ゲノム情報解析基礎（2019 年 5 月 13 日）課題（提出期限 2019 年 5 月 27 日）

受講 ID（5 桁）：

学生証番号（ない人は空欄で可）：

名前：

課題 1：DFAST への入力として与えた out_gapClosed.fa の配列数は 117 個だったが、 実行結果として配列数が 61 個となった。この理由について考えを述べよ。DFAST Help 画面内の記載事項をもとに述べてもよいし、R 上で検証した結果に基づいて述べてもよい。

課題 2：以下の中から興味ある事柄を 1 つ選び、簡単に説明せよ。

- Sequence logos の計算手順について(情報屋さん向け。内部で何を計算しているのかなど)。
  http://www.iu.a.u-tokyo.ac.jp/~kadota/20140430_kadota.pdf
  のスライド 30-35 あたりにも記載あり。

- GC skew について（バクテリア解析屋さん向け）。
  乳酸菌 NGS 連載第 9 回の p3-4 あたりにも記載あり。

- 配列のドットプロット（比較ゲノム屋さん向け）。
  乳酸菌 NGS 連載第 7 回の p106 や、第 13 回の p39-41 あたりにも記載あり。

- GFF3 ファイル annotation.gff からの CDS 情報抽出手段について（チャレンジャー向け）
  様々な戦略が考えられます。例えば、makeTxDbFromGFF 関数が読み込める形式にすべく、annotation.gff の中身をいじる方向性。他には、annotation.gff の中身は変更せずに最終的にうまく 2,311 個の CDS 領域情報を読み込んで TxDb オブジェクトを作成する方向性。GFF3 ファイルを読み込める別の関数を利用する手もあるかもしれません。

感想やコメントなど（optional）：