# RNA-seq演習

高橋　弘喜

2018-03-23

## RNA-seq演習

テストデータを用いて，RNA-seq解析を実際にやってみる．テストデータとして，*Saccharomyces cerevisiae* を対象に取得されたデータを使用する[1]．

リードのマッピング，遺伝子発現比較までを遺伝研のスパコン上で行う．その後の解析は，ローカル環境で統計言語R[2]（cummeRbund[3]）を利用することで，各種統計量の可視化，ヒートマップなどの作成，有意差のあった遺伝子群の抽出などが可能である．

なお，今回の演習で紹介する方法以外にも，各遺伝子のリード数に基づいた解析も数多くなされている．ソフトウエアとしては，HTSeq[4], DESeq[5], DESeq2[6], edgeR[7] などが挙げられる．

### データ準備@スパコン

通常は，取得したデータを遺伝研スパコン上で解析するために，スパコン上へデータ転送を行う．今回は，スパコン上でデータをダウンロードし，解凍して作業を進める．

### ファイル転送

遺伝研スパコンにデータを転送する．

1. FileZilla, WinScpなどのファイル転送ソフトによって，データ転送を行う．
2. scpコマンドによるファイル転送

### 遺伝研スパコンへログイン

Windowsの場合は，TeraTermなどを用いる．Mac, Linuxの場合は，端末を起動して実行する．

```
$ ssh user@gw.ddbj.nig.ac.jp
$ qlogin
```

サーバーログイン後

```
#作業場所の確認
$ pwd
/home/hi-takah

#データのダウンロード
curl -O http://bioinfo.pf.chiba-u.jp/rna-seq-nig/rna-seq_20180323.zip

#ファイルリストの確認
$ ls

#配布データの解凍
$ unzip rna-seq_20180323.zip

#作業場所へ移動
$ cd rna-seq_20180323

#場所の確認
$ pwd
/home/hi-takah/rna-seq_20180323

#ファイルの確認
$ ls
fastq  genes.gtf  idx
```

**fastq の中身確認**

```
$ less fastq/SRR453566_1.fq
@SRR453566.2151832 HWI-ST167:4:2107:11419:39607 length=101
CCAGAATCTAGAAAGTTGTTTGCTGCTACTGTTGCAAAGGCTAAGACCATTGTCTGGAACGGTCCACCAGGTGTTTTCGAATTCG
+
CCCFFFFFHGHHHJJHIJIJJJIJIIJJJJJJJJJJJJJIJJIIJJJJJJIJJJGIJIIBHICECGGH7A=@@DDDCEDDDD
@SRR453566.4926005 HWI-ST167:7:2107:20424:50062 length=101
CGCCATTCTCTTGAAGTACTTGACACAAGTAGAGATGTTTTCTTCATCGGTTTCTTCAGACAAATCCAACATGTGGGAGGAGAAT
+
CCCFFFFFHHHHHJJJIJIJJJIJIJJJJHIJIJJJJIJJJIJJJFGIJJGBHHIJJJGGIIIIIJJIAHHIHHJGACEFBA;@@
```

### データの取得

用意したデータには下記のファイルが含まれている．

```
rna-seq_20180323/
├── fastq
│   ├── SRR453566_1.fq
│   ├── SRR453566_2.fq
│   ├── SRR453567_1.fq
│   ├── SRR453567_2.fq
│   ├── SRR453569_1.fq
│   ├── SRR453569_2.fq
│   ├── SRR453570_1.fq
│   └── SRR453570_2.fq
├── genes.gtf
└── idx
    └── Bowtie2Index
        ├── genome.1.bt2
        ├── genome.2.bt2
        ├── genome.3.bt2
        ├── genome.4.bt2
        ├── genome.fa
        ├── genome.rev.1.bt2
        └── genome.rev.2.bt2
```

| ファイル名 | condition | # of reads |
|---|---|---|
| SRR453566_1.fq | batch 1 | 100,000 |
| SRR453566_2.fq | | 100,000 |
| SRR453567_1.fq | batch 2 | 100,000 |
| SRR453567_2.fq | | 100,000 |
| SRR453569_1.fq | chemo 1 | 100,000 |
| SRR453569_2.fq | | 100,000 |
| SRR453570_1.fq | chemo 2 | 100,000 |
| SRR453570_2.fq | | 100,000 |

### マッピング

今回はTopHat2を用いて，RNA-seqリードをリファレンスゲノムにマッピングする．マッピングにおいて参照ゲノムと遺伝子情報ファイル（gtfファイル）が必要となる．

| 用意すべきファイル | ファイル名 | 備考 |
|---|---|---|
| fastq ファイル | ***.fastq, ***.fq, ***.fastq.gz, ***.fq.gz | paired, single いずれかの RNA-seq データ |
| gtf ファイル | genes.gtf (Saccharomyces cerevisiae (Yeast) Ensembl R64-1-1) | アノテーションファイル（今回は iGenomes[8] より取得） |
| リファレンスゲノムのインデックスファイル | Bowtie2Index | ない場合は bowtie2-build で作成する |

### マッピング (TopHat2)

RNA-seq のマッピングに関しては，多くのソフトウエアが開発されている．いずれも真核生物を対象に実装されている．

- TopHat[9]
- TopHat2[10]
- STAR[11]
- HISAT2[12]

原核生物の場合は，ゲノムシーケンス同様 bowtie2[13] などを用いる．

### コマンドの確認

スパコンにインストール済みのソフトウエア一覧 [14] を参照すると，TopHat2 に関して 6 つのバージョンが使用可能である．

| 名称 | バージョン | PATH |
|---|---|---|
| TopHat2 | 2.1.1 | /usr/local/pkg/tophat2/2.1.1 |
| | | /usr/local/pkg/tophat2/current |
| | 2.1.0 | /usr/local/pkg/tophat2/2.1.0 |
| | 2.0.13 | /usr/local/pkg/tophat2/2.0.13 |
| | 2.0.11 | /usr/local/pkg/tophat2/2.0.11 |
| | 2.0.5 | /usr/local/pkg/tophat2/2.0.5 |
| | 2.0.4 | /usr/local/pkg/tophat2/2.0.4 |
| Cufflinks | 2.2.1 | /usr/local/pkg/Cufflinks/2.2.1 |
| | | /usr/local/pkg/Cufflinks/current |
| | | /usr/local/bin |

| 名称 | バージョン | PATH |
|---|---|---|
| | 2.1.1 | /usr/local/pkg/Cufflinks/2.1.1 |
| | 2.0.1 | /usr/local/pkg/Cufflinks/2.0.1 |
| | 2.0.0 | /usr/local/pkg/Cufflinks/2.0.0 |

コマンドを確認してみる．

```
$ /usr/local/pkg/tophat2/2.1.1/tophat2
tophat:
TopHat maps short sequences from spliced transcripts to whole genomes.

Usage:
    tophat [options] <bowtie_index> <reads1[,reads2,...]> [reads1[,reads2,...]] \
                                    [quals1,[quals2,...]] [quals1[,quals2,...]]

Options:
    -v/--version
    -o/--output-dir                 <string>    [ default: ./tophat_out         ]
    --bowtie1                                   [ default: bowtie2              ]
```

## ライブラリーについて

strand specific のサンプル調整キットを使用している場合は，“–library-type” オプションを用いることで，リードの向きを考慮したマッピングが可能となる．次の cuffdiff においても同様のオプションを指定すればよい．

| Library Type | Examples | Description |
|---|---|---|
| fr-unstranded (default) | Standard Illumina | Reads from the left-most end of the fragment (in transcript coordinates) map to the transcript strand, and the right-most end maps to the opposite strand. |
| fr-firststrand | dUTP, NSR, NNSR | Same as above except we enforce the rule that the right-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand generated during first strand synthesis is sequenced. |

| Library Type | Examples | Description |
|---|---|---|
| fr-secondstrand | Directional Illumina (Ligation), Standard SOLiD | Same as above except we enforce the rule that the left-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand generated during second strand synthesis is sequenced. |

**tophat1.sh**

SRR453566 データを TopHat2 でマッピングを行う.

```
#!/bin/bash
#$ -S /bin/bash
#$ -N enshu_tophat
#$ -pe def_slot 8
#$ -cwd

f=("SRR453566")

tophat="/usr/local/pkg/tophat2/2.1.1/tophat2"

${tophat} -p 8 -o ${f} --GTF genes.gtf \
  idx/Bowtie2Index/genome \
  fastq/${f}_1.fq \
  fastq/${f}_2.fq
```

実行

```
$ qsub -l short -l s_vmem=1G -l mem_req=1G tophat1.sh
$ qstat
```

計算が終わると，SRR453566 ディレクトリが作成され，多くの結果ファイルが格納されている.

```
SRR453566/
├── accepted_hits.bam
├── align_summary.txt
├── deletions.bed
├── insertions.bed
├── junctions.bed
```

```
├──── logs
│     ├──── bam_merge_um.log
│     ├──── bowtie.left_kept_reads.log
│     ├──── bowtie.left_kept_reads.m2g_um.log
│     ├──── bowtie.left_kept_reads.m2g_um_seg1.log
│     ├──── bowtie.left_kept_reads.m2g_um_seg2.log
│     ├──── bowtie.left_kept_reads.m2g_um_seg3.log
│     ├──── bowtie.left_kept_reads.m2g_um_seg4.log
│     ├──── bowtie.right_kept_reads.log
│     ├──── bowtie.right_kept_reads.m2g_um.log
│     ├──── bowtie.right_kept_reads.m2g_um_seg1.log
│     ├──── bowtie.right_kept_reads.m2g_um_seg2.log
│     ├──── bowtie.right_kept_reads.m2g_um_seg3.log
│     ├──── bowtie.right_kept_reads.m2g_um_seg4.log
│     ├──── bowtie_build.log
│     ├──── g2f.err
│     ├──── g2f.out
│     ├──── gtf_juncs.log
│     ├──── juncs_db.log
│     ├──── long_spanning_reads.segs.log
│     ├──── m2g_left_kept_reads.err
│     ├──── m2g_left_kept_reads.out
│     ├──── m2g_right_kept_reads.err
│     ├──── m2g_right_kept_reads.out
│     ├──── prep_reads.log
│     ├──── reports.log
│     ├──── reports.samtools_sort.log0
│     ├──── run.log
│     ├──── segment_juncs.log
│     └──── tophat.log
├──── prep_reads.info
└──── unmapped.bam
```

**tophat.sh**

残りの 3 つのデータについても TopHat2 を実行する．複数サンプルのマッピングには，for ループを使用することができる．

```
#!/bin/bash
#$ -S /bin/bash
#$ -N enshu_tophat
#$ -pe def_slot 8
```

```
#$ -cwd

f=("SRR453567" "SRR453569" "SRR453570")

tophat="/usr/local/pkg/tophat2/2.1.1/tophat2"

for sample in ${f[@]}
do

    ${tophat} -p 8 -o ${sample} --GTF genes.gtf \
        idx/Bowtie2Index/genome \
        fastq/${sample}_1.fq \
        fastq/${sample}_2.fq

done
```

実行

```
$ qsub -l short -l s_vmem=1G -l mem_req=1G tophat.sh
$ qstat
```

### cuffdiff

TopHat2（それ以外のソフトウエアでも可）で得られたマッピング結果（TopHat2 の場合は，accepted_hits.bam）に基づいて，各遺伝子の発現比較を行う．

コマンドの確認

```
$ cuffdiff
cuffdiff v2.2.1 (4237)
-----------------------------
Usage:   cuffdiff [options] <transcripts.gtf> <sample1_hits.sam> <sample2_hits.sam> [
   Supply replicate SAMs as comma separated lists for each condition: sample1_rep1.sa
General Options:
  -o/--output-dir             write all output files to this directory
  -L/--labels                 comma-separated list of condition labels
```

**cuffdiff.sh**

マッピング結果を用いて，cuffdiff による遺伝子発現比較を行う．“-L/--labels” オプションで，各サンプルの名前をカンマ区切り（スペースは不要）で指定する．bam ファイルは，“-L/--labels” オプションと同じ順で記載する必要があるので注意する．反復実験のデータはカンマ区切りとして，サンプル間をスペースで繋ぐ.（n=1 同士の比較解析の場合でも，p 値が算出されるがほとんど意味をなさない.）

```
#!/bin/bash
#$ -S /bin/bash
#$ -N enshu_cuffdiff
#$ -pe def_slot 8
#$ -cwd

bam="accepted_hits.bam"
f=("SRR453566" "SRR453567" "SRR453569" "SRR453570")

cuffdiff -p 8 -o yeast -L batch,chemo \
    genes.gtf \
    ${f[0]}/${bam},${f[1]}/${bam} ${f[2]}/${bam},${f[3]}/${bam}
```

実行

```
$ qsub -l short -l s_vmem=1G -l mem_req=1G cuffdiff.sh
$ qstat
```

結果

下記のように，多くの結果ファイルが出力される．次のステップとしては，R(cummeRbund) を用いることで可視化などが実現できる．yeast ディレクトリそのものを cummeRbund の入力として使用する．

| ファイル名 | 内容 |
|---|---|
| isoforms.fpkm_tracking | Transcript FPKMs |
| genes.fpkm_tracking | Gene FPKMs. Tracks the summed FPKM of transcripts sharing each gene_id |
| cds.fpkm_tracking | Coding sequence FPKMs. Tracks the summed FPKM of transcripts sharing each p_id, independent of tss_id |
| tss_groups.fpkm_tracking | Primary transcript FPKMs. Tracks the summed FPKM of transcripts sharing each tss_id |

| ファイル名 | 内容 |
|---|---|
| isoforms.count_tracking | Transcript counts |
| genes.count_tracking | Gene counts. Tracks the summed counts of transcripts sharing each gene_id |
| cds.count_tracking | Coding sequence counts. Tracks the summed counts of transcripts sharing each p_id, independent of tss_id |
| tss_groups.count_tracking | Primary transcript counts. Tracks the summed counts of transcripts sharing each tss_id |
| isoforms.read_group_tracking | Transcript read group tracking |
| genes.read_group_tracking | Gene read group tracking. Tracks the summed expression and counts of transcripts sharing each gene_id in each replicate |
| cds.read_group_tracking | Coding sequence FPKMs. Tracks the summed expression and counts of transcripts sharing each p_id, independent of tss_id in each replicate |
| tss_groups.read_group_tracking | Primary transcript FPKMs. Tracks the summed expression and counts of transcripts sharing each tss_id in each replicate |
| isoform_exp.diff | Transcript-level differential expression. |
| gene_exp.diff | Gene-level differential expression. Tests differences in the summed FPKM of transcripts sharing each gene_id |
| tss_group_exp.diff | Primary transcript differential expression. Tests differences in the summed FPKM of transcripts sharing each tss_id |
| cds_exp.diff | Coding sequence differential expression. Tests differences in the summed FPKM of transcripts sharing each p_id independent of tss_id |

**gene_exp.diff**

gene_exp.diff に統計検定結果が要約されている．スパコンからダウンロードして，エクセルにて閲覧する．

```
$ less yeast/gene_exp.diff
test_id gene_id gene    locus   sample_1        sample_2        status  value_1 value
       p_value q_value significant
15S_rRNA        15S_rRNA        15S_rRNA        MT:6545-8194    batch   chemo   NOTES
inf    0       1
       1       no
```

```
21S_rRNA        21S_rRNA        21S_rRNA        MT:58008-62447  batch   chemo   NOTES
       1       no
HRA1    HRA1    HRA1    I:94686-99868   batch   chemo   NOTEST  41.7801 49.416  0.242
ICR1    ICR1    ICR1    IX:393883-397082        batch   chemo   NOTEST  7.4821  3.437
1.12192         0       1       1
       no
LSR1    LSR1    LSR1    II:680687-681862        batch   chemo   NOTEST  15.346  47.18
NME1    NME1    NME1    XIV:585586-585926       batch   chemo   NOTEST  0       0
PWR1    PWR1    PWR1    IX:393883-397082        batch   chemo   NOTEST  10.1823 0
inf     0       1       1       no
Q0010   Q0010   Q0010   MT:3951-4415    batch   chemo   NOTEST  0       0       0
```

### その他

講習会のデータ作成に用いたツール.

1. seqtk[15]: リードのサンプリング
2. SRA Toolkit[16]: SRA から sra データの取得，fastq への変換

### 参考文献


1. Nookaew, I. *et al.* A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in saccharomyces cerevisiae. *Nucleic Acids Res* **40,** 10084–10097

2. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2018).

3. Goff, L., Trapnell, C. & Kelley, D. *cummeRbund: Analysis, exploration, manipulation, and visualization of cufflinks high-throughput sequencing data.* (2013).

4. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169

5. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11,** R106 (2010).

6. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol* **15,** 550 (2014).

7. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140

8. iGenomes. Available at: http://jp.support.illumina.com/sequencing/sequencing_software/igenome.html.

9. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with rna-seq. *Bioinformatics* **25,** 1105–1111

10. Kim, D. *et al.* TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36

11. Dobin, A. *et al.* STAR: Ultrafast universal rna-seq aligner. *Bioinformatics* **29,** 15–21

12. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods* **12,** 357–360

13. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nat Methods* **9,** 357–359

14. 利用可能オープンソースソフトウェア. Available at: https://sc2.ddbj.nig.ac.jp/index.php/ja-avail-oss.

15. Seqtk. Available at: https://github.com/lh3/seqtk.

16. SRA toolkit. Available at: https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/.