refseqR : operaciones computacionales comunes con registros de la colección RefSeq (NCBI)

Jose V. Die Ramón
Dept. Genética - ETSIAM
Universidad de Córdoba

@jdieramon

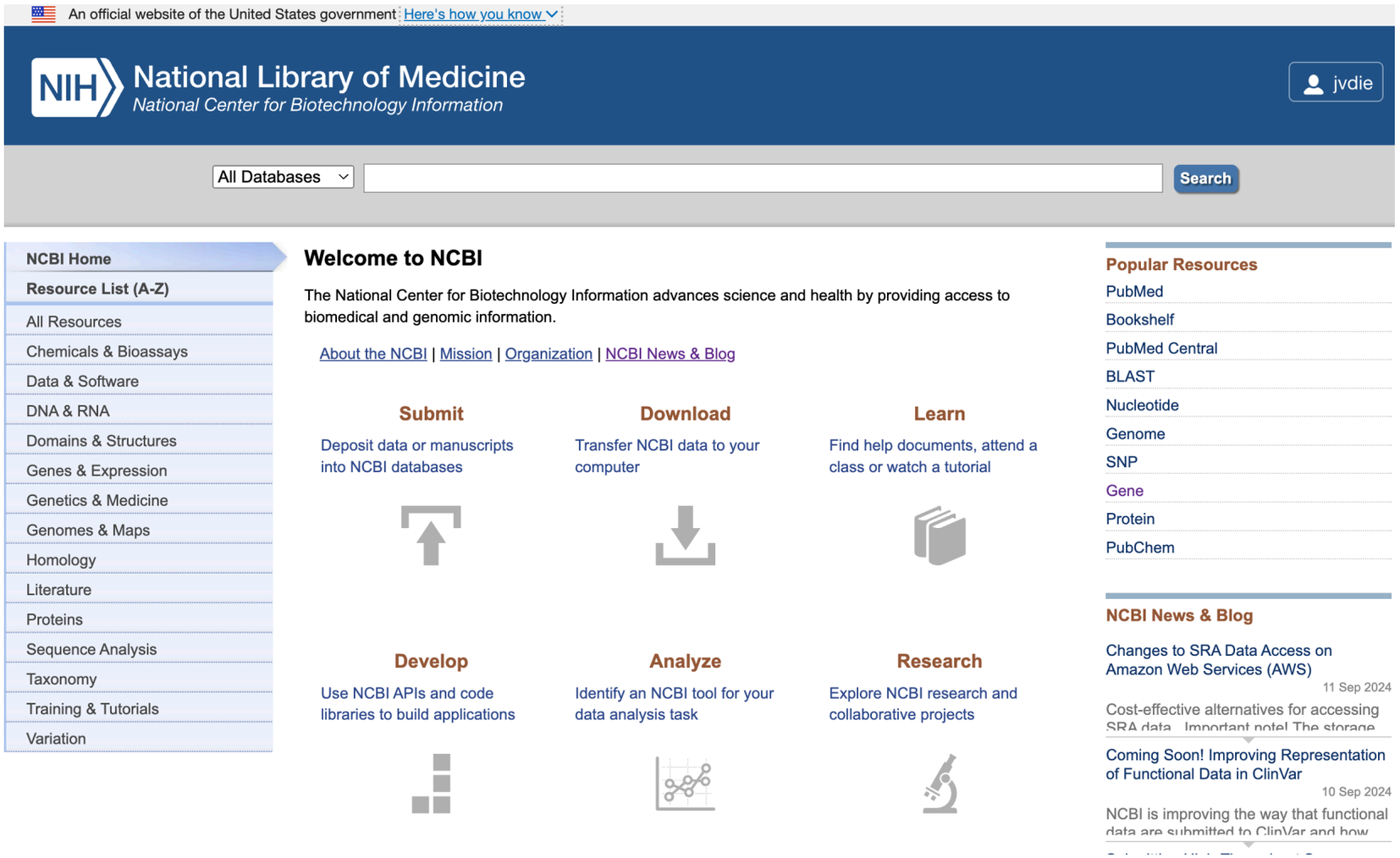
About me

- 2009 - PhD in Plant Genetics (UCO)
- 2012 - 2017 US Department of Agriculture
- 2017 - Dept. Genetics (UCO)
- 2018 - Visiting Bioinformaticians Program (NCBI)

- Broad Interests : intersection of Genomics & Data Science, molecular breeding , **R**

@jdieramon

National Center for Biotechnology Information



To develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease.

Fuente : https://www.ncbi.nlm.nih.gov/home/about/mission/

National Center for Biotechnology Information

RefSeq RefSeq Search

About RefSeq

The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses. [more...]

RefSeq genomes are copies of selected assembled genomes available in GenBank. RefSeq transcript and protein records are generated by several processes including:

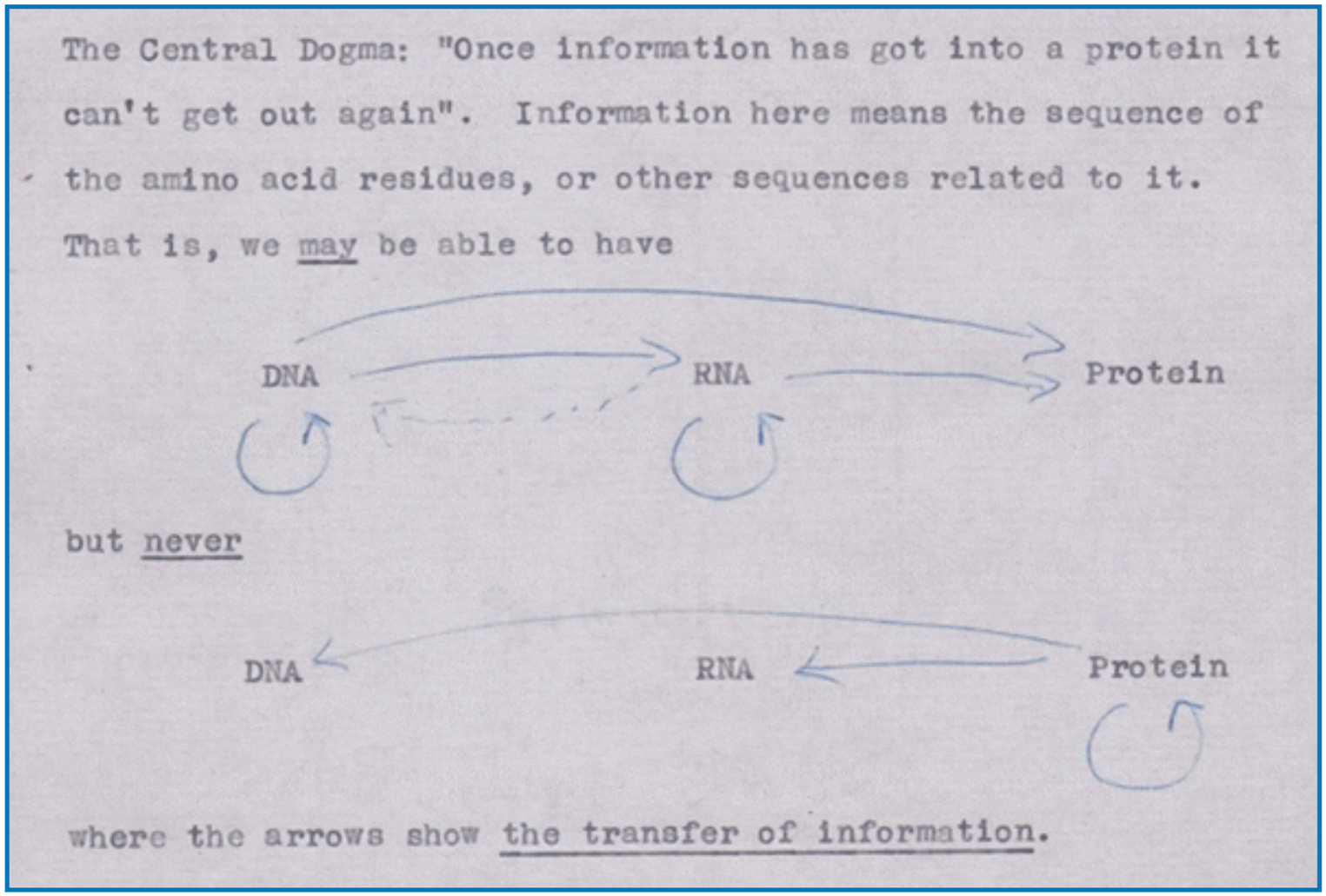
- Computation
  - Eukaryotic Genome Annotation Pipeline
  - Prokaryotic Genome Annotation Pipeline
- Manual curation
- Propagation from annotated genomes that are submitted to members of the International Nucleotide Sequence Database Collaboration (INSDC)

Scope

NCBI provides RefSeqs for taxonomically diverse organisms including archaea, bacteria, eukaryotes, and viruses. References sequences are provided for genomes, transcripts, and proteins. Some targeted loci projects are included in RefSeq including: RefSeqGene, fungal ITS, and rRNA loci. New or updated records are added to the collection as data become publicly available.

The lecture that changed Biology

Crick en Cold Spring Harbor Symposium (1963)
Crédito : Cold Spring Harbor Laboratory



The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have

DNA → RNA → Protein

but never

DNA ← RNA ← Protein

where the arrows show the transfer of information.

Crédito : Wellcome Library, London

Flow of molecular information

NATURE VOL. 227 AUGUST 8 1970

Central Dogma of Molecular Biology

by
FRANCIS CRICK
MRC Laboratory of Molecular Biology,
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

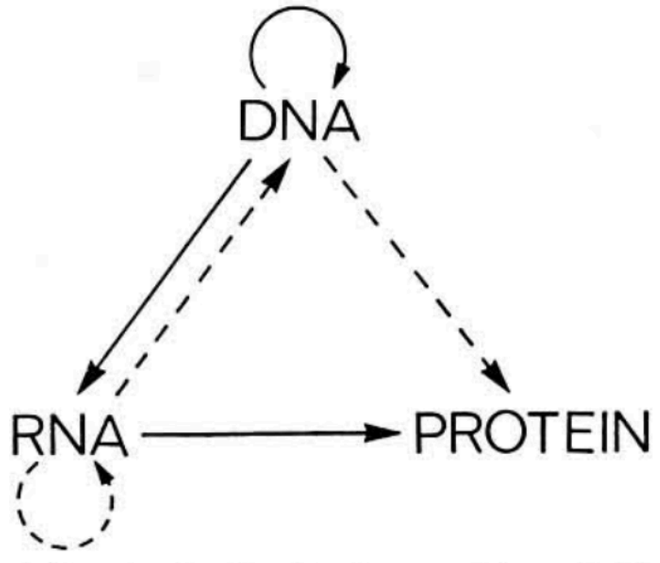
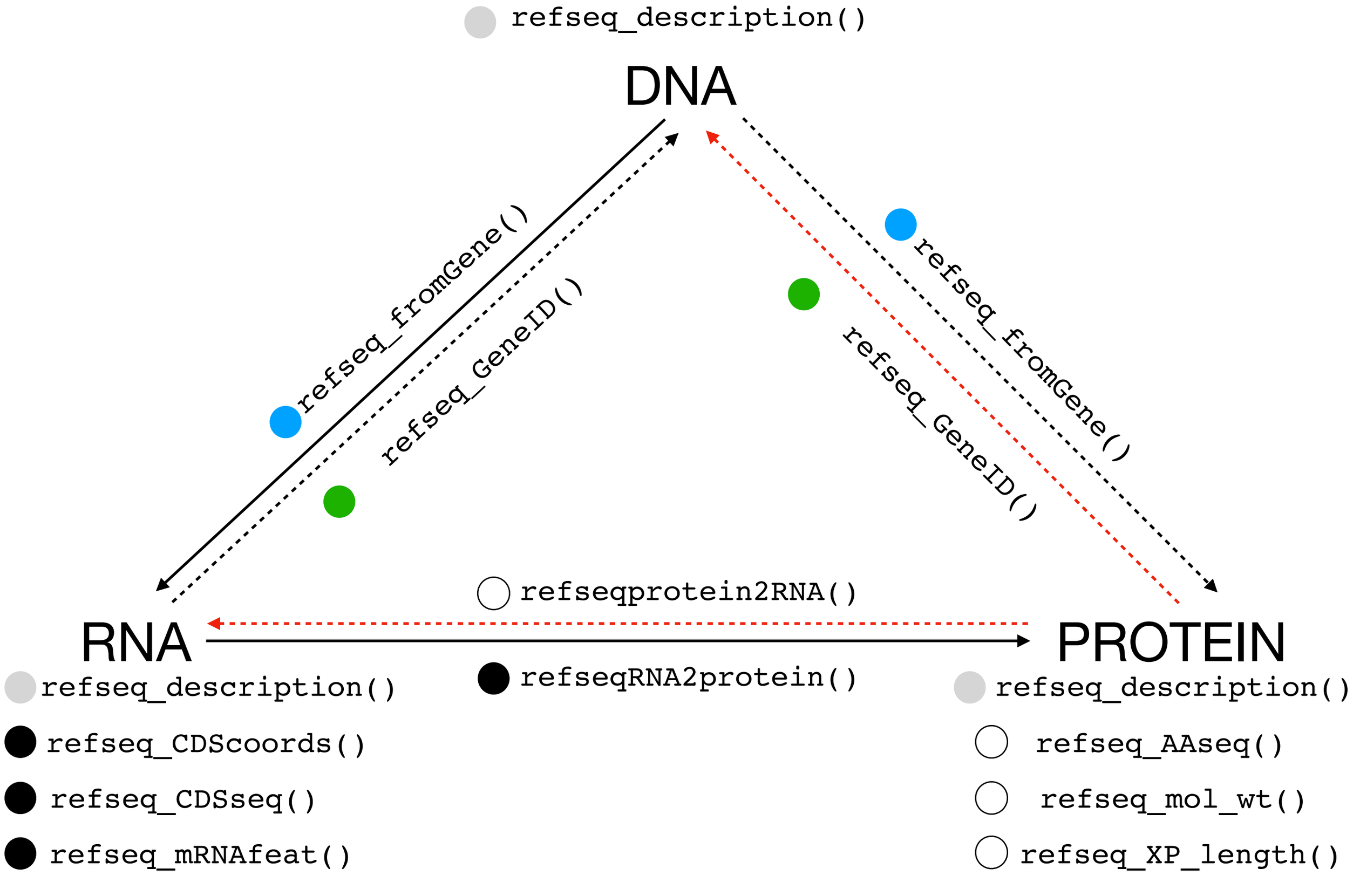


Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.

Our implementation: the ***refseqR*** package



Our implementation: the ***refseqR*** package

README License MIT license

devel version 1.1.4 | lifecycle stable | repo status Active

CRAN 1.1.5 | license MIT | cran checks ok

DOI 10.5281/zenodo.1189508

refseqR

Common computational operations working with RefSeq (GenBank accessions, NCBI).

Installation

Get the released version from CRAN:

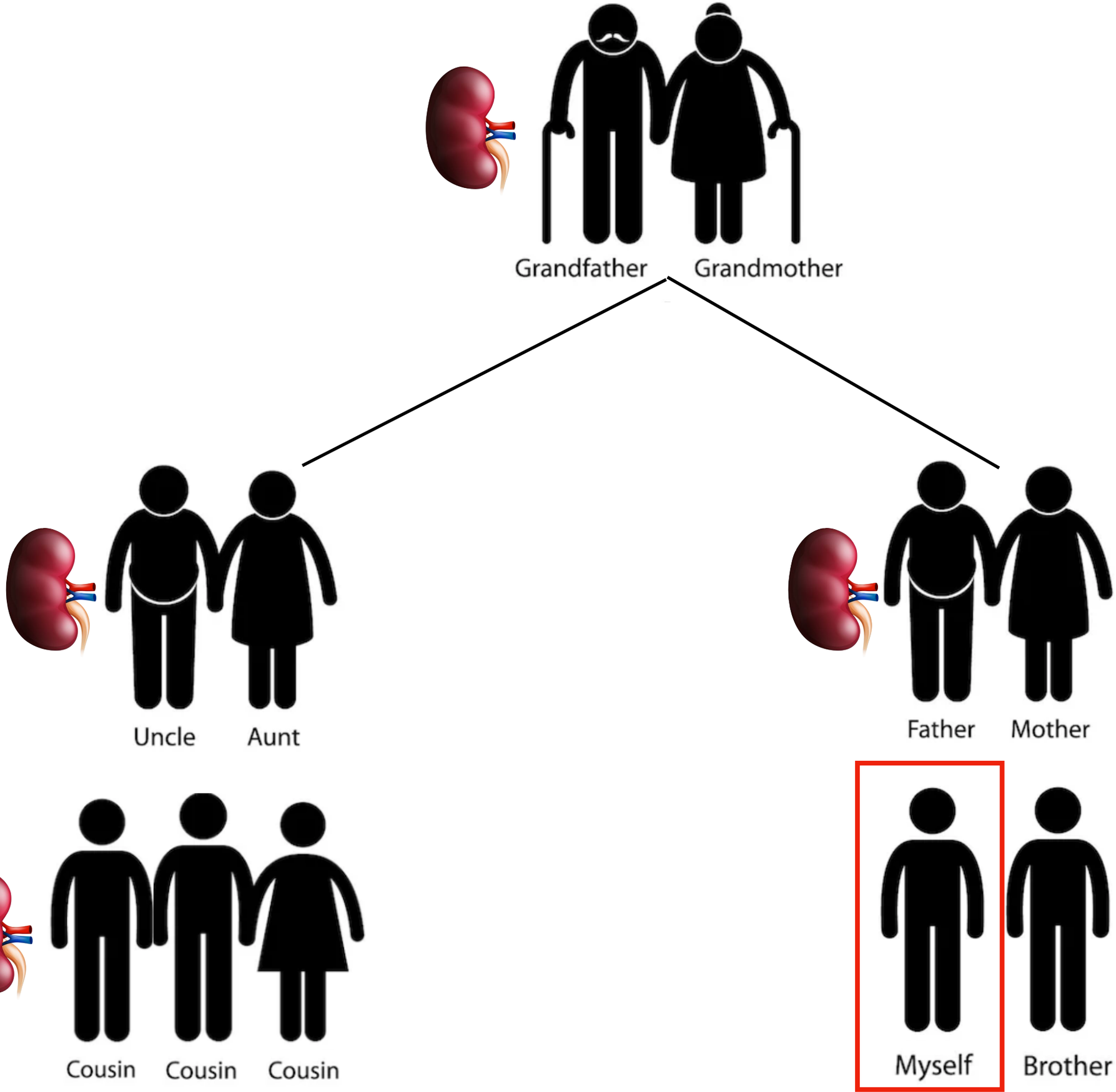
```
install.packages("refseqR")
```

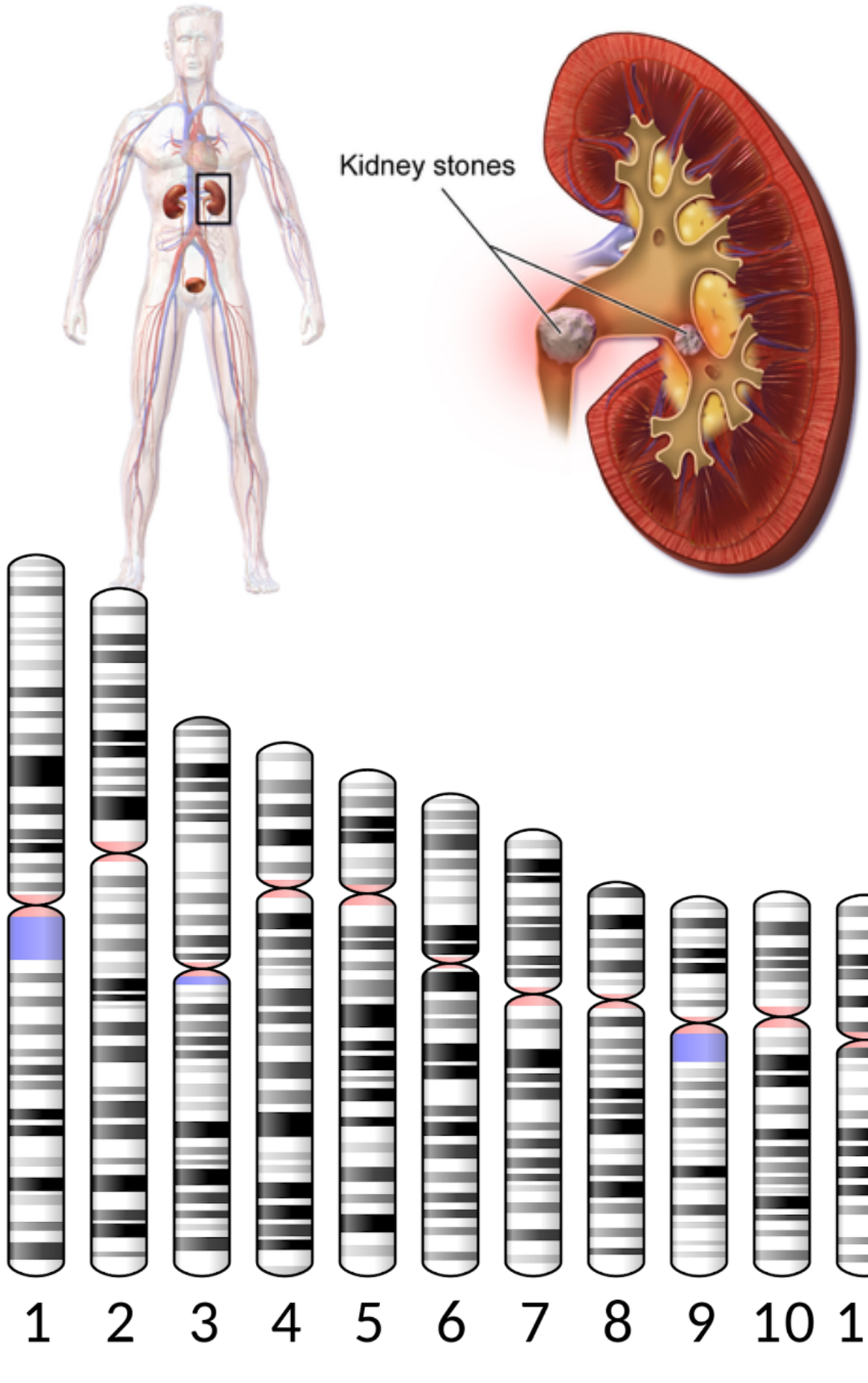
Or the development version from github:

```
devtools::install_github("jdieramon/refseqR")
```

- Builds upon ***rentrez***, ***IRanges***, ***Biostrings***
- Input can be any character string object (first argument)
- Provides output for interoperability and integration with Biocoductor objects

A hidden truth in my Family





Catéter Doble Jota @cateterdoblej · Jun 17, 2017

Cálculo renal. Imagen del Department of Chemical and Biomolecular Engineering, University of Houston, TX

3 35 50

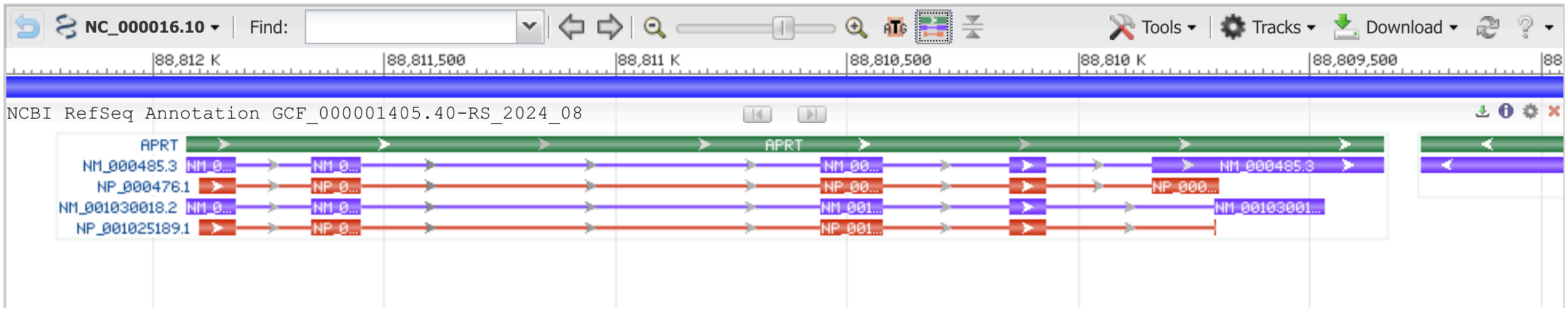
APRT adenine phosphoribosyltransferase [*Homo sapiens* (human)]

Download Datasets

Gene ID: 353, updated on 28-Oct-2024

Summary

Official Symbol	APRT provided by HGNC
Official Full Name	adenine phosphoribosyltransferase provided by HGNC
Primary source	HGNC:HGNC:626
See related	Ensembl:ENSG00000198931 MIM:102600; AllianceGenome:HGNC:626
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	AMP; APRTD
Summary	Adenine phosphoribosyltransferase belongs to the purine/pyrimidine phosphoribosyltransferase family. A conserved feature of this gene is the distribution of CpG dinucleotides. This enzyme catalyzes the formation of AMP and inorganic pyrophosphate from adenine and 5-phosphoribosyl-1-pyrophosphate (PRPP). It also produces adenine as a by-product of the polyamine biosynthesis pathway. A homozygous deficiency in this enzyme causes 2,8-dihydroxyadenine urolithiasis. Two transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]
Expression	Ubiquitous expression in colon (RPKM 38.8), appendix (RPKM 37.9) and 25 other tissues See more
Orthologs	mouse all



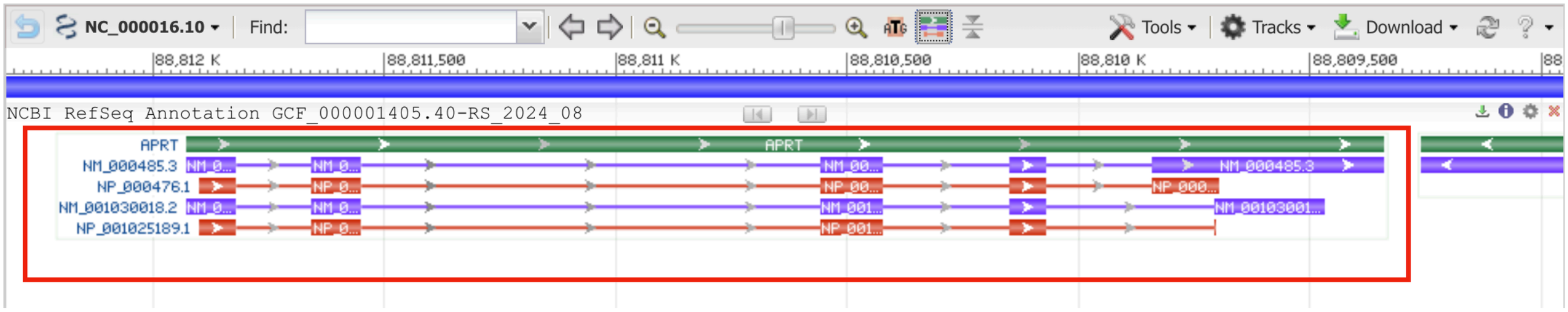
APRT adenine phosphoribosyltransferase [*Homo sapiens* (human)]

Download Datasets

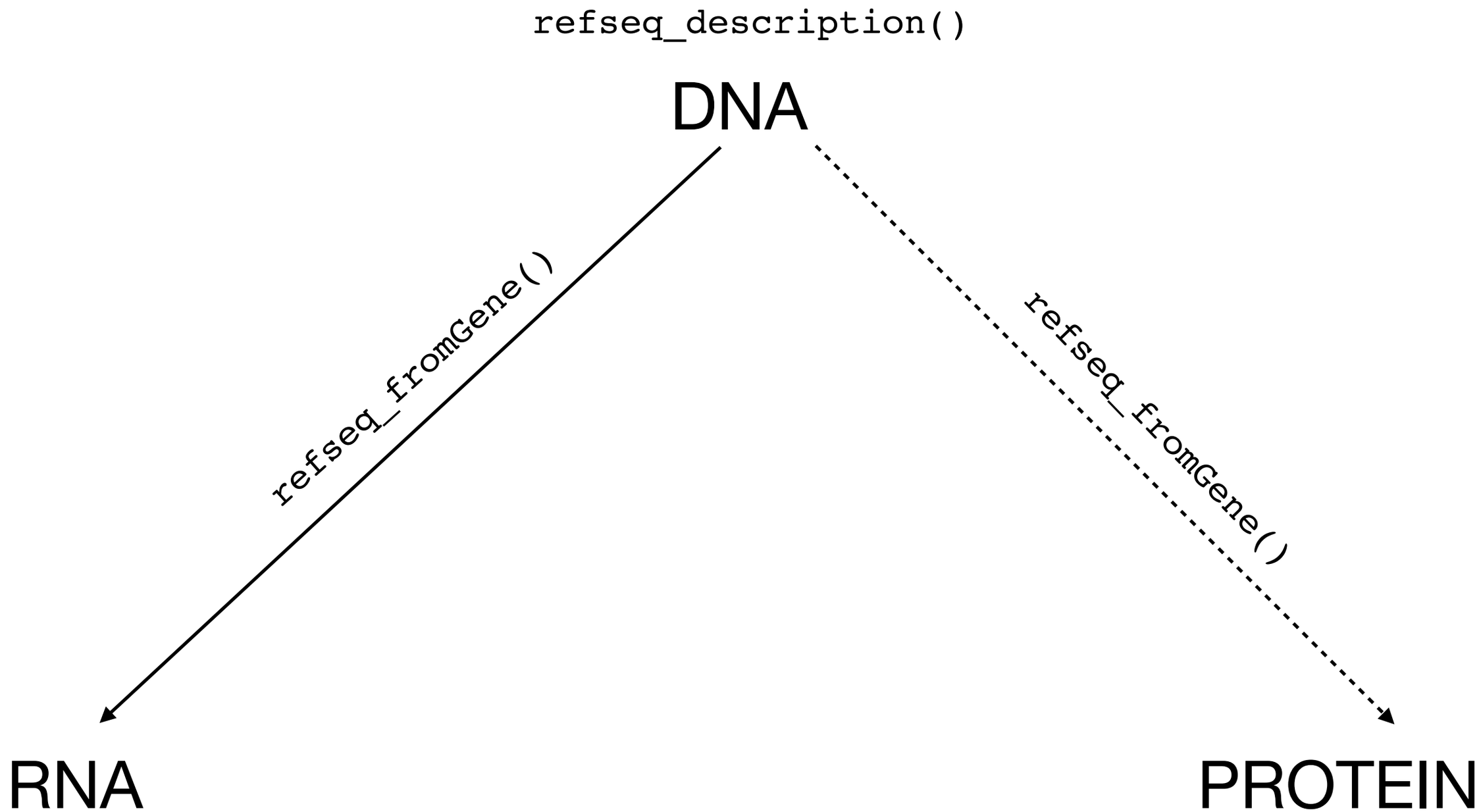
Gene ID: 353, updated on 28-Oct-2024

Summary

Official Symbol	APRT provided by HGNC
Official Full Name	adenine phosphoribosyltransferase provided by HGNC
Primary source	HGNC:HGNC:626
See related	Ensembl:ENSG00000198931 MIM:102600; AllianceGenome:HGNC:626
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	AMP; APRTD
Summary	Adenine phosphoribosyltransferase belongs to the purine/pyrimidine phosphoribosyltransferase family. A conserved feature of this gene is the distribution of CpG dinucleotides. This enzyme catalyzes the formation of AMP and inorganic pyrophosphate from adenine and 5-phosphoribosyl-1-pyrophosphate (PRPP). It also produces adenine as a by-product of the polyamine biosynthesis pathway. A homozygous deficiency in this enzyme causes 2,8-dihydroxyadenine urolithiasis. Two transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]
Expression	Ubiquitous expression in colon (RPKM 38.8), appendix (RPKM 37.9) and 25 other tissues See more
Orthologs	mouse all



GeneID accessions



GeneID accessions

`refseq_fromGene ()`

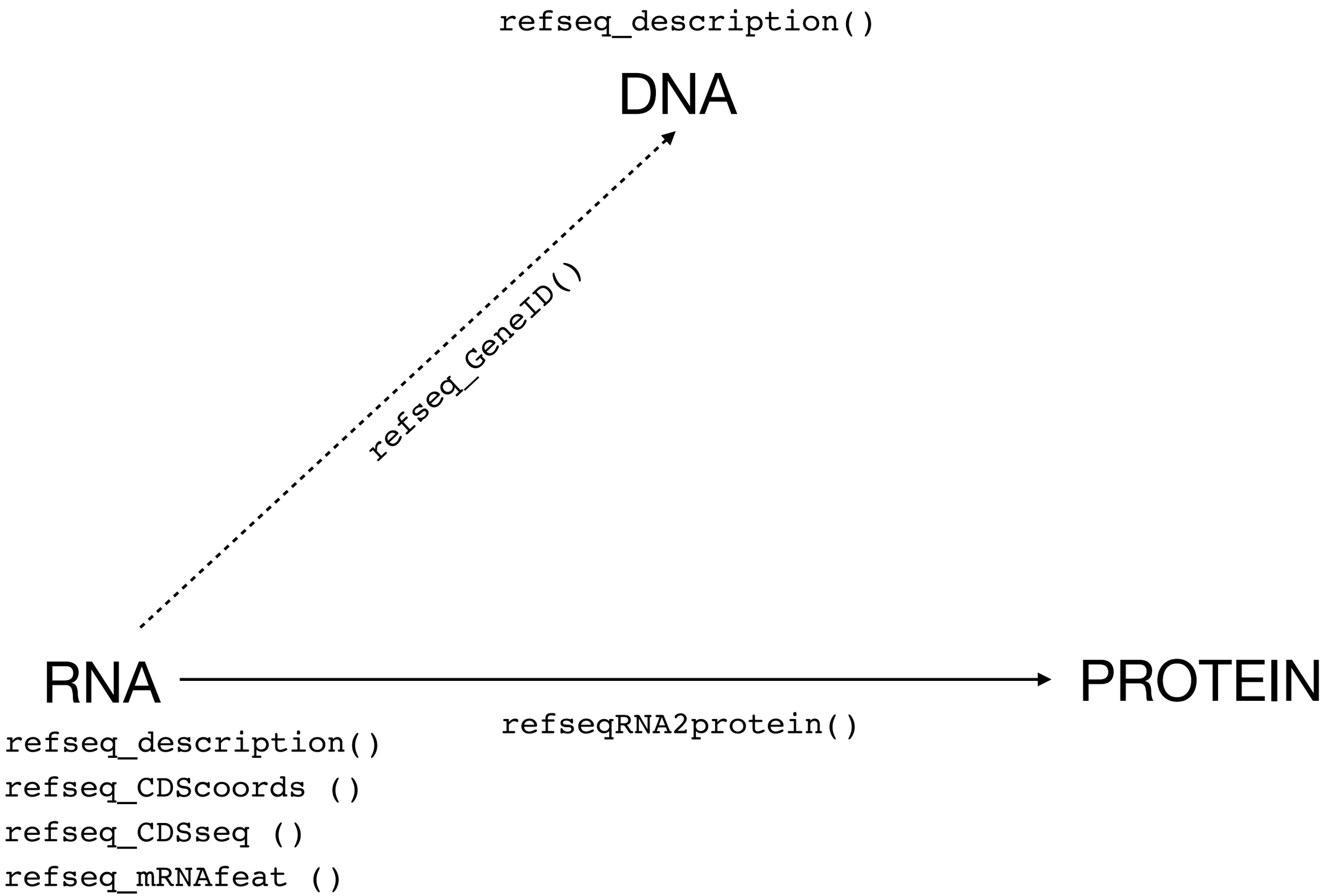
`refseq_description()`

```
> refseq_fromGene(GeneID = "353", sequence = "transcript")
[1] "NM_001030018" "NM_000485"

> refseq_fromGene(GeneID = "353", sequence = "protein")
[1] "NP_001025189" "NP_000476"
```

```
> refseq_description(id = "353")
[1] "adenine phosphoribosyltransferase"
```

mRNA accessions



mRNA accessions

```
> refseq_GeneID(accession = "NM_001030018", db = "nuccore")
[1] "353"

> refseq_description(id = "NM_001030018")
[1] "adenine phosphoribosyltransferase"

> refseq_CDSseq(transcript = "NM_001030018")
DNAStringSet object of length 1:
    width seq                                                 names
[1]   405 ATGGCCGACTCCGAGCTGCAGCTG...ATCTGCTGGCCACTGGTGTATGA NM_001030018.2
```

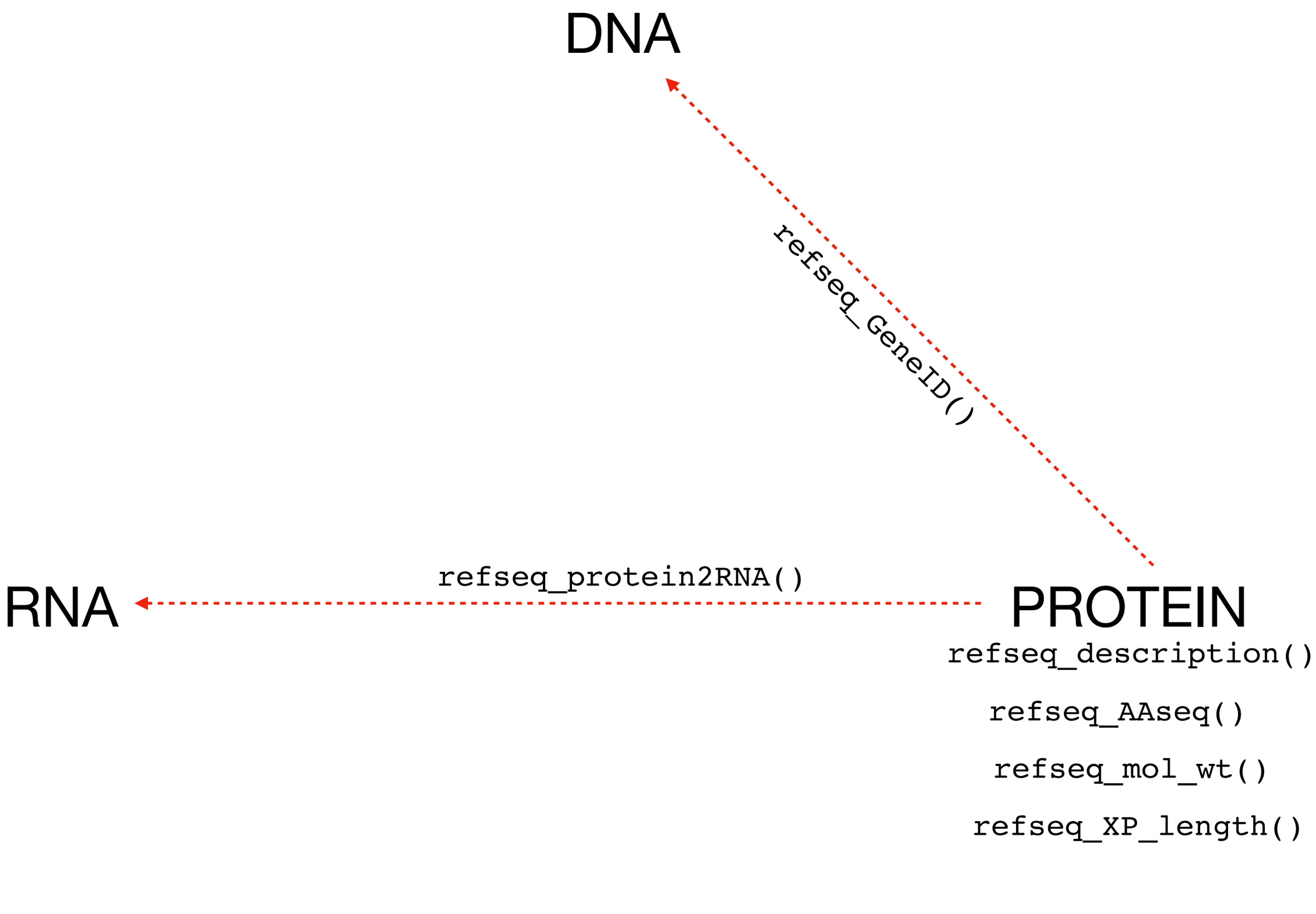
mRNA accessions

```
> refseq_CDScoords(transcript = "NM_001030018")
IRanges object with 1 range and 0 metadata columns:
                    start       end     width
                <integer> <integer> <integer>
  NM_001030018.2        30       434       405
```

```
> refseq_mRNAfeat(transcript ="NM_001030018",
+                 feat = c("caption", "moltype", "sourcedb", "slen"))
# A tibble: 1 × 4
  caption      moltype sourcedb slen
  <chr>        <chr>   <chr>    <chr>
1 NM_001030018 rna     refseq   667
```

```
> refseq_RNA2protein(transcript = "NM_001030018")
[1] "NP_001025189"
```

protein accessions



protein accessions

```
> refseq_description(id = "NP_001025189")
[1] "adenine phosphoribosyltransferase"


> refseq_AAlen(protein = "NP_001025189")
[1] 134
> refseq_AAmol_wt(protein = "NP_001025189")
[1] 14426


> refseq_AAseq(accession = "NP_001025189")
AAStringSet object of length 1:
    width seq                                               names
[1]   134 MADSELQLVEQRIRSFPDFP...ALEPGQRVVVVDDLLATGV NP_001025189


> refseq_protein2RNA(protein = "NP_001025189")
[1] "NM_001030018"


> refseq_GeneID(accession = "NP_001025189", db = "protein")
[1] "353"
```

Vectorization

```
> transcript = c("NM_001030018", "NM_001388492", "NM_000492")
> feat = c("caption", "moltype", "sourcedb", "slen", "title")
> refseq_mRNAfeat(transcript, feat)
# A tibble: 3 × 5
  caption      moltype sourcedb slen  title
  <chr>        <chr>   <chr>    <chr> <chr>
1 NM_001030018 rna     refseq   667   Homo sapiens adenine phosphoribosyltransferase (APRT), t…
2 NM_001388492 rna     refseq   13472 Homo sapiens huntingtin (HTT), transcript variant 1, mRNA
3 NM_000492    rna     refseq   6070  Homo sapiens CF transmembrane conductance regulator (CFT…
```

```
> # Get the protein ids from a set of transcript accessions
> transcript = c("NM_001030018", "NM_001388492", "NM_000492")
> sapply(transcript, function(x) refseq_RNA2protein(x), USE.NAMES = FALSE)
[1] "NP_001025189" "NP_001375421" "NP_000483"
```

Availability

refseqR: Common Computational Operations Working with RefSeq Entries (GenBank)

Fetches NCBI data (RefSeq <https://www.ncbi.nlm.nih.gov/refseq/> database) and provides an environment to extract information at the level of gene, mRNA or protein accessions.

Version:	1.1.5
Imports:	IRanges, rentrez, tibble, Biostrings
Suggests:	knitr, rmarkdown
Published:	2024-10-30
DOI:	10.32614/CRAN.package.refseqR
Author:	Jose V. Die [aut, cre], Lluís Revilla Sancho [ctb]
Maintainer:	Jose V. Die <jose.die at uco.es>
BugReports:	https://github.com/jdieramon/refseqR/issues
License:	MIT + file LICENSE
URL:	https://github.com/jdieramon/refseqR
NeedsCompilation:	no
CRAN checks:	refseqR results

Documentation:

Reference manual:	refseqR.pdf
Vignettes:	Working with the RefSeq database (source, R code)

Downloads:

Package source:	refseqR_1.1.5.tar.gz
Windows binaries:	r-devel: refseqR_1.1.5.zip, r-release: refseqR_1.1.5.zip, r-oldrel: refseqR_1.1.5.zip
macOS binaries:	r-release (arm64): refseqR_1.1.5.tgz, r-oldrel (arm64): refseqR_1.1.5.tgz, r-release (x86_64): refseqR_1.1.5.tgz, r-oldrel (x86_64): refseqR_1.1.5.tgz
Old sources:	refseqR archive

Linking:

Please use the canonical form `https://CRAN.R-project.org/package=refseqR` to link to this page.

Availability

Bioinformatics Advances, 2024, **00**, vbae122
https://doi.org/10.1093/bioadv/vbae122
Advance Access Publication Date: 21 August 2024
Application Note

Data and text mining

`refseqR`: an R package for common computational operations with records on RefSeq collection

Jose V. Die [1,*]

[1]Department of Genetics—ETSIAM, University of Cordoba, Córdoba, 14071, Spain

*Corresponding author. Department of Genetics—ETSIAM, University of Cordoba, Campus de Rabanales, Córdoba, 14071, Spain. E-mail: jose.die@uco.es

Associate Editor: Shanfeng Zhu

Abstract

Summary: We introduce `refseqR`, an R package that offers a user-friendly solution, enabling common computational operations on RefSeq entries (GenBank, NCBI). The package is specifically designed to interact with records curated from the RefSeq database. Most importantly, the interoperability and integration with several Bioconductor objects allow connections to be applied to other projects.

Availability and implementation: The package `refseqR` is implemented in R and published under the MIT open-source license. The source code, documentation, and usage instructions are available on CRAN (https://CRAN.R-project.org/package=refseqR).

External Dependencies

National Library of Medicine
National Center for Biotechnology Information

Log in

Search NCBI

Search NCBI

Search

Some portions of our website are experiencing intermittent outages, and E-utilities requests may take longer than normal to complete. We apologize for the inconvenience and thank you for your patience as we resolve the issue.

CLOSE

GenBank Release 263.0 Now Available!

GenBank release 263.0 (10/19/2024) is now available on the NCBI FTP site. This release has 36.50 trillion bases and 5.13 billion records.

The current release has:

- 251,998,350 traditional records containing 4,250,942,573,681 base pairs of sequence data
- 3,745,772,758 WGS records containing 31,362,454,467,668 base pairs of sequence data
- 948,733,596 bulk-oriented TSA records containing 812,661,461,811 base pairs of sequence data
- 187,349,395 bulk-oriented TLS records containing 77,037,504,468 base pairs of sequence data

Continue reading →

OCTOBER 24, 2024 GENBANK

External Dependencies

CRAN Package Check Results for Package refseqR

Last updated on 2024-11-02 16:49:22 CET.

Flavor	Version	$T_{install}$	T_{check}	T_{total}	Status	Flags
r-devel-linux-x86_64-debian-clang	1.1.5	11.91	123.91	135.82	OK	
r-devel-linux-x86_64-debian-gcc	1.1.5	10.73	91.35	102.08	OK	
r-devel-linux-x86_64-fedora-clang	1.1.5			234.10	OK	
r-devel-linux-x86_64-fedora-gcc	1.1.5			236.46	OK	
r-devel-windows-x86_64	1.1.5	19.00	180.00	199.00	OK	
r-patched-linux-x86_64	1.1.5	13.55	112.64	126.19	OK	
r-release-linux-x86_64	1.1.5	12.77	112.29	125.06	OK	
r-release-macos-arm64	1.1.5			118.00	OK	
r-release-macos-x86_64	1.1.5			130.00	OK	
r-release-windows-x86_64	1.1.5	14.00	179.00	193.00	OK	
r-oldrel-macos-arm64	1.1.5			111.00	OK	
r-oldrel-macos-x86_64	1.1.5			127.00	OK	
r-oldrel-windows-x86_64	1.1.5	17.00	182.00	199.00	OK	

Acknowledgements

Lluís Revilla Sancho

T-cell Immunology and Vaccines Group

IrsiCaixa

Sean Davis

School of Medicine

University of Colorado

AGR-114 Mejora Genética Vegetal

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Andalucía se mueve con Europa

Unión Europea
Fondo Social Europeo