# 農学生命情報科学特論 I(2014 年 6 月 18 日)課題の回答とコメント

受講 ID(5 桁):

学生証番号(ない人は空欄で可):

名前:

**課題:small RNA-seq リードのアダプター配列除去を自在に行う。入力ファイルは SRR609266.fastq.gz、テンプレートコードは rcode_adapter.txt。**

**Q1:アダプター配列除去後、18-44 塩基の範囲内にあるリードのみを抽出するためにはどこをどう変更すればよいか示せ。**

```
in_f <- "SRR609266.fastq.gz"          #入力ファイル名を指定してin_fに格納(RNA
out_f <- "hoge4.fasta.gz"             #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 2                   #許容するミスマッチ数を指定
#param_nBases <- 0                     #許容するACGT以外の文字数(実質的にはNの
param_range <- 20:30                  #配列長の範囲を指定
```

20:30 のところを 18:44 とするのが正解です。あるいはそれに準ずる指定方法であれば OK としています。

**Q2:アダプター配列除去後、18-44 塩基の範囲に含まれる総リード数を示せ。ただし、param_mismatch オプションは 2 のままとする。**

11,673,096 リードが正解です。オリジナルのリード数は 11,928,428 個であるため、この数値を記載していたヒトは間違いです。11,691,441 個や 11,653,225 個と記載している人が何人かいました。11,653,225 個は param_nBases を有効にした場合の結果ですね。これについては特に言及していませんでしたが OK です。11,691,441 個はよくわかりませんでしたが、付随した記載内容的には手続きはちゃんとできているようですので、それらも一応正解にしてあります。

**Q3:アダプター配列除去条件の違いについて考察せよ。**

間違ったことを書いていなければよしとしています。このデータの場合は「フィルタリング条件の多少の違いはマッピング結果にそれほど大きな影響を及ぼさない」のような考察に一票。Mac の場合は、gzip 圧縮ファイル状態での読み込みはうまくいかないようです。解凍したものを入力として与えればうまく読み込めると思います。

自由記載欄