復習問題

先に `~/data/IU` に移動せよ

1. アダプター除去後の２つのリードファイル etec_1.cut.fq, etec_2.cut.fq を、それぞれ single end read のデータと見なしてマッピングし、paired end としてマップした場合と比較してみよう。

1) etec_1.cut.fq, etec_2.cut.fq を single end read のデータと見なして bowtie2 で etec をリファレンスとしてマッピングし、結果をファイル etec_bowtie2_single.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。
2) 出力ファイルの行数を、etec_bowtie2.sam と比較せよ。
3) それぞれのファイルの先頭 20 行を head で出力して比較し、以下の点について違いを論ぜよ。
   a) ヘッダ行
   b) 出力されるリードの並び
   c) フラグ値
   d) リードがマップされた位置
   e) ペアとなるリードがマップされた位置

2. 再び etec_1.cut.fq と etec_2.cut.fq を paired end として etec に対してマッピングするが、その際オプションとして –I 100 –X 200 を指定しよう。

1) これらのオプションはどういう意味を持っているか。
2) このコマンドを、出力ファイルを etec_bowtie2_X200.sam として実行せよ。
3) 出力ファイルの行数は、etec_bowtie2.sam と比べて変化したか。
4) ファイルの内容を以下のコマンドで比較し、どこが変わったか検討せよ。ただし、diff は 2 つのファイルを行ごとに比較して異なる行を出力するコマンドで、'<'で始まる行は最初のファイル、'>'で始まる行は 2 番目のファイルのみに出現する行を示す。また、less の-S オプションは、長い行を折り返さずに表示することを指示する。

```
$ diff etec_bowtie2.sam etec_bowtie2_X200.sam | less –S
```

3. samtools view の機能を使って etec_bowtie2_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。また、マッピングクオリティが 20 以上という条件をつけると数が変化するか。

	染色体名	開始位置-終了位置	遺伝子名
1)	ETEC_chr	336 - 2798	thrA
2)	ETEC_chr	4518271 - 4522299	rpoB

4. etec_bowtie2_sorted.bam から、samtools view -f を使ってペアが存在して両方ともがマップされていないリードを抽出して数を数えよ（与える FLAG 値がいくつになるのかを準備編テキスト 99 ページの表から考えよう）。

5. 以下のアライメントを表す CIGAR 文字列を作成せよ。

```
reference  ATGA-TGGTGTCGA
read       ATGGGTGGAG--GA
```

6. 同じディレクトリにある GTF 形式のファイル sox6.gtf に関して、以下の問いに UNIX コマンド grep, wc, sort, awk を用いて答えよ。

1) トランスクリプト NM_001145811.1 に関する行のみを抜き出し、sox6_tr1.gtf として保存せよ。以下 2)-4)はこのファイルを対象に調べよ。
2) このトランスクリプトにはいくつの exon が含まれているか。
3) このトランスクリプトに関する情報の各行を、開始位置（リファレンス配列上左端の位置）が転写される向きの順に並ぶように並べ替えよ。
4) このトランスクリプトの CDS の長さの和を計算せよ。
5) sox6.gtf には何種類のトランスクリプトが含まれているか。（ヒント : transcript_id のカラムを抜き出し、ユニークな行の数を数える。ユニークな行は sort –u を用いて抽出できる）

解答

1.
1) bowtie2 を実行するコマンド
```
$ bowtie2 -x etec -U etec_1.cut.fq,etec_2.cut.fq -S etec_bowtie2_single.sam
```

2) 行数のカウント
```
$ wc etec_bowtie2.sam etec_bowtie2_single.sam
```

行数はどちらも 100009 行で同じ。なお、行数のうち 9 行はヘッダ行、残りの 100000 行がマッピング結果で、各リードについて必ず 1 行のマッピング結果がある。

3) 各ファイル先頭 20 行の表示
```
$ head -20 etec_bowtie2.sam etec_bowtie2_single.sam
```

`head` で先頭 20 行を表示した結果から、`etec_bowtie2.sam`(以下 ***paired*** と呼ぶ)と、`etec_bowtue2_single.sam`(以下 ***single*** と呼ぶ)との間に以下のような違いが観察される。

a) ヘッダ行は@PG 行の CL（コマンドライン）のみが異なり、あとは同じ。
b) 1 カラム目のリード配列名が、***paired*** では各リードにつき 2 行続けて出力されるのに対して、***single*** では 1 行ずつしか出力されていない。***paired*** では 2 つのファイルが同時に読み込まれ、対応するリードが対として扱われているのに対して、***single*** では各ファイルが独立なものとして順次処理されている。
c) 2 カラム目のフラグは、***single*** では `0, 4, 16` の値をとるのに対し、***paired*** では `89, 73, 133` などの値をとっている。シングルの場合は、1 となるフラグは 4(セグメントがマップされなかった)または 16(逆鎖にマップされた)のいずれかのみであるのに対し、ペアの場合にはより多くの情報が格納されるため、異なる値となる。
d) 4 カラム目（マップされた位置）は、異なっている場合と同じ場合とがある。5 カラム目（マッピングクオリティ;MAPQ）が 42 になっているときには 4 カラム目が同じになっている点に注意しよう。MAPQ が高い場合は、ユニークにマップされたことを意味しており、位置はつねに同じになるが、MAPQ が低い場合は実際には複数箇所にマップされており、その中の一つがランダムに選ばれている。そこで、ペアで照合する際に別の位置にマップされたと考えられる。なお、相補鎖がマッチした場合は、10 カラム目は相補鎖の配列が、11 カラム目はクオリティ値が逆向きに表示されている。
e) 7-9 カラム目の、ペアとなるフラグメントに関する情報が、***single*** の方では出力されない（すべて* 0 0 となっている）。

2.
1) オプション `-I 100 -X 200` は、リード対をマップしたときのフラグメント長が 100 から 200 の 間の値であることを指示する（デフォルトは 0 から 500）。

2) bowtie2 を実行するコマンド
```
$ bowtie2 -x etec -1 etec_1.cut.fq -2 etec_2.cut.fq -S etec_bowtie2_X200.sam
    -I 100 -X 200
```

3)
```
$ wc etec_bowtie2.sam etec_bowtie2_X200.sam
```

行数はいずれも 100009 行で同じ。すなわち、リード対についての条件を変えてもデフォルトではすべての行が出力されるので、行数は変わらない。条件を満たすかどうかはフラグの値で表される。

4) 2 つの SAM ファイルの違いを表示
```
$ diff etec_bowtie2.sam etec_bowtie2_X200.sam | less -S
```

`diff` コマンドで表示した `etec_bowtie2.sam`（以下 ***default*** と呼ぶ)と、`etec_bowtie2_X200.sam`(以下 ***X200*** と呼ぶ)とで異なる行について、一般的に以下のような特徴が観察される（若干の例外はある）。

- 異なる行においては、2 カラム目（フラグ）の値が ***X200*** の方が ***default*** より 2 小さくなっている。ペアリードの間隔や向きが正しくマップされたかどうかは、フラグの 2 ビット目（2 進数の 10、すなわち 10 進数の 2）で示される。***X200*** の方が間隔に対する条件が厳しいため、***default*** で正しくマップされたと判定されたものが、***X200*** では正しくないと判定されることがあり、その場合にフラグの 2 ビット目が 1 から 0 に変化した結果、値が 2 小さくなった。
- 異なる行においては、9 カラム目の絶対値（フラグメントの長さ）が 200 より大きいか 100 より小さくなっている。そのような場合において、`-I 100 -X 200` の条件を満たさなくなるのでフラグの値が変化した。

3.
1)
```
$ samtools view etec_bowtie2_sorted.bam ETEC_chr:336-2798 | wc
```

2)
```
$ samtools view etec_bowtie2_sorted.bam ETEC_chr:4518271-4522299 | wc
```

数はそれぞれ 6 個と 195 個。
マッピングクオリティが 20 以上という条件を付けるには、コマンドに`-q 20` を加える。この場合は、結果は変わらない。

4.
```
$ samtools view -f 13 etec_bowtie2_sorted.bam | wc
```
ペアリードがあり(1)、自身がマップされていない(4)、相手もマップされていない(8)　フラグは合計 13。 このビット全てが立っているフラグは全て含まれる。矛盾しないフラグは以下の 2 通り
```
01001101 = 77
PAIRED,UNMAP,MUNMAP,READ1
10001101 = 141
PAIRED,UNMAP,MUNMAP,READ2
```

5. `4M1I5M2D2M`

6.
1)

```
$ grep 'NM_001145811¥.1' sox6.gtf > sox6_tr1.gtf
```
「.」は正規表現で「任意の 1 文字」を表すので、それを打ち消して「.」という文字のみにマッチさせるために「.」の前に¥（バックスラッシュ）をつけている。ただし、この場合はつけなくても結果は変わらない。

2) 15 個
```
$ grep exon sox6_tr1.gtf | wc
```

3)
```
$ sort –k 4,4nr sox6_tr1.gtf
```

4) 2,403 bp
```
$ awk '$3=="CDS"{sum+=($5-$4+1)} END{print sum}' sox6_tr1.gtf
```

5)　4 種類
```
$ awk '{print $16}' sox6.gtf | sort –u | wc
```