GenomicFeatures

October 5, 2010

`CpG.mm9` *Locations of CpG islands*

Description

Locations of CpG islands in the mouse genome (build mm9).

Usage

```
data(CpG.mm9)
```

Format

A data frame with 15991 observations on the following 4 variables.

`chromosome` chromosome name as a character vector

`start` interval start points

`end` interval end points

`ID` an identifier

Source

The UCSC Genome Browser

`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. 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

```
data(geneHuman)
```

Format

A data frame with 56722 observations on the following 12 variables.

`name` The name of the gene.

`chrom` The name of the chromosome the gene is located on.

`strand` The strand the gene is coded on, `"+"`, or `"-"`.

`txStart` Transcription start site.

`txEnd` Transcription stop site.

`cdsStart` Start position of the coding sequence.

`cdsEnd` End position of the coding sequence.

`exonCount` The number of exons.

`exonStarts` A comma separated list of the exon start positions.

`exonEnds` A comma separated list of exon stop positions.

`proteinID` An ID for the protein produced, missing values are coded as NA.

`alignID` Unique identifier of each gene and RNA alignment pair, apparently redundant with `name`.

Details

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

Source

This table was taken directly from the knownGene table in the UCSC database for hg18, 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.

Examples

```
data(geneHuman)
str(geneHuman)
transcripts(geneHuman)
```

`geneMouse` *UCSC Gene Predictions for mm9*

Description

Gene coordinates and annotations for M. musculus from UCSC. Coordinates are relative to the mm9 build and are in nucleotides from the 5' end of the positive `"+"` strand. 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

```
data(geneMouse)
```

Format

A data frame with 49409 observations on the following 12 variables.

`name` The name of the gene.

`chrom` The name of the chromosome the gene is located on.

`strand` The strand the gene is coded on, `"+"`, or `"-"`.

`txStart` Transcription start site.

`txEnd` Transcription stop site.

`cdsStart` Start position of the coding sequence.

`cdsEnd` End position of the coding sequence.

`exonCount` The number of exons.

`exonStarts` A comma separated list of the exon start positions.

`exonEnds` A comma separated list of exon stop positions.

`proteinID` An ID for the protein produced, missing values are coded as NA.

`alignID` Unique identifier of each gene and RNA alignment pair, apparently redundant with `name`.

Details

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

Source

This table was taken directly from the knownGene table in the UCSC database for mm9, 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.

Examples

```
data(geneMouse)
str(geneMouse)
transcripts(geneMouse)
```

`isochores.mm8` *Isochore boundaries for Mus musculus (build mm9).*

Description

Isochore boundaries for Mus musculus (build mm9). Isochores are large segments of the genome such that within-segment variability in GC content is substantially lower than between-segment variability. These isochores are computationally predicted by IsoFinder (see below).

Usage

```
data(isochores.mm8)
```

Format

A data frame with 32894 observations on the following 4 variables.

`Begin` isochore starts.

`End` isochore ends.

`GC` GC content in isochore.

`chromosome` chromosome identifier.

Source

`http://bioinfo2.ugr.es/isochores/`

`regions` *Functions that compute genomic regions of interest.*

Description

Functions that compute genomic regions of interest such as promotor, upstream regions etc, from the genomic locations provided in data like `geneMouse`.

Usage

```
transcripts(genes, proximal = 500, distal = 10000)
exons(genes)
introns(genes)
```

Arguments

`genes`	A data.frame like that provided by `geneMouse`.
`proximal`	The number of bases on either side of TSS and 3'-end for the promoter and end region, respectively.
`distal`	The number of bases on either side for upstream/downstream, i.e. enhancer/silencer regions.

Details

The assumption made for introns is that there must be more than one exon, and that the introns are between the end of one exon and before the start of the next exon.

Value

All of these functions return a `RangedData` object with a `gene` column with the UCSC ID of the gene. For `transcripts`, each element corresponds to a transcript, and there are columns for each type of region (promoter, threeprime, upstream, and downstream). For `exons`, each element corresponds to an exon. For `introns`, each element corresponds to an intron.

Author(s)

M. Lawrence.

Examples

```
data(geneHuman)
## promoter 300bp up and down from TSS (threeprime from TES)
transcripts(geneHuman, proximal = 300)
```

Index