ex2: Transcript-based Mapping with Bowtie2

マウス Mus musculus のRNA-seqを行った。ライブラリは1種類のみで、single end (片側 Read1のみ) 75bpシークエンスを行った。これらのリードをマウスmRNAリファレンスにマッピングさせたい。

戦略：bowtie2でmRNAリファレンスにマッピング。

Data

データファイルは、~/data/SS 以下に保存してある。

Input reads

- IlluminaReads1.fq

Reference

- minimouse_mRNA.fa

Setup

Setup environment

ex2 ディレクトリをつくり、以下の解析はその下で作業しよう。

Sequence reads

''less'' などのコマンドで、 シーケンスファイル（IlluminaReads1.fq）の内容を確認する。

注）本番の解析では、リード数の確認、フォーマットの確認、クオリティの確認などを行う。必要であればアダプター配列の除去、低クオリティ部位のトリムも行う。

Reference sequence and annotation files

''minimouse_mRNA.fa'' の内容をless などで確認する。

Create index of reference

$ bowtie2-build reference.fasta output_basename

- reference.fasta : referenceのfastaファイル。今回の場合は minimouse_mRNA.fa（のパス）
- output_basename : 生成されるインデックスファイル群のbase name。

たとえば

```
bowtie2-build Data/RefSeq.MM9.cds.nr.fasta myref
```

を実行すると、

```
myref.1.bt2  myref.4.bt2
myref.2.bt2  myref.rev.1.bt2
myref.3.bt2  myref.rev.2.bt2
```

の6つのファイルができる。

Run Bowtie2

bowtie2でマッピングしよう。

```
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]
```

bowtie2には様々なオプションがあるが今回は最低限のオプションだけを設定して実行する。どのようなオプションが利用可能かは、''bowtie2 -h'' で確認できる。また開発者ホームページに詳細な解説がある。本番の解析では、適切なオプションを適切なパラメータで実行しなければいけない。実際は、いくつかパラメータを振って試行錯誤することになる。

```
$  bowtie2 -p 4 -x RefSeq.MM9.cds.nr -U mouse_200k.left.fq -S out.sam
```

- out.sam がマッピング結果 SAM format
- -p は使うCPUコア数。使用するコンピュータにあわせて設定する。

コマンドを実行するとしばらくして、

```
200000 reads; of these:
  200000 (100.00%) were unpaired; of these:
    114740 (57.37%) aligned 0 times
    68238 (34.12%) aligned exactly 1 time
    17022 (8.51%) aligned >1 times
42.63% overall alignment rate
```

のようなレポートが表示されて終了する。マッピング率など有用な情報なので、テキストファイルにコピー＆ペーストして保存しておくと良い。

Inspect Results

計算が終わったら、どのようなファイルが生成されたか確認する。 (''ls -l''など)

out.sam の内容を確認しよう (''less, head, tail''など).最初の約2万行はヘッダで、アライメントはそのあとに続く。

SAM to BAM

mapping結果を可視化したりカウントしたり、様々な下流解析を行うために、SAMファイルをsort済のBAMに変換する。そしてインデクシングする。SAM <=> BAM の変換は、NGS解析ではよく行う作業なので必ず身に付けること。

```
$ samtools view -bS out.sam > out.bam
$ samtools sort out.bam out.sorted
# => out.sorted.bam が生成される
$ samtools index out.sorted.bam
# => out.sorted.bam.bai が生成される
```

(optional) Count by transcript

samtoolsを使って、transcriptごとにカウントする簡易な方法を紹介する。

amtoolsのサブコマンド idxstats　は　reference sequenceのエントリー毎にマップされたリード数を集計する。今回は各シークエンスエントリーが各トランスクリプトに相当するので、これを利用するとtranscriptごとのカウント情報が得られる。

```
$ samtools idxstats out.sorted.bam
```