第1部の掟

①今年度のリンク先の、②第1部の該当項目をクリックして、そこを利用してください。例えば、③統計解析の初日(2016.07.20)はこちらをクリック

(Rで)塩基配列解析

～NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス～
(last modified 2016/06/13, since 2011)

What's nev

- このウェ
  従って フ
  は基本的
  系的にま
- 私(門田)
  演依頼・
  ん研究系

- 基本的な利用法 (last modified 2015/04/03)
- サンプルデータ (last modified 2016/05/13)
- 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2016 (last
- 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
- 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 20
- 書籍 | トランスクリプトーム解析 | について (last modified 2014/05/12)
- 書籍 | トラ
- 書籍 | トラ
- 書籍 | トラ
- 書籍 | トラ
- 書籍 | トラ
- 書籍 | トラ
- 書籍 | トラ

①

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2016

平成28年度NGSハンズオン講習会の実習で用いるリンク先、データファイル、コピペ用コード集などはここで示します。

- はじめに(講習会参加者必読) (last modified 2016/06/13)
- 事前準備 | Bio-Linux 8とRのインストール状況確認(2016.07.19) (last modified 2016/06/07)
- 第1部 | 統計解析 | について (last modified 2016/06/02)
- 第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20) (last modified 2016/05/07)
- 第1部 | 統計解析 | トランスクリプトーム解析1(2016.07.21) (last modified 2016/05/23)
- 第1部 | 統計解析 | トランスクリプトーム解析2(2016.07.22) (last modified 2016/06/02)
- 第2部 | NGS解析(初～中級) | について (last modified 2016/05/26)
- 第2部 | NGS解析(初～中級) | NGS解析基礎(2016.07.25) (last modified 2016/04/26)
- 第2部 | NGS解析(初～中級) | ゲノムReseq、変異解析(2016.07.26) (last modified 2016/04/26)
- 第2部 | NGS解析(初～中級) | RNA-seq(2016.07.27) (last modified 2016/04/26)
- 第2部 | NGS解析(初～中級) | ChIP-seq(2016.07.28) (last modified 2016/04/26)
- 第3部 | NGS解析(中～上級) | について (last modified 2016/06/13)
- 第3部 | NGS解析(中～上級) | Linux環境でのデータ解析: JavaやRの利用法(2016.08.01) (last modified 2016/06/13)
- 第3部 | NGS解析(中～上級) | Linux環境でのデータ解析: マッピング、トリミング、アセンブリ(2016.08.02) (last modified 2016/06/13)
- 第3部 | NGS解析(中～上級) | クラウド環境との連携、ロングリードデータの解析(2016.08.03) (last modified 2016/04/26)
- 第3部 | NGS解析(中～上級) | トランスクリプトームアセンブリ、発現量推定(2016.08.04) (last modified 2016/04/26)

③

②

第1部の掟

①この日の講義資料や解析データ(hoge.zip)はここにあります。適宜アップデートしているので、最終更新日に気をつけてください

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res., 2013
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T): Tanizawa et al., BMC Genomics, 2015
    - DRR024500* (PacBioデータ; 後に問題が判明し削除済み)
    - DRR024501 (paired-end MiSeqデータ)
    - DRR054113 (PacBioデータ; 再登録後のもの)
    - DRR054114 (PacBioデータ; 再登録後のもの)
    - DRR054115 (PacBioデータ; 再登録後のもの)
    - DRR054116 (PacBioデータ; 再登録後のもの)
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res., 2008
  - Platanus: Kajitani et al., Genome Res., 2014
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25MB)
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

第1部の掟

①この部分からページ下部に移動することが、②のスライドをめくっていくことに相当します

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20) ①

当日は講義資料の②スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res., 2013
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T): Tanizawa et al., BMC Genomics, 2015
    - DRR024500* (PacBioデータ; 後に問題が判明し削除済み)
    - DRR024501 (paired-end MiSeqデータ)
    - DRR054113 (PacBioデータ; 再登録後のもの)
    - DRR054114 (PacBioデータ; 再登録後のもの)
    - DRR054115 (PacBioデータ; 再登録後のもの)
    - DRR054116 (PacBioデータ; 再登録後のもの)
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res., 2008
  - Platanus: Kajitani et al., Genome Res., 2014
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25MB)
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

第1部の掟

例えば②は、①のスライド6の補足情報です。そして③は、①のスライド24の補足情報となります。ウェブ資料中でスライド番号を明記していないのは、スライドの挿入・追加・削除による番号のずれの修正で疲弊するのを避けるためで…した

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res., 2013
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T): Tanizawa et al., BMC Genomics, 2015
    - DRR024500* (PacBioデータ; 後に問題が判明し削除済み)
    - DRR024501 (paired-end MiSeqデータ)
    - DRR054113 (PacBioデータ; 再登録後のもの)
    - DRR054114 (PacBioデータ; 再登録後のもの)
    - DRR054115 (PacBioデータ; 再登録後のもの)
    - DRR054116 (PacBioデータ; 再登録後のもの)
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res., 2008
  - Platanus: Kajitani et al., Genome Res., 2014
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセンブリ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25MB)
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

第1部の掟

ここが最も重要！①のスライド35-36に相当するのが、②の部分

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の①スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res.
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T
    - DRR024500* (PacBioデータ; 後に
    - DRR024501 (paired-end MiSeqデー
    - DRR054113 (PacBioデータ; 再登録
    - DRR054114 (PacBioデータ; 再登録
    - DRR054115 (PacBioデータ; 再登録
    - DRR054116 (PacBioデータ; 再登録
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res.,
  - Platanus: Kajitani et al., Genome Res., 201
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25ME
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイ
- k-mer解析(k=1)

②

- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings」の例題7は、以下と同じ。入力ファイルはout_gapClosed.fa (約2.4MB)

```
in_f <- "out_gapClosed.fa"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt"                      #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

#本番
hoge <- alphabetFrequency(fasta)          #A,C,G,T,..の数を各配列ごとにカウントした結果を
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をobj
#out <- colSums(hoge[, obj])              #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#tmp
```

- 一気に結果を得る

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                       #パッケージの読み込み
```

第1部の掟

①赤下線部分にも書いてあります。何が言いたいかというと…

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res.
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T
    - DRR024500* (PacBioデータ; 後に
    - DRR024501 (paired-end MiSeqデー
    - DRR054113 (PacBioデータ; 再登録
    - DRR054114 (PacBioデータ; 再登録
    - DRR054115 (PacBioデータ; 再登録
    - DRR054116 (PacBioデータ; 再登録
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res.,
  - Platanus: Kajitani et al., Genome Res., 201
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25ME
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイ
- k-mer解析(k=1)

- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings」の例題7は、以下と同じ。入力ファイルはout_gapClosed.fa (約2.4MB)

①

```
in_f <- "out_gapClosed.fa"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt"                      #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

#本番
hoge <- alphabetFrequency(fasta)          #A,C,G,T,..の数を各配列ごとにカウントした結果を
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をobj
#out <- colSums(hoge[, obj])              #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#tmp
```

- 一気に結果を得る

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                       #パッケージの読み込み
```

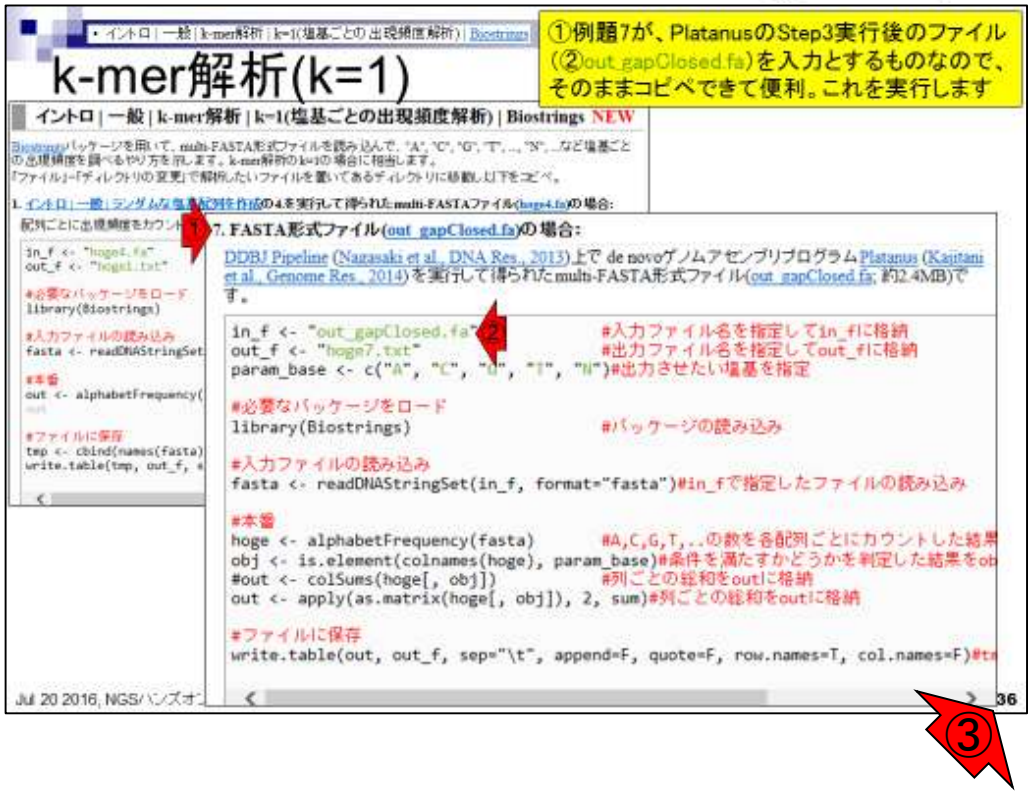
第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の1スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

①のスライド35-36がそれぞれ②と③ですが、ここで書いてある通りにやらなくてよい（やるな）ということです！



第1部の掟

①のスライド35-36は、やったふりをする。実際には②を眺め、③の（特に例題番号）記述内容を確認し…

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の[illegible]ド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res.,
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T
    - DRR024500* (PacBioデータ; 後に
    - DRR024501 (paired-end MiSeqデー
    - DRR054113 (PacBioデータ; 再登録
    - DRR054114 (PacBioデータ; 再登録
    - DRR054115 (PacBioデータ; 再登録
    - DRR054116 (PacBioデータ; 再登録
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res.,
  - Platanus: Kajitani et al., Genome Res., 201
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25ME
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイ
- k-mer解析(k=1)

- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings」の例題7は、以下と同じ。入力ファイルはout_gapClosed.fa (約2.4MB)

```
in_f <- "out_gapClosed.fa"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt"                      #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

#本番
hoge <- alphabetFrequency(fasta)          #A,C,G,T,..の数を各配列ごとにカウントした結果を
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をobj
#out <- colSums(hoge[, obj])              #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#tmp
```

- 一気に結果を得る

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                       #パッケージの読み込み
```

第1部の掟

①のスライド35-36は、やったふりをする。実際には②を眺め、③の（特に例題番号）記述内容を確認し…④の赤枠内をコピペして効率的に進めていきましょう

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res.,
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T
    - DRR024500* (PacBioデータ; 後に
    - DRR024501 (paired-end MiSeqデー
    - DRR054113 (PacBioデータ; 再登録
    - DRR054114 (PacBioデータ; 再登録
    - DRR054115 (PacBioデータ; 再登
    - DRR054116 (PacBioデータ; 再登
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res.,
  - Platanus: Kajitani et al., Genome Res., 201
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25ME
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイ
- k-mer解析(k=1)

- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings」の例題7は、以下と同じ。入力ファイルはout_gapClosed.fa (約2.4MB)

```
in_f <- "out_gapClosed.fa"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt"                          #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

#本番
hoge <- alphabetFrequency(fasta)              #A,C,G,T,..の数を各配列ごとにカウントした結果を
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をobj
#out <- colSums(hoge[, obj])                  #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#tmp
```

- 一気に結果を得る

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                           #パッケージの読み込み
```

第1部の掟

尚、①の例題番号と、②出力ファイルのhogeX.txtの番号は（基本的に）同じです。これも確認手段として有効利用しましょう

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- NGSデータ解析戦略
  - DDBJ Pipeline: Nagasaki et al., DNA Res.
  - Illumina BaseSpace
  - Galaxy: Goecks et al., Genome Biol., 2010
- NGSデータ
  - Lactobacillus hokkaidonensis LOOC260(T
    - DRR024500* (PacBioデータ; 後に
    - DRR024501 (paired-end MiSeqデー
    - DRR054113 (PacBioデータ; 再登録
    - DRR054114 (PacBioデータ; 再登録
    - DRR054115 (PacBioデータ; 再登録
    - DRR054116 (PacBioデータ; 再登録
- de novoアセンブリ
  - Velvet: Zerbino and Birney, Genome Res.,
  - Platanus: Kajitani et al., Genome Res., 201
- 日本乳酸菌学会誌のNGS連載第6回ゲノムアセ
  - 原稿PDF
  - ウェブ資料PDF
    - Windows用(2016.03.29版; 約25ME
    - Macintosh用
- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイ
- k-mer解析(k=1)

- DDBJ PipelineでPlatanus
  - Platanus ver. 1.2.2を実行したzip圧縮ファイル(platanusResult.zip; 約2.2MB)
- k-mer解析(k=1)

塩基ごとの出現頻度解析は、k-mer解析のk=1の場合に相当します。「イントロ | 一般 | k-mer解析 | k=1(塩基ごとの出現頻度解析) | Biostrings」の例題7は、以下と同じ。入力ファイルはout_gapClosed.fa (約2.4MB)

```
in_f <- "out_gapClosed.fa"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge7.txt"                          #出力ファイル名を指定してout_fに格納
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

#本番
hoge <- alphabetFrequency(fasta)              #A,C,G,T,..の数を各配列ごとにカウントした結果を
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をobj
#out <- colSums(hoge[, obj])                  #列ごとの総和をoutに格納
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納

#ファイルに保存
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#tmp
```

- 一気に結果を得る

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                           #パッケージの読み込み
```

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の[illegible]ド25あたりからやる予定です。スライド24あたりまで

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

ここはそれほど重要ではありませんが念のため。①の②スライド45では、rcode1.txtというテキストファイルを開いてもらっています。このとき、エディタによっては改行コードが崩れていたり、OSの違いによってコピペ実行時に不具合が生じるかもしれませんので…

一気に結果を得る

実際の利用時は、hogeフォルダ直下にある①rcode1.txtのように、無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペ

rcode1.txt

```
paran_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                    #パッケージの読み込み

##################
###  Step 1
##################
in_f <- "out_contig.fa"                #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt"            #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta)       #A,C,G,T,..の数を各配列ごとにカウントした
obj <- is.element(colnames(hoge), paran_base)#条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="¥t", append=F, quote=F, row.names=T, col.names=F)#t

##################
###  Step 2
##################
in_f <- "out_scaffold.fa"              #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt"            #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta)       #A,C,G,T,..の数を各配列ごとにカウントした
obj <- is.element(colnames(hoge), paran_base)#条件を満たすかどうかを判定した結果
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="¥t", append=F, quote=F, row.names=T, col.names=F)#t

##################
###  Step 3
##################
```

45

念のため、②のところでもコピペ用コードを提供していますので、必要に応じてご利用ください

第1部の掟

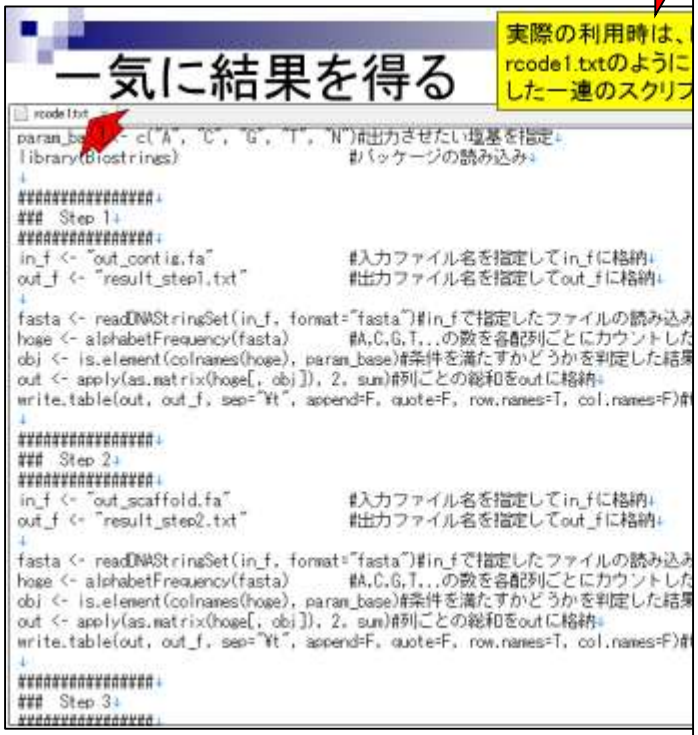
第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

①

②



- 一気に結果を得る

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                     #パッケージの読み込み

################
###  Step 1
################
in_f <- "out_contig.fa"                 #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt"             #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta)        #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をo
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#o

################
###  Step 2
################
in_f <- "out_scaffold.fa"               #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt"             #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
```

- 配列数

rcode2.txtは、配列数をカウントする必要最小限のコード。ファイルの中身は以下と同じ。

```
library(Biostrings)                     #パッケージの読み込み
```

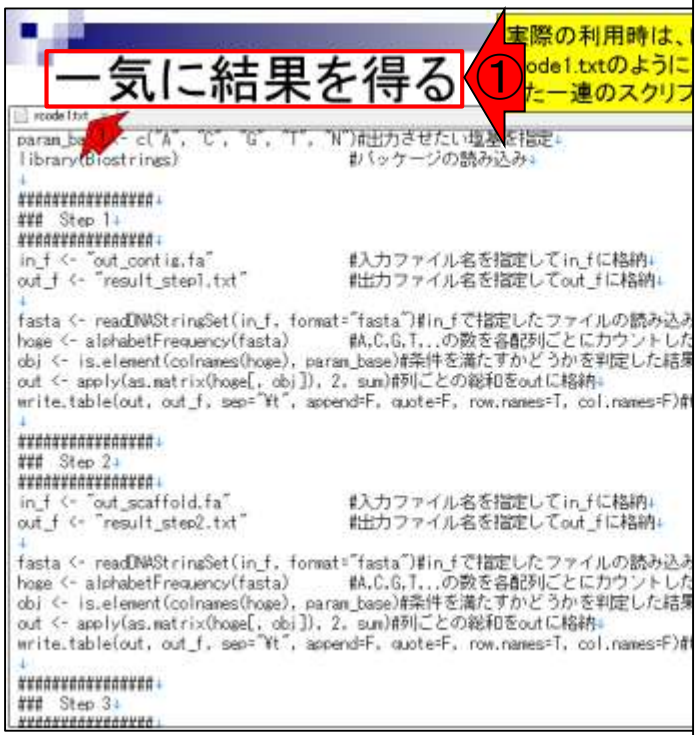
第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

赤枠で示すように、①スライドのタイトルで対応づけを**ある程度**行えます。場所がよくわからなくなったら、「(Rで)塩基配列解析」のウェブサイト内で文字列検索してみてもいいでしょう



- 一気に結果を得る ①

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                     #パッケージの読み込み

################
###  Step 1
################
in_f <- "out_contig.fa"                 #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt"             #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta)        #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をo
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#o

################
###  Step 2
################
in_f <- "out_scaffold.fa"               #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt"             #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
```

- 配列数

rcode2.txtは、配列数をカウントする必要最小限のコード。ファイルの中身は以下と同じ。

```
library(Biostrings)                     #パッケージの読み込み
```

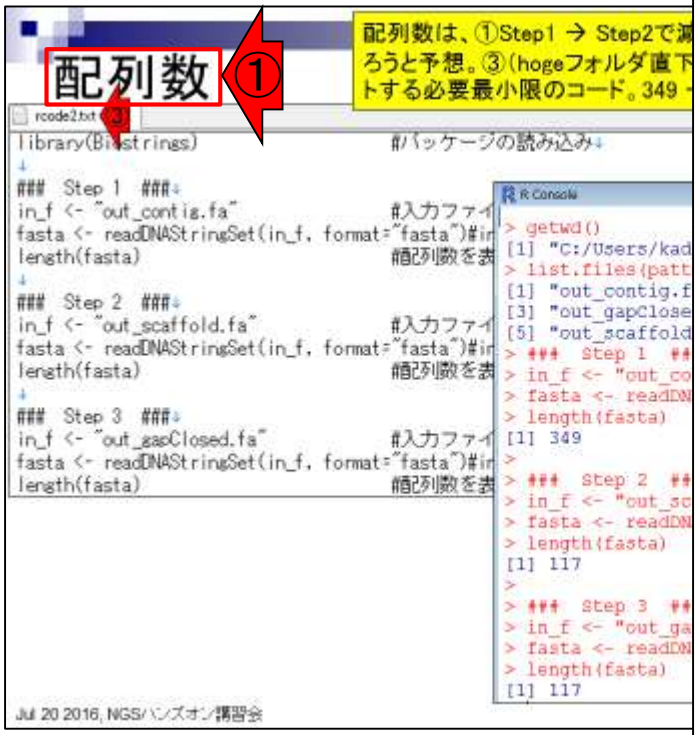
第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料のスライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

①スライド50についても、②でコピペ用コードを提供しています。この場合、タイトルが「配列数」と短いので、キーワード検索は厳しいかもしれません。しかし、③の次の内容なので、普通にページを下にスクロールすれば見つけられます



③

- 一気に結果を得る

実際の利用時は、rcode1.txtのような無駄なコメントを除いてスリムにした一連のスクリプトを作成しておき、一気にコピペする。ファイルの中身は以下と同じ。

```
param_base <- c("A", "C", "G", "T", "N")#出力させたい塩基を指定
library(Biostrings)                     #パッケージの読み込み

################
###  Step 1
################
in_f <- "out_contig.fa"                 #入力ファイル名を指定してin_fに格納
out_f <- "result_step1.txt"             #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
hoge <- alphabetFrequency(fasta)        #A,C,G,T,..の数を各配列ごとにカウントした結果
obj <- is.element(colnames(hoge), param_base)#条件を満たすかどうかを判定した結果をo
out <- apply(as.matrix(hoge[, obj]), 2, sum)#列ごとの総和をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=T, col.names=F)#o

################
###  Step 2
################
in_f <- "out_scaffold.fa"               #入力ファイル名を指定してin_fに格納
out_f <- "result_step2.txt"             #出力ファイル名を指定してout_fに格納

fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
```

②

- 配列数

rcode2.txtは、配列数をカウントする必要最小限のコード。ファイルの中身は以下と同じ。

```
library(Biostrings)                     #パッケージの読み込み
```

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の1スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

掟に従って、ページをスクロールしていかないと…、①の②スライド120のところで必ずつまづきます。なぜならここでコピペしてもらうコードは、どこか他のところに例題があるわけではなく

②のところにしかないからです！

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の1ページ目から25あたりからやる予定です。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

- k=6, 8, 10, 12
必要最小限のコードにして、k値の違い(k=6, 8, 10, and 12)による影響の全体像を把握。

```
in_f <- "sample34_ngs.fasta"                #入力ファイル名を指定してin_fに格納
library(Biostrings)                         #パッケージの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

param_kmer <- 6                             #k-merのkの値を指定
hoge <- oligonucleotideFrequency(fasta, width=param_kmer)#k連続塩基の出現頻度情報を
out <- colSums(hoge)                        #列ごとの総和をoutに格納
kmer <- out[out > 0]                        #1回以上出現したk-merのみをkmerに格納
length(kmer)                                #1回以上出現したk-merの種類数を表示
table(kmer)                                 #(k-merの種類は問わずに)k-merの出現回数分布を
median(kmer)                                #出現回数の中央値(median)を表示

param_kmer <- 8                             #k-merのkの値を指定
hoge <- oligonucleotideFrequency(fasta, width=param_kmer)#k連続塩基の出現頻度情報を
out <- colSums(hoge)                        #列ごとの総和をoutに格納
kmer <- out[out > 0]                        #1回以上出現したk-merのみをkmerに格納
length(kmer)                                #1回以上出現したk-merの種類数を表示
table(kmer)                                 #(k-merの種類は問わずに)k-merの出現回数分布を
median(kmer)                                #出現回数の中央値(median)を表示

param_kmer <- 10                            #k-merのkの値を指定
hoge <- oligonucleotideFrequency(fasta, width=param_kmer)#k連続塩基の出現頻度情報を
```

- 作図(k=6, 8, 10)
上記のk=6, 8, 10の結果より、横軸の最大値は39であることがわかっているが、ヒストグラム全体の変遷が把握しづらいので1から20の範囲を指定している。このあたりは結果を眺めながら自分の好みに合わせて微調整する。

```
in_f <- "sample34_ngs.fasta"                #入力ファイル名を指定してin_fに格納
param_fig <- c(450, 210)                    #ヒストグラム描画時の横幅と縦幅を指定(単位は
```

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の1スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

次は、作業ディレクトリについてです。②スライド54からは、どこで作業するかは明記していません。しかし、hoge4.faがどこに存在するかくらいは、自分で適切に判断しましょう。デフォルトはhogeフォルダ直下にありますので、作業ディレクトリの変更をするなり、現在の作業ディレクトリ中にhoge4.faをコピーするなりしましょう。例えばhogeフォルダ内にあるplatanusResultフォルダ中で作業を行いたいのなら、そこにはデフォルトではhoge4.faはないので、自分でコピーすればいいでしょ、ということです。

第1部の掟

第1部 | 統計解析 | ゲノム解析、塩基配列解析(2016.07.20)

当日は講義資料の1スライド25あたりからやる予定です。スライド24あたりまでは自習。

- 講義資料PDF(2016.06.08版; 約5MB)
- hoge.zip(2016.05.13版; 約2MB)

続、作業ディレクトリ。②スライド72のところは、入力ファイルはありません。出力ファイルが生成されるだけです。つまり自分がどこで作業を行っているかを適切に把握できていればそれでいいのです