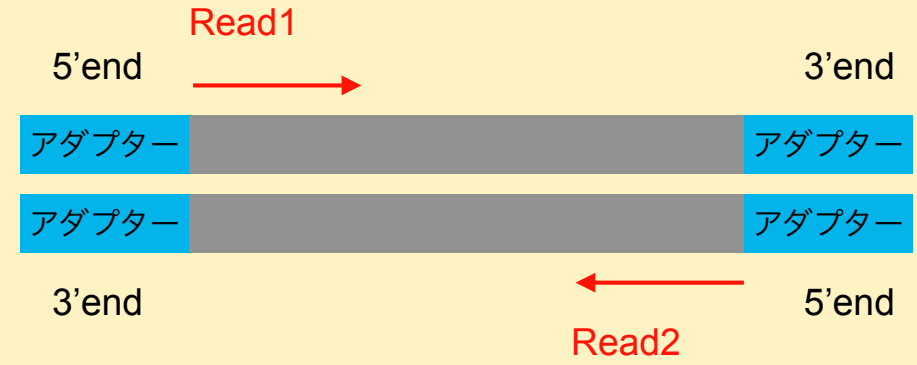
2025/11/28

遺伝研スパコンでの初歩的な解析の実践（RNA-seq解析等）

国立遺伝学研究所 望月孝子

RNA-seq とは

RNA-seq は、次世代シーケンサーを用いて RNA（主に mRNA）の配列を網羅的に読み取り、その情報から遺伝子発現量や転写の特徴を解析する手法。



現在は、ストランド特異的（strand-specific）ライブラリ作製が主流

Read1 / Read2 が “元の RNA のどちらの鎖（sense / antisense）を反映しているか” が、ライブラリ調製キットによって決まります。

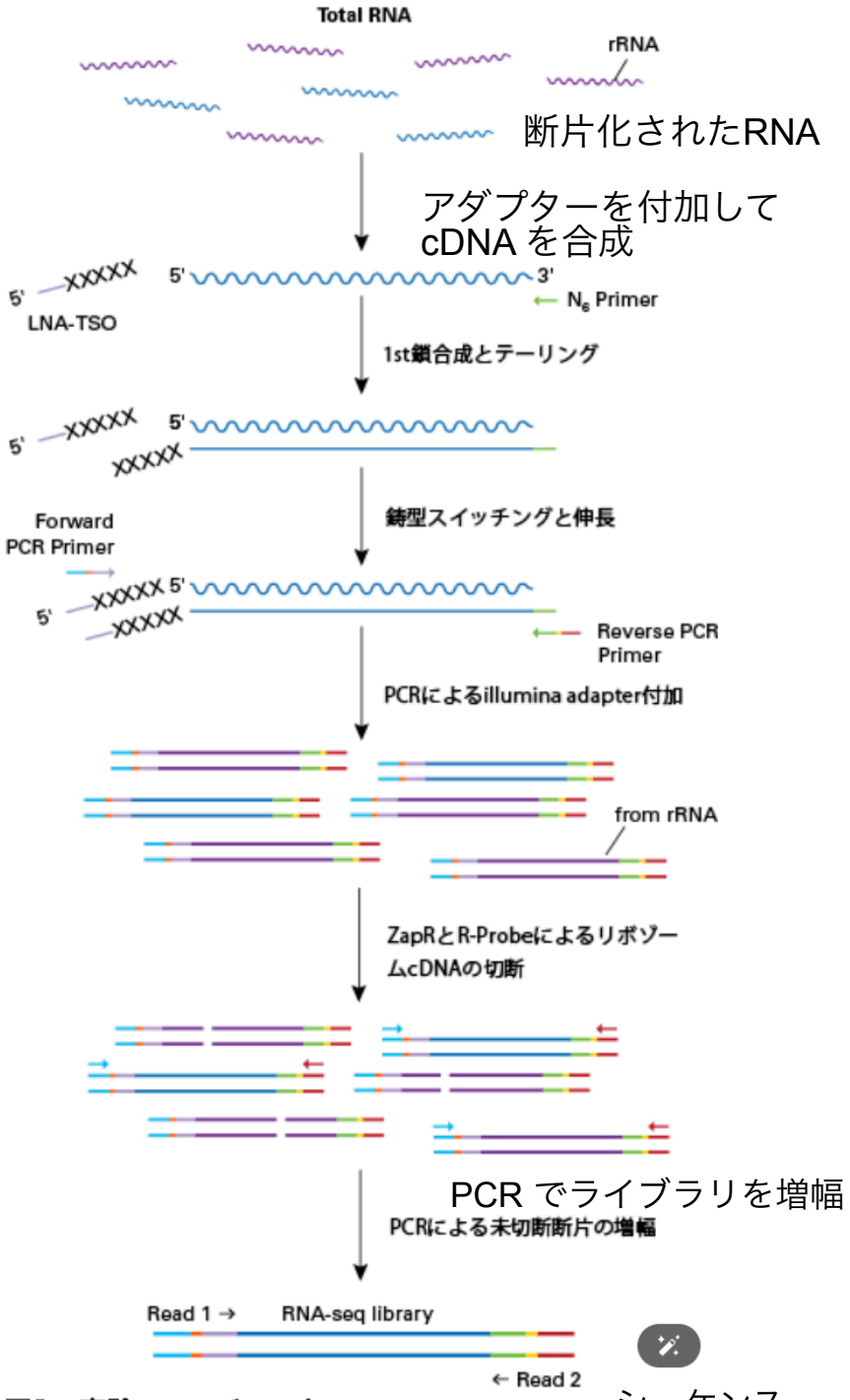
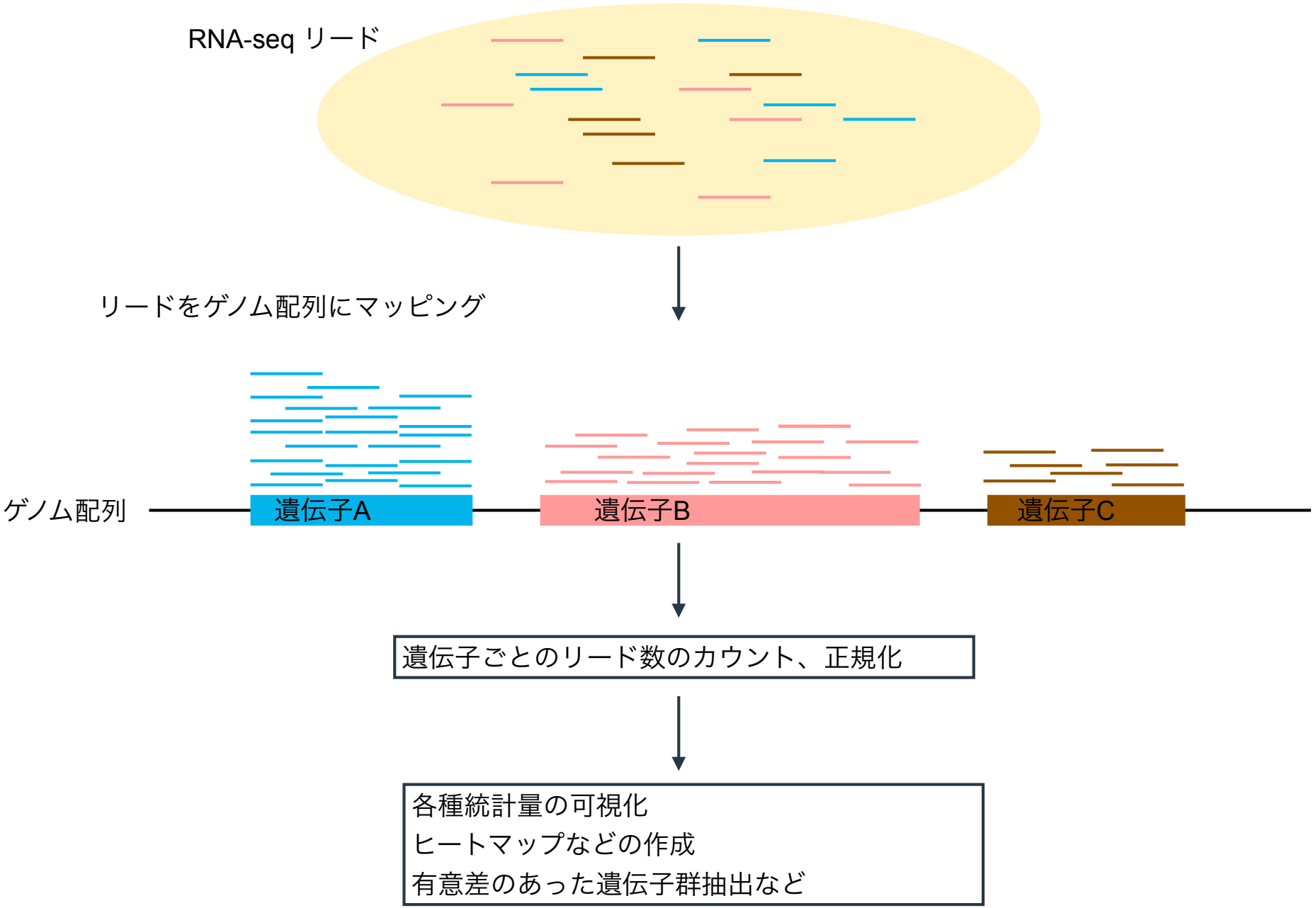


図1．実験フローチャート

SMARTer® Stranded Total RNA-Seq Kit - Pico Input Mammalian

https://catalog.takara-bio.co.jp/product/basic_info.php?unitid=U100009165

RNA-seq による発現量解析



DDBJ Sequence Read Archive (SRA)

INSDC (International Nucleotide Sequence Database Collaboration) の 一員として、塩基配列データを収集・公開。

NCBI (米) および ENA/EBI (欧) との間でデータの共有。

次世代シーケンサーからの出力データ (生データ) は、Sequece Read Archive に登録されています。

https://www.ddbj.nig.ac.jp/dra/index.html

実験をする前に検索すると、欲しいデータが見つかるかもしれません。

このスライドは、遺伝研 谷澤助教のスライドを参考にさせて頂きました。

本日の講習 ゲノム配列を利用した RNA-seq 発現量解析

本日
の講習

RNA-seq
(Fastq file)

リファレンスゲノムへマッピング
HISAT2

発現量の算出 Stringtie

計算力が必要なのでスパコンでの実行が好ましい

講習内では、
sbatch を実行しないでください。

その後の解析….

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

統計言語 R
ballgown, edgeR, DEseq/DEseq2 など

ローカルPC でもできる

配布データのコピー

ホームディレクトリ に移動

```
$ cd
```

ホームディレクトリ に20251128 というファイル/ディレクトリがないかを確認。

```
$ ls 20251128
ls: cannot access '20251128': No such file or directory
```

あれば、講習時間だけファイル/ディレクトリ 名前を一時的に変更してください。

```
$ mv 20251128 20251128_tmp
```

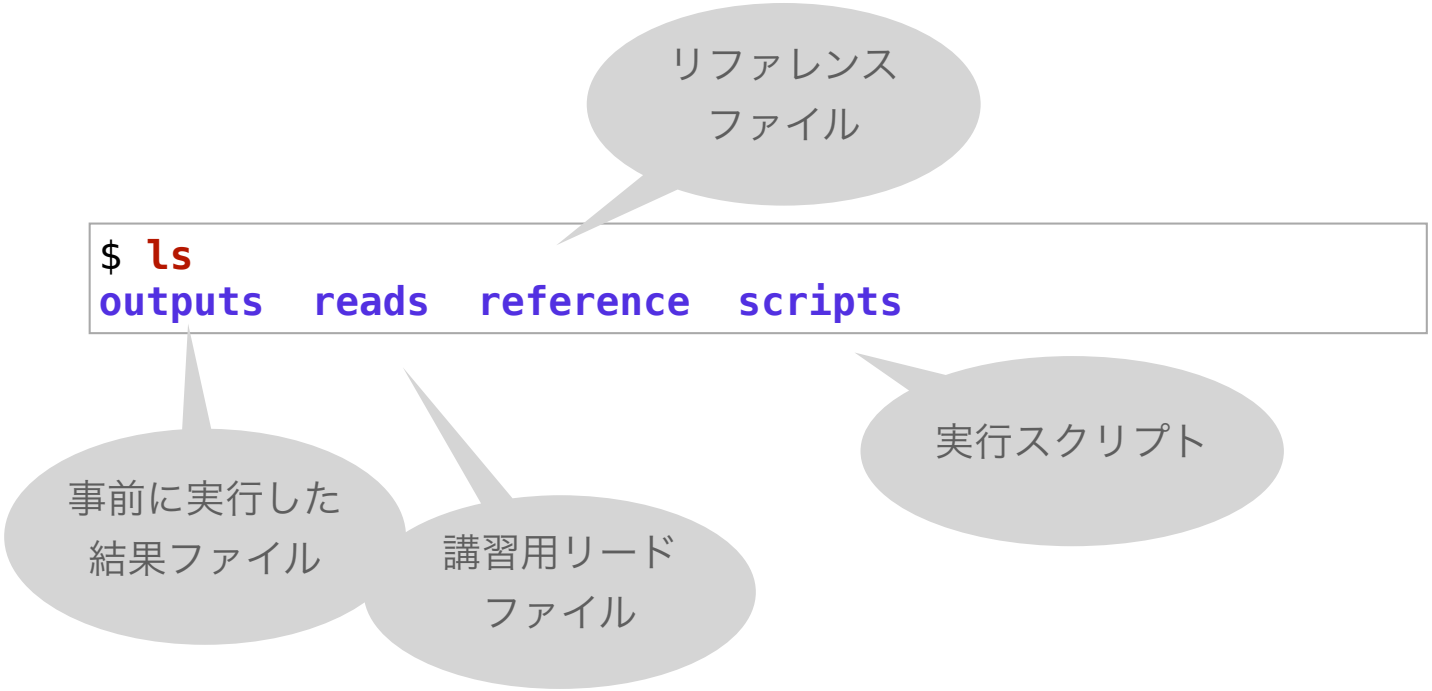
講習データをコピー

```
$ cp -r /home/koshu3/20251128 .
```

本日の講習はこちらのディレクトリで

```
$ cd 20251128
```

配布データの確認 (1)



それぞれのディレクトリ の中身は

```
$ ≈
$ ls outputs/hisat2_index
$ ls outputs/hisat2
$ ls outputs/stringtie
$ ls reads
$ ls reference
$ ls scripts
```

配布データの確認 (2)

配布データの構成

```
20251128
  |____ outputs  # 解析結果
              |___hisat2
              |___hisat2_index
              |___stringtie
  |----- reads      # リードファイル格納用
            |___SRR23499137_1.fastq.gz
            |___SRR23499137_2.fastq.gz
            |___SRR23499142_1.fastq.gz
            |___SRR23499142_2.fastq.gz
  |----- reference   # リファレンスファイル
            |___GCA_000269885.1_ASM26988v1_genomic.fna
            |___GCA_000269885.1_ASM26988v1_genomic.gff
  |----- scripts     # スクリプト
            |___hisat2.sh
            |___hisat2_index.sh
            |___stringtie.sh
```

講習用 RNA-seq データ

Metabolic Engineering
Volume 82, March 2024, Pages 201-215

Engineering *Saccharomyces cerevisiae* for fast vitamin-independent aerobic growth

Anja K. Ehrmann [a 1], Anna K. Wronska [b 1], Thomas Perli [b 1], Erik A.F. de Hulster [b], Marijke A.H. Luttik [b], Marcel van den Broek [b], Clara Carqueija Cardoso [b], Jack T. Pronk [b], Jean-Marc Daran [b]

[a] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Building 220, 2800 Kgs, Lyngby, Denmark
[b] Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 HZ, Delft, the Netherlands

Saccharomyces cerevisiae CEN.PK113-7D

SRR23499142

野生型（ビタミン要求性のある通常の実験株）
株：CEN.PK113-7D

SRR23499137

改変株（ビタミン非依存）
株：IMX2816

TruSeq stranded mRNA kit (illumina) を使用
Read 1 -> アンチセンス (−) 鎖にマップ
Read 2 -> センス鎖 (+) にマップ

```
$ ls reads/

SRR23499137_1.fastq.gz SRR23499137_2.fastq.gz   SRR23499142_1.fastq.gz SRR23499142_2.fastq.gz
```

Paired-end データ

Paired-end データ

FASTQ フォーマット　4行で1配列の情報を表す。

```
$ zcat reads/SRR23499137_1.fastq.gz | more

@SRR23499137.1 A00709:427:H3N7WDSX5:4:1101:2555:1000 length=150
NCGGTAGAAGTTGGTAGAGCAGAGGAGACTGTTTCTTGGGACACGGTCGAAGAGGCAGCACTGGAAGAGTGAGCCTCGCTAGTGGAGGAAGC
+SRR23499137.1 A00709:427:H3N7WDSX5:4:1101:2555:1000 length=150
#FFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

1行目: @ の後ろにその配列のID
2行目: 配列
3行目: + を記載する。(配列のID を記載してもしなくてもよい)
4行目: その配列のクオリティ値

クオリティ値はアスキーコードで表示
アスキー値 - 33 が クオリティ値

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                    |   |      |           |           |               |
33                    59  64     73          88          104             126
 0.....................26...31.....40
                       -5....0.....9.....................40
                            0.....9.....................40
                              3.....9.....................41
 0.2...................26...31.....41
 0.....................20.....30.....40.....50
 0.....................20.....30.....40.....50...55
 0.....................20.....30.....40.....50.....................................93
```

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
N - Nanopore      Phred+33,  Duplex reads typically (0, 50)
E - ElemBio AVITI Phred+33,  raw reads typically (0, 55)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

<- SRA

クオリティ値 @ の場合

64 - 33 = **31**

https://en.wikipedia.org/wiki/FASTQ_format

講習用 リファレンスファイル　ゲノム配列 fasta

ゲノム
配列

```
$ ls reference/
GCA_000269885.1_ASM26988v1_genomic.fna  GCA_000269885.1_ASM26988v1_genomic.gff
```

ファイルの中身を確認

```
$ more reference/GCA_000269885.1_ASM26988v1_genomic.fna
>CM001522.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome I, whole genome shotgun sequence
TTCAACCCTGACGAACCTGTCTCTCAACTTACCCTCCATTACCCTACCTCCCCACTCGTTACCCTGTCTCATTCAACCTT
ACCACTCCCAACCACCATCCATCTCTCTACTTACTACCACCAACCCACCGTCCACCATAACCGTTACCCTCCAATTACCC
```

FASTA ヘッダの出力

染色体16本分の配列
+Unplaced 配列

```
$ grep ">" reference/GCA_000269885.1_ASM26988v1_genomic.fna
>CM001522.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome I, whole genome shotgun sequence
>CM001523.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome II, whole genome shotgun sequence
>CM001524.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome III, whole genome shotgun sequence
>CM001525.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome IV, whole genome shotgun sequence
>CM001526.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome V, whole genome shotgun sequence
>CM001527.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome VI, whole genome shotgun sequence
>CM001528.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome VII, whole genome shotgun sequence
>CM001529.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome VIII, whole genome shotgun sequence
>CM001530.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome IX, whole genome shotgun sequence
>CM001531.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome X, whole genome shotgun sequence
>CM001532.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome XI, whole genome shotgun sequence
>CM001533.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome XII, whole genome shotgun sequence
>CM001534.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome XIII, whole genome shotgun sequence
>CM001535.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome XIV, whole genome shotgun sequence
>CM001536.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome XV, whole genome shotgun sequence
>CM001537.1 Saccharomyces cerevisiae CEN.PK113-7D chromosome XVI, whole genome shotgun sequence
.
.
```

講習用 リファレンスファイル　アノテーションファイル

アノテーション
ファイル

```
$ ls reference/
GCA_000269885.1_ASM26988v1_genomic.fna  GCA_000269885.1_ASM26988v1_genomic.gff
```

ファイルの中身を確認

```
$ more reference/GCA_000269885.1_ASM26988v1_genomic.gff
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM26988v1
#!genome-build-accession NCBI_Assembly:GCA_000269885.1
##sequence-region CM001522.1 1 223219
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=889517
CM001522.1  Genbank region  1  223219.+.ID=CM001522.1:1..223219;Dbxref=taxon:889517;Name
CM001522.1  Genbank gene 3162532998.+.ID=gene-CENPK1137D_4927;Name=CENPK1137D_4927;Note
CM001522.1  Genbank mRNA 3162532998.+.ID=rna-mrna.CENPK1137D_4927;Parent=gene-CENPK1137
CM001522.1  Genbank exon 3162532998.+.ID=exon-mrna.CENPK1137D_4927-1;Parent=rna-mrna.CE
CM001522.1  Genbank CDS  3162532998.+0ID=cds-EIW12309.1;Parent=rna-mrna.CENPK1137D_4927
```

GFF フォーマット

遺伝子アノテーションのフォーマット

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM26988v1
#!genome-build-accession NCBI_Assembly:GCA_000269885.1
##sequence-region CM001522.1 1 223219
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=889517
CM001522.1  Genbank  region  1      223219  .  +  .  ID=CM001522.1:1..223219;Dbxref=taxon:889517;Name=I;chromoson
CM001522.1  Genbank  gene    31625  32998   .  +  .  ID=gene-CENPK1137D_4927;Name=CENPK1137D_4927;Note=Simi
CM001522.1  Genbank  mRNA    31625  32998   .  +  .  ID=rna-mrna.CENPK1137D_4927;Parent=gene-CENPK1137D_4927;g
CM001522.1  Genbank  exon    31625  32998   .  +  .  ID=exon-mrna.CENPK1137D_4927-1;Parent=rna-mrna.CENPK1137D
CM001522.1  Genbank  CDS     31625  32998   .  +  0  ID=cds-EIW12309.1;Parent=rna-mrna.CENPK1137D_4927;Dbxref=N
```

タブ区切りフォーマット。値がない場合は、”.” が設定される。

1. seqname：染色体 or スキャフォールドの名前
2. source：アノテーションを生成したプログラムまたはデータソースの名前
3. feature：フィーチャータイプ（mRNA, gene, exon, CDS ….）
4. start：スタートポジション (1bp ～)
5. end：エンドポジション (1bp ～)
6. score：スコア
7. strand：+(forward)、-(reverse)または '.'
8. frame：翻訳フレーム（0, 1, 2）
9. attribute：追加情報。セミコロンで区切られたタグと値のペアのリスト。

リファレンスゲノムへリードをマッピング

ステップ

1. リファレンスゲノムのインデックスを作成

   hisat2_index.sh

2. リードをリファレンスゲノムへマッピング

   hisat2.sh　　2サンプル分をアレイジョブで同時実行

スクリプト

```
$ ls scripts/hisat2*
scripts/hisat2.sh  scripts/hisat2_index.sh
```

マッピング用インデックスの作成　　スクリプトの確認

```
$ more scripts/hisat2_index.sh
#!/bin/bash
#SBATCH --partition=rome
#SBATCH -output=%x_%j.out
#SBATCH --error=%x_%j.err
#SBATCH --ntasks=1
#SBATCH -cpus-per-task=1
#SBATCH --mem=10G
#SBATCH -time=10:00:00

GENOME=./reference/GCA_000269885.1_ASM26988v1_genomic.fna
INDEX=./reference/GCA_000269885.1_ASM26988v1_genomic.fna

apptainer exec -B ${HOME} /usr/local/biotools/h/hisat2:2.2.1--h503566f_8 \
  hisat2-build ${GENOME} ${INDEX}
```

リファレンスゲノムの
インデックス化

sbatch オプション

#SBATCH --partition=rome　　# 使用する計算ノード（CPUノード rome）

#SBATCH --output=%x_%j.out　　# 標準出力の保存先

%x = ジョブ名（スクリプト名）

%j = ジョブID

#SBATCH --error=%x_%j.err　　# 標準エラーの保存先（エラー内容が入る）

#SBATCH --ntasks=1　　# 実行するタスク数（通常は1）

#SBATCH --cpus-per-task=1　　# 1タスクが使うCPUコア数

#SBATCH --mem=10G　　# 使用メモリ量（10GB を確保）

#SBATCH --time=10:00:00　　# 最大実行時間。超えるとジョブは終了

apptainer オプション

-B ${HOME}　　# ホームディレクトリをコンテナ内に見えるようにする

マッピング用インデックスの作成　　オプションの確認

```
$ apptainer exec /usr/local/biotools/h/hisat2:2.2.1--h503566f_8 hisat2-build -h

HISAT2 version 2.2.1 by Daehwan Kim (infphilo@gmail.com, http://www.ccb.jhu.edu/people/infphilo)
Usage: hisat2-build [options]* <reference_in> <ht2_index_base>
    reference_in            comma-separated list of files with ref sequences
    hisat2_index_base       write ht2 data to files with this dir/basename
Options:
    -c                      reference sequences given on cmd line (as
                            <reference_in>)
    --large-index           force generated index to be 'large', even if ref
                            has fewer than 4 billion nucleotides
    -a/--noauto             disable automatic -p/--bmax/--dcv memory-fitting
    -p <int>                number of threads
    --bmax <int>            max bucket sz for blockwise suffix-array builder
    --bmaxdivn <int>        max bucket sz as divisor of ref len (default: 4)
    --dcv <int>             diff-cover period for blockwise (default: 1024)
    --nodc                  disable diff-cover (algorithm becomes quadratic)
    -r/--noref              don't build .3/.4.ht2 (packed reference) portion
    -3/--justref            just build .3/.4.ht2 (packed reference) portion
    -o/--offrate <int>      SA is sampled every 2^offRate BWT chars (default: 5)
    -t/--ftabchars <int>    # of chars consumed in initial lookup (default: 10)
    --localoffrate <int>    SA (local) is sampled every 2^offRate BWT chars (default: 3)
    --localftabchars <int>  # of chars consumed in initial lookup in a local index (default: 6)
    --snp <path>            SNP file name
    --haplotype <path>      haplotype file name
    --ss <path>             Splice site file name
    --exon <path>           Exon file name
    --repeat-ref <path>     Repeat reference file name
    --repeat-info <path>    Repeat information file name
    --repeat-snp <path>     Repeat snp file name
    --repeat-haplotype <path>   Repeat haplotype file name
    --seed <int>            seed for random number generator
    -q/--quiet              disable verbose output (for debugging)
    -h/--help               print detailed description of tool and its options
    --usage                 print this usage message
    --version               print version information and quit
```

マッピング用インデックスの作成　実行

実行

講習中はqsub を実行しないでください。

```
$ sbatch scripts/hisat2_index.sh
Submitted batch job 10265055
```

ステータスの確認

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          10265055      rome hisat2_i   koshu3 PD       0:00      1 (None)
```

実行待機

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          10265055      rome hisat2_i   koshu3  R       0:02      1 at145
```

実行中

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
```

ジョブが終了すると該当ジョブID が表示されなくなる。

マッピング用インデックスの作成　　実行結果の確認

```
$ ls
hisat2_index.sh_10265055.err  hisat2_index.sh_10265055.out  outputs  reads  reference  scripts
```

ログの確認

```
$ more hisat2_index.sh_10265055.err
Settings:
  Output files: "./reference/
GCA_000269885.1_ASM26988v1_genomic.fna.*.
ht2"
  Line rate: 6 (line is 64 bytes)
  Lines per side: 1 (side is 64 bytes)
  Offset rate: 4 (one in 16)
  FTable chars: 10
   .
   .
```

```
$ more hisat2_index.sh_10265055.out
Building DifferenceCoverSample
  Building sPrime
  Building sPrimeOrder
  V-Sorting samples
  V-Sorting samples time: 00:00:00
  Allocating rank array
  Ranking v-sort output
  Ranking v-sort output time: 00:00:00
  Invoking Larsson-Sadakane on ranks
      .
      .
Returning block of 1956549 for bucket 7
```

インデックスが作成された

```
$ ls -l reference/
total 38848
-rw-r--r-- 1 koshu3 koshu 12157033 Nov 15 20:41 GCA_000269885.1_ASM26988v1_genomic.fna
-rw-r--r-- 1 koshu3 koshu  8035055 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.1.ht2
-rw-r--r-- 1 koshu3 koshu  2868972 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.2.ht2
-rw-r--r-- 1 koshu3 koshu     5318 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.3.ht2
-rw-r--r-- 1 koshu3 koshu  2868966 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.4.ht2
-rw-r--r-- 1 koshu3 koshu  5604379 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.5.ht2
-rw-r--r-- 1 koshu3 koshu  2919388 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.6.ht2
-rw-r--r-- 1 koshu3 koshu       12 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.7.ht2
-rw-r--r-- 1 koshu3 koshu        8 Nov 19 20:05 GCA_000269885.1_ASM26988v1_genomic.fna.8.ht2
-rw-r--r-- 1 koshu3 koshu  5292953 Nov 15 20:41 GCA_000269885.1_ASM26988v1_genomic.gff
```

事前実行した結果で確認

```
$ ls outputs/hisat2_index/
$ more outputs/hisat2_index/hisat2_index.sh_10265055.err
$ more outputs/hisat2_index/hisat2_index.sh_10265055.out
```

マッピング　　スクリプトの確認

```
$ more scripts/hisat2.sh
#!/bin/bash
#SBATCH --partition=rome                # CPUノード
#SBATCH --output=%x_%A_%a.out           # 標準出力　%A=ArrayID %a=タスク番号
#SBATCH --error=%x_%A_%a.err            # 標準エラー
#SBATCH --ntasks=1                      # タスク数
#SBATCH --cpus-per-task=4               # 各タスクで使用するCPUスレッド数
#SBATCH --mem=16G                       # メモリ
#SBATCH --time=10:00:00                 # 最大実行時間
#SBATCH --array=0-1                     # アレイジョブ（例：2サンプル分）
#

# IMX2816: SRR23499137
# wild: SRR23499142
ACESSIONS=(23499137 23499142)

NUM=${ACESSIONS[${SLURM_ARRAY_TASK_ID}]}
PREFIX=SRR${NUM}

# read file
DIR=./reads/
QUERY1_1=${DIR}${PREFIX}"_1.fastq.gz"
QUERY1_2=${DIR}${PREFIX}"_2.fastq.gz"

apptainer exec -B ${HOME} /usr/local/biotools/h/hisat2:2.2.1--h503566f_8 \
   hisat2 -p ${SLURM_CPUS_PER_TASK} -x ./reference/GCA_000269885.1_ASM26988v1_genomic.fna --dta \
   -1 ${QUERY1_1} -2 ${QUERY1_2} \
   --rna-strandness RF \
   -S ${PREFIX}.sam

# convert sam to bam
# sort by position
apptainer exec -B ${HOME} /usr/local/biotools/s/samtools:1.22.1--h96c455f_0 \
   samtools sort -@ ${SLURM_CPUS_PER_TASK} ${PREFIX}.sam -o ${PREFIX}.sorted.bam
```

アレイジョブを指定

#SBATCH –array=0-6%2
とするとアレイジョブの同時実行数を制限できる

アレイジョブのタスクID

—dta: reports alignments tailored for transcript assemblers

```
$ apptainer exec /usr/local/biotools/h/hisat2:2.2.1--h503566f_8 hisat2 -h

HISAT2 version 2.2.1 by Daehwan Kim (infphilo@gmail.com, www.ccb.jhu.edu/people/infphilo)
Usage:
  hisat2 [options]* -x <ht2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <ht2-idx>  Index filename prefix (minus trailing .X.ht2).
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.

Options (defaults in parentheses):

 Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  --qseq             query input files are in Illumina's qseq format
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 <m1>, <m2>, <r> are sequences themselves, not files
  -s/--skip <int>    skip the first <int> reads/pairs in the input (none)
  -u/--upto <int>    stop after first <int> reads/pairs (no limit)
  -5/--trim5 <int>   trim <int> bases from 5'/left end of reads (0)
  -3/--trim3 <int>   trim <int> bases from 3'/right end of reads (0)
  --phred33          qualities are Phred+33 (default)
  --phred64          qualities are Phred+64
  --int-quals        qualities encoded as space-delimited integers

 Presets:                 Same as:
   --fast                 --no-repeat-index
   --sensitive            --bowtie2-dp 1 -k 30 --score-min L,0,-0.5
   --very-sensitive       --bowtie2-dp 2 -k 50 --score-min L,0,-1

 Alignment:
  --bowtie2-dp <int> use Bowtie2's dynamic programming alignment algorithm (0) - 0: no dynamic programming, 1:
conditional dynamic programming, and 2: unconditional dynamic programming (slowest)
  --n-ceil <func>    func for max # non-A/C/G/Ts permitted in aln (L,0,0.15)
```

samtools sort オプションの確認

```
$ apptainer exec /usr/local/biotools/s/samtools:1.22.1--h96c455f_0 samtools sort -h

sort: invalid option -- 'h'
Usage: samtools sort [options...] [in.bam]
Options:
  -l INT     Set compression level, from 0 (uncompressed) to 9 (best)
  -u         Output uncompressed data (equivalent to -l 0)
  -m INT     Set maximum memory per thread; suffix K/M/G recognized [768M]
  -M         Use minimiser for clustering unaligned/unplaced reads
  -R         Do not use reverse strand (only compatible with -M)
  -K INT     Kmer size to use for minimiser [20]
  -I FILE    Order minimisers by their position in FILE FASTA
  -w INT     Window size for minimiser indexing via -I ref.fa [100]
  -H         Squash homopolymers when computing minimiser
  -n         Sort by read name (natural): cannot be used with samtools index
  -N         Sort by read name (ASCII): cannot be used with samtools index
  -t TAG     Sort by value of TAG. Uses position as secondary index (or read name if -n is set)
  -o FILE    Write final output to FILE rather than standard output
  -T PREFIX  Write temporary files to PREFIX.nnnn.bam
      --no-PG
               Do not add a PG line
      --template-coordinate
               Sort by template-coordinate
      --input-fmt-option OPT[=VAL]
               Specify a single input file format option in the form
               of OPTION or OPTION=VALUE
  -O, --output-fmt FORMAT[,OPT[=VAL]]...
               Specify output format (SAM, BAM, CRAM)
      --output-fmt-option OPT[=VAL]
               Specify a single output file format option in the form
               of OPTION or OPTION=VALUE
      --reference FILE
               Reference sequence FASTA FILE [null]
  -@, --threads INT
               Number of additional threads to use [0]
      --write-index
               Automatically index the output files [off]
      --verbosity INT
               Set level of verbosity
(base) [koshu3@a001 scripts]$ apptainer exec /usr/local/biotools/s/samtools:1.22.1--h96c455f_0 samtools sort --help
sort: unrecognized option '--help'
Usage: samtools sort [options...] [in.bam]
Options:
```

SAM フォーマット/ BAM フォーマット

```
$ more SRR23499137.sam
@HD     VN:1.0  SO:unsorted
@SQ     SN:CM001522.1 LN:223219
@SQ     SN:CM001523.1 LN:806679
@SQ     SN:CM001524.1 LN:318651
@SQ     SN:CM001525.1 LN:1524084
@SQ     SN:CM001526.1 LN:568439
.
.
@PG     ID:hisat2       PN:hisat2       VN:2.2.1        CL:"/usr/local/bin/hisat2-align-s --wrapper basic-0 -p 4 -x ./reference/GCA_000269885.1_ASM26988v
SRR23499137.7 99    CM001536.1    1060273    60    1S149M        =    1060378    256    GCCCGCATGATTATTCACATATCATTTACAATAACA
SRR23499137.7 147   CM001536.1    1060378    60    150M          =    1060273    -256   AGATAATTCAGCGGTGCTAGAGGATGTAGCAGAG
SRR23499137.3 99    CM001525.1    824693     60    31M1I118M     =    824758     215    ACCATCGCATTATATTAGTTCAAACTAAACCTTTTT
```

@ヘッダ行

HD: ヘッダ行 SAMフォーマットのバージョンなど

SQ: リファレンスの情報

PG ツールの実行情報

各フィールドの説明

QNAME	リード名
FLG	アラインメント情報。参考 https://broadinstitute.github.io/picard/explain-flags.html
RNAME	マップされたリファレンス名
POS	マップポジション
MAPQ	マッピングスコア
CIGAR	マッピングの状況 ex) M アライメントマッチ I リファレンスにインサーションあり など
RNEXT	ペアエンドの場合、ペアのリード名（=: QNAME）。
PNEXT	ペアエンドの場合、ペアのマップされた開始位置。
TLEN	ペアエンドのリード間の距離。
SEQ	FASTQ の塩基配列データ
QUAL	FASTQ のクオリティデータ。

BAM は SAM をバイナリ形式にしたファイル

マッピング　実行

講習中はqsub を実行しないでください。

実行

```
$ sbatch scripts/hisat2.sh
Submitted batch job 10265964
```

ステータスの確認

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
      10265964_[0-1]      rome hisat2.s   koshu3 PD       0:00      1 (Resources)
```

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
      10265964_1      rome hisat2.s   koshu3  R       0:16      1 at145
      10265964_0      rome hisat2.s   koshu3  R       0:19      1 at142
```

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
```

マッピング　　実行結果の確認

講習中はqsub を実行しないでください。

```
$ ls -l
total 14324208
-rw-r--r-- 1 koshu3 koshu 7730914274 Nov 19 20:14 SRR23499137.sam
-rw-r--r-- 1 koshu3 koshu  616515715 Nov 19 20:15 SRR23499137.sorted.bam
-rw-r--r-- 1 koshu3 koshu 5858017344 Nov 19 20:13 SRR23499142.sam
-rw-r--r-- 1 koshu3 koshu  462471048 Nov 19 20:14 SRR23499142.sorted.bam
-rw-r--r-- 1 koshu3 koshu        685 Nov 19 20:14 hisat2.sh_10265964_0.err
-rw-r--r-- 1 koshu3 koshu          0 Nov 19 20:12 hisat2.sh_10265964_0.out
-rw-r--r-- 1 koshu3 koshu        683 Nov 19 20:14 hisat2.sh_10265964_1.err
-rw-r--r-- 1 koshu3 koshu          0 Nov 19 20:12 hisat2.sh_10265964_1.out
-rw-r--r-- 1 koshu3 koshu       2442 Nov 19 20:05 hisat2_index.sh_10265055.err
-rw-r--r-- 1 koshu3 koshu       4107 Nov 19 20:05 hisat2_index.sh_10265055.out
drwxr-xr-x 5 koshu3 koshu       4096 Nov 15 20:41 outputs
drwxr-xr-x 2 koshu3 koshu       4096 Nov 19 19:57 reads
drwxr-xr-x 2 koshu3 koshu       4096 Nov 19 20:05 reference
drwxr-xr-x 2 koshu3 koshu       4096 Nov 15 20:41 scripts
```

事前実行した結果で確認

$ ls -al outputs/hisat2/

マッピング　実行結果　ログの確認

```
$ more hisat2.sh_10265964_0.err
8698699 reads; of these:
  8698699 (100.00%) were paired; of these:
    667830 (7.68%) aligned concordantly 0 times
    8012664 (92.11%) aligned concordantly exactly 1 time
    18205 (0.21%) aligned concordantly >1 times
    ----
    667830 pairs aligned concordantly 0 times; of these:
      67194 (10.06%) aligned discordantly 1 time
    ----
    600636 pairs aligned 0 times concordantly or discordantly; of these:
      1201272 mates make up the pairs; of these:
        837762 (69.74%) aligned 0 times
        359403 (29.92%) aligned exactly 1 time
        4107 (0.34%) aligned >1 times
95.18% overall alignment rate
[bam_sort_core] merging from 2 files and 4 in-memory blocks...
```

事前実行した結果で確認

$ more outputs/hisat2/hisat2.sh_10265964_0.err

```
$ more hisat2.sh_10265964_1.err
6593231 reads; of these:
  6593231 (100.00%) were paired; of these:
    443901 (6.73%) aligned concordantly 0 times
    6138041 (93.10%) aligned concordantly exactly 1 time
    11289 (0.17%) aligned concordantly >1 times
    ----
    443901 pairs aligned concordantly 0 times; of these:
      41654 (9.38%) aligned discordantly 1 time
    ----
    402247 pairs aligned 0 times concordantly or discordantly; of these:
      804494 mates make up the pairs; of these:
        514431 (63.94%) aligned 0 times
        287516 (35.74%) aligned exactly 1 time
        2547 (0.32%) aligned >1 times
96.10% overall alignment rate
[bam_sort_core] merging from 1 files and 4 in-memory blocks...
```

事前実行した結果で確認

$ more outputs/hisat2/hisat2.sh_10265964_1.err

マッピング　実行結果　アライメントファイル (sam)の確認

```
$ more SRR23499137.sam
@HD   VN:1.0   SO:unsorted
@SQ   SN:CM001522.1  LN:223219
@SQ   SN:CM001523.1  LN:806679
@SQ   SN:CM001524.1  LN:318651
@SQ   SN:CM001525.1  LN:1524084
@SQ   SN:CM001526.1  LN:568439
.
.
@PG   ID:hisat2   PN:hisat2   VN:2.2.1CL:"/usr/local/bin/hisat2-align-s --wrapper basic-0 -p 4 -x ./refe
SRR23499137.7  99 CM001536.1  1060273601S149M=1060378256GCCCGCATGATTATTCACATATCATTTACAATAACATGACGGCAGCAA
SRR23499137.7  147   CM001536.1  106037860150M=1060273-256AGATAATTCAGCGGTGCTAGAGGATGTAGCAGAGGAAGAAGTTTCA
SRR23499137.3  99 CM001525.1  8246936031M1I118M=824758215ACCATCGCATTATATTAGTTCAAACTAAACCTTTTTTTTTCTTGCGG
```

事前実行した結果で確認

$ more outputs/hisat2/SRR23499137.sam

```
$ more SRR23499142.sam
@HD   VN:1.0   SO:unsorted
@SQ   SN:CM001522.1  LN:223219
@SQ   SN:CM001523.1  LN:806679
@SQ   SN:CM001524.1  LN:318651
@SQ   SN:CM001525.1  LN:1524084
@SQ   SN:CM001526.1  LN:568439
.
.
@PG   ID:hisat2   PN:hisat2   VN:2.2.1CL:"/usr/local/bin/hisat2-align-s --wrapper basic-0 -p 4 -x ./refer
SRR23499142.2  153   CM001533.1  2119260150M=211920TGTCAATATTAAAACGCGAATGCTTCGTTTCCGGCTGTTCAGGTGGAATATAA
SRR23499142.2  69 CM001533.1  21192 0*=211920NACCATCGTGAACCAAAGCGGTTCTCAAAACAACCTTCAAAGCATCTTCGATAGTAACTT
SRR23499142.7  137   CM001525.1  92781760150M=9278170TATCTTTAACTAATGACGACTTGAACCCTAATGTTAGAGACCCCATCGTTAA
```

事前実行した結果で確認

$ more outputs/hisat2/SRR453569.sam

発現量の算出　　スクリプトの確認

```
$ more scripts/stringtie.sh
#!/bin/bash
#SBATCH --partition=rome                # CPUノード
#SBATCH --output=%x_%A_%a.out           # 標準出力　%A=ArrayID %a=タスク番号
#SBATCH --error=%x_%A_%a.err            # 標準エラー
#SBATCH --ntasks=1                      # タスク数
#SBATCH --cpus-per-task=4               # 各タスクで使用するCPUスレッド数
#SBATCH --mem=16G                       # メモリ
#SBATCH --time=10:00:00                 # 最大実行時間
#

# IMX2816: SRR23499137
# wild: SRR23499142
ACESSIONS=(23499137 23499142)

for NUM in ${ACESSIONS[@]}
do
  PREFIX=SRR${NUM}
  BAM=${PREFIX}".sorted.bam"
  apptainer exec -B ${HOME} /usr/local/biotools/s/stringtie:3.0.1--h00789bb_0 stringtie \
     -e -B -p ${SLURM_CPUS_PER_TASK} \
     --rf \
     -G reference/GCA_000269885.1_ASM26988v1_genomic.gff \
     -o ballgown/$PREFIX/${PREFIX}.out.gtf \
     -A ${PREFIX}.gene_abund.tab $BAM
done
```

```
-e only estimate the abundance of given reference transcripts (requires -G)
-B enable output of Ballgown table files which will be created in the
    same directory as the output GTF (requires -G, -o recommended)
-p number of threads (CPUs) to use (default: 1)
--rf assume stranded library fr-firststrand
-G reference annotation to use for guiding the assembly process (GTF/GFF)
-o output path/file name for the assembled transcripts GTF (default: stdout)
-A gene abundance estimation output file
```

Stringtie オプションの確認

```
$ apptainer exec /usr/local/biotools/s/stringtie:3.0.1--h00789bb_0 stringtie -h

StringTie v3.0.1 usage:

stringtie <in.bam ..> [-G <guide_gff>] [-l <prefix>] [-o <out.gtf>] [-p <cpus>]
 [-v] [-a <min_anchor_len>] [-m <min_len>] [-j <min_anchor_cov>] [-f <min_iso>]
 [-c <min_bundle_cov>] [-g <bdist>] [-u] [-L] [-e] [--viral] [-E <err_margin>]
 [--ptf <f_tab>] [-x <seqid,..>] [-A <gene_abund.out>] [-h] {-B|-b <dir_path>}
 [--mix] [--conservative] [--rf] [--fr]
Assemble RNA-Seq alignments into potential transcripts.
Options:
 --version : print just the version at stdout and exit
 --conservative : conservative transcript assembly, same as -t -c 1.5 -f 0.05
 --mix : both short and long read data alignments are provided
        (long read alignments must be the 2nd BAM/CRAM input file)
 --rf : assume stranded library fr-firststrand
 --fr : assume stranded library fr-secondstrand
 -G reference annotation to use for guiding the assembly process (GTF/GFF)
 --ptf : load point-features from a given 4 column feature file <f_tab>
 -o output path/file name for the assembled transcripts GTF (default: stdout)
 -l name prefix for output transcripts (default: STRG)
 -f minimum isoform fraction (default: 0.01)
 -L long reads processing; also enforces -s 1.5 -g 0 (default:false)
 -R if long reads are provided, just clean and collapse the reads but
    do not assemble
 -m minimum assembled transcript length (default: 200)
 -a minimum anchor length for junctions (default: 10)
 -j minimum junction coverage (default: 1)
 -t disable trimming of predicted transcripts based on coverage
    (default: coverage trimming is enabled)
 -c minimum reads per bp coverage to consider for multi-exon transcript
    (default: 1)
 -s minimum reads per bp coverage to consider for single-exon transcript
    (default: 4.75)
 -v verbose (log bundle processing details)
 -g maximum gap allowed between read mappings (default: 50)
 -M fraction of bundle allowed to be covered by multi-hit reads (default:1)
```

発現量の算出

講習中はqsub を実行しないでください。

実行

```
$ sbatch scripts/stringtie.sh
Submitted batch job 10269484
```

ステータスの確認

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          10269484      rome stringti   koshu3 PD       0:00      1 (Resources)
```

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          10269484      rome stringti   koshu3  R       0:01      1 at142
```

```
$ squeue -u koshu3
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
```

発現量の算出

事前実行した結果で確認
$ ls outputs/stringtie

```
$ ls -l
total 14325076
-rw-r--r-- 1 koshu3 koshu     440307 Nov 19 20:39 SRR23499137.gene_abund.tab
-rw-r--r-- 1 koshu3 koshu 7730914274 Nov 19 20:14 SRR23499137.sam
-rw-r--r-- 1 koshu3 koshu  616515715 Nov 19 20:15 SRR23499137.sorted.bam
-rw-r--r-- 1 koshu3 koshu     439637 Nov 19 20:40 SRR23499142.gene_abund.tab
-rw-r--r-- 1 koshu3 koshu 5858017344 Nov 19 20:13 SRR23499142.sam
-rw-r--r-- 1 koshu3 koshu  462471048 Nov 19 20:14 SRR23499142.sorted.bam
drwxr-xr-x 4 koshu3 koshu       4096 Nov 19 20:39 ballgown
-rw-r--r-- 1 koshu3 koshu        685 Nov 19 20:14 hisat2.sh_10265964_0.err
-rw-r--r-- 1 koshu3 koshu          0 Nov 19 20:12 hisat2.sh_10265964_0.out
-rw-r--r-- 1 koshu3 koshu        683 Nov 19 20:14 hisat2.sh_10265964_1.err
-rw-r--r-- 1 koshu3 koshu          0 Nov 19 20:12 hisat2.sh_10265964_1.out
-rw-r--r-- 1 koshu3 koshu       2442 Nov 19 20:05 hisat2_index.sh_10265055.err
-rw-r--r-- 1 koshu3 koshu       4107 Nov 19 20:05 hisat2_index.sh_10265055.out
drwxr-xr-x 5 koshu3 koshu       4096 Nov 15 20:41 outputs
drwxr-xr-x 2 koshu3 koshu       4096 Nov 19 19:57 reads
drwxr-xr-x 2 koshu3 koshu       4096 Nov 19 20:05 reference
drwxr-xr-x 2 koshu3 koshu       4096 Nov 15 20:41 scripts
-rw-r--r-- 1 koshu3 koshu          0 Nov 19 20:39 stringtie.sh_10269484_4294967294.err
-rw-r--r-- 1 koshu3 koshu          0 Nov 19 20:39 stringtie.sh_10269484_4294967294.out
```

事前実行した結果で確認
$ ls outputs/stringtie/ballgown

```
$ ls ballgown/*
ballgown/SRR23499137:
SRR23499137.out.gtf  e2t.ctab  e_data.ctab  i2t.ctab  i_data.ctab  t_data.ctab

ballgown/SRR23499142:
SRR23499142.out.gtf  e2t.ctab  e_data.ctab  i2t.ctab  i_data.ctab  t_data.ctab
```

次のステップとして、R などを用いることで可視化などができる。

結果ファイルの確認　SRR23499137.gene_abund.tab ファイル

```
$ more SRR23499137.gene_abund.tab
```

Gene ID	Gene Name	Reference	Strand	Start	End	Coverage	FPKM	TPM
gene-CENPK1137D_4927	-	CM001522.1	+	31625	32998	123.764192	58.407735	67.673571
gene-CENPK1137D_4938	-	CM001522.1	+	33506	34759	149.295056	70.456454	81.633705
gene-CENPK1137D_4949	-	CM001522.1	+	35211	36359	690.818103	326.016113	377.735493
gene-CENPK1137D_4960	-	CM001522.1	+	36565	37203	137.860720	65.060275	75.381474
gene-CENPK1137D_4971	-	CM001522.1	+	37520	39028	90.736912	42.821251	49.614438
gene-CENPK1137D_4982	-	CM001522.1	+	39315	41957	56.524782	26.675603	30.907436
gene-CENPK1137D_4993	-	CM001522.1	+	42233	42775	95.276243	44.963486	52.096519
gene-CENPK1137D_4997	-	CM001522.1	-	42935	45076	1584.073763	747.568091	866.162715
gene-CENPK1137D_4962	-	CM001522.1	-	113772	114773	125.382236	59.171335	68.558309
gene-CENPK1137D_4963	-	CM001522.1	+	115077	118472	29.298587	13.826811	16.020304

発現量のノーマライズ

FPKM: Fragments Per Kilobase of exon per Million reads mapped
TPM: Transcripts Per kilobase Milion

FPKMもTPMも以下の二つで補正するが、補正する順番が異なる。

(1) 総リード数での補正（総リード数 100万）

(2) 遺伝子長での補正（遺伝子長 1000b）

FPKM (1) -> (2)
TPM (2) -> (1)

GTF format とは

```
# /usr/local/bin/stringtie -e -B -p 4 --rf -G reference/GCA_000269885.1_ASM26988v1_genomic.gff -o ballgown/SRR23499137/SRR234
# StringTie version 3.0.1
CM001522.1  StringTie  transcript  31625  32998  1000  +  .  gene_id "gene-CENPK1137D_4927"; transcript_id "rna-mrna.CE
CM001522.1  StringTie  exon        31625  32998  1000  +  .  gene_id "gene-CENPK1137D_4927"; transcript_id "rna-mrna.CE
CM001522.1  StringTie  transcript  33506  34759  1000  +  .  gene_id "gene-CENPK1137D_4938"; transcript_id "rna-mrna.CE
CM001522.1  StringTie  exon        33506  34759  1000  +  .  gene_id "gene-CENPK1137D_4938"; transcript_id "rna-mrna.CE
CM001522.1  StringTie  transcript  35211  36359  1000  +  .  gene_id "gene-CENPK1137D_4949"; transcript_id "rna-mrna.CE
CM001522.1  StringTie  exon        35211  36359  1000  +  .  gene_id "gene-CENPK1137D_4949"; transcript_id "rna-mrna.CE
```

タブ区切りフォーマット。値がない場合は、”.” が設定される。

1. seqname：染色体 or スキャフォールドの名前
2. source：アノテーションを生成したプログラムまたはデータソースの名前
3. feature：フィーチャータイプ（mRNA, gene, exon, CDS ….）
4. start：スタートポジション (1bp ～)
5. end：エンドポジション (1bp ～)
6. score：スコア
7. strand：+(forward)、-(reverse)または '.'
8. frame：翻訳フレーム（0, 1, 2）
9. attribute：追加情報。セミコロンで区切られたタグと値のペアのリスト。

GFF では、gene_id=XXXXX; の形式に対して、GTF では、gene_id “XXXXXX”; と記載していく。

Thank you for your attention !