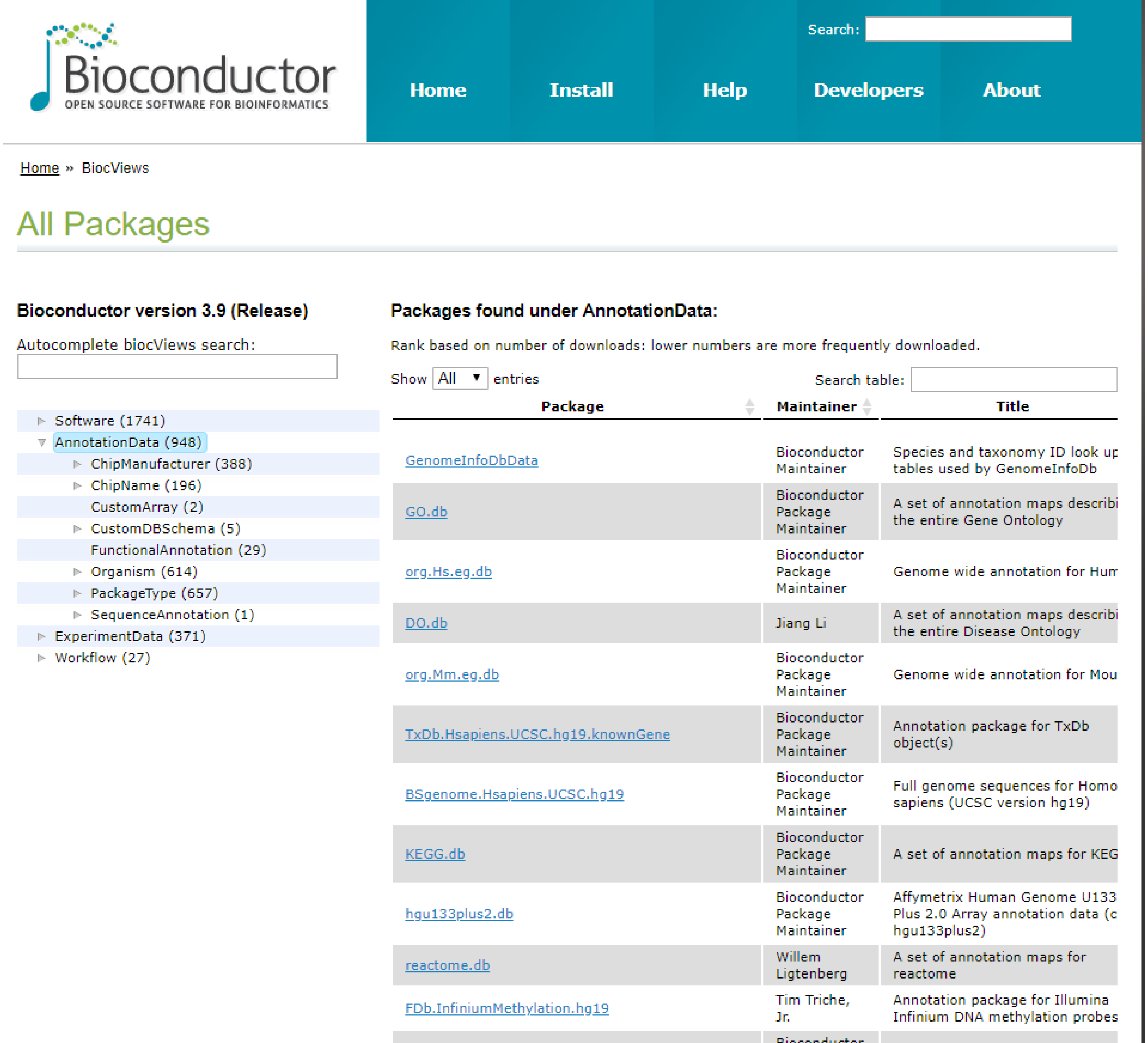
### Packages found under Software:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries

Search table:

| Package | Maintainer | Title | Rank |
|---|---|---|---|
| BiocGenerics | Bioconductor Package Maintainer | S4 generic functions used in Bioconductor | 1 |
| IRanges | Bioconductor Package Maintainer | Foundation of integer range manipulation in Bioconductor | 2 |
| Biobase | Bioconductor Package Maintainer | Biobase: Base functions for Bioconductor | 3 |
| S4Vectors | Bioconductor Package Maintainer | Foundation of vector-like and list-like containers in Bioconductor | 4 |
| AnnotationDbi | Bioconductor Package Maintainer | Manipulation of SQLite-based annotations in Bioconductor | 5 |
| zlibbioc | Bioconductor Package Maintainer | An R packaged zlib-1.2.5 | 6 |
| BiocParallel | Bioconductor Package Maintainer | Bioconductor facilities for parallel evaluation | 7 |
| XVector | Hervé Pagès | Foundation of external vector representation and manipulation in Bioconductor | 8 |
|  | Bioconductor | Representation and |  |



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home Install Help Developers About

Home » BiocViews

# All Packages

## Bioconductor version 3.9 (Release)

Autocomplete biocViews search:

- Software (1741)
- AnnotationData (948)
  - ChipManufacturer (388)
  - ChipName (196)
  - CustomArray (2)
  - CustomDBSchema (5)
  - FunctionalAnnotation (29)
  - Organism (614)
  - PackageType (657)
  - SequenceAnnotation (1)
- ExperimentData (371)
- Workflow (27)

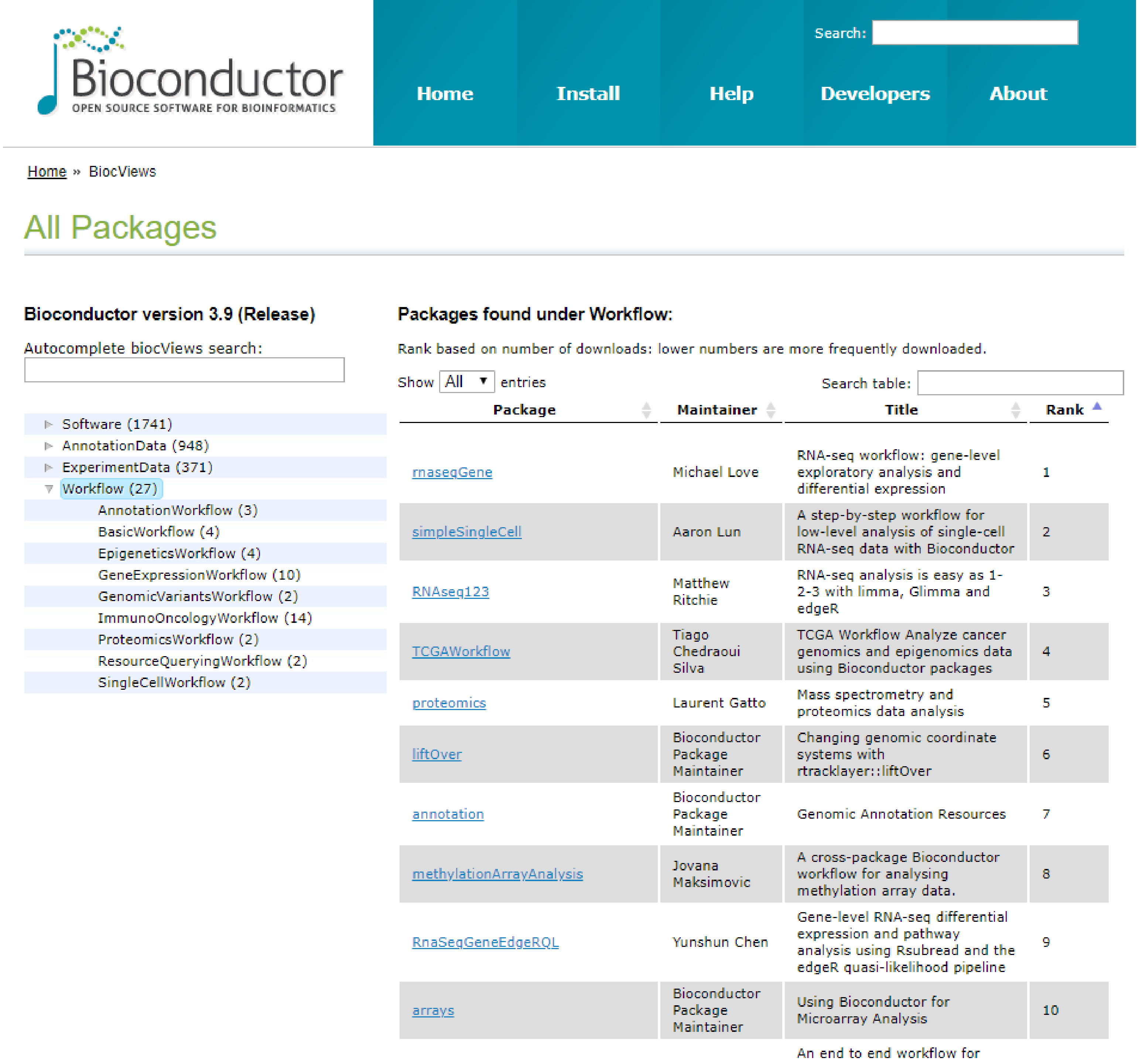
## Packages found under AnnotationData:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries

Search table:

| Package | Maintainer | Title |
|---|---|---|
| GenomeInfoDbData | Bioconductor Maintainer | Species and taxonomy ID look up tables used by GenomeInfoDb |
| GO.db | Bioconductor Package Maintainer | A set of annotation maps describi the entire Gene Ontology |
| org.Hs.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Hum |
| DO.db | Jiang Li | A set of annotation maps describi the entire Disease Ontology |
| org.Mm.eg.db | Bioconductor Package Maintainer | Genome wide annotation for Mou |
| TxDb.Hsapiens.UCSC.hg19.knownGene | Bioconductor Package Maintainer | Annotation package for TxDb object(s) |
| BSgenome.Hsapiens.UCSC.hg19 | Bioconductor Package Maintainer | Full genome sequences for Homo sapiens (UCSC version hg19) |
| KEGG.db | Bioconductor Package Maintainer | A set of annotation maps for KEG |
| hgu133plus2.db | Bioconductor Package Maintainer | Affymetrix Human Genome U133 Plus 2.0 Array annotation data (c hgu133plus2) |
| reactome.db | Willem Ligtenberg | A set of annotation maps for reactome |
| FDb.InfiniumMethylation.hg19 | Tim Triche, Jr. | Annotation package for Illumina Infinium DNA methylation probes |
|  | Bioconductor |  |



Home » BiocViews

# All Packages

### Bioconductor version 3.9 (Release)

Autocomplete biocViews search:

- Software (1741)
- AnnotationData (948)
- ExperimentData (371)
- Workflow (27)
  - AnnotationWorkflow (3)
  - BasicWorkflow (4)
  - EpigeneticsWorkflow (4)
  - GeneExpressionWorkflow (10)
  - GenomicVariantsWorkflow (2)
  - ImmunoOncologyWorkflow (14)
  - ProteomicsWorkflow (2)
  - ResourceQueryingWorkflow (2)
  - SingleCellWorkflow (2)

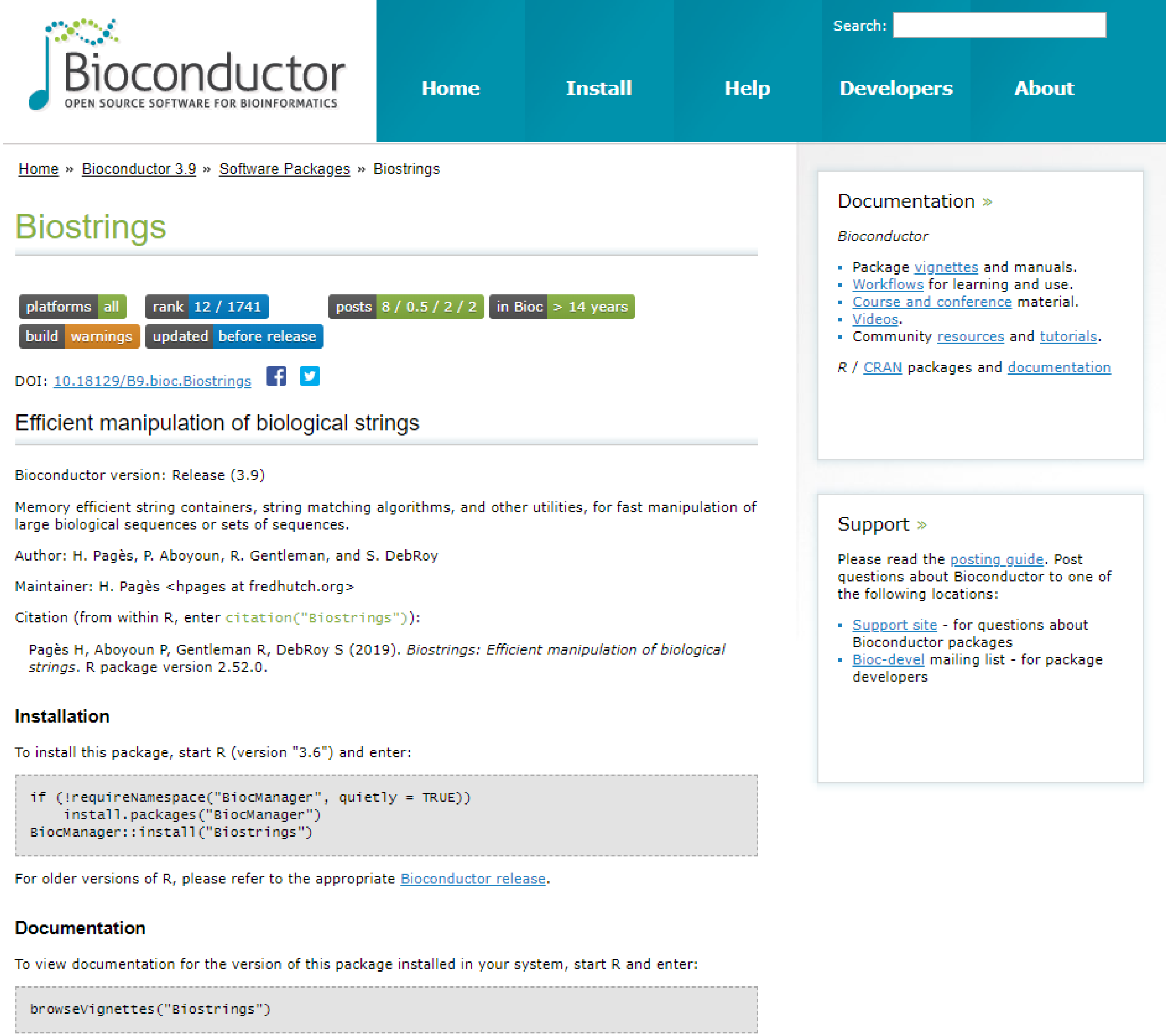
### Packages found under Workflow:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries

Search table:

| Package | Maintainer | Title | Rank |
|---|---|---|---|
| rnaseqGene | Michael Love | RNA-seq workflow: gene-level exploratory analysis and differential expression | 1 |
| simpleSingleCell | Aaron Lun | A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor | 2 |
| RNAseq123 | Matthew Ritchie | RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR | 3 |
| TCGAWorkflow | Tiago Chedraoui Silva | TCGA Workflow Analyze cancer genomics and epigenomics data using Bioconductor packages | 4 |
| proteomics | Laurent Gatto | Mass spectrometry and proteomics data analysis | 5 |
| liftOver | Bioconductor Package Maintainer | Changing genomic coordinate systems with rtracklayer::liftOver | 6 |
| annotation | Bioconductor Package Maintainer | Genomic Annotation Resources | 7 |
| methylationArrayAnalysis | Jovana Maksimovic | A cross-package Bioconductor workflow for analysing methylation array data. | 8 |
| RnaSeqGeneEdgeRQL | Yunshun Chen | Gene-level RNA-seq differential expression and pathway analysis using Rsubread and the edgeR quasi-likelihood pipeline | 9 |
| arrays | Bioconductor Package Maintainer | Using Bioconductor for Microarray Analysis | 10 |
| | | An end to end workflow for | |



Home » Bioconductor 3.9 » Software Packages » Biostrings

# Biostrings

DOI: 10.18129/B9.bioc.Biostrings

## Efficient manipulation of biological strings

Bioconductor version: Release (3.9)

Memory efficient string containers, string matching algorithms, and other utilities, for fast manipulation of large biological sequences or sets of sequences.

Author: H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy

Maintainer: H. Pagès <hpages at fredhutch.org>

Citation (from within R, enter `citation("Biostrings")`):

> Pagès H, Aboyoun P, Gentleman R, DebRoy S (2019). *Biostrings: Efficient manipulation of biological strings*. R package version 2.52.0.

### Installation

To install this package, start R (version "3.6") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("Biostrings")
```

For older versions of R, please refer to the appropriate Bioconductor release.

### Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("Biostrings")
```

## Documentation »

*Bioconductor*

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

*R* / CRAN packages and documentation

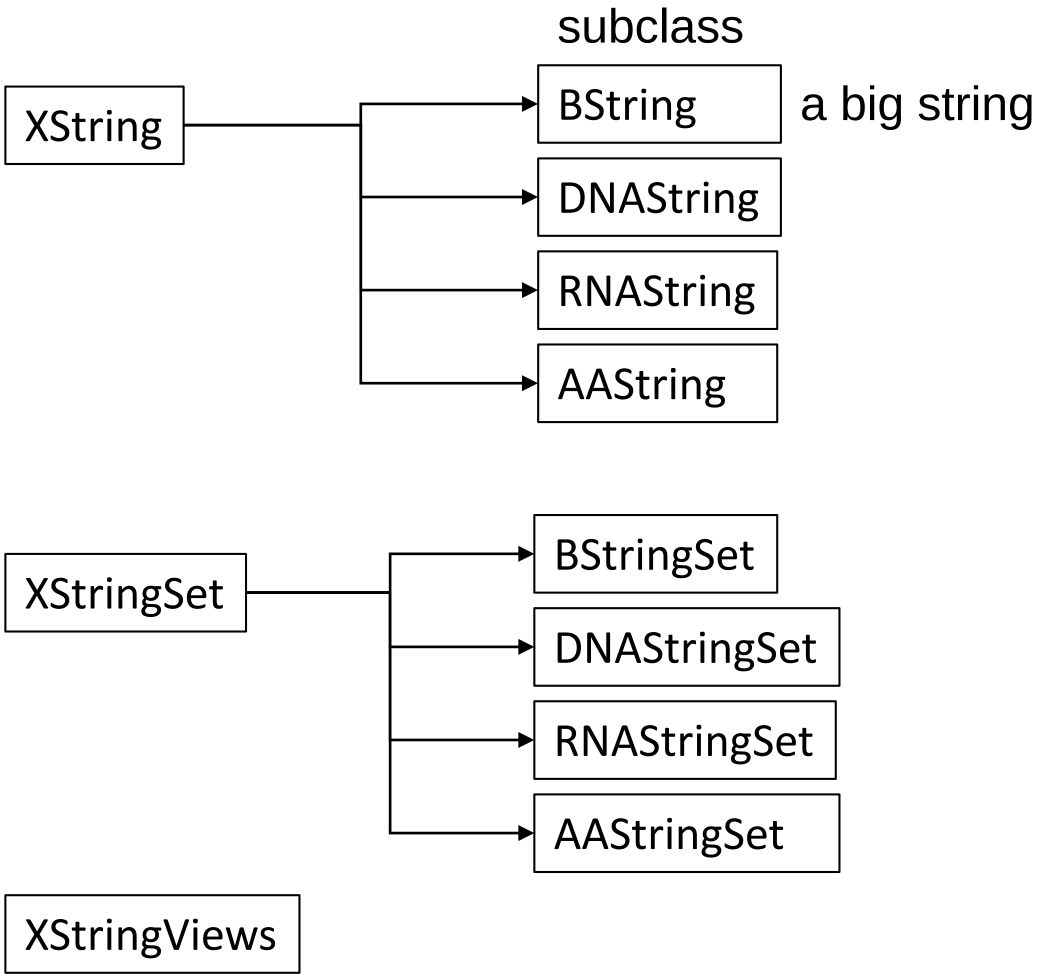
## Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

# The XString Classes

Designed to make manipulation of big strings (DNA, RNA, Proteins)



subclass

XString

BString

a big string

DNAString

RNAString

AAString

XStringSet

BStringSet

DNAStringSet

RNAStringSet

AAStringSet

XStringViews

# Predefined constants

- Some useful predefined constants

```
> DNA_BASES
[1] "A" "C" "G" "T“

> DNA_ALPHABET
 [1] "A" "C" "G" "T" "M" "R" "W" "S" "Y" "K" "V" "H" "D" "B" "N" "-" "+" ".“

> IUPAC_CODE_MAP
     A      C      G      T      M      R      W      S      Y      K      V      H      D      B
   "A"    "C"    "G"    "T"   "AC"   "AG"   "AT"   "CG"   "CT"   "GT"  "ACG"  "ACT"  "AGT"  "CGT"
     N
"ACGT"
```

# The XString Class

- Basic functions  and indexing

```
x0 <- "TTGAAA-CTC-N"
x0
x1 = DNAString(x0)
x1
class(x0)
class(x1)
length(x1)
toString(x1)
complement(x1)
Biostrings::complement(x1)
reverseComplement(x1)
x1[1]
x1[1:3]
subseq(x1, start=3, end=5)
subseq(x1, 3, 5)
alphabetFrequency(x1, baseOnly=TRUE, as.prob=TRUE)
letterFrequency(x1, c("G", "C"), as.prob=TRUE)
```

# Exercise 10-1)

- Generate random DNA sequence of length 30bp and save it in a variable “x0”
  - Use sample, paste functions

```
x0 <- paste(sample(????????), collapse="")
```

- Paste “ATG” in front of the string
- Paste “TAG” at the end of the string
- Convert the string to DNAString class and save it in “x1”
- Get complementary of the sequence
- Translate the DNA sequence

# XStringSet

- Basic functions  and indexing

```
x0 <- c("CTC-NACCAGTAT", "TTGA", "TACCTAGAG")
x1 <- DNAStringSet(x0)
class(x0)
class(x1)
names(x1)
names(x1) <- c("A", "B", "C")
length(x1)
width(x1)
subseq(x1, 2, 4)
x1[[1]]
x1[1]
alphabetFrequency(x1, baseOnly=TRUE, as.prob=TRUE)
letterFrequency(x1, c("G", "C"), as.prob=TRUE)
```

# Exercise 10-2)

- Generate 10 random DNA sequences with length of 30bp and save it in a variable "x0"
    - Attach "ATG" at the start of the sequences
    - Attach "TAG" at the end of the sequences
- Convert the sequences to DNAStringSet class and save it in "x1"
- count "G" and "C" letters
- can you draw a bar graph the GC ratio using ggplot2

# Creating views

- A useful way to view multiple subsequences of a XString

```
x2 <- x1[[1]]
Views(x2, start=1, width=20)
Views(x2, start=1, end=4)
Views(x2, start=c(1,3), end=4)
Views(x2, start=c(1,3,4), width=20)
Views(x2, start=c(1,3,4), width=20)

successiveViews(x2, width=20)
successiveViews(x2, width=rep(20, 2))
successiveViews(x2, width=rep(20, 3))

v <- Views(x2, start=c(1,10), end=c(3,15))
v
gaps(v)
```

# Exercise 10-3)

- Generate random DNA sequence of length 1000bp and save it as a DNAString class in a variable “x0”
- View x0 with 40bp width
- How could you generalize the view code to apply different size of sequences?

# The NCBI sequence database

The National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov)

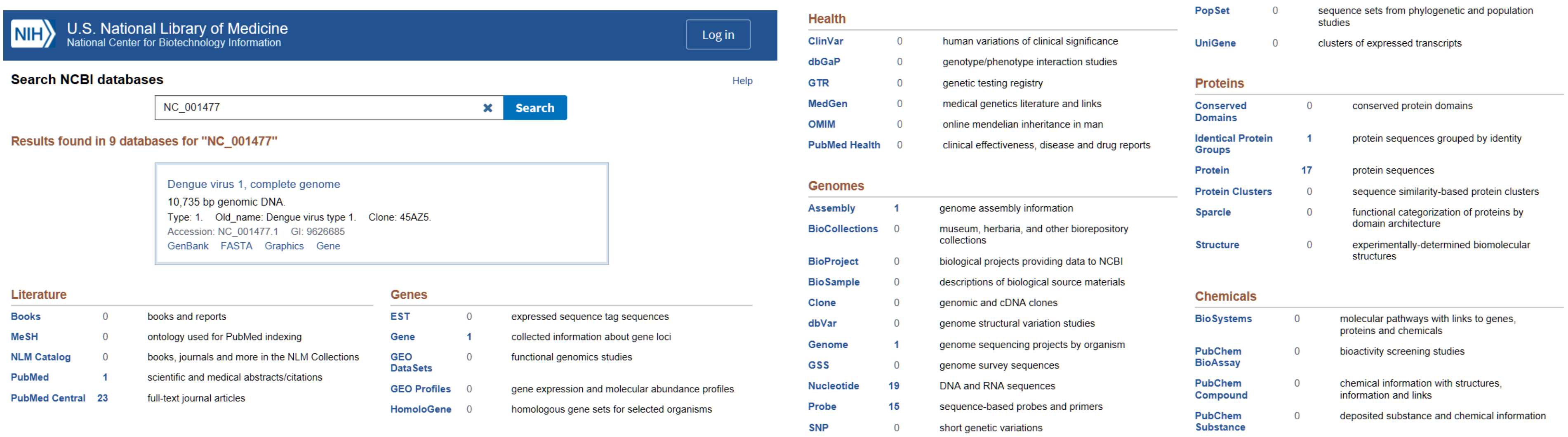
The European Molecular Biology Laboratory (EMBL) Sequence Database (www.ebi.ac.uk/embl)

In Japan, the DNA Data Bank of Japan (DDBJ; www.ddbj.nig.ac.jp).

# Dengue fever

Dengue virus: DEN-1, DEN-2, DEN-3, and DEN-4

Accession no.: NC_001477, NC_001474, NC_001475 and NC_002640



NIH
U.S. National Library of Medicine
National Center for Biotechnology Information
Log in

Search NCBI databases

Results found in 9 databases for "NC_001477"

Dengue virus 1, complete genome
10,735 bp genomic DNA.
Type: 1. Old_name: Dengue virus type 1. Clone: 45AZ5.
Accession: NC_001477.1 GI: 9626685
GenBank FASTA Graphics Gene

**Literature**

| | | |
|---|---|---|
| Books | 0 | books and reports |
| MeSH | 0 | ontology used for PubMed indexing |
| NLM Catalog | 0 | books, journals and more in the NLM Collections |
| PubMed | 1 | scientific and medical abstracts/citations |
| PubMed Central | 23 | full-text journal articles |

**Genes**

| | | |
|---|---|---|
| EST | 0 | expressed sequence tag sequences |
| Gene | 1 | collected information about gene loci |
| GEO DataSets | 0 | functional genomics studies |
| GEO Profiles | 0 | gene expression and molecular abundance profiles |
| HomoloGene | 0 | homologous gene sets for selected organisms |

**Health**

| | | |
|---|---|---|
| ClinVar | 0 | human variations of clinical significance |
| dbGaP | 0 | genotype/phenotype interaction studies |
| GTR | 0 | genetic testing registry |
| MedGen | 0 | medical genetics literature and links |
| OMIM | 0 | online mendelian inheritance in man |
| PubMed Health | 0 | clinical effectiveness, disease and drug reports |

**Genomes**

| | | |
|---|---|---|
| Assembly | 1 | genome assembly information |
| BioCollections | 0 | museum, herbaria, and other biorepository collections |
| BioProject | 0 | biological projects providing data to NCBI |
| BioSample | 0 | descriptions of biological source materials |
| Clone | 0 | genomic and cDNA clones |
| dbVar | 0 | genome structural variation studies |
| Genome | 1 | genome sequencing projects by organism |
| GSS | 0 | genome survey sequences |
| Nucleotide | 19 | DNA and RNA sequences |
| Probe | 15 | sequence-based probes and primers |
| SNP | 0 | short genetic variations |
| PopSet | 0 | sequence sets from phylogenetic and population studies |
| UniGene | 0 | clusters of expressed transcripts |

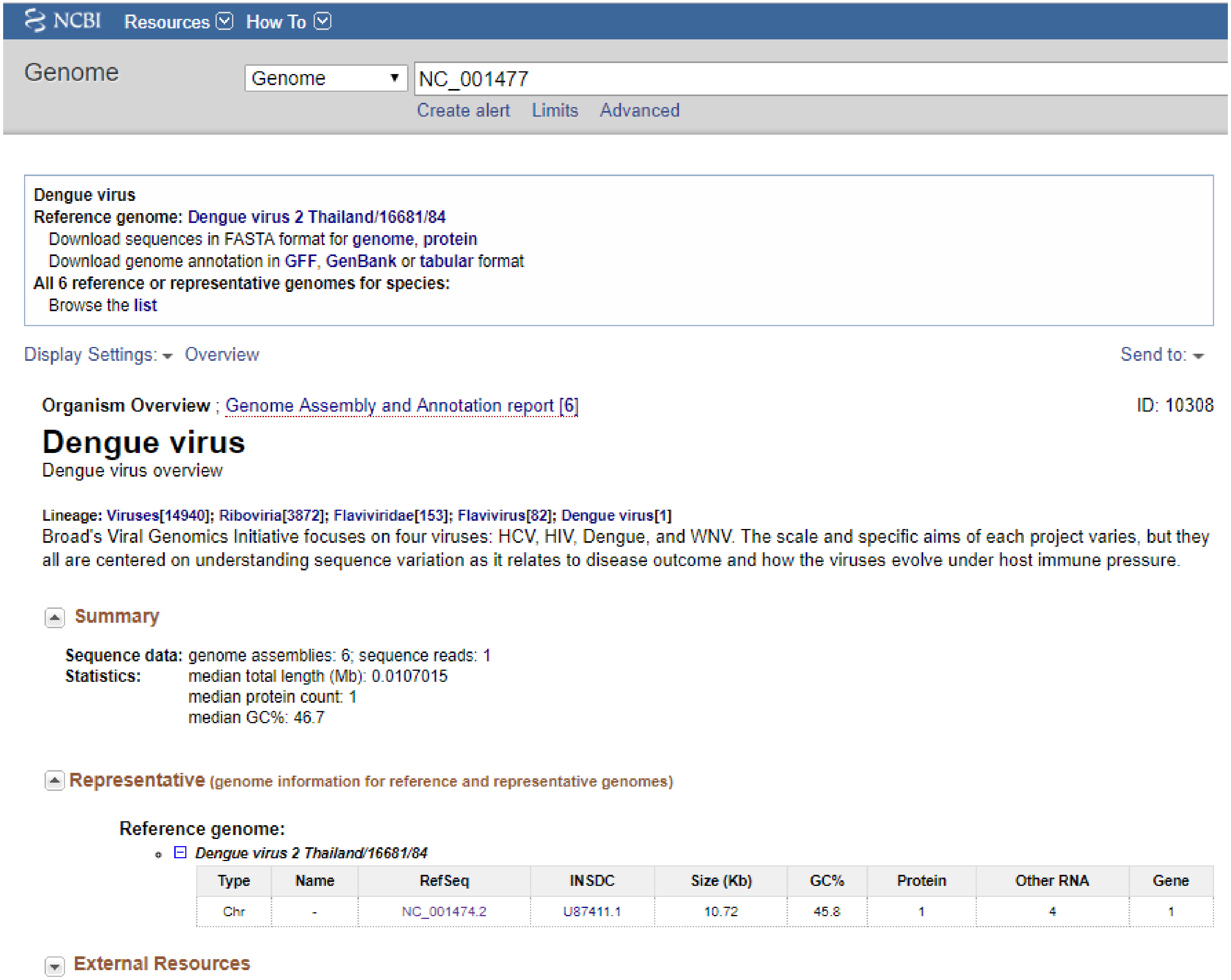
**Proteins**

| | | |
|---|---|---|
| Conserved Domains | 0 | conserved protein domains |
| Identical Protein Groups | 1 | protein sequences grouped by identity |
| Protein | 17 | protein sequences |
| Protein Clusters | 0 | sequence similarity-based protein clusters |
| Sparcle | 0 | functional categorization of proteins by domain architecture |
| Structure | 0 | experimentally-determined biomolecular structures |

**Chemicals**

| | | |
|---|---|---|
| BioSystems | 0 | molecular pathways with links to genes, proteins and chemicals |
| PubChem BioAssay | 0 | bioactivity screening studies |
| PubChem Compound | 0 | chemical information with structures, information and links |
| PubChem Substance | 0 | deposited substance and chemical information |

# Download sequences from NCBI



NC_001477

**Dengue virus**
**Reference genome:** Dengue virus 2 Thailand/16681/84
Download sequences in FASTA format for genome, protein
Download genome annotation in GFF, GenBank or tabular format
**All 6 reference or representative genomes for species:**
Browse the list

Organism Overview ; Genome Assembly and Annotation report [6]

ID: 10308

## Dengue virus

Dengue virus overview

**Lineage:** Viruses[14940]; Riboviria[3872]; Flaviviridae[153]; Flavivirus[82]; Dengue virus[1]
Broad's Viral Genomics Initiative focuses on four viruses: HCV, HIV, Dengue, and WNV. The scale and specific aims of each project varies, but they all are centered on understanding sequence variation as it relates to disease outcome and how the viruses evolve under host immune pressure.

### Summary

**Sequence data:** genome assemblies: 6; sequence reads: 1
**Statistics:** median total length (Mb): 0.0107015
median protein count: 1
median GC%: 46.7

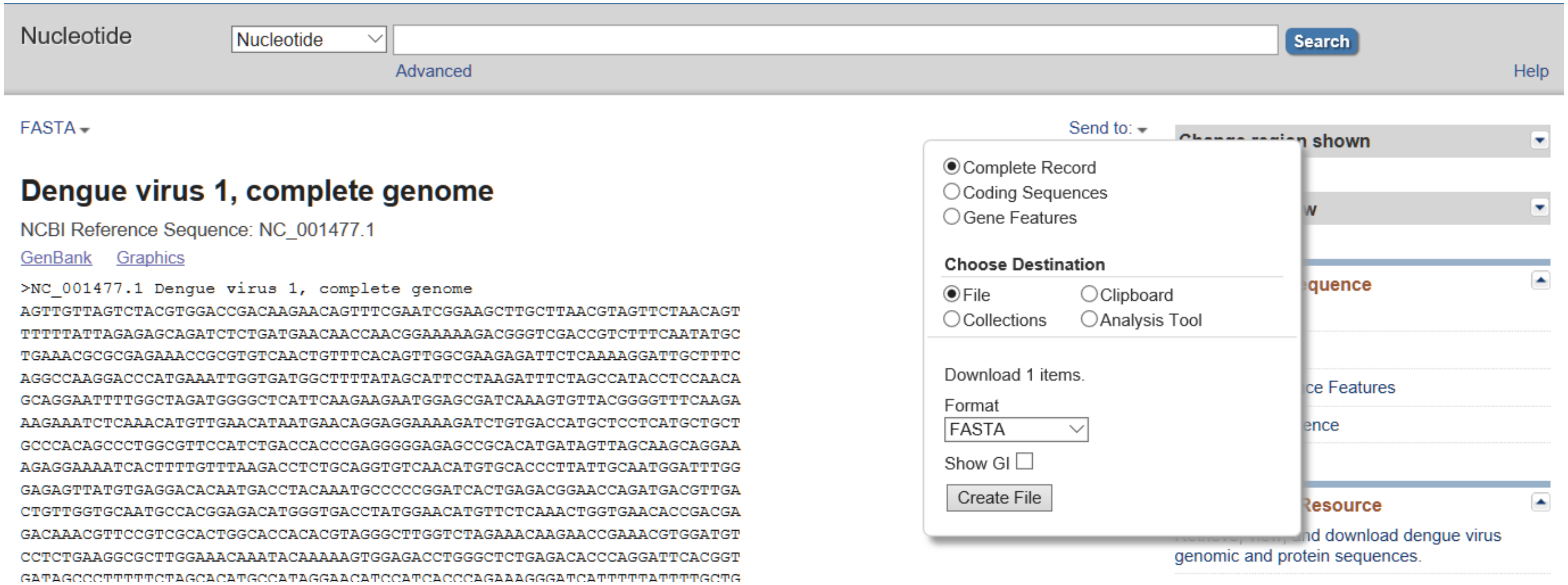
### Representative (genome information for reference and representative genomes)

**Reference genome:**
*Dengue virus 2 Thailand/16681/84*

| Type | Name | RefSeq | INSDC | Size (Kb) | GC% | Protein | Other RNA | Gene |
|---|---|---|---|---|---|---|---|---|
| Chr | - | NC_001474.2 | U87411.1 | 10.72 | 45.8 | 1 | 4 | 1 |

### External Resources

# Retrieving genome sequence data via the NCBI website



## Dengue virus 1, complete genome

NCBI Reference Sequence: NC_001477.1

GenBank Graphics

```
>NC_001477.1 Dengue virus 1, complete genome
AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCGAATCGGAAGCTTGCTTAACGTAGTTCTAACAGT
TTTTTATTAGAGAGCAGATCTCTGATGAACAACCAACGGAAAAAGACGGGTCGACCGTCTTTCAATATGC
TGAAACGCGCGAGAAACCGCGTGTCAACTGTTTCACAGTTGGCGAAGAGATTCTCAAAAGGATTGCTTTC
AGGCCAAGGACCCATGAAATTGGTGATGGCTTTTATAGCATTCCTAAGATTTCTAGCCATACCTCCAACA
GCAGGAATTTTGGCTAGATGGGGCTCATTCAAGAAGAATGGAGCGATCAAAGTGTTACGGGGTTTCAAGA
AAGAAATCTCAAACATGTTGAACATAATGAACAGGAGGAAAAGATCTGTGACCATGCTCCTCATGCTGCT
GCCCACAGCCCTGGCGTTCCATCTGACCACCCGAGGGGGAGAGCCGCACATGATAGTTAGCAAGCAGGAA
AGAGGAAAATCACTTTTGTTTAAGACCTCTGCAGGTGTCAACATGTGCACCCTTATTGCAATGGATTTGG
GAGAGTTATGTGAGGACACAATGACCTACAAATGCCCCCGGATCACTGAGACGGAACCAGATGACGTTGA
CTGTTGGTGCAATGCCACGGAGACATGGGTGACCTATGGAACATGTTCTCAAACTGGTGAACACCGACGA
GACAAACGTTCCGTCGCACTGGCACCACACGTAGGGCTTGGTCTAGAAACAAGAACCGAAACGTGGATGT
CCTCTGAAGGCGCTTGGAAACAAATACAAAAAGTGGAGACCTGGGCTCTGAGACACCCAGGATTCACGGT
GATAGCCCTTTTTCTAGCACATGCCATAGGAACATCCATCACCCAGAAAGGGATCATTTTTATTTTGCTG
```

● Complete Record
○ Coding Sequences
○ Gene Features

**Choose Destination**

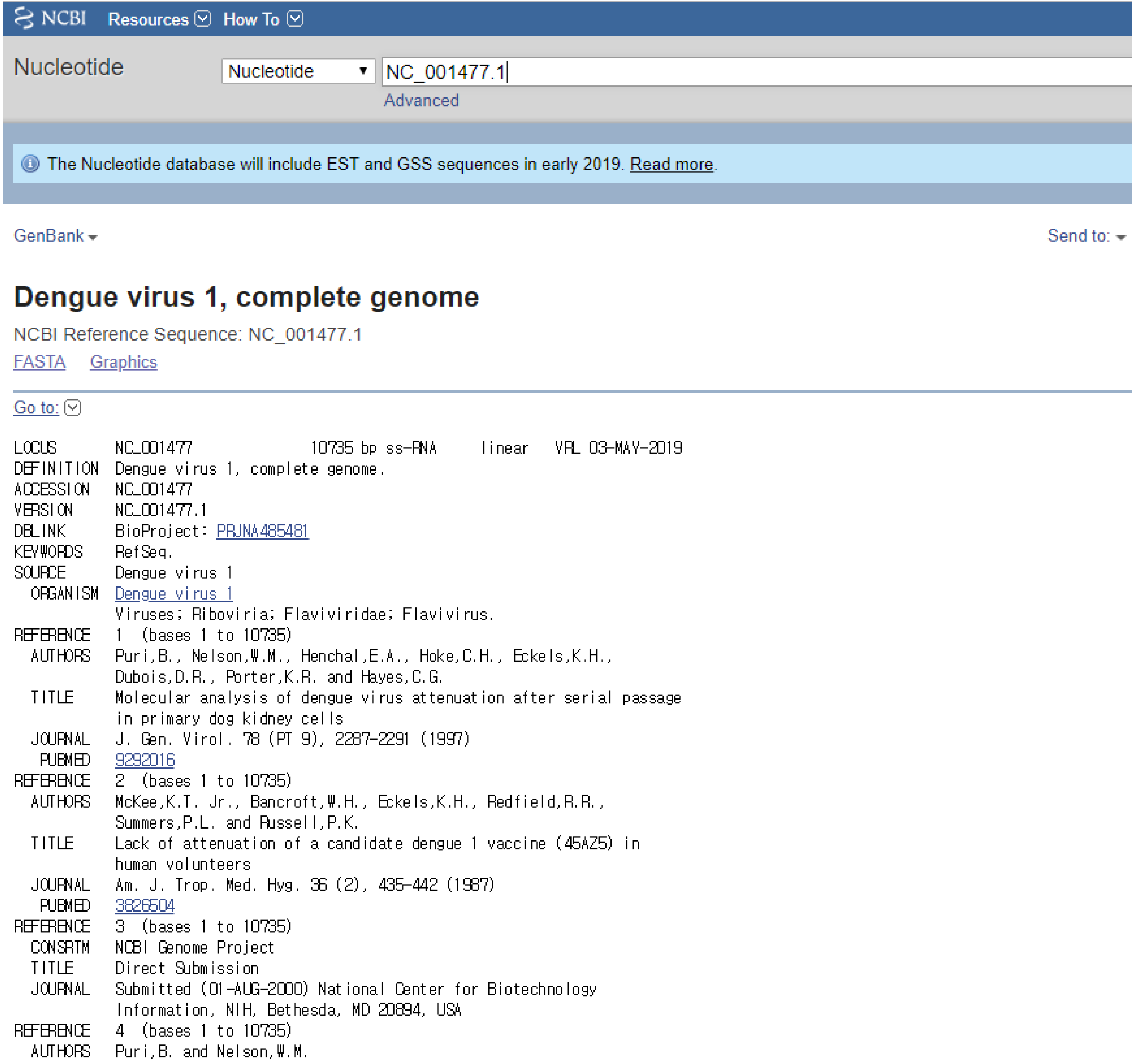
● File ○ Clipboard
○ Collections ○ Analysis Tool

Download 1 items.

Format: FASTA

Show GI □

Create File



The Nucleotide database will include EST and GSS sequences in early 2019. Read more.

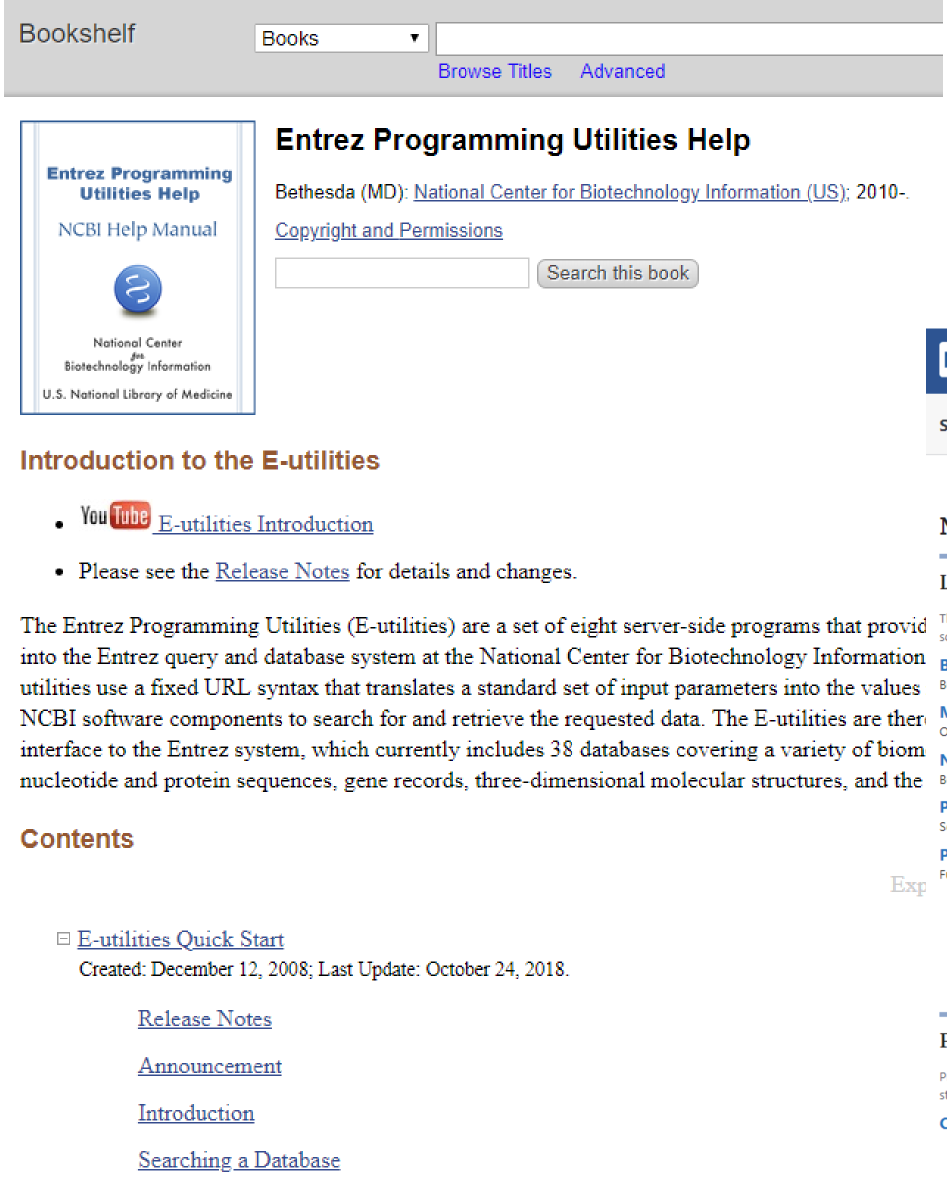
## Dengue virus 1, complete genome

NCBI Reference Sequence: NC_001477.1

```
LOCUS       NC_001477              10735 bp ss-RNA     linear   VRL 03-MAY-2019
DEFINITION  Dengue virus 1, complete genome.
ACCESSION   NC_001477
VERSION     NC_001477.1
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Dengue virus 1
  ORGANISM  Dengue virus 1
            Viruses; Riboviria; Flaviviridae; Flavivirus.
REFERENCE   1  (bases 1 to 10735)
  AUTHORS   Puri,B., Nelson,W.M., Henchal,E.A., Hoke,C.H., Eckels,K.H.,
            Dubois,D.R., Porter,K.R. and Hayes,C.G.
  TITLE     Molecular analysis of dengue virus attenuation after serial passage
            in primary dog kidney cells
  JOURNAL   J. Gen. Virol. 78 (PT 9), 2287-2291 (1997)
   PUBMED   9292016
REFERENCE   2  (bases 1 to 10735)
  AUTHORS   McKee,K.T. Jr., Bancroft,W.H., Eckels,K.H., Redfield,R.R.,
            Summers,P.L. and Russell,P.K.
  TITLE     Lack of attenuation of a candidate dengue 1 vaccine (45AZ5) in
            human volunteers
  JOURNAL   Am. J. Trop. Med. Hyg. 36 (2), 435-442 (1987)
   PUBMED   3826504
REFERENCE   3  (bases 1 to 10735)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (01-AUG-2000) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   4  (bases 1 to 10735)
  AUTHORS   Puri,B. and Nelson,W.M.
```

# Entrez Programming utilities (E-utilities)

**Entrez** is NCBI’s primary text search and retrieval system that integrates the PubMed database of biomedical literature with 38 other literature and molecular databases including DNA and protein sequence, structure, gene, genome, genetic variation and gene expression.



Bookshelf
Books ▾
Browse Titles
Advanced

Entrez Programming Utilities Help
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine

## Entrez Programming Utilities Help

Bethesda (MD): National Center for Biotechnology Information (US); 2010-.

Copyright and Permissions

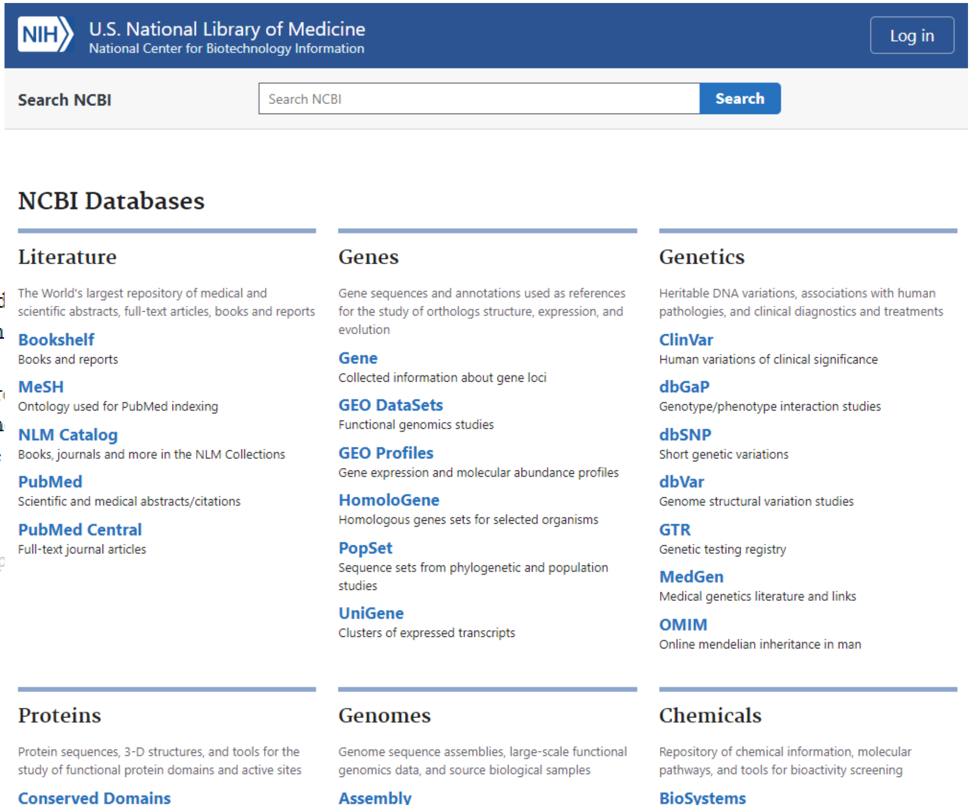
### Introduction to the E-utilities

- YouTube E-utilities Introduction
- Please see the Release Notes for details and changes.

The Entrez Programming Utilities (E-utilities) are a set of eight server-side programs that provid into the Entrez query and database system at the National Center for Biotechnology Information utilities use a fixed URL syntax that translates a standard set of input parameters into the values NCBI software components to search for and retrieve the requested data. The E-utilities are ther interface to the Entrez system, which currently includes 38 databases covering a variety of biom nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the

### Contents

Exp



NIH
U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
Search NCBI
Search

## NCBI Databases

### Literature

The World's largest repository of medical and scientific abstracts, full-text articles, books and reports

**Bookshelf**
Books and reports

**MeSH**
Ontology used for PubMed indexing

**NLM Catalog**
Books, journals and more in the NLM Collections

**PubMed**
Scientific and medical abstracts/citations

**PubMed Central**
Full-text journal articles

### Genes

Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

**Gene**
Collected information about gene loci

**GEO DataSets**
Functional genomics studies

**GEO Profiles**
Gene expression and molecular abundance profiles

**HomoloGene**
Homologous genes sets for selected organisms

**PopSet**
Sequence sets from phylogenetic and population studies

**UniGene**
Clusters of expressed transcripts

### Genetics

Heritable DNA variations, associations with human pathologies, and clinical diagnostics and treatments

**ClinVar**
Human variations of clinical significance

**dbGaP**
Genotype/phenotype interaction studies

**dbSNP**
Short genetic variations

**dbVar**
Genome structural variation studies

**GTR**
Genetic testing registry

**MedGen**
Medical genetics literature and links

**OMIM**
Online mendelian inheritance in man

### Proteins

Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

**Conserved Domains**

### Genomes

Genome sequence assemblies, large-scale functional genomics data, and source biological samples

**Assembly**

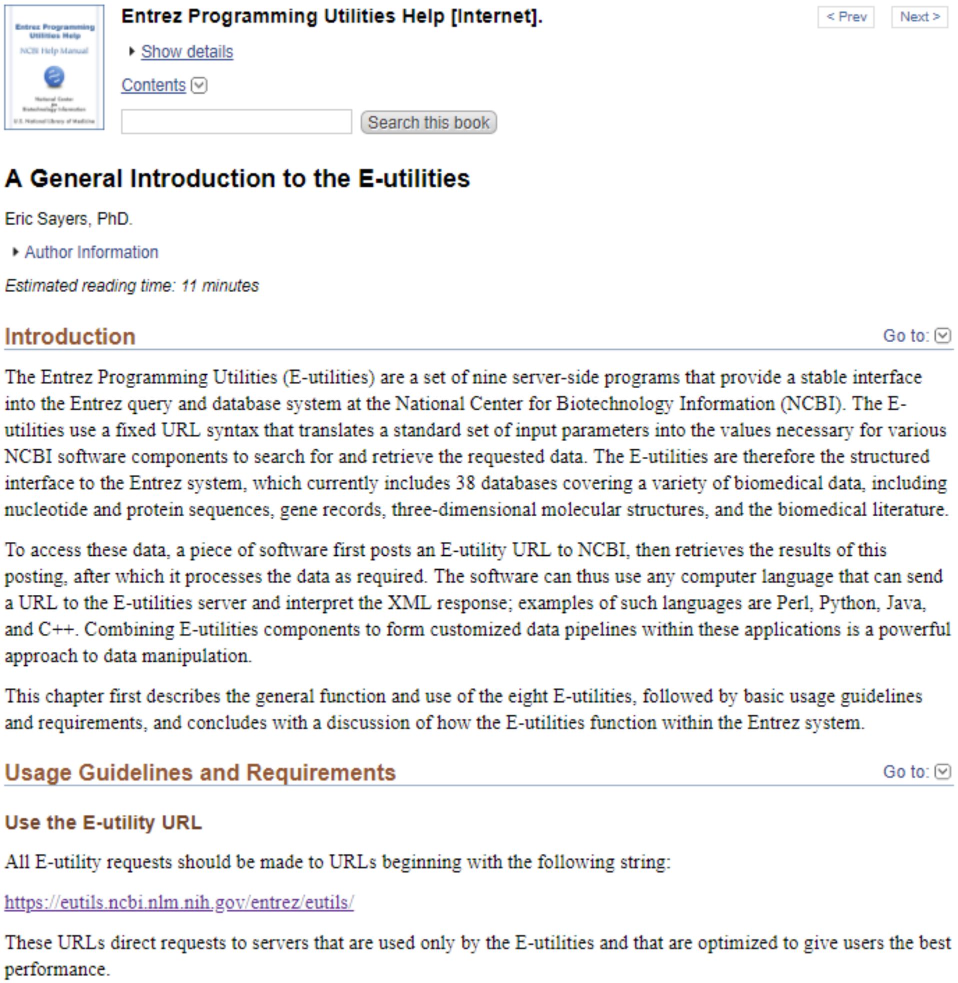
### Chemicals

Repository of chemical information, molecular pathways, and tools for bioactivity screening

**BioSystems**

# A General Introduction to the E-utilities

https://www.ncbi.nlm.nih.gov/books/NBK25497/table/chapter2.T._entrez_unique_identifiers_ui/?report=objectonly



**Entrez Programming Utilities Help [Internet].**

## A General Introduction to the E-utilities

Eric Sayers, PhD.

Author Information

*Estimated reading time: 11 minutes*

### Introduction

The Entrez Programming Utilities (E-utilities) are a set of nine server-side programs that provide a stable interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI). The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data. The E-utilities are therefore the structured interface to the Entrez system, which currently includes 38 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature.

To access these data, a piece of software first posts an E-utility URL to NCBI, then retrieves the results of this posting, after which it processes the data as required. The software can thus use any computer language that can send a URL to the E-utilities server and interpret the XML response; examples of such languages are Perl, Python, Java, and C++. Combining E-utilities components to form customized data pipelines within these applications is a powerful approach to data manipulation.

This chapter first describes the general function and use of the eight E-utilities, followed by basic usage guidelines and requirements, and concludes with a discussion of how the E-utilities function within the Entrez system.

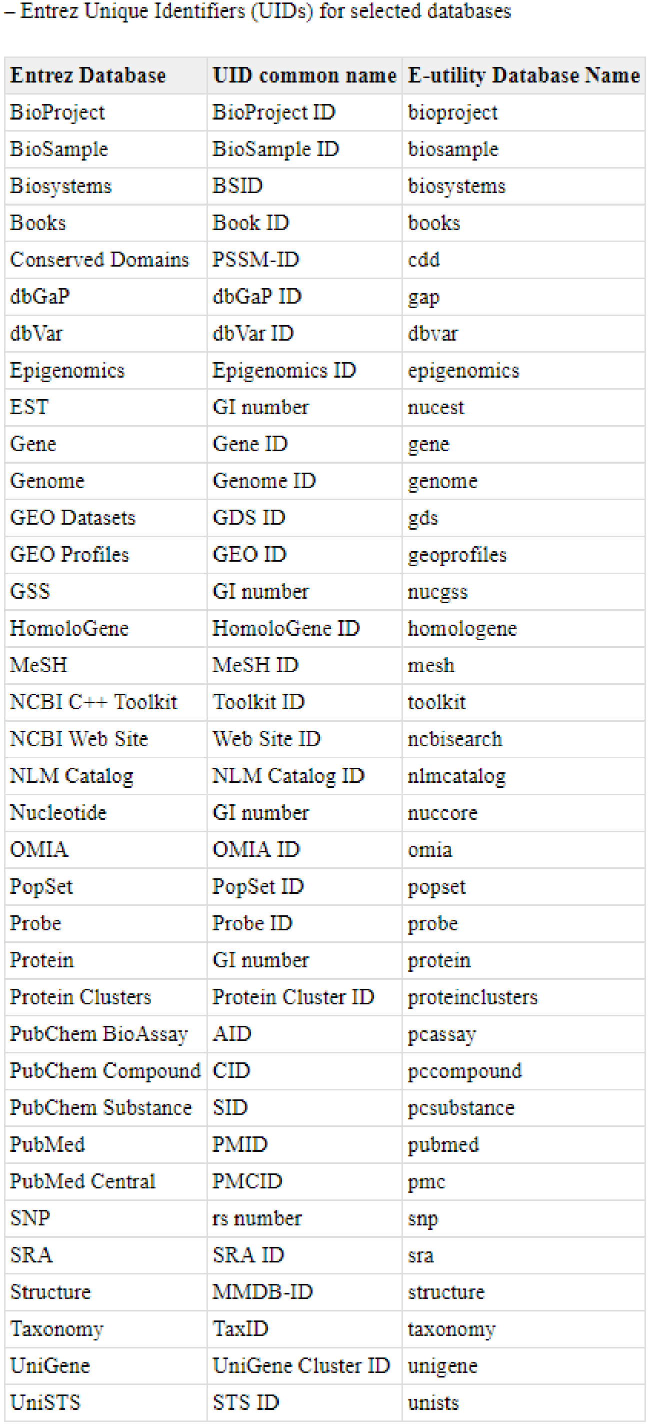
### Usage Guidelines and Requirements

#### Use the E-utility URL

All E-utility requests should be made to URLs beginning with the following string:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/

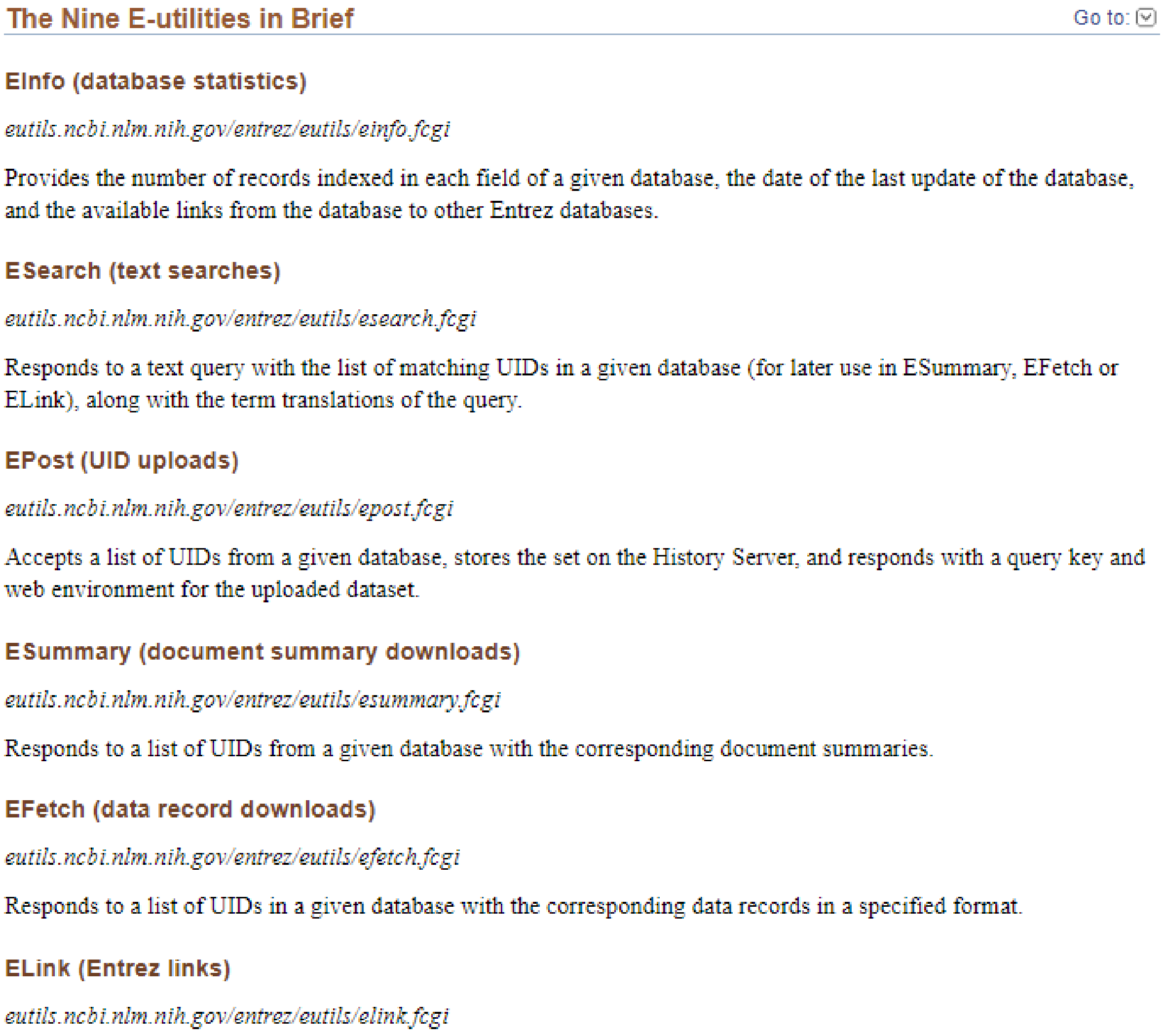
These URLs direct requests to servers that are used only by the E-utilities and that are optimized to give users the best performance.



– Entrez Unique Identifiers (UIDs) for selected databases

| Entrez Database | UID common name | E-utility Database Name |
|---|---|---|
| BioProject | BioProject ID | bioproject |
| BioSample | BioSample ID | biosample |
| Biosystems | BSID | biosystems |
| Books | Book ID | books |
| Conserved Domains | PSSM-ID | cdd |
| dbGaP | dbGaP ID | gap |
| dbVar | dbVar ID | dbvar |
| Epigenomics | Epigenomics ID | epigenomics |
| EST | GI number | nucest |
| Gene | Gene ID | gene |
| Genome | Genome ID | genome |
| GEO Datasets | GDS ID | gds |
| GEO Profiles | GEO ID | geoprofiles |
| GSS | GI number | nucgss |
| HomoloGene | HomoloGene ID | homologene |
| MeSH | MeSH ID | mesh |
| NCBI C++ Toolkit | Toolkit ID | toolkit |
| NCBI Web Site | Web Site ID | ncbisearch |
| NLM Catalog | NLM Catalog ID | nlmcatalog |
| Nucleotide | GI number | nuccore |
| OMIA | OMIA ID | omia |
| PopSet | PopSet ID | popset |
| Probe | Probe ID | probe |
| Protein | GI number | protein |
| Protein Clusters | Protein Cluster ID | proteinclusters |
| PubChem BioAssay | AID | pcassay |
| PubChem Compound | CID | pccompound |
| PubChem Substance | SID | pcsubstance |
| PubMed | PMID | pubmed |
| PubMed Central | PMCID | pmc |
| SNP | rs number | snp |
| SRA | SRA ID | sra |
| Structure | MMDB-ID | structure |
| Taxonomy | TaxID | taxonomy |
| UniGene | UniGene Cluster ID | unigene |
| UniSTS | STS ID | unists |

# Basic usage of E-utilities



## The Nine E-utilities in Brief

### EInfo (database statistics)

*eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi*

Provides the number of records indexed in each field of a given database, the date of the last update of the database, and the available links from the database to other Entrez databases.

### ESearch (text searches)

*eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi*

Responds to a text query with the list of matching UIDs in a given database (for later use in ESummary, EFetch or ELink), along with the term translations of the query.

### EPost (UID uploads)

*eutils.ncbi.nlm.nih.gov/entrez/eutils/epost.fcgi*

Accepts a list of UIDs from a given database, stores the set on the History Server, and responds with a query key and web environment for the uploaded dataset.

### ESummary (document summary downloads)

*eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi*

Responds to a list of UIDs from a given database with the corresponding document summaries.

### EFetch (data record downloads)

*eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi*

Responds to a list of UIDs in a given database with the corresponding data records in a specified format.

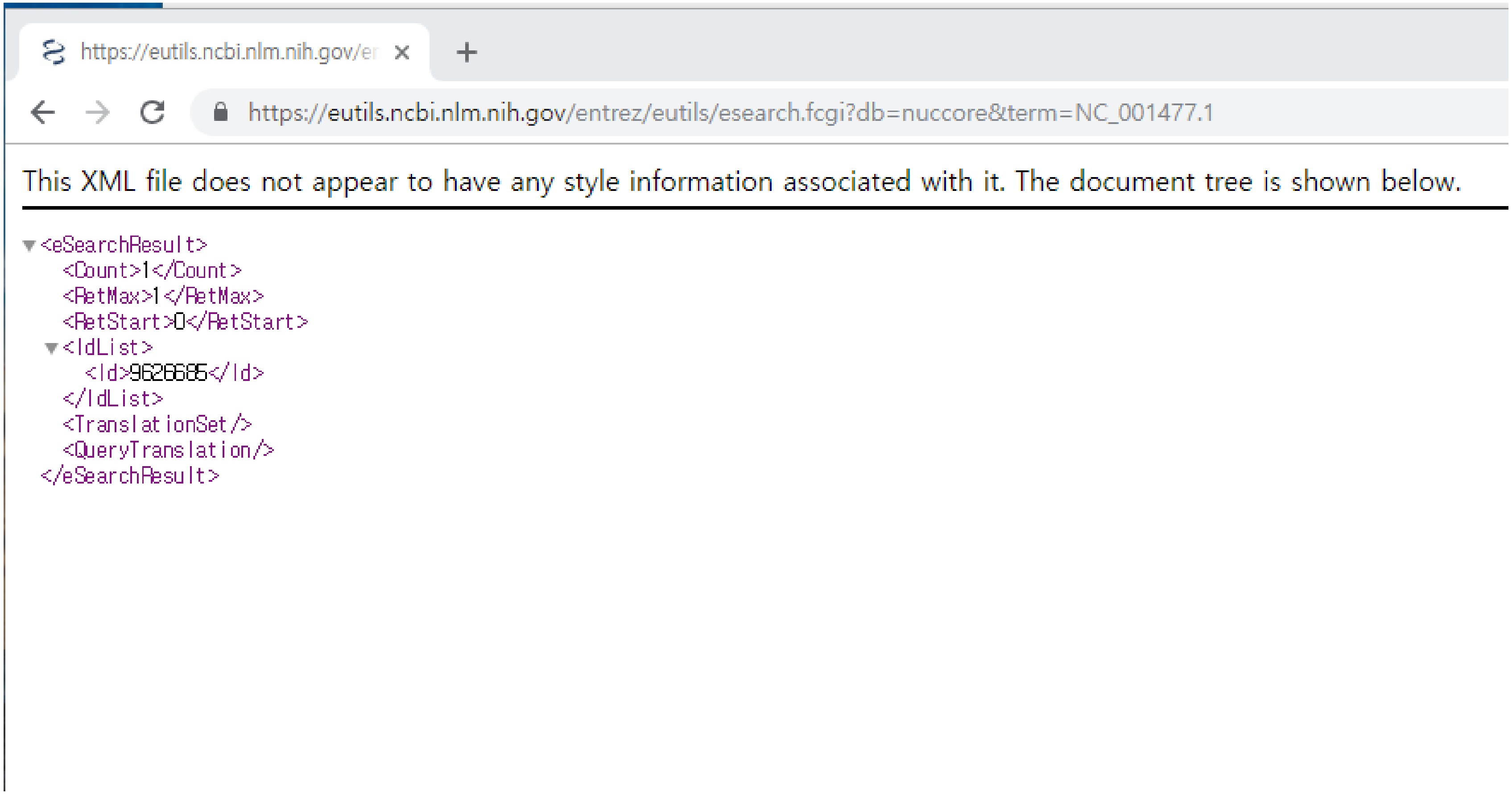
### ELink (Entrez links)

*eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi*

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/

```
esearch.fcgi?db=<database>&term=<query>
esummary.fcgi?db=<database>&id=<uid_list>
efetch.fcgi?db=<database>&id=<uid_list>&rettype=<retrieval_type>&retmode=<retrieval_mode>
```

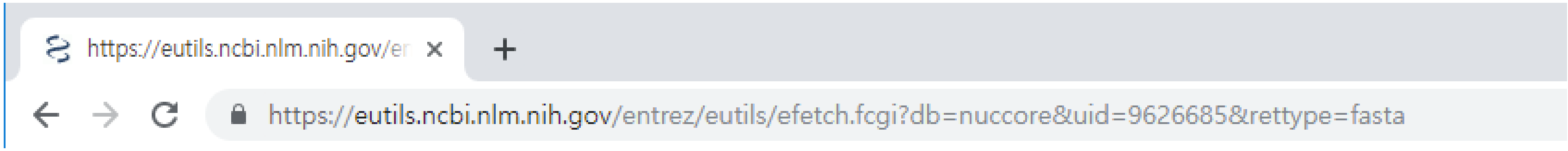
# Exercise 10-4)
# Download NC_001477.1 sequence



https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=nuccore&term=NC_001477.1

This XML file does not appear to have any style information associated with it. The document tree is shown below.

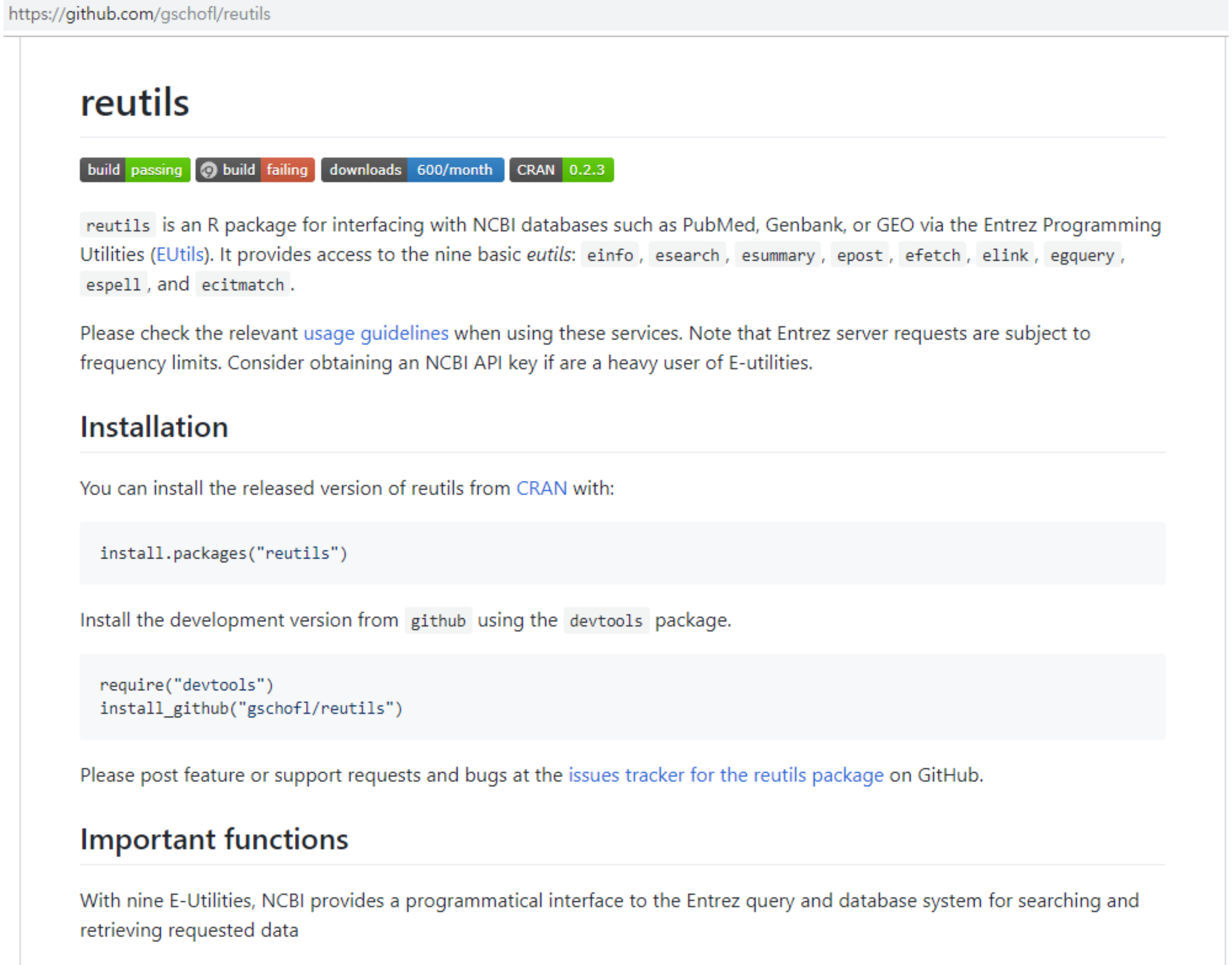
```
▼<eSearchResult>
   <Count>1</Count>
   <RetMax>1</RetMax>
   <RetStart>0</RetStart>
 ▼<IdList>
     <Id>9626685</Id>
   </IdList>
   <TranslationSet/>
   <QueryTranslation/>
 </eSearchResult>
```



https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&uid=9626685&rettype=fasta

# reutils

- https://github.com/gschofl/reutils



https://github.com/gschofl/reutils

## reutils

build passing | build failing | downloads 600/month | CRAN 0.2.3

`reutils` is an R package for interfacing with NCBI databases such as PubMed, Genbank, or GEO via the Entrez Programming Utilities (EUtils). It provides access to the nine basic *eutils*: `einfo`, `esearch`, `esummary`, `epost`, `efetch`, `elink`, `egquery`, `espell`, and `ecitmatch`.

Please check the relevant usage guidelines when using these services. Note that Entrez server requests are subject to frequency limits. Consider obtaining an NCBI API key if are a heavy user of E-utilities.

### Installation

You can install the released version of reutils from CRAN with:

```
install.packages("reutils")
```

Install the development version from `github` using the `devtools` package.

```
require("devtools")
install_github("gschofl/reutils")
```

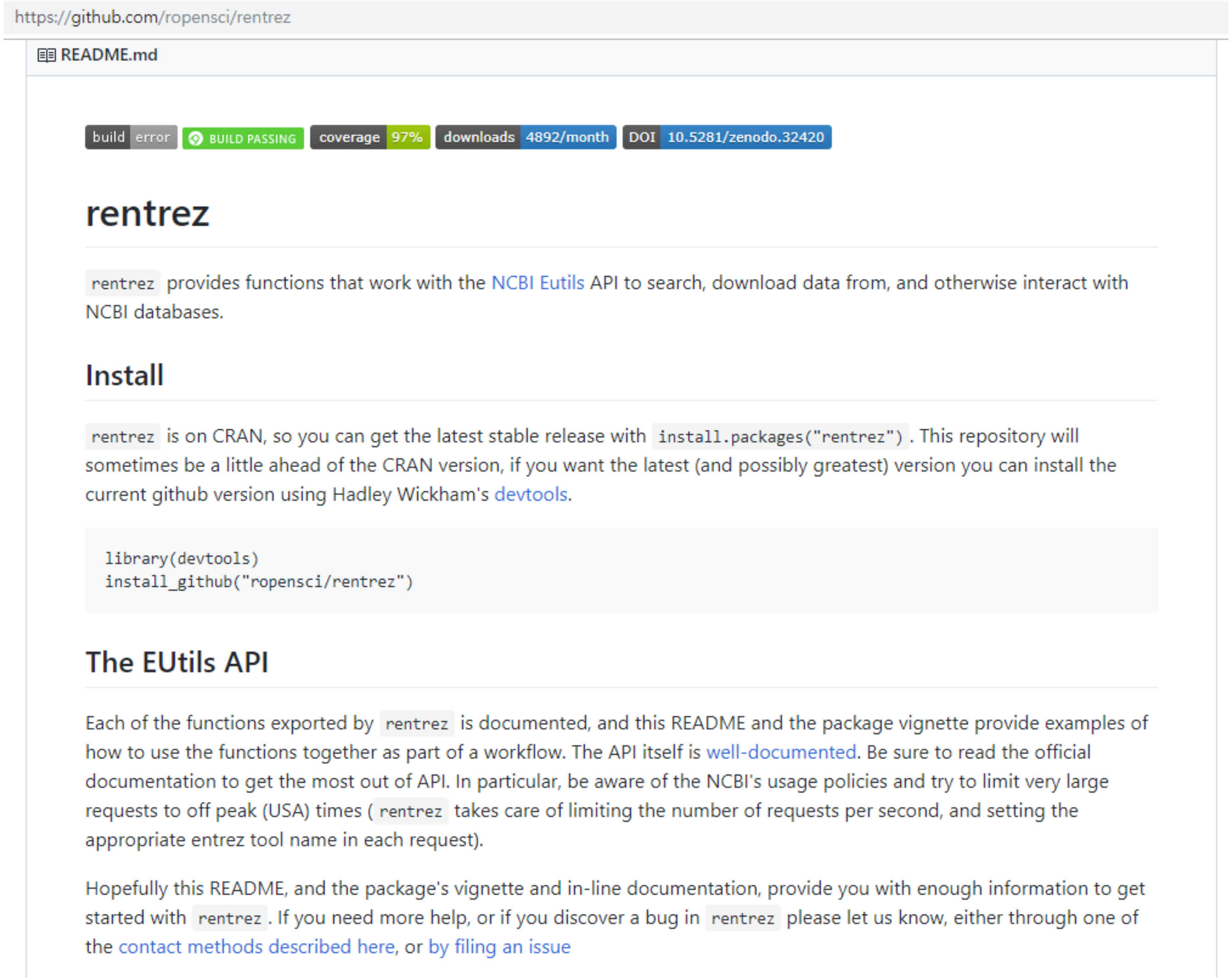
Please post feature or support requests and bugs at the issues tracker for the reutils package on GitHub.

### Important functions

With nine E-Utilities, NCBI provides a programmatical interface to the Entrez query and database system for searching and retrieving requested data

# rentrez

- https://github.com/ropensci/rentrez



https://github.com/ropensci/rentrez

README.md

## rentrez

`rentrez` provides functions that work with the NCBI Eutils API to search, download data from, and otherwise interact with NCBI databases.

### Install

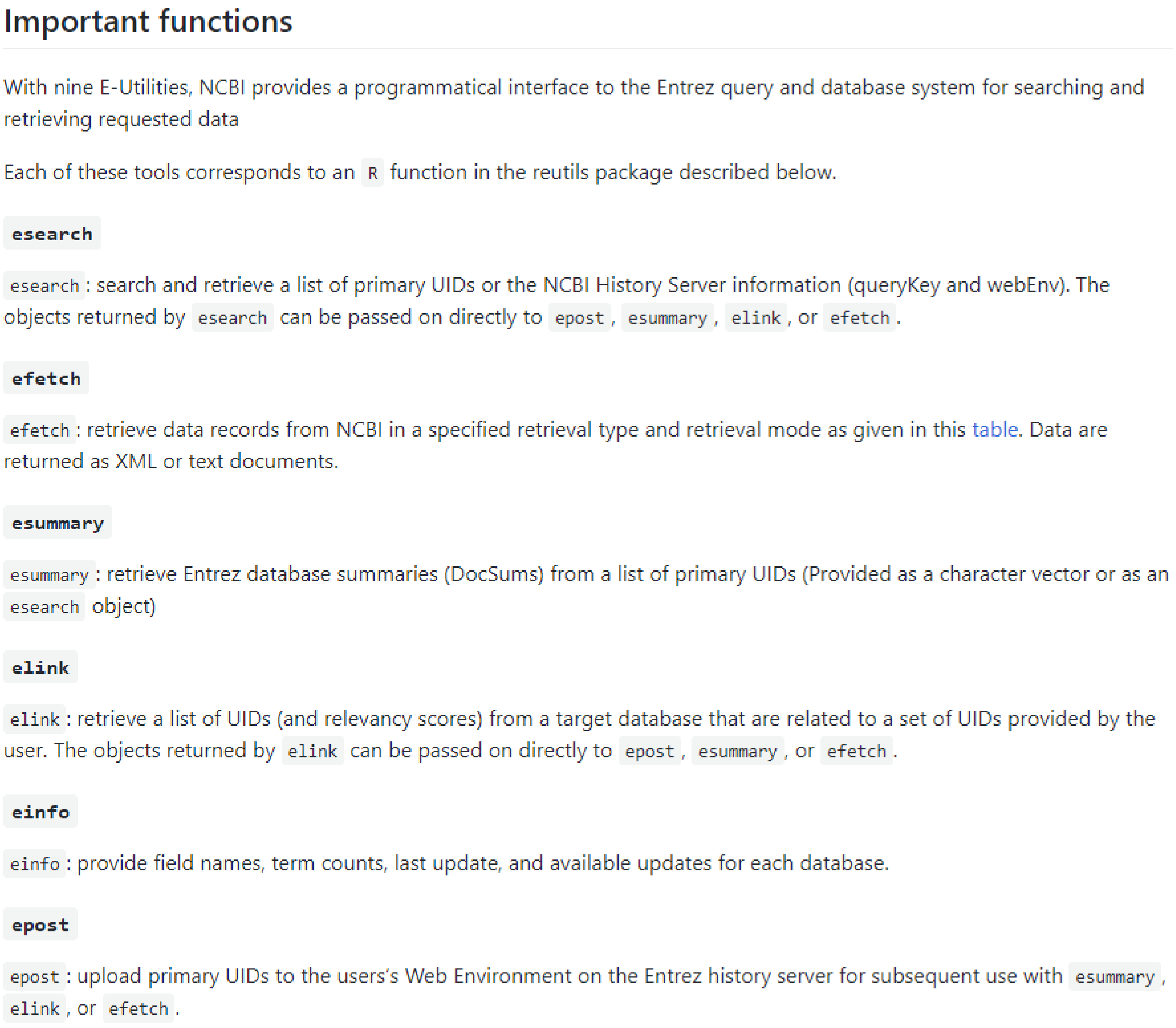
`rentrez` is on CRAN, so you can get the latest stable release with `install.packages("rentrez")`. This repository will sometimes be a little ahead of the CRAN version, if you want the latest (and possibly greatest) version you can install the current github version using Hadley Wickham's devtools.

```
library(devtools)
install_github("ropensci/rentrez")
```

### The EUtils API

Each of the functions exported by `rentrez` is documented, and this README and the package vignette provide examples of how to use the functions together as part of a workflow. The API itself is well-documented. Be sure to read the official documentation to get the most out of API. In particular, be aware of the NCBI's usage policies and try to limit very large requests to off peak (USA) times (`rentrez` takes care of limiting the number of requests per second, and setting the appropriate entrez tool name in each request).

Hopefully this README, and the package's vignette and in-line documentation, provide you with enough information to get started with `rentrez`. If you need more help, or if you discover a bug in `rentrez` please let us know, either through one of the contact methods described here, or by filing an issue



## Important functions

With nine E-Utilities, NCBI provides a programmatical interface to the Entrez query and database system for searching and retrieving requested data

Each of these tools corresponds to an `R` function in the reutils package described below.

### `esearch`

`esearch`: search and retrieve a list of primary UIDs or the NCBI History Server information (queryKey and webEnv). The objects returned by `esearch` can be passed on directly to `epost`, `esummary`, `elink`, or `efetch`.

### `efetch`

`efetch`: retrieve data records from NCBI in a specified retrieval type and retrieval mode as given in this table. Data are returned as XML or text documents.

### `esummary`

`esummary`: retrieve Entrez database summaries (DocSums) from a list of primary UIDs (Provided as a character vector or as an `esearch` object)

### `elink`

`elink`: retrieve a list of UIDs (and relevancy scores) from a target database that are related to a set of UIDs provided by the user. The objects returned by `elink` can be passed on directly to `epost`, `esummary`, or `efetch`.

### `einfo`

`einfo`: provide field names, term counts, last update, and available updates for each database.

### `epost`

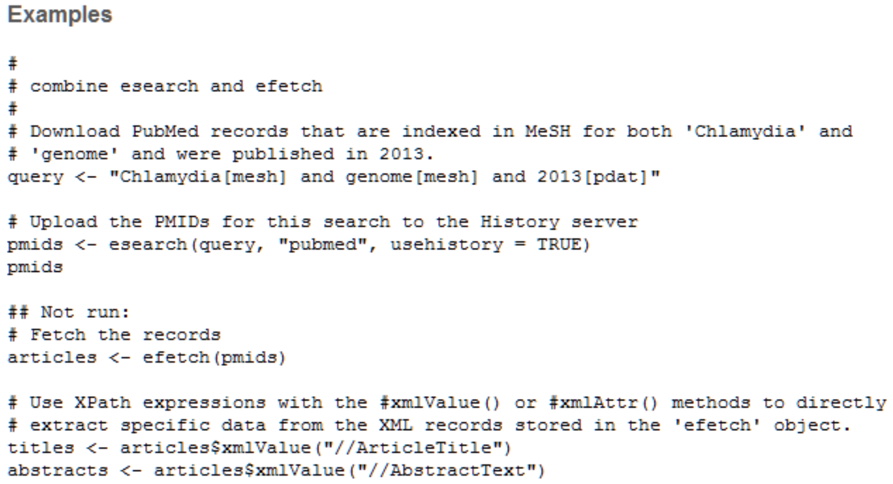
`epost`: upload primary UIDs to the users's Web Environment on the Entrez history server for subsequent use with `esummary`, `elink`, or `efetch`.

# Download NC_001477.1 sequence

```
library(rentrez)

nuc_search <- entrez_search(db = "nuccore", term = "NC_001477.1")
nuc_fetech <- entrez_fetch(db = "nuccore", id=nuc_search$ids, rettype = "fasta")

nuc_search <- entrez_search(db = "nuccore", term = "NC_001477.1", use_history = TRUE)
nuc_fetech <- entrez_fetch(db = "nuccore", web_history = nuc_search$web_history, rettype = "fasta")

write.table(nuc_fetech, file="nc_001477.fasta", quote=F, row.names=F, col.names=F)
mydna <- readDNAStringSet("nc_001477.fasta")
```

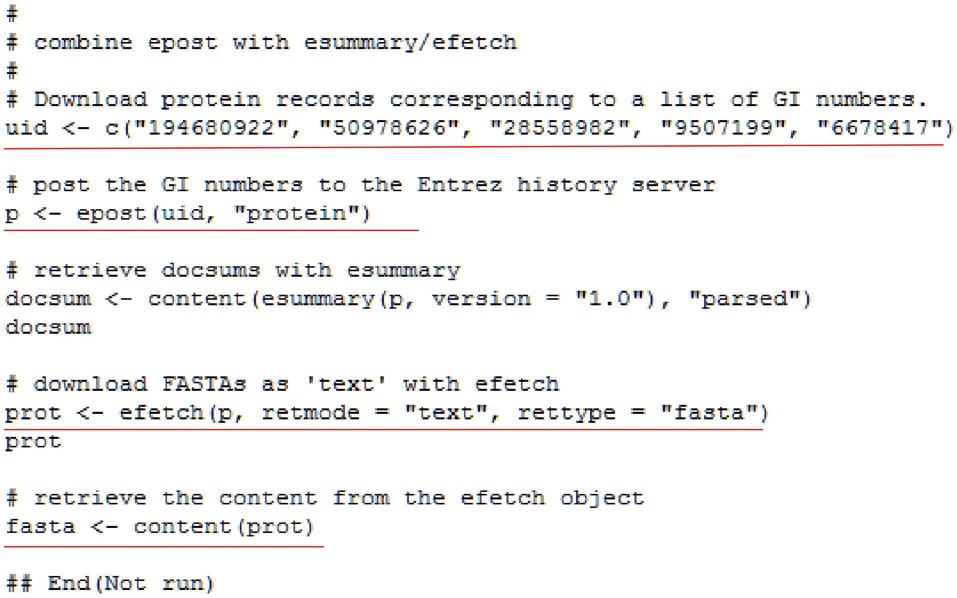
# reutils package

An interface to NCBI databases such as PubMed, GenBank, or GEO powered by the Entrez Programming Utilities



### Examples

```
#
# combine esearch and efetch
#
# Download PubMed records that are indexed in MeSH for both 'Chlamydia' and
# 'genome' and were published in 2013.
query <- "Chlamydia[mesh] and genome[mesh] and 2013[pdat]"

# Upload the PMIDs for this search to the History server
pmids <- esearch(query, "pubmed", usehistory = TRUE)
pmids

## Not run:
# Fetch the records
articles <- efetch(pmids)

# Use XPath expressions with the #xmlValue() or #xmlAttr() methods to directly
# extract specific data from the XML records stored in the 'efetch' object.
titles <- articles$xmlValue("//ArticleTitle")
abstracts <- articles$xmlValue("//AbstractText")
```



```
#
# combine epost with esummary/efetch
#
# Download protein records corresponding to a list of GI numbers.
uid <- c("194680922", "50978626", "28558982", "9507199", "6678417")

# post the GI numbers to the Entrez history server
p <- epost(uid, "protein")

# retrieve docsums with esummary
docsum <- content(esummary(p, version = "1.0"), "parsed")
docsum

# download FASTAs as 'text' with efetch
prot <- efetch(p, retmode = "text", rettype = "fasta")
prot

# retrieve the content from the efetch object
fasta <- content(prot)

## End(Not run)
```

# Download NC_001477.1 sequence

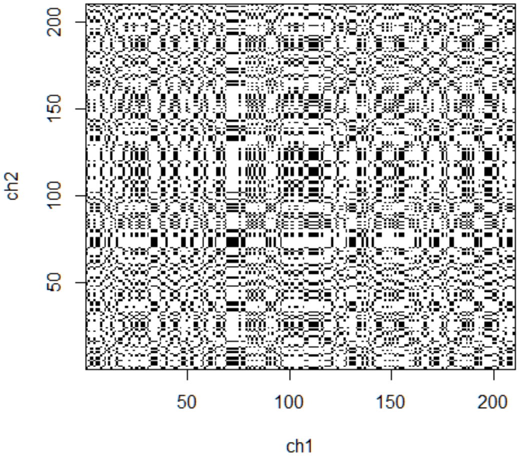
Dengue virus: DEN-1, DEN-2, DEN-3, and DEN-4
Accession no.: NC_001477, NC_001474, NC_001475 and NC_002640

```
acc <- c("NC_001477", "NC_001474", "NC_001475", "NC_002640")
ep <- epost(acc, "nuccore")
ef <- efetch(ep, retmode = "text", rettype = "fasta")
nc <- content(ef)
nc

## write the sequences to a file
write.table(nc, file="den.fasta", quote=F, col.names=F, row.names=F)

## read the sequences
den.seqs <- readDNAStringSet("den.fasta")
```

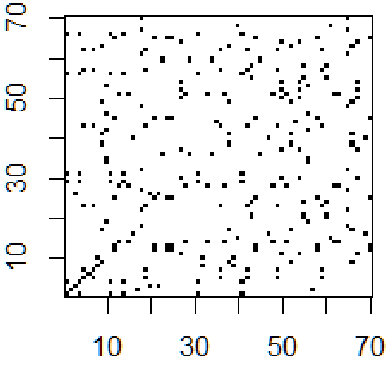
# DENGUE Sequence

```
## GC contents
letterFrequency(den.seqs, letters="GC", as.prob=T)

## see all base contents
alphabetFrequency(den.seqs, baseOnly=T, as.prob=T)
alphabetFrequency(den.seqs, baseOnly=T, as.prob=T, collapse=T)
```

```
## Biological Sequences Retrieval and Analysis
library(seqinr)

## convert DNAString to string
str1 <- toString(den.seqs[[1]])
str2 <- toString(den.seqs[[2]])
str1
ch1 <- s2c(str1)[1:210]
ch2 <- s2c(str2)[1:210]

dotPlot(ch1, ch2)
```



```
aa1 <- Biostrings::translate(den.seqs[[1]])
aa2 <- Biostrings::translate(den.seqs[[2]])
dotPlot(s2c(toString(aa1))[1:70],
s2c(toString(aa2))[1:70])
```



# Next

- Sequence analysis in R II
- Install Bioconductor packages
  - DECIPHER
- 다음시간 5/22 (수) 중회의실