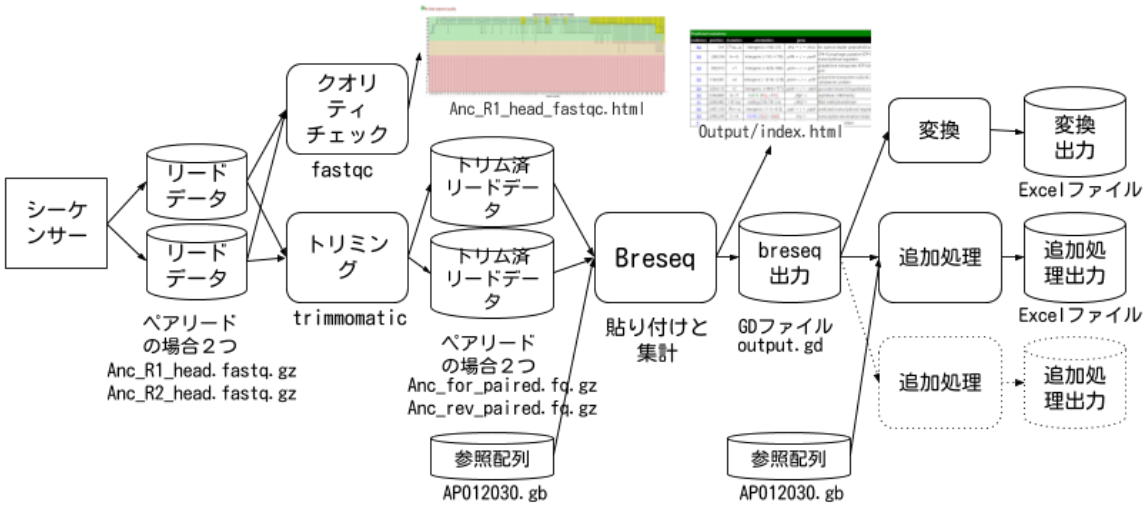
第１２回　Breseqの処理手順(続)

2019-06-10付　RNA-seqの流れと同じ内容

2020-01-20付け　Breseqの処理手順

2020-01-27　　山内 長承

Breseq処理の流れ（全体図）



トリミング再実験

前回、trimmomaticのサイトがアクセスできずダウンロードできなかったため、今回もう一度挑戦する。前回資料参照。

前回資料の訂正：　Breseqの入力として、トリミング後のforwardのみ

```
breseq/bin/breseq -o Output -r AP012030.gb Anc_for_paired.fq.gz > breseq.out
```

としていたが、これは間違いで、forwardとreverseの両方を入力する必要がある。

```
breseq/bin/breseq -o Output -r AP012030.gb Anc_for_paired.fq.gz Anc_rev_paired.fq.gz >
breseq.out
```

追加処理

Breseqから得られたGD（Gene Difference）ファイルには書かれていない情報（主に、変異個所が含まれている遺伝子/CDSに関する情報）を追加する。具体的には、一方から参照配列のアノテーション情報AP012030.gb を読み込み、他方でBreseqから得られたGDファイルを読み込み、GD上の変異の位置（全ゲノム内の塩基位置）から、その位置を含むCDSを見つける。そのCDS情報をもとにして、

- GBファイルのアノテーションからCDSに関する情報を抽出してリストする。

- CDSの位置範囲情報から対応する配列を切り出し、変異を適用して、変異後のCDS配列を作ってリストする。

  SNPの場合は変異部分を置き換える。INSの場合は追加する。DELの場合は削除する。MODの場合は置き換える。
- 変異がstop codonを生成する場合、変異後のCDSを最初に現れるstop codonまでに切る。
- 変異の部分のcodonを求めて、codonの変異をリストする。

実際の処理

プログラムをパッケージ化していないので、必要なプログラムファイルをダウンロードした後、入出力ファイルを書き直す必要がある。

ダウンロード：　https://pepper.is.sci.toho-u.ac.jp/DL　に置いておいた　ProcessGDX.zip　をダウンロードフォルダにダウンロードする。

展開・コピー：　ProcessGDX.zipファイルをクリックしてzip圧縮を展開すると、同じディレクトリに、新しく ProcessGDX ディレクトリができ、その中に３つのファイル、ProcessGDX.py（処理本体）、ParseGD.py（GDファイルの読込み）、ReadCDS.py（GenBankファイルのCDSの読込み）が作られる。この３つのファイルを、前回の作業ディレクトリ　Breseq-Seminar にコピーする。

ProcessGDX.py の中で、入力ファイル名と出力ファイル名が指定されている（プログラムに書き込まれていて、決め打ちになっている）ので、相当するファイル名（特に入力ファイル名、ディレクトリ）に合わせてプログラムを修正する。（今回は Output/output/output.gd を入力としてある。）

処理：　ProcessGDX.py を起動する。具体的には

```
python ProcessGDX.py
```

とする。まもなく　ParseGD と表示され、GDファイルの読込みが始まる。これが非常に時間がかかるので、進行状況を入力の1000行ごとに表示しているが、今回のoutput.gdのデータでは 54811 行あるので、その完了まで待たなければならない。（手元の非力MacBookAirでは小一時間かかった。）　入力が終了すると、変異ごとの処理を行い0～9が表示され（Ancなので全部で9個の変異がある）出力ファイルが生成される。

- Output_GD_trimmed_new.xlsx　（メインで扱うExcelの表）
- Output_GD_trimmed_new.pickle　（上記と同じデータを、Pandasのデータフレームのままの形でセーブしたファイル）
- Output.xlsx　（GDファイルを読んだ内容を、Excelに出力した表）
- Output.pickle　（上記と同じデータを、Pandasのデータフレームのままの形でセーブしたファイル）

２回目以降は、もしOutput.pickleファイルが存在すればそこからスタートする（つまり時間のかかるGDファイルの読込みをスキップし、読込み済みのデータフレームからスタートする）。もし、GDファイルの内容が変更されていれば、面倒だがOutput.pickleファイルを削除してからprocessGDX.pyの処理を行うこと。さもないと、Output.pickleに残っている前の状態のGDのデータを元に処理を行ってしまう。

もしハンズオン中に時間が足りなければ、上記処理を済ませた結果のファイル　Output.zip　を　https://pepper.is.sci.toho-u.ac.jp/DL　からダウンロードフォルダにダウンロードし、クリックしてzipを展開すると、フォルダ　Output　が出来てその中に上記のうちOutput.xlsxを除いた３つのファイルが入っている。その３つのファイルを作業ディレクトリ　Breseq-Seminar にコピーする。その上で、

```
python ProcessGDX.py
```

とすると、上で説明した「２回目以降」と同じ状態になっているので、GDファイルの読込みをスキップして、短時間で処理が終わる。

生成された出力　Output_GD_trimmed_new.xlsx　の内容を確認する。