

配布するUSBメモリ中のhogeフォルダを
デスクトップにコピーしておいてください。

NGSハンズオン講習会

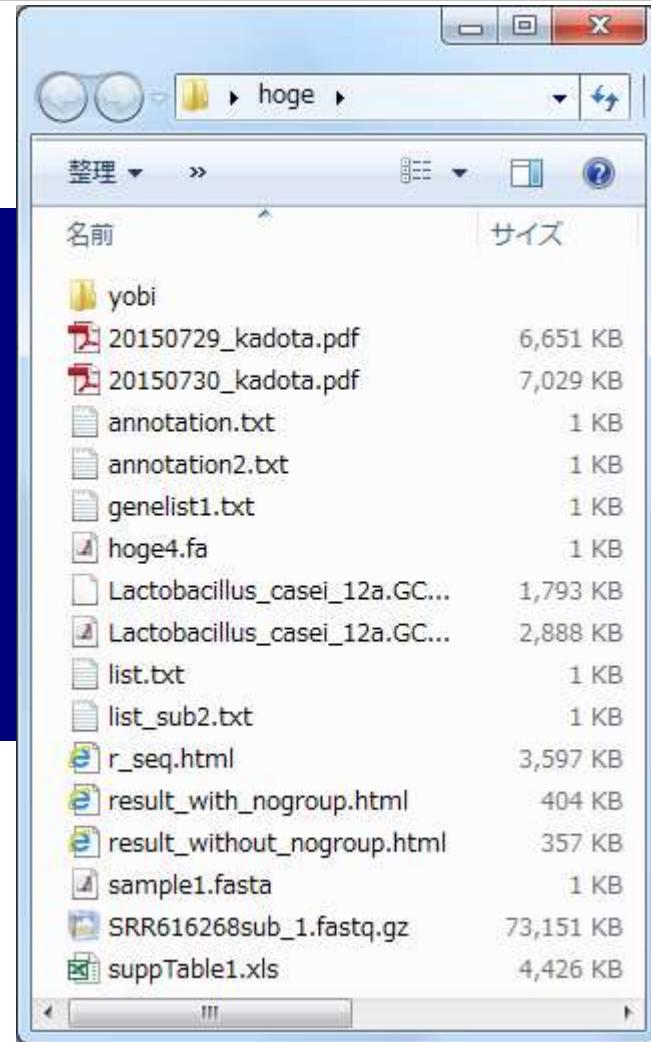
Bioconductorの利用法

東京大学・大学院農学生命科学研究所
アグリバイオインフォマティクス教育研究プログラム

門田幸二(かどた こうじ)

kadota@iu.a.u-tokyo.ac.jp

<http://www.iu.a.u-tokyo.ac.jp/~kadota/>



Rが導く未来
は明るいのだ
ろうか…



Contents(全体)

- 7月22日(水):84→**83名**。Bio-Linux 8とRのインストール状況確認。基本自習(門田・寺田先生)
- 7月23日(木):92→**90名**。Linux基礎。LinuxコマンドなどUNIXの基礎の理解(門田)
- 7月24日(金):85→**83名**。スクリプト言語。シェルスクリプト(アメリカエフ株式会社 服部恵美先生)
- 7月27日(月):93→**91名**。スクリプト言語。Perl(アメリカエフ 服部先生)
- 7月28日(火):91→**90名**。スクリプト言語。Python(アメリカエフ 服部先生)
- 7月29日(水):94→**88名**。データ解析環境R(門田)
 - R基礎1(初級):R言語の基礎(インストールから利用まで)
 - R基礎2(初級):ファイルの読み込み、行列演算の基本
 - R各種パッケージ(中級):パッケージのインストール法と代表的なパッケージの利用法
- 7月30日(木):96→**91名**。データ解析環境R(門田)
 - Bioconductorの利用法1(中級):データの型やバージョンの違い
 - Bioconductorの利用法2(中級):FASTA/FASTQファイルの各種解析
- 8月3日(月):89→**84名**。NGS解析。基礎(アメリカエフ 山口昌雄先生)
- 8月4日(火):85→**80名**。NGS解析。ゲノムReseq、変異解析(アメリカエフ 山口先生)
- 8月5日(水):86 →**81名**。NGS解析。RNA-seq、統計解析(前半:山口先生、後半:門田)
- 8月6日(木):104 →**98名**。NGS解析。ChIP-seq(理研 森岡勝樹先生)

Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

パッケージ

R起動直後に「?関数名」と打ち込んで
も、使用法を記したウェブページが開
かずにエラーが出ることがあります。

R R Console

```
R version 3.1.3 (2015-03-09) -- "Smooth Sidewalk"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、「license()」あるいは「licence()」と入力し$

R は多くの貢献者による共同プロジェクトです。
詳しくは「contributors()」と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
「citation()」と入力してください。

'demo()'と入力すればデモをみることができます。
'help()'とすればオンラインヘルプが出ます。
'help.start()'で HTML ブラウザによるヘルプがみられます。
'q()'と入力すれば R を終了します。

> ?subseq
No documentation for 'subseq' in specified packages and libraries:
you could try '??subseq'
> ?alphabetFrequency
No documentation for 'alphabetFrequency' in specified packages and$ you could try '??alphabetFrequency'
```

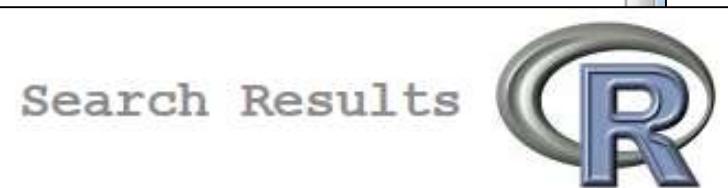
パッケージ

①「??alphabetFrequency」と打ち込むように勧められているので打ってみる。検索結果のウェブページが表示されるので、
②それっぽい関数名のところをクリック。

R Console

```
> ?subseq
No documentation for 'subseq' in specified packages and libraries:
you could try '??subseq'
> ?alphabetFrequency
No documentation for 'alphabetFrequency' in specified packages and libraries:
you could try '??alphabetFrequency'
> ??alphabetFrequency
starting httpd help server ... done
> |
```

 ①



The search string was "alphabetFrequency"

Help pages:

Biostrings::class::MultipleAlignment	MultipleAlignment objects
Biostrings::letterFrequency	Calculate the frequency of letters in a biological sequence, or the consensus matrix of a set of sequences
GenomicAlignments::stackStringsFromBam	Stack the read sequences stored in a BAM file on a region of interest
ShortRead::QualityScore-class	Quality scores for short reads and their alignments

 ②

パッケージ

`letterFrequency {Biostrings}`

alphabetFrequency関数はBiostringsというパッケージから提供されているものだと読み解く。「??関数名」は、関数名は既知だがどのパッケージから提供されているものかを知りたい場合などに利用する。

Calculate the frequency of letters in a biological sequence, or the consensus matrix of a set of sequences

Description

Given a biological sequence (or a set of biological sequences), the `alphabetFrequency` function computes the frequency of each letter of the relevant [alphabet](#).

`letterFrequency` is similar, but more compact if one is only interested in certain letters. It can also tabulate letters "in common".

`letterFrequencyInSlidingView` is a more specialized version of `letterFrequency` for (non-masked) [XString](#) objects. It tallies the requested letter frequencies for a fixed-width view, or window, that is conceptually slid along the entire input sequence.

The `consensusMatrix` function computes the consensus matrix of a set of sequences, and the `consensusString` function creates the consensus sequence from the consensus matrix based upon specified criteria.

In this man page we call "DNA input" (or "RNA input") an [XString](#), [XStringSet](#), [XStringViews](#) or [MaskedXString](#) object of base type DNA (or RNA).

Usage

```
alphabetFrequency(x, as.prob=FALSE, ...)
hasOnlyBaseLetters(x)
uniqueLetters(x)
```

パッケージ

multi-FASTAファイルを読み込んで様々な解析ができるのは、①Biostringsや②seqinrなどの塩基配列解析用パッケージのおかげです。

- ・ イントロ | 一般 | 任意の位置の塩基を置換 (last modified 2013/09/12)
- ・ イントロ | 一般 | 指定した範囲の配列を取得 (last modified 2015/04/06) NEW
- ・ イントロ | 一般 | 指定したID(染色体やdescription)を取得 (last modified 2014/03/10)
①
- ・ イントロ | 一般 | 翻訳配列(translate)を取得(基礎) | Biostrings (last modified 2015/03/09)
- ・ イントロ | 一般 | 翻訳配列(translate)を取得(応用) | seqinr(Charif 2005) (last modified 2015/03/09)
- ・ イントロ | 一般 | 相補鎖(complement)を取得 (last modified 2013/06/14)
②
- ・ イントロ | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
- ・ イントロ | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
- ・ イントロ | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
- ・ イントロ | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
- ・ イントロ | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
- ・ イントロ | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)
- ・ イントロ | 一般 | 逆相補鎖(reverse complement)を取得 (last modified 2013/06/14)

イントロ | 一般 | 翻訳配列(translate)を取得(基礎) | Biostrings

Biostringsパッケージを用いて塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。翻訳のための遺伝コード(genetic code)は、Standard Genetic Codeだそうです。もちろん生物種?によって多少違い(variants)があるようで、"Standard", "SGC0", "Vertebrate Mitochondrial", "SGC1"などいろいろ選べるようです。「ファイル」→「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. FASTA形式ファイル(sample1.fasta)の場合:

multi-FASTAではないsingle-FASTA

```
in_f <- "sample1.fasta"  
out_f <- "hoge1.fasta"  
  
#必要なパッケージをロード  
library(Biostrings)  
  
#入力ファイルの読み込み  
fasta <- readDNAStringSet
```

イントロ | 一般 | 翻訳配列(translate)を取得(応用) | seqinr(Charif_2005)

seqinrパッケージを用いて塩基配列を読み込んでアミノ酸配列に翻訳するやり方を示します。本気で翻訳配列を取得する場合はこちらの利用をお勧めします。翻訳できないコードはアミノ酸X(不明なアミノ酸)に変換してくれたり、translate関数のオプションとしてambiguous=Tとすると、翻訳できるものは可能な限り翻訳してくれます(高橋 広夫 氏提供情報)。lapply関数を用いるやり方(高橋 広夫 氏提供情報)とsapply関数を用いるやり方(甲斐 政親 氏提供情報)を示します。

「ファイル」→「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. FASTA形式ファイル(sample1.fasta)の場合:

multi-FASTAではないsingle-FASTA形式ファイルです。

```
in_f <- "sample1.fasta"  
out_f <- "hoge1.fasta"  
  
#必要なパッケージをロード  
library(seqinr)  
  
#入力ファイルの読み込み  
hoge <- read.fasta(in_f, seqtype="DNA")#in_fで指定したファイルの読み込み  
hoge #確認してるだけです
```

#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_fに格納
#パッケージの読み込み
#確認してるだけです

パッケージ

R R Console

```
> library(Biostrings) ←  
要求されたパッケージ BiocGenerics をロード中です  
要求されたパッケージ parallel をロード中です  
  
次のパッケージを付け加えます: 'BiocGenerics'
```

Biostringsというパッケージをlibrary関数を用いて読み込むことによって、alphabetFrequencyのようなBiostringsが提供する関数群を利用できるのです。ここでは、意図的に「library(Biostrings)」を2回実行して、2回目は何も表示されないということを思い出させています。実際には1回のみで大丈夫です。「?alp」まで打ってからTabキーを押すなどして「タブ補完」テクを有効利用。

The following objects are masked from 'package:parallel':

```
clusterApply, clusterApplyLB, clusterCall, clusterExport, clusterMap, parApply, parCapply, parLapply, parLapplyLB, parRapply, parSapply, parSapplyLB
```

The following object is masked from 'package:stats':

```
xtabs
```

The following objects are masked from 'package:base':

```
anyDuplicated, append, as.data.frame, as.vector, colnames, do.call, duplicated, eval, evalq, Filter, get, intersect, is.unsorted, lapply, Map, mapply, mget, order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int, rownames, setdiff, sort, table, tapply, union, unique, unlist, unsplit
```

```
要求されたパッケージ S4Vectors をロード中です  
要求されたパッケージ stats4 をロード中です  
要求されたパッケージ IRanges をロード中です  
要求されたパッケージ XVector をロード中です  
> library(Biostrings)  
> ?alphabetFrequency
```

letterFrequency {Biostrings}

R Documentation

Calculate the frequency of letters in a biological sequence, or the consensus matrix of a set of sequences

Description

Given a biological sequence (or a set of biological sequences), the alphabetFrequency function computes the frequency of each letter of the relevant alphabet.

letterFrequency is similar, but more compact if one is only interested in certain letters. It can also tabulate letters "in common".

letterFrequencyInSlidingView is a more specialized version of letterFrequency for (non-masked) XString objects. It tallies the requested letter frequencies for a fixed-width view, or window, that is conceptually slid along the entire input sequence.

The consensusMatrix function computes the consensus matrix of a set of sequences, and the consensusString function creates the consensus sequence from the consensus matrix based upon specified criteria.

In this man page we call "DNA input" (or "RNA input") an XString, XStringSet, XStringViews or MaskedXString object of base type DNA (or RNA).

Usage

```
alphabetFrequency(x, as.prob=FALSE, ...)  
hasOnlyBaseLetters(x)  
uniqueLetters(x)
```

R本体とパッケージの関係

「R本体」と「パッケージ」の関係は、「パソコン」と「ソフト」、「Microsoft EXCEL」と「アドイン」、「Cytoscape」と「プラグイン」のようなものという理解でよい。

- パソコンを購入しただけの状態では、できることが限られています。
 - 通常は、Officeやウイルス撃退ソフトなどをインストールして利用します。
- Linuxをインストールしただけの状態では、できることが限られています。
 - 通常は、マッピングなど各種プログラムをインストールして利用します。
- R本体をインストールしただけの状態では、できることが限られています。
 - NGS解析を行う各種パッケージ(またはライブラリ)をインストールして利用します。

CRANとBioconductor

CRAN提供パッケージは生命科学を含む様々な分野で利用される。NGS解析は、主にBioconductor提供パッケージを利用。

■ R上で利用可能なパッケージの2大リポジトリ(貯蔵庫)

- CRAN (The Comprehensive R Archive Network): 6,878パッケージ
- Bioconductor: 1,024パッケージ

- 作図 | ROC曲線 | 基礎編 | [7. 図の重ね書き\(new\)](#) (last modified 2015/02/15) NEW
- 作図 | ROC曲線 | 基礎編 | [8. 凡例を追加\(legend\)](#) (last modified 2015/02/15) NEW
- 作図 | ROC曲線 | 応用編 (last modified 2015/02/07) NEW
- 作図 | [Splicing Graphs](#) (last modified 2015/02/07)
- パイプライン | [について](#) (last modified 2015/02/07)
- パイプライン | ゲノム | 発現変動 | 2篇
- パイプライン | ゲノム | 機能解析 | 2篇
- パイプライン | ゲノム | 機能解析 | 2篇
- パイプライン | ゲノム | small RNA | 2篇
- [リンク集](#) (modified 2012/03/29)

リンク集

- [R](#)
- [Bioconductor](#) [Gentleman et al., Genome Biol., 2004](#)
- [CRAN](#)
- [RjpWiki](#)
- [R Tips\(竹澤様\)](#)
- [BioEdit](#)(フリーの配列編集ソフト)
- [BioMart](#): [Smedley et al., BMC Genomics, 2009](#)
- [DDBJ Read Annotation Pipeline](#): [Nagasaki et al., DNA Res., 2013](#)
- [EMBOSS explorer](#) (EMBOSSのウェブ版)
- [Biostar](#): [Parnell et al., PLoS Comput Biol., 2011](#)
- [SEQanswers](#): [Li et al., Bioinformatics, 2012](#)
- [NGS WikiBook](#): [Li et al., Brief Bioinform., 2013](#)
- [HT Sequence Analysis with R and Bioconductor](#)

定期的にバージョンアップ

■ 近年のリリース頻度

□ R本体 (<http://www.r-project.org/>)

- 2015-06-18にver. 3.2.1をリリース
- 2015-04-16にver. 3.2.0をリリース
- 2015-03-09にver. 3.1.3をリリース
- 2014-10-31にver. 3.1.2をリリース
- ...
- 2012-03-30にver. 2.15.0をリリース
- ...

□ Bioconductor (<http://bioconductor.org/>)は半年ごとにリリース

- 2015-04にver. 3.1をリリース (R ver. 3.2.1で動作確認)、提供パッケージ数: 1,024
- 2014-10にver. 3.0をリリース (R ver. 3.1.1で動作確認)、提供パッケージ数: 934
- 2014-04にver. 2.14をリリース (R ver. 3.1.0で動作確認)、提供パッケージ数: 824
- 2013-10にver. 2.13をリリース (R ver. 3.0で動作確認)、提供パッケージ数: 750
- 2013-04にver. 2.12をリリース (R ver. 3.0で動作確認)、提供パッケージ数: 672
- 2012-10にver. 2.11をリリース (R ver. 2.15.1で動作確認)、提供パッケージ数: 608
- 2012-04にver. 2.10をリリース (R ver. 2.15.0で動作確認)、提供パッケージ数: 553
- ...

バグの修正や新たな機能がどんどん追加されている。最新版の利用をお勧め。毎年5月と11月ごろにバージョンアップするとよいだろう。



Bioconductor

Bioconductorに関する総説(Review)。ゲノム配列やアノテーションパッケージもBioconductorから提供されており、それらに関する言及もあり。

Nat Methods. 2015 Feb;12(2):115-21. doi: 10.1038/nmeth.3252.

Orchestrating high-throughput genomic analysis with Bioconductor.

Huber W¹, Carey VJ², Gentleman R³, Anders S¹, Carlson M⁴, Carvalho BS⁵, Bravo HC⁶, Davis S⁷, Gatto L⁸,
Girke T⁹, Gottardo R¹⁰, Hahne F¹¹, Hansen KD¹², Irizarry RA¹³, Lawrence M³, Love MI¹³, MacDonald J¹⁴,
Obenchain V⁴, Oleś AK¹, Pagès H⁴, Reyes A¹, Shannon P⁴, Smyth GK¹⁵, Tenenbaum D⁴, Waldron L¹⁶,
Morgan M⁴.

Author information

Abstract

Bioconductor is an open-source, open-development software project for the analysis and comprehension of high-throughput data in genomics and molecular biology. The project aims to enable interdisciplinary research, collaboration and rapid development of scientific software. Based on the statistical programming language R, Bioconductor comprises 934 interoperable packages contributed by a large, diverse community of scientists. Packages cover a range of bioinformatic and statistical applications. They undergo formal initial review and continuous automated testing. We present an overview for prospective users and contributors.

パッケージのインストール

- ・ [はじめに \(last modified 2015/03/31\)](#)
- ・ [参考資料 \(講義、講習会、本など\) \(last modified 2015/07/07\) NEW](#)
- ・ [過去のお知らせ \(last modified 2015/07/08\) NEW](#)
- ・ [インストールについて \(last modified 2015/04/04\)](#)
- ・ インストール | R本体 | 最新版 | [Win用 \(last modified 2015/03/22\) 推奨](#)
- ・ インストール | R本体 | 最新版 | [Mac用 \(last modified 2015/04/22\) 推奨](#)
- ・ インストール | R本体 | 過去版 | [Win用 \(last modified 2015/03/22\)](#)
- ・ インストール | R本体 | 過去版 | [Mac用 \(last modified 2015/03/22\)](#)
- ・ インストール | Rパッケージ | [ほぼ全て\(20GB以上?!\) \(last modified 2015/05/25\)](#)
- ・ インストール | Rパッケージ | [必要最小限プラスアルファ\(数GB?!\) \(last modified 2015/06/02\) 推奨](#)
- ・ インストール | Rパッケージ | [必要最小限プラスアルファ\(アグリバイア居室のみ\) \(last modified 2015/06/16\)](#)
- ・ インストール | Rパッケージ | [必要最小限\(数GB?!\) \(last modified 2015/05/25\)](#)
- ・ インストール | Rパッケージ | [個別 \(last modified 2015/06/10\)](#)
- ・ [\(削除予定\)Rのインストールと起動 \(last modified 2015/04/02\)](#)
- ・ [\(削除予定\)個別パッケージのインストール \(last modified 2015/02/20\)](#)
- ・ [基本的な利用法 \(last modified 2015/04/03\)](#)
- ・ [サンプルデータ \(last modified 2015/06/15\)](#)
- ・ バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | [NGSハンズオン講習会2015 \(last modified 2015/04/02\)](#)
- ・ バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | [速習コース2014 \(last modified 2015/04/02\)](#)
- ・ [書籍 | ランスクリプトーム解析 | について \(last modified 2014/05/12\)](#)
- ・ [書籍 | ランスクリプトーム解析 | 2.3.1 RNA-seqデータ\(FASTQファイル\) \(last modified 2014/04/15\)](#)
- ・ [書籍 | ランスクリプトーム解析 | 2.3.2 リファレンス配列 \(last modified 2014/04/16\)](#)

「必要最小限プラスアルファ」の推奨
インストール手順を行えば、「(Rで)塩基配列解析」で利用する多くのパッケージがインストールされます。

パッケージのインストール

インストール | Rパッケージ | 必要最小限プラスアルファ(数GB?!) NEW

(Rで)塩基配列解析、(Rで)マイクロアレイデータ解析中で利用するパッケージ、プラスアルファのパッケージをインストールするやり方です。Rパッケージの2大リポジトリであるCRANとBioconductorから提供されているパッケージ群のうち、一部のインストールに相当しますので、相当短時間でインストールが完了します。

1. R本体を起動



2. CRANから提供されているパッケージ群のインストール

以下を「Rコンソール画面上」でコピー＆ペースト。どこからダウンロードするか？と聞かれるので、その場合は自分のいる場所から近いサイトを指定しましょう。

#(Rで)塩基配列解析で主に利用

```
install.packages("limma")
install.packages("samr")
install.packages("seqinr")
```

#(Rで)マイクロアレイデータ解析でも利用
#(Rで)マイクロアレイデータ解析でも利用
#(Rで)マイクロアレイデータ解析でも利用

#(Rで)マイクロアレイデータ解析で利用

```
install.packages("cclust")
install.packages("class")
install.packages("e1071")
install.packages("GeneCycle")
install.packages("gptk")
install.packages("GSA")
install.packages("mixOmics")
install.packages("pvclust")
install.packages("RobLoxBioC")
install.packages("som")
install.packages("st")
install.packages("varSelRF")
```

#アグリバイオの他の講義科目で利用予定

```
install.packages("ape")
install.packages("cluster")
install.packages("fields")
```

パッケージのインストール

```
biocLite("DEGseq", suppressUpdates=TRUE)
biocLite("DESeq", suppressUpdates=TRUE)
biocLite("DESeq2", suppressUpdates=TRUE)
biocLite("DiffBind", suppressUpdates=TRUE)
biocLite("doMC", suppressUpdates=TRUE)
biocLite("EBSeq", suppressUpdates=TRUE)
biocLite("EDASeq", suppressUpdates=TRUE)
biocLite("edgeR", suppressUpdates=TRUE)
biocLite("GenomicAlignments", suppressUpdates=TRUE)
```

①ゲノム情報のパッケージ群

([BSgenome](#)...)はBioconductorから提供されています。ここでは計6パッケージをインストールしています。②例えば赤線部分は、マウスのmm10というバージョンのゲノム配列情報を含むパッケージの名前([BSgenome.Mmusculus.UCSC.mm10](#))に相当します。biocLiteという関数を用いて該当パッケージをインストールしています。



4. Bioconductorから提供されているパッケージ群のインストール

ゲノム配列パッケージです。一つ一つの容量が尋常でないため、必要に応じてテキストエディタなどに予めコピペしておき、いらないゲノムパッケージを削除してからお使いください。

```
source("http://bioconductor.org/biocLite.R")#おまじない
biocLite("BSgenome.Athaliana.TAIR.TAIR9", suppressUpdates=TRUE)#シロイヌナズナゲノム
biocLite("BSgenome.Celegans.UCSC.ce6", suppressUpdates=TRUE)#線虫ゲノム
biocLite("BSgenome.Drerio.UCSC.danRer7", suppressUpdates=TRUE)#ゼブラフィッシュゲノム
biocLite("BSgenome.Hsapiens.NCBI.GRCh38", suppressUpdates=TRUE)#ヒトゲノム(GRCh38)
biocLite("BSgenome.Hsapiens.UCSC.hg19", suppressUpdates=TRUE)#ヒトゲノム(hg19)
biocLite("BSgenome.Mmusculus.UCSC.mm10", suppressUpdates=TRUE)#マウスゲノム(mm10)
```



Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content項目の結果の違い → Rで検証

Bioconductor概観

- ・ 作図 | ROC曲線 | 基礎編 | 7. 図の重ね書き(new) (last modified 2015/02/15) NEW
- ・ 作図 | ROC曲線 | 基礎編 | 8. 凡例を追加(legend) (last modified 2015/02/15) NEW
- ・ 作図 | ROC曲線 | 応用編 (last modified 2015/02/07) NEW
- ・ 作図 | SplicingGraphs (la
- ・ パイプライン | について (
- ・ パイプライン | ゲノム | 発
- ・ パイプライン | ゲノム | 機
- ・ パイプライン | ゲノム | 機
- ・ パイプライン | ゲノム | sm
- ・ リンク集 (modified 2015/02/15)

リンク集

- ・ [R](#)
- ・ [Bioconductor: Gentleman et al., Genome Biol., 2004](#)
- ・ [CRAN](#) ②
- ・ [RjpWIKI](#)
- ・ [R Tips\(竹澤様\)](#)
- ・ [BioEdit\(フリー\)](#)
- ・ [BioMart: Smed](#)
- ・ [DDBJ Read An](#)
- ・ [EMBOSS expla](#)
- ・ [Biostar: Parnell](#)
- ・ [SEQanswers: L](#)
- ・ [NGS WikiBook](#)
- ・ [HT Sequence A](#)

The screenshot shows the Bioconductor website's 'About' page. The header features the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. The main content includes sections on the history and tools of Bioconductor, news, and various ways to get involved (Install, Learn, Use, Develop). Red numbered arrows highlight specific parts of the page:

- ① Points to the 'Link Collection' link in the sidebar.
- ② Points to the 'CRAN' link in the sidebar.
- ③ Points to the 'Install' section in the main content area.

About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 934 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

News

- Nature Methods Orchestrating high-throughput genomic analysis with Bioconductor ([abstract](#); full-text free with registration) and other recent [literature citations](#).
- Read our latest [newsletter](#).
- Updated [course material](#) and [videos](#).
- Use the [support site](#) to get help installing, learning and using Bioconductor.

Install »

Get started with Bioconductor

- [Install Bioconductor](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master Bioconductor tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with Bioconductor

- [Software, Annotation, and Experiment packages](#)
- [Amazon Machine Image](#)
- [Latest release announcement](#)
- [Support site](#)

Develop »

Contribute to Bioconductor

- [Use Bioc 'devel'](#)
- ['Devel' Software, Annotation and Experiment packages](#)
- [Package guidelines](#)
- [New package submission](#)
- [Developer resources](#)
- [Build reports](#)

PubMed上で「R Bioconductor」でキーワード検索し原著論文があるパッケージのみ探すのも一つの戦略ですが、原著論文公開前のパッケージも見つかります。

Bioconductor概観

ネットワーク環境にもよりますが、数秒～数分以内にこのような画面になります。

The screenshot shows a web browser window displaying the Bioconductor BiocViews software page. The URL in the address bar is http://bioconductor.org/packages/release/BiocViews.html#_Software. The page has a blue header with the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. A search bar is also present in the header. The main content area is titled "All Packages" and displays a table of packages found under Software. The table has columns for Package, Maintainer, and Title. The packages listed are: a4, a4Base, a4Classif, a4Core, a4Preproc, a4Reporting, ABarray, and ABSSeq. The "a4" package is highlighted in the table.

Bioconductor version 3.0 (Release)

Autocomplete biocViews search:

- Software (936)
 - AssayDomain (299)
 - BiologicalQuestion (261)
 - Infrastructure (187)
 - ResearchField (193)
 - StatisticalMethod (261)
 - Technology (591)
 - WorkflowStep (477)
- AnnotationData (895)
- ExperimentData (223)

Packages found under Software:

Show All entries	Search table:	
Package	Maintainer	Title
a4	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Umbrella Package
a4Base	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Base Package
a4Classif	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Classification Package
a4Core	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Core Package
a4Preproc	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Preprocessing Package
a4Reporting	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Reporting Package
ABarray	Yongming Andrew Sun	Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microarray (AB1700) gene expression data.
ABSSeq	Wentao Yang	ABSSeq: a new RNA-Seq analysis method based on absolute expression differences and generalized Poisson

Bioconductor概観

基本的には左側のカテゴリ分けのところを眺めますが、Biostringsなど何を行うパッケージかがある程度分かっているものから逆引きして感覚をつかんでおくとよいでしょう。

The screenshot shows the Bioconductor BiocViews software page. At the top, there's a navigation bar with links for Home, Install, Help, Developers, and About. A search bar is also present. Below the navigation bar, the page title is "All Packages". A sub-header indicates "Bioconductor version 3.0 (Release)". There is an "Autocomplete biocViews search:" input field. To the right of the search field is a table titled "Packages found under Software:". The table has columns for Package, Maintainer, and Title. The packages listed are: a4, a4Base, a4Classif, a4Core, a4Preproc, a4Reporting, ABarray, and ABSSeq. Each row provides details about the package, such as its maintainer and a brief description. A large black arrow points from the text "左側のカテゴリ分けのところを眺めます" to the "Software (936)" category in the sidebar. Another large black arrow points from the text "逆引きして感覚をつかんでおくとよいでしょう。" to the search table.

「RNA-seq」など、フリーワード検索をやってもいいとは思いますが、経験上あまりうまく引っかかってこないので私はやりません。

Package	Maintainer	Title
a4	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Umbrella Package
a4Base	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Base Package
a4Classif	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Classification Package
a4Core	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Core Package
a4Preproc	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Preprocessing Package
a4Reporting	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Reporting Package
ABarray	Yongming Andrew Sun	Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microarray (AB1700) gene expression data.
ABSSeq	Wentao Yang	ABSSeq: a new RNA-Seq analysis method based on absolute expression differences and generalized Poisson

「biostrings」と打つとすぐリストアップされる。

Bioconductor概観

The screenshot shows a web browser window for the Bioconductor BiocViews software. The URL is http://bioconductor.org/packages/release/BiocViews.html#_Software. The page title is "Bioconductor - BiocViews". The navigation menu includes Home, Install, Help, Developers, and About. A search bar at the top right contains the placeholder "Search:". Below it, a search table displays results for "biostrings". The table has columns for Package, Maintainer, and Title. The "Package" column lists "Biostrings" and "BSgenome". The "Title" column provides a brief description of each package: "String objects representing biological sequences, and matching algorithms" for Biostrings and "Infrastructure for Biostrings-based genome data packages" for BSgenome. A red arrow points to the "Biostrings" entry in the table. On the left, a sidebar shows a tree view of categories under "Software": AssayDomain (299), BiologicalQuestion (261), Infrastructure (187), ResearchField (193), StatisticalMethod (261), Technology (591), WorkflowStep (477), AnnotationData (895), and ExperimentData (223). The "Software" category is currently selected.

biostrings

biostrings

All Packages

Bioconductor version 3.0 (Release)

Autocomplete biocViews search:

Packages found under Software:

Show All entries	Search table: biostrings	
Package	Maintainer	Title
Biostrings	H. Pages	String objects representing biological sequences, and matching algorithms
BSgenome	H. Pages	Infrastructure for Biostrings-based genome data packages

Showing 1 to 2 of 2 entries (filtered from 939 total entries)

Previous Next

Software (936)

- ▶ AssayDomain (299)
- ▶ BiologicalQuestion (261)
- ▶ Infrastructure (187)
- ▶ ResearchField (193)
- ▶ StatisticalMethod (261)
- ▶ Technology (591)
- ▶ WorkflowStep (477)
- ▶ AnnotationData (895)
- ▶ ExperimentData (223)

BioconductorのBiostringsページのページに飛びます。

Bioconductor概観

The screenshot shows a web browser window with the URL <http://bioconductor.org/packages/release/bioc/html/Biostrings.html>. The page title is "Bioconductor - Biostrings". The main content area displays information about the Biostrings package, including its purpose (string objects for biological sequences), version (Release 3.0), authors (H. Pages, P. Aboyoun, R. Gentleman, S. DebRoy), maintainer (H. Pages <hpages at fhcrc.org>), citation details, installation instructions, and documentation links. A sidebar on the right contains sections for "Workflows" (listing various Bioconductor workflows) and "Mailing Lists". A yellow callout box in the top right corner points to the "BioconductorのBiostringsページのページに飛びます。" text.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Search:

Home » Bioconductor 3.0 » Software Packages » Biostrings

Biostrings

String objects representing biological sequences, and matching algorithms

Bioconductor version: Release (3.0)

Memory efficient string containers, string matching algorithms, and other utilities, for fast manipulation of large biological sequences or sets of sequences.

Author: H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy

Maintainer: H. Pages <hpages@fhcrc.org>

Citation (from within R, enter `citation("Biostrings")`):

Pages H, Aboyoun P, Gentleman R and DebRoy S. *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.34.1.

Installation

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("Biostrings")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

Workflows

Common Bioconductor workflows include:

- [Oligonucleotide Arrays](#)
- [High-throughput Sequencing](#)
- [Counting Reads for Differential Expression](#) (parathyroideSE vignette)
- [Annotation](#)
- [Annotating Variants](#)
- [Annotating Ranges](#)
- [Flow Cytometry](#) and other assays
- [Candidate Binding Sites for Known Transcription Factors](#)
- [Cloud-enabled cis-eQTL search and annotation](#)
- [RNA-Seq workflow: gene-level exploratory analysis and differential expression](#)
- [Changing genomic coordinate systems with rtracklayer::liftOver](#)
- [Mass spectrometry and proteomics data analysis](#)

Mailing Lists

Post questions about Bioconductor packages to our mailing lists. Read the [posting guide](#) before posting!

Bioconductor概観

The screenshot shows the documentation page for the Biostrings package. At the top, there's a navigation bar with back, forward, and search icons, and the URL <http://bioconductor.org/packages/release/bioc/html/Biostrings.html>. Below the URL is a "Documentation" section with a link to "browseVignettes('Biostrings')".

The main content area has a "Details" section. Under "biocViews", there is a red box highlighting the following keywords: [DataImport](#), [DataRepresentation](#), [Genetics](#), [Infrastructure](#), [SequenceMatching](#), [Sequencing](#), and [Software](#). A large red arrow points from the text "ちょっと下のほうに移動。biocViewsのところで見えるキーワードっぽいのがさきほどのカテゴリ分けに相当。" to this highlighted section.

Other sections visible include "PDF", "R Script", "Text", "NEWS", "Version", "In", "License", "Depends", "Imports", "Suggests", "System Requirements", and "URL". The "URL" section lists several R packages: [altcdfenvs](#), [Basic4Cseq](#), [BRAIN](#), [BSgenome](#), [ChIPpeakAnno](#), [ChIPsim](#), [cleaver](#), [CRISPRseek](#), [DASiR](#), [DECIPHER](#), [deepSNV](#), [FDb.FANTOM4.promoters.hg19](#), [GeneRegionScan](#), [genomes](#), [GenomicAlignments](#), [GOTHic](#), [harbChIP](#), [hiReadsProcessor](#), [iPAC](#), [JASPAR2014](#), [kebabs](#), [methVisual](#), [minfi](#), [MotifDb](#), [motifRG](#), [oligo](#), [oneChannelGUI](#).

BioconductorのBiostringsパッケージのページで、ちょっと下のほうに移動。biocViewsのところで見えるキーワードっぽいのがさきほどのカテゴリ分けに相当。例えば、DataRepresentationをクリックするとカテゴリ分けの階層関係がわかる。

Bioconductor概観

DataRepresentationのカテゴリに含まれるのは34パッケージであることが分かる。赤枠部分がそのリスト。Biostringsは、①大分類はSoftware、②中分類はInfrastructureとなっており、その下の階層の③DataRepresentationに含まれていることがわかる。

bioconductor.org/packages/release/BiocViews.html#_DataRepreser

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Search:

Home » BiocViews

All Packages

Bioconductor version 3.0 (Release)

Autocomplete biocViews search:

①

②

③

Packages found under DataRepresentation:

Package	Maintainer	Title
AtlasRDF	James Malone	Gene Expression Atlas query and gene set enrichment package.
BaseSpaceR	Adrian Alexa	R SDK for BaseSpace RESTful API
bigmemoryExtras	Peter M. Haverty	An extension of the bigmemory package with added safety, convenience, and a factor class.
Biostrings	H. Pages	String objects representing biological sequences, and matching algorithms
BSgenome	H. Pages	Infrastructure for Biostrings-based genome data packages
cummeRbund	Loyal A. Goff	Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data.
flowPlots	N. Hawkins	flowPlots: analysis plots and data class for gated flow cytometry data
flowWorkspace	Greg Finak, Mike Jiang	Import flowJo Workspaces into BioConductor and replicate flowJo gating with flowCore

Bioconductor概観

①大分類のSoftware、②中分類のInfrastructureも存在する。Biostringsは塩基配列の切り出しや文字列検索系関数も提供しているので、③SequenceMatchingがあるのも妥当。

The screenshot shows the Bioconductor Biostrings package documentation page at <http://bioconductor.org/packages/release/bioc/html/Biostrings.html>. The page includes a navigation bar, a documentation section, and a details section.

Documentation:

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("Biostrings")
```

Details:

biocViews: [DataImport](#), [DataRepresentation](#), [Genetics](#), [Infrastructure](#), [SequenceMatching](#), [Sequencing](#), [Software](#)

Version: 2.34.1

In: Bioconductor BioC 1.6 (R-2.1) or earlier since

License: Artistic-2.0

Depends: R (>= 2.8.0), methods, [BiocGenerics](#)(>= 0.11.3), [S4Vectors](#)(>= 0.2.2), [IRanges](#)(>= 1.99.27), [XVector](#)(>= 0.5.8)

Imports: graphics, methods, stats, utils, [BiocGenerics](#), [IRanges](#), [XVector](#), [zlibbioc](#), [BSgenome](#)(>= 1.13.14), [BSgenome.Celegans.UCSC.ce2](#)(>= 1.3.11), [BSgenome.Dmelanoqaster.UCSC.dm3](#)(>= 1.3.11), [BSgenome.Hsapiens.UCSC.hg18](#), [drosophila2probe](#), [hgu95av2probe](#), [hgu133aprobe](#), [GenomicFeatures](#)(>= 1.3.14), [hgu95av2cdf](#), [affy](#)(>= 1.41.3), [affydata](#)(>= 1.11.5), [RUnit](#)

System Requirements

URL: [altcdfenvs](#), [Basic4Cseq](#), [BRAIN](#), [BSgenome](#), [ChIPpeakAnno](#), [ChIPsim](#), [cleaver](#), [CRISPRseek](#), [DASiR](#), [DECIPHER](#), [deepSNV](#), [FDb.FANTOM4.promoters.hg19](#), [GeneRegionScan](#), [genomes](#), [GenomicAlignments](#), [GOTHic](#), [harbChIP](#), [hiReadsProcessor](#), [iPAC](#), [JASPAR2014](#), [kebabs](#), [methVisual](#), [minfi](#), [MotifDb](#), [motifRG](#), [oligo](#), [oneChannelGUI](#)

Red arrows point from the numbered labels to the following sections in the details table:

- ① points to the "biocViews" row.
- ② points to the "Depends" row.
- ③ points to the "Imports" row.

Bioconductor概観

①SequenceMatchingに含まれる17パッケージの一部しか表示されていないが、(Rで)塩基配列解析中にはない②CRISPR関連のパッケージなども存在することに気づく。また、③ゲノム配列パッケージBSgenomeもあることがわかる。

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Search:

Home » BiocViews

All Packages

Bioconductor version 3.0 (Release)

Autocomplete biocViews search:

Package	Maintainer	Title
Biostrings	H. Pages	String objects representing biological sequences, and matching algorithms
BSgenome	H. Pages	Infrastructure for Biostrings-based genome data packages
cleanUpdTSeq	Sarah Sheppard; Jianhong Ou; Lihua Julie Zhu	This package classifies putative polyadenylation sites as true or false/internally oligodT primed.
cobindR	Manuela Benary	Finding Co-occurring motifs of transcription factor binding sites
CRISPRseek	Lihua Julie Zhu	Design of target-specific guide RNAs in CRISPR-Cas9, genome-editing systems
dagLogo	Jianhong Ou	dagLogo
FunciSNP	Simon G. Coetzee	Integrating Functional Non-coding Datasets with Genetic Association Studies to Identify Candidate Regulatory SNPs
hapFabia	Sepp Hochreiter	hapFabia: Identification of very short segments of identity by descent (IBD) characterized by rare variants in large sequencing data

Packages found under SequenceMatching:

Show All entries Search table:

Package	Maintainer	Title
Biostrings	H. Pages	String objects representing biological sequences, and matching algorithms
BSgenome	H. Pages	Infrastructure for Biostrings-based genome data packages
cleanUpdTSeq	Sarah Sheppard; Jianhong Ou; Lihua Julie Zhu	This package classifies putative polyadenylation sites as true or false/internally oligodT primed.
cobindR	Manuela Benary	Finding Co-occurring motifs of transcription factor binding sites
CRISPRseek	Lihua Julie Zhu	Design of target-specific guide RNAs in CRISPR-Cas9, genome-editing systems
dagLogo	Jianhong Ou	dagLogo
FunciSNP	Simon G. Coetzee	Integrating Functional Non-coding Datasets with Genetic Association Studies to Identify Candidate Regulatory SNPs
hapFabia	Sepp Hochreiter	hapFabia: Identification of very short segments of identity by descent (IBD) characterized by rare variants in large sequencing data

Bioconductor概観

ゲノム配列パッケージBSgenomeは、①内部的にBiostringsパッケージを利用していることがわかる。

1

BSgenome

Infrastructure for Biostrings-based genome data packages

Bioconductor version: Release (3.0)

Infrastructure shared by all the Biostrings-based genome data packages

Author: Herve Pages

Maintainer: H. Pages <hpages at fhcrc.org>

Citation (from within R, enter `citation("BSgenome")`):

Pages H. *BSgenome: Infrastructure for Biostrings-based genome data packages*. R package version 1.34.1.

Installation

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("BSgenome")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

Workflows »

Common Bioconductor workflows include:

- [Oligonucleotide Arrays](#)
- [High-throughput Sequencing](#)
- [Counting Reads for Differential Expression](#) (parathyroideSE vignette)
- [Annotation](#)
- [Annotating Variants](#)
- [Annotating Ranges](#)
- [Flow Cytometry and other assays](#)
- [Candidate Binding Sites for Known Transcription Factors](#)
- [Cloud-enabled cis-eQTL search and annotation](#)
- [RNA-Seq workflow: gene-level exploratory analysis and differential expression](#)
- [Changing genomic coordinate systems with rtracklayer::liftOver](#)
- [Mass spectrometry and proteomics data analysis](#)

Mailing Lists »

Post questions about Bioconductor packages to our mailing lists. Read the [posting guide](#) before posting!

Bioconductor概観

The screenshot shows a web browser displaying the documentation for the BSgenome package. The URL is <http://bioconductor.org/packages/release/bioc/html/BSgenome.html>. The page includes sections for Documentation, Details, and various package metadata like biocViews, Version, In, License, Depends, Imports, Suggests, System Requirements, and URL.

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("BSgenome")
```

Details

biocViews: Annotation, DataRepresentation, Genetics, Infrastructure, SNP, SequenceMatching, Software

Version: 1.34.1

In Bioconductor: BioC 1.9 (R-2.4)

since

License: Artistic-2.0

Depends: R (>= 2.8.0), methods, BiocGenerics(>= 0.1.2), S4Vectors(>= 0.0.7), IRanges(>= 1.99.1), GenomeInfoDb(>= 1.1.4), GenomicRanges(>= 1.17.15), Biostrings(>= 2.33.3), rtracklayer(>= 1.25.8)

Imports: methods, stats, BiocGenerics, S4Vectors, IRanges, XVector, GenomeInfoDb, GenomicRanges, Biostrings, Rsamtools, rtracklayer

Suggests: BiocInstaller, BSgenome.Celegans.UCSC.ce2(>= 1.3.11), BSgenome.Hsapiens.UCSC.hg19(>= 1.3.11), BSgenome.Hsapiens.UCSC.hg19.masked, BSgenome.Rnorvegicus.UCSC.rn5, SNPLocs.Hsapiens.dbSNP.20100427, hqu95av2probe, Biobase, RUnit

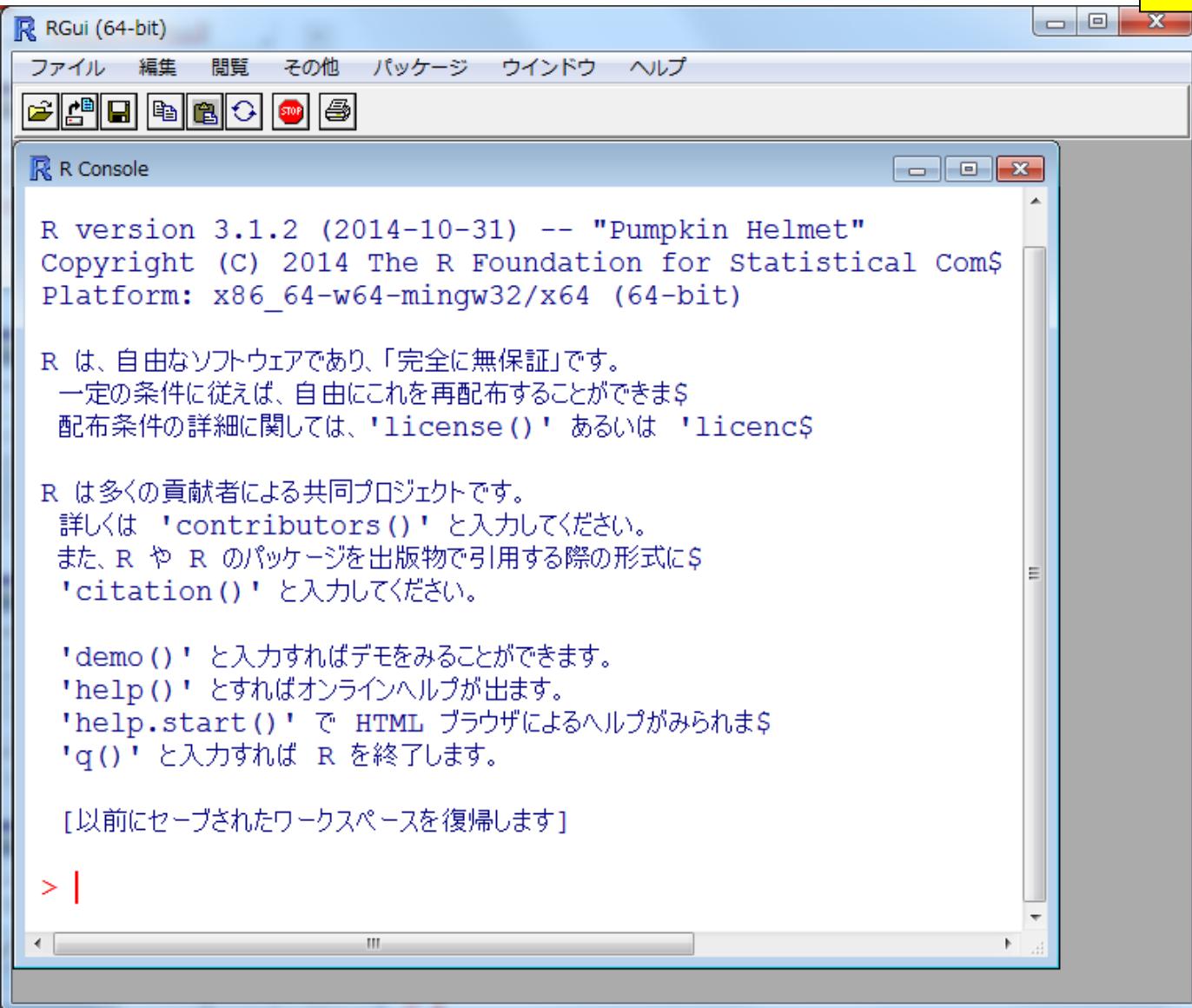
System Requirements

URL: BSgenome.Alvyrata.JGI.v1, BSgenome.Amelifera.BeeBase.assembly4, BSgenome.Amelifera.UCSC.apiMel2, BSgenome.Amelifera.UCSC.apiMel2.masked, BSgenome.Athaliana.TAIR.04232008, BSgenome.Athaliana.TAIR.TAIR9, BSgenome.Btaurus.UCSC.bosTau3, BSgenome.Btaurus.UCSC.bosTau3.masked, BSgenome.Btaurus.UCSC.bosTau4, BSgenome.Btaurus.UCSC.bosTau4.masked, BSgenome.Btaurus.UCSC.bosTau6, BSgenome.Btaurus.UCSC.bosTau6.masked,

ゲノム配列パッケージBSgenomeのちょっと下のほうに移動。Depends(依存)のところにBiostringsが存在することがわかる。BSgenomeパッケージを利用したい場合には、予めこれらの依存関係のあるパッケージ群のインストールが完了している必要がある。推奨手順通りにパッケージのインストールをしていればBSgenomeを問題なく利用できるはずである。

Bioconductor概観

BSgenomeを問題なく利用できるか
は、library(BSgenome)が通るかどうかで判断。R Guiの新規画面を起動。



Bioconductor概観

library(BSgenome)の実行結果中にエラーメッセージが出ていなければOK

R GUI (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ

R Console

```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます
配布条件の詳細に関しては、'license()' あるいは 'licenc$

R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の形式に '$
'citation()' と入力してください。

'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみられます
'q()' と入力すれば R を終了します。

[以前にセーブされたワークスペースを復帰します]

> library(BSgenome)
```

R Console

```
> library(BSgenome)
要求されたパッケージ BiocGenerics をロード中です
要求されたパッケージ parallel をロード中です

次のパッケージを付け加えます: 'BiocGenerics'

The following objects are masked from 'package:parallel$

clusterApply, clusterApplyLB, clusterCall,
clusterEvalQ, clusterExport, clusterMap,
parApply, parCapply, parLapply, parLapplyLB,
parRapply, parSapply, parSapplyLB

The following object is masked from 'package:stats$

xtabs

The following objects are masked from 'package:base$

anyDuplicated, append, as.data.frame,
as.vector, cbind, colnames, do.call,
duplicated, eval, evalq, Filter, Find, get,
intersect, is.unsorted, lapply, Map, mapply,
match, mget, order, paste, pmax, pmax.int,
pmin, pmin.int, Position, rank, rbind,
Reduce, rep.int, rownames, sapply, setdiff,
sort, table, tapply, union, unique, unlist

要求されたパッケージ IRanges をロード中です
要求されたパッケージ GenomicRanges をロード中です
要求されたパッケージ GenomeInfoDb をロード中です
要求されたパッケージ Biostrings をロード中です
要求されたパッケージ XVector をロード中です

> |
```

library(XXX)をやったときに、XXXパッケージ内部で利用するDependsやImportsにリストアップされているパッケージも同時にロードしている(読み込んでいる)。

Bioconductor概観

http://bioconductor.org/packages/release/bioc/html/BSgenome.html

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("BSgenome")
```

[PDF](#) [R Script](#) Efficient genome searching with Biostrings and the BSgenome data packages
[PDF](#) [R Script](#) How to forge a BSgenome data package
[PDF](#) Reference Manual
[Text](#) NEWS

Details

biocViews [Annotation](#), [DataRepresentation](#), [Genetics](#), [Infrastructure](#), [SNP](#), [SequenceMatching](#), [Software](#)
Version 1.34.1
In Bioconductor BioC 1.9 (R-2.4)
since
License Artistic-2.0
Depends R (>= 2.8.0), methods, [BiocGenerics](#)(>= 0.1.2), [S4Vectors](#)(>= 0.0.7), [IRanges](#)(>= 1.99.1), [GenomeInfoDb](#)(>= 1.1.4), [GenomicRanges](#)(>= 1.17.15), [Biostrings](#)(>= 2.33), [tracklayer](#)(>= 1.25.8)
Imports methods, stats, [BiocGenerics](#), [S4Vectors](#), [IRanges](#), [XVector](#), [GenomeInfoDb](#), [GenomicRanges](#), [Biostrings](#), [Rsamtools](#), [tracklayer](#), [BiocInstaller](#), [BSgenome.Celegans.UCSC.ce2](#)(>= 1.3.11), [BSgenome.Hsapiens.UCSC.hg19](#)(>= 1.3.11), [BSgenome.Hsapiens.UCSC.hg19.maske](#), [BSgenome.Rnorvegicus.UCSC.rn5](#), [SNPLocs.Hsapiens.dbSNP.20100427](#), [hqu95av2pro](#), [Biobase](#), [RUnit](#)
Suggests
System Requirements
URL [BSgenome.Alvratia.JGI.v1](#), [BSgenome.Amelifera.BeeBase.assembly4](#), [BSgenome.Amelifera.UCSC.apiMel2](#), [BSgenome.Amelifera.UCSC.apiMel2.masked](#), [BSgenome.Athaliana.TAIR.04232008](#), [BSgenome.Athaliana.TAIR.TAIR9](#), [BSgenome.Btaurus.UCSC.bosTau3](#), [BSgenome.Btaurus.UCSC.bosTau3.masked](#), [BSgenome.Btaurus.UCSC.bosTau4](#), [BSgenome.Btaurus.UCSC.bosTau4.masked](#), [BSgenome.Btaurus.UCSC.bosTau6](#), [BSgenome.Btaurus.UCSC.bosTau6.masked](#)

> library(BSgenome)

要求されたパッケージ BiocGenerics をロード中です ←
要求されたパッケージ parallel をロード中です ←

次のパッケージを付け加えます: 'BiocGenerics'

The following objects are masked from 'package:parallel\$'

clusterApply, clusterApplyLB, clusterCall,
clusterEvalQ, clusterExport, clusterMap,
parApply, parCapply, parLapply, parLapplyLB,
parRapply, parSapply, parSapplyLB

The following object is masked from 'package:stats\$':

xtabs

The following objects are masked from 'package:base\$':

anyDuplicated, append, as.data.frame,
as.vector, cbind, colnames, do.call,
duplicated, eval, evalq, Filter, Find, get,
intersect, is.unsorted, lapply, Map, mapply,
match, mget, order, paste, pmax, pmax.int,
pmin, pmin.int, Position, rank, rbind,
Reduce, rep.int, rownames, sapply, setdiff,
sort, table, tapply, union, unique, unlist

要求されたパッケージ IRanges をロード中です ←

要求されたパッケージ GenomicRanges をロード中です ←

要求されたパッケージ GenomeInfoDb をロード中です ←

要求されたパッケージ Biostrings をロード中です ←

要求されたパッケージ XVector をロード中です ←

> |

Bioconductor概観

もう一度library(BSgenome)をやってもメッセージは出ません。エラーメッセージが出ていなければ特に気にする必要はありません。

```
R Console

The following object is masked from 'package:stats':
  xtabs

The following objects are masked from 'package:base':
  anyDuplicated, append, as.data.frame,
  as.vector, cbind, colnames, do.call,
  duplicated, eval, evalq, Filter, Find, get,
  intersect, is.unsorted, lapply, Map, mapply,
  match, mget, order, paste, pmax, pmax.int,
  pmin, pmin.int, Position, rank, rbind,
  Reduce, rep.int, rownames, sapply, setdiff,
  sort, table, tapply, union, unique, unlist

要求されたパッケージ IRanges をロード中です
要求されたパッケージ GenomicRanges をロード中です
要求されたパッケージ GenomeInfoDb をロード中です
要求されたパッケージ Biostrings をロード中です
要求されたパッケージ XVector をロード中です
> library(BSgenome)
> |
```

Bioconductor概観

先にBSgenomeが内部的に利用しているBiostringsパッケージのロードを行っておくと、若干見栄えが異なるが、エラーメッセージが出ていないことさえ確認できれば問題ない。

R Gui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ

R Console

```
R version 3.1.2 (2014-10-31)
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32

R は、自由なソフトウェアであり、「完全に無保証」の状態で提供されています。一定の条件に従えば、自由にこれを再配布することができます。配布条件の詳細に関しては、「license」を参照してください。
```

R は多くの貢献者による共同プロジェクトです。詳しくは「contributors()」と入力してください。また、R や R のパッケージを出版物で引用する場合は「citation()」と入力してください。

'demo()' と入力すればデモをみることができます。'help()' とすればオンラインヘルプが出ます。'help.start()' で HTML ブラウザを開くことができます。'q()' と入力すれば R を終了します。

[以前にセーブされたワークスペースを復帰します]

```
> library(Biostrings)
```

R Console

```
clusterExport, clusterMap, parApply, parCapply, parLapply,
parLapplyLB, parRapply, parSapply, parSapplyLB

The following object is masked from 'package:stats':
xtabs

The following objects are masked from 'package:base':
anyDuplicated, append, as.data.frame, as.vector, cbind,
colnames, do.call, duplicated, eval, evalq, Filter, Find,
get, intersect, is.unsorted, lapply, Map, mapply, match,
mget, order, paste, pmax, pmax.int, pmin, pmin.int,
Position, rank, rbind, Reduce, rep.int, rownames, sapply,
setdiff, sort, table, tapply, union, unique, unlist
```

要求されたパッケージ IRanges をロード中です
要求されたパッケージ XVector をロード中です

```
> library(BSgenome)
```

要求されたパッケージ GenomicRanges をロード中です
要求されたパッケージ GenomeInfoDb をロード中です

```
> library(BSgenome)
> |
```

Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

BSgenome利用の意義

- 解析 | 機能解析 | パスウェイ(Pathway)解析 | SeqGSEA(Wang 2014) (last modified 2014/12/19)
- 解析 | 菌叢解析 | について (last modified 2014/12/19)
- 解析 | 菌叢解析 | phyloseq(McMurdie 2013) (last modified 2014/05/29)
- 解析 | エクソーム解析 | について (last modified 2014/05/29)
- 解析 | ChIP-seq | について (last modified 2015/02/19)
- 解析 | ChIP-seq | DiffBind(Ross-Innes 2012) (last modified 2015/02/19)
- 解析 | ChIP-seq | ChIPseqR(Humburg 2011) (last modified 2015/02/19)
- 解析 | ChIP-seq | chipseq (last modified 2011/12/19)
- 解析 | ChIP-seq | PICS(Zhang 2011) (last modified 2011/12/19)

解析 | ChIP-seq | について

このあたりはほとんどノータッチです。SraTailor(Oki et al., 2014)は、実験医学2014年12月号の「Close Up実験法」中で日本語による解説記事があります(沖 真弥氏提供情報)。2015年2月に調査した結果をリストアップします。

R用:

- ChIPsim: Zhang et al., PLoS Comput. Biol., 2008
- PeakSeq法: Rozowsky et al., Nat Biotechnol., 2009
- CSAR: Kaufmann et al., PLoS Biol., 2009
- rMAT: Droit et al., Bioinformatics, 2010
- ChIPpeakAnno: Zhu et al., BMC Bioinformatics, 2010
- PICS: Zhang et al., Biometrics, 2011
- ChIPseqR: Humburg et al., BMC Bioinformatics, 2011
- DiffBind: Ross-Innes et al., Nature, 2012
- MEDIPS: Lienhard et al., Bioinformatics, 2014
- DSS: Feng et al., Nucleic Acids Res., 2014
- methylSig: Park et al., Bioinformatics, 2014

R以外:

- bwtool: Pohl and Beato, Bioinformatics, 2014
- SraTailor: Oki et al., Genes Cells., 2014

Review、ガイドライン、バイブル系:

- ガイドライン: Bailey et al., PLoS Comput Biol., 2013
- Review: Robinson et al., Front Genet., 2014

ゲノム配列情報はUCSCやEnsemblなどのウェブサイトから取得するのが一般的ではあるが、Rの生物種ごとに提供されているBSgenomeで取得、あるいは取り扱うことも可能。ChIP-seq用パッケージMEDIPSはBSgenomeを利用。

BSgenome

プロモーター配列取得ではなく、
①ゲノム配列取得のほうです！

- ・ イントロ | 一般 | [任意の長さの連続塩基の出現頻度情報を取得](#) (last modified 2013/06/14)
- ・ イントロ | 一般 | [Tips | 任意の拡張子でファイルを保存](#) (last modified 2013/09/26)
- ・ イントロ | 一般 | [Tips | 拡張子は同じで任意の文字を追加して保存](#) (last modified 2013/09/26)
- ・ イントロ | 一般 | [配列取得 | ゲノム配列 | 公共DBから](#) (last modified 2014/05/28)
- ・ イントロ | 一般 | [配列取得 | ゲノム配列 | BSgenome](#) (last modified 2015/02/18) NEW
- ・ イントロ | 一般 | [配列取得 | プロモーター配列 | 公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | [配列取得 | プロモーター配列 | BSgenome](#) (last modified 2014/04/25)
- ・ イントロ | 一般 | [配列取得 | プロモーター配列 | GenomicFeatures\(Lawrence 2013\)](#) (last modified 2014/04/23)
- ・ イントロ | 一般 | [配列取得 | トランスクレプション](#)
- ・ イントロ | 一般 | [配列取得 | トランスクレプション](#)
- ・ イントロ | NGS | [様々なプラットフォーム](#)
- ・ イントロ | NGS | [qPCRやmicroarrayなど](#)



イントロ | 一般 | 配列取得 | ゲノム配列 | BSgenome NEW

BSgenomeパッケージを用いて様々な生物種のゲノム配列を取得するやり方を示します。ミヤマハタザオ (*A. lyrata*)、セイヨウミツバチ (*A. mellifera*)、シロイヌナズナ (*A. thaliana*)、ウシ (*B. taurus*)、線虫 (*C. elegans*)、犬 (*C. familiaris*)、キイロショウジョウバエ (*D. melanogaster*)、ゼブラフィッシュ (*D. rerio*)、大腸菌 (*E. coli*)、イトヨリ (*G. aculeatus*)、セキショクヤケイ (*G. gallus*)、ヒト (*H. sapiens*)、アカゲザル (*M. mulatta*)、マウス (*M. musculus*)、チンパンジー (*P. troglodytes*)、ラット (*R. norvegicus*)、出芽酵母 (*S. cerevisiae*)、トキソプラズマ (*T. gondii*)と実際に様々な生物種が利用可能であることがわかります。getSeq関数はBSgenomeオブジェクト中の「single sequences」というあたりにリストアップされているchr...というものを全て抽出しています。したがって、例えばマウスゲノムは「chr1」以外に「chr1_random」や「chrUn_random」なども等価に取扱っている点に注意してください。
「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに移動し以下をコピペ。

1. 利用可能な生物種とRにインストール済みの生物種をリストアップしたい場合:

```
#必要なパッケージをロード
library(BSgenome) #パッケージの読み込み

#本番（利用可能なパッケージをリストアップ；インストール済みとは限らない）
available.genomes() #このパッケージ中で利用可能なゲノムをリストアップ

#本番（インストール済みの生物種をリストアップ）
installed.genomes() #インストール済みの生物種をリストアップ

#後処理（パッケージ名でだいたいわかるがproviderやversionを分割して表示したい場合）
installed.genomes(splitNameParts=TRUE) #インストール済みの生物種をリストアップ
```

BSgenome

イントロ | 一般 | 配列取得 | ゲノム配列 | BSgenome NEW

[BSgenome](#)パッケージを用いて様々な生物種のゲノム配列を取得するやり方を示します。ミヤマバ
lyrata)、セイヨウミツバチ (*A. mellifera*)、シロイヌナズナ (*A. thaliana*)、ウシ (*B. taurus*)、線虫 (*C. elegans*)、
(*C. familiaris*)、キイロショウジョウバエ (*D. melanogaster*)、ゼブラフィッシュ (*D. rerio*)、大腸菌 (*E. coli*)
(*G. aculeatus*)、セキショクヤケイ (*G. gallus*)、ヒト (*H. sapiens*)、アカゲザル (*M. mulatta*)、マウス (*M. musculus*)、チ
ンパンジー (*P. troglodytes*)、ラット (*R. norvegicus*)、出芽酵母 (*S. cerevisiae*)、トキソプラズマ (*T. gondii*)と実際に様々な
な生物種が利用可能であることがわかります。`getSeq`関数は `BSgenome`オブジェクト中の「single sequences」という
あたりにリストアップされている `chr...` というものを全て抽出しています。したがって、例えばマウスゲノムは「`chr1`」
以外に「`chr1_random`」や「`chrUn_random`」なども等価に取扱っている。

黒枠部分のコードをコピペ。R ver. 3.1.3
(Bioconductor ver. 3.0)で利用可能な生物
種のパッケージ名をリストアップ。71個あ
ることが分かる。Rのバージョンが古いと
パッケージ数は少なくなる。例えばR ver.
3.2.0 (Bioconductor ver. 3.1)では77個。

「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに
「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに

1. 利用可能な生物種とRにインストール済みの生物種をリストアップ

```
#必要なパッケージをロード
library(BSgenome)
#本番（利用可能なパッケージをリストアップ；インストール
available.genomes() #このパッケージ

#本番（インストール済みの生物種をリストアップ）
installed.genomes() #インストール済みの生物種をリストアップ

#後処理（パッケージ名でだいたいわかるがproviderやversionを確認）
installed.genomes(splitNameParts=TRUE) #インストール済みの生物種をリストアップ
```

```
[56] "BSgenome.Ptroglodytes.UCSC.panTro2"
[57] "BSgenome.Ptroglodytes.UCSC.panTro2.masked"
[58] "BSgenome.Ptroglodytes.UCSC.panTro3"
[59] "BSgenome.Ptroglodytes.UCSC.panTro3.masked"
[60] "BSgenome.Rnorvegicus.UCSC.rn4"
[61] "BSgenome.Rnorvegicus.UCSC.rn4.masked"
[62] "BSgenome.Rnorvegicus.UCSC.rn5"
[63] "BSgenome.Rnorvegicus.UCSC.rn5.masked"
[64] "BSgenome.Scerevisiae.UCSC.sacCer1"
[65] "BSgenome.Scerevisiae.UCSC.sacCer2"
[66] "BSgenome.Scerevisiae.UCSC.sacCer3"
[67] "BSgenome.Sscrofa.UCSC.susScr3"
[68] "BSgenome.Sscrofa.UCSC.susScr3.masked"
[69] "BSgenome.Tgondii.ToxoDB.7.0"
[70] "BSgenome.Tguttata.UCSC.taeGut1"
[71] "BSgenome.Tguttata.UCSC.taeGut1.masked"
> |
```

```
> available.genomes()
```

```
[1] "BSgenome.Alyrata.JGI.v1"
[2] "BSgenome.Amellifera.BeeBase.asse
[3] "BSgenome.Amellifera.UCSC.apIMel2
[4] "BSgenome.Amellifera.UCSC.apIMel2
[5] "BSgenome.Athaliana.TAIR.04232008
[6] "BSgenome.Athaliana.TAIR.TAIR9"
[7] "BSgenome.Btaurus.UCSC.bosTau3"
[8] "BSgenome.Btaurus.UCSC.bosTau3.ma
[9] "BSgenome.Btaurus.UCSC.bosTau4"
[10] "BSgenome.Btaurus.UCSC.bosTau4.ma
[11] "BSgenome.Btaurus.UCSC.bosTau6"
[12] "BSgenome.Btaurus.UCSC.bosTau6.ma
[13] "BSgenome.Celegans.UCSC.ce10"
[14] "BSgenome.Celegans.UCSC.ce2"
[15] "BSgenome.Celegans.UCSC.ce6"
[16] "BSgenome.Cfamiliaris.UCSC.canFam
[17] "BSgenome.Cfamiliaris.UCSC.canFam
[18] "BSgenome.Cfamiliaris.UCSC.canFam
[19] "BSgenome.Cfamiliaris.UCSC.canFam
[20] "BSgenome.Dmelanogaster.UCSC.dm2"
[21] "BSgenome.Dmelanogaster.UCSC.dm2.
[22] "BSgenome.Dmelanogaster.UCSC.dm3"
[23] "BSgenome.Dmelanogaster.UCSC.dm3.
[24] "BSgenome.Drerio.UCSC.danRer5"
[25] "BSgenome.Drerio.UCSC.danRer5.mas
[26] "BSgenome.Drerio.UCSC.danRer6"
[27] "BSgenome.Drerio.UCSC.danRer6.mas
[28] "BSgenome.Drerio.UCSC.danRer7"
```

赤枠のヒトゲノムは4バージョン(hg17, 18, 19, GRCh38)を利用可能。2013年12月にリリースされた①最新版(GRCh38)のRパッケージも利用可能。

```
[29] "BSgenome.Hsapiens.NCBI.GRCh38"
[30] "BSgenome.Hsapiens.NCBI.20080805"
[31] "BSgenome.Gaculeatus.UCSC.gasAcul"
[32] "BSgenome.Gaculeatus.UCSC.gasAcul.masked"
[33] "BSgenome.Ggallus.UCSC.galGal3"
[34] "BSgenome.Ggallus.UCSC.galGal3.masked"
[35] "BSgenome.Ggallus.UCSC.galGal4"
[36] "BSgenome.Ggallus.UCSC.galGal4.masked"
[37] "BSgenome.Hsapiens.NCBI.GRCh38" ①
[38] "BSgenome.Hsapiens.UCSC.hg17"
[39] "BSgenome.Hsapiens.UCSC.hg17.masked"
[40] "BSgenome.Hsapiens.UCSC.hg18"
[41] "BSgenome.Hsapiens.UCSC.hg18.masked"
[42] "BSgenome.Hsapiens.UCSC.hg19"
[43] "BSgenome.Hsapiens.UCSC.hg19.masked"
[44] "BSgenome.Mfuro.UCSC.musFur1"
[45] "BSgenome.Mmulatta.UCSC.rheMac2"
[46] "BSgenome.Mmulatta.UCSC.rheMac2.masked"
[47] "BSgenome.Mmulatta.UCSC.rheMac3"
[48] "BSgenome.Mmulatta.UCSC.rheMac3.masked"
[49] "BSgenome.Mmusculus.UCSC.mm10"
[50] "BSgenome.Mmusculus.UCSC.mm10.masked"
[51] "BSgenome.Mmusculus.UCSC.mm8"
[52] "BSgenome.Mmusculus.UCSC.mm8.masked"
[53] "BSgenome.Mmusculus.UCSC.mm9"
[54] "BSgenome.Mmusculus.UCSC.mm9.masked"
[55] "BSgenome.Osativa.MSU.MSU7"
```

BSgenome

イントロ | 一般 | 配列取得 | ゲノム配列 | BSgenome NT

BSgenomeパッケージを用いて様々な生物種のゲノム配列を取得するやり方を示します。人間(Homo sapiens)、マウス(Mus musculus)、大鼠(Rattus norvegicus)、酵母(Saccharomyces cerevisiae)、カエル(Xenopus laevis)、カブトムシ(Drosophila melanogaster)、ヒト細胞株(Hela)など、多くの生物種が利用可能になります。getSeq関数はBSgenomeオブジェクト中の「single sequences」というあたりにリストアップされているchr...というものを全て抽出しています。したがって、例えばマウスゲノムは「chr1」以外に「chr1_random」や「chrUn_random」なども等価に取扱っている点に注意してください。

「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに移動し以下をコピペ。

1. 利用可能な生物種とRにインストール済みの生物種をリストアップしたい場合:

```
#必要なパッケージをロード
library(BSgenome) #パッケージの読み込み

#本番（利用可能なパッケージをリストアップ；インストール済みとは別）
available.genomes() #このパッケージ中に含まれる生物種をリストアップ

#本番（インストール済みの生物種をリストアップ）
installed.genomes() #インストール済みの生物種をリストアップ

#後処理（パッケージ名でだいたいわかるがproviderやversionを確認）
installed.genomes(splitNameParts=TRUE) #インストール済みの生物種を確認
```

```
R Console
[68] "BSgenome.Sscrofa.UCSC.susScr3.masked"
[69] "BSgenome.Tgondii.ToxoDB.7.0"
[70] "BSgenome.Tguttata.UCSC.taeGut1"
[71] "BSgenome.Tguttata.UCSC.taeGut1.masked"
> installed.genomes() #インストール済みの生物種をリストアップ
[1] "BSgenome.Athaliana.TAIR.TAIR9"
[2] "BSgenome.Celegans.UCSC.ce2"
[3] "BSgenome.Celegans.UCSC.ce6"
[4] "BSgenome.Drerio.UCSC.danRer7"
[5] "BSgenome.Hsapiens.NCBI.GRCh38" #1
[6] "BSgenome.Hsapiens.UCSC.hg19"
[7] "BSgenome.Mmusculus.UCSC.mm10"
> |
```

Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

BSgenome

イントロ | 一般 | 配列取得 | ゲノム配列 | BSgenome NEW

BSgenomeパッケージを用いて様々な生物種のゲノム配列を取得するやり方を示します。 lyrata)、セイヨウミツバチ (A. mellifera)、シロイヌナズナ (A. thaliana)、ウシ (B. taurus)、(C. familiaris)、キイロショウジョウバエ (D. melanogaster)、ゼブラフィッシュ (D. rerio)、(G. aculeatus)、セキショクヤケイ (G. gallus)、ヒト (H. sapiens)、アカゲザル (M. mulatta)、マウス (M. musculus)、チンパンジー (P. troglodytes)、ラット (R. norvegicus)、出芽酵母 (S. cerevisiae)、トキソプラズマ (T. gondii) と実際に様々な生物種が利用可能であることがわかります。getSeq関数はBSgenomeオブジェクト中の「single sequences」というあたりリストアップされているchrと、それを全て抽出してしまってからデータ例 (ゲノム) が残る「chr1」

以外に「chr1_random」

①

9. インストール済みのヒト ("BSgenome.Hsapiens.NCBI.GRCh38") のゲノム配列を multi-FASTA ファイルで保存したい場合:

2013年12月にリリースされた Genome Reference Consortium GRCh38 です。R ver. 3.1.0 と Bioconductor ver. 2.14 以上の環境で実行可能です。

1. 利用可能な生物種とR

```
#必要なパッケージを
library(BSgenome)

#本番 (利用可能なパ
available.genomes

#本番 (インストール
installed.genomes

#後処理 (パッケージ
installed.genomes
```

```
out_f <- "hoge9.fasta" #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定

#必要なパッケージをロード
library(param, character.only=T) #paramで指定したパッケージの読み込み

#前処理 (paramで指定したパッケージ中のオブジェクト名をgenomeに統一)
#tmp <- unlist(strsplit(param, ".", fixed=TRUE))[2] #paramで指定した文字列からオブジェクト名を取得し
tmp <- ls(paste("package", param, sep=":")) #paramで指定したパッケージで利用可能なオブジェクト名を取得
genome <- eval(parse(text=tmp)) #文字列tmpをRオブジェクトとしてgenomeに格納 (パッケージ中には)
#確認してるだけです

#本番
fasta <- getSeq(genome) #ゲノム塩基配列情報を抽出した結果をfastaに格納
names(fasta) <- seqnames(genome) #description情報を追加している

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50) #fastaの中身を指定したファイル名で保存
```

①例題9。2013年12月にリリースされたヒトゲノム最新版(GRCh38)のRパッケージを入力、multi-FASTAファイルを出力として得る。作業ディレクトリはどこでもよいが基本はデスクトップ上のhoge。数分かかるが、約3.3GBのファイルが生成される。決してテキストエディタで開かないで!

BSgenome

①出力ファイルの内容は、fastaオブジェクトに格納されている。慣れれば②fastaオブジェクトの中身を眺めるほうが全体像をつかみやすい。

9. インストール済みのヒト("BSgenome.Hsapiens.NCBI.GRCh38")のゲノム配列をmulti-FASTAファイルで保存したい場合:

2013年12月にリリースされたGenome Reference Consortium GRCh38です。R ver. 3.1.0とBioconductor ver. 2.14以上の環境で実行可能です。

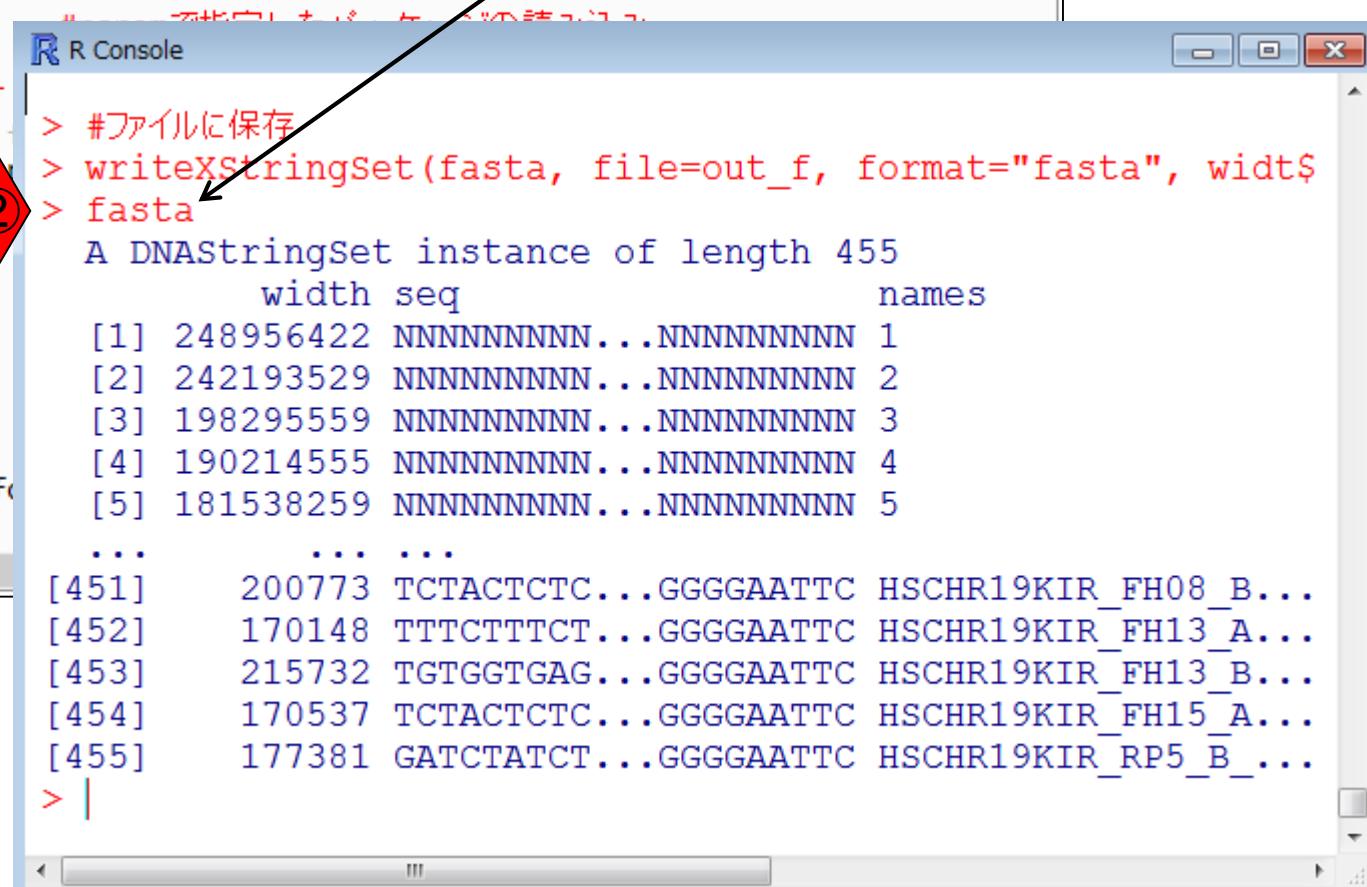
```
out_f <- "hoge9.fasta"          #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定

#必要なパッケージをロード
library(param, character.only=T)

#前処理(paramで指定したパッケージ中のオブジェクトをリスト化)
tmp <- unlist(strsplit(param, "."))
tmp <- ls(paste("package", param, sep=""))
genome <- eval(parse(text=tmp))
genome

#本番
fasta <- getSeq(genome)
names(fasta) <- seqnames(genome)

#ファイルに保存
writeXStringSet(fasta, file=out_f, fo
```



The screenshot shows the R Console window with the following session history:

```
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta", width=60)
> fasta
```

The output shows a DNAStringSet instance of length 455 with the following structure:

	width	seq	names
[1]	248956422	NNNNNNNN...NNNNNNNN	1
[2]	242193529	NNNNNNNN...NNNNNNNN	2
[3]	198295559	NNNNNNNN...NNNNNNNN	3
[4]	190214555	NNNNNNNN...NNNNNNNN	4
[5]	181538259	NNNNNNNN...NNNNNNNN	5
...	
[451]	200773	TCTACTCTC...GGGAAATT	HSCHR19KIR_FH08_B...
[452]	170148	TTTCTTTCT...GGGAAATT	HSCHR19KIR_FH13_A...
[453]	215732	TGTGGTGAG...GGGAAATT	HSCHR19KIR_FH13_B...
[454]	170537	TCTACTCTC...GGGAAATT	HSCHR19KIR_FH15_A...
[455]	177381	GATCTATCT...GGGAAATT	HSCHR19KIR_RP5_B...

BSgenome

①1～22番染色体のみ取扱いたい場合。
染色体番号の数が大きくなるほど配列長
が短くなっている傾向が一目瞭然ですね。

9. インストール済みのヒト("BSgenome.Hsapiens.NCBI.GRCh38")のゲノム配列をmulti-FASTAファイルで保存したい場合:

2013年12月にリリースされたGenome Reference Consortium GRCh38です。R ver. 3.1.0とBioconductor ver. 2.14以上の環境で実行可能です。

```
out_f <- "hoge9.fasta"          #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定
```

#必要なパッケージをロード

```
library(param, character.only=T)
```

#前処理(paramで指定したパッケージ中のオ

```
#tmp <- unlist(strsplit(param, "."))
tmp <- ls(paste("package", param, sep=""))
genome <- eval(parse(text=tmp))
genome
```

#本番

```
fasta <- getSeq(genome)
names(fasta) <- seqnames(genome)
```

#ファイルに保存

```
writeXStringSet(fasta, file=out_f, fo
```

R Console

	width	seq	names
[1]	248956422	NNNNNNNNNN...NNNNNNNNNN	1
[2]	242193529	NNNNNNNNNN...NNNNNNNNNN	2
[3]	198295559	NNNNNNNNNN...NNNNNNNNNN	3
[4]	190214555	NNNNNNNNNN...NNNNNNNNNN	4
[5]	181538259	NNNNNNNNNN...NNNNNNNNNN	5
...	
[18]	80373285	NNNNNNNNNN...NNNNNNNNNN	18
[19]	58617616	NNNNNNNNNN...NNNNNNNNNN	19
[20]	64444167	NNNNNNNNNN...NNNNNNNNNN	20
[21]	46709983	NNNNNNNNNN...NNNNNNNNNN	21
[22]	50818468	NNNNNNNNNN...NNNNNNNNNN	22

BSgenome

①X, Y, およびミトコンドリア配列も含めたい場合。②配列の並びの確認は試行錯誤。最初からわかっていたわけではありません。R画面上で眺めるほうが、全体像を把握しやすい。

9. インストール済みのヒト("BSgenome.Hsapiens.NCBI.GRCh38")のゲノム配列をmulti-FASTAファイルで保存したい場合:

2013年12月にリリースされたGenome Reference Consortium GRCh38です。R ver. 3.1.0とBioconductor ver. 2.14以上の環境で実行可能です。

```
out_f <- "hoge9.fasta"          #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定

#必要なパッケージをロード
library(param, character.only=T)

#前処理(paramで指定したパッケージ中のオ
#tmp <- unlist(strsplit(param, "."))
#tmp <- ls(paste("package", param, sep=""))
#genome <- eval(parse(text=tmp))
#genome

#本番
fastas <- getSeq(genome)
names(fasta) <- seqnames(genome)

#ファイルに保存
writeXStringSet(fasta, file=out_f, fo
```

width	seq	names
[1]	248956422 NNNNNNNNNN...NNNNNNNNNN	1
[2]	242193529 NNNNNNNNNN...NNNNNNNNNN	2
[3]	198295559 NNNNNNNNNN...NNNNNNNNNN	3
[4]	190214555 NNNNNNNNNN...NNNNNNNNNN	4
[5]	181538259 NNNNNNNNNN...NNNNNNNNNN	5
...	...	
[21]	46709983 NNNNNNNNNN...NNNNNNNNNN	21
[22]	50818468 NNNNNNNNNN...NNNNNNNNNN	22
[23]	156040895 NNNNNNNNNN...NNNNNNNNNN	X
[24]	57227415 NNNNNNNNNN...NNNNNNNNNN	Y
[25]	16569 GATCACAGGT...ATCACGATG	MT

BSgenome

9. インストール済みのヒト("BSgenome.Hsapiens.NCBI.GRCh38")のゲノム

2013年12月にリリースされたGenome Reference Consortium GRCh38です。R ver. 3.1.0とBioconductor ver. 2.14以上との環境で実行可能です。

```
out_f <- "hoge9.fasta"          #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定

#必要なパッケージをロード
library(param, character.only=T)

#前処理(paramで指定したパッケージ中のオブジェクト)
#tmp <- unlist(strsplit(param, ".", -1))
#tmp <- ls(paste("package", param, sep=""))
#genome <- eval(parse(text=tmp))
#genome

#本番
fastas <- getSeq(genome)
names(fastas) <- seqnames(genome)

#ファイルに保存
writeXStringSet(fasta, file=out_f, fo
```

X, Y, およびミトコンドリア配列までのサブセットをhoge10.fastaで保存したい場合。①上矢印キーを何回か押して、ファイルに保存するためのコマンドを出し、②と③の水色下線部分を変更すればよい。

R Console window showing DNA sequence data and R code.

```
R Console
[1] 46709983 NNNNNNNNNN...NNNNNNNNN 21
[2] 50818468 NNNNNNNNNN...NNNNNNNNN 22
> fasta[1:25]
A DNAStringSet instance of length 25
width seq names
[1] 248956422 NNNNNNNNNN...NNNNNNNNN 1
[2] 242193529 NNNNNNNNNN...NNNNNNNNN 2
[3] 198295559 NNNNNNNNNN...NNNNNNNNN 3
[4] 190214555 NNNNNNNNNN...NNNNNNNNN 4
[5] 181538259 NNNNNNNNNN...NNNNNNNNN 5
...
[21] 46709983 NNNNNNNNNN...NNNNNNNNN 21
[22] 50818468 NNNNNNNNNN...NNNNNNNNN 22
[23] 156040895 NNNNNNNNNN...NNNNNNNNN X
[24] 57227415 NNNNNNNNNN...NNNNNNNNN Y
[25] 16569 GATCACAGGT...ATCACCGATG MT
> writeXStringSet(fasta, file=out_f, format="fasta", width=100)
```

The code in the R console is:

```
> writeXStringSet(fasta, file=out_f, format="fasta", width=100)
```

Annotations with red arrows:

- ① A red arrow points from the top of the R console window to the command line area, indicating where to enter the command.
- ② A red arrow points to the underlined word "fasta" in the command line.
- ③ A red arrow points to the underlined word "file" in the command line.

BSgenome

②と③をこんな感じで変更。実行後にhoge9.fastaよりも若干ファイルサイズの小さいhoge10.fastaが生成されていることが確認できるはず。決してテキストエディタで開かないで!

9. インストール済みのヒト("BSgenome.Hsapiens.NCBI.GRCh38")のゲノム配列をmulti-FASTAファイルで保存したい場合:

2013年12月にリリースされたGenome Reference Consortium GRCh38です。R ver. 3.1.0とBioconductor ver. 2.14以上の環境で実行可能です。

```
out_f <- "hoge9.fasta"          #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定

#必要なパッケージをロード
library(param, character.only=T)

#前処理(paramで指定したパッケージ中のオ
#tmp <- unlist(strsplit(param, ".",
#tmp <- ls(paste("package", param, se
genome <- eval(parse(text=tmp)))
genome

#本番
fasta <- getSeq(genome)
names(fasta) <- seqnames(genome)

#ファイルに保存
writeXStringSet(fasta, file=out_f, fo
```

The screenshot shows the R Console window with the following content:

```
[22] 50818468 NNNNNNNNNN...NNNNNNNN 22
> fasta[1:25]
  A DNAStringSet instance of length 25
      width seq                         names
[1] 248956422 NNNNNNNNNN...NNNNNNNN 1
[2] 242193529 NNNNNNNNNN...NNNNNNNN 2
[3] 198295559 NNNNNNNNNN...NNNNNNNN 3
[4] 190214555 NNNNNNNNNN...NNNNNNNN 4
[5] 181538259 NNNNNNNNNN...NNNNNNNN 5
...
[21] 46709983 NNNNNNNNNN...NNNNNNNN 21
[22] 50818468 NNNNNNNNNN...NNNNNNNN 22
[23] 156040895 NNNNNNNNNN...NNNNNNNN X
[24] 57227415 NNNNNNNNNN...NNNNNNNN Y
[25] 16569 GATCACAGGT...ATCACGATG MT
> writeXStringSet(fasta[1:25], file="hoge10.fasta", forma$
```

Two red arrows point to specific parts of the code:

- ② points to the line `writeXStringSet(fasta[1:25], file="hoge10.fasta", forma$`.
- ③ points to the line `> |`.

①例題10。②様々な記述形式があります。やらないでいいです。決してテキストエディタで開かないで!

BSgenome

イントロ | 一般 | 配列取得 | ゲノム配列 | BSgenome NEW

BSgenomeパッケージを用いて様々な生物種のゲノム配列を取得するやり方を示します。ミヤマハタザオ (*A. lyra*) サイコウミツバチ (*A. mellifera*) シロイヌナズナ (*A. thaliana*) ウミノヒルメ (*B. taurus*) 緑茶 (*C. sativa*) 犬 (*D. melanogaster*) ハムスター (*E. carpio*) ハムガム (*F. rubra*) ハムサ (*G. aculeatus*) ハムレット (*H. sapiens*) ハムレット (*I. maculatus*) ハムレット (*M. musculus*) ハムレット (*O. latipes*) ハムレット (*S. pombe*) ハムレット (*T. brucei*) ハムレット (*Z. mays*)

10. インストール済みのヒト ("BSgenome.Hsapiens.NCBI.GRCh38") のゲノム配列のmulti-FASTAファイルで保存したい場合:

一部を抽出して保存するやり方です。このパッケージ中の染色体の並びが既知 (chr1, 2, ..., chr22, chrX, chrY, and MT) であるという前提です。

```
out_f <- "hoge10.fasta"          #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定
param_range <- 1:25               #抽出したい範囲を指定
```

#必要なパッケージを読み込む
library(param, character.only=T)
#前処理(paramで指定したパッケージ中の情報をtmpに格納)
#tmp <- unlist(strsplit(param, "."))
tmp <- ls(paste("package", param, sep="."))
genome <- eval(parse(text=tmp))
genome

#本番
fasta <- getSeq(genome)
names(fasta) <- seqnames(genome)

#後処理(フィルタリング)
obj <- param_range
fasta <- fasta[obj]
fasta

R Console

	width	seq	names
[1]	248956422	NNNNNNNNNNNN...NNNNNNNNNN	1
[2]	242193529	NNNNNNNNNNNN...NNNNNNNNNN	2
[3]	198295559	NNNNNNNNNNNN...NNNNNNNNNN	3
[4]	190214555	NNNNNNNNNNNN...NNNNNNNNNN	4
[5]	181538259	NNNNNNNNNNNN...NNNNNNNNNN	5
...	
[21]	46709983	NNNNNNNNNNNN...NNNNNNNNNN	21
[22]	50818468	NNNNNNNNNNNN...NNNNNNNNNN	22
[23]	156040895	NNNNNNNNNNNN...NNNNNNNNNN	X
[24]	57227415	NNNNNNNNNNNN...NNNNNNNNNN	Y
[25]	16569	GATCACAGGTC...CATCACGATG	MT
>			
> #ファイルに保存			
> writeXStringSet(fasta, file=out_f, format="fasta", width=\$			
>			

BSgenome

9. インストール済みのヒト ("BSgenome.Hsapiens.NCBI.GRCh38")

2013年12月にリリースされたGenome Reference Consortium GRCh38のデータをRで利用するためのパッケージです。

```
out_f <- "hoge9.fasta"                      #出力ファイル名を指定してout_fに格納
param <- "BSgenome.Hsapiens.NCBI.GRCh38" #パッケージ名を指定

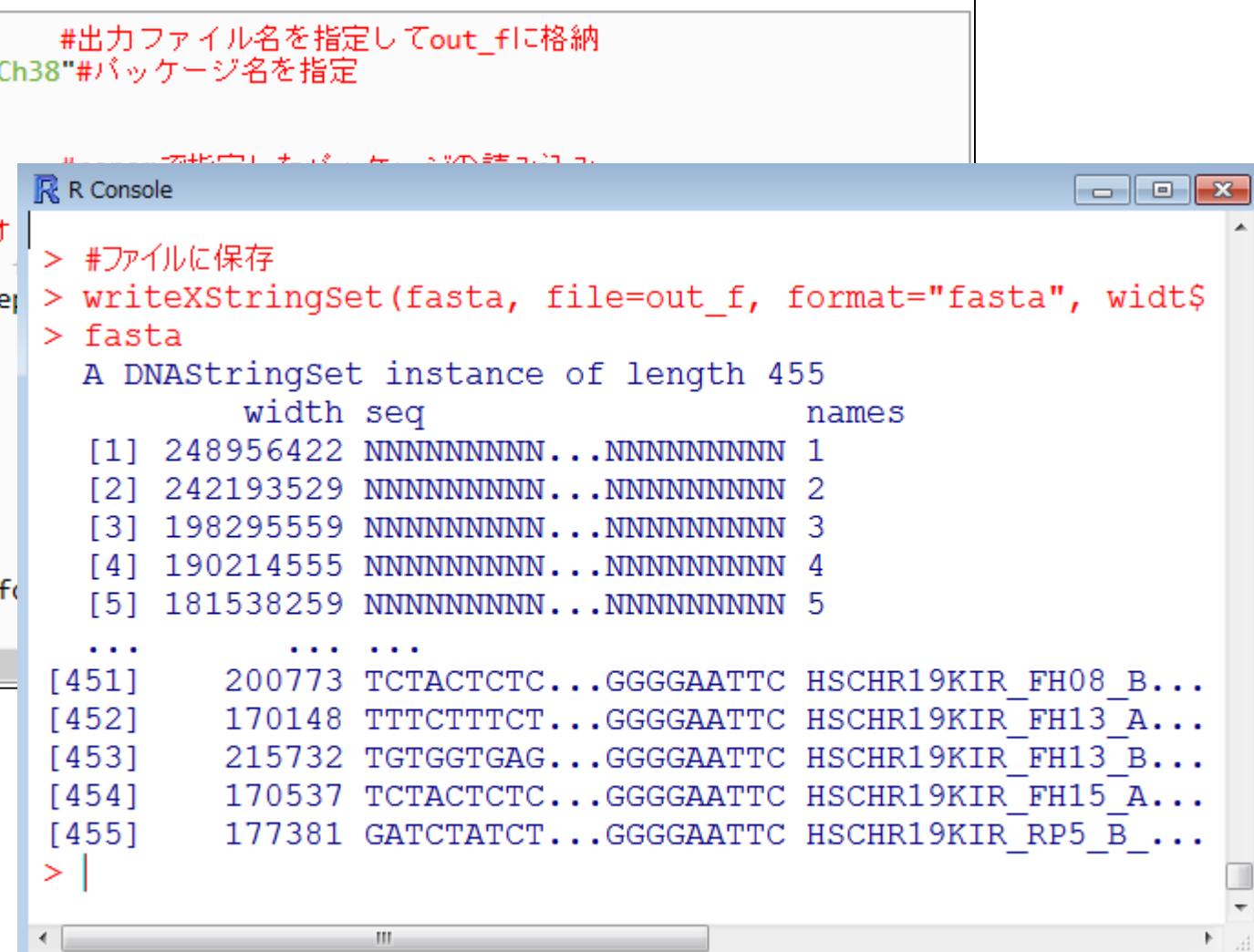
#必要なパッケージをロード
library(param, character.only=T)

#前処理(paramで指定したパッケージ中のオブジェクトを読み込む)
tmp <- unlist(strsplit(param, "."))
tmp <- ls(paste("package", param, sep=""))
genome <- eval(parse(text=tmp))
genome

#本番
fasta <- getSeq(genome)
names(fasta) <- seqnames(genome)

#ファイルに保存
writeXStringSet(fasta, file=out_f, fo
```

例題9。26番目以降の配列は、ヒトゲノムの一部ではあるものの、まだ割り当てられる染色体が定まっていないものたちです。メタゲノム解析などでヒトゲノムにマップされないリードのみ取扱いたい場合には、利用可能な全配列をマッチング時のリファレンスとして用いるのが自然だと思います。



R Console window showing DNA sequence data:

```
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta", width=10)
> fasta
A DNAStringSet instance of length 455
      width seq           names
[1] 248956422 NNNNNNNNN...NNNNNNNNNN 1
[2] 242193529 NNNNNNNNN...NNNNNNNNNN 2
[3] 198295559 NNNNNNNNN...NNNNNNNNNN 3
[4] 190214555 NNNNNNNNN...NNNNNNNNNN 4
[5] 181538259 NNNNNNNNN...NNNNNNNNNN 5
...
[451] 200773 TCTACTCTC...GGGAAATTG HSCHR19KIR_FH08_B...
[452] 170148 TTTCTTTCT...GGGAAATTG HSCHR19KIR_FH13_A...
[453] 215732 TGTGGTGAG...GGGAAATTG HSCHR19KIR_FH13_B...
[454] 170537 TCTACTCTC...GGGAAATTG HSCHR19KIR_FH15_A...
[455] 177381 GATCTATCT...GGGAAATTG HSCHR19KIR_RP5_B...
```

Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

①ゲノム配列と②アノテーション情報があれば、任意の領域の配列を取り出すことができます。

プロモーター配列取得

- ・ イントロ | 一般 | Tips | [拡張子は同じで任意の文字を追加して保存](#) (last modified 2013/09/26)
- ・ イントロ | 一般 | 配列取得 | ゲノム配列 | [公共DBから](#) (last modified 2014/05/28)
- ・ イントロ | 一般 | 配列取得 | ゲノム配列 | [BSgenome](#) (last modified 2015/02/19) NEW
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [BSgenomeとTxDbから](#) (modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20)
- ・ イントロ | 一般 | 配列取得 | トランск립トーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランск립トーム配列 | [biomaRt\(Durinck 2009\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | NG | NG
- ・ イントロ | NG | NG

イントロ | 一般 | 配列取得 | プロモーター配列 | BSgenomeとTxDbから

NEW



ゲノム配列([BSgenome](#))パッケージとアノテーション情報([TxDb](#))パッケージを用いて様々な生物種のプロモーター配列(転写開始点近傍配列; 上流配列)を取得するやり方を示します。2014年4月リリースの Bioconductor 2.14 以降の推奨手順では、ゲノムのパッケージ(例: [BSgenome.Hsapiens.UCSC.hg19](#))と対応するアノテーションパッケージ(例: [TxDb.Hsapiens.UCSC.hg19.knownGene](#))の両方を読み込ませる必要がありますので、2015年2月に記述内容を大幅に変更しました。ヒトなどの主要なパッケージ以外はおそらくデフォルトではインストールされていませんので、「パッケージがインストールされていません」的なエラーが出た場合は、[個別パッケージのインストール](#)を参考にして予め利用したいパッケージのインストールを行ってから再挑戦してください。出力はmulti-FASTAファイルです。現状では、ゼブラフィッシュ([danRer7](#))はゲノムパッケージ([BSgenome.Drerio.UCSC.danRer7](#))は存在しますが、対応する TxDb パッケージが存在しないので、どこかから GFF ファイルを取得して makeTranscriptDbFromGFF 関数などを利用して TranscriptDb オブジェクトを得るなどする必要があります。

プロモーター配列取得

1. ヒト(hg19)の場合:

ゲノムパッケージ([BSgenome.Hsapiens.UCSC.hg19](#))と対応するアノテーションパッケージ([TxDb.Hsapiens.UCSC.hg19.knownGene](#))を読み込んで、転写開始点上流1000塩基分を取得する。

```
out_f <- "hoge1.fasta" #出力ファイル名を指定してout_fに格納
param_bsgenome <- "BSgenome.Hsapiens.UCSC.hg19" #パッケージ名を指定(BSgenome系のゲノムパッケージ)
param_txdb <- "TxDb.Hsapiens.UCSC.hg19.knownGene" #パッケージ名を指定(TxDB系のアノテーションパッケージ)
param_upstream <- 1000 #転写開始点上流の塩基配列数を指定
```

```
#前処理(指定したパッケージ中のオブジェクト)
library(param_bsgenome, character.only=TRUE)
tmp <- ls(paste("package", param_bsgenome))
genome <- eval(parse(text=tmp))

library(param_txdb, character.only=TRUE)
tmp <- ls(paste("package", param_txdb))
txdb <- eval(parse(text=tmp))

#本番
gn <- sort(genes(txdb))
hoge <- flank(gn, width=param_upstream)
fasta <- getSeq(genome, hoge)
fasta

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta")
```

**例題1。作業ディレクトリはどこでもよいが
基本はデスクトップ上のhoge。転写開始点上流1000塩基を取得したい場合。
width列は配列長に相当。出力ファイルを見なくてもうまくいっているだろうと思える。**

R Console

```
> fasta <- getSeq(genome, hoge) #指定した範囲の$DNAStringSetを取得
> fasta #確認してるだけ
A DNAStringSet instance of length 23056
  width seq names
  [1] 1000 ACACATGCTA...TTTCGTTAA 100287102
  [2] 1000 GCTATTATCA...AATAACTCT 79501
  [3] 1000 GAATTAGGCT...AAGGCGGGG 643837
  [4] 1000 CGGGGAGGCC...CTTGGGCCA 148398
  [5] 1000 CGGCAGGGCT...AGCGGCGGG 339451
  ...
  [23052] 1000 GGTGAGCCAA...ATCTCAGCC 283788
  [23053] 1000 AGCCCTCCAC...CTCTCCAAC 100507412
  [23054] 1000 CGGGGCCAG...CCTGGCTGC 728410
  [23055] 1000 CGGGGCCAG...CCTGGCTGC 100653046
  [23056] 1000 CAGGCTGAGC...CTCACCGCG 100288687
>
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta", width=1000)
```

プロモーター配列取得

3. 線虫(ce6)の場合:

ゲノムパッケージ([BSgenome.Celegans.UCSC.ce6](#))と対応するアノテーションパッケージ([TxDb.Celegans.UCSC.ce6.ensGene](#))を読み込んで、転写開始点上流500塩基から下流20塩基までの範囲を取得するやり方です。

```
out_f <- "hoge3.fasta"
param_bsgenome <- "BSgenome.Celegans.UCSC.ce6" #パッケージ名を指定(BSgenome系のゲノムパッ
param_txdb <- "TxDb.Celegans.UCSC.ce6.ensGene" #パッケージ名を指定(TxDB系のアノテーション
param_upstream <- 500
param_downstream <- 20
```

```
#前処理(指定したパッケージ中のオブジェクト)
library(param_bsgenome, character.only=TRUE)
tmp <- ls(paste("package", param_bsgenome))
genome <- eval(parse(text=tmp))

library(param_txdb, character.only=T)
tmp <- ls(paste("package", param_txdb))
txdb <- eval(parse(text=tmp))

#本番
gn <- sort(genes(txdb))
hoge <- promoters(gn, upstream=param_upstream,
                  downstream=param_downstream)
fasta <- getSeq(genome, hoge)
fasta
```

R Console

```
> fasta <- getSeq(genome, hoge) #指定した範囲の$  

> fasta #確認してるだけ$  

A DNAStringSet instance of length 27928  

width seq names  

[1] 520 TTGCGCGTAA...CCGCGTCAC Y74C9A.2  

[2] 520 GCAGATAATT...ACGATGGAA Y74C9A.1  

[3] 520 AGAGGAATT...CGAGACAAA Y48G1C.12  

[4] 520 TTACCTCCAG...CCCGATCCC Y48G1C.4  

[5] 520 CTTCATCCCC...GCAGCAGCA Y48G1C.2  

... ...  

[27924] 520 GGGTGTTACA...ATAAAAAAAG MTCE.29  

[27925] 520 ATATCTGCAG...ATAATTCAAG MTCE.30  

[27926] 520 GTAGTATT...TTATATTAA MTCE.32  

[27927] 520 TTTATATTAT...TGTTTAAAT MTCE.33  

[27928] 520 GATTAATATT...TTAATAAAA MTCE.36  

>  

> #ファイルに保存  

> writeXStringSet(fasta, file=out_f, format="fasta", width=50)  

>
```

例題3。線虫はBSgenomeとTxDbの両方が提供されています。予め2つのパッケージを個別にインストールしておけば、上流500～下流20塩基の範囲の配列を取得できる。

Tips: sessionInfo

R R Console

①

②

```
> sessionInfo()
```

R version 3.2.0 (2015-04-16)

Platform: x86_64-w64-mingw32/x64 (64-bit)

Running under: Windows 7 x64 (build 7601) Service Pack 1

locale:

```
[1] LC_COLLATE=Japanese_Japan.932  LC_CTYPE=Japanese_Japan.932  
[3] LC_MONETARY=Japanese_Japan.932 LC_NUMERIC=C  
[5] LC_TIME=Japanese_Japan.932
```

attached base packages:

```
[1] stats4    parallel   stats      graphics   grDevices  utils      datasets  methods  
[9] base
```

other attached packages:

```
[1] TxDb.Celegans.UCSC.ce6.ensGene_3.1.2          BSgenome.Celegans.UCSC.ce6_1.4.0  
[3] BSgenome.Hsapiens.UCSC.hg19_1.4.0              BSgenome_1.36.0  
[5] rtracklayer_1.28.3                            Biostrings_2.36.1  
[7] XVector_0.8.0                                 RCurl_1.95-4.6  
[9] bitops_1.0-6                                  TxDb.Hsapiens.UCSC.hg19.knownGene_3.1.2  
[11] GenomicFeatures_1.20.1                      AnnotationDbi_1.30.1  
[13] Biobase_2.28.0                                GenomicRanges_1.20.3  
[15] GenomeInfoDb_1.4.0                           IRanges_2.2.1  
[17] S4Vectors_0.6.0                             BiocGenerics_0.14.0
```

loaded via a namespace (and not attached):

自分の解析環境を知る。エラーが出てどうにもならない場合は、sessionInfo()実行結果をメール添付か本文中に記載して助けを求める。見る側は、②R本体のバージョン(この場合、3.2.0)や、パッケージが読み込まれているか、あるいはパッケージのバージョンをチェックする。

Tips: sessionInfo

R R Console

```
> sessionInfo()
R version 3.2.0 (2015-04-16)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

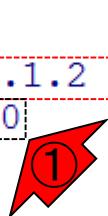
locale:
[1] LC_COLLATE=Japanese_Japan.932   LC_CTYPE=Japanese_Japan.932
[3] LC_MONETARY=Japanese_Japan.932  LC_NUMERIC=C
[5] LC_TIME=Japanese_Japan.932

attached base packages:
[1] stats4    parallel  stats      graphics  grDevices utils      datasets  methods
[9] base

other attached packages:
[1] TxDb.Celegans.UCSC.ce6.ensGene_3.1.2
[3] BSgenome.Hsapiens.UCSC.hg19_1.4.0
[5] rtracklayer_1.28.3
[7] XVector_0.8.0
[9] bitops_1.0-6
[11] GenomicFeatures_1.20.1
[13] Biobase_2.28.0
[15] GenomeInfoDb_1.4.0
[17] S4Vectors_0.6.0

loaded via a namespace (and not attached):
```

これは、うまくプロモーター配列を取得できたときのR環境。①例題1のヒト(*H. sapiens*)プロモーター配列を取得したときと、②例題3の線虫(*C. elegans*)プロモーター配列を取得したときに読み込んだパッケージが表示されているのがわかる。



パッケージのインストール

- ・インストール | R本体 | 過去版 | Mac用 (last modified 2015/03/22)
- ・インストール | Rパッケージ | ほぼ全て(20GB以上?!) (last modified 2015/05/25)
- ・インストール | Rパッケージ | 必要最小限プラスアルファ(数GB?!) (last modified 2015/07/24) 推奨
- ・インストール | Rパッケージ | 必要最小限プラスアルファ(アグリバイアス) (居室のみ) (last modified 2015/05/25)
- ・インストール | Rパッケージ | 必要最小限(数GB?!) (last modified 2015/05/25)
- ・インストール
- ・削除予定
- ・削除予定
- ・基本的な利用

インストール | Rパッケージ | 必要最小限プラスアルファ(数GB?!) NEW

(Rで)塩基配列解析、(Rで)マイクロアレイデータ解析 中で利用するパッケージ、プラスアルファのパッケージをインストールするやり方です。Rパッケージの2大リポジトリであるCRANとBioconductorから提供されているパッケージ群のうち、一部のインストールに相当しますので、相当短時間でインストールが完了します。

SAFEではなくsafeパッケージの間違い
(_)_m limmaの取得先が間違ってCRA
迫力しました。

1. R本体を起動

2. CRANから提供されているパッケージ

以下を「Rコンソール画面上」でコピー
合は自分のいる場所から近いサイトを

```
#(Rで)塩基配列解析で主に利用
install.packages("PoissonSe")
install.packages("samr")
install.packages("seqinr")

#(Rで)マイクロアレイデータ解析
```

```
biocLite("SRAdb", suppressUpdates=TRUE)
biocLite("TCC", suppressUpdates=TRUE) #②マイクロアレイデータ解析でも利用
biocLite("TxDb.Celegans.UCSC.ce6.ensGene", suppressUpdates=TRUE) #①
biocLite("TxDb.Hsapiens.UCSC.hg19.knownGene", suppressUpdates=TRUE)
biocLite("TxDb.Hsapiens.UCSC.hg38.knownGene", suppressUpdates=TRUE)
biocLite("TxDb.Mmusculus.UCSC.mm10.knownGene", suppressUpdates=TRUE)
biocLite("TxDb.Rnorvegicus.UCSC.rn5.refGene", suppressUpdates=TRUE)
#biocLite("yeastRNASeq", suppressUpdates=TRUE)
```

```
#(Rで)マイクロアレイデータ解析で利用
biocLite("affy", suppressUpdates=TRUE)
biocLite("agilo", suppressUpdates=TRUE)
```

4. Bioconductorから提供されているパッケージ群のインストール

ゲノム配列パッケージです。一つ一つの容量が尋常でないため、必要に応じてテキストエディタなどに予めコピペしておき、いらないゲノムパッケージを削除してからお使いください。

```
source("http://bioconductor.org/biocLite.R")#おまじない
biocLite("BSgenome.Athaliana.TAIR.TAIR10", suppressUpdates=TRUE) #シロイヌナズナゲノム
biocLite("BSgenome.Celegans.UCSC.ce6", suppressUpdates=TRUE) #線虫ゲノム
biocLite("BSgenome.Drerio.UCSC.danRer7", suppressUpdates=TRUE) #ゼブラフィッシュゲノム
biocLite("BSgenome.Hsapiens.NCBI.GRCh38", suppressUpdates=TRUE) #ヒトゲノム(GRCh38)
biocLite("BSgenome.Hsapiens.UCSC.hg19", suppressUpdates=TRUE) #ヒトゲノム(hg19)
biocLite("BSgenome.Mmusculus.UCSC.mm10", suppressUpdates=TRUE) #マウスゲノム(mm10)
```

「必要最小限プラスアルファ」の推奨
インストール手順通りにインストール
したヒトの多くはうまくいっている。

全体像

リポジトリ(パッケージ提供元)	パッケージ名	R本体	
		R ver. 3.1.3	R ver. 3.2.0
Bioconductor	TxDb.Celegans.UCSC.ce6.ensGene		3.1.2
Bioconductor	BSgenome.Celegans.UCSC.ce6		1.4.0
CRAN	XML		
CRAN	RCurl		1.95–4.6
CRAN	checkmate		–

R ver. 3.1.3でこけた

①例題3を門田のR ver. 3.1.3でやった結果。②XMLおよび③RCurlパッケージがないためにエラー祭りとなつた(爆)

R R Console

```

> out_f <- "hoge3.fasta" #出力ファイル名を指定してお$  

> param_bsgenome <- "BSgenome.Celegans.UCSC.ce6" #パッケージ名を指定$  

> param_txdb <- "TxDb.Celegans.UCSC.ce6.ensGene" #パッケージ名を指定$  

> param_upstream <- 500 #転写開始点上流の塩基配列$  

> param_downstream <- 20 #転写開始点下流の塩基配列$  

>  

> #前処理(指定したパッケージ中のオブジェクト名をgenomeおよびtxdbに$  

> library(param_bsgenome, character.only=T) #指定したパッケージの読み$  

    要求されたパッケージ BSgenome をロード中です  

    要求されたパッケージ rtracklayer をロード中です  

Error in loadNamespace(j <- i[[1L]], c(lib.loc, .libPaths()), versi$  

  'XML' という名前のパッケージはありません  

  エラー: パッケージ 'rtracklayer' をロードできませんでした  

> tmp <- ls(paste("package", param_bsgenome, sep=":")) #指定したパッ$  

  以下にエラー as.environment(pos) :  

    検索リストに "package:BSgenome.Celegans.UCSC.ce6" という項目はあ$  

> genome <- eval(parse(text=tmp)) #文字列tmpをRオブジェクト$  

  以下にエラー eval(parse(text = tmp)) :  

    引数 'expr' の評価中にエラーが起きました (関数 'eval' に対するメ$  

>  

> library(param_txdb, character.only=T) #指定したパッケージの読み$  

    要求されたパッケージ GenomicFeatures をロード中です  

Error in loadNamespace(j <- i[[1L]], c(lib.loc, .libPaths()), versi$  

  'RCurl' という名前のパッケージはありません  

  エラー: パッケージ 'GenomicFeatures' をロードできませんでした

```

XMLのインストール

R R Console

```
> source("http://bioconductor.org/biocLite.R")#
Bioconductor version 3.0 (BiocInstaller 1.16.5), ?biocLite for help
A new version of Bioconductor is available after installing the
most recent version of R; see http://bioconductor.org/install
> biocLite("XML", suppressUpdates=TRUE)①
BioC_mirror: http://bioconductor.org
Using Bioconductor version 3.0 (BiocInstaller 1.16.5), R version
3.1.3.
Installing package(s) 'XML'
URL 'http://cran.fhcrc.org/bin/windows/contrib/3.1/XML_3.98-1.3.zip'
Content type 'application/zip' length 4293309 bytes (4.1 MB)
開かれた URL
downloaded 4.1 MB

package 'XML' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\kadota\AppData\Local\Temp\RtmpER54eM\downloaded_p$
```

①XMLパッケージのインストール。XML自体はCRANから提供されているが、CRANパッケージ用のinstall.packages関数ではなく、Bioconductor用のbiocLite関数でもCRANパッケージのXMLをインストール可能であるというTipsを兼ねて紹介。

RCurlのインストール

①RCurlパッケージのインストール。
RCurlもCRANから提供されているが、
biocLite関数でインストール可能。

R R Console

```
> source("http://bioconductor.org/biocLite.R") #おまじない
Bioconductor version 3.0 (BiocInstaller 1.16.5), ?biocLite for help
A new version of Bioconductor is available after installing the
most recent version of R; see http://bioconductor.org/install
> biocLite("XML", suppressUpdates=TRUE)
BioC_mirror: http://bioconductor.org
Using Bioconductor version 3.0 (BiocInstaller 1.16.5), R version
3.1.3.
Installing package(s) 'XML'
URL 'http://cran.fhcrc.org/bin/windows/contrib/3.1/XML_3.98-1.3.zip'
Content type 'application/zip' length 4293309 bytes (4.1 MB)
開かれた URL
downloaded 4.1 MB

package 'XML' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
      C:\Users\kadota\AppData\Local\Temp\RtmpER54eM\downloaded_p$
```

1

RCurlのインストール

①RCurlパッケージのインストール。
RCurlもCRANから提供されているが、
biocLite関数でインストール可能。

R R Console

```
package 'XML' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\kadota\AppData\Local\Temp\RtmpER54eM\downloaded_p$ 
> biocLite("RCurl", suppressUpdates=TRUE)
BioC_mirror: http://bioconductor.org
Using Bioconductor version 3.0 (BiocInstaller 1.16.5), R version
  3.1.3.
Installing package(s) 'RCurl'
 URL 'http://cran.fhcrc.org/bin/windows/contrib/3.1/RCurl_1.95-4.7$'
Content type 'application/zip' length 2858866 bytes (2.7 MB)
  開かれた URL
downloaded 2.7 MB

package 'RCurl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\kadota\AppData\Local\Temp\RtmpER54eM\downloaded_p$ 
> |
```

R ver. 3.1.3で再挑戦

意気揚々と、再度例題3をコピペ。しかし
 ①今度はcheckmateというパッケージがないからダメだと言われていることがわかる。

R Console

```

> out_f <- "hoge3.fasta" #出力ファイル名を指定して$  

> param_bsgenome <- "BSgenome.Celegans.UCSC.ce6" #パッケージ名を指$  

> param_txdb <- "TxDb.Celegans.UCSC.ce6.ensGene" #パッケージ名を指$  

> param_upstream <- 500 #転写開始点上流の塩基配列$  

> param_downstream <- 20 #転写開始点下流の塩基配列$  

>  

> #前処理(指定したパッケージ中のオブジェクト名をgenomeおよびtxdbに$  

> library(param_bsgenome, character.only=T) #指定したパッケージの読み$  

    要求されたパッケージ BSgenome をロード中です  

    要求されたパッケージ rtracklayer をロード中です  

Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = $  

  'checkmate' という名前のパッケージはありません  

エラー: パッケージ 'rtracklayer' をロードできませんでした  

> tmp <- ls(paste("package", param_bsgenome, sep=":")) #指定したパ$  

以下にエラー as.environment(pos) :  

  検索リストに "package:BSgenome.Celegans.UCSC.ce6" という項目は$  

> genome <- eval(parse(text=tmp)) #文字列tmpをRオブジェクト$  

以下にエラー eval(parse(text = tmp)) :  

  引数 'expr' の評価中にエラーが起きました (関数 'eval' に対する$  

>  

> library(param_txdb, character.only=T) #指定したパッケージの読み$  

    要求されたパッケージ GenomicFeatures をロード中です  

Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = $
```

checkmateのインストール

①checkmateパッケージのインストール。
checkmateもCRANから提供されている
が、biocLite関数でインストール可能。

R R Console

```
> fasta <- getSeq(genome, hoge)          #指定した範囲の塩基配列情報
以下にエラー (function (classes, fdef, mtable) :
  unable to find an inherited method for function 'getSeq' for s$ 
> fasta                                     #確認してるだけです
エラー: オブジェクト 'fasta' がありません
>
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta", width=50) #fas$ 
以下にエラー is(x, "XstringSet") : オブジェクト 'fasta' がありません
> biocLite("checkmate", suppressUpdates=TRUE) ①
BioC_mirror: http://bioconductor.org
Using Bioconductor version 3.0 (BiocInstaller 1.16.5), R version
 3.1.3.
Installing package(s) 'checkmate'
URL 'http://cran.fhcrc.org/bin/windows/contrib/3.1/checkmate_1.6.$
Content type 'application/zip' length 393202 bytes (383 KB)
開かれた URL
downloaded 383 KB

package 'checkmate' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\kadota\AppData\Local\Temp\RtmpER54eM\downloaded_p$
```

R ver. 3.1.3で再々挑戦



R R Console

```
> out_f <- "hoge3.fasta" #出力ファイル名を指定して$  
> param_bsgenome <- "BSgenome.Celegans.UCSC.ce6" #パッケージ名を指$  
> param_txdb <- "TxDb.Celegans.UCSC.ce6.ensGene" #パッケージ名を指$  
> param_upstream <- 500 #転写開始点上流の塩基配列$  
> param_downstream <- 20 #転写開始点下流の塩基配列$  
>  
> #前処理(指定したパッケージ中のオブジェクト名を  
> library(param_bsgenome, character.  
要求されたパッケージ BSgenome をロード中です  
要求されたパッケージ rtracklayer をロード中  
> tmp <- ls(paste("package", param_bsgenome))  
> genome <- eval(parse(text=tmp))  
>  
> library(param_txdb, character.  
要求されたパッケージ GenomicFeatures をロード中  
> tmp <- ls(paste("package", param_txdb))  
> txdb <- eval(parse(text=tmp))  
>  
> #本番  
> gn <- sort(genes(txdb))  
> hoge <- promoters(gn, upstream=param_upstream,  
+ downstream=param_downstream)  
> fasta <- getSeq(genome, hoge)  
> fasta
```

R R Console

```
> fasta #確認してるだけ$  
A DNAStringSet instance of length 27928  
  width seq names $  
[1] 520 TTGCGCGTAAAATAT...TCTTGCCGCGTCAC Y74C9A.2  
[2] 520 GCAGATAATTGAGGA...TCGAAACGATGGAA Y74C9A.1  
[3] 520 AGAGGAATTTCACCG...TGCAGACGAGACAAA Y48G1C.12  
[4] 520 TTACCTCCAGTGATT...AAATTCCCGATCCC Y48G1C.4  
[5] 520 CTTCATTCACACT...TGGAAAGCAGCAGCA Y48G1C.2  
... ...  
[27924] 520 GGGTGTACACTATG...TTAGTATAAAAAAG MTCE.29  
[27925] 520 ATATCTGCAGTATT...TTAGTATAATTCAAG MTCE.30  
[27926] 520 GTAGTATTTCGAA...ATAGTTATATTAA MTCE.32  
[27927] 520 TTTATATTACGG...TTACTTGTAAAT MTCE.33  
[27928] 520 GATTAATATTTTG...TTAGTTAATAAAA MTCE.36  
>  
> #ファイルに保存  
> writeXStringSet(fasta, file=out_f, format="fasta", width=50)  
> |
```

この程度は...

バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | NGSハンズオン講習会2015 NEW

NGSハンズオン講習会を2015年7月22日-8月6日の11日間で実施します。予定通り受講申込多数のため、予備日(8月26日、27日、28日)も実施することになっています。



はじめに(全)

- ・ 講習
基本
ズオ
- ・ 平成

2015年7月22日(2015.07.22):Bio-Linux 8とRのインストール状況確認(門田幸二、寺田透)

書籍「日本乳酸菌学会誌」について示した通りのPC環境を構築しておきましょう。連載第1-3回、および第4回のウェブ資料W6-5までの予習は必須です。Rについても同様です。一週間程度はきっちり時間をかけて予習しておきましょう。7/22は、以下に示すようなことができる(わかる)ようになっていることの確認を自分でしてもらう日です。門田自身全てを完全に把握しているわけではありませんし、ウェブ資料のページ数も膨大ですので、どこにどういうことが書かれていたかの全体像の俯瞰やチェックリストという位置づけであります(頻繁に更新しているのでときどきリロードしましょう):

- Bio-Linux 8

• R

1. インストールについてをよく読み、ここに書いてある手順に従って2015年4月4日以降にインストールを行った
2. インストールについてで書いてある内容はBio-Linux8(ゲストOS)とは無関係であり、WindowsやMacintosh(つまりホストOS)上で行う作業である
3. ファイルの拡張子(.txtや.docxなど)はちゃんと表示されている
4. Rの起動と終了ができる。終了時に表示されるメッセージにこうろたえない
5. R本体だけでなくRパッケージ群のインストールもちゃんと行った
6. Rパッケージ群のインストール確認も行い、エラーが出ないことを確認した
7. library関数を用いたRパッケージのロード中に、別のパッケージがないことに起因するエラーメッセージが出ることもあるが、必要なパッケージを個別にインストールするやり方を知っている
8. 基本的な利用法をよく読み、予習を行った
9. 作業ディレクトリの変更ができる
10. 例題ファイルのダウンロード時に、拡張子が勝手に変わることがあるので注意する
11. 慣れない場合は、getwd()とlist.files()を打ち込むことで、作業ディレクトリと入力ファイルの存在確認を行う
12. エラーに遭遇した際、「ありがちなミス1-4」に当てはまっていないかどうか自分で確認

①

sessionInfo (R 3.1.3)

R R Console

① > sessionInfo()
②

```
> sessionInfo()
R version 3.1.3 (2015-03-09)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

locale:
[1] LC_COLLATE=Japanese_Japan.932  LC_CTYPE=Japanese_Japan.932
[3] LC_MONETARY=Japanese_Japan.932 LC_NUMERIC=C
[5] LC_TIME=Japanese_Japan.932

attached base packages:
[1] stats4     parallel   stats      graphics   grDevices  utils      datasets   methods
[9] base

other attached packages:
[1] TxDb.Celegans.UCSC.ce6.ensGene_3.0.0  GenomicFeatures_1.18.7
[3] BSgenome.Celegans.UCSC.ce6_1.4.0        BSgenome_1.34.1
[5] rtracklayer_1.26.3                      BiocInstaller_1.16.5
[7] Biostrings_2.34.1                       XVector_0.6.0
[9] AnnotationDbi_1.28.2                    Biobase_2.26.0
[11] GenomicRanges_1.18.4                   GenomeInfoDb_1.2.4
[13] IRanges_2.0.1                           S4Vectors_0.4.0
[15] BiocGenerics_0.12.1

loaded via a namespace (and not attached):
```

sessionInfo (R 3.1.3)

特に①～③のパッケージは、④2015年7月24日にインストールしたばかりであり、前述のR ver. 3.2.0実行結果よりも後にインストールしたものである。

R R Console

other attached packages:

```
[1] TxDb.Celegans.UCSC.ce6.ensGene_3.0.0 GenomicFeatures_1.18.7
[3] BSgenome.Celegans.UCSC.ce6_1.4.0   BSgenome_1.34.1
[5] rtracklayer_1.26.3      BiocInstaller_1.16.5
[7] Biostrings_2.34.1       XVector_0.6.0
[9] AnnotationDbi_1.28.2    Biobase_2.26.0
[11] GenomicRanges_1.18.4   GenomeInfoDb_1.2.4
[13] IRanges_2.0.1          S4Vectors_0.4.0
[15] BiocGenerics_0.12.1
```

loaded via a namespace (and not attached):

```
[1] base64enc_0.1-2           BatchJobs_1.6
[4] BiocParallel_1.0.3         biomaRt_2.22.0
[7] brew_1.0-6                 checkmate_1.6.1
[10] crayon_1.3.0              DBI_0.3.1
[13] fail_1.2                  foreach_1.4.2
[16] iterators_1.0.7            magrittr_1.5
[19] RCurl_1.95-4.7             Rsamtools_1.18.3
[22] sendmailR_1.2-1            stringi_0.4-1
[25] testthat_0.10.0            tools_3.1.3
[28] zlibbioc_1.12.0
> date()
[1] "Fri Jul 24 21:21:59 2015"
```

④

BBmisc_1.9
bitops_1.0-6
codetools_0.2-11
digest_0.6.8
GenomicAlignments_1.2.2
memoise_0.2.1
RSQLite_1.0.0
stringr_1.0.0
XML_3.98-1.3

①

③

②

全体像

リポジトリ(パッケージ提供元)	パッケージ名	R本体	
		R ver. 3.1.3	R ver. 3.2.0
Bioconductor	TxDb.Celegans.UCSC.ce6.ensGene	3.0.0	3.1.2
Bioconductor	BSgenome.Celegans.UCSC.ce6	1.4.0	1.4.0
CRAN	XML	3.98–1.3	
CRAN	RCurl	1.95–4.7	1.95–4.6
CRAN	checkmate	1.6.1	–

sessionInfo (R 3.2.0)

BSgenomeパッケージはver. 1.4.0と同じ。しかし
TxDbパッケージは、ver. 3.1.2 (R ver. 3.2.0)とver.
3.0.0 (R ver. 3.1.3)で異なっていることがわかる。

R R Console

① > sessionInfo()
② R version 3.2.0 (2015-04-16)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

locale:
[1] LC_COLLATE=Japanese_Japan.932 LC_CTYPE=Japanese_Japan.932
[3] LC_MONETARY=Japanese_Japan.932 LC_NUMERIC=C
[5] LC_TIME=Japanese_Japan.932

attached base packages:
[1] stats4 parallel stats graphics grDevices utils datasets methods
[9] base

other attached packages:
[1] TxDb.Celegans.UCSC.ce6.ensGene_3.1.2 BSgenome.Celegans.UCSC.ce6_1.4.0
[3] BSgenome.Hsapiens.UCSC.hg19_1.4.0 BSgenome_1.36.0
[5] rtracklayer_1.28.3 Biostrings_2.36.1
[7] XVector_0.8.0 RCurl_1.95-4.6
[9] bitops_1.0-6 TxDb.Hsapiens.UCSC.hg19.knownGene_3.1.2
[11] GenomicFeatures_1.20.1 AnnotationDbi_1.30.1
[13] Biobase_2.28.0 GenomicRanges_1.20.3
[15] GenomeInfoDb_1.4.0 IRanges_2.2.1
[17] S4Vectors_0.6.0 BiocGenerics_0.14.0

loaded via a namespace (and not attached):

sessionInfo (R 3.2.0)

R R Console

other attached packages:

```
[1] TxDb.Celegans.UCSC.ce6.ensGene_3.1.2      BSgenome.Celegans.UCSC.ce6_1.4.0
[3] BSgenome.Hsapiens.UCSC.hg19_1.4.0          BSgenome_1.36.0
[5] rtracklayer_1.28.3                          Biostrings_2.36.1
[7] XVector_0.8.0                             RCurl_1.95-4.6 ②
[9] bitops_1.0-6                               TxDb.Hsapiens.UCSC.hg19.knownGene_3.1.2
[11] GenomicFeatures_1.20.1                     AnnotationDbi_1.30.1
[13] Biobase_2.28.0                           GenomicRanges_1.20.3
[15] GenomeInfoDb_1.4.0                        IRanges_2.2.1
[17] S4Vectors_0.6.0                          BiocGenerics_0.14.0
```

loaded via a namespace (and not attached):

```
[1] zlibbioc_1.14.0           GenomicAlignments_1.4.1 BiocParallel_1.2.1
[4] tools_3.2.0                DBI_0.3.1                 lambda.r_1.1.7
[7] futile.logger_1.4.1        futile.options_1.0.0    biomaRt_2.24.0
[10] RSQLite_1.0.0              Rsamtools_1.20.2       XML_3.98-1.1 ①
```

①XML、および②RCurlのバージョン。
checkmateパッケージは、存在の有無は不明であるが、少なくともロードされておらず、
R ver. 3.2.0では必要とされていないようだ。

①赤枠がごく最近(2015年7月24日)インストールしたCRAN提供パッケージたち。バージョン番号がR ver. 3.2.0のものに比べて大きいので、確かに最新版なのだろう。

全体像を把握

リポジトリ(パッケージ提供元)	パッケージ名	R本体	
		R ver. 3.1.3	R ver. 3.2.0
Bioconductor	TxDb.Celegans.UCSC.ce6.ensGene	3.0.0	3.1.2
Bioconductor	BSgenome.Celegans.UCSC.ce6	1.4.0	1.4.0
CRAN	XML	3.98-1.3	3.98-1.1
CRAN	RCurl	1.95-4.7	1.95-4.6
CRAN	checkmate	1.6.1	-

全体像を把握

これらの結果から、2つの結論が導き出される。①CRAN提供パッケージは、R本体のバージョンに関係なく、常に最新版がインストールされる。②R本体のバージョンが上がると必要とされないパッケージが出ることもある(checkmate)

リポジトリ(パッケージ提供元)	パッケージ名	R本体	
		R ver. 3.1.3	R ver. 3.2.0
Bioconductor	TxDb.Celegans.UCSC.ce6.ensGene	3.0.0	3.1.2
Bioconductor	BSgenome.Celegans.UCSC.ce6	1.4.0	1.4.0
CRAN	XML	3.98-1.3	3.98-1.1
CRAN	RCurl	1.95-4.7	1.95-4.6
CRAN	checkmate	1.6.1	-

R ver. 3.1.3で再挑戦

R R Console

```
> out_f <- "hoge3.fasta" #出力ファイル名を指定して$  
> param_bsgenome <- "BSgenome.Celegans.UCSC.ce6" #パッケージ名を指$  
> param_txdb <- "TxDb.Celegans.UCSC.ce6.ensGene" #パッケージ名を指$  
> param_upstream <- 500 #転写開始点上流の塩基配列$  
> param_downstream <- 20 #転写開始点下流の塩基配列$  
>  
> #前処理(指定したパッケージ中のオブジェクト名をgenomeおよびtxdbに$  
> library(param_bsgenome, character.only=T) #指定したパッケージの読み$  
要求されたパッケージ BSGenome をロード中です  
要求されたパッケージ rtracklayer をロード中です  
Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = $  
'checkmate' という名前のパッケージはありません  
エラー: パッケージ 'rtracklayer' をロードできませんでした  
> tmp <- ls(paste("package", param_bsgenome, sep=":")) #指定したパッ$  
以下にエラー as.environment(pos) :  
  検索リストに "package:BSgenome.Celegans.UCSC.ce6" という項目は$  
> genome <- eval(parse(text=tmp)) #文字列tmpをRオブジェクト$  
以下にエラー eval(parse(text = tmp)) :  
  引数 'expr' の評価中にエラーが起きました (関数 'eval' に対する$  
  
>  
> library(param_txdb, character.only=T) #指定したパッケージの読み$  
要求されたパッケージ GenomicFeatures をロード中です  
Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = $
```

ここで、もう一度「R ver. 3.1.3で再挑戦」したときのエラーメッセージを読み返す。①黒枠は、BSgenome.Celegans.UCSC.ce6パッケージのロード部分。

①

R ver. 3.1.3で再挑戦

R Console

```

> out_f <- "hoge3.fasta" #出力ファイル名を
> param_bsgenome <- "BSgenome.Celegans.UCSC.ce6" #/パッケージ名
> param_txdb <- "TxDb.Celegans.UCSC.ce6.ensGene" #/パッケージ名
> param_upstream <- 500 #転写開始点上流
> param_downstream <- 20 #転写開始点下流
>
> #前処理(指定したパッケージ中のオブジェクト名をgenomeおよびtxdbに$)
> library(param_bsgenome, character.only=T) #指定したパッケージの読み込み
要求されたパッケージ BSgenome をロード中です
要求されたパッケージ rtracklayer をロード中です
Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = $)
  'checkmate' という名前のパッケージはありません
エラー: パッケージ 'rtracklayer' をロードできませんでした
> tmp <- ls(paste("package", param_bsgenome, sep=":")) #指定したパッケージ名
以下にエラー as.environment(pos) :
  検索リストに "package:BSgenome.Celegans.UCSC.ce6" という項目は $
> genome <- eval(parse(text=tmp)) #文字列tmpをRオブジェクトに
以下にエラー eval(parse(text = tmp)) :
  引数 'expr' の評価中にエラーが起きました (関数 'eval' に対する $
>
> library(param_txdb, character.only=T) #指定したパッケージの読み込み
要求されたパッケージ GenomicFeatures をロード中です
Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = $)

```

文脈から、以下のように理解する。①
 BSgenome.Celegans.UCSC.ce6は内部的にrtracklayerを利用しているため、
 rtracklayerもロードしようとしている。しかし、
 ②rtracklayerは内部的にcheckmateを利用している。このためcheckmateもロードしようとしたが、checkmateが未インストールのためこけた、というのが事の顛末。

①

②

依存関係を含めると…

複雑。rtracklayer ver. 1.28.3ではcheckmateパッケージを内部的に使わなくなったのだろう、と理解する。たとえBsgenomeとTxDbのバージョンがR本体の間で不变であったとしても、内部的に用いるパッケージのバージョンや、それに起因する実行結果の違いもありうる。

リポジトリ(パッケージ提供元)	パッケージ名	R本体	
		R ver. 3.1.3	R ver. 3.2.0
Bioconductor	TxDb.Celegans.UCSC.ce6.ensGene	3.0.0	3.1.2
CRAN	RCurl	1.95–4.7	1.95–4.6
Bioconductor	Bsgenome.Celegans.UCSC.ce6	1.4.0	1.4.0
Bioconductor	rtracklayer	1.26.3	1.28.3
CRAN	XML	3.98–1.3	3.98–1.1
CRAN	checkmate	1.6.1	–

①

現実的な対策

私は少しの結果の違いは(突き詰めていっても不毛なので)気にしない。論文中に記載するのはR本体と主要なパッケージ(この場合BSgenomeとTxDb)のみが基本。

リポジトリ(パッケージ提供元)	パッケージ名	R本体	
		R ver. 3.1.3	R ver. 3.2.0
Bioconductor	TxDb.Celegans.UCSC.ce6.ensGene	3.0.0	3.1.2
CRAN	RCurl	1.95–4.7	1.95–4.6
Bioconductor	BSgenome.Celegans.UCSC.ce6	1.4.0	1.4.0
Bioconductor	rtracklayer	1.26.3	1.28.3
CRAN	XML	3.98–1.3	3.98–1.1
CRAN	checkmate	1.6.1	–

R本体のバージョンは重要

■ 近年のリリース頻度

- R本体 (<http://www.r-project.org/>)
 - 2015-06-18にver. 3.2.1をリリース
 - 2015-04-16にver. 3.2.0をリリース
 - 2015-03-09にver. 3.1.3をリリース
 - 2014-10-31にver. 3.1.2をリリース
 - ...
 - 2012-03-30にver. 2.15.0をリリース
 - ...

□ Bioconductor (<http://bioconductor.org/>)は半年ごとにリリース

- 2015-04にver. 3.1をリリース (R ver. 3.2.1で動作確認)、提供パッケージ数 : 1,024
- 2014-10にver. 3.0をリリース (R ver. 3.1.1で動作確認)、提供パッケージ数 : 934
- 2014-04にver. 2.14をリリース (R ver. 3.1.0で動作確認)、提供パッケージ数 : 824
- 2013-10にver. 2.13をリリース (R ver. 3.0で動作確認)、提供パッケージ数 : 750
- 2013-04にver. 2.12をリリース (R ver. 3.0で動作確認)、提供パッケージ数 : 672
- 2012-10にver. 2.11をリリース (R ver. 2.15.1で動作確認)、提供パッケージ数 : 608
- 2012-04にver. 2.10をリリース (R ver. 2.15.0で動作確認)、提供パッケージ数 : 553
- ...

R本体のバージョンがわかれば、Bioconductor提供パッケージのバージョンも概ね定まる。例えば、①R ver. 3.1.2では、どのタイミングで(例えば2015年8月に)パッケージのインストールを行おうが、Bioconductor ver. 3.0で提供されているパッケージのバージョンとなる。決して2015年4月に公開されたBioconductor ver. 3.1で提供されているパッケージがインストールされることはない。

①

Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

プロモーター配列取得

- ・ イントロ | 一般 | Tips | [拡張子は同じで任意の文字を追加して保存](#) (last modified 2013/09/26)
- ・ イントロ | 一般 | 配列取得 | ゲノム配列 | [公共DBから](#) (last modified 2014/05/28)
- ・ イントロ | 一般 | 配列取得 | ゲノム配列 | [BSgenome](#) (last modified 2015/02/19) NEW
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [BSgenomeとTxDbから](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [公共DBから](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トランскриპтомум配列 | [PublicDB](#) (last modified 2014/04/02)
- ・ イントロ | 一般 | 配列取得 | トранскриптом配列 | [Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トранскриптом配列 | [TranscriptDb](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トранскриптом配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20) NEW
- ・ イントロ | 一般 | 配列取得 | トранскриптом配列 | [PublicDB](#) (last modified 2014/04/02)

multi-FASTA形式のゲノム配列ファイルとGFF3形式などの一般的なアノテーションファイルがあれば、それを利用することもできる。②例題7の乳酸菌のプロモーター配列取得を示します。

イントロ | 一般 | 配列取得 | プロモーター配列 | GenomicFeatures(Lawrence_2013) NEW

[GenomicFeatures](#)パッケージを主に用いてプロモーター配列(転写開始点近傍配列)を得るやり方を示します。

ここでは、[イントロ | 一般 | 配列取得 | ゲノム配列\(BSgenomeから\)](#)で指定可能なゲノムと[イントロ | NGS | アノテーション情報取得 | TranscriptDb | GenomicFeatures\(Lawrence 2013\)](#)で作成可能なTranscriptDbオブジェクトを入力として、getPromoterSeqという関数を用いて、転写開始点から任意の[上流xxx塩基, 下流yyy塩基]分の塩基配列を取得して、FASTA形式ファイルで保存するやり方を示しています。multi-FASTA形式のゲノム配列ファイルとGFF3形式のアノテーションファイルのみからプロモータ配列を取得するやり方も示しています。

1. ヒト (2) 7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモーター配列を取得する場合:

[Ensembl](#) (Flicek et al., Nucleic Acids Res., 2014)から提供されている *Lactobacillus casei* 12Aの multi-FASTA形式ゲノム配列ファイル(*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa)と GFF3形式のアノテーションファイル(*Lactobacillus casei* 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3)を読み込むやり方です。

```
in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル名
out_f <- "hoge7.fasta" #出力ファイル名を指定してout_fに格納
param_upstream <- 500 #転写開始点上流の塩基配列数を指定
param_downstream <- 20 #転写開始点下流の塩基配列数を指定

#必要なパッケージをロード
library(Rsamtools) #パッケージの読み込み
library(Biostrings) #パッケージの読み込み
library(GenomicFeatures) #パッケージの読み込み
```

プロモーター配列取得

2015年3月5-6日のHPCIハンズオン講習会時のスライド(2015.03.05)です。このときはR ver. 3.1.2でうまくいった。

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモーター配列を取得する場合:

[Ensembl](#) (Flieck et al., Nucleic Acids Res., 2014)から提供されている [Lactobacillus casei 12A](#) の multi-FASTA形式ゲノム配列ファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3) と GFF3形式のアノテーションファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa) を読み込むやり方です。

`makeTranscriptDbFromGFF` 関数実行時に用いている `useGenesAsTranscripts=T` は、原核生物(prokaryotes)に代表される「転写物 = 遺伝子」の場合に指定します。デフォルトは `F` です。

```
in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル名を指定
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル名を指定
out_f <- "hoge7.fasta" #出力ファイル名を指定してout_fに格納
param_upstream <- 500 #転写開始点上流の塩基配列数を指定
param_downstream <- 20 #転写開始点下流の塩基配列数を指定
```

```
#必要なパッケージをロード
library(Rsamtools)
library(Biostrings)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDb
                                   format="gff3", useGenesAsTranscripts=T) #確認してるだ
txdb

#前処理(欲しい領域の座標情報取得)
gn <- sort(genes(txdb)) #遺伝子の座標
hoge <- promoters(gn, upstream=param_upstream, #指定
                  downstream=param_downstream) #指定した範囲
hoge
obj <- (ranges(hoge)@start >= 0) #条件を満たす
```

R Console

```
> #本番(配列取得)
> fasta <- getSeq(FaFile(in_f1), hoge) #指定した範囲の塩基配列情報を取得
> names(fasta) <- names(hoge) #description情報を追加して$を付けて確認するだけです
> fasta
A DNAStringSet instance of length 2680
  width seq
  [1] 520 CTTGAAAGCCTTGAAA...ATTTACGATCACCCG names
      gene:LCA12A_0618
  [2] 520 GAACTGATAATGTGCG...AACAAATCGAAATCAC names
      gene:LCA12A_0619
  [3] 520 GTTTCACGATCAATCG...ACTGGATCACTTGGT names
      gene:LCA12A_0620
  [4] 520 CCACAGCATTGGCTGT...GGACAAGAAAGAAC names
      gene:LCA12A_0621
  [5] 520 ACTGATCATTATGACC...TGATGCCAAGAAAG names
      gene:LCA12A_0622
...
[2676] 520 AAGTGACCGTTGCTTA...TGCATCTATCACTGC names
      gene:LCA12A_0612
[2677] 520 ACACAGAAACTTTATG...AACCTTTCAAGGCAG names
      gene:LCA12A_0613
[2678] 520 TGGCGCAAGTCATATCG...AAATCGTCAAATCAA names
      gene:LCA12A_0614
[2679] 520 CGAATTTCCTGTGATT...TGGCGCTTTCGGCGG names
      gene:LCA12A_0615
[2680] 520 GCGCCAATTACACGCTC...AACCAAACGGACTTT names
      gene:LCA12A_0616
>
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta", width=50) #fast$を付けて確認する
> |
```

プロモーター配列取得

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモ

Ensembl (Flieck et al., Nucleic Acids Res., 2014)から提供されている [Lactobacillus casei 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3](#) と [Lactobacillus casei 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa](#) を読み込むやり方です。

`makeTranscriptDbFromGFF` 関数実行時に用いている `useGenesAsTranscripts=T` は、原核生物(prokaryotes)に代表される「転写物 = 遺伝子」の場合に指定します。デフォルトは F です。

```

in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル名を指定
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル名を指定
out_f <- "hoge7.fasta" #出力ファイル名を指定してout_fに格納
param_upstream <- 500 #転写開始点上流の塩基配列数を指定
param_downstream <- 20 #転写開始点下流の塩基配列数を指定

#必要なパッケージをロード
library(Rsamtools) #パッケージの読み込み
library(Biostrings) #パッケージの読み込み
library(GenomicFeatures) #パッケージの読み込み

#前処理(アノテーション情報取得)
txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクトを取得してtxdbに格納
                                    format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクトを取得してtxdbに格納
txdb #確認してるだけです

#前処理(欲しい領域の座標情報取得)
gn <- sort(genes(txdb)) #遺伝子の座標情報を取得
hoge <- promoters(gn, upstream=param_upstream, #指定した範囲の座標情報を取得
                  downstream=param_downstream) #指定した範囲の座標情報を取得
hoge #確認してるだけです
obj <- (ranges(hoge)$start >= 0) #条件を満たすかどうかを判定した結果をobjに格納(座標が0よりも大きい)

```

当時(2015年3月; R ver. 3.1.2)は、GFF3形式のアノテーションファイル読み込み時に①`makeTranscriptDbFromGFF` 関数、および②`useGenesAsTranscripts` オプション(デフォルトは F)を利用していた。

2015年7月3日にR ver. 3.2.0で黒枠部分をコピペすると警告メッセージが出た

プロモーター配列取得

```

in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル名
out_f <- "hoge7.fasta" #出力ファイル名を指定してout_fに格納
param_upstream <- 500 #転写開始点上流の塩基配列数を指定
param_downstream <- 20 #転写開始点下流の塩基配列数を指定

#必要なパッケージをロード
library(Rsamtools) #パッケージの読み込み
library(Biostrings) #パッケージの読み込み
library(GenomicFeatures) #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクトを取得してtxdbに格納
                                    format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクトを取得してtxdbに格納
txdb #確認しているだけです

```

```

#前処理(欲しい領域の座標情報を取得)
gn <- sort(genes(txdb))
hoge <- promoters(gn, upstream=param_upstream,
                  downstream=param_downstream)
obj <- (ranges(hoge))@start

```

R Console

```

> #前処理(アノテーション情報を取得)
> txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクトを$ #確認しているだけです
+                                     format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクトを$ #確認しているだけです
Prepare the 'metadata' data frame ... metadata: OK
Warning messages:
1: 'makeTranscriptDbFromGFF' is deprecated.
Use 'makeTxDbFromGFF' instead. ①
See help("Deprecated")
2: 'useGenesAsTranscripts' is ignored and deprecated
3: In matchCircularity(seqlevels(gr), circ_seqs) :
  None of the strings in your circ_seqs argument match your seqnames.
> txdb
TxDb object:

```



プロモーター配列取得

R R Console

```
> #前処理(アノテーション情報を取得)
> txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクト
+                                     format="gff3", useGenesAsTranscripts=T) #TxDbオブジェクト
Prepare the 'metadata' data frame ... metadata: OK
Warning messages:
1: 'makeTranscriptDbFromGFF' is deprecated.
  Use 'makeTxDbFromGFF' instead.
See help("Deprecated")
2: 'useGenesAsTranscripts' is ignored and deprecated
3: In matchCircularity(seqlevels(gr), circ_seqs) :
  None of the strings in your circ_seqs argument match your seqnames.
> txdb
#確認してるだけです
```

```
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3
# Organism: NA
# miRBase build ID: NA
# Genome: NA
# transcript_nrow: 2799
# exon_nrow: 2800
# cds_nrow: 2681
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2015-07-25 15:30:24 +0900 (Sat, 25 Jul 2015)
# GenomicFeatures version at creation time: 1.20.1
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
```

警告メッセージの中身。①

makeTranscriptDbFromGFF関数は削除予定だ。makeTxDbFromGFF関数を使用せよ。②useGenesAsTranscriptsオプションを与えても無視するし、このオプション自体も削除予定だよ。といった具合。R ver. 3.2.0では、今のところ③で示すようにうまく読み込めてはいる。

①

②

③

プロモーター配列取得

- ・ イントロ | 一般 | 配列取得 | ゲノム配列 | [BSgenome](#) (last modified 2015/04/22)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [公共DBから](#) (last modified 2014/04/07)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [BSgenomeとTxDbから](#) (last modified 2015/02/20)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20)
- ・ イントロ | 一般 | 配列取得 | リンスクリプトーム配列 | [公共DBから](#) (last modified 2015/05/09)
- ・ イントロ | 一般 | 配列取得 | [TranscriptDb](#) (last modified 2012/01/01)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) NEW
- ・ イントロ | 一般 | 配列取得 | ゲノム配列(BSgenomeから)で指定可能なゲノムと [NGS | アノテーション情報取得](#)
- ・ イントロ | 一般 | 配列取得 | [TranscriptDb](#) (last modified 2012/01/01) で作成可能なTranscriptDbオブジェクトを確認する

イントロ | 一般 | 配列取得 | プロモーター配列 | GenomicFeatures(Lawrence 2013) NEW

[GenomicFeatures](#)パッケージを主に用いてプロモーター配列(転写開始点近傍配列)を得るやり方を示します。

ここでは、[イントロ | 一般 | 配列取得 | ゲノム配列\(BSgenomeから\)](#)で指定可能なゲノムと [NGS | アノテーション情報取得](#) TranscriptDb (last modified 2012/01/01) で作成可能なTranscriptDbオブジェクトを確認する

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモーター配列を取得する場合:

Ensembl (Flieck et al., Nucleic Acids Res., 2014)から提供されている *Lactobacillus casei* 12Aの multi-FASTA形式ゲノム配列ファイル(*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa)と GFF3形式のアノテーションファイル(*Lactobacillus casei* 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3)を読み込むやり方です。

```
in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル
out_f <- "hoge7.fasta" #出力ファイル名を指定してout_fに格納
param_upstream <- 500 #転写開始点上流の塩基配列数を指定
param_downstream <- 20 #転写開始点下流の塩基配列数を指定

#必要なパッケージをロード
library(Rsamtools) #パッケージの読み込み
library(Biostrings) #パッケージの読み込み
library(GenomicFeatures) #パッケージの読み込み

#前処理(アノテーション情報を取得)
#txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクトを取得してtxdbに格納
#                                    format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクトを取得してtxdbに格納
txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDbオブジェクトを取得してtxdbに格納
txdb #確認しているだけです
```

2015年7月25日現在。③R ver. 3.1.2で正常動作したコードを(#のコメントアウトで)残しつつ、④R ver. 3.2.0で動作確認済みのコードがデフォルトとなるようにしている。

④

①

③

R ver. 3.2.0で実行

makeTranscriptDbFromGFF関数
やuseGenesAsTranscriptsオプションを使っていないので、①それ系の警告は出でていないことがわかる。

```
> #前処理(アノテーション情報を取得)
> #txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブ$
> #           format="gff3", useGenesAsTranscripts=T) #Trans$#
> txdb <- makeTxDbFromGFF(in_f2, format="gff3") #Transcript$#
Prepare the 'metadata' data frame ... metadata: OK
Warning message:
```

In matchCircularity(seqlevels(...))
None of the strings in your

```
> txdb
TxDb object: ①
# Db type: TxDb
# Supporting package: GenomicFe
# Data source: Lactobacillus_ca
# Organism: NA
# miRBase build ID: NA
# Genome: NA
# transcript_nrow: 2799
# exon_nrow: 2800
```

> fasta #確認してるだけで\$
A DNAStringSet instance of length 2726
 width seq
 [1] 520 CTTGAAAGCCTTGA...TTTACGATCACCCG LCA12A_0618
 [2] 520 GAACTGATAATGTG...ACAATCGAAATCAC LCA12A_0619
 [3] 520 GTTTCACGATCAAT...CTGGATCACTTGGT LCA12A_0620
 [4] 520 CCACAGCATTGGCT...GACAAGAAAGAAC LCA12A_0621
 [5] 520 ACTGATCATTATGA...GATCGCCAAGAAC LCA12A_0622
 ...
 [2722] 520 AAGTGACCGTTGCT...GCATCTATCACTGC LCA12A_0612
 [2723] 520 ACACAGAAAATTAA...ACCTTTCAAGGCAG LCA12A_0613
 [2724] 520 TGCACAAAGTCATAT...AATCGTCAAATCAA LCA12A_0614
 [2725] 520 CGAATTTCCTGTGA...GCGCGTTCCGGCGG LCA12A_0615
 [2726] 520 GCGCCAATTACACGC...ACCAAAACGGACTTT LCA12A_0616
>
> #ファイルに保存
> writeXStringSet(fasta, file=out_f, format="fasta", width\$#
> |

R ver. 3.1.2で実行

```
in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chr"
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome"
out_f <- "hoge7.fasta"
param_upstream <- 500
param_downstream <- 20

#出力ファイル名を指定してout_fに格納
#転写開始点上流の塩基配列数を指定
#転写開始点下流の塩基配列数を指定
```

```
#必要なパッケージをロード
library(Rsamtools)
library(Biostrings)
library(GenomicFeatures)
```

#前処理(アノテーション情報を取得)

```
#txdb <- makeTranscriptDbFromGFF(in_f1,
#                                     format="gff3", useGenesAsTranscripts=T)
txdb <- makeTxDbFromGFF(in_f2, format="gff3")
```

#前処理(欲しい領域の座標情報を取得)

```
gn <- sort(genes(txdb))
hoge <- promoters(gn, upstream=param_upstream,
                  downstream=param_downstream)
```



The following object is masked from 'package:GenomeInfoDb':

species

警告メッセージ:

1: パッケージ 'GenomicFeatures' はバージョン 3.1.3 の R の下で\$
 2: パッケージ 'AnnotationDbi' はバージョン 3.1.3 の R の下で造\$
 >
 > #前処理(アノテーション情報を取得)
 > #txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクト)
 > # format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブ
 > txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDbオブ
 エラー: 関数 "makeTxDbFromGFF" を見つけることができませんでした
 > txdb #確認してるだけです
 エラー: オブジェクト 'txdb' がありません
 > package.version("GenomicFeatures")
 [1] "1.18.7"
 > packageVersion("GenomicFeatures")
 [1] '1.18.7'
 >

R ver. 3.1.2(正確にはGenomicFeatures ver. 1.18.7)当時は、まだmakeTxDbFromGFF関数が実装されていないので、①makeTxDbFromGFF関数を見つけることができないのは当たり前。

R ver. 3.1.3で実行

```
in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chr"
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome"
out_f <- "hoge7.fasta"
param_upstream <- 500
param_downstream <- 20

#出力ファイル名を指定してout_fに格納
#転写開始点上流の塩基配列数を指定
#転写開始点下流の塩基配列数を指定
```

```
#必要なパッケージをロード
library(Rsamtools)
library(Biostrings)
library(GenomicFeatures)
```

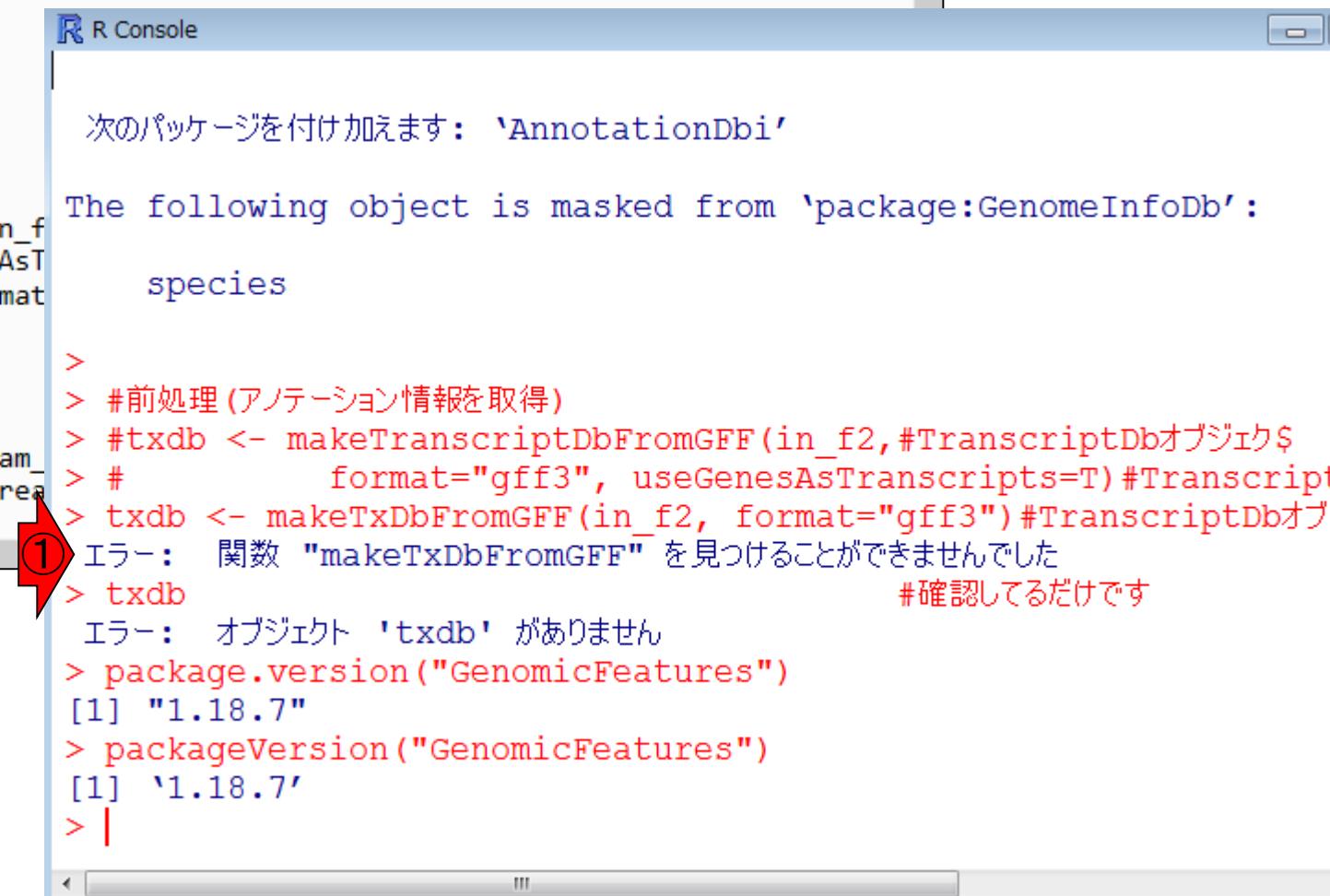
#前処理(アノテーション情報を取得)

```
#txdb <- makeTranscriptDbFromGFF(in_f1,
#                                     format="gff3", useGenesAsTranscripts=T)
txdb <- makeTxDbFromGFF(in_f2, format="gff3")
```

#前処理(欲しい領域の座標情報取得)

```
gn <- sort(genes(txdb))
hoge <- promoters(gn, upstream=param_upstream,
                  downstream=param_downstream)
```

R ver. 3.1.3(正確にはGenomicFeatures ver. 1.18.7)当時は、まだmakeTxDbFromGFF関数が実装されていないので、①makeTxDbFromGFF関数を見つけることができないのは当たり前。



The screenshot shows the R Console window with the following text:

```

R Console

次のパッケージを付け加えます: 'AnnotationDbi'

The following object is masked from 'package:GenomeInfoDb':
  species

>
> #前処理(アノテーション情報を取得)
> #txdb <- makeTranscriptDbFromGFF(in_f1, #TranscriptDbオブジェクト)
> #                                     format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクト
> txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDbオブジェクト
エラー: 関数 "makeTxDbFromGFF" を見つけることができませんでした
> txdb                                         #確認してるだけです
エラー: オブジェクト 'txdb' がありません
> package.version("GenomicFeatures")
[1] "1.18.7"
> packageVersion("GenomicFeatures")
[1] '1.18.7'
>
```

A red arrow points to the first error message: "エラー: 関数 \"makeTxDbFromGFF\" を見つけることができませんでした".

定期的にバージョンアップ

■ 近年のリリース頻度

- R本体 (<http://www.r-project.org/>)

- 2015-06-18にver. 3.2.1をリリース
- 2015-04-16にver. 3.2.0をリリース
- 2015-03-09にver. 3.1.3をリリース
- 2014-10-31にver. 3.1.2をリリース
- ...
- 2012-03-30にver. 2.15.0をリリース
- ...

- Bioconductor (<http://bioconductor.org/>)は半年ごとにリリース

- 2015-04にver. 3.1をリリース (R ver. 3.2.1で動作確認)、提供パッケージ数 : 1,024
- 2014-10にver. 3.0をリリース (R ver. 3.1.1で動作確認)、提供パッケージ数 : 934
- 2014-04にver. 2.14をリリース (R ver. 3.1.0で動作確認)、提供パッケージ数 : 824
- 2013-10にver. 2.13をリリース (R ver. 3.0で動作確認)、提供パッケージ数 : 750
- 2013-04にver. 2.12をリリース (R ver. 3.0で動作確認)、提供パッケージ数 : 672
- 2012-10にver. 2.11をリリース (R ver. 2.15.1で動作確認)、提供パッケージ数 : 608
- 2012-04にver. 2.10をリリース (R ver. 2.15.0で動作確認)、提供パッケージ数 : 553
- ...

バグの修正や新たな機能がどんどん追加されている。最新版の利用をお勧め。推奨の毎年5月と11月ごろにバージョンアップをちゃんとやったヒトは、R ver. 3.2.0以上なはず。

Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

データの型

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]を用いて、

Ensembl (Flieck et al., Nucleic Acids Res., 2014)から提供されている配列ファイル(*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.core注釈ファイル(*Lactobacillus casei* 12a.GCA_000309565.2.25.gff3)

```

in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.gff3"
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.dna.core注釈"
out_f <- "hoge7.fasta" #出力ファイル名
param_upstream <- 500 #転写開始位置
param_downstream <- 20 #転写終了位置

#必要なパッケージをロード
library(Rsamtools)
library(Biostrings)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDB名
#                                     format="gff3", useGenesAsTranscript)
txdb <- makeTxDbFromGFF(in_f2, format="gff3")#確認してます
txdb #確認します

#前処理(欲しい領域の座標情報を取得)
gn <- sort(genes(txdb)) #遺伝子
hoge <- promoters(gn, upstream=param_upstream,
                  downstream=param_downstream) #指定します
hoge #確認します

```

簡単にいって、「型 = 見た目」です。例えば、①makeTxDbFromGFF関数は、(GFF3形式の)アノテーションファイルを読み込んだものを②TxDbという形式のオブジェクトとして格納しています。重要なのは、細かい中身の理解ではなく、ちゃんと読み込んでいるっぽいことを確認すること

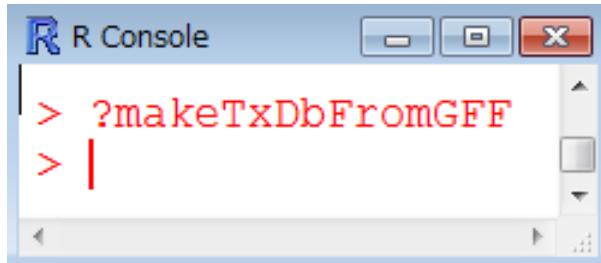
R Console

```

> txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDB名
Prepare the 'metadata' data frame ... metadata: OK
Warning message:
In matchCircularity(seqlevels(gr), circ_seqs) :
  None of the strings in your circ_seqs argument match
> txdb #確認してます
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: Lactobacillus_casei_12a.GCA_000309565.$
# Organism: NA
# miRBase build ID: NA
# Genome: NA
# transcript_nrow: 2799
# exon_nrow: 2800
# cds_nrow: 2681
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2015-07-25 17:25:09 +0900 (Sat, 25 Jul 2015)
# GenomicFeatures version at creation time: 1.20.1
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
>

```

データの型



makeTxDbFromGFF {GenomicFeatures}

R Documentation

Make a TxDb object from annotations available as a GFF3 or GTF file

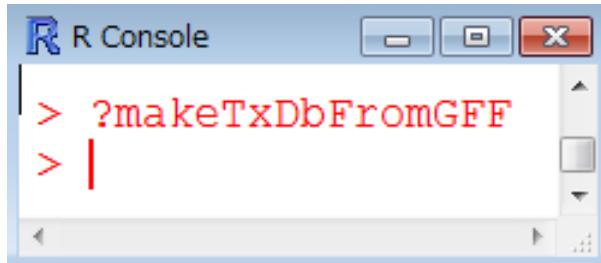
Description

The `makeTxDbFromGFF` function allows the user to make a [TxDb](#) object from transcript annotations available as a GFF3 or GTF file.

Usage

```
makeTxDbFromGFF(file,
                  format=c("gff3", "gtf"),
                  exonRankAttributeName=NA,
                  gffGeneIdAttributeName=NA,
                  chrominfo=NA,
                  dataSource=NA,
                  organism=NA,
                  circ_seqs=DEFAULT_CIRC_SEQS,
                  miRBaseBuild=NA,
                  useGenesAsTranscripts=FALSE,
                  gffTxName="mRNA",
                  species=NA)
```

データの型



①これはGenomicFeaturesパッケージが提供している関数。②GFF3形式以外にGTF形式も読み込めるようだ。③GTF形式ファイルの場合はformat="gtf"と書けばよさそう

makeTxDbFromGFF {GenomicFeatures}

R Documentation

①
Make a TxDb object from annotations available as a GFF3 or GTF file
②

Description

The `makeTxDbFromGFF` function allows the user to make a [TxDb](#) object from transcript annotations available as a GFF3 or GTF file.

Usage

makeTxDbFromGFF(file,
 format=c("gff3", "gtf"),
 exonRankAttributeName=NA,
 gffGeneIdAttributeName=NA,
 chrominfo=NA,
 dataSource=NA,
 organism=NA,
 circ_seqs=DEFAULT_CIRC_SEQS,
 miRBaseBuild=NA,
 useGenesAsTranscripts=FALSE,
 gffTxName="mRNA",
 species=NA)

③

データの型

`makeTxDbFromGFF` is a convenience function that file to the [makeTxDbFromGRanges](#) function.

Value

A [TxDb](#) object. ①

Author(s)

M. Carlson and H. Pages

See Also

- [makeTxDbFromGRanges](#), which `makeTxDbFromGFF` is based on, for making a [TxDb](#) object from a [GRanges](#) object. ②
- The [import](#) function in the **rtracklayer** package (also used by `makeTxDbFromGFF` internally). ③
- [makeTxDbFromUCSC](#) and [makeTxDbFromBiomart](#) for convenient ways to make a [TxDb](#) object from UCSC or BioMart online resources. ④

htmlマニュアルの下部に移動。①`makeTxDbFromGFF`関数実行結果として得られるものはTxDbという形式のオブジェクト。②もちろんR上でアノテーション情報を格納する形式はTxDbだけではなく、GRangesという形式も存在する。GRanges → TxDbへの型変換は`makeTxDbFromGRanges`関数で可能であることがわかる。アノテーション情報源は、③UCSCや④BioMartが有名だが、そこからうまく情報を抽出してTxDbオブジェクトを得るための関数も存在することが分かる。こんな感じで関数の知識や利用の幅を広げていく。

library(help=パッケージ名)

R R Console

```
> library(help=GenomicFeatures)
> |
```

R パッケージ 'GenomicFeatures' のドキュメント

パッケージ 'GenomicFeatures' の情報

記述:

Package:	DEFAULT_CIRC_SEQS	character vector: strings that are usually circular chromosomes
Title:	FeatureDb-class	FeatureDb objects
	TxDb-class	TxDb objects
	asBED, TxDb-method	Coerce to file format structures
Version:	extractTranscriptSeqs	Extract transcript sequences from chromosomes
Author:	extractUpstreamSeqs	Extract sequences upstream of a set of genes or transcripts
	features	Extract simple features from a FeatureDb object
	getPromoterSeq	Get gene promoter sequences
	id2name	Map internal ids to external names for a given feature type
	makeFeatureDbFromUCSC	Making a FeatureDb object from annotations available at the UCSC Genome Browser
	makeTxDb	Making a TxDb object from user supplied annotations

①指定したパッケージ中で利用可能な関数を概観できます。②下のほうに移動すると、③など(Rで)塩基配列解析で紹介していない様々な関数があることがわかります。

パッケージ概観

②Bioconductor (ver. 3.1) 上での GenomicFeatures のトップページ。③ダウンロード数はトップ5%であることがわかる

- ・ イントロ | 一般 | 配列取得 | ゲノム配列 | [BSgenome](#) (last modified 2015/04/22)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [公共DBから](#) (last modified 2014/04/07)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [BSgenomeとTxDbから](#) (last modified 2015/02/20)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/02/20)
- ・ イントロ | 一般 | 配列取得 | リンクスクリプトーム配列 | [公共DBから](#) (last modified 2015/05/09)
- ・ イントロ | 一般 | 配列取得 | リンクスクリプトーム配列 | [BSgenomeから](#) (last modified 2015/05/09)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/05/09)
- ・ イントロ | 一般 | 配列取得 | リンクスクリプトーム配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/05/09)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) NEW

① (1) イントロ | 一般 | 配列取得 | プロモーター配列 | GenomicFeatures(Lawrence_2013) NEW

② (2) GenomicFeatures パッケージを主に用いてプロモーター配列(転写開始点近傍配列)を得るやり方を示します。

ここまでで、**①** イントロ | 一般 | 配列取得 | ゲノム配列(BSgenomeから)で指定可能なゲノムと **②** イントロ | NGS | アノテーション情報取得 | Transcripts が確認できました。

xxx塩基、下は、どの生状の指定が Chr4, Chr5, genome@se 形式のアノ

③ (3) ヒト ("BSgenome")

Home » Bioconductor 3.1 » Software Packages » GenomicFeatures

GenomicFeatures

platforms all downloads top 5% posts 32 / 2 / 3 / 8 in Bioc 5.5 years
 build ok commits 15.17 test coverage 70%

Tools for making and manipulating transcript centric annotations

Bioconductor version: Release (3.1)
 A set of tools and methods for making and manipulating transcript centric annotations. With these tools

パッケージ概観

Author: M. Carlson, H. Pages, P. Aboyoun, S. Falcon, M.

Maintainer: Bioconductor Package Maintainer <maintainer@bioconductor.org>

少し下のほうに移動。①原著論文はきっちり引用。②パッケージのインストール法。③パッケージのドキュメント(マニュアルではない)をローカル環境で開く方法。④具体的なやり方。④の中身は⑤と同じだが、ネットワークかローカルかの違いがある。⑥利用可能な関数マニュアルはここでも見られる。

Citation (from within R, enter `citation("GenomicFeatures")`):

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M and Carey V (2013). "Software for Computing and Annotating Genomic Ranges." *PLoS Computational Biology*, 9. <http://dx.doi.org/10.1371/journal.pcbi.1003118>, <http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003118>



Installation

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("GenomicFeatures")
```



Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("GenomicFeatures")
```



```
R Console
> browseVignettes("GenomicFeatures")
starting httpd help server ... done
> |
```



目的をおさらい

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロ

Ensembl (Flieck et al., Nucleic Acids Res., 2014)から提供されている *Lactobacillus casei* 12Aのゲノム配列ファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.chromosome.1.fna) とアノテーションファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.gff3)

R Console

```
in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.1.fna"
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.gff3"
out_f <- "hoge7.fasta" #出力ファイル名
param_upstream <- 500 #転写開始位置
param_downstream <- 20 #転写終了位置

#必要なパッケージをロード
library(Rsamtools) #パッケージ名
library(Biostrings) #パッケージ名
library(GenomicFeatures) #パッケージ名

#前処理(アノテーション情報を取得)
txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクトの作成
#                                     format="gff3", useGenesAsTranscripts=TRUE)
txdb <- makeTxDbFromGFF(in_f2, format="gff3") #確認
#txdb #確認してます

#前処理(欲しい領域の座標情報を取得)
gn <- sort(genes(txdb)) #遺伝子リスト
hoge <- promoters(gn, upstream=param_upstream,
                  downstream=param_downstream) #指定した範囲
hoge #確認
```

乳酸菌*L. casei* 12A株のゲノム配列とアノテーションファイルを読み込んで、全遺伝子の上流500塩基、および下流20塩基分の塩基配列を抽出してプロモーター解析(するための入力データを取得)したい! 黒枠までで行ったことは、アノテーションファイルを読み込んでtxdbオブジェクトに格納したところまで。

```
> txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDbオブジェクトの作成
Prepare the 'metadata' data frame ... metadata: OK
Warning message:
In matchCircularity(seqlevels(gr), circ_seqs) :
  None of the strings in your circ_seqs argument match
> txdb #確認してます
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: Lactobacillus_casei_12a.GCA_000309565.2.25
# Organism: NA
# miRBase build ID: NA
# Genome: NA
# transcript_nrow: 2799
# exon_nrow: 2800
# cds_nrow: 2681
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2015-07-25 17:25:09 +0900 (Sat, 25 Jul 2015)
# GenomicFeatures version at creation time: 1.20.1
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
> |
```

残りのコードを概説

#前処理(アノテーション情報を取得)

```
#txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクト
#           format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクト
txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDbオブジェクト
txdb #確認してるだけ
```



#前処理(欲しい領域の座標情報取得)

```
gn <- sort(genes(txdb)) #遺伝子の座標情報
hoge <- promoters(gn, upstream=param_upstream, #指定した範囲の座標
                  downstream=param_downstream) #指定した範囲の座標
hoge #確認してるだけ
obj <- (ranges(hoge)$start >= 0) #条件を満たすかどうかを判定した結果をobjに格納(座標が0よりも大きい)
hoge <- hoge[obj] #objがTRUEとなる要素のみ抽出した結果をhogeに格納
hoge #確認してるだけ
```



#本番(配列取得)

```
fasta <- getSeq(FaFile(in_f1), hoge) #指定した範囲の塩基配列情報を取得
names(fasta) <- names(hoge) #description情報を追加している
fasta #確認してるだけ
```



#ファイルに保存

```
writeXStringSet(fasta, file=out_f, format="fasta", width=50) #fastaの中身を指定したファイル名で保存
```

①TxDbオブジェクト(オブジェクト名txdb)を出発点として、欲しい領域(上流500塩基、下流20塩基)情報を取得。②ゲノム配列情報FaFile(in_f1)と欲しい領域情報hogeを入力として塩基配列を取得。③FASTA形式ファイルで保存。このあたりは、H26年度受講生追跡調査の「配列情報を使った次の解析へのinput fileの加工が難関だと感じている」に対する回答。特にプロモーター解析などは「…で、その入力ファイルをどうやって作るの?」かが実務担当者の直面する課題。

データの型

#前処理(アノテーション情報を取得)

```
#txdb <- makeTranscriptDbFromGFF(in_f1,
#                                     format="gff3", useGenesAsTxns=TRUE)
txdb <- makeTxdbFromGFF(in_f2, format="gff3")
txdb
```

#前処理(欲しい領域の座標情報を取得)

```
gn <- sort(genes)
```

②

```
hoge <- promoters(gn, upstream=param_upstream, #指定した範囲の座標情報を取得
                   downstream=param_downstream) #指定した範囲の座標情報を取得
```

```
hoge
obj <- (ranges(hoge)$start >= 0)
hoge <- hoge[obj]
hoge
```

①黒枠部分は感覚的にpromotersという関数を用いて転写開始点上流と下流の範囲を取得した結果をhogeに格納してるんだろうな、くらいは分かりますよね。そしてなぜ②のところでtxdbを使っていないのか?という素朴な疑問があるでしょう。理由はpromoters関数が入力としてTxDbオブジェクトを想定していないからです(おそらく2014年10月リリースのBioconductor ver. 3.0までは正解)。2015年4月リリースのBioconductor ver. 3.1では、GenomicFeaturesパッケージから提供されているpromoters関数はTxDbに対応済み。

#指定した範囲の座標情報を取得

①

#確認してるだけです

#条件を満たすかどうかを判定した結果をobjに格納(座標が0よりも小)

#objがTRUEとなる要素のみ抽出した結果をhogeに格納

#確認してるだけです

#本番(配列取得)

```
fasta <- getSeq(FaFile(in_f1), hoge)
names(fasta) <- names(hoge)
fasta
```

#指定した範囲の塩基配列情報を取得

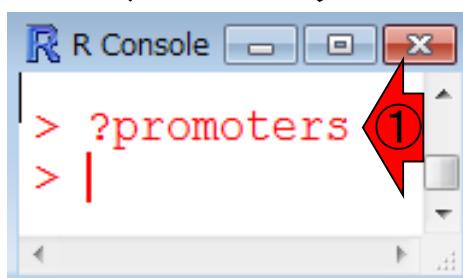
#description情報を追加している

#確認してるだけです

#ファイルに保存

```
writeXStringSet(fasta, file=out_f, format="fasta", width=50) #fastaの中身を指定したファイル名で保存
```

データの型



①「?promoters」。門田のR ver. 3.2.0環境では、`promoters`という同じ名前の関数が3つのパッケージ(`IRanges`, `GenomicFeatures`, and `GenomicRanges`)から提供されていることがわかる。②`GenomicRanges`パッケージから提供されている`promoters`関数は入力としてTxDbオブジェクトを受け付けないことがわかる。しかし、③`GenomicFeatures`パッケージから提供されている`promoters`関数は、TxDbを入力としていることが2015年7月26日に判明(爆)。

Help on topic 'promoters' was found in the following packages:

[Intra range transformations of a Ranges, Views, RangesList, or MaskCollection object](#)

(in package `IRanges` in library C:/Users/kadota/Documents/R/win-library/3.2)

[Extract genomic features from an object](#) ③

(in package `GenomicFeatures` in library C:/Users/kadota/Documents/R/win-library/3.2)

[Intra range transformations of a GRanges or GRangesList object](#)

(in package `GenomicRanges` in library C:/Users/kadota/Documents/R/win-library/3.2)

②

③

データの型

③GenomicFeaturesパッケージから提供されているpromoters関数は、④TxDbを入力としていることが2015年7月26日に判明(爆)。

Help on topic 'promoters' was found in the following packages:

[Intra range transformations of a Ranges, Views, RangesList, or MaskCollection object](#)
(in package [IRanges](#) in library C:/Users/kadota/Documents/R/win-library/3.2)

[Extract genomic features from an object](#) ③
(in package [GenomicFeatures](#) in library C:/Users/kadota/Documents/R/win-library/3.2)

[Intra range transformations of a GRanges or GRangesList object](#)
(in package [Genom](#)

transcripts {GenomicFeatures}

R Documentation

Extract genomic features from an object

Description

④

Generic functions to extract genomic features from an object. This page documents the methods for [TxDb](#) objects only.

Usage

```
transcripts(x, ...)  
## S4 method for signature 'TxDb'  
transcripts(x, vals=NULL, columns=c("tx_id", "tx_name"))  
  
exons(x, ...)  
## S4 method for signature 'TxDb'  
exons(x, vals=NULL, columns="exon_id")  
  
cds(x, ...)  
## S4 method for signature 'TxDb'  
cds(x, vals=NULL, columns="cds_id")  
  
genes(x, ...)  
## S4 method for signature 'TxDb'  
genes(x, vals=NULL, columns="gene_id", single.strand.genes.only=TRUE)  
  
## S4 method for signature 'TxDb'  
promoters(x, upstream=2000, downstream=200, ...)
```

⑤

プロモーター配列取得

②例題8に、③TxDbオブジェクトを入力としてpromoters関数を実行するコードを追加。1つ上の行は過去の遺物(legacy)として残しているだけ。(Rで)塩基配列解析自体がデカすぎて、もはや全体の動作確認もままならないし、修正・変更もしきれていないのが現状

- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [BSgenomeとTxDbから](#) (last modified 2015/04/21)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/04/21)
- ・ イントロ | 一般 | 配列取得 | トランск립トーム配列 | [公共DBから](#) (last modified 2015/04/21)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/04/21)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/04/21)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/04/21)
- ・ イントロ | 一般 | 配列取得 | プロモーター配列 | [GenomicFeatures\(Lawrence 2013\)](#) (last modified 2015/04/21)

[GenomicFeatures](#)パッケージを主に用いてプロモーター配列(転写開始点近傍配列)を得るやり方を示します。

ここでは、[TranscriptDb](#)

xxx塩基、下流0塩基は、どの生物種に状の指定が可能
Chr4, Chr5, ChrM
genome@seqinfo
形式のアノテーション

1. ヒト ("BSgenome
ヒトゲノム ver 1.0")

8. 乳酸菌ゲノム(*Lactobacillus casei* 12a)の[上流100塩基、下流0塩基]のプロモーター配列を取得する場合:

7.と基本的に同じですが、少なくとも2015年4月リリースのBioconductor ver. 3.1以降で、[GenomicFeatures](#)パッケージが提供するpromoters関数がTxDbオブジェクトをそのまま読み込めることが判明したので、それを反映させています。

```

in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル
out_f <- "hoge8.fasta" #出力ファイル名を指定してout_fに格納
param_upstream <- 100 #転写開始点上流の塩基配列数を指定
param_downstream <- 0 #転写開始点下流の塩基配列数を指定

#必要なパッケージをロード
library(Rsamtools) #パッケージの読み込み
library(Biostrings) #パッケージの読み込み
library(GenomicFeatures) #パッケージの読み込み

#前処理(アノテーション情報を取得)
#txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクトを取得してtxdbに格納
#                                    format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクトを取得してtxdbに格納
txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDbオブジェクトを取得してtxdbに格納
txdb #確認してるだけです

#前処理(欲しい領域の座標情報を取得)
#gn <- sort(genes(txdb)) #遺伝子の座標情報を取得
hoge <- promoters(txdb, upstream=param_upstream, #指定した範囲の座標情報を取得
                  downstream=param_downstream) #指定した範囲の座標情報を取得
hoge #確認してるだけです

```

②

①

③

遺言1

- R本体とパッケージのバージョンアップを毎年5月と11月に行うべし
 - 依存関係など様々な問題はあるものの、落ち着いて必要なパッケージを個別にインストールしていけば大丈夫
 - 第一義的に重要なのは、Linuxでプログラムのインストールができることと同様、library(XXX)の部分でこけないようにすること。

- [はじめに \(last modified 2015/03/31\)](#)
- [参考資料\(講義、講習会、本など\) \(last modified 2015/07/07\) NEW](#)
- [過去のお知らせ \(last modified 2015/07/08\) NEW](#)
- [インストールについて \(last modified 2015/04/04\)](#)
- インストール | R本体 | 最新版 | [Win用 \(last modified 2015/03/22\) 推奨](#)
- インストール | R本体 | 最新版 | [Mac用 \(last modified 2015/04/22\) 推奨](#)
- インストール | R本体 | 過去版 | [Win用 \(last modified 2015/03/22\)](#)
- インストール | R本体 | 過去版 | [Mac用 \(last modified 2015/03/22\)](#)
- インストール | Rパッケージ | [ほぼ全て\(20GB以上?!\) \(last modified 2015/05/25\)](#)
- インストール | Rパッケージ | [必要最小限プラスアルファ\(数GB?!\) \(last modified 2015/07/24\) 推奨 NEW](#)
- インストール | Rパッケージ | [必要最小限プラスアルファ\(アグリバイ[†]居室のみ\) \(last modified 2015/06/16\)](#)
- インストール | Rパッケージ | [必要最小限\(数GB?!\) \(last modified 2015/05/25\)](#)
- インストール | Rパッケージ | [個別 \(last modified 2015/06/10\)](#)

②

①

遺言2

R最新版でうまくいったなら、過去版のエラー原因究明は時間の無駄!

- ①R ver. 3.1.2のときはmakeTranscriptDbFromGFF関数でうまくいくはずだが、エラーが出る現象に遭遇した(2015.07.07のアグリバイオ大学院講義)。しかし、②R ver. 3.2.1でmakeTxDbFromGFF関数で正常動作することがわかり、その時点で思考停止

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモーター配列を取得する場合:

[Ensembl](#) (Flieck et al., Nucleic Acids Res., 2014)から提供されている *Lactobacillus casei* 12A の multi-FASTA形式ゲノム配列ファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa) と GFF3形式のアノテーションファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3) を読み込むやり方です。

```

in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル
out_f <- "hoge7.fasta" #出力ファイル名を指定してout_fに格納
param_upstream <- 500 #転写開始点上流の塩基配列数を指定
param_downstream <- 20 #転写開始点下流の塩基配列数を指定

#必要なパッケージをロード
library(Rsamtools) #パッケージの読み込み
library(Biostrings) #パッケージの読み込み
library(GenomicFeatures) #パッケージの読み込み

#前処理(アノテーション情報を取得)
#txdb <- makeTranscriptDbFromGFF(in_f2, #TranscriptDbオブジェクトを取得してtxdbに格納
#                                    #format="gff3", useGenesAsTranscripts=T) #TranscriptDbオブジェクトを取得してtxdbに格納
① txdb <- makeTxDbFromGFF(in_f2, format="gff3") #TranscriptDbオブジェクトを取得してtxdbに格納
       #確認してるだけです

#前処理(欲しい領域の座標情報を取得)
gn <- sort(genes(txdb)) #遺伝子の座標情報を取得
hoge <- promoters(gn, upstream=param_upstream, #指定した範囲の座標情報を取得
                  downstream=param_downstream) #指定した範囲の座標情報を取得
hoge #確認してるだけです

```

②

①

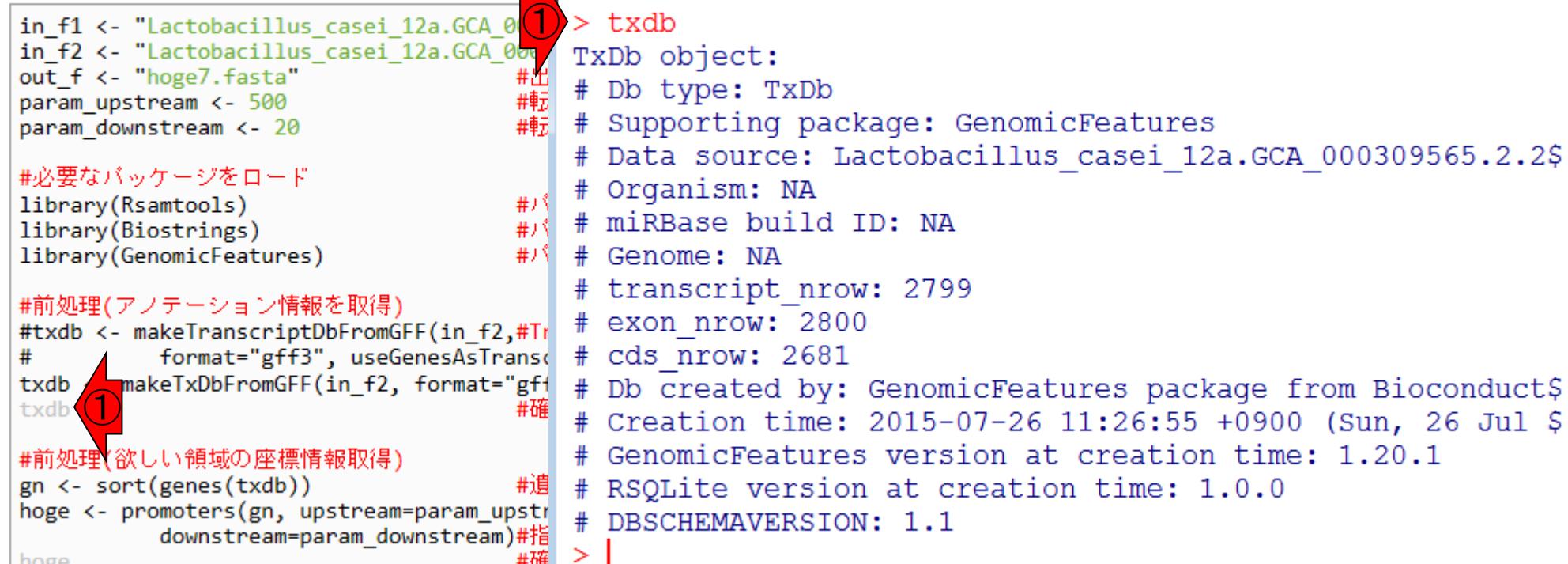
遺言3

■ 得られた結果を無条件に受け入れるな！様々な角度で検証せよ。

- 実はプログラムのバグで、GFF3形式ファイルの読み込みに失敗していただけ、みたいなオチもよくあるだろう。フリーソフトの利用は自己責任である。

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモーター配列を取得する場合:

Ensembl (Flieck et al., Nucleic Acids Res., 2014)から提供されている *Lactobacillus casei* 12A の multi-FASTA 形式ゲノム配列ファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.dna_chromosome.Chromosome.fna) と GFF3 形式のアノテーションファイル (*Lactobacillus casei* 12a.GCA_000309565.gff3)



The screenshot shows an R console window with the title "R Console". The code in the console reads a GFF3 file ("in_f1") and creates a TxDb object ("txdb"). The TxDb object is a Txdb object with various attributes: Db type: Txdb, Supporting package: GenomicFeatures, Data source: Lactobacillus_casei_12a.GCA_000309565.2.25, Organism: NA, miRBase build ID: NA, Genome: NA, transcript_nrow: 2799, exon_nrow: 2800, cds_nrow: 2681, Db created by: GenomicFeatures package from Bioconduct\$, Creation time: 2015-07-26 11:26:55 +0900 (Sun, 26 Jul \$, GenomicFeatures version at creation time: 1.20.1, RSQLite version at creation time: 1.0.0, DBSCHEMAVERSION: 1.1.

```

in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.gff3"
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.dna_chromosome.Chromosome.fna"
out_f <- "hoge7.fasta" #出力用ファイル名
param_upstream <- 500 #転写開始位置からの上流距離
param_downstream <- 20 #転写終了位置からの下流距離

#必要なパッケージをロード
library(Rsamtools)
library(Biostrings)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
#txdb <- makeTranscriptDbFromGFF(in_f2, #Tr #アノテーション情報を取得
#                                    format="gff3", useGenesAsTranscripts=TRUE)
txdb <- makeTxDbFromGFF(in_f2, format="gff3") #確認用
txdb #確認用

#前処理(欲しい領域の座標情報を取得)
gn <- sort(genes(txdb)) #遺伝子情報を取得
hoge <- promoters(gn, upstream=param_upstream, downstream=param_downstream) #指定した領域のプロモーター情報を取得
hoge #確認用

```

Rを用いた検証例

7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモーター配列を取得

Ensembl (Flieck et al., Nucleic Acids Res., 2014)から提供されている *Lactobacillus casei* 12A の multi-FASTA 形式のノム配列ファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa) と GFF3 形式のアノテーションファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3) を読み込むやり方です。

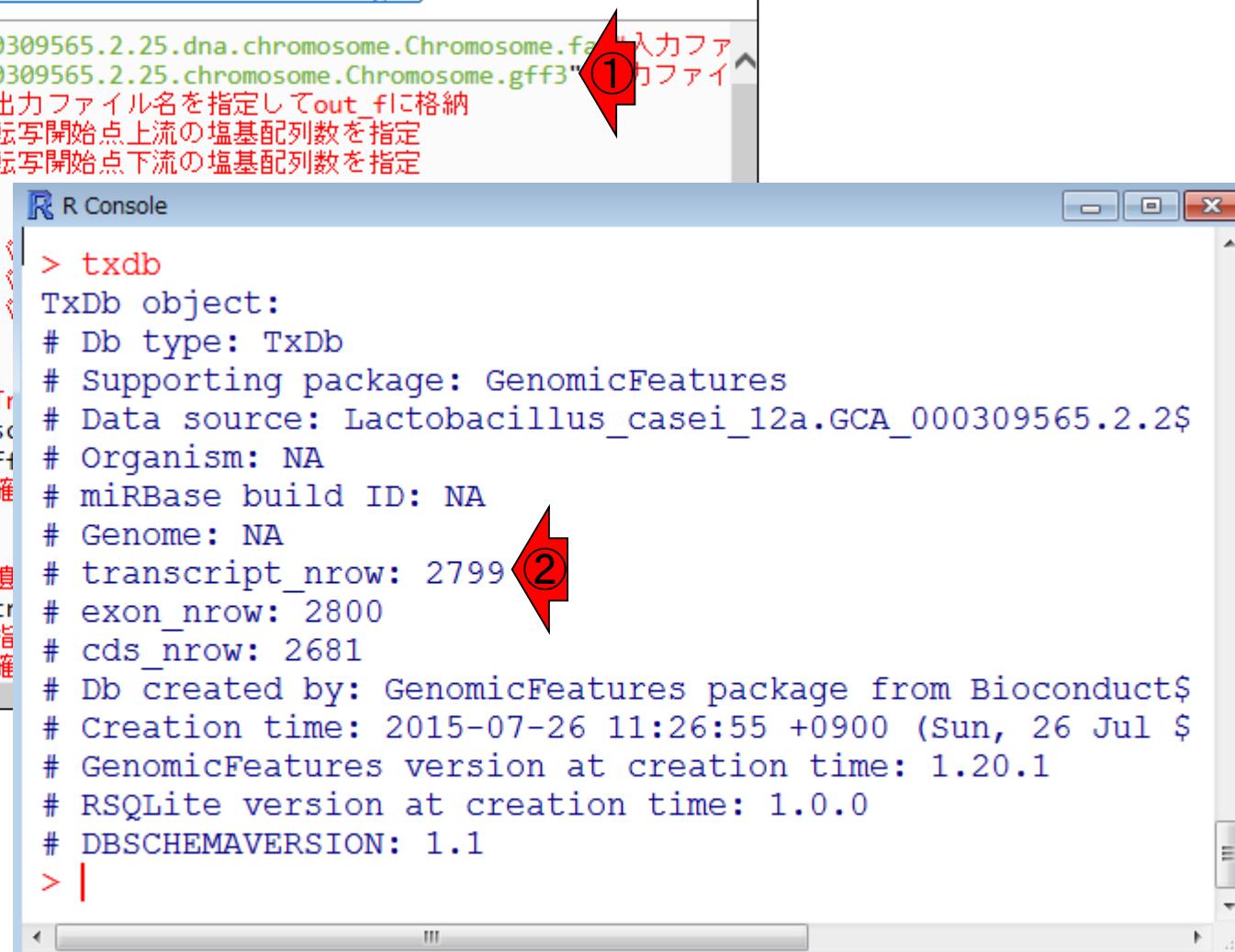
```
in_f1 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa"
in_f2 <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3"
out_f <- "hoge7.fasta"
param_upstream <- 500
param_downstream <- 20

#必要なパッケージをロード
library(Rsamtools)
library(Biostrings)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
#txdb <- makeTranscriptDbFromGFF(in_f2, #Tr
#    format="gff3", useGenesAsTransc
txdb <- makeTxDbFromGFF(in_f2, format="gff3")
txdb #確認

#前処理(欲しい領域の座標情報を取得)
gn <- sort(genes(txdb)) #遺伝子
hoge <- promoters(gn, upstream=param_upstr
                      downstream=param_downstream) #指定
hoge #確認
```

txdbオブジェクトの元データは、①の GFF3 ファイル。②乳酸菌 *L. casei* 12A の遺伝子数 (= 転写物数) は、2,799 個となっている。これが正しいかを検証する。



The screenshot shows the R Console window with the following output:

```
R R Console
> txdb
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: Lactobacillus_casei_12a.GCA_000309565.2.25
# Organism: NA
# miRBase build ID: NA
# Genome: NA
# transcript_nrow: 2799
# exon_nrow: 2800
# cds_nrow: 2681
# Db created by: GenomicFeatures package from Bioconduct
# Creation time: 2015-07-26 11:26:55 +0900 (Sun, 26 Jul 2015)
# GenomicFeatures version at creation time: 1.20.1
# RSQLite version at creation time: 1.0.0
# DBSCHEMAVERSION: 1.1
>
```

Two red arrows point to specific parts of the output:

- ① Points to the line "# transcript_nrow: 2799".
- ② Points to the line "# Organism: NA".

Rを用いた検証例

(9列目の)"ID=gene"という文字列を含む行数を調べてみると、いのでは、という思想のもので検証する。

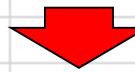
7. 乳酸菌ゲノム(*Lactobacillus casei* 12A)の[上流500塩基, 下流20塩基]のプロモーター配列を取得する場合:

Ensembl (Flieck et al., Nucleic Acids Res., 2014)から提供されている *Lactobacillus casei* 12A の multi-FASTA 形式ゲノム配列ファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa) と GFF3 形式のアノテーションファイル (*Lactobacillus casei* 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3) を読み込むやり方です。

```
in_f1 <- "Lactobacillus casei 12a.GCA_000309565.2.25.dna.chromosome.Chromosome.fa" #入力ファイル
in_f2 <- "Lactobacillus casei 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル
out_f <- "hoge7.fasta" #出力ファイル名を指定してout_fに格納
```

```
param ##gff-version 3
param ##sequence-region Chromosome 1 2907892
```

#必要	Chromosome	ena	gene	1	1350	.	+	ID=gene:LCA12A_0617;assembly_name=GCA_000309565.2;b
library	Chromosome	ena	gene	1523	2662	.	+	ID=gene:LCA12A_0618;assembly_name=GCA_000309565.2;b
library	Chromosome	ena	gene	3240	3452	.	+	ID=gene:LCA12A_0619;assembly_name=GCA_000309565.2;b
#前処理	Chromosome	ena	gene	3449	4564	.	+	ID=gene:LCA12A_0620;assembly_name=GCA_000309565.2;b
#txdb	Chromosome	ena	gene	4817	6778	.	+	ID=gene:LCA12A_0621;assembly_name=GCA_000309565.2;b
#txdb	Chromosome	ena	gene	6840	9461	.	+	ID=gene:LCA12A_0622;assembly_name=GCA_000309565.2;b
#前処理	Chromosome	ena	gene	9566	10270	.	-	ID=gene:LCA12A_0623;assembly_name=GCA_000309565.2;b
gn <- hoge	Chromosome	ensembl	CDS	1	1350	.	+	0 ID=CDS:EKP96483;Parent=transcript:EKP96483;assembly_n
hoge <	Chromosome	ensembl	exon	1	1350	.	+	Name=EKP96483-1;Parent=transcript:EKP96483;assembly_n
hoge <	Chromosome	ensembl	CDS	1523	2662	.	+	0 ID=CDS:EKP96484;Parent=transcript:EKP96484;assembly_n
hoge <	Chromosome	ensembl	exon	1523	2662	.	+	Name=EKP96484-1;Parent=transcript:EKP96484;assembly_n
hoge <	Chromosome	ensembl	CDS	3240	3452	.	+	0 ID=CDS:EKP96485;Parent=transcript:EKP96485;assembly_n
hoge <	Chromosome	ensembl	exon	3240	3452	.	+	Name=EKP96485-1;Parent=transcript:EKP96485;assembly_n
hoge <	Chromosome	ensembl	CDS	3449	4564	.	+	0 ID=CDS:EKP96486;Parent=transcript:EKP96486;assembly_n
hoge <	Chromosome	ensembl	exon	3449	4564	.	+	Name=EKP96486-1;Parent=transcript:EKP96486;assembly_n
hoge <	Chromosome	ensembl	CDS	4817	6778	.	+	0 ID=CDS:EKP96487;Parent=transcript:EKP96487;assembly_n
hoge <	Chromosome	ensembl	exon	4817	6778	.	+	Name=EKP96487-1;Parent=transcript:EKP96487;assembly_n
hoge <	Chromosome	ensembl	CDS	6840	9461	.	+	0 ID=CDS:EKP96488;Parent=transcript:EKP96488;assembly_n



Rを用いた検証例

- ・ 書籍 | 日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール
- ・ イントロ | 一般 | ランダムに行を抽出 (last modified 2014/07/17)
- ・ イントロ | 一般 | 任意の文字列を行の最初に挿入 (last modified 2014/07/17)
- ・ イントロ | 一般 | 任意のキーワードを含む行を抽出(基礎) (last modified 2014/04/11)
- ・ イントロ | 一般 | ランダムな塩基配列を生成 (last modified 2014/06/16)
- ・ イントロ | 一般 | 任意の長さの可変性塩基配列を作成 (last modified 2015/02)

② 14. GFF3形式のタブ区切りテキストファイル(Lactobacillus casei 12a.GCA_000309565.2.25.chromosome.Chromosome.gff3)に対して、"ID=gene"という文字列が含まれる行全体を出力したい場合:

乳酸菌ゲノム(Lactobacillus casei 12A)のアノテーションファイルです。4.をベースに作成。

```

in_f <- "Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3" #入力ファイル名を
out_f <- "hoge14.txt" #出力ファイル名を指定してout_fに格納
param <- "ID=gene" #検索したい文字列を指定

#入力ファイルの読み込み
data <- readLines(in_f)
length(data)

#本番(リストファイル中の要素一つ一つに対して)
hoge <- sapply(param, grep, x=data)
hoge <- unique(hoge)
out <- data[hoge]
length(out)

#ファイルに保存
writeLines(out, out_f)

```

txdbオブジェクトの元データは、①のGFF3ファイル。乳酸菌L. casei 12Aの遺伝子数(=転写物数)は、2,799個となっている。Rで③"ID=gene"を含む行数も④2,799個だったので、大丈夫だろうと判断する。

R Console

```

# > data <- readLines(in_f) #in_fで指定
# > length(data) #オブジェクト
# [1] 11081 #リスト
# >
# > #本番(リストファイル中の要素一つ一つに対して、要素)
# > hoge <- sapply(param, grep, x=data) #得られるhoge
# > hoge <- unique(hoge) #hogeで指定
# > out <- data[hoge] #オブジェクト
# > length(out) #outの中身
# [1] 2799
# >
# > #ファイルに保存
# > writeLines(out, out_f)
# > |

```

Linuxを用いた検証例

iu@bielinux[~/Desktop/mac_share]

```
iu@bielinux[mac_share] pwd  
/home/iu/Desktop/mac_share
```

```
iu@bielinux[mac_share] ls
```

Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3

```
iu@bielinux[mac_share] mv Lactobacillus_casei_12a.GCA_000309565.2.25.ch  
romosome.Chromosome.gff3 hoge.gff3
```

```
iu@bielinux[mac_share] ls
```

hoge.gff3

```
iu@bielinux[mac_share] grep -c ID=gene hoge.gff3
```

2799

```
iu@bielinux[mac_share] ■
```

Linuxをある程度使えるヒトは、普通grepコマンドを使います。①のmvコマンドは、ファイル名が長いので、hoge.gff3にrenameしているだけです。②の-cは、③で指定した文字列と一致した行数を表示させるオプション。

[3:08 午後]

[3:08 午後]

[3:08 午後]

[3:08 午後]



Linuxを用いた検証例

iu@bielinux[~/Desktop/mac_share]

```
iu@bielinux[mac_share] pwd  
/home/iu/Desktop/mac_share  
iu@bielinux[mac_share] ls
```

Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3

```
iu@bielinux[mac_share] mv Lactobacillus_casei_12a.GCA_000309565.2.25.chromosome.Chromosome.gff3 hoge.gff3
```

```
iu@bielinux[mac_share] ls  
hoge.gff3
```

```
iu@bielinux[mac_share] grep -c ID=gene hoge.gff3  
2799
```

```
iu@bielinux[mac_share] grep -cv ID=gene hoge.gff3  
8282
```

```
iu@bielinux[mac_share] grep -c gene hoge.gff3  
5598
```

```
iu@bielinux[mac_share] grep -c CDS hoge.gff3  
2681
```

```
iu@bielinux[mac_share] grep -c exon hoge.gff3  
2807
```

```
iu@bielinux[mac_share] grep -c ^Chromosome hoge.gff3  
11079
```

iu@bielinux[mac_share] ■

①-vは、一致しないものを表示するオプション。
-cvと併用することで、一致しない行数を表示。
②Linuxだと他の気になるキーワードも簡単に
調査可能。③^をつけることで「Chromosomeを
含む行数」ではなく「行頭にChromosomeを含
む行数」をカウントすることもできる。

[3:19 午後]

[3:20 午後]

[3:20 午後]



Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

FASTQファイル解析

①ファイル形式の変換(FASTQ → FASTA)や、FastQCのいくつかの項目と同じ解析をRでも実行可能である:②Overrepresented sequences、③Per base sequence content(リードのpositionごとの出現確率)、④Kmer Content。

- イントロ | NGS | 読み込み | FASTQ形式 | 基礎 | ④ modified 2015/06/24)
- イントロ | NGS | 読み込み | FASTQ形式 | 応用 | (last modified 2015/06/18)
- イントロ | NGS | 読み込み | FASTQ形式 | description行の記述を整形 | (last mo
- イントロ | NGS | 読み込み | Illuminaの * seq.txt | (last modified 2013/06/13)
- イントロ | NGS | 読み込み | Illuminaの * qseq.txt | (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | について | (last modified 2014/06/09)
- イントロ | ファイル形式の変換 | BAM --> BED | (last modified 2014/06/21)
- イントロ | ファイル形式の変換 | FASTQ --> FASTA | ① modified 2015/06/15)
- イントロ | ファイル形式の変換 | Genbank --> FASTA | (last modified 2014/03/10)
- イントロ | ファイル形式の変換 | qseq --> FASTA | (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | qseq --> Illumina FASTQ | (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | qseq --> Sanger FASTQ | (last modified 2013/08/19)
- 前処理 | クオリティコントロール | について | (last modified 2014/06/25)
- 前処理 | クオリティチェック | QuasR(Gaidatzis 2015) | ③ modified 2015/06/15)
- 前処理 | クオリティチェック | qrqc | (last modified 2014/06/17)
- 前処理 | クオリティチェック | PHREDスコアに変換 | (last modified 2013/06/18)
- 前処理 | クオリティチェック | 配列長分布を調べる | (last modified 2015/06/22)
- 前処理 | クオリティチェック | Overrepresented sequences | ShortRead(Morgan 2009) | ② modified 2015/07/21)
- 前処理 | トリミング | ポリA配列除去 | ShortRead(Morgan 2009) | (last modified 2014/06/11)
- 前処理 | トリミング | アダプター配列除去(基礎) | QuasR(Gaidatzis 2015) | (last modified 2015/06/26)推奨
- 前処理 | トリミング | アダプター配列除去(基礎) | girafe(Toedling 2010) | (last modified 2014/06/11)
- 前処理 | トリミング | アダプター配列除去(基礎) | ShortRead(Morgan 2009) | (last modified 2014/06/21)
- 前処理 | トリミング | アダプター配列除去(応用) | QuasR(Gaidatzis 2015) | (last modified 2015/06/26)推奨
- 前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan 2009) | (last modified 2014/06/18)
- 前処理 | トリミング | 指定した末端塩基数だけ除去 | (last modified 2015/06/29) NEW

FastQC実行結果

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

✖ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CCCCGGTATATTTCGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCA	14383	1.4383	No Hit
GGCCTATTCACTGCGGCTGACCTTGCAGTCAGCACCCCTTCTCCGAAGT	11044	1.1044	No Hit
GTGCTTTCACCTTCCCTCACGGTACTGGTTACTATCGGTCACTAGGG	8892	0.8892000000000001	No Hit
CCCGGTATATTTCGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCAC	8474	0.8474	No Hit
GTCACTAGGGAGTATTTAGCCTGGGAGATGGTCCTCCCGGATTCCGACG	8189	0.8189	No Hit
GCGGCATTCTCACTTCTAACCGCTCCAGCCGTCTCACGATCGACCTTC	8132	0.8132	No Hit
GTCCAGTCTACAACCCCGAGAAGCAAGCTTCGGTTGGCTTCCC	6663	0.6663	No Hit
GTCGGTTGCGGTACGGTAGTTATTCTCACTAGAAAGCTTCTGGC	6411	0.6411	No Hit
GGTCACTTGGTTTGGGTCTACATCTGCTTACTCATCGCCCTGTTAGA	5502	0.5502	No Hit
GCGGCATTCTCACTTCTAACCGCTCCAGCCGTCTCACGATCAACCTTC	4845	0.4845000000000004	No Hit
CCCTCCATCGCTAAACAAAATAAAACTAGTGCAGGAATCTAACCTGCTT	4395	0.4394999999999995	No Hit
CCGGTATATTTCGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCACT	4385	0.4385	No Hit
CCCGCGTCTGCCGCCGCCAGCTATGTATTCACTGACAAGCAATACACTG	4366	0.4366	No Hit
CCACAGTTGGTATTATGCTTAGCCCCGGTATATTTCGGCGCAGTGCC	4314	0.4314	No Hit
CTGGGCTGTTCCCTTCAACAATGGACCTTATCGCTCACTGTCTGACTC	4113	0.4113000000000005	No Hit
CCGCCGTACTCAGGATCCTGGACGGAGGGTTCGACGTTCGCTTACAGGG	4081	0.4081	No Hit
CCGGCATTCTCACTTCTAACCGCTCCAGCCGTCTCACGATCGACCTCA	3846	0.3846	No Hit
GTAGGTCACTTGGTTGGGTCTACATCTGCTTACTCATCGCCCTGTT	3823	0.3823	No Hit

①頻出する配列をリストアップ。②トップは「CCCCGGTATA…」という50塩基の配列で14,383回出現。Percentageは1.4383%。全部で100万リードなので妥当。オリジナル107 bpのうち最初の50 bpで解析している。復習

subseqとtable関数の併用で同じ結果が得られる例として提示。これもただの復習。

ShortRead実行結果

- 前処理 | クオリティチェック | 配列長分布を調べる (last modified 2015/06/22)
- 前処理 | クオリティチェック | Overrepresented sequences | ShortRead(Morgan 2009) ①
- 前処理 | トリミング | ポリA配列除去 | ShortRead(Morgan 2009) (last modified 2014/06/11)
- 前処理 | トリミング | マダガスカル配列除去(基礎) | OverR(Caidatzic 2015) (last modified 2015/06/22)
- 前処理
- 前処理
- 前処理

前処理 | クオリティチェック | Overrepresented sequences | ShortRead

ShortRead

じょうな組
「ファイル」

1. gzip圧縮

small RNA

```
in_f <-  
out_f <-  
#必要なパッケージをロード  
library()  
#入力ファイルの読み込み  
fastq <-  
sread(fa
```

乳酸菌RNA-seqデータSRR616268の最初の100万リード分(約73MB)です。長さは全て107 bpです。3.と基本的に同じですが、FastQCのデフォルトと同じく最初の500万分のみで解析オススメの方です。「インポート」→「一般」→「指定した範囲の配列を取得」のテクニック

R Console

```
> head(out) #確認してるだけ$  
fasta  
CCCCGGTATATTTCGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCA  
14383  
GGCCTATTCACTGCGGCTGACCTTGCAGCACCCCTTCTTCCGAAGT  
11044  
GTGCTTTCACCTTCCCTCACGGTACTGGTTCACTATCGTCACTAGGG  
8892  
CCCGGTATATTTCGGCGCAGTGCCACTCGACTAGTGAGCTATTACGCAC  
8474  
GTCACTAGGGAGTATTAGCCTTGGGAGATGGTCCTCCGGATTCCGACG  
8189  
GCCGGCATTCTCACTTCTAACAGCGCTCCAGCCGTCCACGATCGACCTC  
8132  
>  
> #ファイルに保存  
> tmp <- cbind(names(out), out) #保存したい情報$  
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row
```

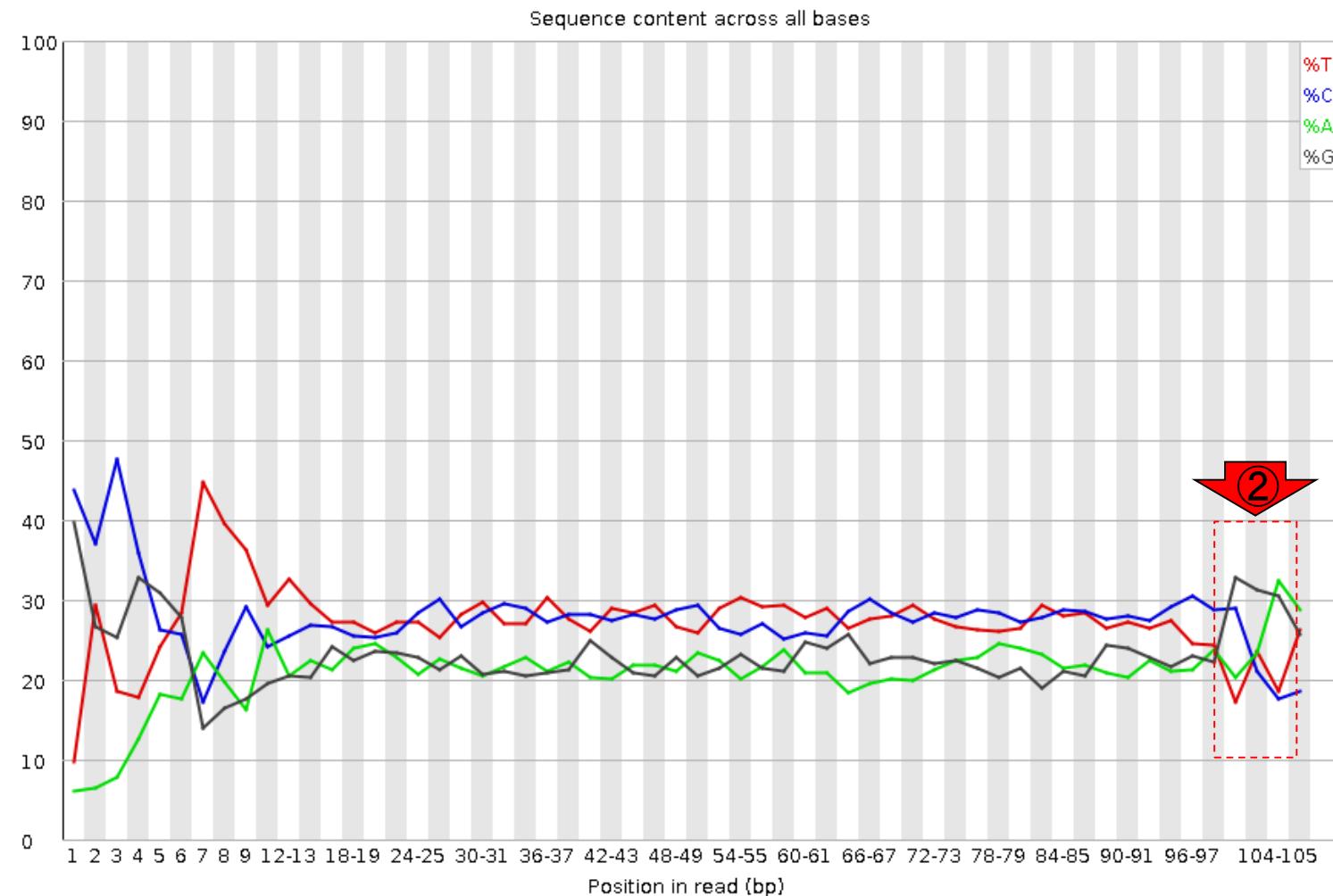
FastQC実行結果

FastQC Report

Summary

- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)
- [Kmer Content](#)

✖ Per base sequence content



QuasR実行結果

- ・インストール | ファイル形式の変換 | [qseq --> Sanger FASTQ](#) (last modified 2013/08/19)
- ・前処理 | クオリティコントロール | について (last modified 2015/06/25)
- ・前処理 | クオリティチェック | [QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/15) ①
- ・前処理 | クオリティチェック | [qrqc](#) (last modified 2014/06/17)
- ・前処理 | クオリティチェック | [PHREDスコアに変換](#) (last modified 2013/06/18)
- ・前処理
- ・前処理
- ・前処理

①QuasRというパッケージは、Quality Control部分だけで比較すると、FastQCの簡易版のような位置づけ。②例題4実行結果の [hoge4.pdf](#) の一部が右下プロットで、FastQCのPer base sequence contentと同じもの。③の赤枠部分に注目!

前処理 | クオリティチェック | QuasR(Gaidatzis_2015)

QuasRパッケージを用いてQCLレポートファイルを出力するやり方を示します。[FastQC](#)のR版のような [qrqc](#)よりも相当ストイックな出力結果です。

「ファイル」② 4. FASTQ形式ファイル([SRR616268sub_1.fastq.gz](#))の場合:

1. サンプルデータ

[SRR037439](#)
al., 2010).

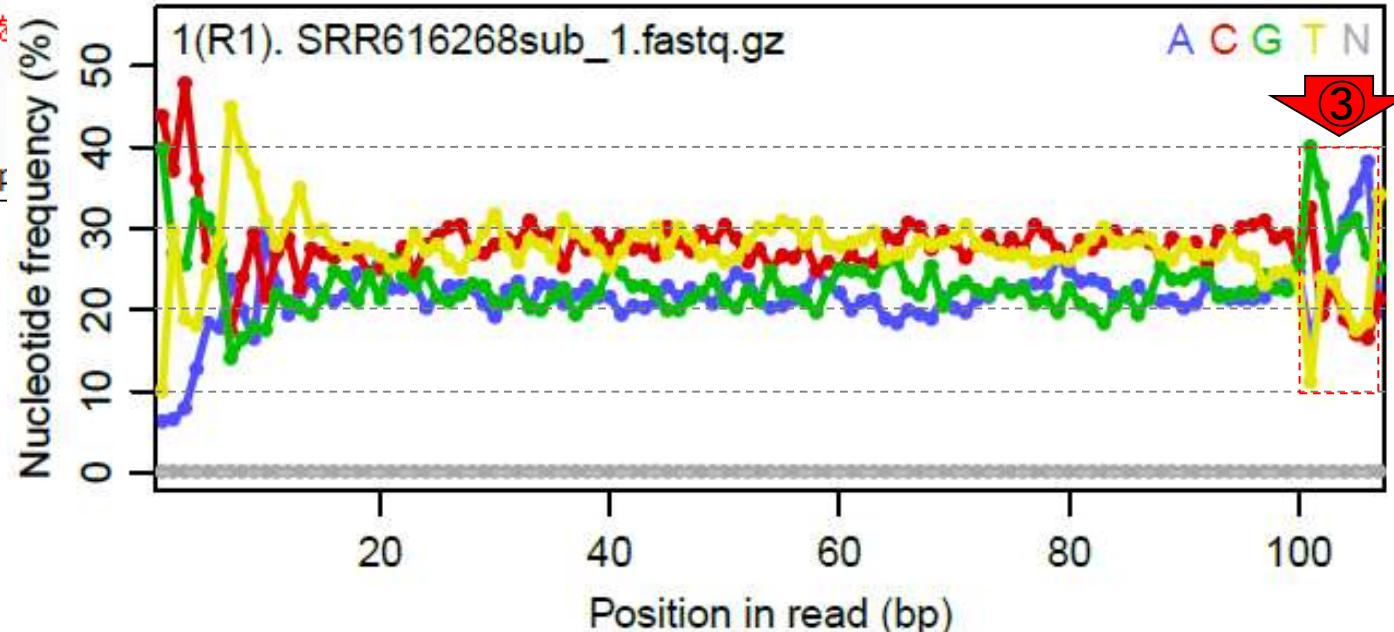
```
in_f <- "SRR616268sub_1.fastq.gz"
out_f <- "hoge4.pdf"
```

```
#必要なパッケージを
library(QuasR)

#本番
qQCReport(in_f, r
```

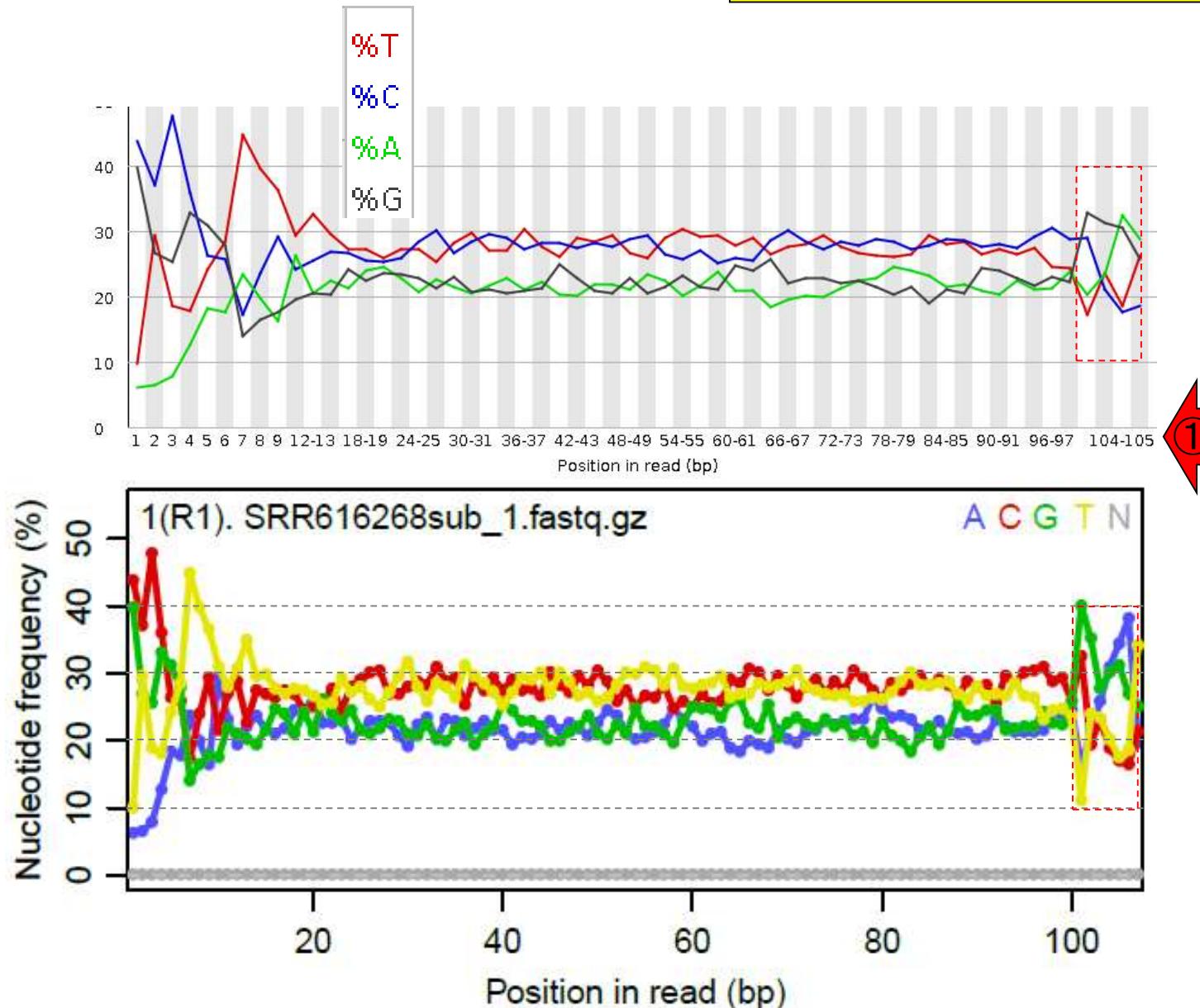
乳酸菌RNA-seqデータSRR616268の最初の100万リード分(約73MB)です。

`in_f <- "SRR616268sub_1.fastq.gz"` #入力ファイル名を指定してin_fに格納
`out_f <- "hoge4.pdf"` #出力ファイル名を指定してout_fに格納



FastQC vs. QuasR

①FastQCのほうの横軸の数値をよく眺めると、こちらは2塩基分づつ平均化したものをプロットしているのではないかと思うに至る



Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

FastQCマニュアル

①fastqc2コマンドのマニュアルを表示。予めマニュアル全体を眺めて、`--nogroup`オプションを用いれば平均化(grouping)せず塩基ごとに表示できることが分かったうえで、②「grep -A 5 nogroup」とパイプでつなげて必要最小限の情報を表示

iu@bielinux[~]



iu@bielinux[iu] fastqc2 -v

[10:46午後]

FastQC v0.11.3

iu@bielinux[iu] fastqc2 -h | grep -A 5 nogroup

[10:46午後]

--nogroup

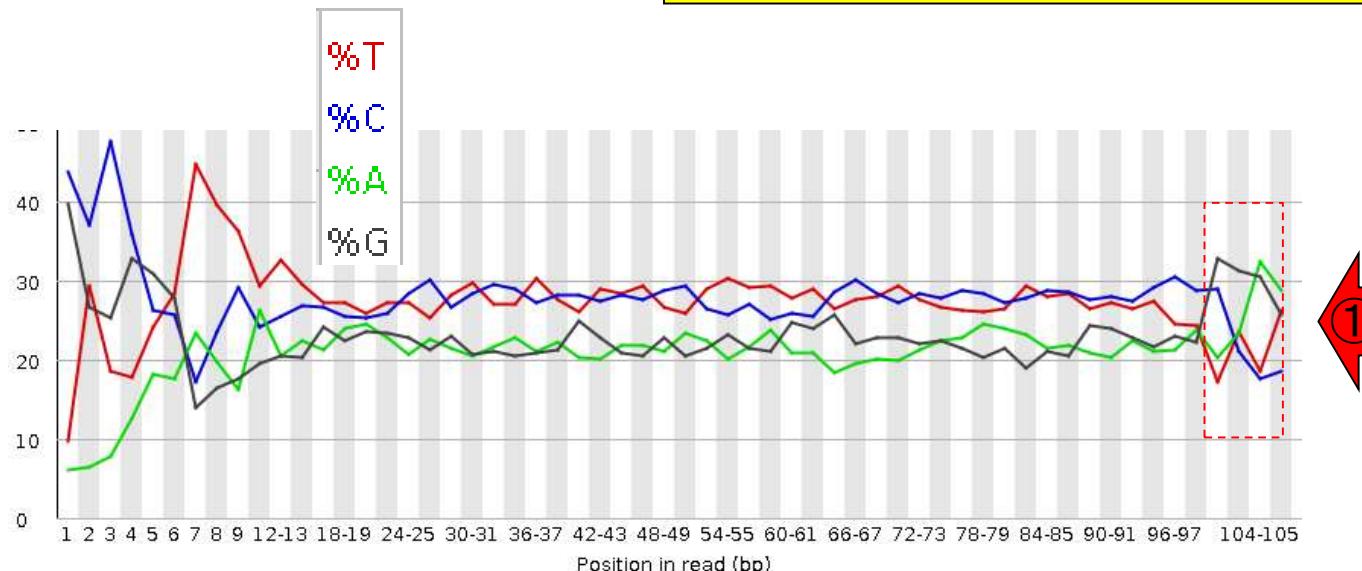
Disable grouping of bases for reads >50bp. All reports will show data for every base in the read. WARNING: Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned!

iu@bielinux[iu] ■

[10:46午後]

FastQC

①このFastQCレポートhtmlは、②FastQCをデフォルトオプションで実行した結果。③得られたhtmlファイル名をresult_without_nogroup.htmlに変更。



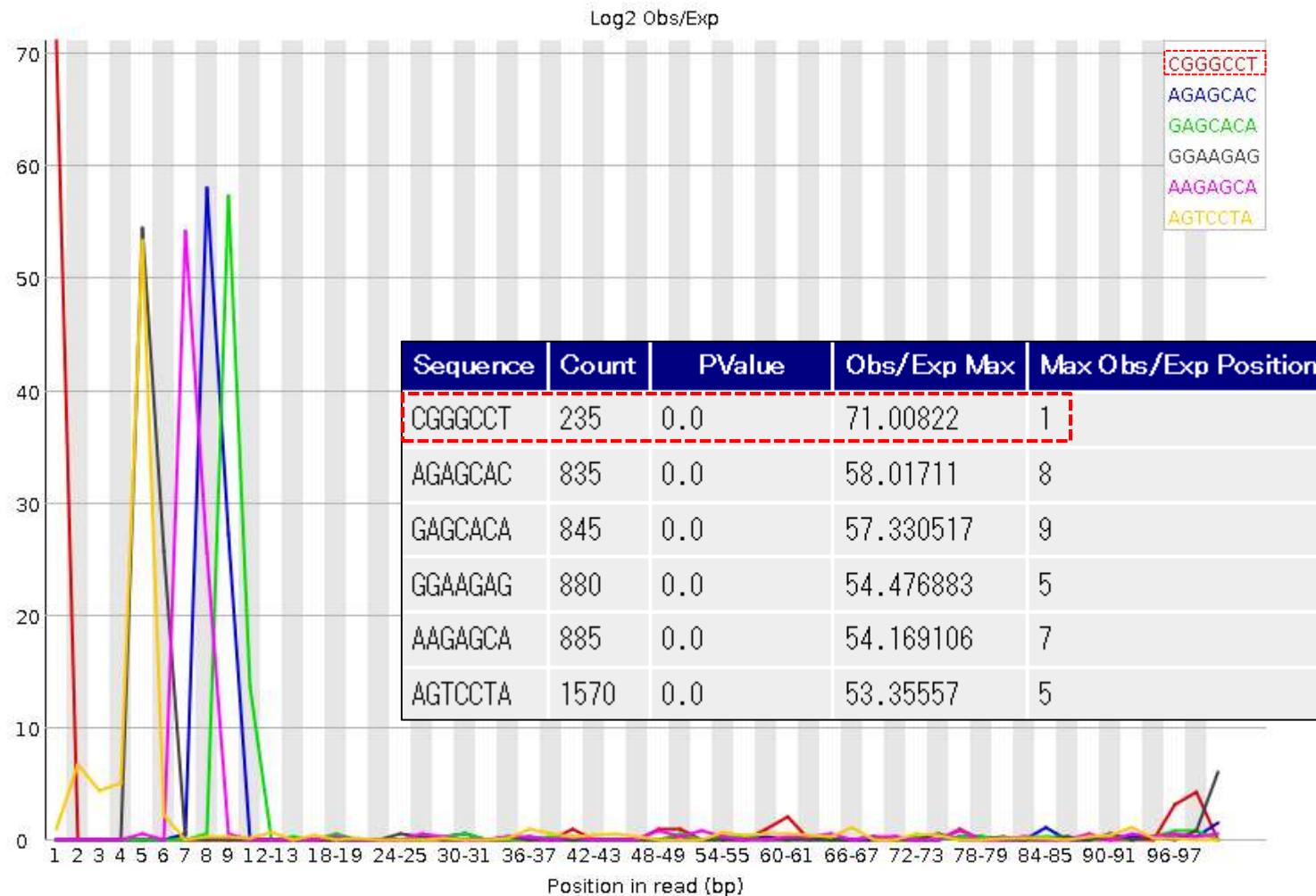
```
iu@bielinux[~/Desktop/mac_share]
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -la SRR616268sub_1.fastq.gz
-rwxrwxrwx 1 root root 74906576 7月 13 17:06 SRR616268sub_1.fastq.gz
iu@bielinux[mac_share] fastqc2 -q SRR616268sub_1.fastq.gz
iu@bielinux[mac_share] ls -la *.html
-rwxrwxrwx 1 root root 365019 7月 26 23:06 SRR616268sub_1_fastqc.html
iu@bielinux[mac_share] mv SRR616268sub_1_fastqc.html result_without_nogroup.html
iu@bielinux[mac_share] date
2015年 7月 26日 日曜日 23:07:50 JST
iu@bielinux[mac_share]
```

①Kmer Contentの項目はこんな感じでした。(2015.06.23の講義資料も参照のこと)

FastQC

Summary

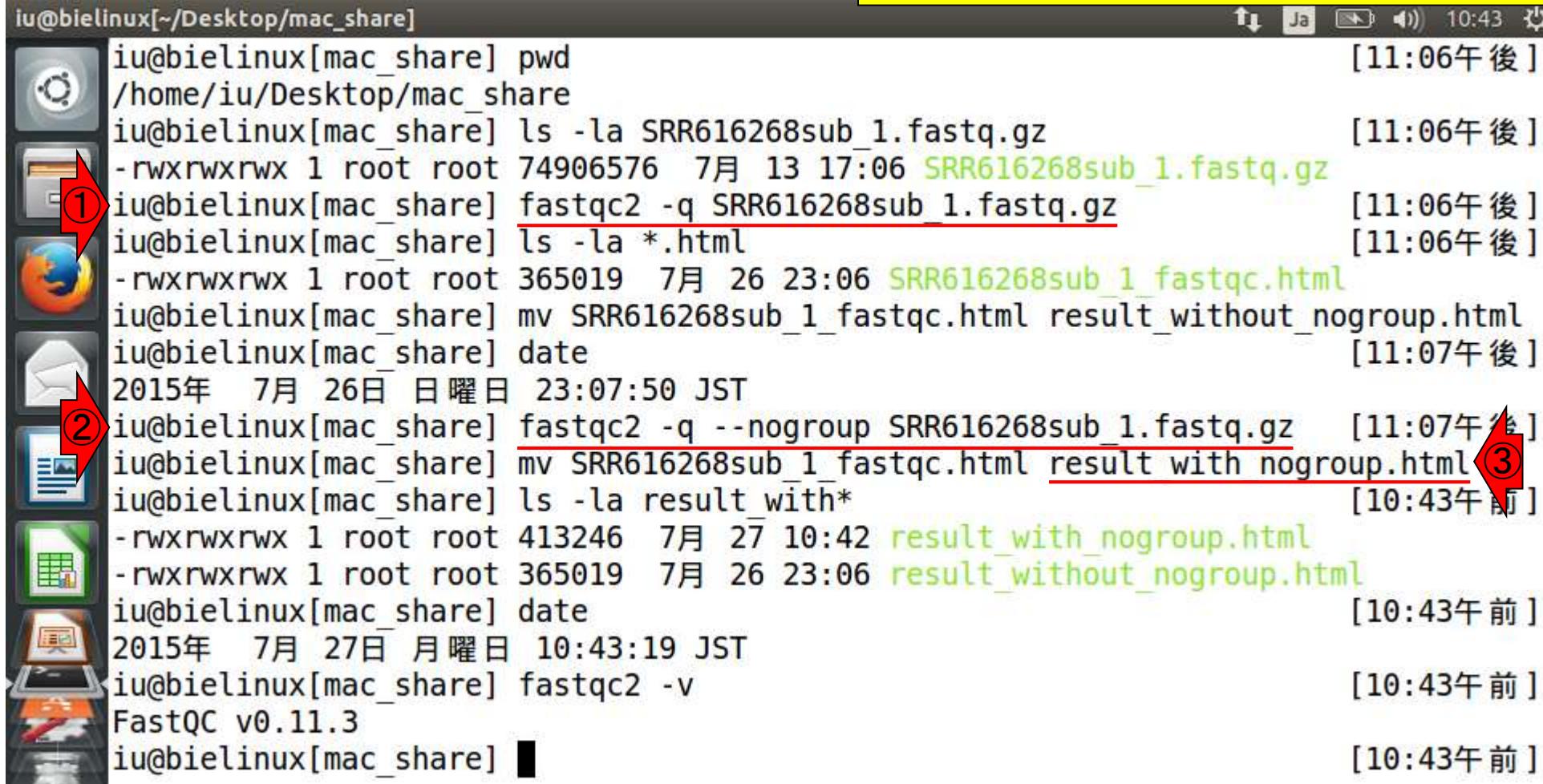
- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)
- [Kmer Content](#) ①



FastQC

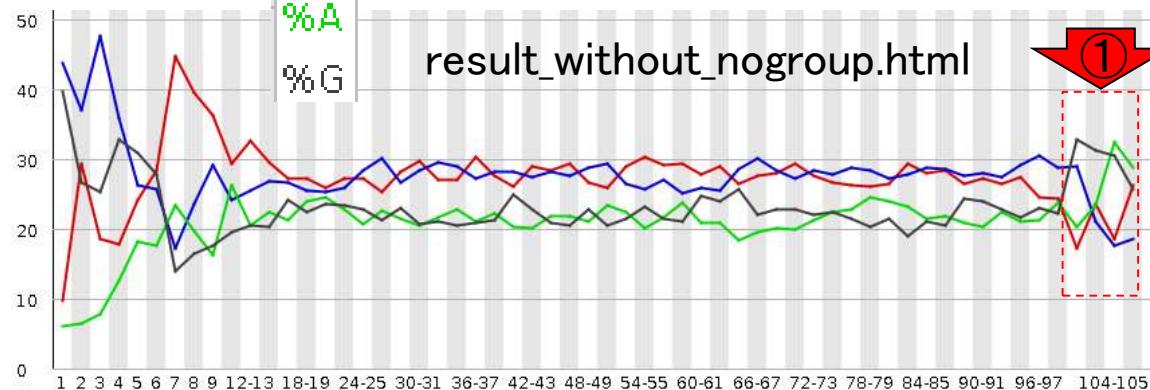
FastQC (ver. 0.11.3)を①デフォルトで実行、②--nogroupオプションつきで実行。③得られたhtmlファイル名をresult_with_nogroup.htmlに変更。

```
iu@bielinux[~/Desktop/mac_share]                               [11:06午後]
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -la SRR616268sub_1.fastq.gz      [11:06午後]
-rwxrwxrwx 1 root root 74906576 7月 13 17:06 SRR616268sub_1.fastq.gz
iu@bielinux[mac_share] fastqc2 -q SRR616268sub_1.fastq.gz    [11:06午後]
iu@bielinux[mac_share] ls -la *.html                         [11:06午後]
-rwxrwxrwx 1 root root 365019 7月 26 23:06 SRR616268sub_1_fastqc.html
iu@bielinux[mac_share] mv SRR616268sub_1_fastqc.html result_without_nogroup.html
iu@bielinux[mac_share] date                                [11:07午後]
2015年 7月 26日 日曜日 23:07:50 JST
iu@bielinux[mac_share] fastqc2 -q --nogroup SRR616268sub_1.fastq.gz  [11:07午後]
iu@bielinux[mac_share] mv SRR616268sub_1_fastqc.html result_with_nogroup.html ③
iu@bielinux[mac_share] ls -la result_with*                 [10:43午前]
-rwxrwxrwx 1 root root 413246 7月 27 10:42 result_with_nogroup.html
-rwxrwxrwx 1 root root 365019 7月 26 23:06 result_without_nogroup.html
iu@bielinux[mac_share] date                                [10:43午前]
2015年 7月 27日 月曜日 10:43:19 JST
iu@bielinux[mac_share] fastqc2 -v                          [10:43午前]
FastQC v0.11.3
iu@bielinux[mac_share] █                                [10:43午前]
```

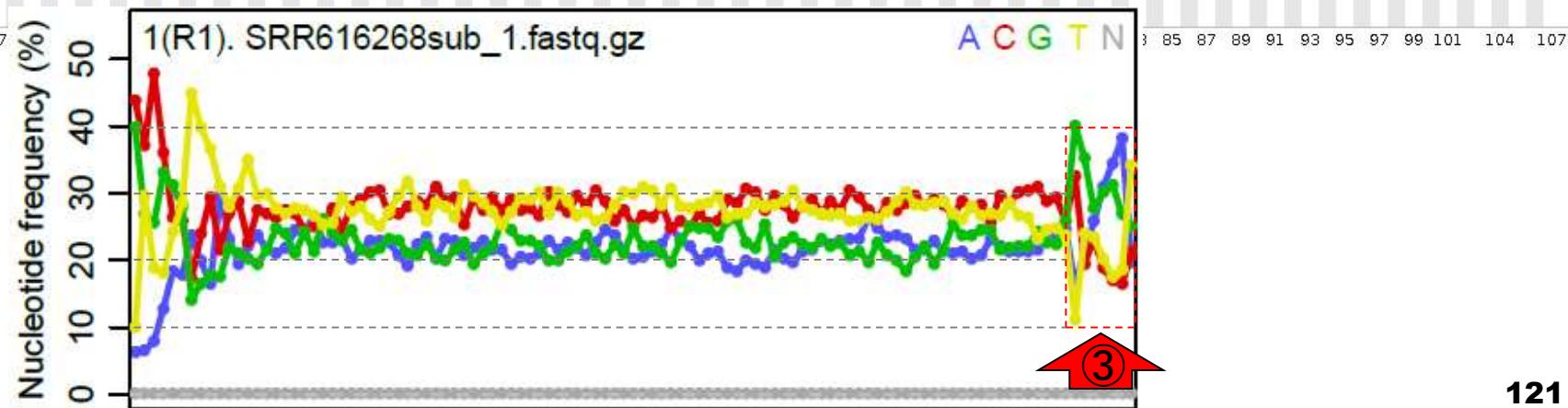
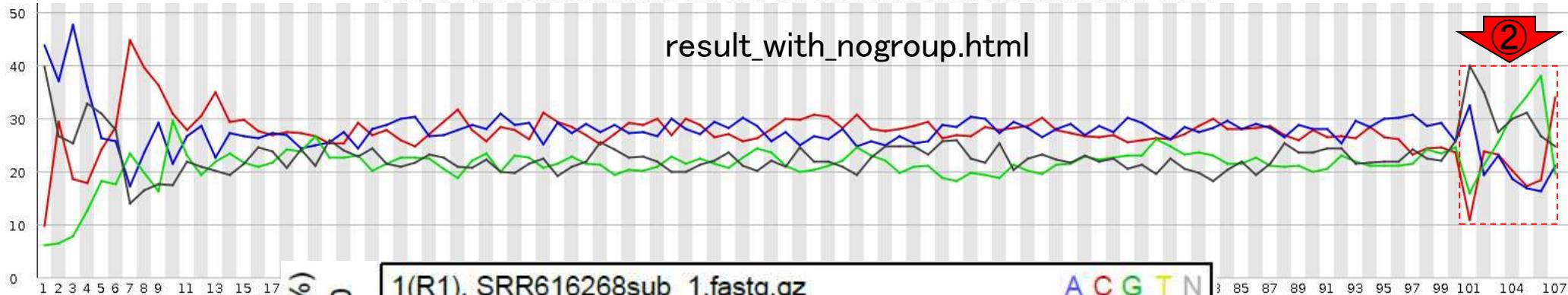


FastQC

①FastQCデフォルトの結果。②FastQC --nogroupオプションつきの結果。Groupingされないので、配列長が長くなるほど、どんどん横長になる。③QuasRデフォルトの結果



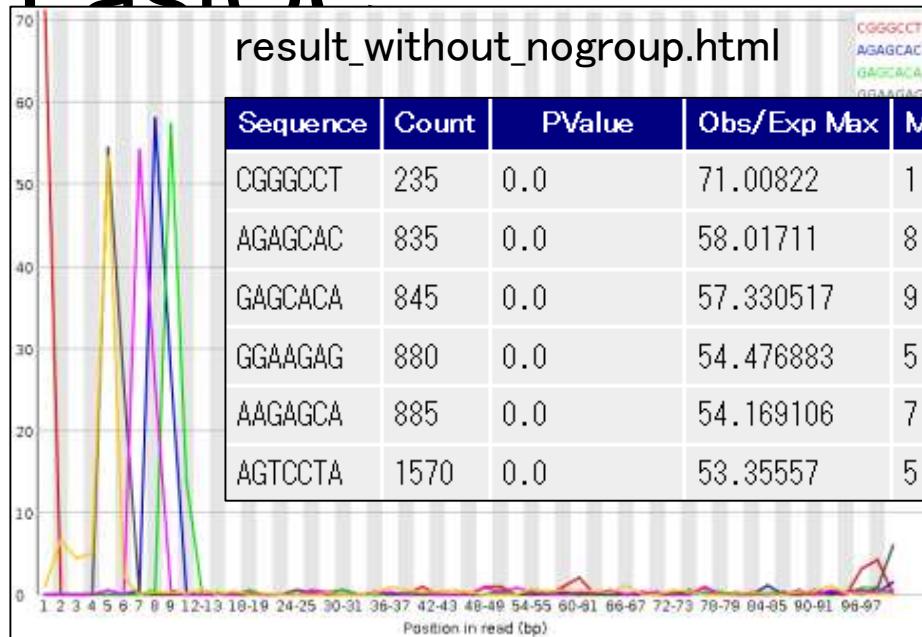
result_without_nogroup.html



Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

FactOC



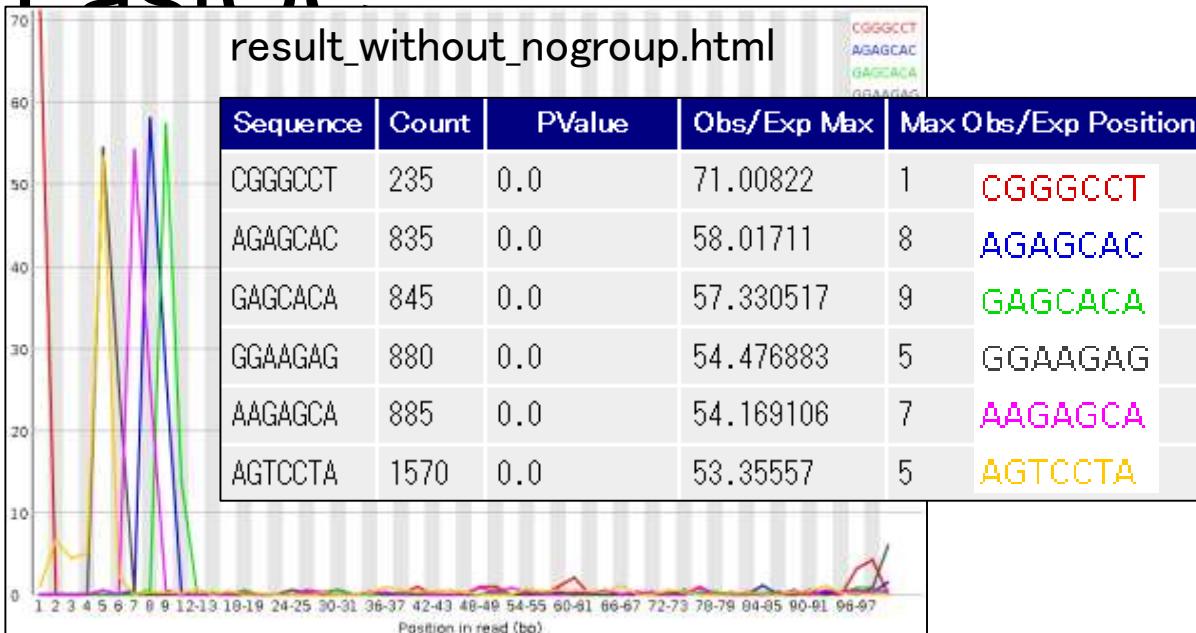
Kmer Contentの項目を比較。違いは--nogroup の有無のみ…何か変。結論としては、--nogroupをつけないときの結果は(門田の感覚では)バグ

result_with_nogroup.html

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CTCGGAG	395	0.0	86.890045	101 CTCGGAG
CTACAAT	215	0.0	84.51276	101 CTACAAT
AGGTGGA	445	0.0	83.98286	100 AGGTGGA
GATGAGT	960	0.0	82.54421	101 GATGAGT
TCGGACC	235	0.0	81.66475	95 TCGGACC
CCCTAGG	100	0.0	80.80512	95 CCCTAGG

FastQC

result_without_nogroup.html



開発者側の論理としては、デフォルト(--nogroupオプションをつけない)だとKmer Content解析結果も塩基をgroupingする、という思想なのかもしれない。ただしこれは実質的に5'側が上位に来る傾向となり、3'側に除去すべきアダプターやプライマー配列が存在していてもそれに気づけない(実際門田がこれにハマった)。

iu@bielinux[~]

[10:46午後]



iu@bielinux[iu] fastqc2 -v

FastQC v0.11.3

① 1 iu@bielinux[iu] fastqc2 -h | grep -A 5 nogroup [10:46午後]
--nogroup Disable grouping of bases for reads >50bp. All reports will
show data for every base in the read. WARNING: Using this
option will cause fastqc to crash and burn if you use it on
really long reads, and your plots may end up a ridiculous size.
You have been warned!

iu@bielinux[iu] ■

[10:46午後]

様々な状況証拠を収集

■ リファレンス配列がある場合(マッピング)

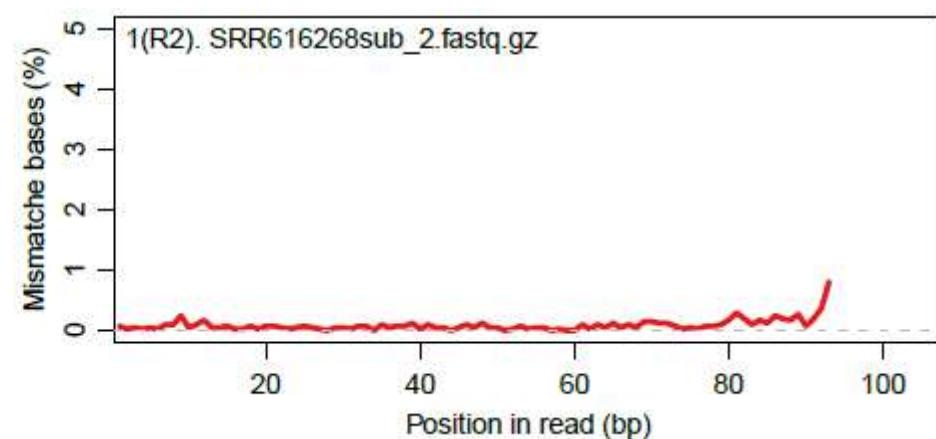
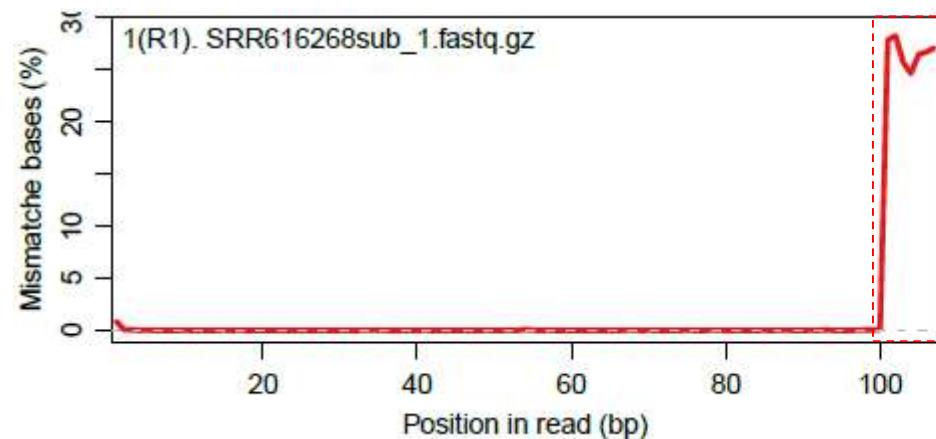
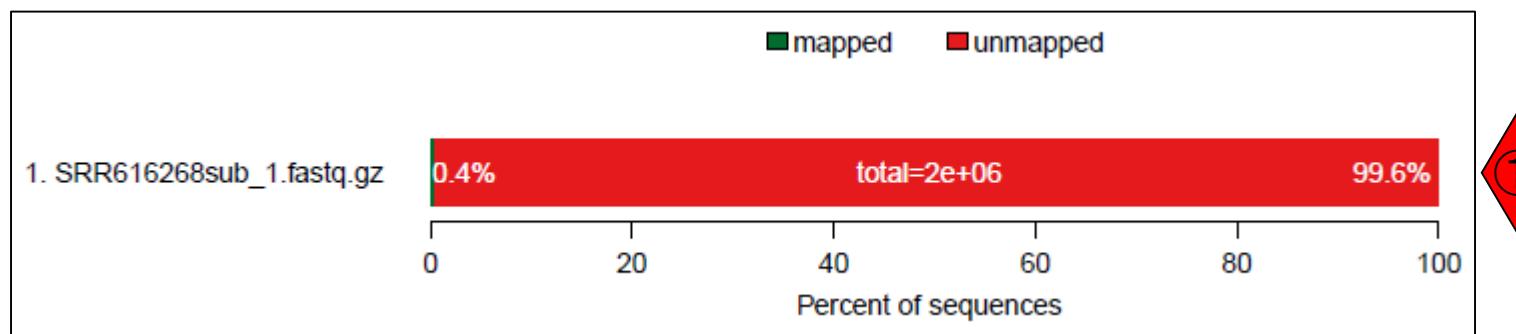
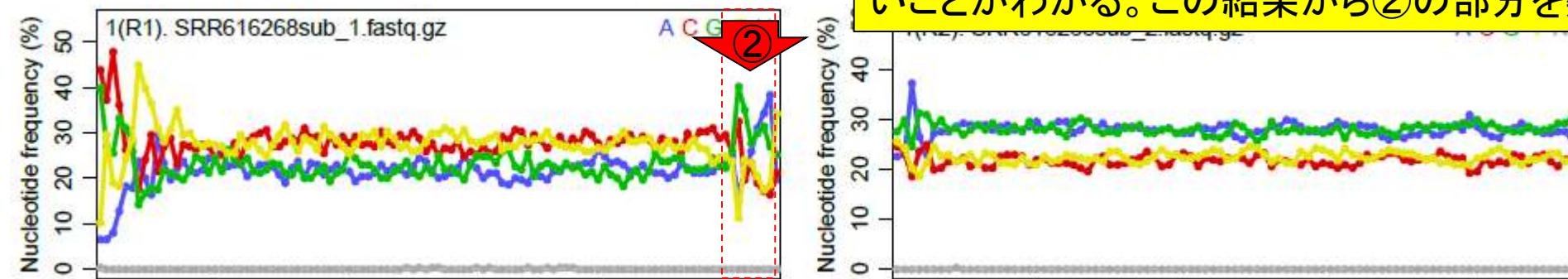
1. 診断:(比較的多めのミスマッチ数で)マッピングを行い、マップされたものの中でミスマッチがあったポジションの分布を眺める。もし3'側にミスマッチの偏りがあれば、それはプライマー/アダプター配列由来塩基という判断を下せる。
2. 対策:該当部分の塩基を一定数トリム
3. 検証:マップ率が向上するかどうかを調べる。

■ リファレンス配列がない場合(アセンブル)

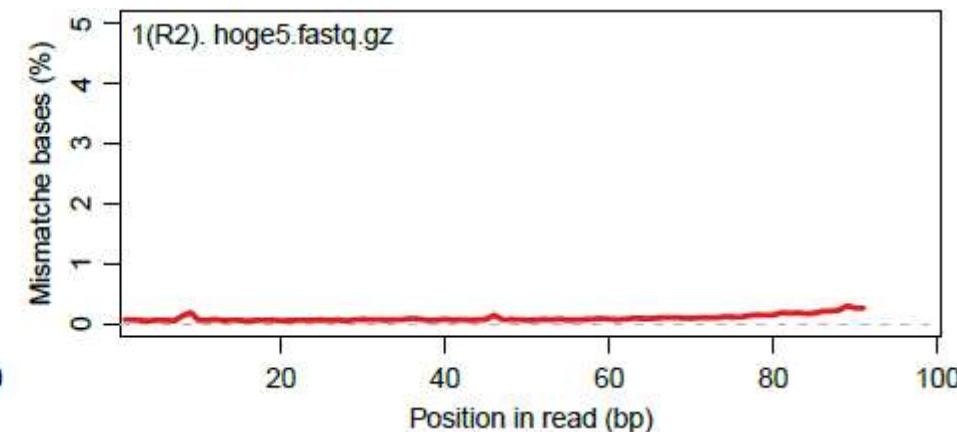
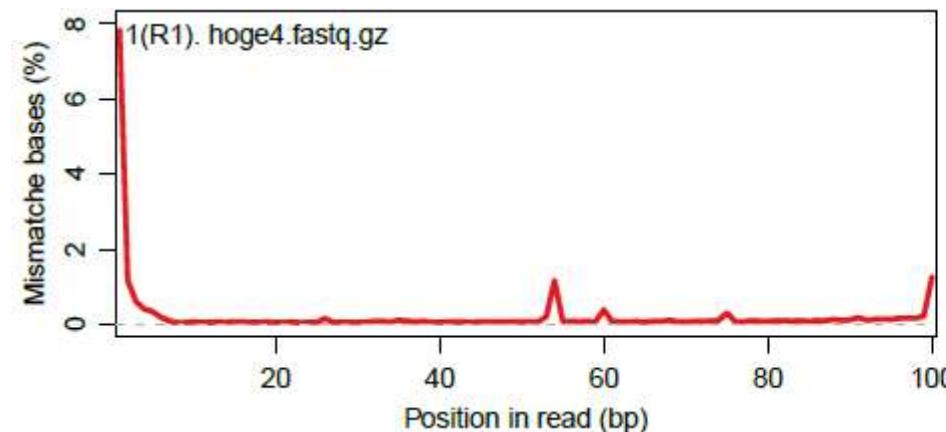
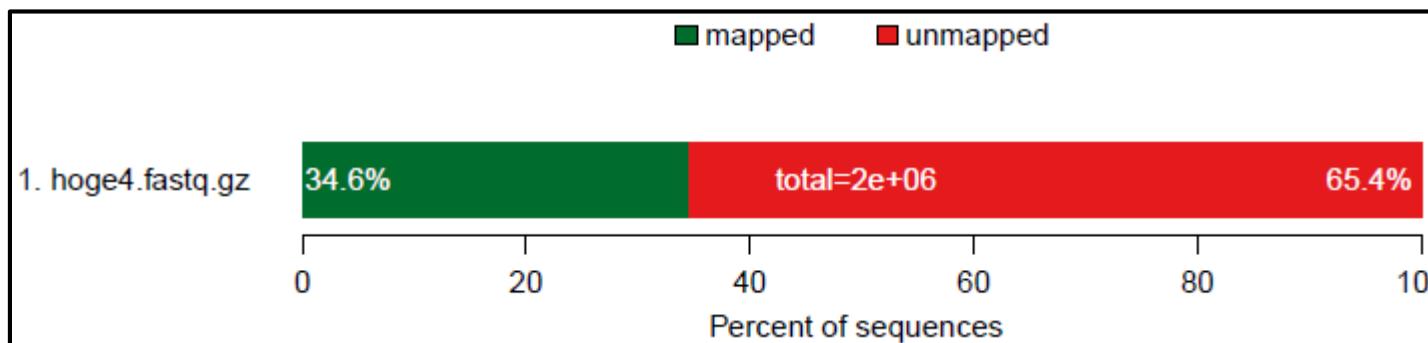
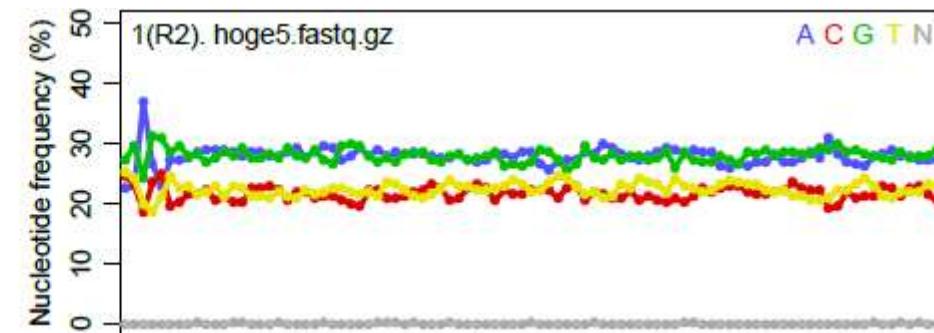
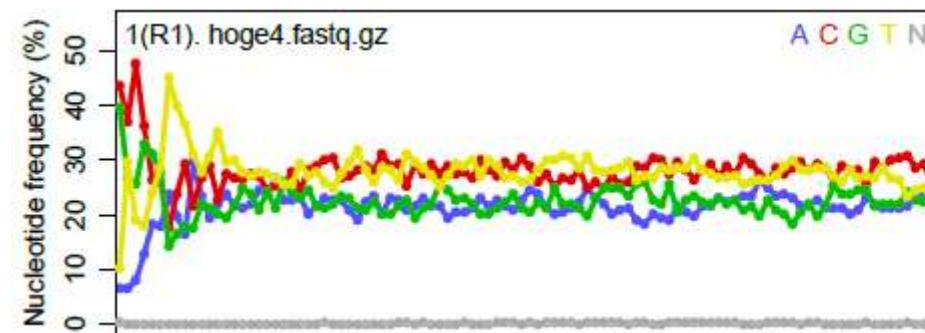
1. 診断: paired-endデータの場合は、single-endデータとしてde novo assembleするとある程度アセンブルされる(一定数のコンティグが得られる)一方、paired-endデータで実行すると途端にアセンブルされなくなる。
2. 対策: 1塩基づつトリムしてはアセンブル、みたいなことを繰り返して、一定範囲のデータを得る。
3. 検証: アセンブルが劇的に改善されるところが出てくる(はず)。

QuasR(トリム前)

QuasRはマッピングもできる。乳酸菌ゲノム配列にマッピングした結果のPDFファイルの一部を表示。
 ①ほとんどマップされていない。100万リード中、わずか0.4% (約4,000リード)しかマップされていないことがわかる。この結果から②の部分を疑う。



QuasR(トリム後)



末端塩基のトリム

- 前処理|トリミング|アダプター配列除去(応用)|[QuasR\(Gaidatzis 2015\)](#) (last modified 2015/06/29) NEW
- 前処理|トリミング|アダプター配列除去(応用)|[ShortRead\(Morgan 2009\)](#) (last modified 2014/03/03)
- 前処理|トリミング|指定した末端塩基数だけ除去 (last modified 2015/06/29) NEW
- 前処理|フィルタリング|PHREDスコアが低い塩基をNに置換 (last modified 2014/03/03)
- 前処理|フィルタリング|PHREDスコアが低い配列(リード)を除去 (last modified 2014/08/08)
- 前処理|フィルタリング|指定した末端塩基数だけ除去 (last modified 2014/08/08)
- 前処理|ファイル操作|3'末端を指定塩基数分だけトリムするやり方を示します。
- 前処理|ファイル操作|「ファイル」→「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。
- 前処理|ファイル操作|1. multi-FASTQ
- 前処理|ファイル操作|2. FASTQ形式ファイル([SRR616268sub_1.fastq.gz](#))の場合:

①

前処理|トリミング|指定した末端塩基数だけ除去 NEW

1. multi-FASTQ

```
## イントロ | 一般
in_f <- "hoge.fq"
out_f <- "hoge4.fq"
param_trim <- 7
```

#必要なパッケージをロード

```
library(Biostrings)
library(ShortRead)
```

#入力ファイルの読み込み

```
fastq <- readFastq(in_f)
sread(fastq)
```

#本番

```
hoge <- width(sread(fastq)) - param_trim
```

4. FASTQ形式ファイル([SRR616268sub_1.fastq.gz](#))の場合:

乳酸菌RNA-seqデータSRR616268の最初の100万リード分(約73MB; 74,906,576 bytes)です。paired-endのforward側です。長さは全て107 bpです。3'側の7 bp分をトリムするので、出力ファイル中の長さは100 bpになります。

```
in_f <- "SRR616268sub_1.fastq.gz"
out_f <- "hoge4.fastq.gz"
param_trim <- 7
```

#必要なパッケージをロード

```
library(Biostrings)
library(ShortRead)
```

#入力ファイルの読み込み

```
fastq <- readFastq(in_f)
sread(fastq)
```

#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_fに格納
#3'末端のトリムしたい塩基数を指定

#パッケージの読み込み
#パッケージの読み込み

#in_fで指定したファイルの読み込み
#配列情報を表示

#本番

```
hoge <- width(sread(fastq)) - param_trim #トリム後のend位置情報を取得
hoge[hoge < 1] <- 1 #トリム後のend位置が1未満の場合には1にしている
hoge1 <- DNAStringSet(sread(fastq), start=1, end=hoge) #sread(fastq)から指定した範囲を抽出
hoge2 <- BStringSet(quality(quality(fastq)), start=1, end=hoge) #quality(quality(fastq))
fastq <- ShortRead(hoge1, hoge2, id(fastq)) #ShortRead()というクラスオブジェクトを作成
```

②

末端塩基のトリム

①トリム前のリード情報。②7塩基をトリムするので、③赤枠部分をトリムした結果が返される(はず)。

4. FASTQ形式ファイル(SRR616268sub_1.fastq.gz)の場合:

乳酸菌RNA-seqデータSRR616268の最初の100万リード分(約73MB; 74,906,576 bytes)です。paired-endのforward側です。長さは全て107 bpです。3'側の7 bp分をトリムするので、出力ファイル中の長さは100 bpになります。

```
in_f <- "SRR616268sub_1.fastq.gz"      #入力ファイル名を指定してin_fに格納
out_f <- "hoge4.fastq.gz"                #出力ファイル名を指定してout_fに格納
param_trim <- 7                          #3'末端のトリムしたい塩基数を指定
```

#必要なパッケージをロード

```
library(Biostrings)
library(ShortRead)
```

#入力ファイルの読み込み

```
fastq <- readFastq(in_f)
sread(fastq)
```

#本番

```
hoge <- width(sread(fastq))
hoge[hoge < 1] <- 1
hoge1 <- DNAStringSet(sread(fastq))
hoge2 <- BStringSet(quality(fastq))
fastq <- ShortRead(hoge1, hoge2)
```

The screenshot shows an R console window titled 'R Console'. The code being run is:

```
> fastq <- readFastq(in_f)
> sread(fastq)
```

The output shows the result of the command:

```
A DNAStringSet instance of length 1000000
width seq
[1] 107 AGCCCGACTTCGTCCCTGCTC...TTCTGAGGGAACCTTCTCTAAAC
[2] 107 GATCTGGGCTGTTCCCCTTCG...CTGAATTCACTACCTCCGAAA
[3] 107 CCGGTATATTTGGCGCAGTG...GTTGTCTGTGCAACGATCTTAC
[4] 107 CTTCGATAACGCCAAGAACTT...GCCATCTGTGGCCATAAAGACC
[5] 107 CCCCGGTATATTTGGCGCAG...TAGTTGTCTGTGCAAGGGTCCAT
...
[999996] 107 CCCCGGTATATTTGGCGCAG...TAGTTGTCTGTGCAAGGGTTAC
[999997] 107 TTGGGTCTACATCTGCTTACT...AACGTAACCTGCCGGGTGAAAT
[999998] 107 CGTCCATCCGGCTCTCGTA...CGTTCTGAACCCAGCCAAGAGT
[999999] 107 CTAGGGAGTATTAGCCTTGGG...AGGGTTCGACGTTCTGGACA
[1000000] 107 GCCTTGTCAATCAAGGTGAGCA...TCTTTCACCTTGACAGCAGCT
```

Annotations with red numbers 1, 2, and 3 point to specific parts of the code and output:

- Annotation 1 points to the command `sread(fastq)` in the R console.
- Annotation 2 points to the value `7` in the R code `param_trim <- 7`.
- Annotation 3 points to the last 7 bases of the sequence for the first read, which are highlighted with a red dashed box.

末端塩基のトリム

①トリム後のリード情報。赤枠部分の塩基がなくなっているのが分かる。また、②配列長も100 bpになっていることがわかる。

#入力ファイルの読み込み

```
fastq <- readFastq(in_f)
sread(fastq)
```

#in_fで指定したファイルの読み込み
#配列情報を表示

#本番

```
hoge <- width(sread(fastq)) - param_trim #トリム後のend位置情報を取得
hoge[hoge < 1] <- 1 #トリム後のend位置が1未満の場合には1にしている
hoge1 <- DNAStringSet(sread(fastq), start=1, end=hoge) #sread(fastq)から指定した範囲を抽出
hoge2 <- BStringSet(quality(quality(fastq)), start=1, end=hoge) #quality(quality(fastq))
fastq <- ShortReadQ(hoge1, hoge2, id(fastq)) #ShortReadQというクラスオブジェクトを作成し
sread(fastq)
quality(fastq)
```

①

①

#ファイルに保存

```
writeFastq(fastq, out_f, compre
```

R Console

```
> sread(fastq) #配列情報を表示
A DNAStringSet instance of length 1000000
  width seq
  [1] 100 AGCCCGACTTCGTCCTGCTC... CCAACCATTCTGAGGGAACCTT
  [2] 100 GATCTGGGCTGTTCCCCTTCG... AGTTTATCTGAATTCACTAACC
  [3] 100 CCGGTATATTTCGGCCGCAGTG... CATCCTAGTTGTCTGTGCAAACG
  [4] 100 CTTCGATAACGCCAAGAACTT... CAAAGGTGCCATCTGTGGCCAT
  [5] 100 CCCCGGTATATTTCGGCCGCAG... AACATCCTAGTTGTCTGTGCAA
  ...
  [999996] 100 CCCCGGTATATTTCGGCCGCAG... AACATCCTAGTTGTCTGTGCAA
  [999997] 100 TTCGGGTCTACATCTGCTTACT... GCAAGCAAACGTAACCTGCCGG
  [999998] 100 CGTCCATCCCGGTCTCTCGTA... CTCACGACGTTCTGAACCCAGC
  [999999] 100 CTAGGGAGTATTAGCCTGGG... TGGACGGAGGGTTCGACGTTTC
  [1000000] 100 GCCTTGTCAATCAAGGTGAGCA... GCATCCATTTTACCTTGAC
> | ②
```

Tips:例外処理

#入力ファイルの読み込み

```
fastq <- readFastq(in_f)  
sread(fastq)
```

#in_fで指定したファイル
#配列情報を表示

#本番

```
hoge <- width(sread(fastq)) - param_trim#① ム後のend位置情報を取得  
hoge[hoge < 1] <- 1 # ム後のend位置が1未満の場合には1にしている  
hoge1 <- DNAStringSet(sread(fastq), start=1, end=hoge)#sread(fastq)から指定した範囲を抽出  
hoge2 <- BStringSet(quality(quality(fastq)), start=1, end=hoge)#quality(quality(fastq)  
fastq <- ShortReadQ(hoge1, hoge2, id(fastq))#ShortReadQというクラスオブジェクトを作成し  
sread(fastq) #配列情報を表示  
quality(fastq) #quality情報を表示
```

#ファイルに保存

```
writeFastq(fastq, out_f, compress=T) #fastqの中身を指定したファイル名で保存
```

実用上は、①このあたりの例外処理が重要。
このプログラムはリードごとにトリム後の配列長情報をhogeに予め格納している。そしてトリム後の配列長がもし1 bp未満になつたら、1 bpは残すような処理をしている。このあたりの想定外への対処が、いわゆる「bug fix」。

①

ム後のend位置情報を取得

ム後のend位置が1未満の場合には1にしている

sread(fastq)から指定した範囲を抽出

quality(quality(fastq))

ShortReadQというクラスオブジェクトを作成し

#配列情報を表示

#quality情報を表示

R Console

```
> length(hoge)
```

```
[1] 1000000
```

```
> hoge[1:10]
```

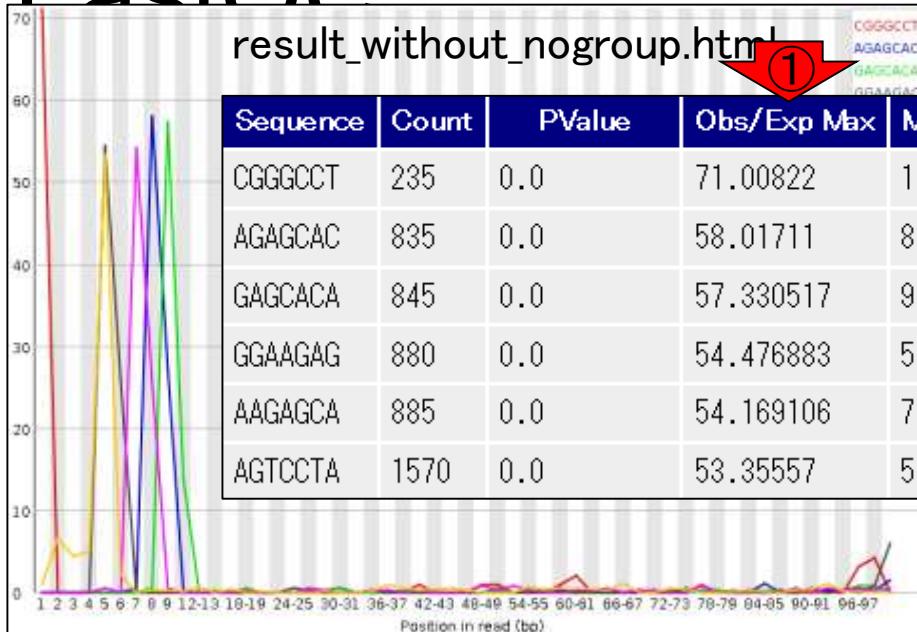
```
[1] 100 100 100 100 100 100 100 100 100 100
```

```
> |
```

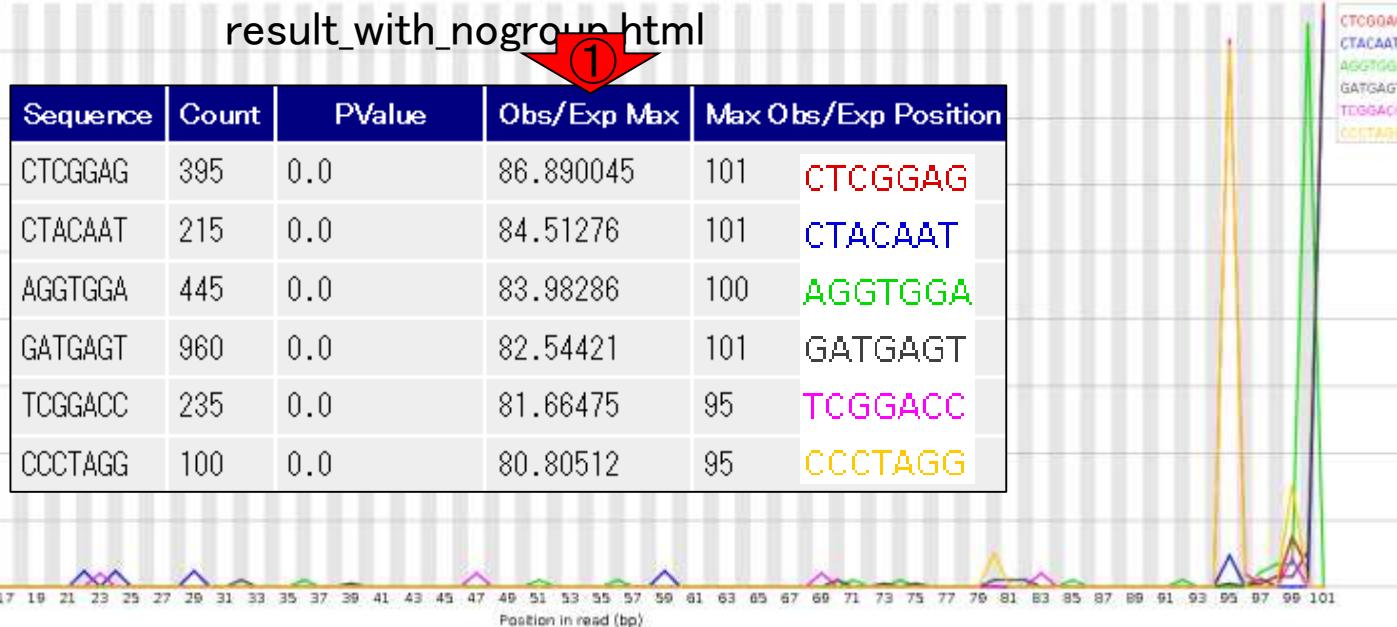
Contents

- パッケージ
 - CRANとBioconductor
 - 推奨パッケージインストール手順のおさらい
 - Bioconductor概観 → ゲノム配列パッケージ(BSgenome)
 - ヒトゲノム情報パッケージの解析
- プロモーター配列取得(sessionInfo、バージョンの違い)
 - Rパッケージ(ゲノムとアノテーションパッケージの併用)
 - FASTA形式ファイルとGFF3形式ファイル
 - データの型
- FASTQファイルの各種解析(LinuxとRを相補的に活用)
 - Linux (FastQC)とR (ShortRead)で同じ: Overrepresented sequences 項目
 - Linux (FastQC)とR (QuasR)で異なる見栄え: Per base sequence content 項目。
FastQCに--nogroupというオプションがあることを知る。
 - FastQCのオプション(デフォルトと--nogroupあり)の違いによるKmer Content 項目の結果の違い → Rで検証

FactOC



Kmer Contentの項目で、①Obs/Exp Max やプロファイルの図を(若干数値が代わりますが)Rで再現することができます。



参考資料

(Rで)塩基配列解析

~NGS、RNA-seq、ゲノム、トランскриプトーム、正規化、発現変動、統計、
(last modified 2015/07/26 since 2011)

What's new

- このウェブページは、アグリバイオバイオインフォマティクス教育研究プログラムの大学院講義科目「農学生命情報科学特論I」(東京大学)で使用される教材をまとめたものです。

- はじめに (last modified 2015/03/31)
- 参考資料(講義、講習会、本など) ① modified 2015/07/07 NEW
- 過去のお知らせ (last modified 2015/07/08) NEW
- インストール

参考資料(講義、講習会、本など) NEW

基本的に私門田の個人ページに記載してあるものです。かなり古い講演資料などの情報をもとに勉強されている方

講習会、講義、講演資料

- 門田幸二「講義資料」,アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:農学生命情報科学特論I第4回, 東京大学(東京), 2015.07.07

内容: 教科書の3.3節周辺。FastQC (ver. 0.11.3)の--nogroupオプションの有無とKmer Contentの項目の挙動の違い。様々な角度で動作確認および検証することの重要性。adapters/primers除去後の乳酸菌paired-end RNA-seqデータのマッピング。アノテーション情報がないときとアノテーションファイル(GFF3)を利用したときのカウント情報取得実例。RPKMの基本的な考え方。配列長とカウント数の関係。原著論文(Blekhman et al., 2010)の公共カウントデータを利用した各種解析。エクセルファイルをRで読み込むこと、サブセット抽出や整形のテクニック。サンプル間クラスタリング結果での実験デザインの説明や発現変動解析結果の予想。TCCパッケージを用いた発現変動解析と結果の解釈。同一群同士と異なる群間の発現変動解析結果を示して、分布やモデルの意味を理解。倍率変化の危険性やFDRの結果との比較を通じて、なぜ先人が倍率変化のみで信頼性の高い結果を得てきたのかなど。2コマ(2×90 min)分。

- 門田幸二「講義資料(2015.07.01版)」,アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:農学生命情報科学特論I第3回, 東京大学(東京), 2015.06.30

内容: 教科書の2.3節周辺。NAの取り扱い。paired-endの取り扱いとして、アダプター/プライマー(adapters/primers)配列の除去や、Nの数およびリード長でのフィルタリングの基礎(QuasR)と応用(ShortRead)。アセンブル(ゲノム用とトランскриプトーム用)。Rockhopperを用いた乳酸菌RNA-seqデータのアセンブル(single-end (SE)と paired-end(PE)両方)。マッピングの基礎。仮想データ作成とQuasRを用いたマッピング。出力ファイル形式(SAM/BAMとBED)。bowtieを用いたマッピング時のオプション変更と結果の関係。カウント情報取得。マップされた領域情報取得などを单一・複数のリードファイルでデモ。QuasRを用いた乳酸菌RNA-seq paired-endデータのマッピング。QCLレポートの概観とMismatch basesのポジション分布を眺めて、アセンブルとマッピングの精度向上へのヒントを得る。2コマ(2×90 min)分。

- 門田幸二「講義資料」,アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:農学生命情報科学特論I第2回, 東京大学(東京), 2015.06.23

アグリバイオ大学院講義の枠組みで、FastQC、マッピング、アセンブル、Kmer ContentのRでの検証などについて講義しています。2015.07.07, 2015.06.30, 2015.06.23などでキーワード検索。