GITC 2018 春 準備編　UNIX・R・NGS の基礎　演習

復習問題１　UNIX 基本コマンド　（*印は応用問題につき、時間に余裕があればトライしてみてください）

1. ターミナルを起動して以下のコマンドを実行せよ。

1-1. ~/data/1_unix ディレクトリに移動せよ。

1-2. カレントディレクトリ（現在のディレクトリ）の名前を確認せよ。

1-3. カレントディレクトリの内容を表示せよ（オプション"-R"を使うとディレクトリを辿りながら全てのファイルを表示する）。

（使用するコマンド：cd, pwd, ls）

2. ~/data/1_unix/sprot ディレクトリ内には複数の FASTA ファイル（拡張子が.fasta のファイル）がある。

2-1. それらの FASTA ファイル全てを、ワイルドカード「＊」を使って ~/unixtest/FASTA-EX ディレクトリにコピーせよ。ディレクトリがない場合は新規に作成すること。

2-2. 正しくコピーされたかを確認するために、~/unixtest/FASTA-EX ディレクトリの内容を表示せよ。

（使用するコマンド：mkdir, cp, cd（必要であれば）, ls）

3. ~/unixtest/FASTA-EX ディレクトリに移動せよ。

3-1. **2** でコピーした全ての FASTA ファイル内にある配列名（「>」で始まる行）を、grep コマンドを使用して抜き出せ。

3-2. 上記の grep コマンドの出力をパイプ「｜」で less コマンドに送り、どのように表示されているか確認せよ。

3-3. 複数のファイルに対して grep を実行すると、結果行にファイル名も付加される。 grep において、ファイル名を表示しないオプションを man コマンドで調べよ。調べる際には"filename"をキーワードとして検索すること。

3-4. 調べたオプションを使って、ファイル名が付加されない grep 結果を確認せよ。

（使用するコマンド：grep, less, man）

4. ~/unixtest/FASTA-EX にコピーした FASTA ファイルから、生物種ごとに Multi-FASTA 形式のファイルを作成することを考える（Multi-FASTA とは、1 ファイルの中に複数の FASTA 形式の配列が含まれるファイル）。

4-1. どの生物種の配列であるかは、ファイル名に含まれる文字：HUMAN, MOUSE, DROME によってわかる。cat コマンドおよびワイルドカードを使用して、ファイル名に「HUMAN」を含む全てのファイルの内容を画面に出力してみよ。

4-2. 上記の出力をリダイレクト「>」を使って、human.fasta というファイルに書き出せ。

4-3. less コマンドで human.fasta の内容を確認せよ。

*4-4. 同様に「MOUSE」→ mouse.fasta として、「DROME」→ drome.fasta として作成せよ。

（使用するコマンド：cat, less）

***5.** FASTA-EX ディレクトリのバックアップを作成してから削除する。

5-1. ディレクトリ全体の圧縮アーカイブを~/unixtest/FASTA-EX.tar.gz として tar コマンドで作成せよ(tar のオプションは"zcvf"を使用)。

5-2. 作成した、圧縮アーカイブの中身を確認せよ(tar のオプションは"ztvf"を使用)。

5-3. 確認できたら FASTA-EX ディレクトリを削除しせよ。

（使用するコマンド：tar, rm）

***6.** 圧縮アーカイブ FASTA-EX.tar.gz は自分自身のみ読み書きできるよう chmod コマンドを使用して（グループとその他は読み書きできないように)アクセス権を変更せよ。

***7.** UNIX にはファイル検索に用いる find という非常に便利なコマンドがある。man コマンドを使用して find コマンドの使用方法を確認し、この find コマンドを使用してホームディレクトリ配下のどこかにある"SYYM_DROME.sprot"を探せ。

復習問題２　エディタとスクリプト

この演習では、以下のフォルダのファイルを使用する。
~/data/2_editor/

１）テキストに記載している （演習）引数、および（演習）引数の使用　のスライドを読むこと。

２）example4.sh を作成せよ。新規にファイルを作成してもよいし、(もしあるなら) exapmle3.sh を修正したあと、別名で保存しても良い。
ファイルの編集には Emacs を使用せよ。

３）$ ./exapmle4.sh GO
として、example4.sh を実行し、実行結果を確認せよ。
実行できない場合、実行権が付与されているかを確認せよ。

この example4.sh は、実行した際に .tmp という中間ファイルが作成される。
中間ファイルは必要ないので、これを最終的に消したい。
そうなるように example4.sh を編集せよ。

編集した後のスクリプトを実行し、 ls コマンドで .tmp ファイルが存在しないことを確認せよ。

復習問題３　R

（基本演算編）

R で標準で使える women データを使って、以下の問題を考えよう。

1. まずコンソール上で women とタイプしてデータを表示せよ。women データは身長(height)が inch、体重(weight)が pound であらわされている。これを、身長を cm、体重を kg の単位に直したい。以下の手順でこれを行え。

(a) women から身長の列を抜き出し、これを cm に変換して、h という変数に代入せよ。ただし、1 inch=2.54cm である。

(b) women から体重の列を抜き出し、これを kg に変換して、w という変数に代入せよ。ただし、1 pound=0.454kg である。

(c) data.frame(height=h, weight=w)によって新しいデータフレームを作り、women2 という変数に代入せよ。`

2. 前問で作成した women2 の各行について、その身長と体重からボディマス指数(BMI)を計算せよ。ただし、体重 w kg、身長 h m （cm ではない）の人の BMI は w/h^2 で定義される。

（統計解析編）

R 入門の講義で用いた cancer データを使って、以下の問題を考えよう。変数 cancer が残っていない場合は、作業ディレクトリを~/data/3_R に変更してから read.table を使って読み込むこと（テキスト「データフレームの読込み１」）。

3. (a) 男性で喫煙歴がある人のデータを抜き出し、結果を cancer.subset という変数に代入せよ。何人いるか。

(b) それらの人の gene1 の発現データを取り出し、その平均値を計算せよ。

(c) 抽出した結果をタブ区切りテキストとして cancer.subset.txt というファイルに保存せよ。

4. (a) gene1 と gene2 の散布図を作成し、gender によって点を色分けせよ。

(b) 散布図に回帰直線を引け。この回帰直線へのあてはめは、有意水準を 0. 01 として有意であると言えるか。

*5. テキストでは、gene1 の発現量(gene1)が性別(gender)によって違いがあるという結論が t 検定で得られ、また癌のステージ(stage)によっても違いがあるという結論が分散分析から得られた。ただし、癌のステージの中で明らかな違いがあるのは stage III のみであった（これは boxpot (gene1 ~ stage, cancer)で確認できる）。

そこで、gender の効果を考慮しつつ stage の比較を行うため、Trellis Plot の技法を使ったプロットを作成してみよう。これを行う lattice package は R に標準で含まれているが、使う際に

はライブラリのロードが必要である。

```
> library(lattice)
> bwplot(gene1 ~ stage | gender, cancer)
```

bwplot は boxplot と同様に箱ひげ図を作成するが、特定の因子によって条件付けしたプロットを作成できる。ここでは、gender によってまず被験者を female と male に分けて、そのそれぞれで箱ひげ図を作成している。この結果から stage の gene1 の発現量への効果について、どのような結論が得られるか考察せよ。

*6. (a) cancer データから各患者の gene1 から gene6 の発現量を抜き出した部分データフレームを作成し、変数 expr に代入せよ。

(b) 各遺伝子間の発現量の相関（散布図を描いたときに傾きを持つ直線上に分布する傾向）の強さは相関係数によって表される。相関係数は-1 から 1 までの値を取り、0 が無相関を表す。相関係数が負の値のときは、傾きが負、すなわち一方が大きくなれば他方が小さくなる関係を表す。R では、行列の各カラム間の相関係数は、cor 関数によって一度に計算できる。これを用いて expr の各カラム間の相関係数を計算せよ。

(c) 相関係数の絶対値が 0.5 以上のときに強い相関があるとして、gene1～gene6 を、発現の相関の強さによっていくつかのグループに分けることができるかを検討せよ。ただし、絶対値をとるのは abs 関数で行える。

（関数の作成）

7. 与えられたベクトルに対し、二乗平均平方根(root mean square)を計算する関数を RMS という名前で作成せよ。ただし、二乗平均平方根は、ベクトルの各要素を二乗した値の平均値の平方根であり、与えられた値（ベクトル）の平方根をとる関数は sqrt である。また、関数はエディタを使って作成すること。作成した関数を使って RMS(1:5)を計算せよ。

*8. 「関数の作成(2)」のスライドで使用した plotAll 関数について考えよう。

(a)プログラムのソースコード（plotAll.R）を直接読み込むのではなく、エディタで開いてからマウスでコマンド全体を選択して実行してみよう（「エディタからのコマンドの入力と実行」参照）。これで plotAll 関数が定義される。これを用いて plotAll(cancer[,1:4])を実行せよ。

(b) plotAll 関数は、引数が一つのときは、そのデータフレーム内での総当たりのプロットを作成するが、対角線上とそれ以外とでは異なるコマンドでプロットを作成している。左上のプロット、およびその下の２行１列目のプロットと同じプロットを直接作成する plot コマンドはそれぞれどのようなものか、plotAll.R のプログラムから考えてみよ。ただし、タイトル(main)やラベル(xlab,ylab)をつけるところは難しいので無視してよい。

復習問題４　NGS 基本データフォーマット

~/data/4_format に移動せよ

1. bed ファイル(ex3.bed)と gtf ファイル(ex4.gtf)はヒト染色体上にある遺伝子群について同じ情報を表している。それぞれのファイルの形式の違いに注意しつつ、以下の問に答えよ。
   1）何番染色体にコードされているか。
   2）いくつの遺伝子(重複領域に別名のものもそれぞれ数える)が含まれているか。
   3）遺伝子 BC041449 にエキソンはいくつ含まれているか。
   4）遺伝子 BC041449 の最初のエキソンの開始位置と最後のエキソンの終了位置はそれぞれ何か。ただし、最初の塩基の位置座標は 1 とし、エキソンの開始、終了は転写される向きに沿って考えること
2. bed ファイルはタブ区切りのファイルである。
   1） R を使って ex3.bed からデータを読み込み、変数 bed に代入せよ。また変数 bed の内容を確認せよ。
   2） ex3.bed にはエキソンを一つから最大六つまでもつ遺伝子が含まれている。変数 bed からエキソン数の情報を取り出し、それぞれのエキソン数をもつ遺伝子がいくつずつあるかカウントせよ。ただし、与えられたベクトルの要素の頻度をカウントする関数は table である。
3. Sam ファイル(review_4-2.sam)は paired-end の map 結果である。
   1）ここに上がっている paired-end 数はいくつか。
   2）そのうち正しい paired-end の方向で map しているものはいくつか。

復習問題5　クオリティコントロールと NGS 基本ツール

~/data/5_ngs に移動せよ

1. 2D2L_rep1_R1.fastq と 2D2L_rep1_R2.fastq ファイルはアラビドプシスの発芽・緑化後の芽生えをサンプリング、ライブラリー作製した Paired-end read(76base x2)の RNA-Seq の生リードの fastq ファイルである。これを用いて、以下のパラメータを参考にし、paired-end での cutadapt にかけよ。

```
-q 30
-O 7
-m 50
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
-A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATC
```

   1）Cutadapt の log を見て、pass した pair 数、quality trim された base 数を調べよ。
   2）Cutadapt 処理前後の fastq ファイルを less コマンド等で見比べよ
   3）wc コマンドで cutadapt 前後の read 数を調べよ。
   4）Cutadapt 処理前後の fastq ファイルを fastqc にかけ、cutadapt 処理による、低品質配列が除かれていることを確認せよ。

2. Seqkit を使ってリードファイル ecoli.2.fastq, ecoli.3.fastq の statistic 情報を確認せよ。

3. bowtie2 を使って、リードファイル ecoli.2.fastq, ecoli.3.fastq を、リファレンス：eco にマッピングし、結果をファイル eco_ex.sam に出力せよ。その際、リードファイルはカンマ区切りで複数指定できることを使え。

4. samtools を使って、eco_ex.sam を bam に変換し、eco_ex.bam として保存せよ

5. samtools を使って、eco_ex.bam をソートし、eco_ex_sorted.bam として保存せよ
   現行 samtools は 4.5.の作業は一度にできうるが過程確認のため、今回は個別に行う

6. samtools を使って、eco_ex_sorted.bam にインデックスを作成せよ

7. samtools を使って、eco_ex_sorted.bam から以下の遺伝子にマップされたリードを取り出して数を数えよ。抽出された行を数えるには、 wc コマンドを使うこと。

染色体名	開始位置-終了位置	遺伝子名
chr	337 - 2799	thrA
chr	4179268 - 4183296	rpoB

復習問題6　UNIX によるテキストファイル処理

この演習では、ecoli.htseq を使用する。ファイルは下記のパスにある。

```
~/data/6_text/
```

ecoli.htseq には、アノテーションテーブルを使って、各遺伝子にマッピングされたリード数をカウントしたものが書かれている。

1)　ecoli.htseq に記載されていて、カウントされた値が 100 回以上ある行を標準出力に表示せよ。

2)　そのうち必要なのは b**** という文字列から始まる行である。それらのみを取り出して、ecoli.temp という名前のファイルに保存せよ。

3)　ecoli.temp を、カウントされた回数の多い順にソートし、ecoli.htseq.sorted という名前のファイルに保存せよ。

4)　1)~3)の操作を一つのコマンドで行えるよう、htsort.sh というシェルスクリプトを作成せよ。シェルスクリプトの実行後に中間ファイル(ecoli.temp)を消せることができていればなお良い。

*5)　4)で作成したスクリプトは、ecoli.htseq という特定のファイルにのみ使用できる。上記の処理を他のファイルに対しても行う必要が出てきた。
4)で作成したスクリプトを、引数を 1 つ使用する形に書き換えよ。
例えば、ecoli_other.htseq というファイルに対し、

```
$ ./htsort.sh ecoli_other.htseq
```

というコマンドを実行することで、別のファイルに対しても同じことが行えるようにせよ。