BiostringsCinterfaceDemo

April 20, 2009

R topics documented:

`SolexaSequenceQ-class`
Class "SolexaSequenceQ" illustrates a class that coordinates Solexa sequence and base call quality scores.

Description

This class represents Solexa reads, their names, and corresponding base call quality scores in a coordinate fashion. It is meant as an illustration, and is not a final implementation.

Objects from the Class

Objects from the class are usually created by reading Solexa `s_<lane>_sequence.txt` files. Objects can also be created using the function `SolexaSequenceQ`.

Slots

Use accessors (below) to retrieve information contained in these slots. Note that sequences, etc., should be treated as 'read only'.

`sequences`: Object of class `"DNAStringSet"` containing called read sequences. Reads are all the same length. The Solexa missing base symbol `.` has been translated to the IUPAC standard –.

`.names`: Object of class `"BStringSet"` containing the names of all sequence reads.

`scores`: Object of class `"BStringSet"` containing the ASCII-encoded quality scores of each called base. Decode each nucleotide `nt` with `nt-64` to obtain a Solexa Q value (typically between -40 and 40).

Methods

[`signature(x = "SolexaSequenceQ", i = "ANY", j = "missing")`: select a subset of reads indexed by `i`, returned as a `DNAStringSet` object.

length `signature(x = "SolexaSequenceQ")`: return the number of reads in the object.

names `signature(x = "SolexaSequenceQ")`: return the names of the reads as a `BStringSet` object.

scores `signature(x = "SolexaSequenceQ")`: return the scores of the reads as a `BStringSet` object.

sequences `signature(x = "SolexaSequenceQ")`: return the sequences of the reads as a `DNAStringSet` object.

show `signature(object = "SolexaSequenceQ")`: display the object in a compact fashion.

Author(s)

Martin Morgan <mtmorgan@fhcrc.org>

References

Refer to Solexa documentation for inforamtion on quality score interpretation.

See Also

`link{readSolexaFastq}` for string input; `DNAStringSet`, `BStringSet`.

Examples

```
egFile <- system.file('extdata', 's_1_sequence.txt',
                      package='BiostringsCinterfaceDemo')
obj <- readSolexaFastQ(egFile)
obj
length(obj)
sequences(obj)
scores(obj)
## coordinated subsetting
samp <- obj[sample(length(obj), 10)]
sequences(samp)
names(samp)
```

`alphabetByCycle` *Summarize alphabet use by cycle (nucleotide position)*

Description

This function summarizes nucleotide frequencies per cycle in a `DNAStringSet` containing DNA strings of uniform width.

Usage

```
alphabetByCycle(stringSet, alphabet = Biostrings::alphabet(stringSet))
```

Arguments

`stringSet`	An object of class `DNAStringSet`, with uniform width.
`alphabet`	(Optional) characters represented in the sequence and for which frequencies will be tabulated.

Value

An integer matrix of counts, with rows corresonding to letters of `alphabet` and columsn to cycles `1:width`.

Author(s)

Martin Morgan <mtmorgan@fhcrc.org>

See Also

`DNAStringSet`

Examples

```
example(readSolexaFastQ) # read sequences into 'sq'
alphabetByCycle(sequences(sq))[,2:5] # first five cycles

## specify alpha for scores
alpha <- sapply(33:93, function(i) rawToChar(as.raw(i)))
abc <- alphabetByCycle(scores(sq), alphabet=alpha)
abc[50:61,10:20] ## encoded scores 50:61, cycles 10:20
```

`read.fasta.demo` *Reading FASTA data from a collection of files*

Description

Just some demo functions implemented in C to illustrate the use of the Biostrings C interface for loading character data into an XStringSet object. These functions only support a simplified form of the FASTA format where the records have only 2 lines: one for the description (starting with a '>') and one for the sequences.

Usage

```
read.fasta.demo1(filepaths, desc.prefix=">")
read.fasta.demo2(filepaths, baseClass, desc.prefix=">")
read.fasta.demo3(filepaths, baseClass, desc.prefix=">")
read.fasta.demo3B(filepaths, baseClass, desc.prefix=">")
```

Arguments

filepaths	A character vector containing file paths.
baseClass	Must be the name of one of the direct XString subtypes i.e. "BString", "DNAString", "RNAString" or "AAString". The elements of the XStringSet object returned by the reading function will be of that class. For example with baseClass="DNAString", this XStringSet object will be a DNAStringSet object therefore all its elements will be DNAString objects.
desc.prefix	A single string containing the markup used at the beginning of each description line.

Details

[NO DETAILS FOR NOW]

See Also

XStringSet-class, DNAString-class

Examples

```
file <- system.file("extdata", "fake.fa", package="BiostringsCinterfaceDemo")

## Load the file into a named character vector
x1 <- read.fasta.demo1(file)
x1

## Load the file into a DNAStringSet object
x2 <- read.fasta.demo2(file, "DNAString")
x2

## Load the file into a list of 2 XStringSet objects
x3 <- read.fasta.demo3(file, "DNAString")
x3
x3B <- read.fasta.demo3B(file, "DNAString")
x3B
```

---

readSolexaFastq *Read Solexa fastq and fasta-style files*

Description

These function illustrates how to read a Solexa fasta- and fastq-style file into a DNAStringSet or SolexaSequenceQ object.

Usage

```
readSolexaFastQ(filepaths)
readSolexaFastA(filepaths)
```

Arguments

filepaths	A character vector containing file paths.

Details

Each sequence in a Solexa fastq file consists of four lines. The first and third are identifiers (identical in each record), the second line is the sequence, and the fourth line the ASCII-encoded base quality score. For example:

```
@HWI-EAS88_1_1_1_1001_499
GGACTTTGTAGGATACCCTCGCTTTCCTTCTCCTGT
+HWI-EAS88_1_1_1_1001_499
]]]]]]]]]]]]Y]Y]]]]]]]]]]]]VCHVMPLAS
```

`readSolexaFastq` parses one or more files in this format into a single `SolexaSequenceQ` object.

Solexa fasta files are nearly standard, except that uncalled bases are encoded as `.` instead of `-`.

Value

`readSolexaFastA` returns a `DNAStringSet` class represnting all reads. `readSolexaFastQ` returns a `SolexaSequenceQ` class representing reads, their quality scores, and the read names.

See Also

`SolexaSequenceQ`

Examples

```
aFile <- system.file('extdata', 's_5.fasta',
                     package='BiostringsCinterfaceDemo')
sa <- readSolexaFastQ(aFile)
sa

qFile <- system.file('extdata', 's_1_sequence.txt',
                     package='BiostringsCinterfaceDemo')
sq <- readSolexaFastQ(qFile)
sq
```

`utilities` *Utilities for working with short-read data sets*

Description

These functions provide efficient ways of obtaining information related to short reads.

Usage

```
countLines(filepaths)
```

Arguments

`filepaths`	Character vector of file paths.

Details

`countLines` counts the number of lines in each file represented in its argument.

Value

`countLines` returns an integer vector of line counts per file.

Author(s)

Martin Morgan <mtmorgan@fhcrc.org>

Examples

```
egFile <- system.file('extdata', 's_1_sequence.txt',
                      package='BiostringsCinterfaceDemo')
countLines(egFile)
```

Index