GenomicFeatures.Hsapiens.UCSC.hg18

April 2, 2010

`GenomicFeatures.Hsapiens.UCSC.hg18_dbconn`
Get the connection to the built-in DB

Description

A convenience function for getting a connection object to the annotation DB included in the GenomicFeatures.Hsapiens.UCSC.hg18 package.

Usage

```
GenomicFeatures.Hsapiens.UCSC.hg18_dbconn()
GenomicFeatures.Hsapiens.UCSC.hg18_dbfile()
```

Details

`GenomicFeatures.Hsapiens.UCSC.hg18_dbconn` returns a connection object that was created at load-time and is aimed to hold a permanent connection. It is used internally by some of the functions defined in this package. Don't call `dbDisconnect` on this connection object or you will break these functions.

See Also

`dbGetQuery`, `dbConnect`, `geneHuman`

Examples

```
library(RSQLite)

## Get Human genes in chromosome 1:
chr1_genes <- dbGetQuery(GenomicFeatures.Hsapiens.UCSC.hg18_dbconn(),
                         "SELECT * FROM knownGene WHERE chrom='chr1'")
## Get all the Human genes:
genes <- dbReadTable(GenomicFeatures.Hsapiens.UCSC.hg18_dbconn(),
                     "knownGene", row.names=NULL)
## NOTE: The recommended way to get all the Human genes is to use geneHuman().
```

`geneHuman` *UCSC Gene Predictions for hg18*

Description

Gene coordinates and annotations for H. sapiens from UCSC. Coordinates are relative to the hg18 build and are in nucleotides from the 5' end of the positive (`"+"`) strand. They are always *one-based*, that is, the coordinate of the first (or leftmost) nucleotide in the strand is 1. Each “gene”, or row in the dataset, corresponds to a unique combination of transcript (TSS, TES and exons) and coding sequence (start and end).

Usage

```
geneHuman()
```

Value

A data frame with 66803 observations on the following 12 variables.

1. `name`: The name of the gene.
2. `chrom`: The name of the chromosome the gene is located on.
3. `strand`: The strand the gene is coded on, `"+"`, or `"-"`.
4. `txStart`: Transcription start site.
5. `txEnd`: Transcription stop site.
6. `cdsStart`: Start position of the coding sequence.
7. `cdsEnd`: End position of the coding sequence.
8. `exonCount`: The number of exons.
9. `exonStarts`: A comma separated list of the exon start positions.
10. `exonEnds`: A comma separated list of exon stop positions.
11. `proteinID`: An ID for the protein produced, missing values are coded as NA.
12. `alignID`: Unique identifier of each gene and RNA alignment pair, apparently redundant with `name`.

Note

For genes coded on the negative strand the `txStart` is really the end, and similarly for the coding regions.

Source

This table was obtained by downloading the following database file from UCSC (on Sep 28, 2009): `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownGene.txt.gz` and by translating the start coordinates found in the file from zero-based to one-based.

The knownGene.txt.gz file is a database dump containing the UCSC track called "UCSC Genes" and described here: `http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg18&g=knownGene`

The version of the "UCSC Genes" data set found in the database dump knownGene.txt.gz at the time of the download (Sep 28, 2009) is called "known genes 4" (or "kg4") by the UCSC people.

Hence this is also the version returned by `geneHuman`. The previous version of the data set ("kg3") is also provided thru the `geneHuman.old3` function.

See `http://genome.ucsc.edu/cgi-bin/hgTables` and Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes. Bioinformatics. 2006 May 1;22(9):1036-46.

See `https://lists.soe.ucsc.edu/pipermail/genome/2008-September/017101.html` and `https://lists.soe.ucsc.edu/pipermail/genome-announce/2008-September/000140.html` for the announce of the updated UCSC Genes data set ("known genes 4") and its difference with the previous version.

All the annotations in this package are freely available for public use, except for the Swiss-Prot/UniProt data in the knownGene table, which has the following terms of use:

```
UniProt copyright (c) 2002 - 2004 UniProt consortium

For non-commercial use all databases and documents in the UniProt FTP
directory may be copied and redistributed freely, without advance
permission, provided that this copyright statement is reproduced with
each copy.

For commercial use all databases and documents in the UniProt FTP
directory, except the files

ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/uniprot_sprot.dat.gz

and

ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/uniprot_sprot.xml.gz

may be copied and redistributed freely, without advance permission,
provided that this copyright statement is reproduced with each copy.

More information for commercial users can be found in:
http://www.expasy.org/announce/sp_98.html

From January 1, 2005, all databases and documents in the UniProt FTP
directory may be copied and redistributed freely by all entities,
without advance permission, provided that this copyright statement is
reproduced with each copy.
```

Examples

```
genes <- geneHuman()
str(genes)
transcripts(genes)
```

Index