SNPlocs.Hsapiens.dbSNP141.GRCh38

January 5, 2021

`getSNPlocs` *Accessing the SNPs stored in SNPlocs.Hsapiens.dbSNP141.GRCh38*

Description

Functions for accessing the SNPs stored in the SNPlocs.Hsapiens.dbSNP141.GRCh38 package.

WARNING: All the functions described in this man page are deprecated and will be removed at some point in the future. See ?`snplocs` in the **BSgenome** software package for the new preferred way to access the data stored in this package.

Usage

```
## Count and load all the SNPs for a given chromosome:
getSNPcount()
getSNPlocs(seqname, as.GRanges=FALSE, caching=TRUE)

## Extract SNP information for a set of rs ids:
rsid2loc(rsids, caching=TRUE)
rsid2alleles(rsids, caching=TRUE)
rsidsToGRanges(rsids, caching=TRUE)
```

Arguments

`seqname`	The name of the sequence for which to get the SNP locations and alleles. If `as.GRanges` is `FALSE`, only one sequence can be specified (i.e. `seqname` must be a single string). If `as.GRanges` is `TRUE`, an arbitrary number of sequences can be specified (i.e. `seqname` can be a character vector of arbitrary length).
`as.GRanges`	`TRUE` or `FALSE`. If `TRUE`, then the SNP locations and alleles are returned in a GRanges object. Otherwise (the default), they are returned in a data frame (see below).
`caching`	Should the loaded SNPs be cached in memory for faster further retrieval but at the cost of increased memory usage?
`rsids`	A vector of rs ids. Can be integer or character vector, with or without the `"rs"` prefix. NAs are not allowed.

Details

See SNPlocs.Hsapiens.dbSNP141.GRCh38 for general information about this package.

The SNP data are split by chromosome (1-22, X, Y, MT) i.e. the package contains one data set per chromosome, each of them being a serialized data frame with 1 row per SNP and the 2 following columns:

- `loc`: The 1-based location of the SNP relative to the first base at the 5' end of the plus strand of the reference sequence.
- `alleles`: A raw vector with no NAs which can be converted into a character vector containing the alleles for each SNP represented by an IUPAC nucleotide ambiguity code (see `?IUPAC_CODE_MAP` in the Biostrings package for more information).

Note that those data sets are not intended to be used directly but the user should instead use the `getSNPcount` and `getSNPlocs` convenience wrappers for loading the SNP data. When used with `as.GRanges=FALSE` (the default), `getSNPlocs` returns a data frame with 1 row per SNP and the 3 following columns:

- `RefSNP_id`: RefSNP ID (aka "rs id") with `"rs"` prefix removed. Character vector with no NAs and no duplicates.
- `alleles_as_ambig`: A character vector with no NAs containing the alleles for each SNP represented by an IUPAC nucleotide ambiguity code.
- `loc`: Same as for the 2-col serialized data frame described previously.

Value

`getSNPcount` returns a named integer vector containing the number of SNPs for each sequence in the reference genome.

By default (`as.GRanges=FALSE`), `getSNPlocs` returns the 3-col data frame described above containing the SNP data for the specified chromosome. Otherwise (`as.GRanges=TRUE`), it returns a GRanges object with extra columns `"RefSNP_id"` and `"alleles_as_ambig"`. Note that all the elements (genomic ranges) in this GRanges object have their strand set to `"+"` and that all the sequence lengths are set to `NA`.

`rsid2loc` and `rsid2alleles` both return a named vector (integer vector for the former, character vector for the latter) where each (name, value) pair corresponds to a supplied rs id. For both functions the name in (name, value) is the chromosome of the rs id. The value in (name, value) is the position of the rs id on the chromosome for `rsid2loc`, and a single IUPAC code representing the associated alleles for `rsid2alleles`.

`rsidsToGRanges` returns a GRanges object similar to the one returned by `getSNPlocs` (when used with `as.GRanges=TRUE`) and where each element corresponds to a supplied rs id.

Author(s)

H. Pages

See Also

- `snplocs` in the **BSgenome** software package for the new preferred way to access the data stored in this package.
- SNPlocs.Hsapiens.dbSNP141.GRCh38

SNPlocs.Hsapiens.dbSNP141.GRCh38 *The SNPlocs.Hsapiens.dbSNP141.GRCh38 package*

Description

This package contains SNP locations and alleles for Homo sapiens extracted from dbSNP Build 141.

Details

SNPs from dbSNP were filtered to keep only those satisfying the 3 following criteria:

- The SNP is a single-base substitution i.e. its type is "snp". Other types used by dbSNP are: "in-del", "mixed", "microsatellite", "named-locus", "multinucleotide-polymorphism", etc... All those SNPs were dropped.
- The SNP is marked as notwithdrawn.
- A *single* location on the reference genome (GRCh38) is reported for the SNP, and this location is on chromosomes 1-22, X, Y, or MT.

SNPlocs packages always store the alleles corresponding to the *plus* strand, whatever the strand reported by dbSNP is (which is achieved by storing the complement of the alleles reported by dbSNP for SNPs located on the minus strand). In other words, in a SNPlocs package, all the SNPs are considered to be on the plus strand and everything is reported with respect to that strand.

Note

The source data files used for this package were created by the dbSNP Development Team at NCBI on May 1st, 2014.

The SNPs in this package can be "injected" in BSgenome.Hsapiens.NCBI.GRCh38 or BSgenome.Hsapiens.UCSC.hg38 and will land at the correct location.

See http://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/ for more information about the GRCh38 assembly.

See http://www.ncbi.nlm.nih.gov/snp, the SNP Home at NCBI, for more information about dbSNP.

See ?injectSNPs in the **BSgenome** software package for more information about the SNP injection mechanism.

See http://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38 for the UCSC Genome Browser based on the hg38 assembly. Note hg38 and GRCh38 are the same assemblies (i.e. the 455 genomic sequences in both of them are the same), except that they use different conventions to name the sequences (i.e. for the chromosome and scaffold names).

Author(s)

H. Pages

References

Genome Reference Consortium at NCBI: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

SNP Home at NCBI: http://www.ncbi.nlm.nih.gov/snp

dbSNP Human BUILD 141 announcement: http://www.ncbi.nlm.nih.gov/mailman/pipermail/dbsnp-announce/2014q2/000139.html

UCSC Genome Browser based on the hg38 assembly: http://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38

See Also

- snplocs in the **BSgenome** software package for how to access the data stored in this package.
- IUPAC_CODE_MAP in the **Biostrings** package.
- The GRanges class in the **GenomicRanges** package.
- injectSNPs in the **BSgenome** software package for SNP injection.
- The **VariantAnnotation** software package to annotate variants with respect to location and amino acid coding.

Examples

```
## ---------------------------------------------------------------------
## A. BASIC USAGE
## ---------------------------------------------------------------------
snps <- SNPlocs.Hsapiens.dbSNP141.GRCh38
snpcount(snps)

## Get the locations and alleles of all SNPs on chromosome 22:
ch22snps <- snplocs(snps, "ch22")
dim(ch22snps)
colnames(ch22snps)
head(ch22snps)

## Get the locations and alleles of all SNPs on chromosomes 22 and MT
## as a GRanges object:
snplocs(snps, c("ch22", "chMT"), as.GRanges=TRUE)

## ---------------------------------------------------------------------
## B. EXTRACT SNP INFORMATION FOR A SET OF RS IDS...
## ---------------------------------------------------------------------

## ... and return it in a GRanges object:
my_rsids <- c("rs2639606", "rs75264089", "rs73396229", "rs55871206",
              "rs10932221", "rs56219727", "rs73709730", "rs55838886",
              "rs3734153", "rs79381275", "rs1516535")
gr <- snpid2grange(snps, my_rsids)

## Translate the IUPAC ambiguity codes used to represent the alleles
## into nucleotides:
IUPAC_CODE_MAP[mcols(gr)$alleles_as_ambig]

## ---------------------------------------------------------------------
## C. INJECTION IN THE REFERENCE GENOME
## ---------------------------------------------------------------------
```

```
library(BSgenome.Hsapiens.NCBI.GRCh38)
genome <- BSgenome.Hsapiens.NCBI.GRCh38
genome

genome2 <- injectSNPs(genome, "SNPlocs.Hsapiens.dbSNP141.GRCh38")
genome2  # note the additional line "with SNPs injected from..."

alphabetFrequency(genome[["22"]])
alphabetFrequency(genome2[["22"]])

## Get the number of nucleotides that were modified by this injection:
neditAt(genome[["22"]], genome2[["22"]])  # 737750

## ---------------------------------------------------------------------
## D. SOME BASIC QUALITY CONTROL (WITH SURPRISING RESULTS!)
## ---------------------------------------------------------------------

## Note that dbSNP can assign distinct ids to SNPs located at the same
## position:
any(duplicated(ch22snps$RefSNP_id))  # rs ids are all distinct...
any(duplicated(ch22snps$loc))  # but some locations are repeated!

ch22snps <- ch22snps[order(ch22snps$loc), ]  # sort by location
which(duplicated(ch22snps$loc))[1:2]  # 106, 558
ch22snps[103:108, ]  # rs9617549 and rs199758506 share the same location
                     # (10874444) and alleles (Y, i.e. C/T)

## Also note that not all SNP alleles are consistent with the GRCh38
## genomic sequences i.e. the alleles reported for a given SNP are not
## always compatible with the nucleotide found at the SNP location in
## GRCh38. For example, to get the number of inconsistent SNPs in chr1:
ch1snps <- snplocs(snps, "ch1")
all_alleles <- DNAString(paste(ch1snps$alleles_as_ambig, collapse=""))
nchar(all_alleles)  # 4160510 SNPs on chr1
neditAt(genome[["1"]][ch1snps$loc], all_alleles, fixed=FALSE)
## ==> 57641 SNPs (1.385%) are inconsistent with GRCh38 chr1!
```

Index