ex1: TopHat

キイロショウジョウバエ Drosophila melanogaster のRNA-seqを行った。ライブラリは2種類。それぞれsingle end（インサートの片側だけ読む）で75bpシークエンスした。10万リード得られた。これらのリードをD. melanogasterのゲノムにマッピングしたい。TopHatを用いてsplice-awareなマッピングを行う。

Data

Input reads (''~/data/EX/'' 以下にある）

- C1_10k_Read1.fq
- C2_10k_Read1.fq

Reference

- D. melanogaster genome and annotation (Ensembl BDGP5.25)

Notes

本来はD. melanogaster genome and annotation (Ensembl BDGP5.25) をiGenomes (http://tophat.cbcb.umd.edu/igenomes.html) からダウンロードする。

- ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Drosophila_melanogaster/Ensembl/BDGP5.25/Drosophila_melanogaster_Ensembl_BDGP5.25.tar.gz

ただ、ファイルサイズが比較的大きくダウンロードに時間がかかると思われる。演習用のMacに同じファイルが置いてある (''~/data/EX/'' 以下にある）ので、今回はそれを使って欲しい

- Drosophila_melanogaster_Ensembl_BDGP5.25.tar.gz

Setup

Setup environment

ex1 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。

```
$ mkdir ex1
$ cd ex1
```

dataのコピー

```
$ cp ~/data/EX/C1_10k_Read1.fq ./
$ cp ~/data/EX/C2_10k_Read1.fq ./
$ cp ~/data/EX/Drosophila_melanogaster_Ensembl_BDGP5.25.tar.gz ./
```

Sequence reads

''less'' などのコマンドで、 C1_10k_Read1.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。今回の演習ではスキップ。

Reference sequence and annotation files

Drosophila_melanogaster_Ensembl_BDGP5.25.tar.gz を解凍する

$tar xzvf Drosophila_melanogaster_Ensembl_BDGP5.25.tar.gz

Run tophat

TopHatを実行。

まず、C1_10k_Read1.fq　をマッピングしよう。

```
$ tophat -p 4 -G genes.gtf  -o C1_tophat_out  genome  read.fq
```

上のコマンドが基本形（そのままコピペしても動かない）。genes.gtf, genome, read.fq の部分には適切なファイル名等をいれて実行しよう。以下を参考にしてほしい。

- ''gene.gtf'' はknown transcriptが記録されたgtfファイル。ファイルパスを指定すること。ダウンロードした Drosophila_melanogaster_Ensembl_BDGP5.25　の中のどこかにあるので探してみよう。
- ''genome'' にはbowtie2用のゲノムのインデックスファイルのbase nameを指定する。ダウンロードした Drosophila_melanogaster_Ensembl_BDGP5.25　の中のどこかにあるので探してみよう。
- read.fq にはマッピングしたいシーケンスのファイルをfastqフォーマットで与える。
- -p は使うCPU coreを指定するオプション。使用するコンピュータのスペックに合わせて。
- 発展： --transcriptome-index　オプションは指定した方が良い。初回に作製したbowtie2 indexが2回目以降使い回せる。複数ライブラリを解析する際は大幅に時間の節約になる。今回は無視してよい。
- tophatコマンドのオプションや引数について詳しく知りたいときは、tophat -h としてヘルプ画面を表示させる。
- 解答 -- 自分で動かせるようになるまで見ない。

同様にC2_10k_Read1.fq をマッピング。

```
$ tophat -p 4 -G genes.gtf  -o C2_tophat_out genome C2_10k_Read1.fq
```

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 C1_tophat_out/
total 1048
-rw-r--r--  1 shige staff 1028268 Mar  6 23:10 accepted_hits.bam
-rw-r--r--  1 shige staff      52 Mar  6 23:10 deletions.bed
-rw-r--r--  1 shige staff      54 Mar  6 23:10 insertions.bed
-rw-r--r--  1 shige staff   25506 Mar  6 23:10 junctions.bed
drwxr-xr-x 17 shige staff     578 Mar  6 23:04 logs
-rw-r--r--  1 shige staff      66 Mar  6 23:04 prep_reads.inf
```

prep_reads.info の中身を''less''で確認しよう。

accepted_hits.bam がアライメント結果である。中身を''samtools''で確認しよう。

```
$ samtools view C1_tophat_out/accepted_hits.bam |less
```

IGV

IGV で可視化しよう。

IGVでbamファイルを読むためには、インデクシングをしなければいけない。sort => indexing の段階をふむ。

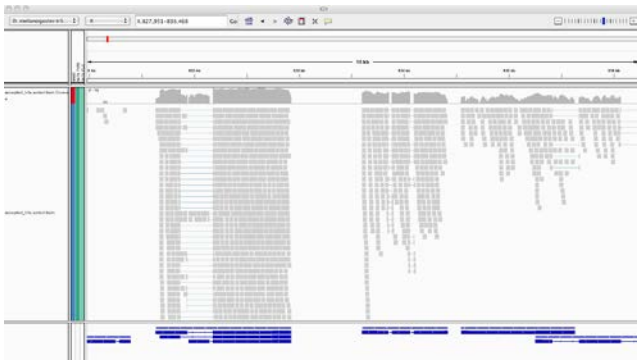
```
$ samtools sort accepted_hits.bam accepted_hits.sorted
# => accepted_hits.sorted.bam ができる
$ samtools index accepted_hits.sorted.bam
# => accepted_hits.sorted.bam.bai ができる
```

1. IGVを立上げる。

2. 左上のプルダウンメニューからDrosophila melanogaster のゲノムを選ぶ。(実は今回使っているリファレンスと同一のバージョンのD. melanogaster のゲノムデータではないが今回の練習ではr5.33を選んで問題はない)

3. メニュー　File > Load from File ... => accepted_hits.sorted.bam を選択

4. 適当な染色体の適当な場所を指定し、適当にズームアップする。（今回はX:830,000付近を見て欲しい）



X:830,000 近辺

注：今回は練習のために、X:830,000　付近にマップされるリードのみを利用しているため、その他の領域ではマッピングはほとんど見られない。