

バッチジョブ、RNA-seq の 各種ツールによる解析

国立遺伝学研究所 大量遺伝情報研究室
望月孝子

この講義では、本講習会で使用する RNA-seq 解析データを作成します。

1. RNA-seq について
2. 本講習会で使用するデータの作成の解説
3. データダウンロード

遺伝研スパコンがメンテナンス中のため、スライドでの講習になります m(_ _)m

1. RNA-seq について

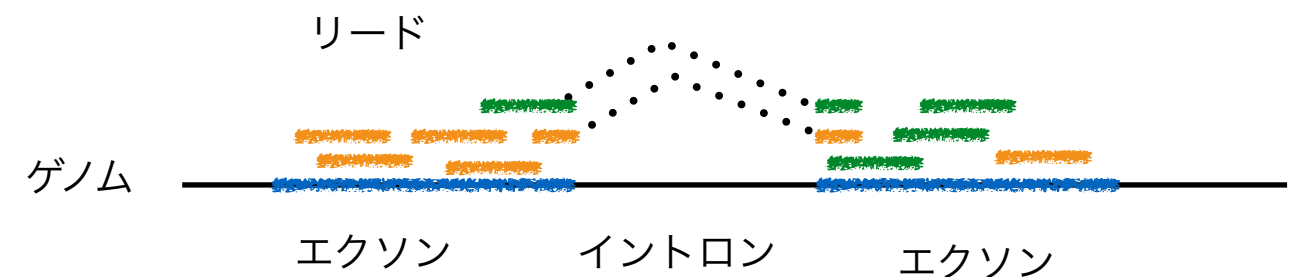
RNA-seq とは

細胞中の mRNA や miRNA の配列をシーケンスして、発現量の定量や新規転写産物の同定を行う手法。

配列のシーケンス



発現量の定量



<https://bi.biopapyrus.jp/rnaseq/>

解析の流れ

1. リードのトリミング
2. エクソン・イントロン構造をもつ場合は、HISAT2 などのソフトウェアでゲノムマッピングを行う。
3. サンプル毎の総リード数の違いや、遺伝子配列長の違いを補正するため、正規化を行う。
4. 遺伝子毎の発現量を同定、比較する。

Sequence Read Archive (SRA)

Illumina HiSeqシリーズ, PacBio RS II/Sequel,
Oxford Nanopore MinION などの出力データをアーカイブ

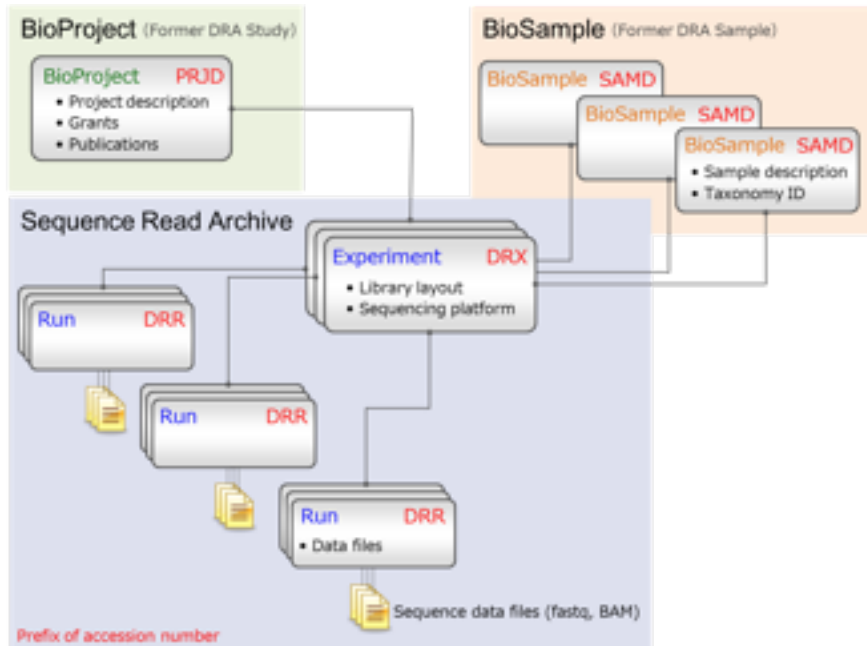


DDBJ SRA は、International Nucleotide Sequence Database Collaboration (INSDC) のメンバーとして、 [NCBI Sequence Read Archive \(SRA\)](#) と [EBI Sequence Read Archive \(ERA\)](#) との国際協力のもと、運営されている。

ご自分の研究で利用できるデータがあるかどうか、実験をする前に SRA を確認することをおすすめします。

SRA: メタデータと配列データ

メタデータ



<https://www.ddbj.nig.ac.jp/dra/submission.html>

BioProject 研究プロジェクト全体の概要。

BioSample 生物学的なサンプルに関する記述。

Experiment BioSample に由来するシーケンス用ライブラリーとシーケンスの手法について記載

Run シーケンス用ライブラリー (Experiment) に由来するデータファイル (SRA/fastq ファイル)

Submission 登録するオブジェクトをとりまとめるオブジェクト。

配列データ

The screenshot shows the DRA Search website. At the top, there's a navigation bar with "DRA Search" and links to "Search Home" and "DRA Home". Below this, the accession number "SRR453566" is displayed along with icons for "FASTQ" and "SRA".

The main content area is divided into two sections. On the left, the "Run Detail" table provides information about the sequencing run:

Run Detail	
Alias	Batch1.1
Instrument model	
Date of run	
Run center	
Number of spots	5,725,730
Number of bases	1,156,597,460

Below the table, there's a section for "READS (joined)" with a "quality" dropdown, a "show" button, and a "rows" dropdown set to "10". It also displays the current page "1" out of "572573" pages.

On the right, the "Navigation" section lists links to different data types:

- Submission: [SRA051410](#) (FTP icon)
- Study: [SRP012047](#)
- Experiment: [SRX135198](#) (FASTQ icon) and [SRA](#) (SRA icon)
- Sample: [SRS307298](#)

A red box highlights the "FASTQ" and "SRA" links under the "Experiment" section.

Below the navigation section, there's a text box containing the following text:

配列データは SRA or fastq フォーマットで提供されています。SRA Toolkits の fastq-dump をインストールして、SRA から配列データをダウンロードする必要があります。(https://ftp-ncbi.sra.nlm.nih.gov/ftp-1/)

配列データは SRA or fastq フォーマットで提供される。SRA の場合は、SRA Toolkit の fastq-dump を使用し fastq に変換する必要がある。(https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/)

DDBJ SRA は、現在、NCBI/EBI の SRA ファイルのftp ミラーリングを停止しています。

DDBJ Services

Login & SubmitContactEnglish

NEWS

DDBJ Center Web SitesGoogle カスタム検索

ホーム検索・解析登録ダウンロードスパコン統計活動講習会センターについて

HOME > News Archive > SRA ファイルが無い NCBI/EBI SRA データのダウンロード

SRA ファイルが無い NCBI/EBI SRA データのダウンロード

作成日: 2018年9月18日

ディスク容量逼迫のため DDBJ Sequence Read Archive (DRA) では2017年4月7日から NCBI/EBI SRA の SRA ファイルのftp ミラーリングを停止しております。DRASearchでのメタデータのミラーリングとインデックス、及び、DRA 自極分のデータ公開は継続しています。

2017年4月7日以降に NCBI/EBI SRA から公開されたデータはメタデータに対応する SRA ファイルの ftp ダウンロードが利用できない場合があります（下記の例を参照）。このようなデータについてはお手数ですが NCBI SRA もしくは EBI SRA (ENA) からのダウンロードをお願いいたします。

- SRX4203001 at DRA
- SRX4203001 at NCBI SRA
- SRX4203001 at EBI SRA

ミラーリングの再開時期については未定です。

SRA ファイル（と SRA ファイルから生成される fastq ファイル）が無い例

DRASearchSearch HomeDRA Home

SRX4203001FASTQSRA

Experiment Detail	
Title	GSM3188536: THX-treated rep1; Homo sapiens; RNA-Seq
Design Description	
Organism	Homo sapiens
Library Description	
Name	
Strategy	RNA-Seq
Source	TRANSCRIPTOMIC
Selection	cDNA
Layout	PAIRED
Orientation	
Nominal Length	
Nominal Sdev	

Navigation	
Submission	SRA721334FTP
Study	SRP150418
Sample	SRS3412724
Run	SRR7300567FASTQSRA

サービス: DRA, DDBJ Center

キーワード: お知らせ

Footer: Policies and Disclaimers | News | FAQs | Sitemap | Calendar | Address | Contact | Last modified: 2018-09-18

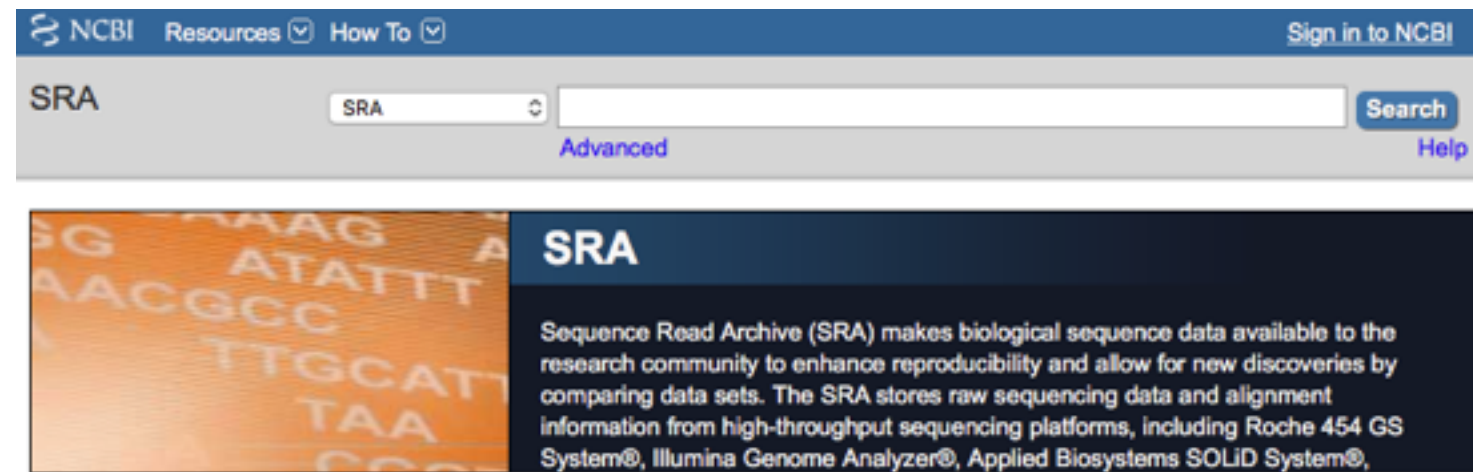
DDBJ SRA にはないデータは、NCBI SRA もしくは EBI SRA からダウンロードしてください。

SRA のデータ検索

DDBJ Search <http://sra.dbcls.jp/index.html>

NCBI SRA

<https://www.ncbi.nlm.nih.gov/sra>



ENA

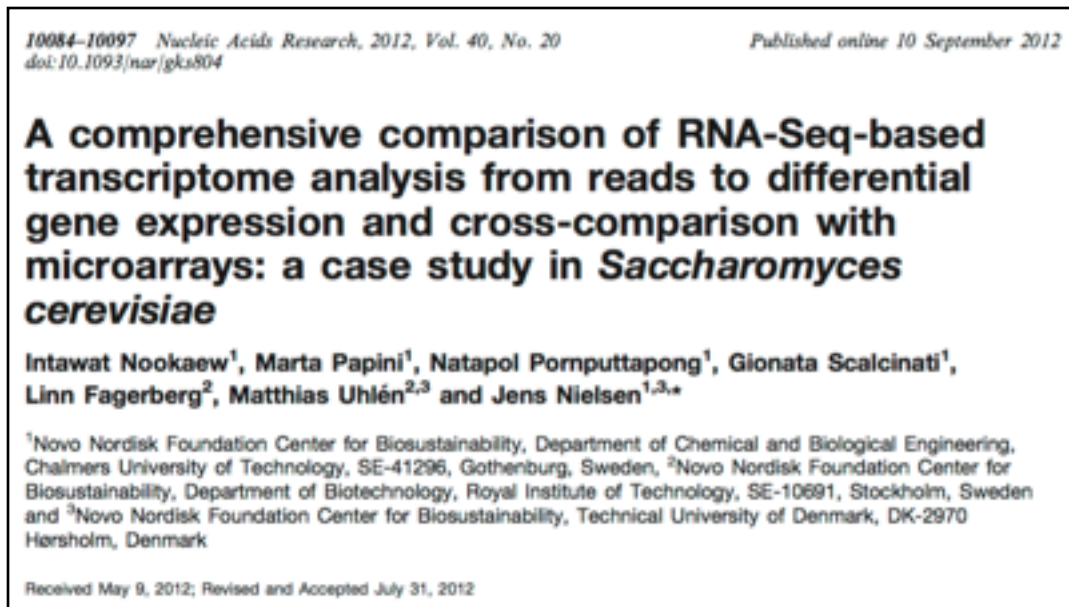
<https://www.ebi.ac.uk/ena/data/warehouse/search>



2. 本講習会で使用するデータの作成

本講習で使うデータ

リード RNA-Seq



Saccharomyces cerevisiae strain CEN.PK 113-7D,
grown under two different conditions (batch and chemostat)

SRA accession

Batch culture: **SRX135198** (three biological triplicate)

SRR453566

SRR453567

SRR453568

chemostat: **SRX135710** (three biological triplicate)

SRR453569

SRR453570

SRR453571

リファレンス

Saccharomyces cerevisiae S288C

RefSeq assembly accession: GCF_000146045.2

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Chr	I	NC_001133.9	BK006935.2	0.23	39.3	94	-	4	2	101	1
	Chr	II	NC_001134.8	BK006936.2	0.81	38.3	415	-	13	4	432	-
	Chr	III	NC_001135.5	BK006937.2	0.32	38.5	168	-	10	4	184	2
	Chr	IV	NC_001136.10	BK006938.2	1.53	37.9	766	-	28	4	799	1
	Chr	V	NC_001137.3	BK006939.2	0.58	38.5	287	-	20	9	317	1
	Chr	VI	NC_001138.5	BK006940.2	0.27	38.7	128	-	10	4	143	1
	Chr	VII	NC_001139.9	BK006941.2	1.09	38.1	539	-	36	10	585	-
	Chr	VIII	NC_001140.6	BK006934.2	0.56	38.5	290	-	11	4	305	-
	Chr	IX	NC_001141.2	BK006942.2	0.44	38.9	213	-	10	3	232	6
	Chr	X	NC_001142.9	BK006943.2	0.75	38.4	362	-	24	6	392	-
	Chr	XI	NC_001143.9	BK006944.2	0.67	38.1	317	-	16	5	338	-
	Chr	XII	NC_001144.5	BK006945.2	1.08	38.5	519	12	21	18	572	2
	Chr	XIII	NC_001145.3	BK006946.2	0.92	38.2	469	-	21	15	505	-
	Chr	XIV	NC_001146.8	BK006947.3	0.78	38.6	398	-	14	6	418	-
	Chr	XV	NC_001147.6	BK006948.2	1.09	38.2	546	-	20	11	579	2
	Chr	XVI	NC_001148.4	BK006949.2	0.95	38.1	472	-	17	6	497	2
		MT	NC_001224.1	KP263414.1	0.09	17.1	19	2	24	1	46	-

Genome size: 12Mb

解析の手順

1. リードとリファレンスの準備

fastq-dump ver. 2.8.2

(<https://github.com/ncbi/sra-tools>)

2. リードクオリティチェック

FastQC ver. 0.11.8

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

3. リードの前処理 (リードトリミング、アダプター配列の除去)

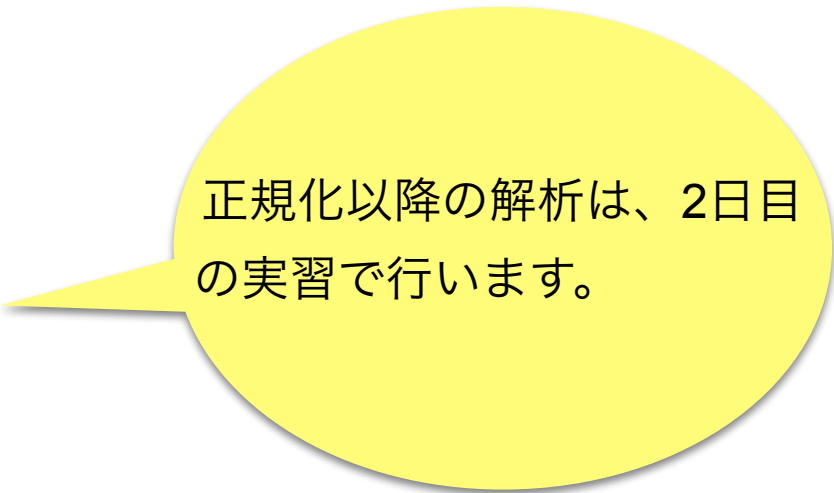
Trimmomatic ver. 0.38 (Bolger *et al.*, 2014)

4. リードをリファレンスゲノムにマッピング

HISAT2 ver. 2.1.0 (Kim *et al.*, 2015)

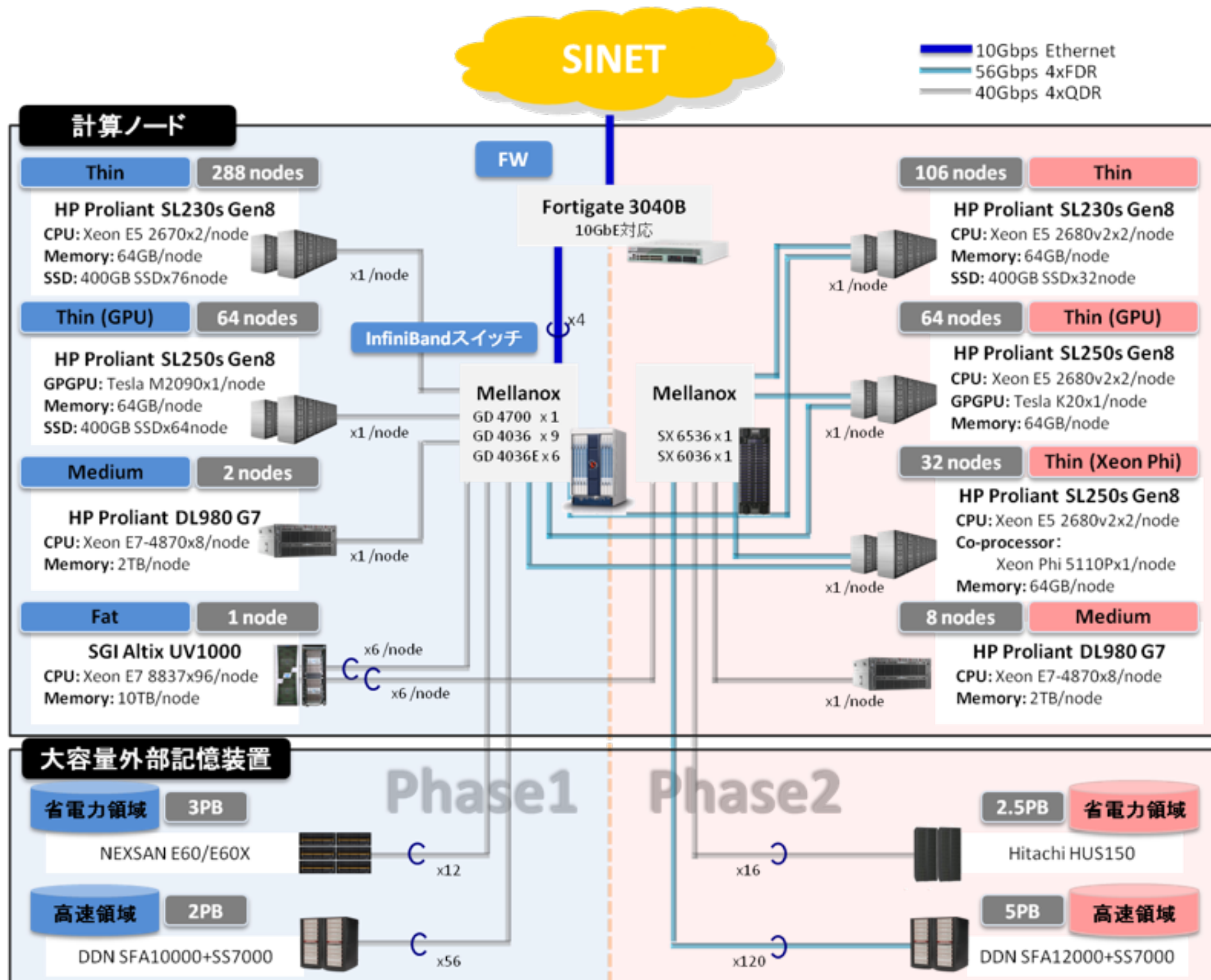
5. 遺伝子毎にリードカウント

featureCounts ver. 1.6.2 (Liao *et al.*, 2014)



正規化以降の解析は、2日目の実習で行います。

遺伝研スパコン システム構成図



遺伝研スパコンでのジョブ投入

コマンドライン \$ qsub test.sh

シェルスクリプト test.sh のヘッダ

```
#$ -S /bin/bash
#$ -pe def_slot 1
#$ -cwd
#$ -l mem_req=4G,s_vmem=4G
```

-S 使用するインタプリタのパス

-pe def_slot 1 ジョブスロット数

-cwd ホームディレクトリではなく、qsubコマンド実行時のディレクトリでジョブを実行。
標準出力 / 標準エラー出力ファイルは、qsubコマンド実行時のディレクトリに出力。

-l 主にキューの選択、メモリ利用上限の変更に使う

s_vmem: ジョブが使用可能な仮想メモリの上限値。(OS に対する宣言)

mem_req: 使用するメモリの量を宣言する。(ジョブ管理システムUGEのジョブリソース管理に対する宣言)

キューの指定: thin ノードの month_hdd.q キューに投入する場合は、キューの指定は不要。

解析の手順

1. リードとリファレンスの準備

fastq-dump ver. 2.8.2

(<https://github.com/ncbi/sra-tools>)

2. リードクオリティチェック

FastQC ver. 0.11.8

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

3. リードの前処理 (リードトリミング、アダプター配列の除去)

Trimmomatic ver. 0.38 (Bolger *et al.*, 2014)

4. リードをリファレンスゲノムにマッピング

HISAT2 ver. 2.1.0 (Kim *et al.*, 2015)

5. 遺伝子毎にリードカウント

featureCounts ver. 1.6.2 (Liao *et al.*, 2014)

1. リードとリファレンスの準備

DDBJ SRA リードファイルのftpアドレスの調べ方

DRASearch Search Home DRA Home

Accession :

Organism : StudyType :

CenterName : Platform :

Keyword :

Show 20 records Sort by Study Search Clear

Data Last Update 2018-10-20

DRASearch にて
アクセッション番号を入力

DRASearch Search Home DRA Home

SRX135198 FASTQ SRA

Experiment Detail	
Title	Batch rep1
Design Description	
Organism	Saccharomyces cerevisiae

Library Description	
Name	PolyA
Strategy	RNA-Seq
Source	TRANSCRIPTOMIC
Selection	unspecified
Layout	PAIRED
Orientation	
Nominal Length	

Navigation	
Submission	SRA051410 ETP
Study	SRP012047
Sample	SRS307298
Run	SRR453566 FASTQ SRA
	SRR453567 FASTQ SRA
	SRR453568 FASTQ SRA

← → ↻ 保護されていない通信 ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX135/SRX135198/SRR453566

/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX135/SRX135198/SRR453566 のインデックス

[親ディレクトリ] ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX135/SRX135198/SRR453566/SRR453566.sra

名前	サイズ	更新日
<input type="checkbox"/> SRR453566.sra	710 MB	2012/08/12 9:00:00

DRASearch にて
アクセッション番号を入力

1. リードとリファレンスの準備

ダウンロード用シェルスクリプト

```
#$ -S /bin/bash
#$ -pe def_slot 1
#$ -cwd
#$ -l mem_req=4G,s_vmem=4G

# リードの取得
mkdir read
cd read

# Batch culture: SRX135198 SRR453566 - SRR453568
ACCESSIONS=`seq 453566 453568`
for NUM in $ACCESSIONS
do
    echo Retrieving SRA file for SRR${NUM}...
    wget ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX135/SRX135198/SRR${NUM}/SRR${NUM}.sra
    echo Converting SRA to FASTQ...
    fastq-dump --split-files SRR${NUM}.sra
    gzip SRR${NUM}_1.fastq
    gzip SRR${NUM}_2.fastq
done

# chemostat: SRX135710 SRR453569 - SRR453571
ACCESSIONS=`seq 453569 453571`
for NUM in $ACCESSIONS
do
    echo Retrieving SRA file for SRR${NUM}...
    wget ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX135/SRX135710/SRR${NUM}/SRR${NUM}.sra
    echo Converting SRA to FASTA...
    fastq-dump --split-files SRR${NUM}.sra
    gzip SRR${NUM}_1.fastq
    gzip SRR${NUM}_2.fastq
done

cd ../

# リファレンスゲノム取得
# Saccharomyces cerevisiae S288C (baker's yeast)
# RefSeqのデータを利用
# https://www.ncbi.nlm.nih.gov/assembly/GCF_000146045.2

mkdir reference
cd reference

# Genomic FASTAファイル
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.fna.gz
# GFFファイル
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.gff.gz
gunzip *.gz

mv GCF_000146045.2_R64_genomic.gff s288c.gff
mv GCF_000146045.2_R64_genomic.fna s288c.fna
```

fastq-dump 遺伝研スパコンに既にインストールされているものを使用している。
—split-files オプション : read1 , read2 でファイルを分ける。

実行前のディレクトリ

```
[xxxxxxx@gw2 20181119]$ ls -al
合計 20
drwxr-xr-x 2 yanakamu yn-nig 12288 10月 29 13:47 2018 ./
drwxr-xr-x 9 yanakamu yn-nig  4096 10月 29 12:21 2018 ../
-rw-r--r-- 1 yanakamu yn-nig  1480 10月 29 13:46 2018 download.sh
```

qsub の実行

```
[yanakamu@gw2 20181119]$ qsub download.sh
Your job 11285156 ("download.sh") has been submitted
```

qstat ジョブのステータス確認

```
[xxxxxxx@gw2 20181119]$ qstat
```

job-ID	prior	name	user	state	submit/start at	queue	jclass	slots	ja-task-ID
11284239	0.25050	QLOGIN	yanakamu	r	10/29/2018 10:57:17	login.q@nt097i		1	
11285156	0.25000	download.s	yanakamu	r	10/29/2018 13:47:50	month_phi.q@nt196i		1	

実行中



ジョブ終了

```
[xxxxxxx@nt097 20181119]$ qstat
```

job-ID	prior	name	user	state	submit/start at	queue	jclass	slots	ja-task-ID
11284239	0.25044	QLOGIN	yanakamu	r	10/29/2018 10:57:17	login.q@nt097i		1	

ジョブが終了すると該当ジョブIDが表示されなくなる。

1. リードとリファレンスの準備

実行結果

実行後のディレクトリ

```
[xxxx@nt097 20181119]$ ls -al
合計 6892
drwxr-xr-x 4 ynakamu yn-nig 12288 10月 29 14:32 2018 ./
drwxr-xr-x 9 ynakamu yn-nig 4096 10月 29 12:21 2018 ../
-rw-r--r-- 1 ynakamu yn-nig 1480 10月 29 13:46 2018 download.sh
-rw-r--r-- 1 ynakamu yn-nig 7016694 10月 29 14:33 2018 download.sh.e11285156
-rw-r--r-- 1 ynakamu yn-nig 846 10月 29 14:26 2018 download.sh.o11285156
-rw-r--r-- 1 ynakamu yn-nig 0 10月 29 13:47 2018 download.sh.pe11285156
-rw-r--r-- 1 ynakamu yn-nig 0 10月 29 13:47 2018 download.sh.po11285156
drwxr-xr-x 2 ynakamu yn-nig 4096 10月 29 14:32 2018 read/
drwxr-xr-x 2 ynakamu yn-nig 4096 10月 29 14:33 2018 reference/
```

結果ファイルが格納されて
いるディレクトリ

標準出力ログ

```
[xxxx@nt097 20181119]$ more download.sh.o11285156
Retrieving SRA file for SRR453566...
Converting SRA to FASTA...
Read 5725730 spots for SRR453566.sra
Written 5725730 spots for SRR453566.sra
Retrieving SRA file for SRR453567...
Converting SRA to FASTA...
Read 7615732 spots for SRR453567.sra
Written 7615732 spots for SRR453567.sra
Retrieving SRA file for SRR453568...
Converting SRA to FASTA...
Read 5565734 spots for SRR453568.sra
Written 5565734 spots for SRR453568.sra
Retrieving SRA file for SRR453569...
Converting SRA to FASTA...
Read 4032514 spots for SRR453569.sra
Written 4032514 spots for SRR453569.sra
Retrieving SRA file for SRR453570...
Converting SRA to FASTA...
Read 6745975 spots for SRR453570.sra
Written 6745975 spots for SRR453570.sra
Retrieving SRA file for SRR453571...
Converting SRA to FASTA...
Read 6163396 spots for SRR453571.sra
Written 6163396 spots for SRR453571.sra
```

標準エラーログ

```
[xxxx@nt097 20181119]$ more download.sh.e11285156
--2018-10-29 13:47:51-- ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX135/SRX135198/SRR453566/SRR453566.sra
=> `SRR453566.sra'
ftp.ddbj.nig.ac.jp をDNSに問いあわせています... 133.39.224.12
ftp.ddbj.nig.ac.jp|133.39.224.12|:21 に接続しています... 接続しました。
anonymous としてログインしています... ログインしました!
==> SYST ... 完了しました。 ==> PWD ... 完了しました。
==> TYPE I ... 完了しました。 ==> CWD (1) /ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX135/SRX135198/SRR453566 ... 完了しました。
==> SIZE SRR453566.sra ... 744076393
==> PASV ... 完了しました。 ==> RETR SRR453566.sra ... 完了しました。
長さ: 744076393 (710M) (確証はありません)

  0K ..... 0% 36.5M 19s
 50K ..... 0% 67.1M 15s
100K ..... 0% 68.9M 13s
150K ..... 0% 68.0M 13s
      :
```

ダウンロードの結果: リードファイル

.sra ファイル (バイナリ) から paired-end の FASTQ ファイルが作成されている。

ペアードエンドリード数	
SRR453566	5,725,730
SRR453567	7,615,732
SRR453568	5,565,734
SRR453569	4,032,514
SRR453570	6,745,975
SRR453571	6,163,396

read1

read2

```
[yanakamu@nt097 20181119]$ zcat read/SRR453566_2.fastq.gz | more  
@SRR453566.1 HWI-ST167:4:1101:1597:1986 length=101  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAAGCCTCCGTTTNNNNNNNNNNN  
+SRR453566.1 HWI-ST167:4:1101:1597:1986 length=101  
#####  
@SRR453566.2 HWI-ST167:4:1101:2535:1992 length=101  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGCCAATCGTGGTNANNANNNNNN  
+SRR453566.2 HWI-ST167:4:1101:2535:1992 length=101  
#####
```

FASTQ フォーマット

4行で1配列の情報を表す。

[illegible]

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されている。Sanger 形式の場合は、アスキー値 - 33 がクオリティ値になる。



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)

<- SRA

1. リードとリファレンスの準備

reference fasta と gff ファイルが作成されている。

reference fasta

```
[yanakamu@nt097 reference]$ more s288c.fna
>NC_001133.9 Saccharomyces cerevisiae S288C chromosome I, complete sequence
ccacaccacacccacaccccacacaccacaccacacaccacacccacacacacacatCCTAACACTACCCTAAC
ACAGCCCTAATCTAACCCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCAT
TCAACCATAACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATC
CAACCCACTGCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATACTGTTCTTCTACCCACCATAT
TGAAACGCTAACAAATGATCGTAAATAACACACACGTGCTTACCCTACCACTTTATACCACCACCACATGCCATACTCAC
```

GFF ファイル

```
[yanakamu@nt097 reference]$ more s288c.gff
##gff-version 3
##gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build R64
#!genome-build-accession NCBI_Assembly:GCF_000146045.2
#!annotation-source SGD R64-2-1
##sequence-region NC_001133.9 1 230218
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=559292
NC_001133.9 RefSeq region 1 230218 . + . ID=id0;Dbxref=taxon:559292;Name=I;chromosome=I;gbkey=Src;genome=chromosome;mol_type=genomic DNA;
NC_001133.9 RefSeq telomere 1 801 . - . ID=id1;Dbxref=SGD:S000028862;Note=TEL01L%3B Telomeric region on the left arm of Chromosome I%3B com
NC_001133.9 RefSeq origin_of_replication 707 776 . + . ID=id2;Dbxref=SGD:S000121252;Note=ARS102~Autonomously Replicating Sequence;gbkey=rep_
NC_001133.9 RefSeq gene 18072169 . - . ID=gene0;Dbxref=GeneID:851229;Name=PAU8;end_range=2169,.;gbkey=Gene;gene=PAU8;gene_biotype=protein_coding
NC_001133.9 RefSeq mRNA 18072169 . - . ID=rna0;Parent=gene0;Dbxref=GeneID:851229,Genbank:NM_001180043.1;Name=NM_001180043.1;end_range=2169,.;gbkey=
NC_001133.9 RefSeq exon 18072169 . - . ID=id3;Parent=rna0;Dbxref=GeneID:851229,Genbank:NM_001180043.1;end_range=2169,.;gbkey=mRNA;gene=PAU8;par
NC_001133.9 RefSeq CDS 18072169 . - 0 ID=cds0;Parent=rna0;Dbxref=SGD:S000002142,Genbank:NP_009332.1;Name=NP_009332.1;Note=hypothetical
```

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 R64
##genome-build-accession NCBI_Assembly:GCF_000146045.2
##annotation-source SGD R64-2-1
##sequence-region NC_001133.9 1 230218
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=559292
NC_001133.9 RefSeq gene 1807 2169 . - . ID=gene0;Dbxref=GeneID:851229;Name=PAU8;end_range=2169,.;gbkey=Gene;gene=PAU8;gene_biotype=p
NC_001133.9 RefSeq mRNA 1807 2169 . - . ID=rna0;Parent=gene0;Dbxref=GeneID:851229,Genbank:NM_001180043.1;Name=NM_001180043.1;end_ra
NC_001133.9 RefSeq exon 1807 2169 . - . ID=id3;Parent=rna0;Dbxref=GeneID:851229,Genbank:NM_001180043.1;end_range=2169,.;gbkey=mRNA;ge
NC_001133.9 RefSeq CDS 1807 2169 . - 0 ID=cds0;Parent=rna0;Dbxref=SGD:S000002142,GeneID:851229,Genbank:NP_009332.1;Name=NP_009332.
NC_001133.9 RefSeq gene 2480 2707 . + . ID=gene1;Dbxref=GeneID:1466426;Name=YAL067W-A;end_range=2707,.;gbkey=Gene;gene_biotype=prote
NC_001133.9 RefSeq mRNA 2480 2707 . + . ID=rna1;Parent=gene1;Dbxref=GeneID:1466426,Genbank:NM_001184582.1;Name=NM_001184582.1;end_r
NC_001133.9 RefSeq exon 2480 2707 . + . ID=id4;Parent=rna1;Dbxref=GeneID:1466426,Genbank:NM_001184582.1;end_range=2707,.;gbkey=mRNA;p
NC_001133.9 RefSeq CDS 2480 2707 . + 0 ID=cds1;Parent=rna1;Dbxref=SGD:S000028593,GeneID:1466426,Genbank:NP_878038.1;Name=NP_878038
NC_001133.9 RefSeq gene 7235 9016 . - . ID=gene2;Dbxref=GeneID:851230;Name=SEO1;end_range=9016,.;gbkey=Gene;gene=SEO1;gene_biotype=p
NC_001133.9 RefSeq mRNA 7235 9016 . - . ID=rna2;Parent=gene2;Dbxref=GeneID:851230,Genbank:NM_001178208.1;Name=NM_001178208.1;end_ra
NC_001133.9 RefSeq exon 7235 9016 . - . ID=id5;Parent=rna2;Dbxref=GeneID:851230,Genbank:NM_001178208.1;end_range=9016,.;gbkey=mRNA;ge
NC_001133.9 RefSeq CDS 7235 9016 . - 0 ID=cds2;Parent=rna2;Dbxref=SGD:S000000062,GeneID:851230,Genbank:NP_009333.1;Name=NP_009333.
NC_001133.9 RefSeq origin_of_replication 7997 8547 . + . ID=id6;Dbxref=SGD:S000121253;Note=ARS103~Autonomously Replicating Sequence%3E
```

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。

解析の手順

1. リードとリファレンスの準備

fastq-dump ver. 2.8.2

(<https://github.com/ncbi/sra-tools>)

2. リードクオリティチェック

FastQC ver. 0.11.8

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

3. リードの前処理 (リードトリミング、アダプター配列の除去)

Trimmomatic ver. 0.38 (Bolger *et al.*, 2014)

4. リードをリファレンスゲノムにマッピング

HISAT2 ver. 2.1.0 (Kim *et al.*, 2015)

5. 遺伝子毎にリードカウント

featureCounts ver. 1.6.2 (Liao *et al.*, 2014)

2. リードクオリティチェック

シェルスクリプト

```
#$ -S /bin/bash
#$ -pe def_slot 1
#$ -cwd
#$ -l mem_req=4G,s_vmem=4G

export PATH=/usr/local/pkg/FastQC/v0.11.8:$PATH

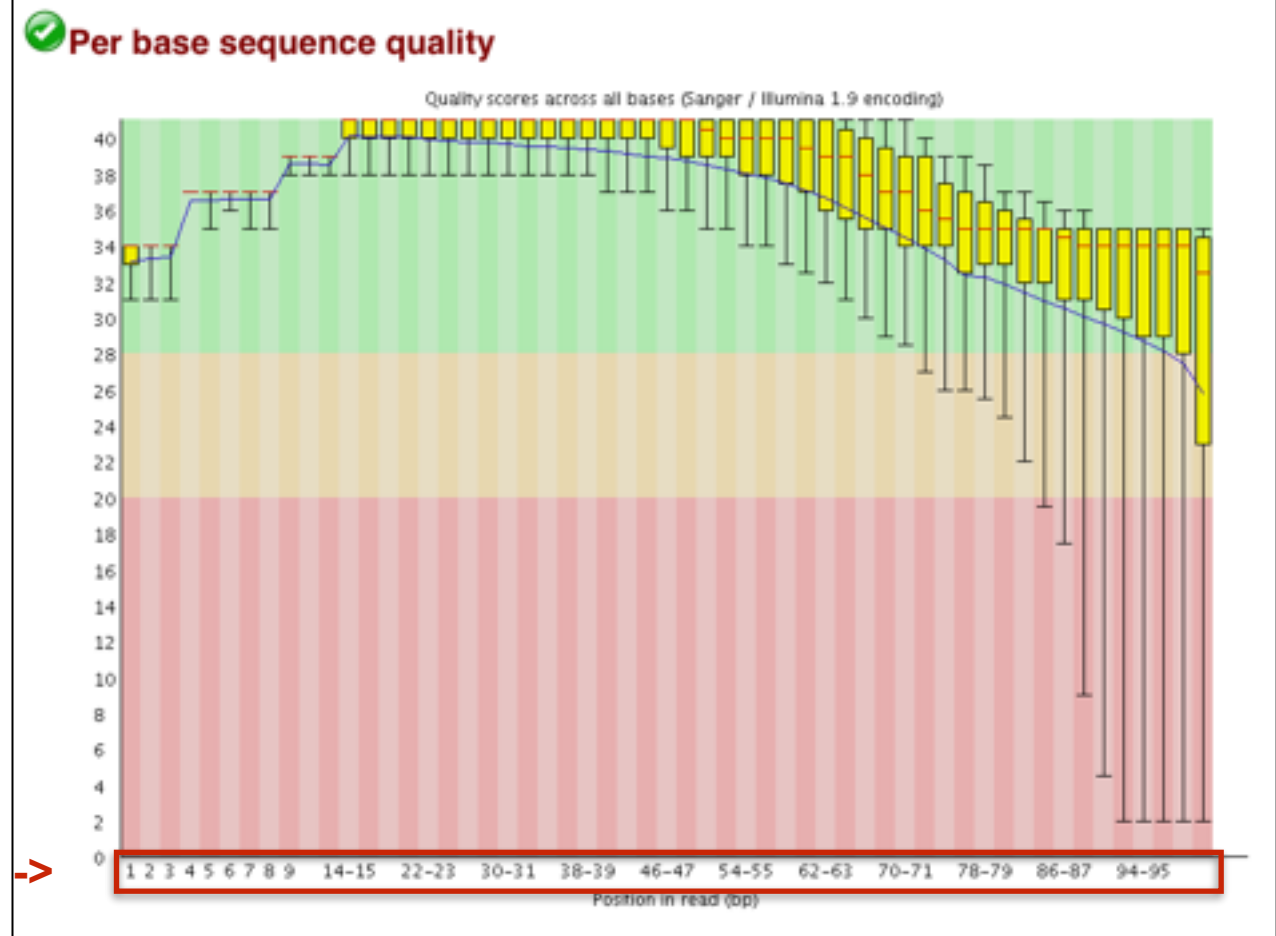
cd read
mkdir FastQC

# Batch culture: SRX135198   SRR453566 - SRR453568
# chemostat: SRX135710 SRR453569 - SRR453571
ACCESSIONS=`seq 453566 453571`
for NUM in $ACCESSIONS
do
    echo FastQC for SRR${NUM}
    fastqc --nogroup -o ./FastQC SRR${NUM}_1.fastq.gz
    fastqc --nogroup -o ./FastQC SRR${NUM}_2.fastq.gzzw
done
```

—nogroup オプションをつけなかった例

—nogroup 結果レポートのグラフ中のリード
ポジションをグループ化せずに表示

リードポジションがグループ化されてしまう。->



2. リードクオリティチェック

実行ログの確認

実行

```
[yanakamu@nt097 20181119]$ qsub fastqc.sh  
Your job 11285829 ("fastqc.sh") has been submitted
```

実行後のディレクトリ

```
[yanakamu@nt097 20181119]$ ls -al  
合計 6900  
drwxr-xr-x 4 yanakamu yn-nig 4096 10月 25 16:51 2018 ./  
drwxr-xr-x 8 yanakamu yn-nig 4096 10月 23 16:55 2018 ../  
-rw-r--r-- 1 yanakamu yn-nig 1480 10月 24 17:35 2018 download.sh  
-rw-r--r-- 1 yanakamu yn-nig 7014855 10月 24 18:21 2018 download.sh.e11278764  
-rw-r--r-- 1 yanakamu yn-nig 846 10月 24 18:14 2018 download.sh.o11278764  
-rw-r--r-- 1 yanakamu yn-nig 0 10月 24 17:36 2018 download.sh.pe11278764  
-rw-r--r-- 1 yanakamu yn-nig 0 10月 24 17:36 2018 download.sh.po11278764  
-rw-r--r-- 1 yanakamu yn-nig 450 10月 25 16:29 2018 fastqc.sh  
-rw-r--r-- 1 yanakamu yn-nig 10740 10月 25 17:01 2018 fastqc.sh.e11279593  
-rw-r--r-- 1 yanakamu yn-nig 642 10月 25 17:01 2018 fastqc.sh.o11279593  
-rw-r--r-- 1 yanakamu yn-nig 0 10月 25 16:51 2018 fastqc.sh.pe11279593  
-rw-r--r-- 1 yanakamu yn-nig 0 10月 25 16:51 2018 fastqc.sh.po11279593  
drwxr-xr-x 3 yanakamu yn-nig 4096 10月 25 16:51 2018 read/  
drwxr-xr-x 2 yanakamu yn-nig 4096 10月 24 18:21 2018 reference/
```

標準出力ログ

```
[yanakamu@nt097 20181119]$ more fastqc.sh.o11279593  
FastQC for SRR453566  
Analysis complete for SRR453566_1.fastq.gz  
Analysis complete for SRR453566_2.fastq.gz  
FastQC for SRR453567  
Analysis complete for SRR453567_1.fastq.gz  
Analysis complete for SRR453567_2.fastq.gz  
FastQC for SRR453568  
Analysis complete for SRR453568_1.fastq.gz  
Analysis complete for SRR453568_2.fastq.gz  
FastQC for SRR453569  
Analysis complete for SRR453569_1.fastq.gz  
Analysis complete for SRR453569_2.fastq.gz  
FastQC for SRR453570  
Analysis complete for SRR453570_1.fastq.gz  
Analysis complete for SRR453570_2.fastq.gz  
FastQC for SRR453571  
Analysis complete for SRR453571_1.fastq.gz  
Analysis complete for SRR453571_2.fastq.gz
```

標準エラーログ

```
yanakamu@nt097 20181119]$ more fastqc.sh.e11279593  
Started analysis of SRR453566_1.fastq.gz  
Approx 5% complete for SRR453566_1.fastq.gz  
Approx 10% complete for SRR453566_1.fastq.gz  
Approx 15% complete for SRR453566_1.fastq.gz  
Approx 20% complete for SRR453566_1.fastq.gz  
Approx 25% complete for SRR453566_1.fastq.gz  
Approx 30% complete for SRR453566_1.fastq.gz  
Approx 35% complete for SRR453566_1.fastq.gz  
Approx 40% complete for SRR453566_1.fastq.gz  
Approx 45% complete for SRR453566_1.fastq.gz  
Approx 50% complete for SRR453566_1.fastq.gz  
Approx 55% complete for SRR453566_1.fastq.gz  
Approx 60% complete for SRR453566_1.fastq.gz  
Approx 65% complete for SRR453566_1.fastq.gz  
Approx 70% complete for SRR453566_1.fastq.gz  
Approx 75% complete for SRR453566_1.fastq.gz  
Approx 80% complete for SRR453566_1.fastq.gz  
Approx 85% complete for SRR453566_1.fastq.gz  
Approx 90% complete for SRR453566_1.fastq.gz  
Approx 95% complete for SRR453566_1.fastq.gz  
:  
:  
:
```

2. リードクオリティチェック

FastQC 結果ファイルの確認

```
[yanakamu@nt097 20181119]$ cd read/
[yanakamu@nt097 read]$ ls -al
合計 10809996
drwxr-xr-x 3 yanakamu yn-nig      4096 10月 25 16:51 2018 ./
drwxr-xr-x 4 yanakamu yn-nig      4096 10月 25 16:51 2018 ../
drwxr-xr-x 2 yanakamu yn-nig      4096 10月 25 17:02 2018 FastQC/
-rw-r--r-- 1 yanakamu yn-nig 744076393 10月 24 17:37 2018 SRR453566.sra
-rw-r--r-- 1 yanakamu yn-nig 495142914 10月 24 17:37 2018 SRR453566_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 525639378 10月 24 17:37 2018 SRR453566_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 995177005 10月 24 17:44 2018 SRR453567.sra
-rw-r--r-- 1 yanakamu yn-nig 669625362 10月 24 17:45 2018 SRR453567_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 695379655 10月 24 17:45 2018 SRR453567_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 721719428 10月 24 17:53 2018 SRR453568.sra
-rw-r--r-- 1 yanakamu yn-nig 478611138 10月 24 17:54 2018 SRR453568_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 511972231 10月 24 17:54 2018 SRR453568_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 525479824 10月 24 18:00 2018 SRR453569.sra
-rw-r--r-- 1 yanakamu yn-nig 352115327 10月 24 18:01 2018 SRR453569_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 370353972 10月 24 18:01 2018 SRR453569_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 882122317 10月 24 18:05 2018 SRR453570.sra
-rw-r--r-- 1 yanakamu yn-nig 585686313 10月 24 18:06 2018 SRR453570_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 611002557 10月 24 18:06 2018 SRR453570_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 802201727 10月 24 18:14 2018 SRR453571.sra
-rw-r--r-- 1 yanakamu yn-nig 537670628 10月 24 18:14 2018 SRR453571_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 565354212 10月 24 18:14 2018 SRR453571_2.fastq.gz
```

FastQC ディレクトリが作成され、その中に
FastQC の結果ファイルが格納されている。

```
[yanakamu@nt097 read]$ ls -al FastQC/
合計 7664
drwxr-xr-x 2 yanakamu yn-nig      4096 10月 25 17:02 2018 ./
drwxr-xr-x 3 yanakamu yn-nig      4096 10月 25 16:51 2018 ../
-rw-r--r-- 1 yanakamu yn-nig 298624 10月 25 16:52 2018 SRR453566_1_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 354707 10月 25 16:52 2018 SRR453566_1_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 302049 10月 25 16:53 2018 SRR453566_2_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 356749 10月 25 16:53 2018 SRR453566_2_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 296474 10月 25 16:54 2018 SRR453567_1_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 352015 10月 25 16:54 2018 SRR453567_1_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 299062 10月 25 16:55 2018 SRR453567_2_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 355650 10月 25 16:55 2018 SRR453567_2_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 294692 10月 25 16:56 2018 SRR453568_1_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 349833 10月 25 16:56 2018 SRR453568_1_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 300201 10月 25 16:56 2018 SRR453568_2_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 354390 10月 25 16:56 2018 SRR453568_2_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 296229 10月 25 16:57 2018 SRR453569_1_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 350171 10月 25 16:57 2018 SRR453569_1_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 298800 10月 25 16:58 2018 SRR453569_2_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 352780 10月 25 16:58 2018 SRR453569_2_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 295745 10月 25 16:59 2018 SRR453570_1_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 350771 10月 25 16:59 2018 SRR453570_1_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 296147 10月 25 17:00 2018 SRR453570_2_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 352601 10月 25 17:00 2018 SRR453570_2_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 293719 10月 25 17:01 2018 SRR453571_1_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 343962 10月 25 17:01 2018 SRR453571_1_fastqc.zip
-rw-r--r-- 1 yanakamu yn-nig 296416 10月 25 17:02 2018 SRR453571_2_fastqc.html
-rw-r--r-- 1 yanakamu yn-nig 348753 10月 25 17:02 2018 SRR453571_2_fastqc.zip
```

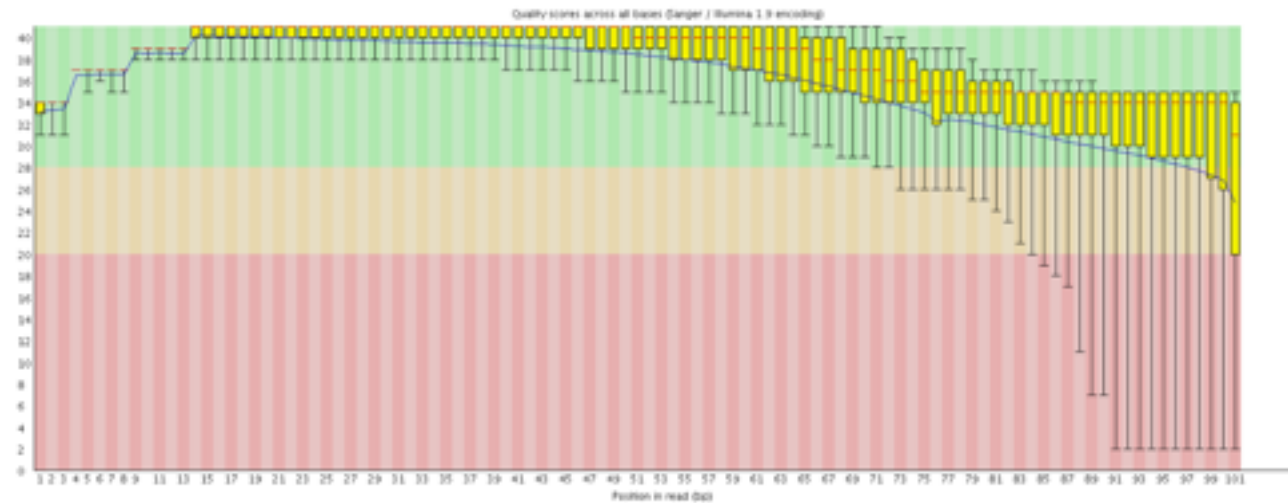
2. リードクオリティチェック

FastQC レポートの確認 (1)

全体を通して、リード2の方がクオリティが低く、アダプターが残っているものもある。

SRR453566_1

✔ Per base sequence quality

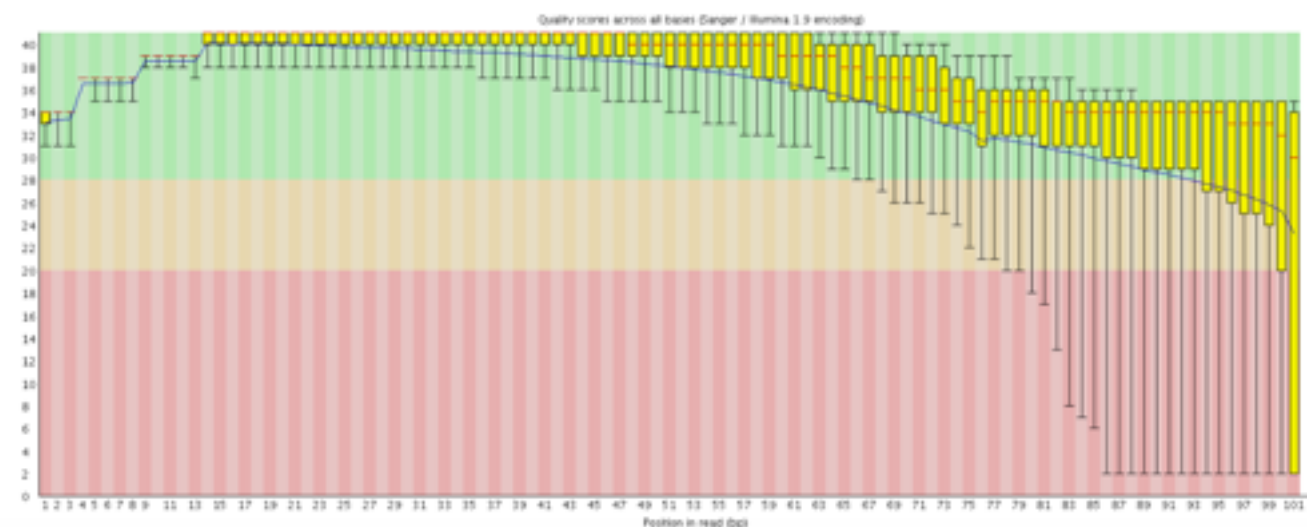


⚠ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	13810	0.24119195281649677	TruSeq Adapter, Index 2 (100% over 50bp)

SRR453567_1

✖ Per base sequence quality

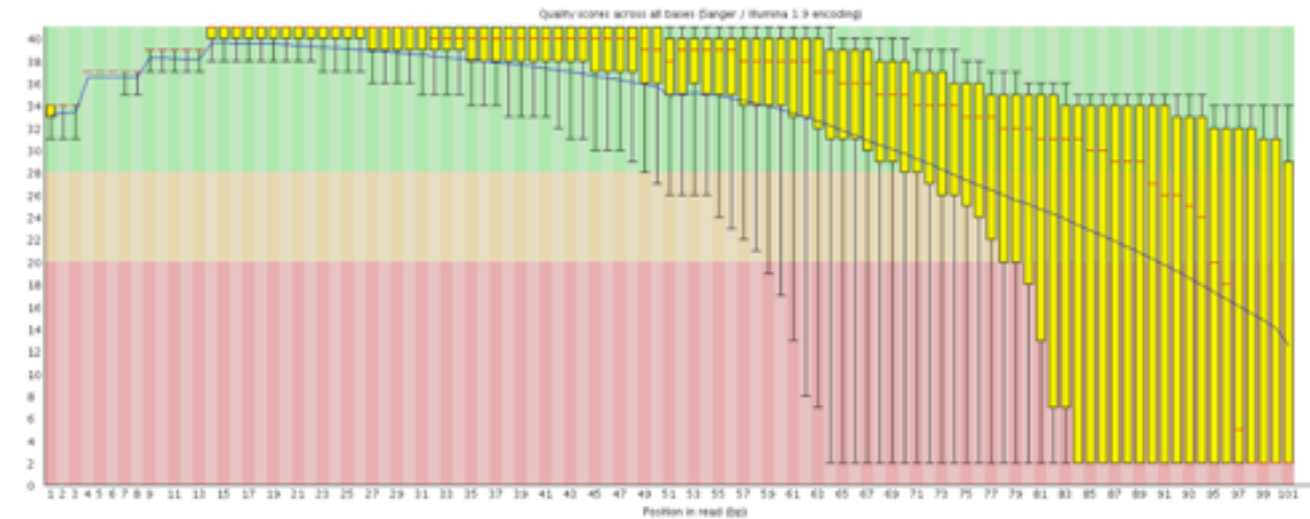


⚠ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTGACCAATCTCGTATGC	13337	0.17512433473236716	TruSeq Adapter, Index 4 (100% over 50bp)

SRR453566_2

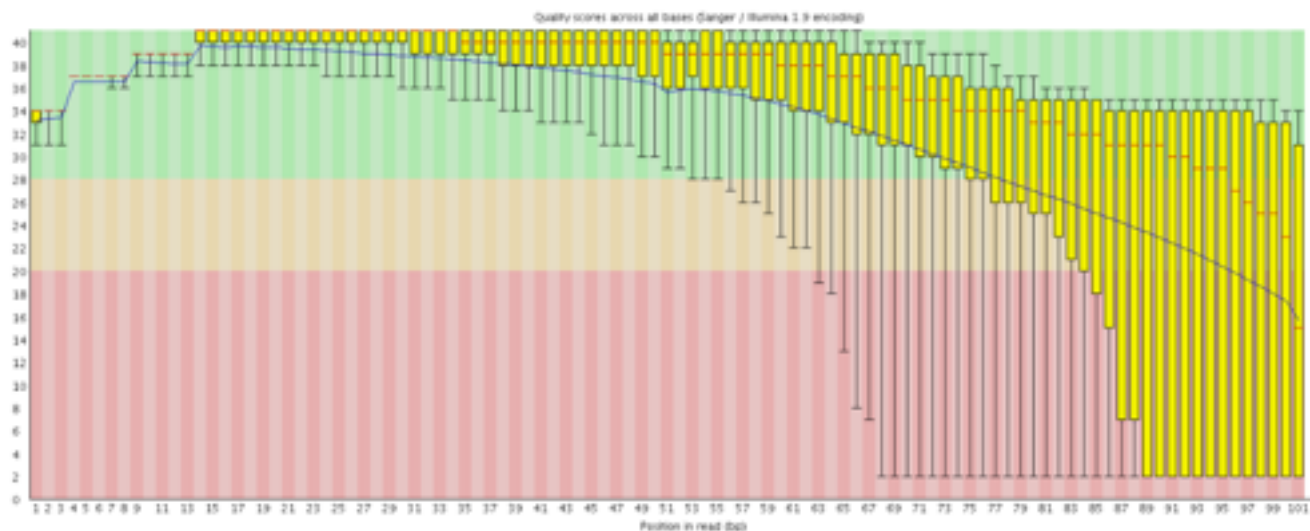
✖ Per base sequence quality



✔ Overrepresented sequences
No overrepresented sequences

SRR453567_2

✖ Per base sequence quality



⚠ Overrepresented sequences

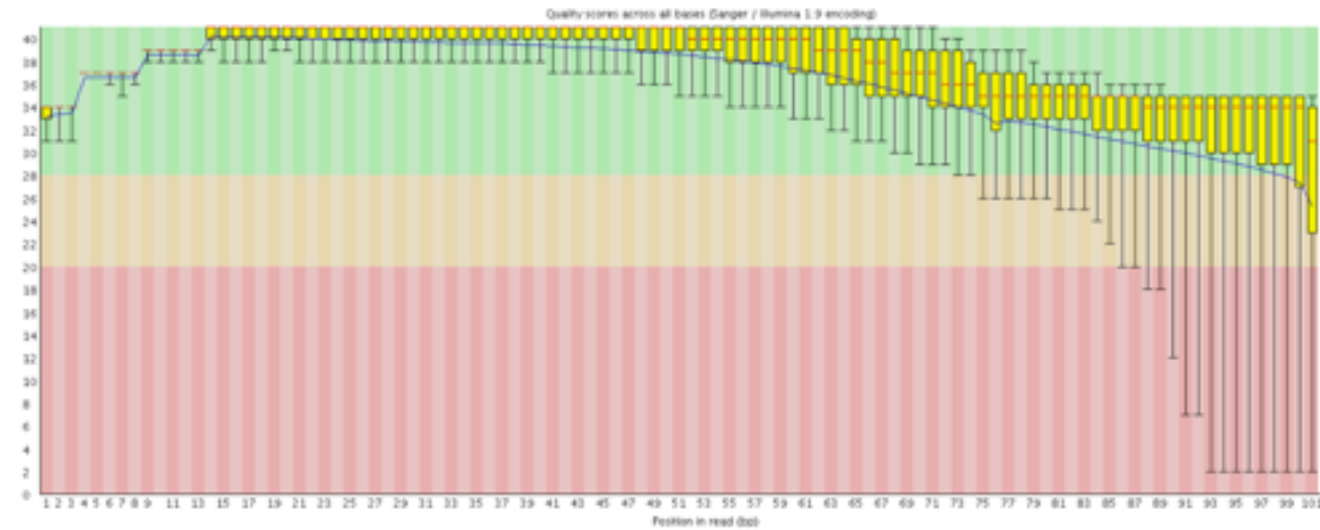
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCG	9153	0.12018542669306115	Illumina Single End PCR Primer 1 (100% over 50bp)

2. リードクオリティチェック

FastQC レポートの確認 (2)

SRR453568_1

✔ Per base sequence quality

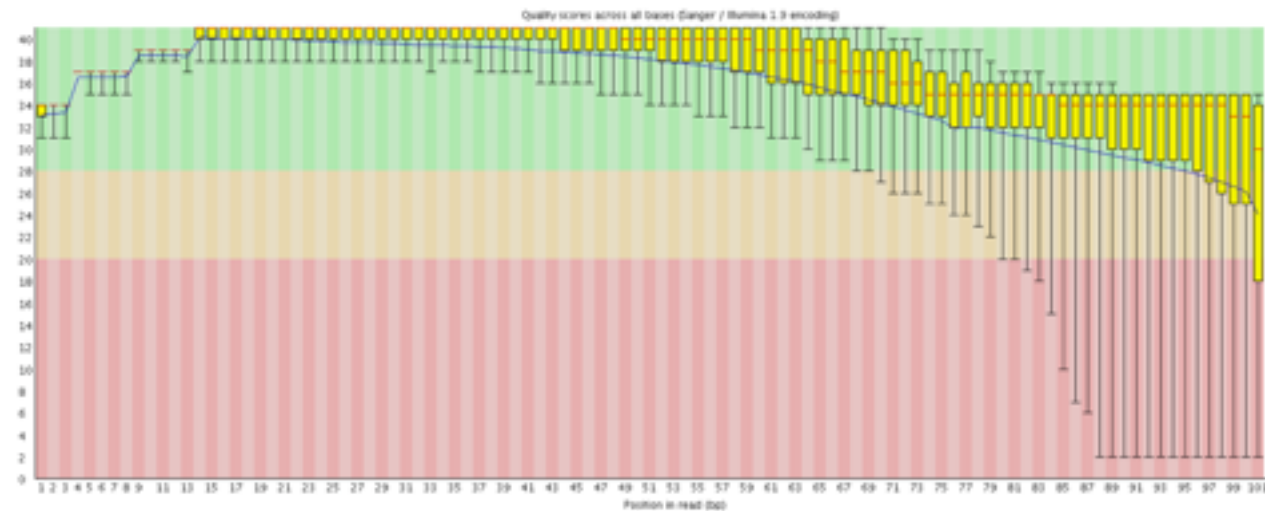


⚠ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC	10856	0.19505064381445467	TruSeq Adapter, Index 5 (100% over 50bp)

SRR453569_1

✔ Per base sequence quality

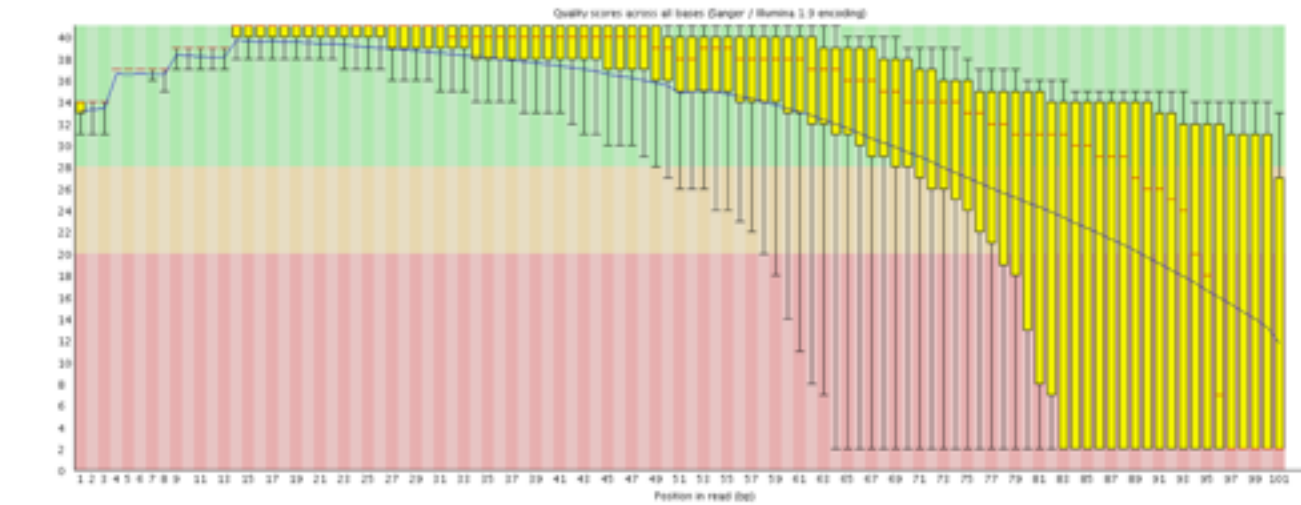


⚠ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCAGATCTCGTATGC	5312	0.1317292388817497	TruSeq Adapter, Index 1 (100% over 50bp)

SRR453568_2

✖ Per base sequence quality

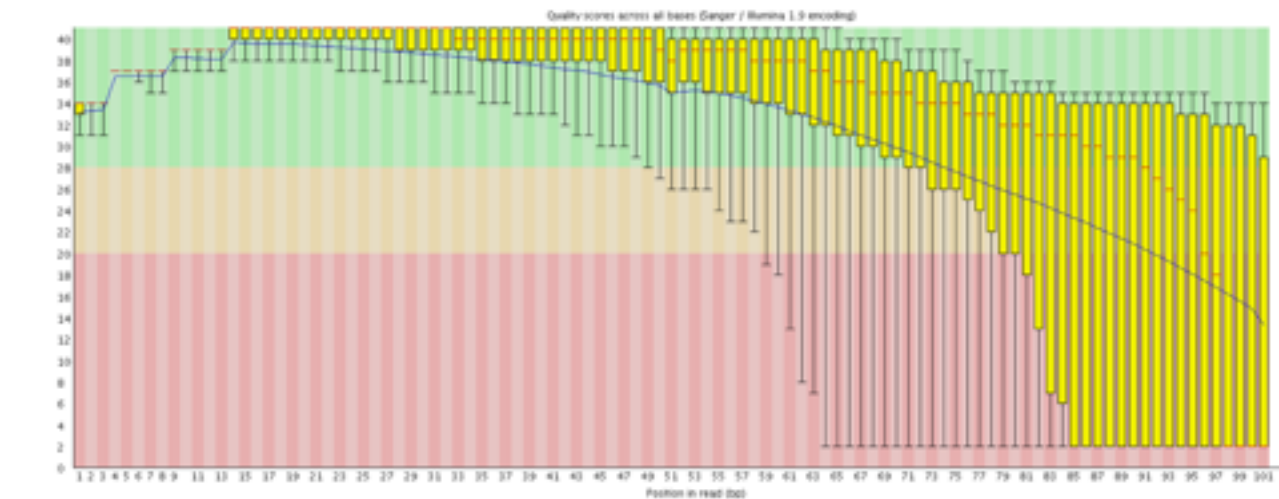


✔ Overrepresented sequences

No overrepresented sequences

SRR453569_2

✖ Per base sequence quality



✔ Overrepresented sequences

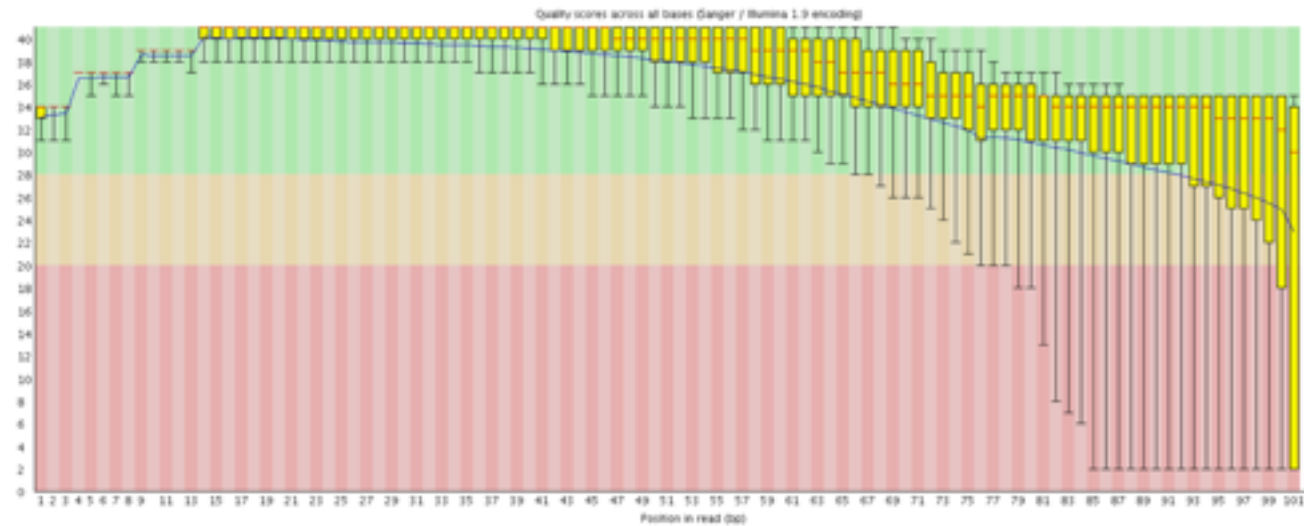
No overrepresented sequences

2. リードクオリティチェック

FastQC レポートの確認 (3)

SRR453570_1

✖ Per base sequence quality

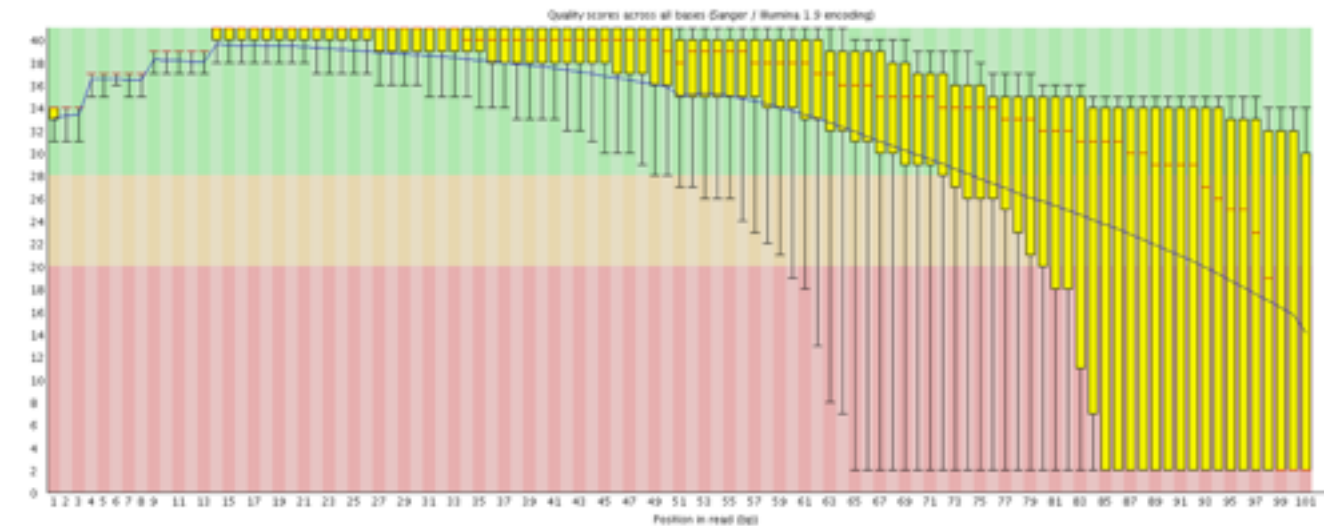


⚠ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAGAGACACACGCTGAACTCCAGTCACTTAGGCATCTCGTATGC	21115	0.3130014564240158	TruSeq Adapter, Index 3 (100% over 50bp)

SRR453570_2

✖ Per base sequence quality

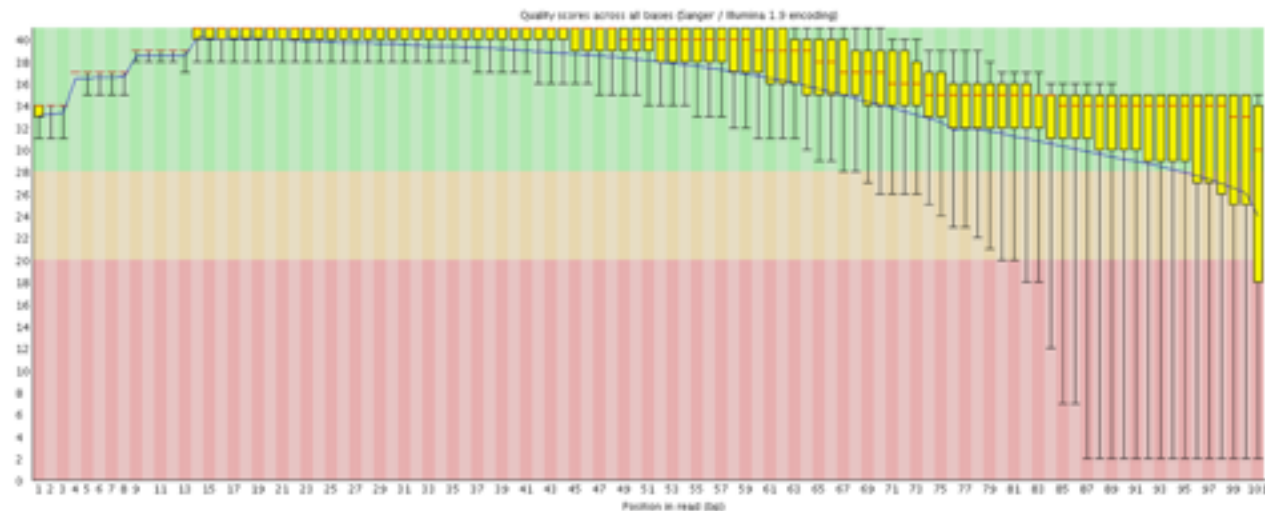


⚠ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAGAGACGCTCGTGTAGGGAAGAGTGTAGATCTCGGTGTCGCCG	11924	0.1767572515462924	Illumina Single End PCR Primer 1 (100% over 50bp)

SRR453571_1

✔ Per base sequence quality

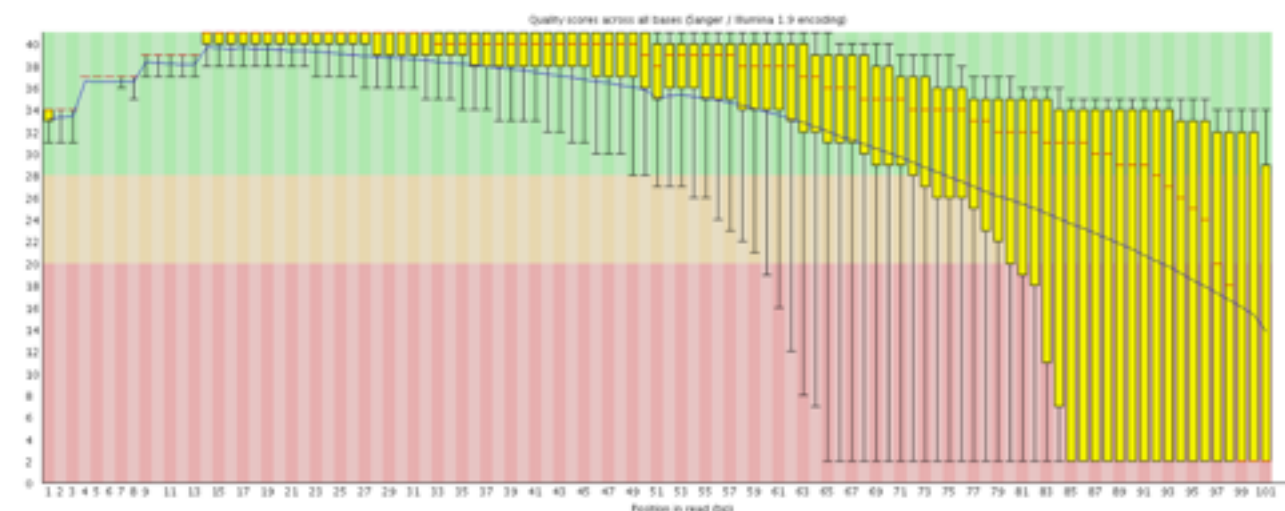


✔ Overrepresented sequences

No overrepresented sequences

SRR453571_2

✖ Per base sequence quality



✔ Overrepresented sequences

No overrepresented sequences

解析の手順

1. リードとリファレンスの準備

fastq-dump ver. 2.8.2

(<https://github.com/ncbi/sra-tools>)

2. リードクオリティチェック

FastQC ver. 0.11.8

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

3. リードの前処理 (リードトリミング、アダプター配列の除去)

Trimmomatic ver. 0.38 (Bolger *et al.*, 2014)

4. リードをリファレンスゲノムにマッピング

HISAT2 ver. 2.1.0 (Kim *et al.*, 2015)

5. 遺伝子毎にリードカウント

featureCounts ver. 1.6.2 (Liao *et al.*, 2014)

3. リードの前処理

リードトリミングとアダプター配列の除去

```
#$ -S /bin/bash
#$ -pe def_slot 4
#$ -cwd
#$ -t 1-6:1
#$ -l mem_req=8G,s_vmem=8G
#$ -l short
```

アレイジョブで実行

各 SRR アクセション毎に ジョブを並列実行させる。

```
# Batch culture: SRX135198 SRR453566 - SRR453568
# chemostat: SRX135710 SRR453569 - SRR453571
ACCESSIONS=(453566 453567 453568 453569 453570 453571)
no=`expr ${SGE_TASK_ID} - 1`
```

アレイジョブのジョブ番号 (ここでは1 - 6) が格納される。

```
NUM=${ACCESSIONS[${no}]}
PREFIX=SRR${NUM}
```

<- 実行する SRR アクセションを指定

```
export PATH=/usr/local/pkg/Trimmomatic/0.38:$PATH
export PATH=/usr/local/pkg/FastQC/v0.11.8:$PATH
```

```
cd read
```

```
java -jar -Xmx512m trimmomatic-0.38.jar \
PE \
-threads ${NSLOTS} \
-phred33 \
-trimlog log_SRR${NUM}.txt \
SRR${NUM}_1.fastq.gz \
SRR${NUM}_2.fastq.gz \
paired_SRR${NUM}_1.trim.fastq.gz \
unpaired_SRR${NUM}_1.trim.fastq.gz \
paired_SRR${NUM}_2.trim.fastq.gz \
unpaired_SRR${NUM}_2.trim.fastq.gz \
ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 \
LEADING:20 \
TRAILING:20 \
SLIDINGWINDOW:4:15 \
MINLEN:36
```

```
fastqc --nogroup -o ./FastQC paired_SRR${NUM}_1.trim.fastq.gz
fastqc --nogroup -o ./FastQC paired_SRR${NUM}_2.trim.fastq.gz
```

Trimmomatic オプション

(詳細はマニュアルをご確認ください。)

アダプター除去

ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>

- fastaWithAdaptersEtc: アダプタ配列
- seedMismatches: ミスマッチ許容数
- palindromeClipThreshold: アダプターが結合したペアエンド間でパリンδροームとしてマッチする塩基数
- simpleClipThreshold: アダプターなどの配列がリードにマッチしていないといけない塩基数

トリミング

LEADING: リードの先頭からトリム位置を探した時の下限クオリティ値。

TRAILING: リードの末端からトリム位置を探した時の下限クオリティ値。

SLIDINGWINDOW: ウィンドウサイズと平均クオリティの設定。

MINLEN: トリミング後に閾値以下の塩基長のリードを除去

3. リードの前処理

実行ログの確認

実行

```
[yanakamu@nt097 20181119]$ qsub trimming.sh
Your job-array 11285953.1-6:1 ("trimming.sh") has been submitted
```

job-ID	prior	name	user	state	submit/start at	queue	jclass	slots	ja-task-ID
11284239	0.25113	QLOGIN	yanakamu	r	10/29/2018 10:57:17	login.q@nt097i		1	
11285953	0.25052	trimming.s	yanakamu	t	10/29/2018 17:25:47	short.q@nt119i		4	1
11285953	0.25052	trimming.s	yanakamu	t	10/29/2018 17:25:47	short.q@nt118i		4	2
11285953	0.25052	trimming.s	yanakamu	t	10/29/2018 17:25:47	short.q@nt136i		4	3
11285953	0.25052	trimming.s	yanakamu	t	10/29/2018 17:25:47	short.q@nt111i		4	4
11285953	0.25052	trimming.s	yanakamu	t	10/29/2018 17:25:47	short.q@nt159i		4	5
11285953	0.25052	trimming.s	yanakamu	t	10/29/2018 17:25:47	short.q@nt109i		4	6

ログの確認

```
[yanakamu@nt097 20181119]$ ls -al trimming.sh*
-rw-r--r-- 1 yanakamu yn-nig 1121 10月 29 17:24 2018 trimming.sh
-rw-r--r-- 1 yanakamu yn-nig 3339 10月 29 17:47 2018 trimming.sh.e11285953.1
-rw-r--r-- 1 yanakamu yn-nig 1919 10月 29 17:53 2018 trimming.sh.e11285953.2
-rw-r--r-- 1 yanakamu yn-nig 3339 10月 29 17:46 2018 trimming.sh.e11285953.3
-rw-r--r-- 1 yanakamu yn-nig 3339 10月 29 17:42 2018 trimming.sh.e11285953.4
-rw-r--r-- 1 yanakamu yn-nig 3339 10月 29 17:53 2018 trimming.sh.e11285953.5
-rw-r--r-- 1 yanakamu yn-nig 1577 10月 29 17:53 2018 trimming.sh.e11285953.6
-rw-r--r-- 1 yanakamu yn-nig 110 10月 29 17:47 2018 trimming.sh.o11285953.1
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.o11285953.2
-rw-r--r-- 1 yanakamu yn-nig 110 10月 29 17:46 2018 trimming.sh.o11285953.3
-rw-r--r-- 1 yanakamu yn-nig 110 10月 29 17:42 2018 trimming.sh.o11285953.4
-rw-r--r-- 1 yanakamu yn-nig 110 10月 29 17:53 2018 trimming.sh.o11285953.5
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.o11285953.6
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.pe11285953.1
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.pe11285953.2
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.pe11285953.3
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.pe11285953.4
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.pe11285953.5
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.pe11285953.6
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.po11285953.1
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.po11285953.2
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.po11285953.3
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.po11285953.4
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.po11285953.5
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 17:25 2018 trimming.sh.po11285953.6
```

標準エラーログ

```
[yanakamu@nt097 20181119]$ more trimming.sh.e11285953.1
TrimmomaticPE: Started with arguments:
  -threads 4 -phred33 -trimlog log1.txt SRR453566_1.fastq.gz
SRR453566_2.fastq.gz paired_SRR453566_1.trim.fastq.gz
unpaired_SRR453566_1.trim.fastq.gz paired_SRR453566_2.trim.fastq.gz
unpaired_SRR453566_2.trim.fastq.gz ILLUMINACLIP:/home/yanaka
mu/tools/Trimmomatic-0.38/adapters/TruSeq3-PE-2.fa:2:30:10 LEADING:20
TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36
Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and
'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
Using Long Clipping Sequence: 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA'
Using Long Clipping Sequence: 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC'
Using Long Clipping Sequence: 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
Using Long Clipping Sequence: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT'
ILLUMINACLIP: Using 1 prefix pairs, 4 forward/reverse sequences, 0
forward only sequences, 0 reverse only sequences
Input Read Pairs: 5725730 Both Surviving: 5115482 (89.34%) Forward
Only Surviving: 514793 (8.99%) Reverse Only Surviving: 46123 (0.81%)
Dropped: 49332 (0.86%)
TrimmomaticPE: Completed successfully
Started analysis of paired_SRR453566_1.trim.fastq.gz
Approx 5% complete for paired_SRR453566_1.trim.fastq.gz
Approx 10% complete for paired_SRR453566_1.trim.fastq.gz
Approx 15% complete for paired_SRR453566_1.trim.fastq.gz
Approx 20% complete for paired_SRR453566_1.trim.fastq.gz
.
.
.
```

標準出力ログ

```
[yanakamu@nt097 20181119]$ more trimming.sh.o11285953.1
Analysis complete for paired_SRR453566_1.trim.fastq.gz
Analysis complete for paired_SRR453566_2.trim.fastq.gz
```

3. リードの前処理 結果ファイルの確認 — トリムされたリードファイル

```
[yanakamu@nt097 20181119]$ ls -al read/*fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 495142914 10月 29 13:49 2018 SRR453566_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 525639378 10月 29 13:49 2018 SRR453566_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 669625362 10月 29 13:56 2018 SRR453567_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 695379655 10月 29 13:56 2018 SRR453567_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 478611138 10月 29 14:05 2018 SRR453568_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 511972231 10月 29 14:05 2018 SRR453568_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 352115327 10月 29 14:12 2018 SRR453569_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 370353972 10月 29 14:12 2018 SRR453569_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 585686313 10月 29 14:18 2018 SRR453570_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 611002557 10月 29 14:18 2018 SRR453570_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 537670628 10月 29 14:26 2018 SRR453571_1.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 565354212 10月 29 14:26 2018 SRR453571_2.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 431708443 10月 29 17:45 2018 paired_SRR453566_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 440362034 10月 29 17:45 2018 paired_SRR453566_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 595869441 10月 29 17:52 2018 paired_SRR453567_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 604446384 10月 29 17:52 2018 paired_SRR453567_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 422719949 10月 29 17:45 2018 paired_SRR453568_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 432431371 10月 29 17:45 2018 paired_SRR453568_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 305437921 10月 29 17:41 2018 paired_SRR453569_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 310510017 10月 29 17:41 2018 paired_SRR453569_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 523262456 10月 29 17:51 2018 paired_SRR453570_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 529538482 10月 29 17:51 2018 paired_SRR453570_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 479333120 10月 29 17:52 2018 paired_SRR453571_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 487805402 10月 29 17:52 2018 paired_SRR453571_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 40206909 10月 29 17:45 2018 unpaired_SRR453566_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 3672126 10月 29 17:45 2018 unpaired_SRR453566_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 39576530 10月 29 17:52 2018 unpaired_SRR453567_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 6403453 10月 29 17:52 2018 unpaired_SRR453567_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 35704720 10月 29 17:45 2018 unpaired_SRR453568_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 3204924 10月 29 17:45 2018 unpaired_SRR453568_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 28865362 10月 29 17:41 2018 unpaired_SRR453569_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 3270768 10月 29 17:41 2018 unpaired_SRR453569_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 32486834 10月 29 17:51 2018 unpaired_SRR453570_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 4570929 10月 29 17:51 2018 unpaired_SRR453570_2.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 32244122 10月 29 17:52 2018 unpaired_SRR453571_1.trim.fastq.gz
-rw-r--r-- 1 yanakamu yn-nig 5595130 10月 29 17:52 2018 unpaired_SRR453571_2.trim.fastq.gz
```

ペアードエンドリード数

	トリミング前	トリミング後
SRR453566	5,725,730	5,115,482 (89%)
SRR453567	7,615,732	6,976,849 (92%)
SRR453568	5,565,734	5,030,887 (90%)
SRR453569	4,032,514	3,592,977 (89%)
SRR453570	6,745,975	6,222,696 (92%)
SRR453571	6,163,396	5,650,460 (92%)

前処理後に、ペアでリードが残ったファイル

前処理後に、ペアでリードが残らなかったファイル

トリムログファイル

```
[yanakamu@nt095 read]$ ls log_*
log_SRR453566.txt  log_SRR453567.txt  log_SRR453568.txt
log_SRR453569.txt  log_SRR453570.txt  log_SRR453571.txt
```

```
[yanakamu@nt095 read]$ more log_SRR453566.txt
SRR453566.1 HWI-ST167:4:1101:1597:1986 length=101 99 2 101 0
SRR453566.1 HWI-ST167:4:1101:1597:1986 length=101 0 0 0 0
SRR453566.2 HWI-ST167:4:1101:2535:1992 length=101 99 2 101 0
SRR453566.2 HWI-ST167:4:1101:2535:1992 length=101 0 0 0 0
SRR453566.3 HWI-ST167:4:1101:2980:1962 length=101 84 2 86 15
SRR453566.3 HWI-ST167:4:1101:2980:1962 length=101 0 0 0 0
SRR453566.4 HWI-ST167:4:1101:4066:1970 length=101 99 2 101 0
SRR453566.4 HWI-ST167:4:1101:4066:1970 length=101 0 0 0 0
SRR453566.5 HWI-ST167:4:1101:4770:1966 length=101 98 2 100 1
```

リードID

トリム後の長さ

from to

read の3'末端からトリムされたbp数

3. リードの前処理

結果ファイルの確認 (2)

— 前処理後のクオリティチェック

```
[yanakamu@nt097 20181119]$ ls -al read/FastQC/paired*  
-rw-r--r-- 1 yanakamu yn-nig 296220 10月 29 17:46 2018 paired_SRR453566_1.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 341928 10月 29 17:46 2018 paired_SRR453566_1.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 298283 10月 29 17:47 2018 paired_SRR453566_2.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 342673 10月 29 17:47 2018 paired_SRR453566_2.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 296528 10月 29 17:53 2018 paired_SRR453567_1.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 342904 10月 29 17:53 2018 paired_SRR453567_1.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 298967 10月 29 17:54 2018 paired_SRR453567_2.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 345005 10月 29 17:54 2018 paired_SRR453567_2.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 294108 10月 29 17:46 2018 paired_SRR453568_1.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 338773 10月 29 17:46 2018 paired_SRR453568_1.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 298655 10月 29 17:46 2018 paired_SRR453568_2.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 344275 10月 29 17:46 2018 paired_SRR453568_2.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 292223 10月 29 17:41 2018 paired_SRR453569_1.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 333430 10月 29 17:41 2018 paired_SRR453569_1.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 296439 10月 29 17:42 2018 paired_SRR453569_2.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 339396 10月 29 17:42 2018 paired_SRR453569_2.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 290457 10月 29 17:52 2018 paired_SRR453570_1.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 334199 10月 29 17:52 2018 paired_SRR453570_1.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 290751 10月 29 17:53 2018 paired_SRR453570_2.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 332052 10月 29 17:53 2018 paired_SRR453570_2.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 293124 10月 29 17:53 2018 paired_SRR453571_1.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 336847 10月 29 17:53 2018 paired_SRR453571_1.trim_fastqc.zip  
-rw-r--r-- 1 yanakamu yn-nig 298035 10月 29 17:54 2018 paired_SRR453571_2.trim_fastqc.html  
-rw-r--r-- 1 yanakamu yn-nig 343223 10月 29 17:54 2018 paired_SRR453571_2.trim_fastqc.zip
```

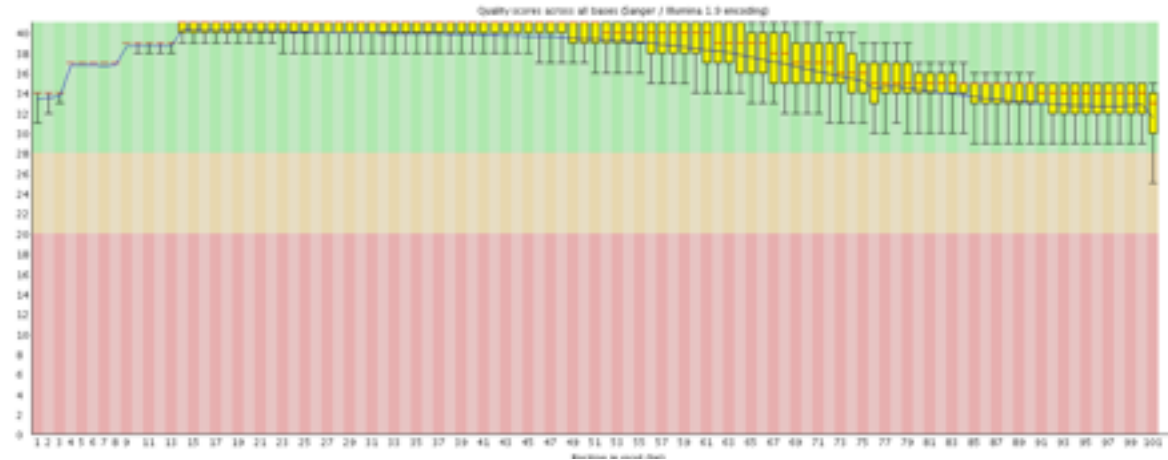
3. リードのトリミング

FastQC レポートの確認 (1)

3'側のクオリティが低い部分とアダプター配列が除去されている。

SRR453566_1

✔ Per base sequence quality

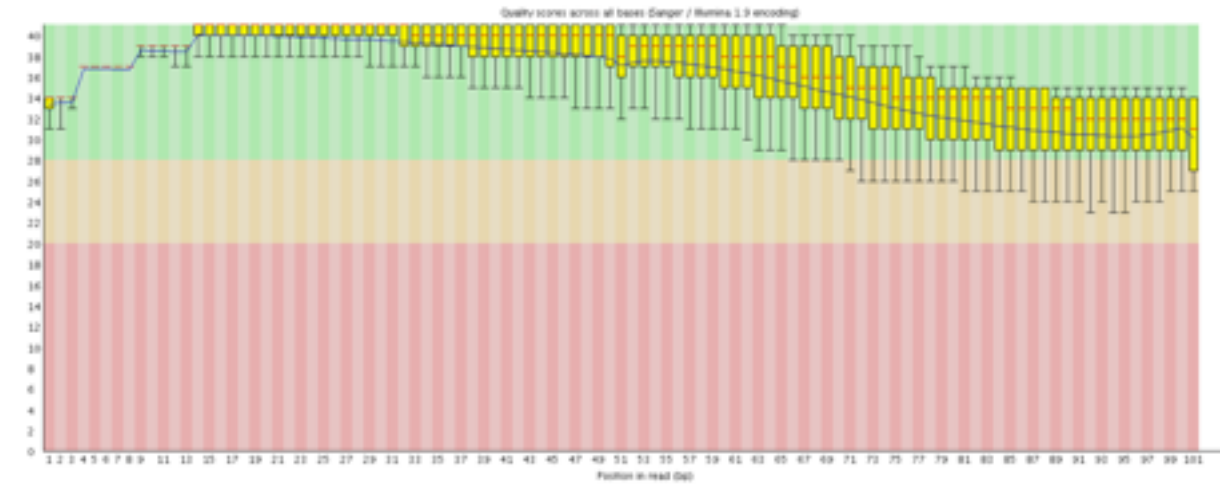


✔ Overrepresented sequences

No overrepresented sequences

SRR453566_2

✔ Per base sequence quality

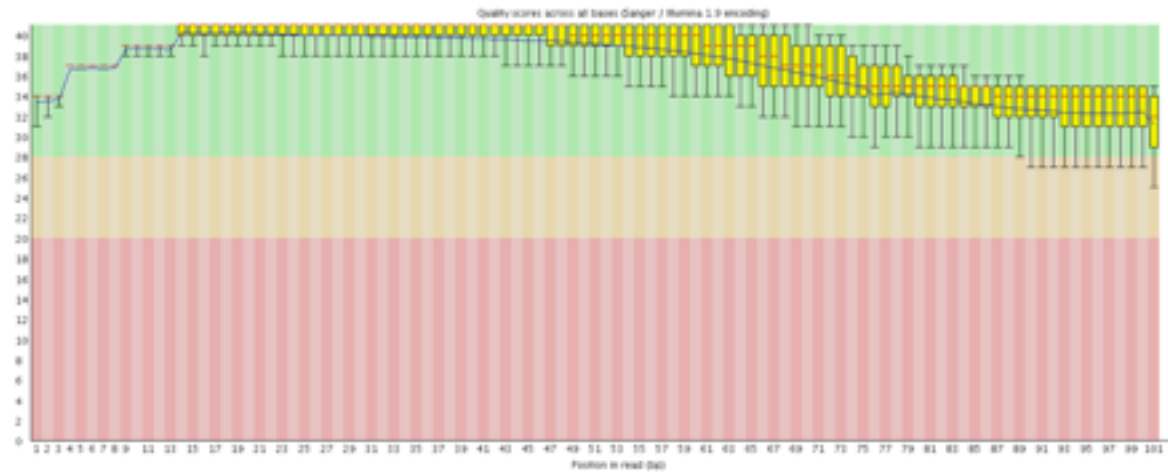


✔ Overrepresented sequences

No overrepresented sequences

SRR453567_1

✔ Per base sequence quality

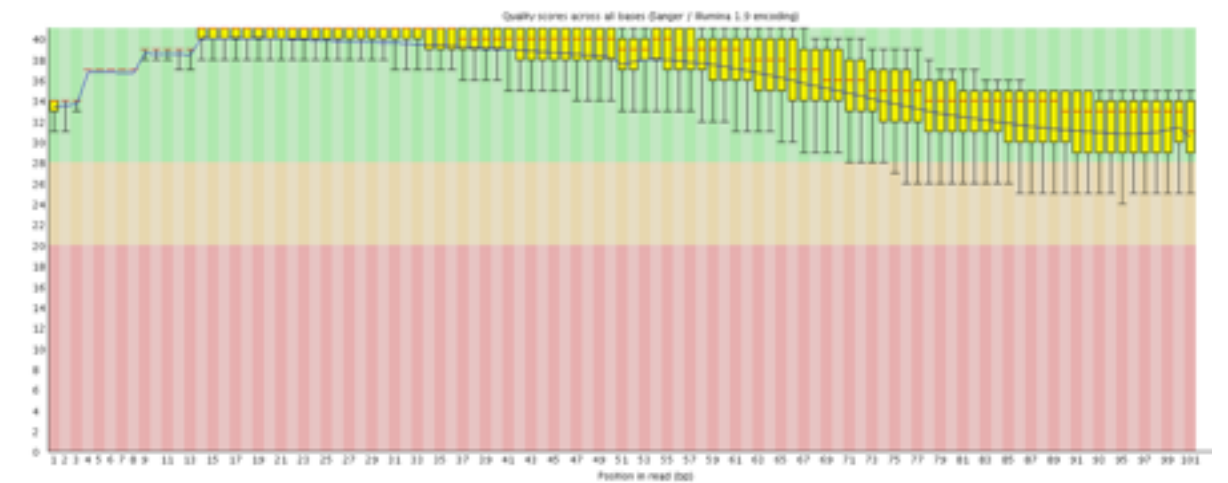


✔ Overrepresented sequences

No overrepresented sequences

SRR453567_2

✔ Per base sequence quality



✔ Overrepresented sequences

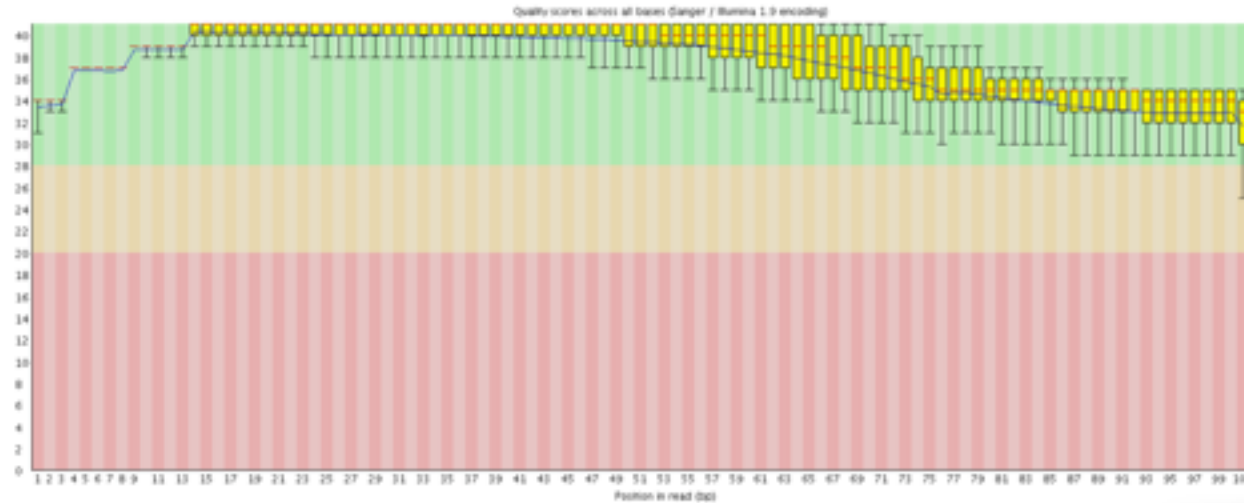
No overrepresented sequences

3. リードのトリミング

FastQC レポートの確認 (2)

SRR453568_1

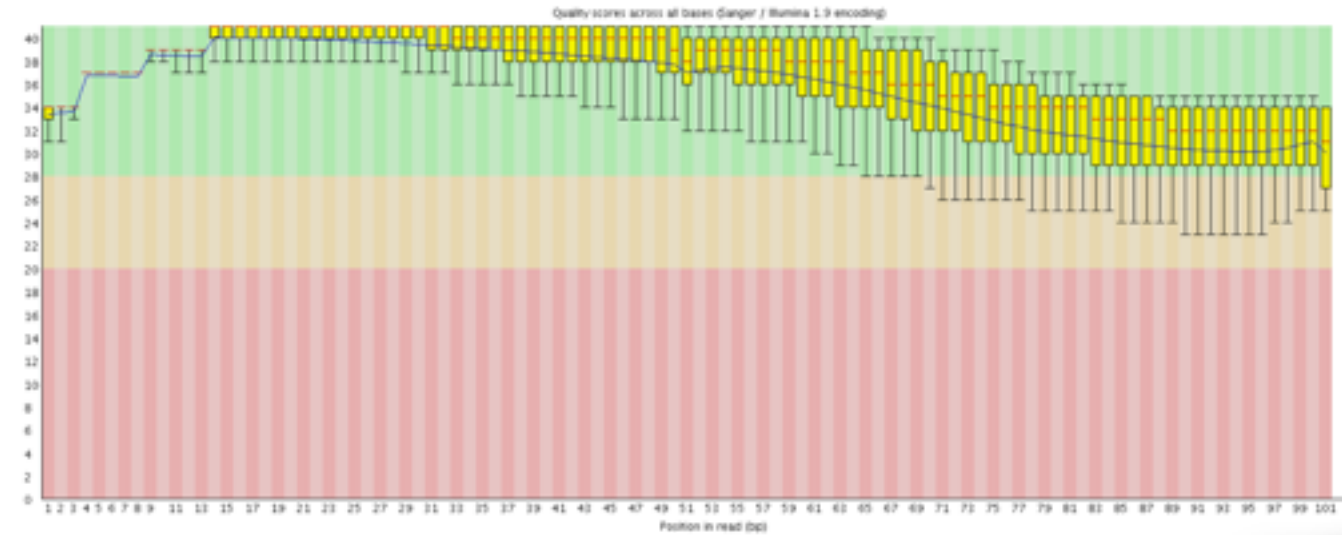
✔ Per base sequence quality



✔ **Overrepresented sequences**
No overrepresented sequences

SRR453568_2

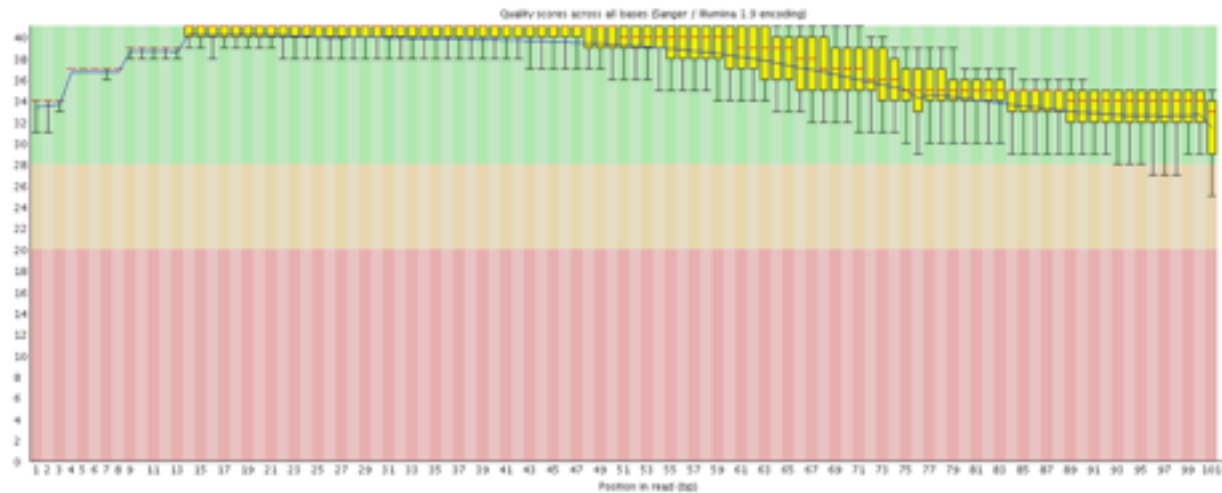
✔ Per base sequence quality



✔ **Overrepresented sequences**
No overrepresented sequences

SRR453569_1

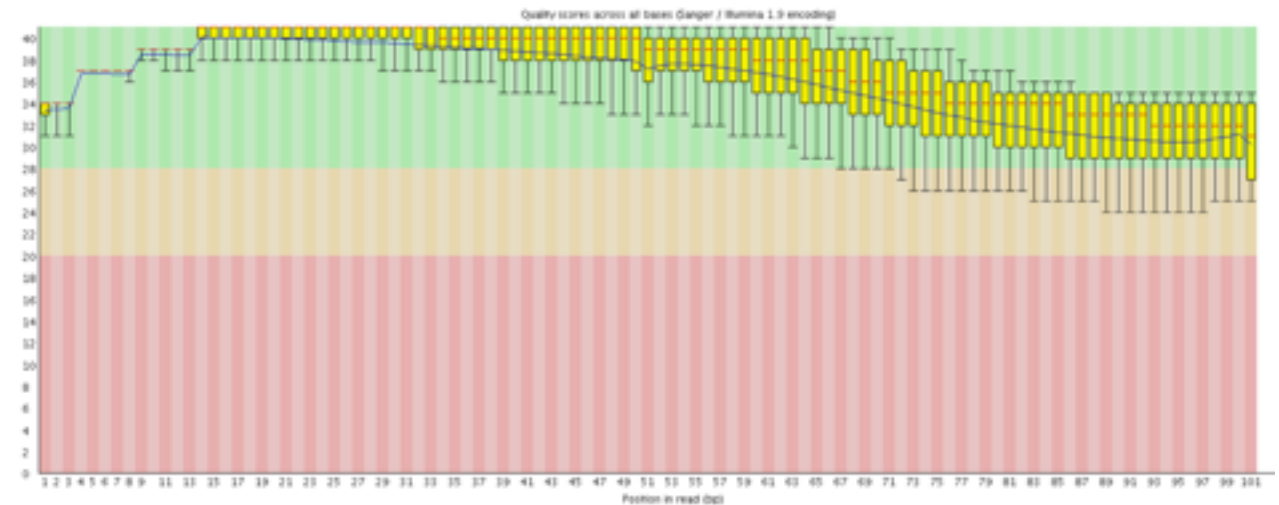
✔ Per base sequence quality



✔ **Overrepresented sequences**
No overrepresented sequences

SRR453569_2

✔ Per base sequence quality



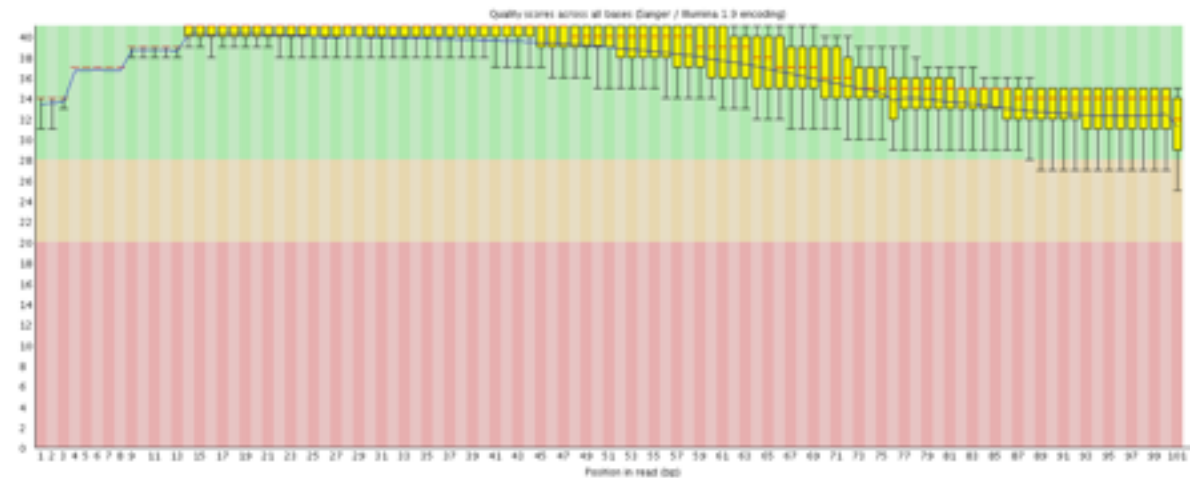
✔ **Overrepresented sequences**
No overrepresented sequences

3. リードのトリミング

FastQC レポートの確認 (3)

SRR453570_1

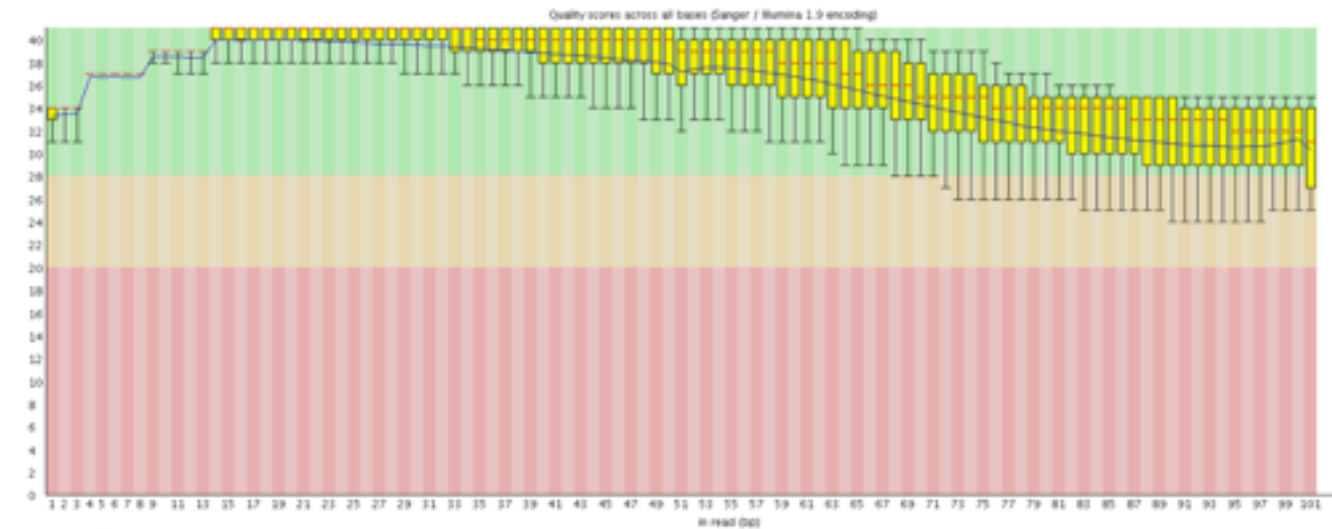
✓ Per base sequence quality



✓ Overrepresented sequences
No overrepresented sequences

SRR4535670_2

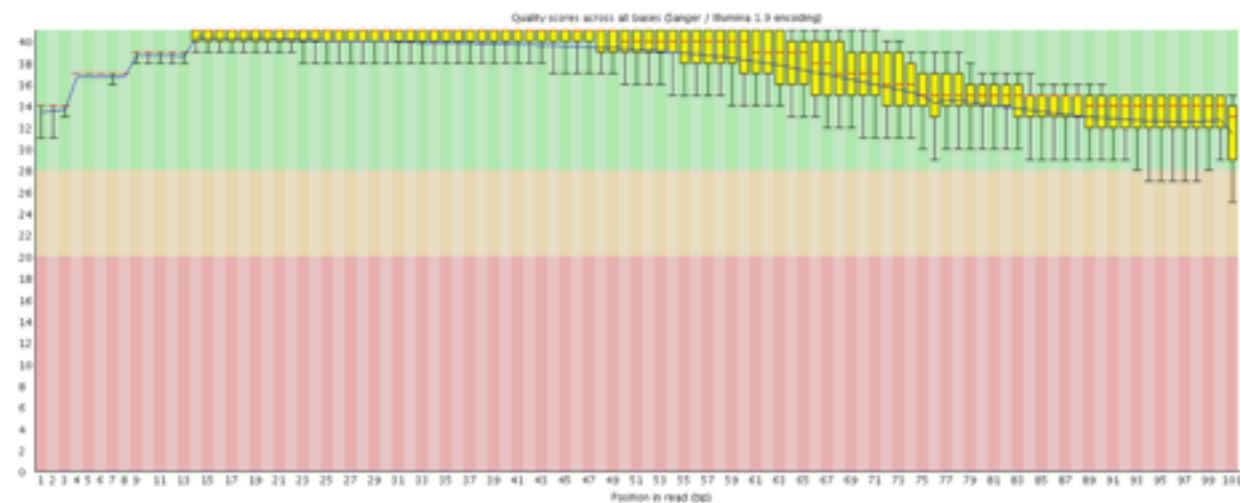
✓ Per base sequence quality



✓ Overrepresented sequences
No overrepresented sequences

SRR453571_1

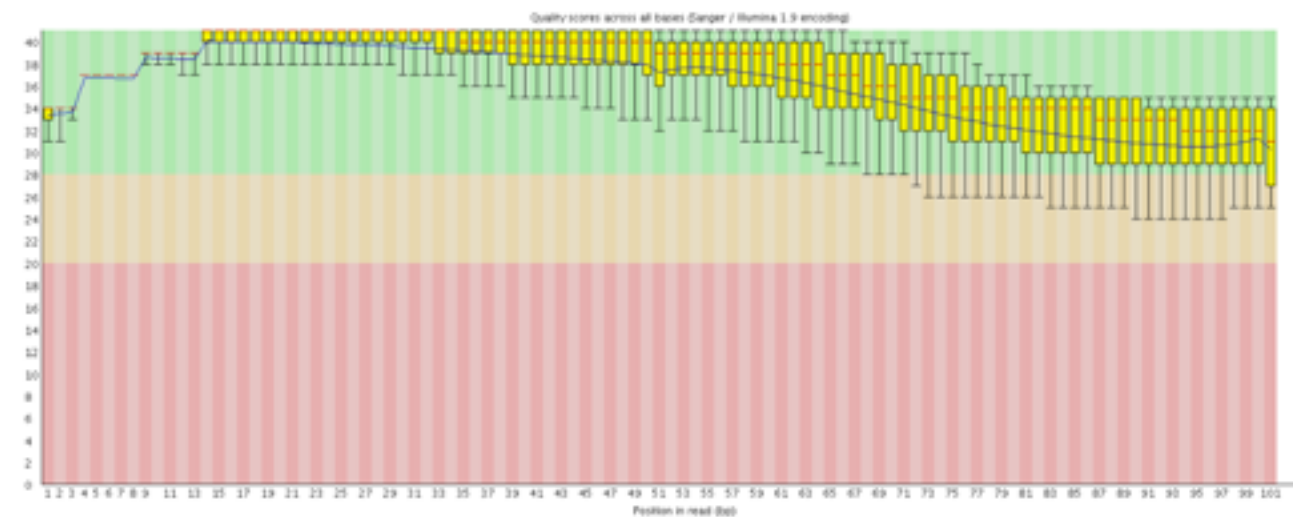
✓ Per base sequence quality



✓ Overrepresented sequences
No overrepresented sequences

SRR453571_2

✓ Per base sequence quality



✓ Overrepresented sequences
No overrepresented sequences

解析の手順

1. リードとリファレンスの準備

fastq-dump ver. 2.8.2

(<https://github.com/ncbi/sra-tools>)

2. リードクオリティチェック

FastQC ver. 0.11.8

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

3. リードのトリミング

Trimmomatic ver. 0.38 (Bolger *et al.*, 2014)

4. リードをリファレンスゲノムにマッピング

HISAT2 ver. 2.1.0 (Kim *et al.*, 2015)

5. 遺伝子毎にリードカウント

featureCounts ver. 1.6.2 (Liao *et al.*, 2014)

4. リードマッピング

index 作成

実行コマンド

```
#$ -S /bin/bash
#$ -pe def_slot 1
#$ -cwd
#$ -l mem_req=3G,s_vmem=3G

export PATH=/usr/local/pkg/hisat2/2.1.0:$PATH

REFERENCE=./reference/s288c.fna
GFF=./reference/s288c.gff

cd reference
mkdir hisat
hisat2-build s288c.fna ./hisat/s288c.fna
```

実行

```
[yanakamu@nt097 20181119]$ qsub index.sh
Your job 11286003 ("index.sh") has been submitted
```

ログの確認

```
[yanakamu@nt097 20181119]$ ls -al index.sh.*
-rw-r--r-- 1 yanakamu yn-nig 2240 10月 29 18:19 2018 index.sh.e11286003
-rw-r--r-- 1 yanakamu yn-nig 3919 10月 29 18:19 2018 index.sh.o11286003
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 18:19 2018 index.sh.pe11286003
-rw-r--r-- 1 yanakamu yn-nig 0 10月 29 18:19 2018 index.sh.po11286003
```

標準エラーログ

```
[yanakamu@nt097 20181119]$ more index.sh.e11286003
Settings:
  Output files: "./hisat/s288c.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
  Strings: unpacked
  Local offset rate: 3 (one in 8)
  Local fTable chars: 6
  Local sequence length: 57344
  Local sequence overlap between two consecutive indexes: 1024
  Endianness: little
  Actual local endianness: little
  Sanity checking: disabled
  Assertions: disabled
  Random seed: 0
  Sizeofs: void*:8, int:4, long:8, size_t:8
  .
  .
  .
  .
```

標準出力ログ

```
[yanakamu@nt097 20181119]$ more index.sh.o11286003
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
  Invoking Larsson-Sadakane on ranks time: 00:00:00
  Sanity-checking and returning
  .
  .
  .
  .
```

結果ファイルの確認

```
[yanakamu@nt097 20181119]$ ls -al reference/hisat/
合計 22324
```

インデックスファイル

```
drwxr-xr-x 2 yanakamu yn-nig 4096 10月 29 18:19 2018 ./
drwxr-xr-x 3 yanakamu yn-nig 4096 10月 29 18:19 2018 ../
-rw-r--r-- 1 yanakamu yn-nig 8248454 10月 29 18:19 2018 s288c.fna.1.ht2
-rw-r--r-- 1 yanakamu yn-nig 3039284 10月 29 18:19 2018 s288c.fna.2.ht2
-rw-r--r-- 1 yanakamu yn-nig 161 10月 29 18:19 2018 s288c.fna.3.ht2
-rw-r--r-- 1 yanakamu yn-nig 3039277 10月 29 18:19 2018 s288c.fna.4.ht2
-rw-r--r-- 1 yanakamu yn-nig 5399069 10月 29 18:19 2018 s288c.fna.5.ht2
-rw-r--r-- 1 yanakamu yn-nig 3092708 10月 29 18:19 2018 s288c.fna.6.ht2
-rw-r--r-- 1 yanakamu yn-nig 12 10月 29 18:19 2018 s288c.fna.7.ht2
-rw-r--r-- 1 yanakamu yn-nig 8 10月 29 18:19 2018 s288c.fna.8.ht2
```

4. リードマッピング

シェルスクリプト

```
#$ -S /bin/bash
#$ -pe def_slot 4
#$ -cwd
#$ -t 1-6:1
#$ -l mem_req=8G,s_vmem=8G
#$ -l short

export PATH=/usr/local/pkg/hisat2/2.1.0:$PATH
export PATH=/usr/local/pkg/samtools/1.7/bin:$PATH

# Batch culture: SRX135198   SRR453566 - SRR453568
# chemostat: SRX135710 SRR453569 - SRR453571
ACCESSIONS=(453566 453567 453568 453569 453570 453571)
no=`expr ${SGE_TASK_ID} - 1`

NUM=${ACCESSIONS[${no}]}
PREFIX=SRR${NUM}

echo HISAT2 for $PREFIX

QUERY1=../read/paired_SRR${NUM}_1.trim.fastq.gz
QUERY2=../read/paired_SRR${NUM}_2.trim.fastq.gz

REFERENCE=../reference/hisat/s288c.fna

cd hisat
hisat2 -p ${NSLOTS} -x ${REFERENCE} -1 ${QUERY1} -2 ${QUERY2} -S ${PREFIX}.sam

samtools view -@ ${NSLOTS} -b ${PREFIX}.sam > ${PREFIX}.bam
samtools sort -@ ${NSLOTS} ${PREFIX}.bam > ${PREFIX}.sorted.bam
```

<- マッピング (samファイルを出力)

<- sam -> bam に変換

<- ポジションでソート

SAM フォーマット

```
@HD VN:1.0 S0:unsorted
@SQ SN:NC_001133.9 LN:230218
@SQ SN:NC_001134.8 LN:813184
@SQ SN:NC_001135.5 LN:316620
@SQ SN:NC_001136.10 LN:1531933
@SQ SN:NC_001137.3 LN:576874
@SQ SN:NC_001138.5 LN:270161
@SQ SN:NC_001139.9 LN:1090940
@SQ SN:NC_001140.6 LN:562643
@SQ SN:NC_001141.2 LN:439888
@SQ SN:NC_001142.9 LN:745751
@SQ SN:NC_001143.9 LN:666816
@SQ SN:NC_001144.5 LN:1078177
@SQ SN:NC_001145.3 LN:924431
@SQ SN:NC_001146.8 LN:784333
@SQ SN:NC_001147.6 LN:1091291
@SQ SN:NC_001148.4 LN:948066
@SQ SN:NC_001224.1 LN:85779
@PG ID:hisat2 PN:hisat2 VN:2.1.0 CL:"/usr/local/pkg/hisat2/2.1.0/hisat2-align-s --wrapper basic-0 -p 4 -x ../reference/hisat/s
SRR453566.24 83 NC_001139.9 727620 60 101M = 727518 -203 AAGGGTAAA... ?DCCDDDDDD... AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z
SRR453566.24 163 NC_001139.9 727518 60 69M = 727620 203 TTAATCAAG... =DFFFFHHHH... AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z
SRR453566.22 99 NC_001142.9 509705 60 101M = 509740 136 CAAAGCGTA... CCCFFFFFFFHG... AS:i:-6 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1
SRR453566.22 147 NC_001142.9 509740 60 101M = 509705 -136 GGTATATTT... @DA@:>>>@C... AS:i:-4 ZS:i:-7 XN:i:0 XM:i:1 XO:i:1
SRR453566.23 99 NC_001134.8 674240 60 101M = 674286 131 TTTTCTTCA... @BCFFFFFFFH... AS:i:-6 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1
SRR453566.23 147 NC_001134.8 674286 60 85M = 674240 -131 AACAAAAGC... ??>C>@5(@>... AS:i:-4 ZS:i:-10 XN:i:0 XM:i:1 XO:i:1
```

@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

QNAME	FLG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	optional fields
-------	-----	-------	-----	------	-------	-------	-------	------	-----	------	-----------------

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 リファレンスにインサクションあり 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。

4.リードマッピング

実行ログの確認

実行

```
[yanakamu@nt097 20181119]$ qsub hisat.sh
Your job-array 11288242.1-6:1 ("hisat.sh") has been submitted
```

ログの確認

```
[yanakamu@nt097 20181119]$ ls -al hisat.sh*
-rw-r--r-- 1 yanakamu yn-nig 992 10月 30 17:01 2018 hisat.sh
-rw-r--r-- 1 yanakamu yn-nig 686 10月 30 17:04 2018 hisat.sh.e11288242.1
-rw-r--r-- 1 yanakamu yn-nig 686 10月 30 17:05 2018 hisat.sh.e11288242.2
-rw-r--r-- 1 yanakamu yn-nig 686 10月 30 17:04 2018 hisat.sh.e11288242.3
-rw-r--r-- 1 yanakamu yn-nig 686 10月 30 17:04 2018 hisat.sh.e11288242.4
-rw-r--r-- 1 yanakamu yn-nig 690 10月 30 17:05 2018 hisat.sh.e11288242.5
-rw-r--r-- 1 yanakamu yn-nig 686 10月 30 17:05 2018 hisat.sh.e11288242.6
-rw-r--r-- 1 yanakamu yn-nig 21 10月 30 17:02 2018 hisat.sh.o11288242.1
-rw-r--r-- 1 yanakamu yn-nig 21 10月 30 17:02 2018 hisat.sh.o11288242.2
-rw-r--r-- 1 yanakamu yn-nig 21 10月 30 17:02 2018 hisat.sh.o11288242.3
-rw-r--r-- 1 yanakamu yn-nig 21 10月 30 17:02 2018 hisat.sh.o11288242.4
-rw-r--r-- 1 yanakamu yn-nig 21 10月 30 17:02 2018 hisat.sh.o11288242.5
-rw-r--r-- 1 yanakamu yn-nig 21 10月 30 17:02 2018 hisat.sh.o11288242.6
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.pe11288242.1
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.pe11288242.2
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.pe11288242.3
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.pe11288242.4
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.pe11288242.5
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.pe11288242.6
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.po11288242.1
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.po11288242.2
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.po11288242.3
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.po11288242.4
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.po11288242.5
-rw-r--r-- 1 yanakamu yn-nig 0 10月 30 17:02 2018 hisat.sh.po11288242.6
```

標準エラーログ

```
[yanakamu@nt097 20181119]$ more hisat.sh.e11288242.1
5115482 reads; of these:
  5115482 (100.00%) were paired; of these:
    323270 (6.32%) aligned concordantly 0 times
    4525002 (88.46%) aligned concordantly exactly 1 time
    267210 (5.22%) aligned concordantly >1 times
  ----
    323270 pairs aligned concordantly 0 times; of these:
      77119 (23.86%) aligned discordantly 1 time
  ----
    246151 pairs aligned 0 times concordantly or discordantly; of these:
      492302 mates make up the pairs; of these:
        341624 (69.39%) aligned 0 times
        135730 (27.57%) aligned exactly 1 time
        14948 (3.04%) aligned >1 times
96.66% overall alignment rate
[bam_sort_core] merging from 0 files and 4 in-memory blocks...
```

標準出力ログ

```
[yanakamu@nt097 20181119]$ more hisat.sh.o11288242.1
HISAT2 for SRR453566
```

マップ率

マップ率	
SRR453566	96.66%
SRR453567	96.81%
SRR453568	97.02%
SRR453569	93.61%
SRR453570	67.98%
SRR453571	94.61%

4. リードマッピング

結果ファイル

```
[yanakamu@nt097 20181119]$ ls -al hisat/  
合計 32763936
```

drwxr-xr-x	2	yanakamu	yn-nig	4096	10月 30 17:06 2018	./
drwxr-xr-x	7	yanakamu	yn-nig	12288	10月 30 17:02 2018	../
-rw-r--r--	1	yanakamu	yn-nig	998681982	10月 30 17:04 2018	SRR453566.bam
-rw-r--r--	1	yanakamu	yn-nig	3638190421	10月 30 17:04 2018	SRR453566.sam
-rw-r--r--	1	yanakamu	yn-nig	709221522	10月 30 17:05 2018	SRR453566.sorted.bam
-rw-r--r--	1	yanakamu	yn-nig	1373197882	10月 30 17:05 2018	SRR453567.bam
-rw-r--r--	1	yanakamu	yn-nig	4995548288	10月 30 17:04 2018	SRR453567.sam
-rw-r--r--	1	yanakamu	yn-nig	959672351	10月 30 17:06 2018	SRR453567.sorted.bam
-rw-r--r--	1	yanakamu	yn-nig	980013680	10月 30 17:04 2018	SRR453568.bam
-rw-r--r--	1	yanakamu	yn-nig	3571579454	10月 30 17:04 2018	SRR453568.sam
-rw-r--r--	1	yanakamu	yn-nig	694677767	10月 30 17:05 2018	SRR453568.sorted.bam
-rw-r--r--	1	yanakamu	yn-nig	699894526	10月 30 17:04 2018	SRR453569.bam
-rw-r--r--	1	yanakamu	yn-nig	2548114661	10月 30 17:03 2018	SRR453569.sam
-rw-r--r--	1	yanakamu	yn-nig	511209451	10月 30 17:04 2018	SRR453569.sorted.bam
-rw-r--r--	1	yanakamu	yn-nig	1143810618	10月 30 17:04 2018	SRR453570.bam
-rw-r--r--	1	yanakamu	yn-nig	3971706758	10月 30 17:04 2018	SRR453570.sam
-rw-r--r--	1	yanakamu	yn-nig	864362830	10月 30 17:05 2018	SRR453570.sorted.bam
-rw-r--r--	1	yanakamu	yn-nig	1102455915	10月 30 17:04 2018	SRR453571.bam
-rw-r--r--	1	yanakamu	yn-nig	4003592729	10月 30 17:04 2018	SRR453571.sam
-rw-r--r--	1	yanakamu	yn-nig	783792693	10月 30 17:05 2018	SRR453571.sorted.bam

sam: マッピング結果のフォーマット

bam: sam のバイナリーファイル

XXXX.sorted.bam: ポジションでソート

解析の手順

1. リードとリファレンスの準備

fastq-dump ver. 2.8.2

(<https://github.com/ncbi/sra-tools>)

2. リードクオリティチェック

FastQC ver. 0.11.8

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

3. リードのトリミング

Trimmomatic ver. 0.38 (Bolger *et al.*, 2014)

4. リードをリファレンスゲノムにマッピング

HISAT2 ver. 2.1.0 (Kim *et al.*, 2015)

5. 遺伝子毎にリードカウント

featureCounts ver. 1.6.2 (Liao *et al.*, 2014)

5. featureカウント

gff の修正

```
#$ -S /bin/bash
#$ -pe def_slot 8
#$ -cwd
#$ -l mem_req=8G,s_vmem=8G

pyenv shell 3.6.1

cd reference
python ../program/add_gene_id s288c.gff s288c_e.gff
```

ダウンロードしてきた gff ファイル (s288c.gff) は、attribute に gene_id が含まれていない。後続処理で、gene 毎のread 数をカウントするために、gene_id タグを追加する。

python プログラムのコード
については、2日目の谷澤さんの講習
で説明があります。

s288c.gff 実行前

```
##gff-version 3
##!gff-spec-version 1.21
##!processor NCBI annotwriter
##!genome-build R64
##!genome-build-accession NCBI_Assembly:GCF_000146045.2
##!annotation-source SGD R64-2-1
##sequence-region NC_001133.9 1 230218
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=559292
NC_001133.9 RefSeq region 1 230218 . + .
NC_001133.9 RefSeq telomere 1 801 . - .
NC_001133.9 RefSeq origin_of_replication 707 776 . + .
NC_001133.9 RefSeq gene 1807 2169 . . .
NC_001133.9 RefSeq mRNA 1807 2169 . - .
NC_001133.9 RefSeq exon 1807 2169 . - .
NC_001133.9 RefSeq CDS 1807 2169 . - 0
NC_001133.9 RefSeq gene 2480 2707 . + .
NC_001133.9 RefSeq mRNA 2480 2707 . + .
NC_001133.9 RefSeq exon 2480 2707 . + .
NC_001133.9 RefSeq CDS 2480 2707 . + 0
NC_001133.9 RefSeq gene 7235 9016 . - .
NC_001133.9 RefSeq mRNA 7235 9016 . - .
NC_001133.9 RefSeq exon 7235 9016 . - .
NC_001133.9 RefSeq CDS 7235 9016 . - 0
NC_001133.9 RefSeq origin_of_replication 7997 8547 . + .
NC_001133.9 RefSeq gene 11565 11951 . . .
NC_001133.9 RefSeq mRNA 11565 11951 . - .
NC_001133.9 RefSeq exon 11565 11951 . - .
NC_001133.9 RefSeq CDS 11565 11951 . - 0
NC_001133.9 RefSeq gene 12046 12426 . + .
NC_001133.9 RefSeq mRNA 12046 12426 . + .
NC_001133.9 RefSeq exon 12046 12426 . + .
NC_001133.9 RefSeq CDS 12046 12426 . + 0
```

gene_id の記載がない。

```
ID=id0;Dbxref=taxon:559292;Name=l;chromosome=l;gbkey=Src;genome=chromosome;mol_type=genomic DNA;strain=S288C
ID=id1;Dbxref=SGD:S000028862;Note=TEL01L%3B Telomeric region on the left arm of Chromosome l%3B composed of an X element
ID=id2;Dbxref=SGD:S000121252;Note=ARS102~Autonomously Replicating Sequence;gbkey=rep_origin
ID=gene0;Dbxref=GeneID:851229;Name=PAU8;end_range=2169,.;gbkey=Gene;gene=PAU8;gene_biotype=protein_coding;locus_tag=PAU8
ID=rna0;Parent=gene0;Dbxref=GeneID:851229,Genbank:NM_001180043.1;Name=NM_001180043.1;end_range=2169,.;gbkey=mRNA
ID=id3;Parent=rna0;Dbxref=GeneID:851229,Genbank:NM_001180043.1;end_range=2169,.;gbkey=mRNA;gene=PAU8;partial=true;product=hypothetical protein
ID=cds0;Parent=rna0;Dbxref=SGD:S000002142,GenelD:851229,Genbank:NP_009332.1;Name=NP_009332.1;Note=hypothetical protein
ID=gene1;Dbxref=GeneID:1466426;Name=YAL067W-A;end_range=2707,.;gbkey=Gene;gene_biotype=protein_coding;locus_tag=YAL067W
ID=rna1;Parent=gene1;Dbxref=GeneID:1466426,Genbank:NM_001184582.1;Name=NM_001184582.1;end_range=2707,.;gbkey=mRNA
ID=id4;Parent=rna1;Dbxref=GeneID:1466426,Genbank:NM_001184582.1;end_range=2707,.;gbkey=mRNA;partial=true;product=hypothetical protein
ID=cds1;Parent=rna1;Dbxref=SGD:S000028593,GenelD:1466426,Genbank:NP_878038.1;Name=NP_878038.1;Note=hypothetical protein
ID=gene2;Dbxref=GeneID:851230;Name=SEO1;end_range=9016,.;gbkey=Gene;gene=SEO1;gene_biotype=protein_coding;locus_tag=SEO1
ID=rna2;Parent=gene2;Dbxref=GeneID:851230,Genbank:NM_001178208.1;Name=NM_001178208.1;end_range=9016,.;gbkey=mRNA
ID=id5;Parent=rna2;Dbxref=GeneID:851230,Genbank:NM_001178208.1;end_range=9016,.;gbkey=mRNA;gene=SEO1;partial=true;product=hypothetical protein
ID=cds2;Parent=rna2;Dbxref=SGD:S000000062,GenelD:851230,Genbank:NP_009333.1;Name=NP_009333.1;Note=Putative permease
ID=id6;Dbxref=SGD:S000121253;Note=ARS103~Autonomously Replicating Sequence%3B replication origin of very weak f
ID=gene3;Dbxref=GeneID:851232;Name=YAL065C;end_range=11951,.;gbkey=Gene;gene_biotype=protein_coding;locus_tag=YAL065C
ID=rna3;Parent=gene3;Dbxref=GeneID:851232,Genbank:NM_001179897.1;Name=NM_001179897.1;end_range=11951,.;gbkey=mRNA
ID=id7;Parent=rna3;Dbxref=GeneID:851232,Genbank:NM_001179897.1;end_range=11951,.;gbkey=mRNA;partial=true;product=hypothetical protein
ID=cds3;Parent=rna3;Dbxref=SGD:S000001817,GenelD:851232,Genbank:NP_009335.1;Name=NP_009335.1;Note=hypothetical protein
ID=gene4;Dbxref=GeneID:851233;Name=YAL064W-B;end_range=12426,.;gbkey=Gene;gene_biotype=protein_coding;locus_tag=YAL064W
ID=rna4;Parent=gene4;Dbxref=GeneID:851233,Genbank:NM_001180042.1;Name=NM_001180042.1;end_range=12426,.;gbkey=mRNA
ID=id8;Parent=rna4;Dbxref=GeneID:851233,Genbank:NM_001180042.1;end_range=12426,.;gbkey=mRNA;partial=true;product=hypothetical protein
ID=cds4;Parent=rna4;Dbxref=SGD:S000002141,GenelD:851233,Genbank:NP_009336.1;Name=NP_009336.1;Note=Fungal-specific
```

5. featureカウント

修正後の gffファイルの確認

s288c.gff 実行後

gene_id が付与された

```
##gff-version 3
##sequence-region NC_001133.9 1 230218
NC_001133.9    annotation    remark    1      230218    .      .      .      gff-version=3;sequence-region=%28%27NC_001133.9%27%2C 0%2C 230218%29,%28%27NC_001134.8%27%2C 0%2C 813184%29,%28%27NC_001140.6%27%2C 0%2C 562643%29,%28%27NC_001141.2%27%2C 0%2C 439888%29,%28%27NC_001142.9%27%2C 0%2C 745751%29,%28%27NC_001143.9%27%2C 0%2C 666816%29,%28%27NC_001144.5%27%2C 0%2C 1078177%29,%28%27NC_001145.3%27%2C 0%2C 924431%29,%28%27NC_001146.8%27%2C 0%2C 784333%29,%28%27NC_001147.6%27%2C 0%2C 10948066%29,%28%27NC_001224.1%27%2C 0%2C 85779%29;species=https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi%3Fid%3D559292
NC_001133.9    RefSeq    region    1      230218    .      +      .      Dbxref=taxon:559292;ID=id0;Name=l;chromosome=l;gbkey=Src;genome=chromosome;mol_type=genomic DNA;strain=S288C
NC_001133.9    RefSeq    telomere  1      801      .      -      .      Dbxref=SGD:S000028862;ID=id1;Note=TEL01L%3B Telomeric region on the left arm of Chromosome l%3B composed of an X element core sequence%2C X element combinatorial repeats%2C and a short terminal stretch of telomeric repeats;gbkey=telomere
NC_001133.9    RefSeq    origin_of_replication 707    776      .      +      .      Dbxref=SGD:S000121252;ID=id2;Note=ARS102%7EAutonomously Replicating Sequence;gbkey=rep_origin
NC_001133.9    RefSeq    gene      1807    2169    .      -      .      Dbxref=GeneID:851229;ID=gene0;Name=PAU8;end_range=2169;gbkey=Gene;gene=PAU8;gene_biotype=protein_coding;gene_id=gene_0001;locus_tag=YAL068C;partial=true;start_range=.,1807
NC_001133.9    RefSeq    mRNA     1807    2169    .      -      .      Dbxref=GeneID:851229,Genbank:NM_001180043.1;ID=rna0;Name=NM_001180043.1;Parent=gene0;end_range=2169;gbkey=mRNA;gene=PAU8;gene_id=gene_0001;partial=true;product=seripauperin PAU8;start_range=.,1807;transcript_id=NM_001180043.1
NC_001133.9    RefSeq    exon     1807    2169    .      -      .      Dbxref=GeneID:851229,Genbank:NM_001180043.1;ID=id3;Parent=rna0;end_range=2169;gbkey=mRNA;gene=PAU8;gene_id=gene_0001;partial=true;product=seripauperin PAU8;start_range=.,1807;transcript_id=NM_001180043.1
NC_001133.9    RefSeq    CDS      1807    2169    .      -      0      Dbxref=SGD:S000002142,GenelID:851229,Genbank:NP_009332.1;ID=cds0;Name=NP_009332.1;Note=hypothetical protein%3B member of the seripauperin multigene family encoded mainly in subtelomeric regions;Parent=rna0;gbkey=CDS;gene=PAU8;gene_id=gene_0001;product=seripauperin PAU8;protein_id=NP_009332.1
NC_001133.9    RefSeq    gene     2480    2707    .      +      .      Dbxref=GeneID:1466426;ID=gene1;Name=YAL067W-A;end_range=2707;gbkey=Gene;gene=YAL067W-A;gene_biotype=protein_coding;gene_id=gene_0002;locus_tag=YAL067W-A;partial=true;start_range=.,2480
NC_001133.9    RefSeq    mRNA     2480    2707    .      +      .      Dbxref=GeneID:1466426,Genbank:NM_001184582.1;ID=rna1;Name=NM_001184582.1;Parent=gene1;end_range=2707;gbkey=mRNA;gene=YAL067W-A;gene_id=gene_0002;partial=true;product=hypothetical protein;start_range=.,2480;transcript_id=NM_001184582.1
NC_001133.9    RefSeq    exon     2480    2707    .      +      .      Dbxref=GeneID:1466426,Genbank:NM_001184582.1;ID=id4;Parent=rna1;end_range=2707;gbkey=mRNA;gene=YAL067W-A;gene_id=gene_0002;partial=true;product=hypothetical protein;start_range=.,2480;transcript_id=NM_001184582.1
NC_001133.9    RefSeq    CDS      2480    2707    .      +      0      Dbxref=SGD:S000028593,GenelID:1466426,Genbank:NP_878038.1;ID=cds1;Name=NP_878038.1;Note=hypothetical protein%3B identified by gene-trapping%2C microarray-based expression analysis%2C and genome-wide homology searching;Parent=rna1;gbkey=CDS;gene=YAL067W-A;gene_id=gene_0002;product=hypothetical protein;protein_id=NP_878038.1
NC_001133.9    RefSeq    gene     7235    9016    .      -      .      Dbxref=GeneID:851230;ID=gene2;Name=SEO1;end_range=9016;gbkey=Gene;gene=SEO1;gene_biotype=protein_coding;gene_id=gene_0003;locus_tag=YAL067C;partial=true;start_range=.,7235
NC_001133.9    RefSeq    mRNA     7235    9016    .      -      .      Dbxref=GeneID:851230,Genbank:NM_001178208.1;ID=rna2;Name=NM_001178208.1;Parent=gene2;end_range=9016;gbkey=mRNA;gene=SEO1;gene_id=gene_0003;partial=true;product=putative permease SEO1;start_range=.,7235;transcript_id=NM_001178208.1
NC_001133.9    RefSeq    exon     7235    9016    .      -      .      Dbxref=GeneID:851230,Genbank:NM_001178208.1;ID=id5;Parent=rna2;end_range=9016;gbkey=mRNA;gene=SEO1;gene_id=gene_0003;partial=true;product=putative permease SEO1;start_range=.,7235;transcript_id=NM_001178208.1
NC_001133.9    RefSeq    CDS      7235    9016    .      -      0      Dbxref=SGD:S000000062,GenelID:851230,Genbank:NP_009333.1;ID=cds2;Name=NP_009333.1;Note=Putative permease%3B member of the allantoate transporter subfamily of the major facilitator superfamily%3B mutation confers resistance to ethionine sulfoxide;Parent=rna2;gbkey=CDS;gene=SEO1;gene_id=gene_0003;product=putative permease SEO1;protein_id=NP_009333.1
NC_001133.9    RefSeq    origin_of_replication 7997    8547    .      +      .      Dbxref=SGD:S000121253;ID=id6;Note=ARS103%7EAutonomously Replicating Sequence%3B replication origin of very weak function;gbkey=rep_origin
NC_001133.9    RefSeq    gene     11565   11951    .      -      .      Dbxref=GeneID:851232;ID=gene3;Name=YAL065C;end_range=11951;gbkey=Gene;gene=YAL065C;gene_biotype=protein_coding;gene_id=gene_0004;locus_tag=YAL065C;partial=true;start_range=.,11565
NC_001133.9    RefSeq    mRNA     11565   11951    .      -      .      Dbxref=GeneID:851232,Genbank:NM_001179897.1;ID=rna3;Name=NM_001179897.1;Parent=gene3;end_range=11951;gbkey=mRNA;gene=YAL065C;gene_id=gene_0004;partial=true;product=hypothetical protein;start_range=.,11565;transcript_id=NM_001179897.1
NC_001133.9    RefSeq    exon     11565   11951    .      -      .      Dbxref=GeneID:851232,Genbank:NM_001179897.1;ID=id7;Parent=rna3;end_range=11951;gbkey=mRNA;gene=YAL065C;gene_id=gene_0004;partial=true;product=hypothetical protein;start_range=.,11565;transcript_id=NM_001179897.1
NC_001133.9    RefSeq    CDS      11565   11951    .      -      0      Dbxref=SGD:S000001817,GenelID:851232,Genbank:NP_009335.1;ID=cds3;Name=NP_009335.1;Note=hypothetical protein%3B shows sequence similarity to FLO1 and other flocculins;Parent=rna3;gbkey=CDS;gene=YAL065C;gene_id=gene_0004;product=hypothetical protein;protein_id=NP_009335.1
```

featurecount 実行とログの確認

オプション

- p ペアエンドのリードではなくフラグメントをカウント
- T CPU コア数
- t 指定された feature type にマップされている フラグメント (リード) をカウントする。
- g meta-feature を指定する
- a アノテーションファイル
- o 出力ファイル

```
[yanakamu@nt097 20181119]$ qsub featurecount.sh
Your job 11288294 ("featurecount.sh") has been submitted
```

標準エラーログ

標準出力ログ

出力なし

5. featureカウント

featurecount 結果ファイル

```
[yanakamu@nt097 20181119]$ ls -al featurecount
合計 440
drwxr-xr-x 2 yanakamu yn-nig  4096 10月 30 18:22 2018 ./
drwxr-xr-x 9 yanakamu yn-nig 12288 10月 30 18:22 2018 ../
-rw-r--r-- 1 yanakamu yn-nig 428718 10月 30 18:22 2018 counts.txt
-rw-r--r-- 1 yanakamu yn-nig   690 10月 30 18:22 2018 counts.txt.summary
```

結果ファイル

counts.txt

```
[yanakamu@nt097 20181119]$ more featurecount/counts.txt
# Program:featureCounts v1.6.2; Command:"featureCounts" "-T" "8" "-p" "-t" "exon" "-g" "gene_id" "-a" "../reference/s288c_e.gff" "-o" "../featurecount/counts.txt" "SRR453566.sorted.bam" "SRR453567.sorted.bam" "SRR453568.sorted.bam" "SRR453569.sorted.bam" "SRR453570.sorted.bam" "SRR453571.sorted.bam"
Geneid  Chr Start   End Strand  Length  SRR453566.sorted.bam  SRR453567.sorted.bam  SRR453568.sorted.bam  SRR453569.sorted.bam  SRR453570.sorted.bam  SRR453571.sorted.bam
SRR453571.sorted.bam
gene_0001  NC_001133.9 18072169-   363 0    2    6    0    0    1
gene_0002  NC_001133.9 24802707+   228 0    0    0    0    0
gene_0003  NC_001133.9 72359016-  17820 0    0    0    0    0
gene_0004  NC_001133.9 11565    11951    -   387 0    0    0    0    0    0
gene_0005  NC_001133.9 12046    12426    +   381 2    8    10   6    7    18
gene_0006  NC_001133.9 13363    13743    -   381 0    0    0    0    0    0
gene_0007  NC_001133.9 21566    21850    +   285 0    0    0    0    0    0
gene_0008  NC_001133.9 22395    22685    -   291 0    0    0    0    0    0
gene_0009  NC_001133.9 24000    27968    -  3969 32  37  33  43  63  84
```

gene_id

リファレンスid

ポジション

strand

長さ

リードカウント数

サマリファイル

counts.txt.summary

```
[yanakamu@nt097 20181119]$ more featurecount/counts.txt.summary
Status  SRR453566.sorted.bam SRR453567.sorted.bam SRR453568.sorted.bam  SRR453569.sorted.bam  SRR453570.sorted.bam  SRR453571.sorted.bam
Assigned 4568791 6257960 4527147 3077143 3873786 4907344
Unassigned_Unmapped 114233 140531 96519 181908 1935999 229888
Unassigned_MappingQuality 0 0 0 0 0 0
Unassigned_Chimera 0 0 0 0 0 0
Unassigned_FragmentLength 0 0 0 0 0 0
Unassigned_Duplicate 0 0 0 0 0 0
Unassigned_MultiMapping 814087 1126629 794069 598424 713577 914200
Unassigned_Secondary 0 0 0 0 0 0
Unassigned_Nonjunction 0 0 0 0 0 0
Unassigned_NoFeatures122392 152780 105634 112410 133839 163247
Unassigned_Overlapping_Length 0 0 0 0 0 0
Unassigned_Ambiguity 31275 37002 23914 20733 33704 33200
```

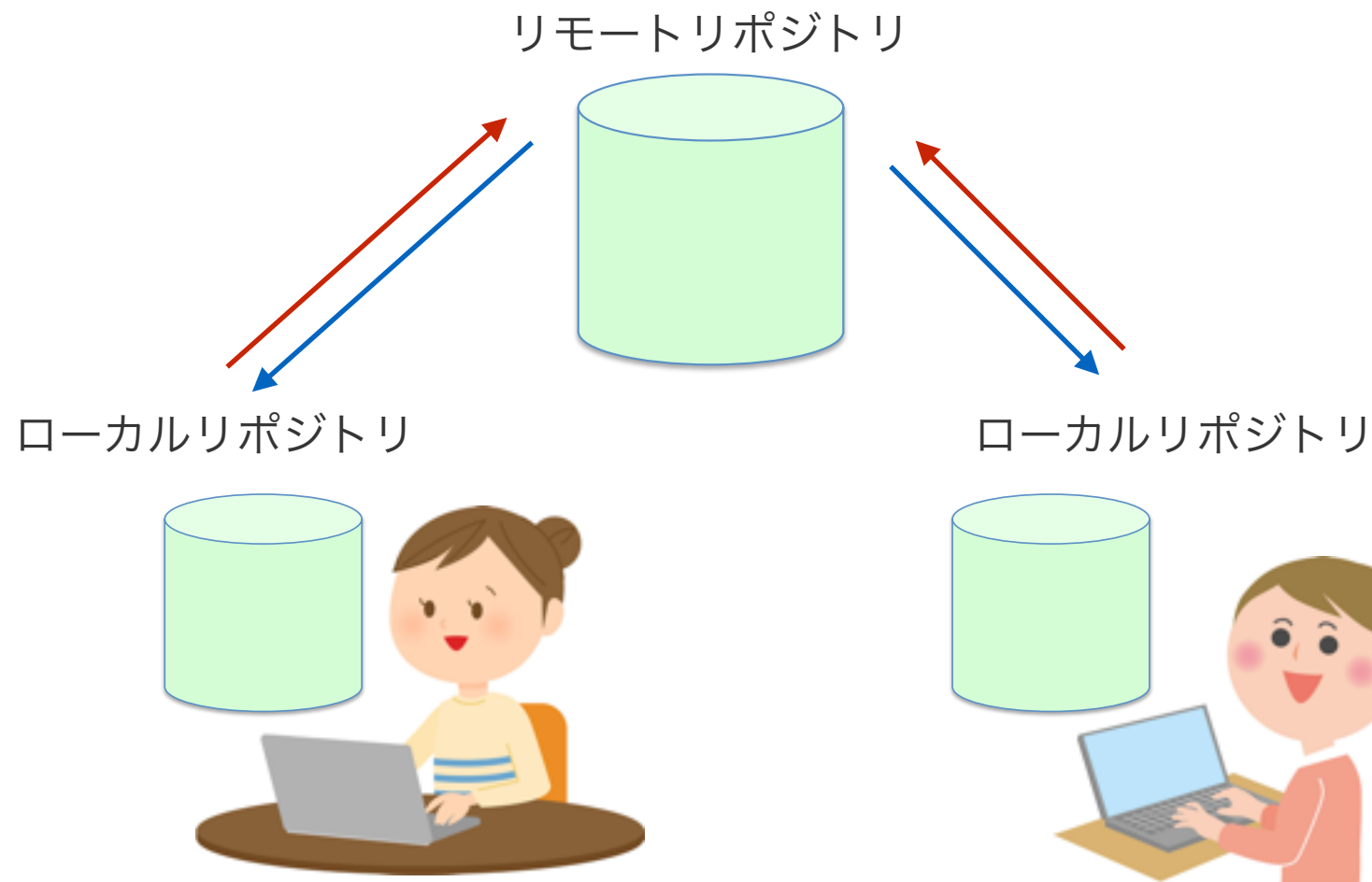
3. データダウンロード

講習データのダウンロード



Git (ギット) とは

プログラムソースなどの変更履歴を管理する分散型のバージョン管理システム



リモートリポジトリを通して、ファイルの共有、変更履歴の管理ができる。

ローカルリポジトリに、リモートリポジトリをコピーし、ローカルでも変更履歴等の管理が行える。

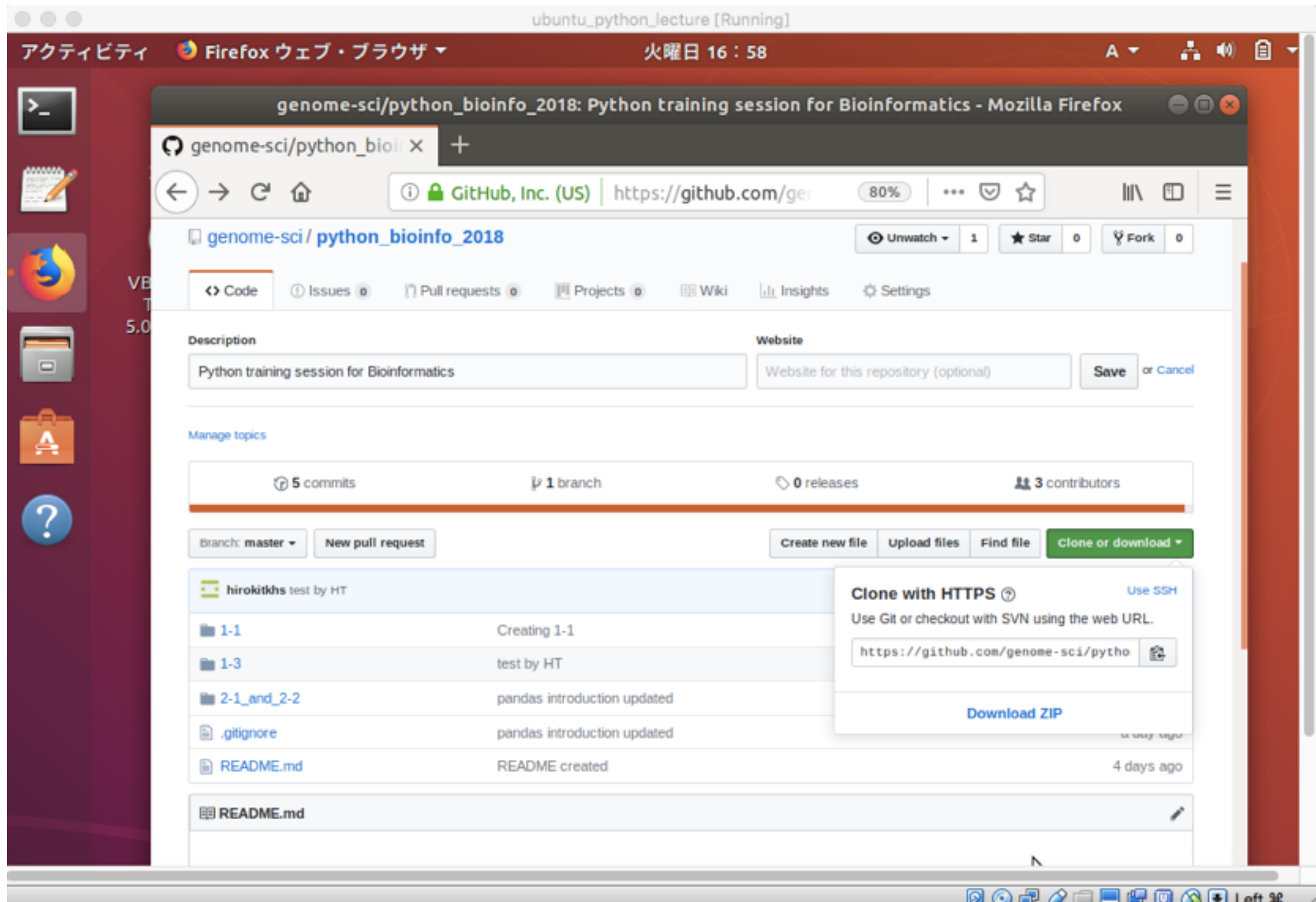
ローカルリポジトリでファイルの修正追加削除を行ったら、リモートリポジトリを更新することで、新しいバージョンを共有することができる。

GitHubとは

GitHub, Inc. が運営するプログラムソースなどのバージョン管理を行うウェブサービス。
無料アカウントでは、登録ファイルが全公開になるのでご注意ください。

VM のブラウザで GitHub にアクセス

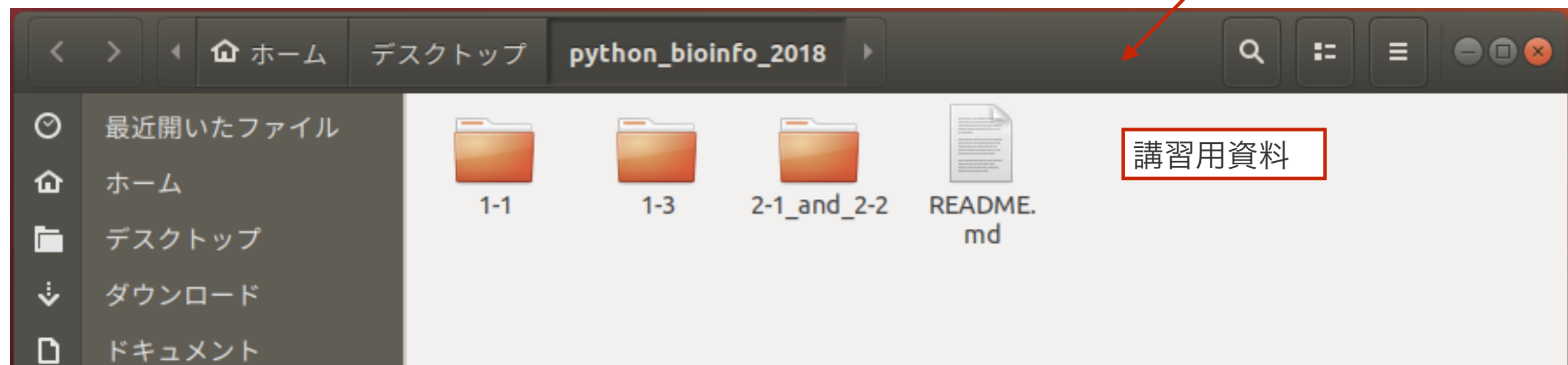
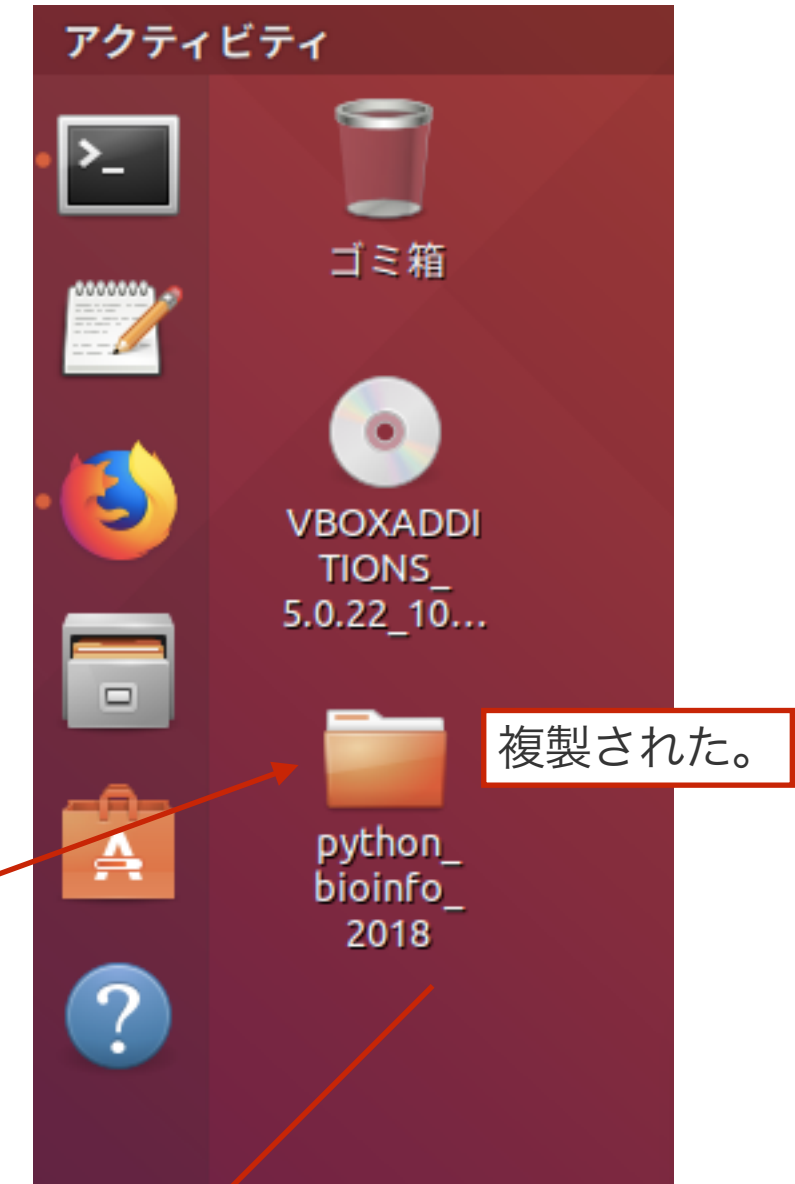
https://github.com/genome-sci/python_bioinfo_2018



リポジトリのコピー (git clone)

```
tm@VirtualBox: ~/デスクトップ
ファイル(F) 編集(E) 表示(V) 検索(S) 端末(T) ヘルプ(H)
tm@VirtualBox:~$ ls
Anaconda3-5.3.0-Linux-x86_64.sh テンプレート ビデオ 公開
anaconda3 デスクトップ ピクチャ
tm@VirtualBox:~/デスクトップ$ git clone https://github.com/genome-sci/python_bioinfo_2018.git
```

```
tm@VirtualBox: ~/デスクトップ
ファイル(F) 編集(E) 表示(V) 検索(S) 端末(T) ヘルプ(H)
tm@VirtualBox:~$ ls
Anaconda3-5.3.0-Linux-x86_64.sh テンプレート ビデオ 公開
anaconda3 デスクトップ ピクチャ
tm@VirtualBox:~/デスクトップ$ git clone https://github.com/genome-sci/python_bioinfo_2018.git
Cloning into 'python_bioinfo_2018'...
remote: Enumerating objects: 32, done.
remote: Counting objects: 100% (32/32), done.
remote: Compressing objects: 100% (25/25), done.
remote: Total 54 (delta 5), reused 32 (delta 5), pack-reused 22
Unpacking objects: 100% (54/54), done.
tm@VirtualBox:~/デスクトップ$
```



ありがとうございました。