Case study 1: Genome-based RNA-Seq pipeline

アラビドプシス(Arabidopsis thaliana)のRNA-seqを行った。ライブラリは2D sample (2days dark conditionで生育させた黄色芽生え)と2D2L sample (その後さらに2days light conditionで生育させた緑化芽生え）でそれぞれsampling duplicateを３つ用意した。シーケンスはpaired-end（インサートの両端を読む）で101bpシークエンスしたものを事前にpre-processingしている。 これらのリードをArabidopsis thalianaのゲノムにマッピングする。TopHatを用いてsplice-awareなマッピングを行う。

Data

Input reads

(ファイルは、~/data/KY/tophat/　にある)

- condition Dark, rep#1:　2D_1_R1.fastq, 2D_1_R2.fastq
- condition Dark, rep#2:　2D_2_R1.fastq, 2D_2_R2.fastq
- condition Dark, rep#3: 2D_3_R1.fastq, 2D_3_R2.fastq
- condition Light, rep#1: 2D2L_1_R1.fastq, 2D2L_1_R2.fastq
- condition Light, rep#2: 2D2L_2_R1.fastq, 2D2L_2_R2.fastq
- condition Light, rep#3: 2D2L_3_R1.fastq, 2D2L_3_R2.fastq

Reference

- （本来ならば、Arabidopsis thaliana genome and annotation (Ensembl) をiGenomes (http://tophat.cbcb.umd.edu/igenomes.html) からダウンロードする　ftp://ussd-ftp.illumina.com/Arabidopsis_thaliana/Ensembl/TAIR10/Arabidopsis_thaliana_Ensembl_TAIR10.tar.gz）
- 今回は、そのままでは実習時間内では計算時間がかかり過ぎるので、今回は演習用にあらかじめChr4のみのデータに限定したgenomeファイル(genome_chr4.fa)とアノテーションファイル(genes_chr4.gtf)およびbowtie2のindexファイル(genome_chr4.fa.*.bt2)を用意してある。（KY/tophat/ディレクトリ）

Software

- tophat (installed)
- cufflinks (installed)
- bowtie2 (installed)
- samtools (installed)

Setup

Setup environment

top_cuff 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 top_cuff
$ cd top_cuff
```

Sequence reads

''less'' などのコマンドで、 2D_1_R1.fastq の内容を確認する。

注）Pre-processing済みであることが分かる。

Run tophat

TopHatを実行。

2D_1_R1.fastq

2D_1_R2.fastq

```
$ tophat -p 4 -G genes.gtf  -o 2D_1 genome.fa 2D_1_R1.fastq 2D_1_R2.fastq
```

*-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。

- オススメ： --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paired-endのデータを用いるが、single readでの解析もできる。

同様に他のもの計６サンプルをマッピング。

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/
```

```
$ $ ls -la 2D_1
total 92072
-rw-r--r--   1 kyamaguc  staff   3761633  3  2 17:14 accepted_hits.bam
-rw-r--r--   1 kyamaguc  staff       557  3  2 17:14 align_summary.txt
-rw-r--r--   1 kyamaguc  staff      5372  3  2 17:14 deletions.bed
-rw-r--r--   1 kyamaguc  staff      2850  3  2 17:14 insertions.bed
-rw-r--r--   1 kyamaguc  staff    407733  3  2 17:14 junctions.bed
drwxr-xr-x  30 kyamaguc  staff      1020  3  2 17:14 logs
-rw-r--r--   1 kyamaguc  staff       176  3  2 17:12 prep_reads.info
-rw-r--r--   1 kyamaguc  staff  33862783  3  2 17:14 unmapped.bam
```

prep_reads.infoやalign_summary.txtの中身を''less''で確認しよう。

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

```
$ samtools view 2D_1/accepted_hits.bam |less
```

IGV

IGV で可視化しよう。

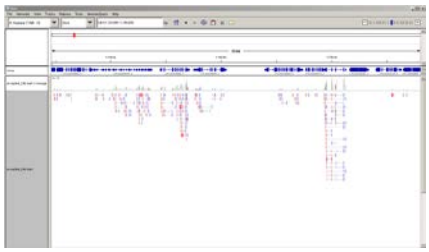
IGVでbamファイルを読むためには、インデクシングをしなければいけない。sort => indexing の段階をふむ。

```
$ 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A.thaliana(TAIR10)を選ぶ。

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

1. 第４染色体の適当な場所を指定し、適当にズームアップする。



X:1.14kb 近辺

Statistics

マップ率（インプットのリードの何%がリファレンスにマップされたか）を調べよう。

Run cufflinks

```
$ cufflinks -o 2D_1 -G genes.gtf accepted_hits.bam
```

- tophatと同じフォルダーに出力させておくのが良いだろう。

less コマンドで確認

```
-rw-r--r--   1 kyamaguc  staff   3761633  3  2 17:14 accepted_hits.bam
-rw-r--r--   1 kyamaguc  staff       557  3  2 17:14 align_summary.txt
-rw-r--r--   1 kyamaguc  staff      5372  3  2 17:14 deletions.bed
-rw-r--r--   1 kyamaguc  staff    422318  3  2 17:24 genes.fpkm_tracking
-rw-r--r--   1 kyamaguc  staff      2850  3  2 17:14 insertions.bed
-rw-r--r--   1 kyamaguc  staff    577476  3  2 17:24 isoforms.fpkm_tracking
-rw-r--r--   1 kyamaguc  staff    407733  3  2 17:14 junctions.bed
drwxr-xr-x  30 kyamaguc  staff      1020  3  2 17:14 logs
-rw-r--r--   1 kyamaguc  staff       176  3  2 17:12 prep_reads.info
-rw-r--r--   1 kyamaguc  staff         0  3  2 17:24 skipped.gtf
-rw-r--r--   1 kyamaguc  staff   8075446  3  2 17:24 transcripts.gtf
-rw-r--r--   1 kyamaguc  staff  33862783  3  2 17:14 unmapped.bam
```

新たにgenges.fpkm_tracking, isoforms.fpkm_trackingなどのファイルができている。

これらを他（2D_2, 2D_3, 2D2L_1, 2D2L_2, 2D2L_3）を含めて、全6sampleに関して行う。

Run cuffmerge

mergeするgtfファイルリストassemble.txtを作成する。以下を参考に自分のマシンに対応したパスで指定

```
~/top_cuff/2D_1/transcripts.gtf
~/top_cuff/2D_2/transcripts.gtf
~/top_cuff/2D_3/transcripts.gtf
~/top_cuff/2D2L_1/transcripts.gtf
~/top_cuff/2D2L_2/transcripts.gtf
~/top_cuff/2D2L_3/transcripts.gtf
```

cuffmergeを実行

```
$ cuffmerge -p 4 -s genome.fa -g genes.gtf assemblies.txt
```

*genome.fa, genes.gtfはパスを指定すること

merged_asmフォルダー下にmerged.gtfファイルが作成された。

less コマンドで確認

Run cuffdiff

GTFに記載の情報のみの解析なら、Run cufflinks, Run cuffmergeの部分はやる必要はない。

```
cuffdiff -p 4 merged.gtf -o 2D_vs_2D2L \
~/top_cuff/2D_1/accepted_hits.bam,~/top_cuff/2D_2/accepted_hits.bam,~/top_cuff/2D_3/accepted_hits.bam \
~/top_cuff/2D2L_1/accepted_hits.bam,~/top_cuff/2D2L_2/accepted_hits.bam,~/top_cuff/2D2L_3/accepted_hits.bam
```

2D_vs_2D2L ディレクトリに結果が出力されるので確認してみよう。

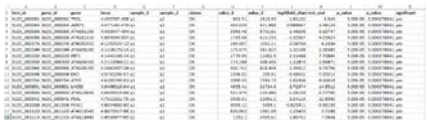
Explore the results

gene level での発現変動に興味があるので、見るべき結果ファイルは、

- gene_exp.diff
- genes.fpkm_tracking

tab区切りテキストなので、Excelに読み込ませることが可能。中身を確認しよう。 Excelのソート機能、フィルター機能を活用しよう。

Q: How many genes are differentially expressed?



Q: Scatter plot やMA plotを書いてみよう

Links

- http://tophat.cbcb.umd.edu/ |TopHat
- http://cufflinks.cbcb.umd.edu/ |CuffLinks

Notes

参考

Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2012).