Package 'ind1KG'

September 23, 2012

Title Data from 1000 Genomes, NA19240 (female) chr6 excerpt

Version 0.1.13

Author VJ Carey <stvjc@channing.harvard.edu>

Maintainer VJ Carey <stvjc@channing.harvard.edu>

Description Elements of samtools/bioc workflow for dealing with personal sequence focusing on identification and interpretation of rare variants

Depends R (>= 2.10.1), chopsticks

Suggests Rsamtools, rtracklayer, GenomicFeatures, org.Hs.eg.db,SNPlocs.Hsapiens.dbSNP.20090506,TxDb.Hsapiens.UCSC.hg18.knownGene, lumiHumanIDMapping, GGBase,GGtools, hmyriB36

License Artistic-2.0

biocViews Genetics, HighThroughputSequencing

R topics documented:

c6snp *SNP metadata used in calling for 1000 genomes pilot data*

Description

SNP metadata used in calling for 1000 genomes pilot data – restricted to chr6

Usage

```
data(c6snp)
```

Format

A data frame with 1143009 observations on the following 16 variables.

`chr` a numeric vector

`chrPosFrom` a numeric vector

`chrPosTo` a numeric vector

`rs` a factor with levels `rs1000` `rs1000009` `rs1000025` `rs10000302` ...

`ChrAllele` a factor with levels `-` `A` `AA` `AAA` `AAAA` `AAAAA` `AAAAAA` `AAAAAAA` ...

`variantAllele` a factor with levels `(CA)11/12/13/14/-` `(G)14/15/16/18/19/20/21/22/23/C` `(G)20/21/22/23/24/25/27/-/G/GGG` `(LAREDELETION)/-/A` ...

`snpAlleleChrOrien` a factor with levels `(A)1/13/15/G` `(A)10/12` `(A)10/14` `(C)10/11` `(CA)10/11/13/14/15` `(CA)10/14/15/16/17/18/20/21` `(CA)10/17/18/19/20/21/22/23/24/25` `(CA)11/12/13/14/-/CACA` `(CA)11/12/13/14/15/16/17` ...

`snp2chrOrien` a numeric vector

`snpClassAbbrev` a factor with levels `Microsatellite` `Named snp dips` `mixed` `multi-base` `single base`

`snpClassCode` a numeric vector

`mapLocType` a numeric vector

`mapLocCnt` a numeric vector

`mapWeight` a numeric vector

`contigLabel` a factor with levels `DR53` `c6_COX` `c6_QBL` `reference`

`unPlacedContig` a factor with levels `NT_113898.1` `NT_113899.1`

Details

Column headings: ========== chr,chrPosFrom, chrPosTo, rs,ChrAllele,variantAllele,snpAlleleChrOrien, snp2chrOrien, snpClassAbbrev,snpClassCode,mapLocType,mapLocCnt,mapWeight, contigLabel,unPlacedContig

Column description: ========== Col1: Chr Col2: chrPosFrom: all chromosome positions are 1 based, that is the first base is counted as 1. Position is for each base, not "interbase". Col3: chrPosTo: Col4: rs Col5: ChrAllele: the base or bases on the chromosomes at the snp position or ranges. Col6: variantAllele: This is the other allele that is not on the chromosome. For ex. Snp is A/C, chromosome has A, variantAllele will be "C". Col7: snpAlleleChrOrien: This is the list of alleles for the snp in the chromosome orientation. Col8: : the alignment orientation between snp flank and the chromosome sequence. #orien: 0 - same; 1 - opposite Col9: snpClassAbbrev: the variation type of the snp. Details at: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpsnpfaq&part=Reports#Reports.Var

Col10: snpClassCode: The numeric code for snp variation class. possible snpClassCode, SnpClassAbbrev and desciptions are below:

1 single base Only single base variation.ex.A/G. 2 dips indel or dips: deletion insertion polymorphism.ex.-/T.ex.ss149071 obs=AA/GCCTG 3 HETEROZYGOUS HETEROZYGOUS 4 Microsatellite Microsatellite 5 Named snp observed field starts with '(', and not class 3 an 4.ex.(Alu) 6 NOVARIATION NOVARIATION 7 mixed If the subsnp's in an rs cluster have different snp class. 8 multi-base Multiple Nucleotide Polymorphism, where all alleles are same length, and length > 1.ex.ss2421179 AT/GA

Col11: mapLocType: The alignment type at snp site. Possible values and meanings are: 1 Insertion on contig: snp is always represented as one base and this one base in the snp sequence is substituted with more than one bases on the contig sequence in the alignment. 2 Contig allele is one base long.snp is always represented as one base and this one base in the snp sequence is substituted with exactly one base on the contig. 3 Deletion on the contig: part of the snp flanking sequence including the snp was absent on the contig sequence in the alignment. 4 In the alignment, part of the snp flanking sequence including snp is replaced with the contig sequence of longer length. 5 In the alignment, part of the snp flanking sequence including snp is replaced with the contig sequence of exactly the same length. 6 In the alignment, part of the snp flanking sequence including snp is replaced with the contig sequence of a shorter length.

Col12: mapLocCnt: the total number of locations the snp maps to within the assembly. Col13: Mapweight: A number that codes for the mapping quality of the snp on each assembly: 1 = snp aligns at exactly one location 2 = snp aligns at two locus on same chromosome 3 = snp aligns at two locus on different chromosomes or more than 3 and less than 10 locations 10= snp aligns at 10 or more 10 locations

Col14: ContigLabel: This is used to show when a snp maps to alternative haplotypes or PAR region. Possible values are: DR53 PAR c22_H2 c5_H2 c6_COX c6_QBL mitochondrial genome reference

col15: unPlacedContig: This field only has value when a snp hits an unplaced contig, there is no chromosome positions for the snp. chrPosFrom and chrPosTo will be NULL. In this case, unPlacedContig will have the actually contig accession that is unplaced on a chromosome.

Source

1000 genomes pilot data

Examples

```
data(c6snp)
c6snp[1:3,]
```

`gw6c6.snp240` *data frame with snp annotation regarding chr6, pertinent to the affy genomewide6 SNP chip*

Description

data frame with snp annotation regarding chr6, pertinent to the affy genomewide6 SNP chip

Usage

```
data(gw6c6.snp240)
```

Format

A data frame with 56271 observations on the following 7 variables.

`man_fsetid` a character vector

`dbsnp_rs_id` a character vector

`physical_pos` a numeric vector

`strand` a numeric vector

`allele_a` a character vector

`allele_b` a character vector

`call240` a numeric vector

Details

an extract from the pd.genomewidesnp.6 metadata

Examples

```
data(gw6c6.snp240)
str(gw6c6.snp240)
```

`n240` *NA19240 – 3 million reads from chromosome 6 derived from 1000 genomes bam file*

Description

NA19240 – 3 million reads from chromosome 6 derived from 1000 genomes bam file

Usage

```
data(n240)
```

Format

The format is:
List of 1
$:List of 13
..$ qname : chr [1:3000000] "EAS254_13:7:88:1639:15041" "EAS139_43:2:31:1128:9551" "EAS254_13:8:68:520:6861
"BGI-FC20AHFAAXX_6_26_477:352" ...
..$ flag : int [1:3000000] 35 35 35 16 35 35 35 0 35 35 ...
..$ rname : Factor w/ 1 level "6": 1 1 1 1 1 1 1 1 1 1 ...
..$ strand: Factor w/ 3 levels "-","+","*": 2 2 2 1 2 2 2 2 2 2 ...
..$ pos : int [1:3000000] 5001 5002 5004 5004 5005 5010 5012 5018 5018 5023 ...
..$ qwidth: int [1:3000000] 51 51 51 36 51 51 36 36 36 45 ...
..$ mapq : int [1:3000000] 0 0 0 0 0 0 0 0 0 0 ...
..$ cigar : chr [1:3000000] "51M" "51M" "51M" "36M" ...
..$ mrnm : Factor w/ 1 level "6": 1 1 1 NA 1 1 1 NA 1 1 ...
..$ mpos : int [1:3000000] 5163 5203 5170 NA 5156 5183 5199 NA 5252 5273 ...
..$ isize : int [1:3000000] 214 253 218 NA 203 225 224 NA 271 296 ...
..$ seq :Formal class 'DNAStringSet' [package "Biostrings"] with 5 slots

.. .. ..@ super :Formal class 'DNAString' [package "Biostrings"] with 6 slots
.. ..@ xdata :Formal class 'RawPtr' [package "IRanges"] with 2 slots
.. ..@ xp :<externalptr>
.. ..@ .link_to_cached_object:<environment: 0xc5354b8>
.. ..@ offset : int 0
.. ..@ length : int 118321039
.. ..@ elementMetadata: NULL
.. ..@ elementType : chr "ANYTHING"
.. ..@ metadata : list()
.. .. ..@ ranges :Formal class 'IRanges' [package "IRanges"] with 6 slots
.. ..@ start : int [1:3000000] 1 52 103 154 190 241 292 328 364 400 ...
.. ..@ width : int [1:3000000] 51 51 51 36 51 51 36 36 36 45 ...
.. ..@ NAMES : NULL
.. ..@ elementMetadata: NULL
.. ..@ elementType : chr "integer"
.. ..@ metadata : list()
.. .. ..@ elementMetadata: NULL
.. .. ..@ elementType : chr "ANYTHING"
.. .. ..@ metadata : list()
..$ qual :Formal class 'PhredQuality' [package "Biostrings"] with 5 slots
.. .. ..@ super :Formal class 'BString' [package "Biostrings"] with 6 slots
.. ..@ xdata :Formal class 'RawPtr' [package "IRanges"] with 2 slots
.. ..@ xp :<externalptr>
.. ..@ .link_to_cached_object:<environment: 0xc5354b8>
.. ..@ offset : int 0
.. ..@ length : int 118321039
.. ..@ elementMetadata: NULL
.. ..@ elementType : chr "ANYTHING"
.. ..@ metadata : list()
.. .. ..@ ranges :Formal class 'IRanges' [package "IRanges"] with 6 slots
.. ..@ start : int [1:3000000] 1 52 103 154 190 241 292 328 364 400 ...
.. ..@ width : int [1:3000000] 51 51 51 36 51 51 36 36 36 45 ...
.. ..@ NAMES : NULL
.. ..@ elementMetadata: NULL
.. ..@ elementType : chr "integer"
.. ..@ metadata : list()
.. .. ..@ elementMetadata: NULL
.. .. ..@ elementType : chr "ANYTHING"
.. .. ..@ metadata : list()

Details

result of Rsamtools scanBam applied to a 3mm line excerpt from the SLX chr6 aligned reads for NA19240

update october 2010 to reflect modifications to cigar handling in Rsamtools

Source

1000 genomes pilot data

Examples

```
data(n240)
sapply(n240[[1]], class)
```

`oregdf` *data frame representing some of the OREGanno track of UCSC browser*

Description

data frame representing some of the OREGanno track of UCSC browser

Usage

```
data(oregdf)
```

Format

A data frame with 86 observations on the following 7 variables.

`space` a factor with levels `chr6`

`start` a numeric vector

`end` a numeric vector

`width` a numeric vector

`name` a factor with levels `OREG0004577` `OREG0004578` ...

`score` a numeric vector

`strand` a factor with levels `+`

Examples

```
data(oregdf)
oregdf[1:5,]
```

`pup240_500k` *data frame on a pileup for 500000 variants noted in NA19240*

Description

data frame on a pileup for 500000 variants noted in NA19240

Usage

```
data(pup240_500k)
```

Format

A data frame with 500000 observations on the following 5 variables.

`V2` a numeric vector

`V3` a character vector

`V4` a character vector

`V5` a numeric vector

`V9` a character vector

Examples

```
data(pup240_500k)
pup240_500k[1:5,]
```

---

pup240_disc *parsed samtools pileup result excerpt*

Description

parsed samtools pileup result excerpt – only locations where the called base differs from the reference

Usage

```
data(pup240_disc)
```

Format

A data frame with 31243 observations on the following 5 variables.

`loc` location

`ref` reference base

`indiv` call for NA19240

`depth` coverage depth at this base

`pileup` pileup

Details

asterisks are to denote deletions...

Examples

```
data(pup240_disc)
pup240_disc[1:4,]
```

rsgdf *slice of information on transcripts, for ind1KG vignette*

Description

slice of information on transcripts, for ind1KG vignette

Usage

```
data(rsgdf)
```

Format

A data frame with 73 observations on the following 13 variables.

`space` a factor with levels `chr6`

`start` a numeric vector

`end` a numeric vector

`width` a numeric vector

`name` a factor with levels `NM_000129` `NM_000904` `NM_001003698` `NM_001003699` ...

`score` a numeric vector

`strand` a factor with levels `+` `-`

`thickStart` a numeric vector

`thickEnd` a numeric vector

`color` a numeric vector

`blockCount` a numeric vector

`blockSizes` a factor with levels `1013,192,167,230,156,80,739,` `1013,192,242,230,156,80,739,` `1026,` `106,119,123,213,90,164,145,137,190,2911,165,798,2949,` ...

`blockStarts` a factor with levels `0,` `0,10224,16062,19437,21152,22478,25840,27681,28397,30622,31385,3498` `0,107199,143191,169690,283829,351818,509940,,` ...

Examples

```
data(rsgdf)
rsgdf[1:4,]
```

yri240_6 *snp.matrix representation of called SNP on chr6 for NA19240*

Description

snp.matrix representation of called SNP on chr6 for NA19240

Usage

```
data(yri240_6)
```

Format

The format is: List of 2 $ hm2 :Formal class 'snp.matrix' [package "snpMatrix"] with 1 slots .. ..@ .Data: raw [1, 1:265955] 01 01 03 02 ... .- attr(*, "dimnames")=List of 2 .. ..$: chr "NA19240" .. ..$: chr [1:265955] "rs4097465" "rs7754266" "rs9393087" "rs12192290" ... $ supp:'data.frame': 265955 obs. of 5 variables: ..$ dbSNPalleles: Factor w/ 6 levels "A/C","A/G","A/T",..: 6 2 5 3 1 4 6 1 4 5 ... ..$ Assignment : Factor w/ 6 levels "A/C","A/G","A/T",..: 6 2 5 3 1 4 6 1 4 5 ... ..$ Chromosome : Factor w/ 1 level "chr6": 1 1 1 1 1 1 1 1 1 ... ..$ Position : int [1:265955] 37012 94609 94901 95272 96774 97749 98500 99536 99694 99750 ... ..$ Strand : Factor w/ 2 levels "+","-": 2 1 1 1 1 1 2 1 1 2 ...

Details

import of hapmap data using read.hapmap.snps

Examples

```
data(yri240_6)
names(yri240_6)
```

Index