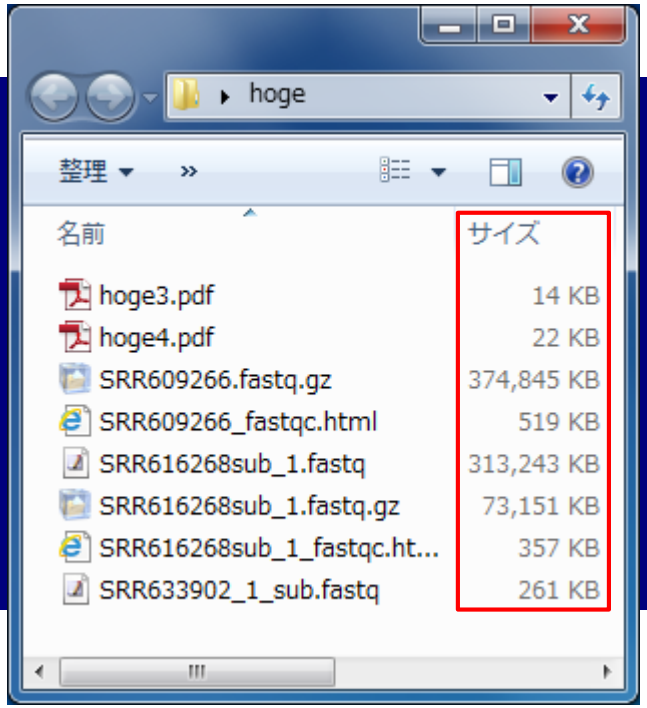
2015.06.18版

USBメモリ中のhogeフォルダをデスクトップにコピーしておいてください。コピー後のファイルサイズが同じになっているかもチェックしてください。



前回（6/9）のhogeフォルダがデスクトップに残っているかもしれないのでご注意ください。

農学生命情報科学特論I 第1回

大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム
門田幸二（かどた　こうじ）
kadota@iu.a.u-tokyo.ac.jp
http://www.iu.a.u-tokyo.ac.jp/~kadota/

講義予定

NGSの普及により、以前は主にゲノム解析系で必要とされていた配列解析のためのスキルがトランスクリプトーム解析においても要求される時代になっています。本科目では、様々な局面で応用可能な配列解析系のスキルアップを目指し、RNA-Seqに基づくトランスクリプトーム解析を題材とした講義を行います。

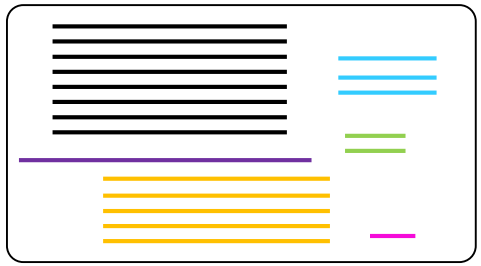
- 第1回（2015年6月16日）
  - データベース、データ取得、ファイル形式、Quality Control
  - 教科書の1.3節周辺
- 第2回（2015年6月23日）
  - Quality Control、フィルタリング、アセンブル
  - 教科書の1.3節、2.3節周辺
- 第3回（2015年6月30日）
  - マッピング、カウント情報取得、クラスタリング
  - 教科書の2.3節周辺
- 第4回（2015年7月7日）
  - データ正規化、実験デザイン、分布（モデル）、発現変動解析
  - 教科書の3.3節周辺

光刺激前(T1)と光刺激後(T2)の状態の数値データを比較して、サンプル(状態)間で発現変動遺伝子(DEG)を同定。

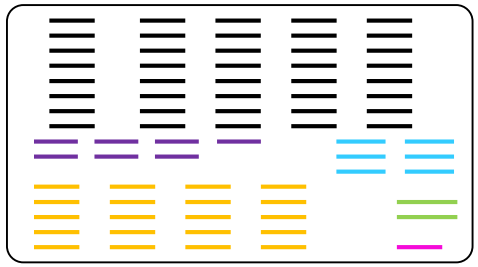
RNA-seq

- 次世代シーケンサー(Illuminaの場合)

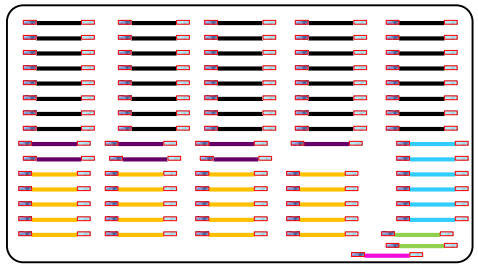
光刺激前(T1)の目のトランスクリプトーム



数百塩基程度に断片化



アダプター配列を両末端に付加



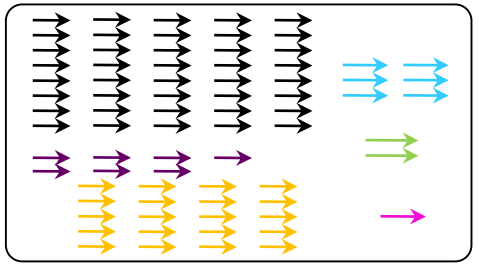
配列決定

・ペアードエンド法

断片配列の両末端が数百塩基以内の対の2種類の配列が得られる

・シングルエンド法

シングルエンド法の場合



Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

Illumina, PacBio, …

様々なNGSプラットフォーム

- イントロ | 一般 | 配列取得 | プロモーター配列 | GenomicFeatures(Lawrence 2013)(last modified 2015/03/04)
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | 公共DBから(last modified 2015/05/09)
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | GenomicFeatures(Lawrence 2013)(last modified 2015/03/02)
- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009)(last modified 2015/02/20)
- イントロ | NGS | 様々なプラットフォーム (last modified 2015/06/13) NEW
- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2014/06/25)
- イントロ | NGS | 配
- イントロ | NGS | 配
- イントロ | NGS | 配
- イントロ | NGS | 配
- イントロ | NGS | フ
- イントロ | NGS | フ
- イントロ | NGS | フ
- イントロ | NGS | フ
- イントロ | NGS | フ

イントロ | NGS | 様々なプラットフォーム NEW

NGS機器(プラットフォーム)もいくつかあります:

- 会社名:製品名
- Illumina: MiSeq, NextSeq 500, HiSeq 3000/4000, HiSeq X Five/Ten, ...
- Pacific Biosciences: PacBio RS II
- Life Technologies: Ion PGM System
- ...
- Roche: GS FLX+ System, GS Junior+ System
- Life Technologies: SOLiD
- ...

Pacific Biosciences (PacBio)について:

PacBio RS II Systemは最長で40,000bp以上(平均でも10,00 bp以上)読めるようですが配列のqualityが若干(85%程度)劣るようです。しかしエラーの入る場所がランダムなようで多数決ルール(majority rule)でエラー補正がかなりうまくいくらしいです(コンセンサス配列の精度は99.999%)。このロングリードでトランスクリプトーム配列決定(新規アイソフォームの発見)をヒト(Sharon et al., 2013)やニワトリ(Thomas et al., 2014)で行った論文などが出始めています。Smart-seq2の実験手順(Picelli et al., Nat Protoc., 2014)なども出ているようですね。葉緑体ゲノムでのアセンブリ性能評価(Ferrarini et al., BMC Genomics, 2013)もなされています。

実際のデータ

PacBioのロングリードデータを眺めてみよう。この生の塩基配列データは日米欧三極のデータベース（DB）から取得可能であるが、ここでは②SRP038897を③「NCBI SRA」というDBで眺める。Let's look SRP038897 by NCBI SRA.

- インストール | Rパッケージ | 個別 (last modified 2015/06/10) NEW
- (削除予定)Rのインストールと起動 (last modified 2015/04/02)
- (削除予定)個別パッケージのインストール (last modified 2015/02/20)
- 基本的な利用法 (last modified 2015/04/03)
- サンプルデータ (last modified 2015/02/15)
- バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | 速習コース (la
- 書籍 | トランスクリプトーム解析 | について (last modified 2014/05/12)
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト

①

サンプルデータ

1. Illumina/36bp/single-end/human (SRA000299) data (Marioni et al., Genome Res., 2008)
   「Kid
   プル
   のも
   デー

   7列
   8列
   9列
   10
   11
   18
   16
   17
   18
   19
   20

2. Illun
   Supp

28. Illumina GAIIx/76bp/paired-end/Drosophila or Illumina HiSeq 2000/100bp/paired-end/Drosophila (SRP009459; GSE33905) data (Graveley et al., Nature, 2011; Brown et al., Nature, 2014)
    ショウジョウバエの様々な組織のデータ(modENCODE)。29 dissected tissue samplesのstrand-specific, paired-endのbiological replicates (duplicates)だそうです。
29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (SRP011435; GSE36469) data (Huang et al., Development, 2012)
    シロイヌナズナの2群間比較用データ: 4 DEX-treated vs. 4 mock-treated
30. PacBio/xxx bp/Human (ERP003225) data (Sharon et al., Nat Biotechnol., 2013)
    ヒトの長鎖RNA-seqデータです。配列長はリードによって異なります。

②

31. PacBio/xxx bp/Chicken (SRP038897 by DRA; SRP038897 by ENA; SRP038897 by SRA) data (Sharon et al., PLoS One, 2014)
    ニワトリの長鎖RNA-seqデータです。配列長はリードによって異なります。
32. k-mer解析用のランダム配列から生成したFASTA形式ファイル(sample32_ref.fastaとsample32_ngs.fasta)です。50塩基の長さのリファレンス配列を生成したのち、20塩基長の部分配列を10リード分だけランダム抽出したものです。塩基の存在比はAが22%, Cが28%, Gが28%, Tが22%にしています。リファレンス配列(仮想ゲノム配列)がsample32_ref.fastaで、10リードからなる仮想NGSデータがsample32_ngs.fastaです。リード長20塩基で10リードなのでトータル200塩基となり、50塩基からなる元のゲノム配列の4倍シーケンスしていることになります(4X coverageに相当)。イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成からと基本的に同じです。

③

```
out_f1 <- "sample32_ref.fasta"          #出力ファイル名を指定してout_f1に格納
```

実際のデータ

SRP038897を眺めているつもりなのに、どんどん知らないIDになっていることにはとりあえず目をつぶって、SRR1177086をクリック。
Click SRR177086.

- サンプルデータ

実際のデータ

NCBI Site map | All databases | Search

Sequence Read Archive

Main | Browse | Search | Download | Submit | Documentation | Software | Trace Archive | Trace Assembly

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

Chicken embryonic heart transcriptome sequencing (SRR1177086)

Metadata | Reads | Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR1177086	1.8M	2.0Gbp	510.0M	48.1%	2014-02-26	public

Quality graph (bigger)

This run has 1 read per spot:

L=1075, σ=685.6, 100%

Legend

Experiment	Library					
SRX475467	Name	Platform	Strategy	Source	Selection	Layout
to BLAST	PacBio embryonic chicken heart cDNA	PacBio SMRT™	OTHER	TRANSCRIPTOMIC	cDNA	SINGLE

Show design

Biosample	Sample Description	Organism	Links
SAMN02650959 (SRS561227)		Gallus gallus	• Model organism or animal sample for Gallus gallus • PRJNA239269 [Gallus gallus]

Bi...	SRA Study	Title
PRJNA239269	SRP038897	Long-read sequencing of chicken transcripts and identification of new transcript isoforms.

Show abstract

②

①

①ここで、再びオリジナルのIDであるSRP038897が見えるようになる。SRPから始まるIDは（最初のとっかりとしては）このデータセットにアクセスするための大元のIDのようなものだと思っておけばよい。②PRJというIDも同じような位置づけという理解で基本的に構わない。You can see the origilan ID (SRP038897) here, but click PRJNA239269.

実際のデータ

論文は、ある研究プロジェクトの成果物という捉え方をすればPRJというIDの位置づけが理解しやすいかも。
Experience is the best teacher!

NCBI Resources How To

BioProject BioProject Search
Advanced Help

Display Settings: Send to:

Gallus gallus (chicken)

Accession: PRJNA239269 ID: 239269

Long-read sequencing of chicken transcripts and identification of new transcript isoforms.

The chicken has long served as an important model organism in many fields, and continues to aid our understanding of animal development. More...

See Genome Information for Gallus gallus

Project Type: Transcriptome or Gene expression

Attributes: Scope: Multiisolate; Material: Transcriptome; Capture type: Sequencing

Relevance: Model Organism

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	
PUBLICATIONS	
PubMed	
PMC	
OTHER DATASETS	
BioSample	

Related information
- BioSample
- Full text in PMC
- Genome
- PubMed

SRA Data Details

Parameter	Value
Data volume, Gbases	2
Data volume, Mbytes	510

Publications:

1. Thomas S *et al.*, "Long-read sequencing of chicken transcripts and identification of new transcript isoforms.", *PLoS One*, 2014 Apr 15;9(4):e94650

Lineage: *Eukaryota*; *Metazoa*; *Chordata*; *Craniata*; *Vertebrata*; *Euteleostomi*; *Archelosauria*; *Archosauria*; *Dinosauria*; *Saurischia*; *Theropoda*; *Coelurosauria*; *Aves*; *Neognathae*; *Galloanserae*; *Galliformes*; *Phasianidae*; *Phasianinae*; *Gallus*; *Gallus gallus* [Taxonomy ID: 9031]

Grants:

"Bench to Bassinet Program" (Grant ID U01-HL098188, National Heart Lung and BLood Institute)

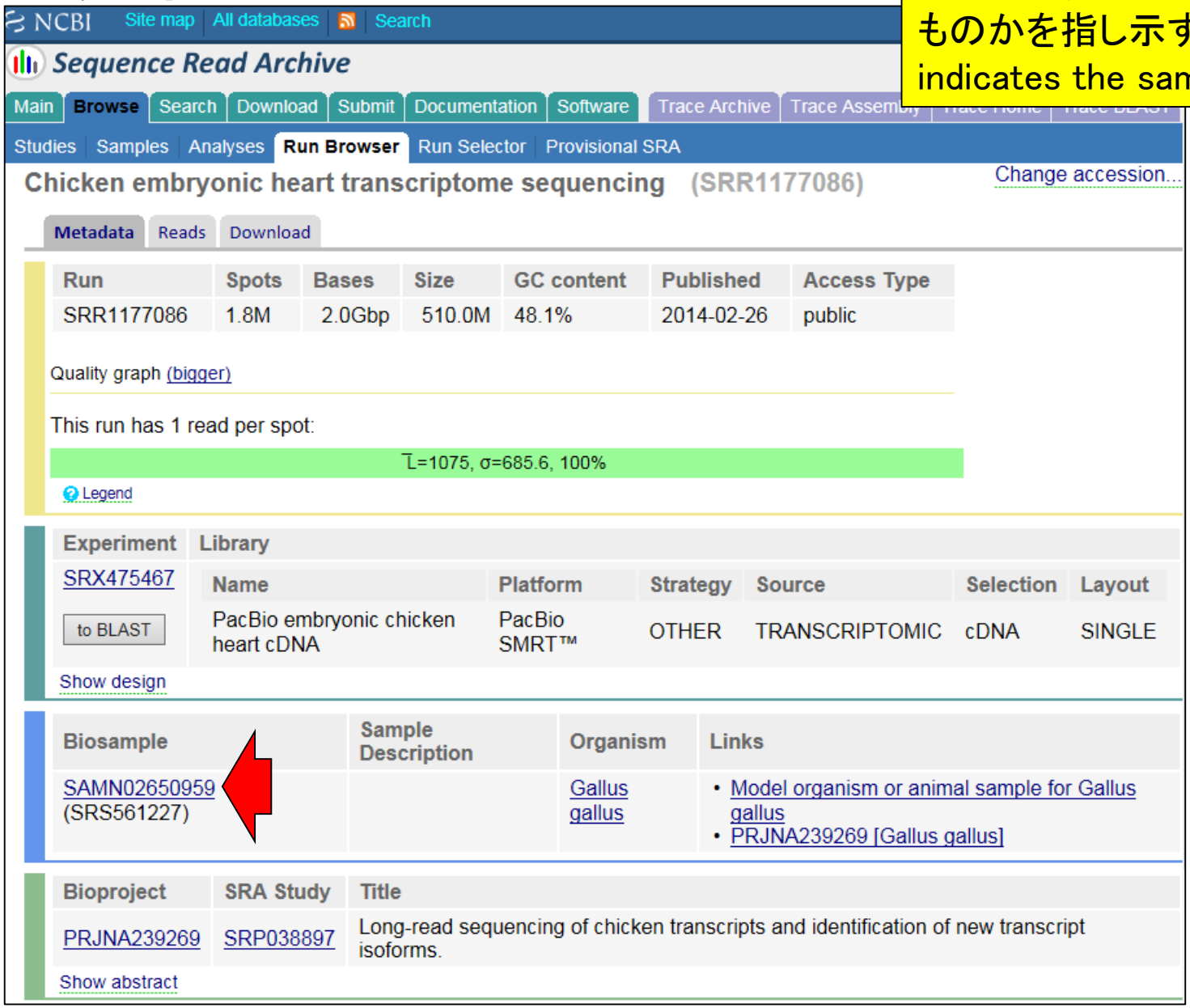
Submission:

Registration date: 24-Feb-2014
Gladstone Institutes

- サンプルデータ

実際のデータ

これは、ニワトリ（Gallus gallusの心臓サンプル）からとられたRNA-seqデータ。③SRSやSAMNは、どのようなサンプルから取得されたものかを指し示すID。The SAMN02650959 indicates the sample ID for chiken (of heart).



実際のデータ

これは、ニワトリ（Gallus gallusの心臓サンプル）からとられたRNA-seqデータ。③SRSやSAMNは、どのようなサンプルから取得されたものかを指し示すID。

Model organism or animal sample for Gallus gallus

Identifiers	BioSample: SAMN02650959; Sample name: Embryonic chicken heart; SRA: SRS561227
Organism	Gallus gallus (chicken) cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Sauropsida; Sauria; Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes; Phasianidae; Phasianinae; Gallus
Package	Model organism or animal; version 1.0
Attributes	**breed** White Leghorn **age** embryo **biomaterial provider** Bruneau lab, Gladstone Institutes **sex** not determined **tissue** heart
BioProject	PRJNA239269 Gallus gallus Retrieve all samples from this project
Submission	Gladstone Institutes, Sean Thomas; 2014-02-24

Accession: SAMN02650959 ID: 2650959
BioProject SRA

実際のデータ

①今見ているものは、SRR1177086というIDのメタデータ。SRRは、あるサンプルについて一回の実験（Run）で得られたもの、というイメージ。ラン（Run）ごとのIDという意味。②アレイデータのサンプル（アレイ）ごとにつけられるGSM IDと同じようなものである。The SRR ID is similar to GSM ID when you learn in Functional Genomics (機能ゲノム学).

NCBI Site map | All databases | Search

Sequence Read Archive

Main | Browse | Search | Download | Submit | Documentation | Software | Trace Archive | Trace Assembly

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

① embryonic heart transcriptome sequencing (SRR1177086)

Metadata | Reads | Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR1177086	1.8M	2.0Gbp	510.0M	48.1%	2014-02-26	public

Quality graph (bigger)

This run has 1 read per spot:

L=1075, σ=685.6, 100%

Legend

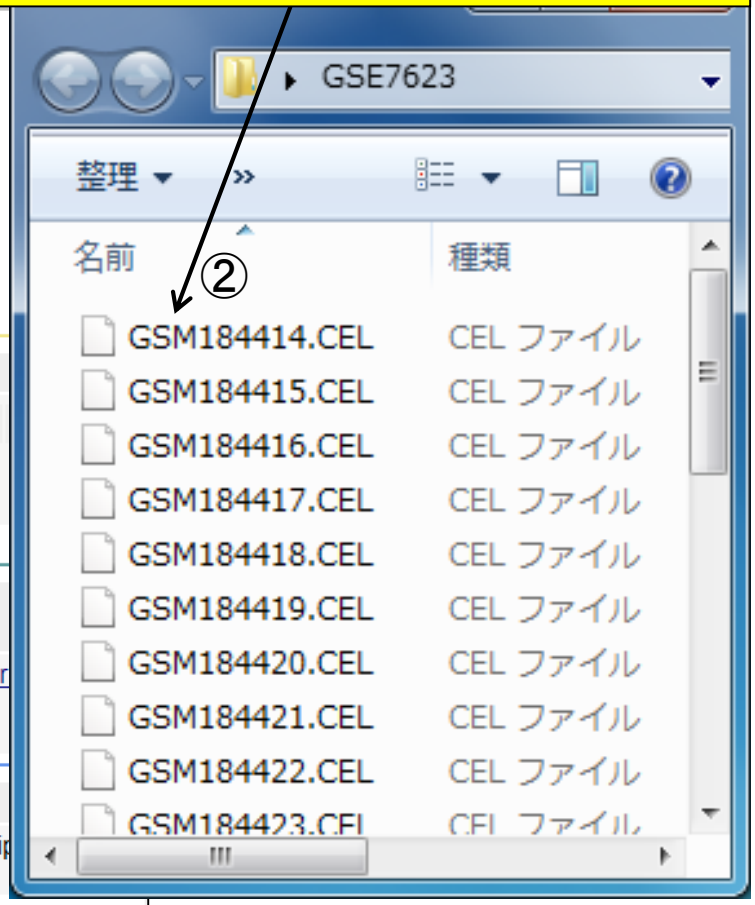
Experiment	Library				
SRX475467	Name	Platform	Strategy	Source	Selection
to BLAST	PacBio embryonic chicken heart cDNA	PacBio SMRT™	OTHER	TRANSCRIPTOMIC	cDNA

Show design

Biosample	Sample Description	Organism	Links
SAMN02650959 (SRS561227)		Gallus gallus	• Model organism or animal sample for gallus • PRJNA239269 [Gallus gallus]

Bioproject	SRA Study	Title
PRJNA239269	SRP038897	Long-read sequencing of chicken transcripts and identification of new transcrip isoforms.

Show abstract



メタデータ

実データを見せると言いながら、説明していたのはメタデータに関する事柄。日米欧三極のDB（SRA, ENA, DRA）で見栄えが異なるが、最低限SRP038897のような大元のIDを頼りにSRRというIDを探すことで、目的のリード情報を眺めたりダウンロードできる。総リード数は、約185万リードであることも分かる。
The understanding of metadata information is important. The number (1,849,778) by black arrow indicates the number of reads in this SRR ID.

NCBI Resources How To

SRA SRA SRP038897 Save search Advanced

Display Settings: Full

SRX475467: Chicken embryonic heart transcript
1 PACBIO_SMRT (PacBio RS) run: 1.8M spots, 2G

Accession: SRX475467
Experiment design: Collecting embryonic chicken
then incubated in Genesis Hova-Bators until tissue
stages 18-20, 25, and 32 and immediately flash fro
TriZol (Life Technologies) by repipetting, and then t
with isopropanol, washed with 75% ethanol and the
cDNA synthesis mRNA were purified using the Stra
hybridized to oligo-dT magnetic beads, separated f
resuspended into the kit's elution buffer. First strand
(Clontech): The first cDNA strand was synthesized
Reverse Transcriptase (Clontech), the CDS III oligo
order to add a consistent 5' site for LD-PCR amplifi
primer: 5'-ATTCTAGAGGCCGAGGCGGCCGACA
SMART IV oligonucleotide: 5'-AAGCAGTGGTATC
AAGCAGTGGTATCAACGCAGAGT-3' Library prep
and four separate size ranges were fractionated: 0-
extracted from the gel and SMRTbell libraries were
(Pacific Biosciences): The cDNA was cleaned using

adapters were then ligated to
SMRTbell templates with cDN
sequentially annealed to the SMRTbell templates using the DNA Polymerase Binding kit (Pacific Biosciences). The MagBead loading kit was used to load annealed templates onto a Pacific Biosciences RS II sequencer, and sequencing was performed for each template library using the DNA Sequencing kit (Pacific Biosciences). Sequences containing both 5' and 3' adapters were identified, and the adapters and poly-A/T sequences were trimmed. The resulting sub-reads were then mapped using GMAP (21) (2012-07-20 release, default settings) to the galGal4 genome assembly (11) .
Submitted by: Gladstone Institute
Study summary: SRP038897 • Long-read sequencing of chicken transcripts and identification of new transcript isoforms. • PRJNA239269 • All experiments • Run Selector (more...)
Sample: SAMN02650959 • Model organism or animal sample for Gallus gallus (more...)
Library: PacBio embryonic chicken heart cDNA (more...)
Platform: PacBio SMRT™ (more...)
Spot descriptor:

1 forward

Total: 1 run, 1.8M spots, 2G bases, 486.4Mb

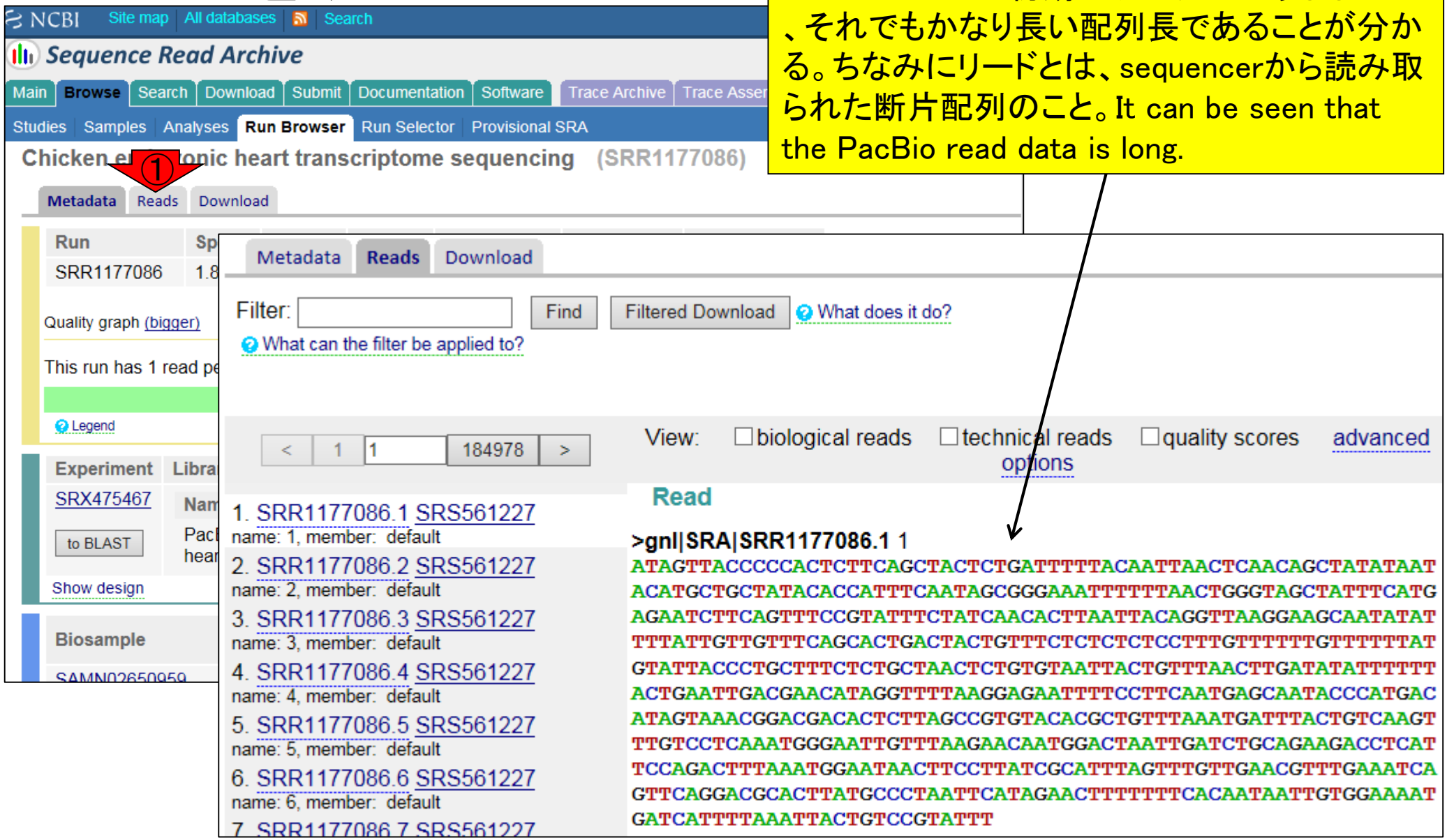
#	Run	# of Spots	# of Bases	Size	Published
1.	SRR1177086	1,849,778	2G	486.4Mb	2014-02-26

ID: 655957

- サンプルデータ

リードを眺める

①Readsタブをクリックすると、全部で1,849,778リードの塩基配列情報を眺めることができる。これはPacBioの初期のデータではあるものの、それでもかなり長い配列長であることが分かる。ちなみにリードとは、sequencerから読み取られた断片配列のこと。It can be seen that the PacBio read data is long.



リードを眺める

クオリティスコアも表示させたい場合。かつては様々なクオリティスコアの流派が存在したが、公共DB上で見られるものは基本的にSanger Quality Scoreというものに統一されている。数値が大きいほどクオリティが高いことを示す。概ね0-60程度の範囲。You can see the quality scores by checking?!,,,here.

Metadata | Reads | Download

Filter: Find | Filtered Download | What does it do?

What can the filter be applied to?

< 1 1 184978 >

View: ☐ biological reads ☐ technical reads ☑ quality scores advanced options

1. SRR1177086.1 SRS561227
name: 1, member: default
2. SRR1177086.2 SRS561227
name: 2, member: default
3. SRR1177086.3 SRS561227
name: 3, member: default
4. SRR1177086.4 SRS561227
name: 4, member: default
5. SRR1177086.5 SRS561227
name: 5, member: default
6. SRR1177086.6 SRS561227
name: 6, member: default
7. SRR1177086.7 SRS561227
name: 7, member: default
8. SRR1177086.8 SRS561227
name: 8, member: default
9. SRR1177086.9 SRS561227
name: 9, member: default
10. SRR1177086.10 SRS561227
name: 10, member: default

Read

>gnl|SRA|SRR1177086.1 1

ATAGTTACCCCCACTCTTCAGCTACTCTGATTTTTACAATTAACTCAACAGCTATATAAT
ACATGCTGCTATACACCATTTCAATAGCGGGAAATTTTTTAACTGGGTAGCTATTTCATG
AGAATCTTCAGTTTCCGTATTTCTATCAACACTTAATTACAGGTTAAGGAAGCAATATAT
TTTATTGTTGTTTCAGCACTGACTACTGTTTCTCTCTCCTTTGTTTTTTGTTTTTTAT
GTATTACCCTGCTTTCTCTGCTAACTCTGTGTAATTACTGTTTAACTTGATATATTTTTT
ACTGAATTGACGAACATAGGTTTTAAGGAGAATTTTCCTTCAATGAGCAATACCCATGAC
ATAGTAAACGGACGACACTCTTAGCCGTGTACACGCTGTTTAAATGATTTACTGTCAAGT
TTGTCCTCAAATGGGAATTGTTTAAGAACAATGGACTAATTGATCTGCAGAAGACCTCAT
TCCAGACTTTAAATGGAATAACTTCCTTATCGCATTTAGTTTGTTGAACGTTTGAAATCA
GTTCAGGACGCACTTATGCCCTAATTCATAGAACTTTTTTTCACAATAATTGTGGAAAAT
GATCATTTTAAATTACTGTCCGTATTT

One channel quality score

61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61

- サンプルデータ

リードを眺める

リードの塩基配列を表示するのに比べて、クオリティスコアを表示するために、かなり大きなスペースを必要とすることがわかる。これはファイルサイズの増加に直結するため、少ない文字数（1文字）でクオリティスコアを表現する必要性が理解できる。NGSデータ解析の事実上の出発地点であるFASTQファイルの形式は、この背景が理解できていればあとは楽勝。Conventional way of displaying quality scores (i.e., number, space, number, …) is resource-hungry. The format of FASTQ is designed for reducing the file size.

name: 1, member: default
2. SRR1177086.2 SRS561227
name: 2, member: default
3. SRR1177086.3 SRS561227
name: 3, member: default
4. SRR1177086.4 SRS561227
name: 4, member: default
5. SRR1177086.5 SRS561227
name: 5, member: default
6. SRR1177086.6 SRS561227
name: 6, member: default
7. SRR1177086.7 SRS561227
name: 7, member: default
8. SRR1177086.8 SRS561227
name: 8, member: default
9. SRR1177086.9 SRS561227
name: 9, member: default
10. SRR1177086.10 SRS561227
name: 10, member: default

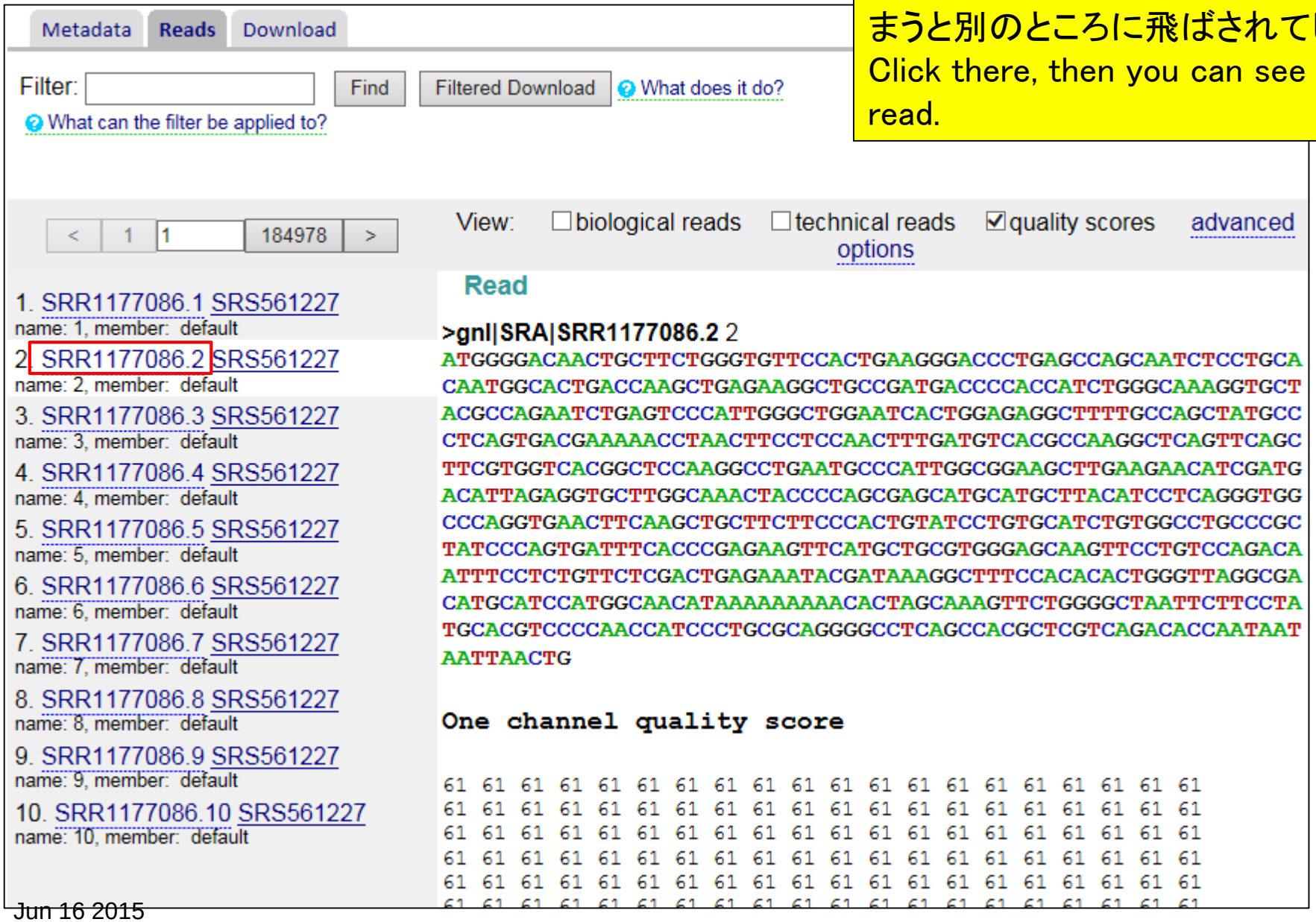
```
>gnl|SRA|SRR1177086.1 1
ATAGTTACCCCCACTCTTCAGCTACTCTGATTTTTACAAT
ACATGCTGCTATACACCATTTCAATAGCGGGAAATTTTTT
AGAATCTTCAGTTTCCGTATTTCTATCAACACTTAATTA
TTTATTGTTGTTTCAGCACTGACTACTGTTTCTCTCTCT
GTATTACCCTGCTTTCTCTGCTAACTCTGTGTAATTACT
ACTGAATTGACGAACATAGGTTTTAAGGAGAATTTTCCTT
ATAGTAAACGGACGACACTCTTAGCCGTGTACACGCTGTT
TTGTCCTCAAATGGGAATTGTTTAAGAACAATGGACTAAT
TCCAGACTTTAAATGGAATAACTTCCTTATCGCATTTAGT
GTTCAGGACGCACTTATGCCCTAATTCATAGAACTTTTTT
GATCATTTTAAATTACTGTCCGTATTT


One channel quality score

61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61  61
```

リードを眺める

2番目のリードを眺める。赤枠のあたりをクリックしないとみられないので注意！赤枠右側の「SRS561227」のあたりをクリックしてしまうと別のところに飛ばされてしまいます。
Click there, then you can see the second read.



リードを眺める

3番目のリードを眺める。PacBioのデータはリードごとに配列長が異なることが分かります。

Metadata | Reads | Download

Filter: Find | Filtered Download | What does it do?

What can the filter be applied to?

< 1 1 184978 >

View: biological reads | technical reads | quality scores | advanced options

1. SRR1177086.1 SRS561227
name: 1, member: default
2. SRR1177086.2 SRS561227
name: 2, member: default
3 SRR1177086.3 SRS561227
name: 3, member: default
4. SRR1177086.4 SRS561227
name: 4, member: default
5. SRR1177086.5 SRS561227
name: 5, member: default
6. SRR1177086.6 SRS561227
name: 6, member: default
7. SRR1177086.7 SRS561227
name: 7, member: default
8. SRR1177086.8 SRS561227
name: 8, member: default
9. SRR1177086.9 SRS561227
name: 9, member: default
10. SRR1177086.10 SRS561227
name: 10, member: default

Read

>gnl|SRA|SRR1177086.3 3

CAGTGGGGGACAACCTGCTCTGGGTGTTCACTGAAGGGAGCCTGACCAGCACTCTCCTGC
ACAATGGCAGCGACACAAAGCTGAGAAGGCTGCCGGACCACCATCTGGGCAAAGGTGGCT
ACCCAGATTGAGTCCATTGGGCTGGAATCACGTGGAGAGGCTTTGCCAAGCTATCTCAAC
GAAAACCCTCTTCCCCTCCTTTGAGTCAGCCATAGGCTCAGTTCAGCTTCGCTGGTCACG
GCTCCAAGGTCCTGATATGCCATTGGGGTAAGGCTGTGAGAACATCGATGACATTAGAAG
GTGCTTTTGGCCCAAACTCAGCGAGCTTGTATGCAGTTACATCCTTCAGGGGTGGTACTC
CAGGAAACTCAAGACTGGCTTTCACACTGGTATCCTGTGCTCTGTTGGCTGCCCGCTATC
CCAGTGATTTCATTCCCAGTAATGTTCTATGCTGCCGTGATGGACAAGTTCCTGTCCAGT
GGCATTTCCTCTGTTATGACTGAGAAATATCATGATAAATGGCGTTCCACACTGGGTTTT
AGGACCATGCATCTCATGGCACACAACAGCTGCCAAGTTTTCTGGGTATGTCTTCCTATG
CAGTCCCCCCAACTCCCCTCGCAGGGGCTCAGCCACCTTGCAGACCAACCAATTAAATAT
TCAATCTGTGAATAAAAAGAAAACAAAGAAAAA

One channel quality score

61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61

リードを眺める

10番目のリードを眺める。PacBioのデータは
リードごとに配列長が異なることが分かります

Metadata | Reads | Download

Filter: Find | Filtered Download | What does it do?

What can the filter be applied to?

< 1 1 184978 >

View: ☐biological reads ☐technical reads ☑quality scores advanced options

1. SRR1177086.1 SRS561227
name: 1, member: default
2. SRR1177086.2 SRS561227
name: 2, member: default
3. SRR1177086.3 SRS561227
name: 3, member: default
4. SRR1177086.4 SRS561227
name: 4, member: default
5. SRR1177086.5 SRS561227
name: 5, member: default
6. SRR1177086.6 SRS561227
name: 6, member: default
7. SRR1177086.7 SRS561227
name: 7, member: default
8. SRR1177086.8 SRS561227
name: 8, member: default
9. SRR1177086.9 SRS561227
name: 9, member: default
10 SRR1177086.10 SRS561227
name: 10, member: default

Read

>gnl|SRA|SRR1177086.10 10
ATTGGGGCCCTTTCGGCTTCTGGCTGCCCAAGGATTGGCGCCCAAGGCCGAAGAAGGGAG
GCCGGGTGCCAGCCGAAGACGGGAGGCAAGGGAAGGGCGCTGGAAGGCCAAGAAGGCGTT
CCTGAAGGAGCCCAGCCACAAGAAGAAGAAGATCCGCACATCACCCACCTTCGGGGAGGC
CCAAGACCCTGCGCCTGCGCGGCAGCCAAATCCCCCGGCAAGAGCGCCCCAGGAGGAACA
AGCTTGGGACCCCATATGCCATTTCAAGTTCCCTTTGACCACAGAATCTGTCAATGAAGA
AGAATAGAGGAGTAACAATACTCTGGGTTTTCATTGTTGATGTCAAGGCAAACAAGCACC
AGATCAAAACAGGCAGTACAAGAAGCTGTATGATATTTGATGGTGGCCAAGGGTCAACAC
CTTAAGTAAGGCCCTGATGGGGAGAAGAAGGTTACGCTCCGACTGGCCTGCCTGACTACG
ATGCGTTGATGTAGCCAACAGATTGGAATCATCTAAACTTGCATCTGCCAAGGACTGTAC
GAAGATAATAAACCACTGTG

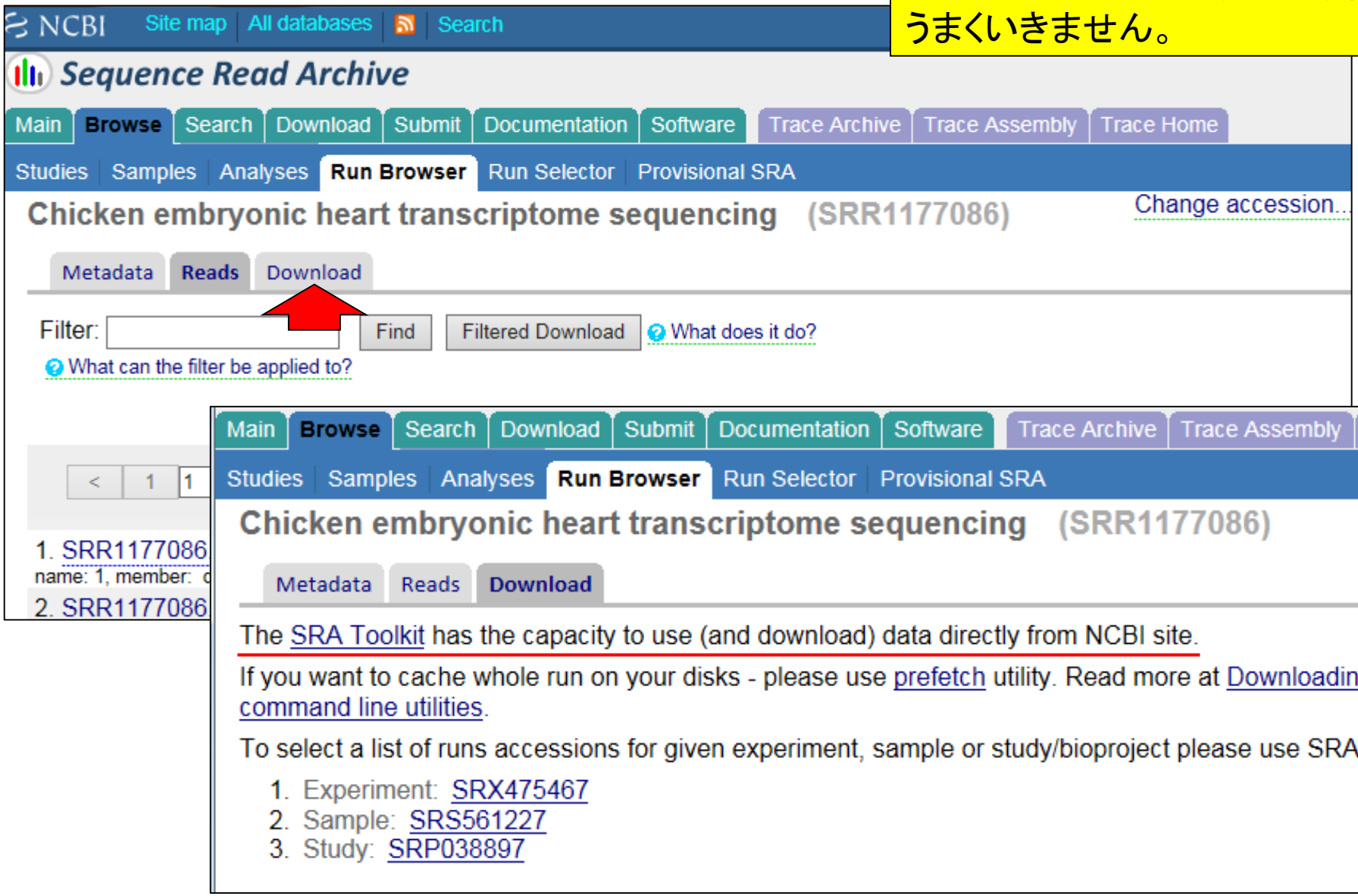
One channel quality score

61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61

- サンプルデータ

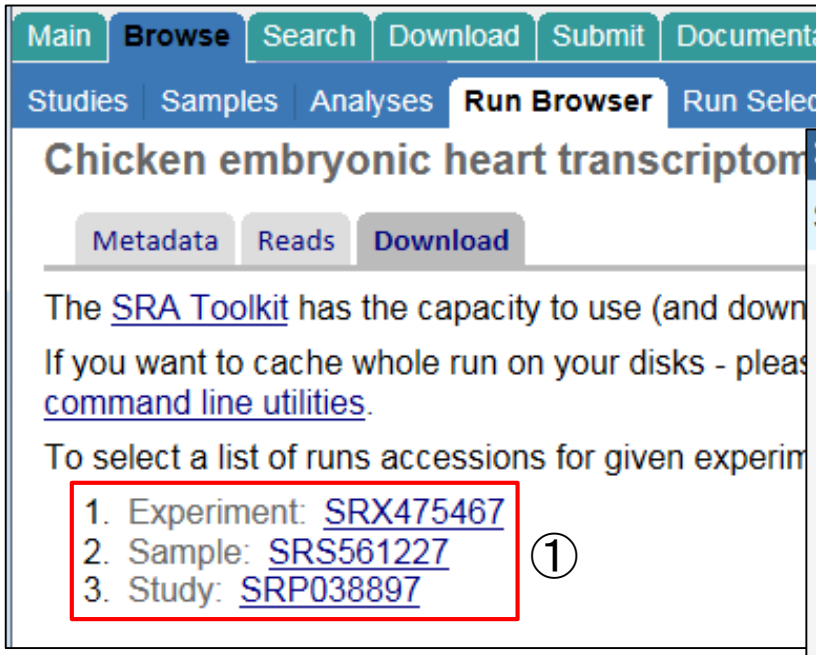
ダウンロード

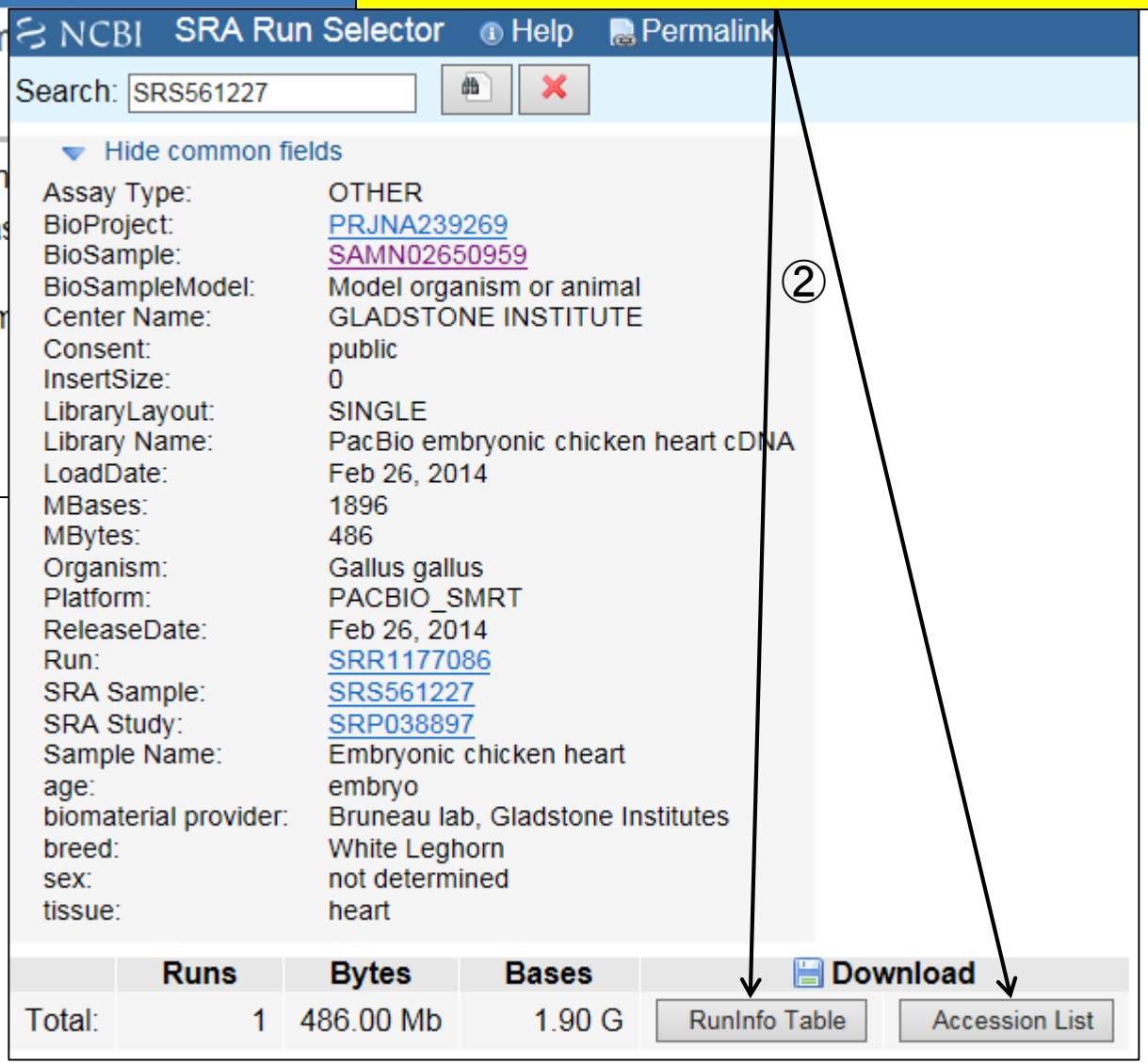
SRR1177086の生データをダウンロードしたいときは、Downloadタブをクリックすればいいのではと思いますが…。結論としてはうまくいきません。



ダウンロード

①赤枠のどのリンク先をクリックしても、右下のような感じになります。そして、②「RunInfo Table」と「Accession List」のいずれもただのテキストファイルであり、目的の生リード情報が含まれるものではない（爆）





ダウンロード

ダウンロードはここから可能。You can download here.
Do NOT click here.

NCBI Resources How To Sign in to NCBI

SRA SRA SRP038897 Search

Save search Advanced Help

Display Settings: Full Send to:

Related information

SRX475467: Chicken embryonic heart transcript
1 PACBIO_SMRT (PacBio RS) run: 1.8M spots, 2G

Accession: SRX475467
Experiment design: Collecting embryonic chicken
then incubated in Genesis Hova-Bators until tissue
stages 18-20, 25, and 32 and immediately flash fro
TriZol (Life Technologies) by repipetting, and then t
with isopropanol, washed with 75% ethanol and the
cDNA synthesis mRNA were purified using the Stra
hybridized to oligo-dT magnetic beads, separated f
resuspended into the kit's elution buffer. First strand
(Clontech): The first cDNA strand was synthesized
Reverse Transcriptase (Clontech), the CDS III oligo
order to add a consistent 5' site for LD-PCR amplifi
primer: 5'-ATTCTAGAGGCCGAGGCGGCCGACA
SMART IV oligonucleotide: 5'-AAGCAGTGGTATC
AAGCAGTGGTATCAACGCAGAGT-3' Library prep
and four separate size ranges were fractionated: 0-
extracted from the gel and SMRTbell libraries were
(Pacific Biosciences): The cDNA was cleaned using

adapters were then ligated to the insert cDNA, exonucleases were added to remove failed ligation products, and SMRTbell templates with cDNA inserts were purified. The sequencing primer and then the polymerase were then sequentially annealed to the SMRTbell templates using the DNA Polymerase Binding kit (Pacific Biosciences). The MagBead loading kit was used to load annealed templates onto a Pacific Biosciences RS II sequencer, and sequencing was performed for each template library using the DNA Sequencing kit (Pacific Biosciences). Sequences containing both 5' and 3' adapters were identified, and the adapters and poly-A/T sequences were trimmed. The resulting sub-reads were then mapped using GMAP (21) (2012-07-20 release, default settings) to the galGal4 genome assembly (11) .

Submitted by: Gladstone Institute
Study summary: SRP038897 • Long-read sequencing of chicken transcripts and identification of new transcript isoforms. • PRJNA239269 • All experiments • Run Selector (more...)
Sample: SAMN02650959 • Model organism or animal sample for Gallus gallus (more...)
Library: PacBio embryonic chicken heart cDNA (more...)
Platform: PacBio SMRT™ (more...)
Spot descriptor:

1 forward

Total: 1 run, 1.8M spots, 2G bases, 486.4Mb

#	Run	# of Spots	# of Bases	Size	Published
1.	SRR1177086	1,849,778	2G	486.4Mb	2014-02-26

ID: 655957

ダウンロード

①ダウンロードはここから可能。②NCBI SRAからダウンロードできる生データの形式は.sraというものです。PacBioの初期の?!データは、総リード数が約185万、③ファイルサイズも約510MBですので、通常の有線LAN環境ではそれほど苦も無くダウンロード可能です。

Sequences containing both 5' and 3' adapters were identified, and the ad
trimmed. The resulting sub-reads were then mapped using GMAP (21) (2
the galGal4 genome assembly (11) .

Submitted by: Gladstone Institute

Study summary: SRP038897 • Long-read sequencing of chicken transcripts and identification of new transcript isoforms. • PRJNA239269 • All experiments • Run Selector (more...)

Sample: SAMN02650959 • Model organism or animal sample for Gallus gallus (more...)

Library: PacBio embryonic chicken heart cDNA (more...)

Platform: PacBio SMRT™ (more...)

Spot descriptor:

1 forward

Total: 1 run, 1.8M spots, 2G bases, 486.4Mb

#	Run	# of Spots	# of Bases	Size	Published
1.	SRR1177086	1,849,778	2G	486.4Mb	2014-02-26

①

ID: 655957

FTP ディレクトリ /sra/sra-instant/reads/ByRun/sra/SRR/SRR117/SRR1177086 / ftp-trace.ncbi.nlm.nih.gov

エクスプローラーでこの FTP サイトを表示するには、Alt キーを押して、[表示] をクリックし、[エクスプローラーで FTP サイトを開く] をクリックしてください。

1 階層上のディレクトリへ

02/26/2014 12:00午前　510,036,379 SRR1177086.sra ②

③

Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

DRA

引き続き、同じPacBioのロングリードデータ（SRP038897）を眺める。次は、日米欧三極の一角であるDDBJ SRA (DRA)で眺める。DDBJはsra形式だけでなくFASTQ形式でも生データを提供しているので、オススメ。

- インストール | Rパッケージ | 個別 (last modified 2015/06/10) NEW
- (削除予定)Rのインストールと起動 (last modified 2015/04/02)
- (削除予定)個別パッケージのインストール (last modified 2015/02/20)
- 基本的な利用法 (last modified 2015/04/03)
- サンプルデータ (last modified 2015/02/15)
- バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | 速習コース (la
- 書籍 | トランスクリプトーム解析 | について (last modified 2014/05/12)
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト

① ②

サンプルデータ

1. Illumina/36bp/single-end/human (SRA000299) data (Marioni et al., Genome Res., 2008)
   「Kid
   プル
   のも
   デー

```
7列
8列
9列
10
11
12
13
14
15
16
17
18
19
20
```

2. Illun
   Supp

28. Illumina GAIIx/76bp/paired-end/Drosophila or Illumina HiSeq 2000/100bp/paired-end/Drosophila (SRP009459; GSE33905) data (Graveley et al., Nature, 2011; Brown et al., Nature, 2014)
    ショウジョウバエの様々な組織のデータ(modENCODE)。29 dissected tissue samplesのstrand-specific, paired-endのbiological replicates (duplicates)だそうです。
29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (SRP011435; GSE36469) data (Huang et al., Development, 2012)
    シロイヌナズナの2群間比較用データ: 4 DEX-treated vs. 4 mock-treated
30. PacBio/xxx bp/Human (ERP003225) data (Sharon et al., Nat Biotechnol., 2013)
    ヒトの長鎖RNA-seqデータ③配列長はリードによって異なります。
31. PacBio/xxx bp/Chicken (SRP038897 by DRA; SRP038897 by ENA; SRP038897 by SRA) data (Sharon et al., PLoS One, 2014)
    ニワトリの長鎖RNA-seqデータです。配列長はリードによって異なります。
32. k-mer解析用のランダム配列から生成したFASTA形式ファイル(sample32_ref.fastaとsample32_ngs.fasta)です。50塩基の長さのリファレンス配列を生成したのち、20塩基長の部分配列を10リード分だけランダム抽出したものです。塩基の存在比はAが22%, Cが28%, Gが28%, Tが22%にしています。リファレンス配列(仮想ゲノム配列)がsample32_ref.fastaで、10リードからなる仮想NGSデータがsample32_ngs.fastaです。リード長20塩基で10リードなのでトータル200塩基となり、50塩基からなる元のゲノム配列の4倍シーケンスしていることになります(4X coverageに相当)。イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成からと基本的に同じです。

```
out_f1 <- "sample32_ref.fasta"          #出力ファイル名を指定してout_f1に格納
```

DRA

Windowsのブラウザ IEで眺める場合は、タブがどんどん増えていきます。ここでは④を押しているが、いろいろポチポチ押して⑤のSRR～というRun IDに辿りつけば、基本的にリードの塩基配列を眺めることができるはずだが、これは眺められない、という例でもある

30. PacBio/xxx bp/Human (ERP003225) data (Sharon et al., Nat Biotechnol., 2013)
ヒトの長鎖RNA-seqデータ…配列長はリードによって異なります。
31. PacBio/xxx bp/Chicken (SRP038897 by DRA; SRP038897 by ENA; SRP038897 by SRA) data (Shar… One, 2014)
32. k-…

③

DRASearch Send Feedback Search

SRP038897

Study Detail	
Title	Long-read sequencing of chicken transcripts and identification of new transcript isoforms.
Study Type	Transcriptome Analysis
	The chicken has long served as an important model organism in many fields, and continues to aid our understanding of animal development.

Navigation
Submission SRA142153 FTP
Experiment SRX475467 FASTQ
Sample SRS561227

④

DRASearch Send Feedback Search Home DRA Home

SRX475467 FASTQ SRA

Experiment Detail	
Title	Chicken embryonic heart transcriptome sequencing
	Collecting embryonic chicken heart RNA Chicken eggs were stored at 4°C after laying and then incubated in Genesis Hova-Bators until tissue was harvested. Hearts were dissected from embryos at HH stages 18-20, 25, and 32 and immediately flash frozen in N2. The hearts were homogenized and suspended in TriZol (Life Technologies) by repipetting, and then total RNA was extracted into

Navigation
Submission SRA142153 FTP
Study SRP038897
Sample SRS561227
Run SRR1177086 FASTQ SRA

⑤

DRA

右下のページの赤枠あたりにリード塩基配列情報が出ることを想定したが、出ない。

DRASearch　Send Feedback　Search Home　DRA Home

SRX475467 FASTQ SRA

Experiment Detail	
Title	Chicken embryonic heart transcriptome sequencing
	Collecting embryonic chicken heart RNA Chicken eggs were stored at 4°C after laying and then incubated in Genesis Hova-Bators until tissue was harvested. Hearts were dissected from embryos at HH stages 18-20, 25, and 32 and immediately flash frozen in N2. The hearts were homogenized and suspended in TriZol (Life

Navigation		
Submission	SRA142153	FTP
Study	SRP038897	
Sample	SRS561227	
Run	SRR1177086	FASTQ SRA

⑤

DRASearch　Send Feedback　Search Home　DRA Home

SRR1177086 FASTQ SRA

Run Detail	
Alias	Chicken embryonic heart transcriptome sequencing
Instrument model	
Date of run	
Run center	
Number of spots	1,849,778
Number of bases	1,989,004,881

Navigation		
Submission	SRA142153	FTP
Study	SRP038897	
Experiment	SRX475467	FASTQ SRA

READS (joined)　quality ☐　show 10 rows　<< < 1 / 184978 Page > >>

DRAでFASTQ取得

比較的新しい論文のデータは、FASTQファイルがまだ生成されてなくてダウンロードができないこともある。1年前はリンクが張られてなかったが、2015年6月13日にはリンクが張られている。

30. PacBio/xxx bp/Human (ERP003225) data (Sharon et al., Nat Biotechnol., 2013)
ヒトの長鎖RNA-seqデータです。配列長はリードによって異なります。
31. PacBio/xxx bp/Chicken (SRP038897 by DRA; SRP038897 by ENA; SRP038897 by SRA) data (
One, 2014)
32. k-

DRASearch Send Feedback Search Home DRA Home

SRP038897

Study Detail	
Title	Long-read sequencing of chicken transcripts and identification of new transcript isoforms.
Study Type	Transcriptome Analysis
	The chicken has long served as an important model organism in many fields, and continues to aid our understanding of animal development.

Navigation
Submission SRA142153 FTP
Experiment SRX475467 FASTQ
Sample SRS561227

DRASearch Send Feedback Search Home DRA Home

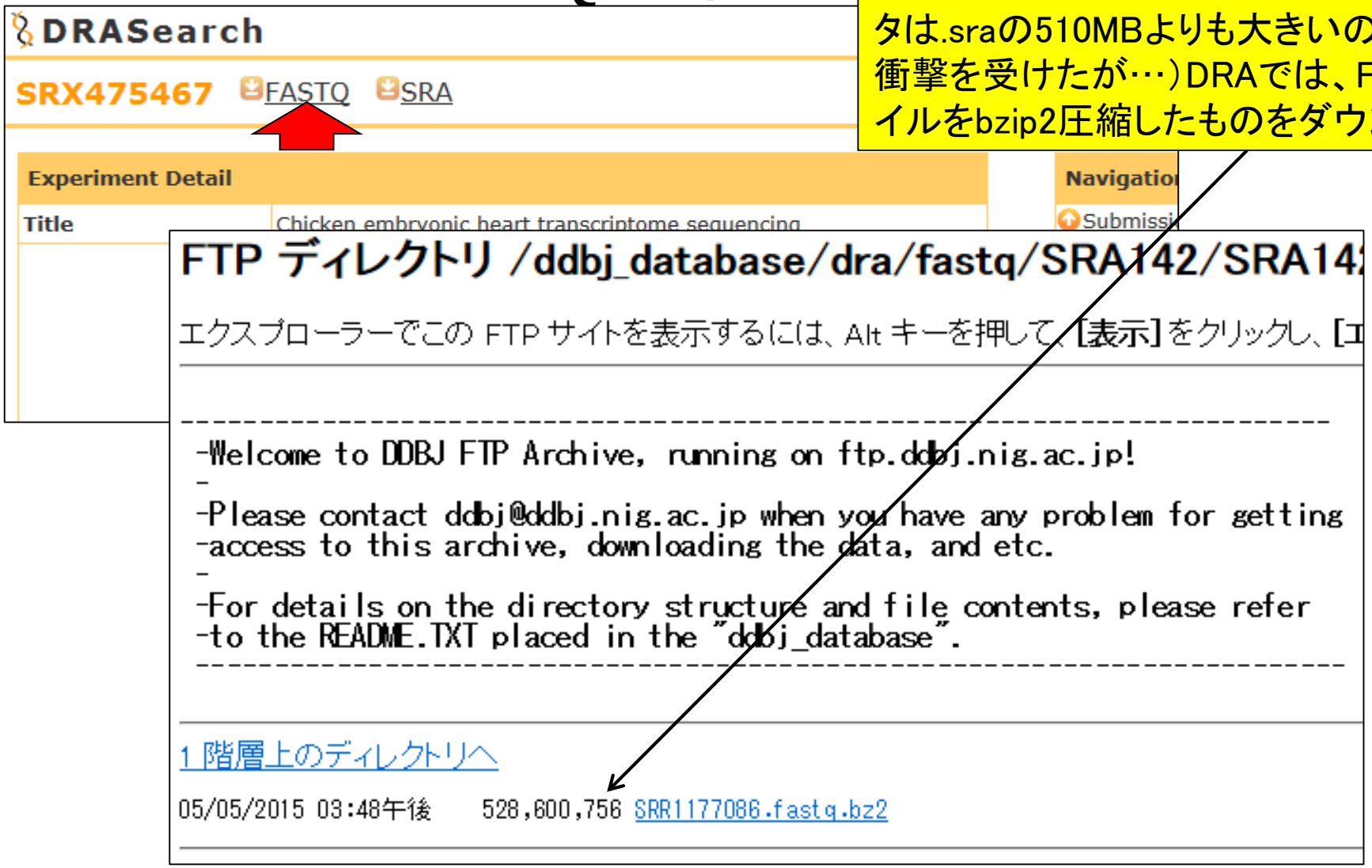
SRX475467 FASTQ SRA

Experiment Detail	
Title	Chicken embryonic heart transcriptome sequencing
	Collecting embryonic chicken heart RNA Chicken eggs were stored at 4°C after laying and then incubated in Genesis Hova-Bators until tissue was harvested. Hearts were dissected from embryos at HH stages 18-20, 25, and 32 and immediately flash frozen in N2. The hearts were homogenized and suspended in TriZol (Life Technologies) by repipetting, and then total RNA was extracted into

Navigation
Submission SRA142153 FTP
Study SRP038897
Sample SRS561227
Run SRR1177086 FASTQ SRA

DRAでFASTQ取得

.sraファイルは、それ自体が圧縮ファイルであり、FASTQ形式ファイルに変換するのが（おそらくLinux上でしかできない?!ので）面倒。（このデータは.sraの510MBよりも大きいので個人的には衝撃を受けたが…）DRAでは、FASTQ形式ファイルをbzip2圧縮したものをダウンロード可能。



Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

GSE → SRP

Illuminaのsingle-endデータ(GSE36469)を眺める。次は、日米欧三極の一角であるEuropean Nucleotide Archive(ENA)で眺める。ENAはgzip圧縮したFASTQ形式ファイルを提供している。また、ENAは1つのデータセット(この場合GSE36469から辿れるSRP011435)に属する複数のSRR IDの全体像を俯瞰するときに便利です。まずは原著論文中のGSE36469からSRP011435という情報を取得する手順の伝授から。

- サンプルデータ

- インストール | Rパッケージ | 個別 (last modified 2015/06/10) NEW
- (削除予定)Rのインストールと起動 (last modified 2015/04/02)
- (削除予定)個別パッケージのインストール (last modified 2015/02/20)
- 基本的な利用法 (last modified 2015/04/03)
- サンプルデータ (last modified 2015/02/15)
- バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | 速習コース
- 書籍 | トランスクリプトーム解析 | について (last modified 2014/05/12)
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト

①

サンプルデータ

1. Illumina/36bp/single-end/human (SRA000299) data (Marioni et al.,
「Kidney 7 samples vs Liver 7 samples」のRNA-seqの遺伝子発現行列データ(SupplementaryTable2.txt)です。サンプルは二つの濃度(1.5 pM and 3 pM)でシーケンスされており、「3 pMのものが5 samples vs. 5 samples」、「1.5 pMのものが
データが

7列目：
8列目：
9列目：
10列目
11列目
12列目
13列目
14列目
15列目
16列目
17列目
18列目
19列目
20列目

②

2. Illumina/
Supplem

27. Illumina HiSeq 2000 (GPL14844)/50bp/single-end/Rat (SRP037986; GSE53960) data (Yu et al., Nat Commun., 2014)
ラットの10組織×雌雄(2種類)×4種類の週齢(2, 6, 21, 104 weeks)×4 biological replicatesの計320サンプルからなるデータ。
28. Illumina GAIIx/76bp/paired-end/Drosophila or Illumina HiSeq 2000/100bp/paired-end/Drosophila (SRP009459; GSE33905) data (Graveley et al., Nature, 2011; Brown et al., Nature, 2014)
ショウジョウバエの様々な組織のデータ(modENCODE)。29 dissected tissue samplesのstrand-specific, paired-endのbiological replicates (duplicates)だそうです。

③

29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (GSE36469) data (Huang et al., Development, 2012)
シロイヌナズナの2群間比較用データ: 4 DEX-treated vs. 4 mock-treated
原著論文中では、GSE36469のみが示されていますが、日米欧三極のDB(SRP011435 by SRA; SRP011435 by DRA; SRP011435 by ENA)からも概観できます。
30. PacBio/xxx bp/Human (ERP003225) data (Sharon et al., Nat Biotechnol., 2013)
ヒトの長鎖RNA-seqデータです。配列長はリードによって異なります。
31. PacBio/xxx bp/Chicken (SRP038897 by DRA; SRP038897 by ENA; SRP038897 by SRA) data (Sharon et al., PLoS One, 2014)
ニワトリの長鎖RNA-seqデータです。配列長はリードによって異なります。

• サンプルデータ

GSE → SRP

③

GSE36469のリンク先は、NCBIのGEOという発現DB。このページの下のほうにSRP011435という情報がある。（GSE36469でググってもある程度いけるが）このSRP IDでENAを概観するのがおススメ

29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (GSE36469) dat

シロ

原著

DRA

DB(SRP011435 by SRA; SRP011435 by

Series GSE36469 Query DataSets for GSE36469

Status	Public on Jul 12, 2012
Title	High-thoughput Illumina RNA sequencing to identify downstream target genes of RABBIT EARS (RBE) in the flowers of Arabidopsis thaliana
Organism	Arabidopsis thaliana
Experiment type	Expression profiling by high throughput sequencing
Summary	In order to identify putative downstream target genes of RBE, we sequenced mRNA from dexamethasone (DEX) and mock treated transgenic Arabidopsis line 35S:GR-RBE (RBE coding region fused to a glucocorticoid receptor domain driven by the constitutive 35S promoter) floral tissues. We compared the results from DEX and mock treatments and focused on the 832 genes whose expression was significantly reduced (P < compared to mock-treated plants. In t (EEP1) as a candidate target of RBE, w molecular and genetic analyses. Regulati normal floral organ formation in Arabidop
Overall design	We used two biological replicates, each hour DEX or mock treated floral tissues to
Contributor(s)	Huang T, López-Giráldez F, Townsend JP,
Citation(s)	Huang T, López-Giráldez F, Townsend JP, expression to effect floral organogenesis. (12):2161-9. PMID: 22573623
Submission date	Mar 13, 2012
Last update date	Dec 19, 2014
Contact name	Francesc Lopez
E-mail	francesc.lopez@yale.edu
Organization name	Yale University
Department	Department of Genetics
Lab	YCGA
Street address	P.O. Box 27386
City	West Haven
State/province	CT
ZIP/Postal code	06516
Country	USA

Platforms (1)	GPL13222	Illumina HiSeq 2000 (Arabidopsis thaliana)
Samples (8) ± More...	GSM894355	mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1
	GSM894356	mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 2
	GSM894357	mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 2, techinical replicate 1

Relations

SRA	SRP011435
BioProject	PRJNA153493

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE36469_LOX_output_combined_final.txt.gz	1.3 Mb	(ftp)(http)	TXT
SRP/SRP011/SRP011435		(ftp)	SRA Study

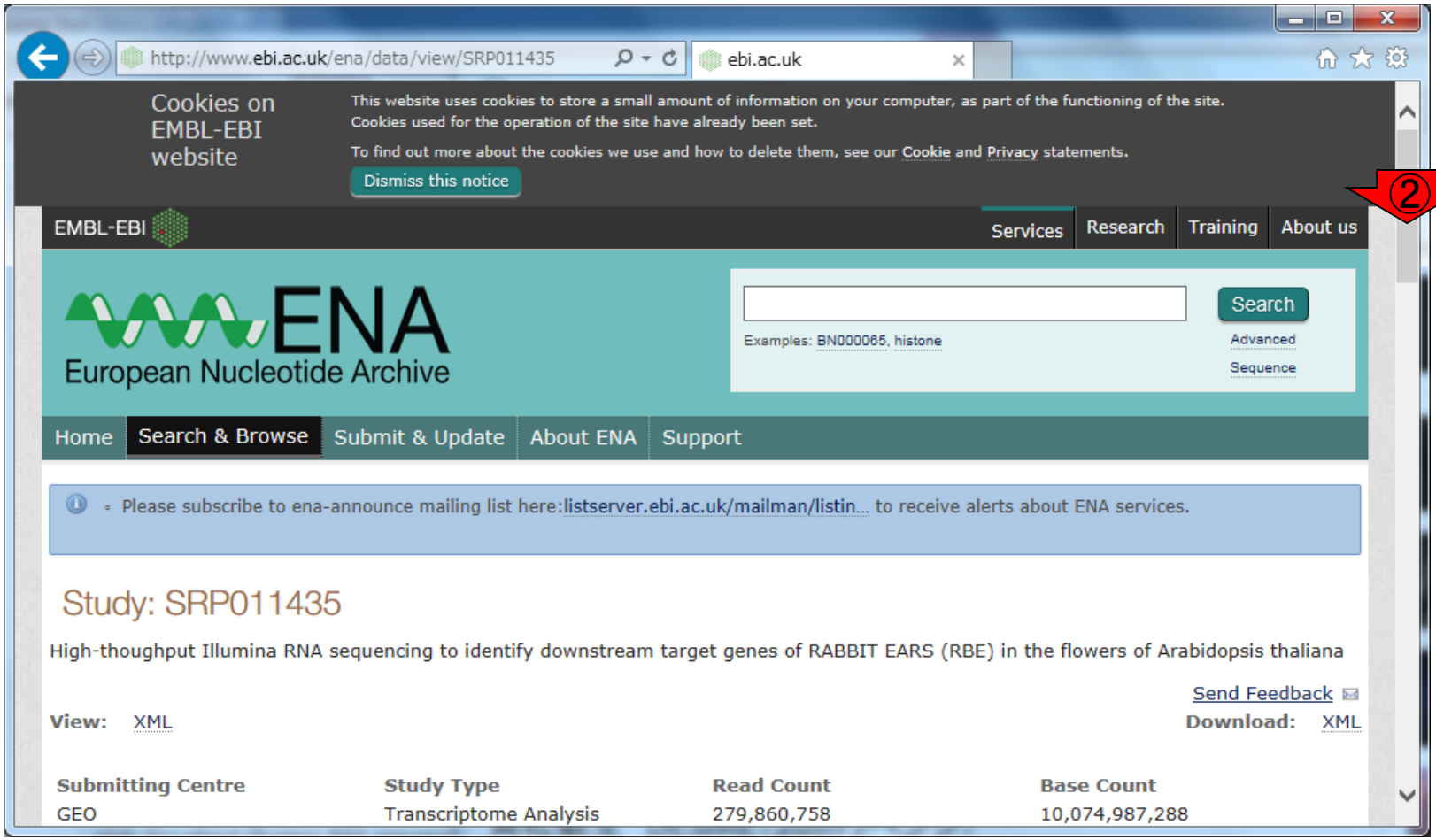
• サンプルデータ

ENA

GSE36469のリンク先は、NCBIのGEOという発現DB。このページの下のほうにSRP011435という情報がある。(GSE36469でググってもある程度いけるが)このSRP IDでENAを概観するのがおススメ

29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (GSE36469) dat
シロイヌナズナの2群間比較用データ: 4 DEX-treated vs. 4 mock-treated
原著論文中では、GSE36469のみが示されていますが、日米欧三極のDB(SRP011435 by SRA; SRP011435 by DRA; SRP011435 by ENA)からも概観できます。

①



②

ENA

SRP011435でENAを概観した結果。それぞれのIDの対応関係が一目瞭然。1つのSRP IDに対して、8つのSRR IDが存在することが分かる。For understanding the overall relationship of various IDs, I recommend you to utilize ENA.

Showing results 1 - 8 of 8 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)
PRJNA153493	SRP011435	SAMN00811243	SRS300127	SRX129185	SRR444595	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811244	SRS300128	SRX129186	SRR444596	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811245	SRS300129	SRX129187	SRR444597	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811246	SRS300130	SRX129188	SRR444598	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811247	SRS300131	SRX129189	SRR444599	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811248	SRS300132	SRX129190	SRR444600	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811249	SRS300133	SRX129191	SRR444601	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811250	SRS300134	SRX129192	SRR444602	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1

ENA

原著論文からもわかるが、①これはsingle-endデータなのでSINGLEと書かれている。②paired-endデータの場合は確かPAIREDとなり、SRR IDごとに2つのファイル(File 1とFile 2)に分かれて提供される。

Showing results 1 - 8 of 8 results

②

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)
PRJNA153493	SRP011435	SAMN00811243	SRS300127	SRX129185	SRR444595	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811244	SRS300128	SRX129186	SRR444596	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811245	SRS300129	SRX129187	SRR444597	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811246	SRS300130	SRX129188	SRR444598	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811247	SRS300131	SRX129189	SRR444599	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811248	SRS300132	SRX129190	SRR444600	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811249	SRS300133	SRX129191	SRR444601	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811250	SRS300134	SRX129192	SRR444602	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1

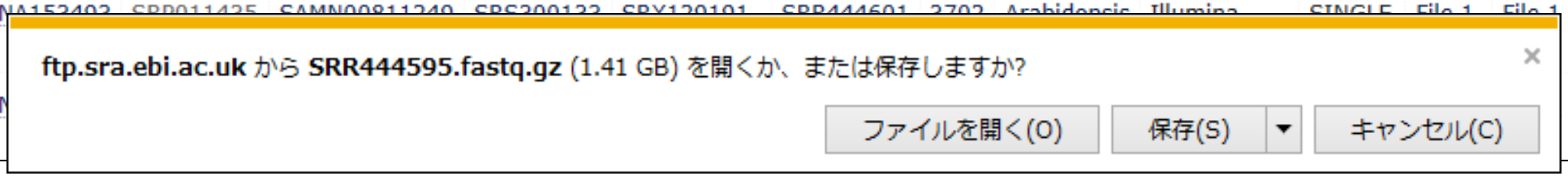
①

ENA

FASTQファイルをダウンロードする場合。ENAではgzip圧縮（拡張子がgz）されたFASTQファイルとして提供されている。1ファイルあたり約1.41GBだから、計8ダウンロードするだけで10GB超になると概算する。

Showing results 1 - 8 of 8 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)
PRJNA153493	SRP011435	SAMN00811243	SRS300127	SRX129185	SRR444595	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811244	SRS300128	SRX129186	SRR444596	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811245	SRS300129	SRX129187	SRR444597	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811246	SRS300130	SRX129188	SRR444598	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811247	SRS300131	SRX129189	SRR444599	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811248	SRS300132	SRX129190	SRR444600	3702	Arabidopsis thaliana	Illumina HiSeq 2000	SINGLE	File 1	File 1
PRJNA153493	SRP011435	SAMN00811249	SRS300133	SRX129191	SRR444601	3702	Arabidopsis	Illumina	SINGLE	File 1	File 1



DRA

SRR44595をENAからダウンロードすると1.41GBであるのに対し、DRAからダウンロードすると1.37GB。

29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (GSE36469) data (Huang et al., Development, 2012)
シロイヌナズナの2群間比較用データ: 4 DEX-treated vs. 4 mock-treated
原著論文中では、GSE36469のみが示されていますが、日米欧三極のDB(SRP011435 by SRA; SRP011435 by DRA; SRP011435 by ENA)からも概観できます。

①

DRASearch　Send Feedback　Search Home　DRA Home

SRP011435

Study Detail	
Title	High-thoughput Illumina RNA sequencing to identify downstream target genes of RABBIT EARS (RBE) in the flowers of Arabidopsis thaliana
Study Type	Transcriptome Analysis
Abstract	In order to identify putative downstream target genes of RBE, we sequenced mRNA from dexamethasone (DEX) and mock treated transgenic Arabidopsis line 35S:GR-RBE (RBE coding region fused to a glucocorticoid receptor domain driven by the constitutive 35S promoter) floral tissues. We compared the resul .. [more]
Description	
Center Name	GEO

Navigation			
Submission	SRA050735	FTP	
Experiment	SRX129185	FASTQ	SRA
	SRX129186	FASTQ	SRA
	SRX129187	FASTQ	SRA
	SRX129188	FASTQ	SRA
	SRX129189	FASTQ	SRA
	SRX129190	FASTQ	SRA
	SRX129191	FASTQ	SRA
	SRX129192	FASTQ	SRA
Sample	SRS300127		
	SRS300128		
	SRS300129		
	SRS300130		
	SRS300131		
	SRS300132		
	SRS300133		
	SRS300134		

②

③

03/25/2013 12:00午前　1,368,387,683 SRR444595.fastq.bz2

SRA

SRR44595をENAからダウンロードすると1.41GB。DRAからダウンロードすると1.37GB。③SRAからダウンロードすると0.92GB (965,960,145 bytes; 1KB = 1,024 bytes)。④リードの概観はこちら

29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (GSE36469) data (Hu
シロイヌナズナの2群間比較用データ: 4 DEX-treated vs. 4 mock-treate
原著論文中では、GSE36469のみが示されていますが、日米欧三極の
DRA; SRP011435 by ENA)からも概観できます。

①

Results: 8

1. GSM894362: mRNA from mock-treated 35S:GR-RBE floral buds, biological replicate 2, techinical replicate 2; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 36.7M spots, 1.3G bases, 735.8Mb downloads
Accession: SRX129192

2. GSM894361: mRNA from mock-treated 35S:GR-RBE floral buds, biological replicate 2, techinical replicate 1; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 38.3M spots, 1.4G bases, 900.8Mb downloads
Accession: SRX129191

3. GSM894360: mRNA from mock-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 2; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 36.1M spots, 1.3G bases, 866.3Mb downloads
Accession: SRX129190

4. GSM894359: mRNA from mock-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 39.7M spots, 1.4G bases, 926.6Mb downloads
Accession: SRX129189

5. GSM894358: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 2, techinical replicate 2; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 27.1M spots, 975.7M bases, 668.4Mb downloads
Accession: SRX129188

6. GSM894357: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 2, techinical replicate 1; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 29.4M spots, 1.1G bases, 721.9Mb downloads
Accession: SRX129187

7. GSM894356: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 2; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 32.1M spots, 1.2G bases, 779.1Mb downloads
Acc 129186

②

8. GSM894355: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 40.4M spots, 1.5G bases, 921.2Mb downloads

SRX129185: GSM894355: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1; Arabidopsis thaliana; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 40.4M spots, 1.5G bases, 921.2Mb downloads

Accession: SRX129185
Experiment design: n/a
Submitted by: GEO
Study summary: SRP011435 • High-thoughput Illumina RNA sequencing to identify downstream target genes of RABBIT EARS (RBE) in the flowers of Arabidopsis thaliana • PRJNA153493 • All experiments • Run Selector (more...)
Sample: SAMN00811243 • mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1 (more...)
Library: GSM894355: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1 (more...)
Platform: Illumina (more...)
Spot descriptor:

1 forward

Experiment attributes:
GEO Accession: GSM894355
Total: 1 run, 40.4M spots, 1.5G bases, 921.2Mb

#	Run	# of Spots	# of Bases	Size	Published
1.	SRR444595	40,422,066	1.5G	921.2Mb	2012-07-13

ID: 14729

④

③

08/12/2012 12:00午前 965,960,145 SRR444595.sra

SRA

リードの概観。これはごく初期のIlluminaの36 bp single-endデータ。

NCBI Site map | All databases | Search

Sequence Read Archive

Main | Browse | Search | Download | Submit | Documentation | Software | Trace Archive | Trace Assembly | Trace Home | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

Change accession..

GSM894355: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1; Arabidopsis thaliana; RNA-Seq (SRR444595)

① Metadata | Reads | Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR444595						

Quality graph (big

This run has 1 r

Legend

Experiment

SRX129185

to BLAST

Biosample

SAMN0081124 (SRS300127)

GSM894355: mRNA from DEX-treated 35S:GR-RBE floral buds, biological replicate 1, techinical replicate 1; Arabidopsis thaliana; RNA-Seq (SRR444595)

Metadata | Reads | Download

Filter: Find | Filtered Download | What does it do?

What can the filter be applied to?

< 1 1 4042207 >

View: ☐ biological reads ☐ technical reads ☑ quality scores advanced options

1. SRR444595.1 SRS300127
name: GA-K_0045:1:1:2896:1080, member: default
x: 2896, y: 1080

2. SRR444595.2 SRS300127
name: GA-K_0045:1:1:3404:1080, member: default
x: 3404, y: 1080

3. SRR444595.3 SRS300127
name: GA-K_0045:1:1:3691:1080, member:

Read

>gnl|SRA|SRR444595.1 GA-K_0045:1:1:2896:1080
GGNGGAGATGTTGTGACTGGTGGTGGTGCTGAGACA

One channel quality score

30 30 2 24 29 30 33 33 33 33 38 38 38 35 33 35 33 35 35 36
32 37 35 32 37 33 29 33 33 30 36 29 38 38 31 23

Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

paired-endデータ概観

IlluminaのPaired-endデータ(GSE42960)をENAで眺める

- インストール | Rパッケージ | 個別 (last modified 2015/06/10) NEW
- (削除予定)Rのインストールと起動 (last modified 2015/04/02)
- (削除予定)個別パッケージのインストール (last modified 2015/02/20)
- 基本的な利用法 (last modified 2015/04/03)
- サンプルデータ (last modified 2015/02/15)
- バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ) | 速習コース (la
- 書籍 | トランスクリプトーム解析 について (last modified 2014/05/12)
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト
- 書籍 | ト

① ②

サンプルデータ

1. Illumina/36bp/single-end/human (SRA000299) data (Marioni et al., Genome Res., 2008)
   「Kidney 7 samples vs Liver 7 samples」のRNA-seqの遺伝子発現行列データ(SupplementaryTable2.txt)です。サンプルは二つの濃度(1.5 pM and 3 pM)でシーケンスされており、「3 pMのものが5 samples vs. 5 samples」、「1.5 pMのものが2 samples vs. 2 samples」という構成です。SupplementaryTable2.txtをエクセルで開くと、7列目以降に発現データがあることがわかります。詳細な情報は以下の通りです(原著論文中のFigure 1からもわかります):

```
7列目：R1L1Kidney (3 pM)
8列目：R1L2Liver (3 pM)
9列目：R1L3Kidney (3 pM)
10列目：R1L4Liver (3 pM)
11列目：R1L6Liver (3 pM)
```

24. Illumina/75bp/single-end/human (SRA061145) data (Wang et al., Nucleic Acids Res., 2013)
    ヒト肺の3群間比較用データ: normal human bronchial epithelial (HBE) cells, human lung cancer A549, and H1299 cells
25. Illumina HiSeq 2000/100bp/paired-end/human (GSE42960) data (Chan et al., Hum. Mol. Genet., 2013)
    ヒトPBMCというサンプルの2群間比較用データ: 未処理群2サンプル (FRDA05-UT and FRDA19.UTB) vs. ニコチンアミド処理群2サンプル(FRDA05-NicoとFRDA19.NB)。原著論文中では、GSE42960のみが示されていますが、日米欧三極のDB(SRP017580 by SRA; SRP017580 by DRA; SRP017580 by ENA) からも概観できます。
    ペアエンドデータのSRR633902_1.fastqを入力として、最初の400リード分を抽出することで、SRR633902_1_sub.fastqを作成しています。writeFastq関数のデフォルトオプションはcompress=Tで、gzip圧縮ファイルを出力します。ここではcompress=Fとして非圧縮ファイルを出力しています。

③

paired-endデータ概観

②paired-endの場合は、1つのSRR IDのリードデータが2つのファイル(*_1.*と*_2.*)に分割される。計4サンプルとはいえ、全部をダウンロードすると、こちらも10GB程度になる

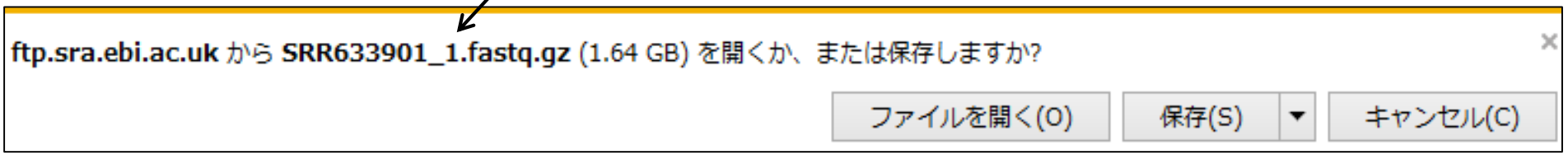
25. Illumina HiSeq 2000/100bp/paired-end/human (GSE42960) data (Chan et a
ヒトPBMCというサンプルの2群間比較用データ：未処理群2サンプル (FRDA05-UT and FRDA19.UTB) vs. ニコチンアミド処理群2サンプル(FRDA05-NicoとFRDA19.NB)。原著論文中では、GSE42960のみが示されていますが、日米欧三極のDB(SRP017580 by SRA; SRP017580 by DRA; SRP017580 by ENA) からも概観できます。

①

Showing results 1 - 4 of 4 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)
PRJNA183943	SRP017580	SAMN01831455	SRS380020	SRX210736	SRR633901	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
PRJNA183943	SRP017580	SAMN01831456	SRS380021	SRX210737	SRR633902	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
PRJNA183943	SRP017580	SAMN01831457	SRS380022	SRX210738	SRR649759	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1
PRJNA183943	SRP017580	SAMN01831458	SRS380023	SRX210739	SRR649760	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1

②



DRAでみた感じ。トータルで約2.92GB。

paired-endデータ概観

25. Illumina HiSeq 2000/100bp/paired-end/human (GSE42960) data (Chan et al., Hum. Mol. Genet., 2013)
ヒトPBMCというサンプルの2群間比較用データ：未処理群2サンプル (FRDA05-UT and FRDA19.UTB) vs. ニコチンアミド処理群2サンプル(FRDA05-NicoとFRDA19.NB)。原著論文中では、GSE42960のみが示されていますが、日米欧三極のDB(SRP017580 by SRA; SRP017580 by DRA; SRP017580 by ENA) からも概観できます。

①

SRP017580

Study Detail	
Title	The effect of nicotinamide on dysregulated genes associated with frataxin deficiency in FRDA.
Study Type	Transcriptome Analysis
Abstract	To investigate the efficacy of nicotinamide treatment using our ex-vivo primary lymphocyte model, we performed high-throughput RNA sequencing on libraries generated from untreated and nicotinamide treated samples. Overall design: PBMC isolated from FRDA affected individuals were cultured to prepare .. [more]
Description	
Center Name	GEO

Navigation			
Submission	SRA062939	FTP	
Experiment	SRX210736	FASTQ	SRA
	SRX210737	FASTQ	SRA
	SRX210738	FASTQ	SRA
	SRX210739	FASTQ	SRA
Sample	SRS380020		
	SRS380021		
	SRS380022		
	SRS380023		

②

04/21/2013 12:00午前	18,928	SRR633901.fastq.bz2
04/21/2013 12:00午前	1,485,828,538	SRR633901_1.fastq.bz2
04/21/2013 12:00午前	1,438,635,248	SRR633901_2.fastq.bz2

SRAでみた感じ。トータルで約2.68GB。

paired-endデータ概観

25. Illumina HiSeq 2000/100bp/paired-end/human (GSE42960) data (Chan et al., Hum. Mol. Genet., 2013)
ヒトPBMCというサンプルの2群間比較用データ：未処理群2サンプル (FRDA05-UT and FRDA19.UTB) vs ニコチンアミド処理群2サンプル(FRDA05-NicoとFRDA19.NB)。原著論文中では、GSE42960のみが示されていますが、日米欧三極のDB(SRP017580 by SRA; SRP017580 by DRA; SRP017580 by ENA) からも概観できます。

①

SRX210736: GSM1054021: FRDA05-UT; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 22.4M spots, 4.5G bases, 2.5Gb downloads

Accession: SRX210736
Experiment design: n/a
Submitted by: GEO
Study summary: SRP017580 • The effect of nicotinamide on dysregulated genes associated with frataxin deficiency in FRDA. • PRJNA183943 • All experiments • Run Selector (more...)
Sample: SAMN01831455 • GSM1054021: FRDA05-UT (more...)
Library: (more...)
Platform: Illumina (more...)
Spot descriptor:

Experiment attributes:
GEO Accession: GSM1054021
Total: 1 run, 22.4M spots, 4.5G bases, 2.5Gb

#	Run	# of Spots	# of Bases	Size	Published
1.	SRR633901	22,419,833	4.5G	2.5Gb	2013-04-03

②

ID: 286859

12/17/2012 12:00午前 2,676,898,597 SRR633901.sra

paired-endデータ概観

①1つのファイルにまとめられているもの?!もあるようです、詳細は不明。ファイルサイズも微妙。

25. Illumina HiSeq 2000/100bp/paired-end/human (GSE42960) data (Chan et al., Hum. Mol. Genet., 2013)
ヒトPBMCというサンプルの2群間比較用データ：未処理群2サンプル (FRDA05-UT and FRDA19.UTB) vs. ニコチンアミド処理群2サンプル(FRDA05-NicoとFRDA19.NB)。原著論文中では、GSE42960のみが示されていますが、日米欧三極のDB(SRP017580 by SRA; SRP017580 by DRA; SRP017580 by ENA) からも概観できます。

Showing results 1 - 4 of 4 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)
PRJNA183943	SRP017580	SAMN01831455	SRS380020	SRX210736	SRR633901	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
PRJNA183943	SRP017580	SAMN01831456	SRS380021	SRX210737	SRR633902	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
PRJNA183943	SRP017580	SAMN01831457	SRS380022	SRX210738	SRR649759	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1
PRJNA183943	SRP017580	SAMN01831458	SRS380023	SRX210739	SRR649760	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1

①

ftp.sra.ebi.ac.uk から SRR649759.fastq.gz (2.08 GB) を開くか、または保存しますか?

ファイルを開く(O) 保存(S) キャンセル(C)

paired-endデータ概観

受講人数の多いアグリバイオでは教えられませんが、個人利用の場合はGalaxyもおススメです。

25. Illumina HiSeq 2000/100bp/paired-end/human (GSE42960) data (Chan et al., Hum. Mol. Genet., 2013)
ヒトPBMCというサンプルの2群間比較用データ：未処理群2サンプル (FRDA05-UT and FRDA19.UTB) vs. ニコチンアミド処理群2サンプル(FRDA05-NicoとFRDA19.NB)。原著論文中では、GSE42960のみが示されていますが、日米欧三極のDB(SRP017580 by SRA; SRP017580 by DRA; SRP017580 by ENA) からも概観できます。

Showing results 1 - 4 of 4 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)
PRJNA183943	SRP017580	SAMN01831455	SRS380020	SRX210736	SRR633901	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
PRJNA183943	SRP017580	SAMN01831456	SRS380021	SRX210737	SRR633902	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2
PRJNA183943	SRP017580	SAMN01831457	SRS380022	SRX210738	SRR649759	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1
PRJNA183943	SRP017580	SAMN01831458	SRS380023	SRX210739	SRR649760	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1

http://dbcls.rois.ac.jp/archives/2715

DBCLS大田先生の講演資料。
Galaxy can also be recommended.

Tips：Galaxyもおススメ

DBCLS
Database Center for Life Science

センターについて
研究開発
サービス
イベント
メンバー
アクセス
お問い合わせ

2015/01/19

統合データベース講習会AJACSadvanced(AJACSa)三島開催のお知らせ

ホーム ＞ イベント ＞ 統合データベース講習会AJACSadvanced(AJACSa)三島開催のお知らせ

ライフサイエンス統合データベースセンター(DBCLS) は、2014年4月より国立遺伝学研究所 (遺伝研) 内にて
「三島ラボ」を構え活動しています。昨年
DBCLS三島ラボが入居しました。 そのこ
催する運びとなりました。この講習会では
講習会AJACS (http://events.bioscienc
開発している各種ツールやデータベース
戦的な使い方についてハンズオンで実習し
ます。終了後には、新しいラボのお披露目

日時: 2015年1月27日(火) 10:00-17:0
http://www.nig.ac.jp/about/map.htm

午前の部

10:00-10:30 統合データベースとDBCLSとは? 坊農 秀雅 (DBCLS)

10:30-11:00 DDBJ 中村 保一 (DDBJ)

11:00-11:40 FANTOM5データの再利用法 粕川 雄也 (RIKEN CLST)

午後の部

13:00-13:50 NGSデータベース検索 仲里 猛留 (DBCLS)

13:50-14:40 遺伝子発現データ解析 小野 浩雅 (DBCLS)

14:40-15:00 休憩(+スパコン見学)

15:00-15:50 CRISPR/Cas9ターゲット配列設計 内藤 雄樹 (DBCLS)

15:50-16:40 galaxyによるNGS解析 大田 達郎 (DBCLS)

16:40-17:00 休憩

http://www.pitagora-galaxy.org/

Tips：ピタゴラギャラクシー

様々な活動に積極的に足を突っ込んでおくのは大事かも。You should learn various strategies for analyzing NGS data.

Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

2014年度までの講義資料
。An old slide used in 2014

公共DB

- DDBJ Sequence Read Archive (DRA; 日)
  - bzip2圧縮FASTQ形式(.fastq.bz2)とSRA-lite形式ファイルでダウンロード可能
- ENA Sequence Read Archive (ERA; 欧)
  - gzip圧縮FASTQ形式ファイル(.fastq.gz)でダウンロード可能
- NCBI Sequence Read Archive (SRA; 米)
  - SRA形式でダウンロード可能

http://trace.ddbj.nig.ac.jp/news/2011-01-05.html

2014年度までの講義資料
。An old slide used in 2014

様々なファイル形式…

- 情報量：SRA−full ＞ SRA−lite ＞ FASTQ ＞ FASTA
  - SRA-full：塩基配列、クオリティ情報、Intensity情報など画像以外の全て
  - SRA-lite：SRA-fullからIntensity情報を除いて軽量化したもの
  - FASTQ：塩基配列とクオリティ情報のみからなるもの
  - FASTA：塩基配列のみからなるもの
  - ファイルサイズ (SRA-full : SRA-lite : FASTQ : FASTA)
    - 6 : 3 : 2 : 1
    - 例：SRA-fullはFASTQの約3倍

http://rgm22.nig.ac.jp/mediawiki−ogareport/index.php/RAW_DATA_archiving/sharing_at_DDBJ

公共DB(Public DB)

- DDBJ Sequence Read Archive (DRA; 日)
  - bzip2圧縮FASTQ形式(.fastq.bz2)
  - SRA形式ファイル(.sra)
- ENA Sequence Read Archive (ERA; 欧)
  - gzip圧縮FASTQ形式ファイル(.fastq.gz)
- NCBI Sequence Read Archive (SRA; 米)
  - SRA形式ファイル(.sra)

昔のSRAは、SRA(-full)とSRA-lite という2つの形式が存在していた。これはリード塩基配列とクオリティ情報の他に、IlluminaのIntensityという情報を含む(SRA-full)か含まないか(SRA-lite)という違い。昔はIntensity情報も保存していたが、実際にはほとんど使われていないということでNCBIがIntensity情報の保存をやめた。それゆえ、SRA-fullとSRA-liteという区別の必要がなくなり、実体としてはSRA-liteのみの情報からなる.sraという拡張子のついたファイルが提供されているのが現在の状況。Current public databases provide mainly two kinds of format (i.e., .fastq and .sra).

DDBJの児玉 悠一先生 提供情報。Thanks to Dr. Kodama, Y.

教科書p11-13

FASTQ形式のファイルサイズはFASTA形式の約2倍。**課題1。**The file size for FASTQ compared to FASTA is roughly two-fold.

ファイル形式(file formats)

- FASTA format
  - 1行目:“>”ではじまる一行のdescription行
  - 2行目:配列情報

```
>SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

- FASTQ format
  - 1行目:“@”ではじまる1行のdescription行
  - 2行目:配列情報
  - 3行目:”+”からはじまる1行(のdescription行)
  - 4行目:クオリティ情報

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

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

教科書p17

一昔前は.sraのほうが圧縮fastqファイルよりもファイルサイズが大きかったのですが、時代の流れは早いモノです（遠い目）。

ファイル形式(file formats)

- SRA format
  - FASTQに相当する情報 ＋ インデックス情報からなる
  - 2015年6月現在、.sraのほうが.fastq.bz2 or .fastq.gzよりもファイルサイズが小さい。主な理由は、圧縮率の向上とアーカイブ（保存）情報の形式による。
    - 64 bits/base (2008) -> 6 bits/base (2012)
  - テキスト形式ではなくバイナリ形式（それゆえ目で見ても意味不明）
    - （後述する）テキスト形式のSAMファイルよりもバイナリ形式のBAMファイルのほうがファイルサイズが小さいように、バイナリ形式にするメリットは大きい。
  - 従来はFASTQ形式のようなsequenceされたリード情報を格納するやり方であったが、ヒトデータならヒトゲノムにマッピングしたBAM形式ファイルのような情報で保持する方向性（リファレンス配列を用いた圧縮; reference-based compression）に大きくシフトしている。
    - FASTQとSRA間でのファイルサイズの優位性に関する議論は無意味になりつつある

DDBJの児玉 悠一先生、中村保一先生 提供情報。
Thanks to Dr. Kodama and Dr. Nakamura

雑感(Miscellaneous)

公共データを気軽に解析できる状況にはありません。公共DB側の動きにもアンテナを張っておかねばなりません。The storage problem is critical.

- 公共DBのストレージ(storage)問題
  - ACGTの情報は2 bitで表現可能。Quality score情報がデータの大部分を占める。
    - Scoreの解像度を落としてファイルサイズを削減する方向性も議論されている
  - データの大元はSRA形式ファイル。DDBJでは、sraファイルを入力としてFASTQを作成したものを提供してくれているが、このFASTQファイル作成は計算量の点でも結構大変。
    - DDBJでsraはすぐに提供されるが、FASTQはなかなか得られない所以
  - また、sraとFASTQ両方だとデータ量が2倍になるので早晩行き詰るかも…

DDBJの中村保一先生 提供情報。Thanks to Dr. Nakamura Y.

DDBJing 講習会

DDBJの講習会(workshop)は2015年6月12日にあったようです。このような講習会を通じて最新情報を入手。

DDBJ
DNA Data Bank of Japan

Google™カスタム検索 Search

DDBJ の紹介 | 利用の手引き | レポート・統計 | FAQ | お問い合わせ

HOME > 利用の手引き > DDBJing 講習会　　最終更新日：2015.6.15.

DDBJing 講習会

次回の DDBJing 講習会は、2015年秋頃に開催予定です。

DDBJing 講習会とは？

DDBJ では、DDBJ が提供する様々なサービスを使っていただくことを "DDBJing" と表現し、DDBJ をより有効に活用して頂くために「 DDBJing 講習会」を開催しています。

講習会の内容

講義

これまでの講習会では，以下の様な内容を取り上げて講義を行いました。今後も様々な内容で講義を行い，最新の情報を皆様にお知らせする予定です。

» データの登録方法

塩基配列データ・次世代シークエンサ出力データなどの登録方法について

- DDBJ Nucleotide Sequence Submission System
- Mass Submission System (MSS)
- DDBJ Sequence Read Archive (DRA)
- DDBJ BioProject Database
- DDBJ BioSample Database
- Japanese Genotype-phenotype Archive (JGA)

» 次世代シークエンサ出力データの解析用ツールの紹介と実習

DBCLSの大田達郎先生 提供情報。Thanks to Dr. Ohta, T.

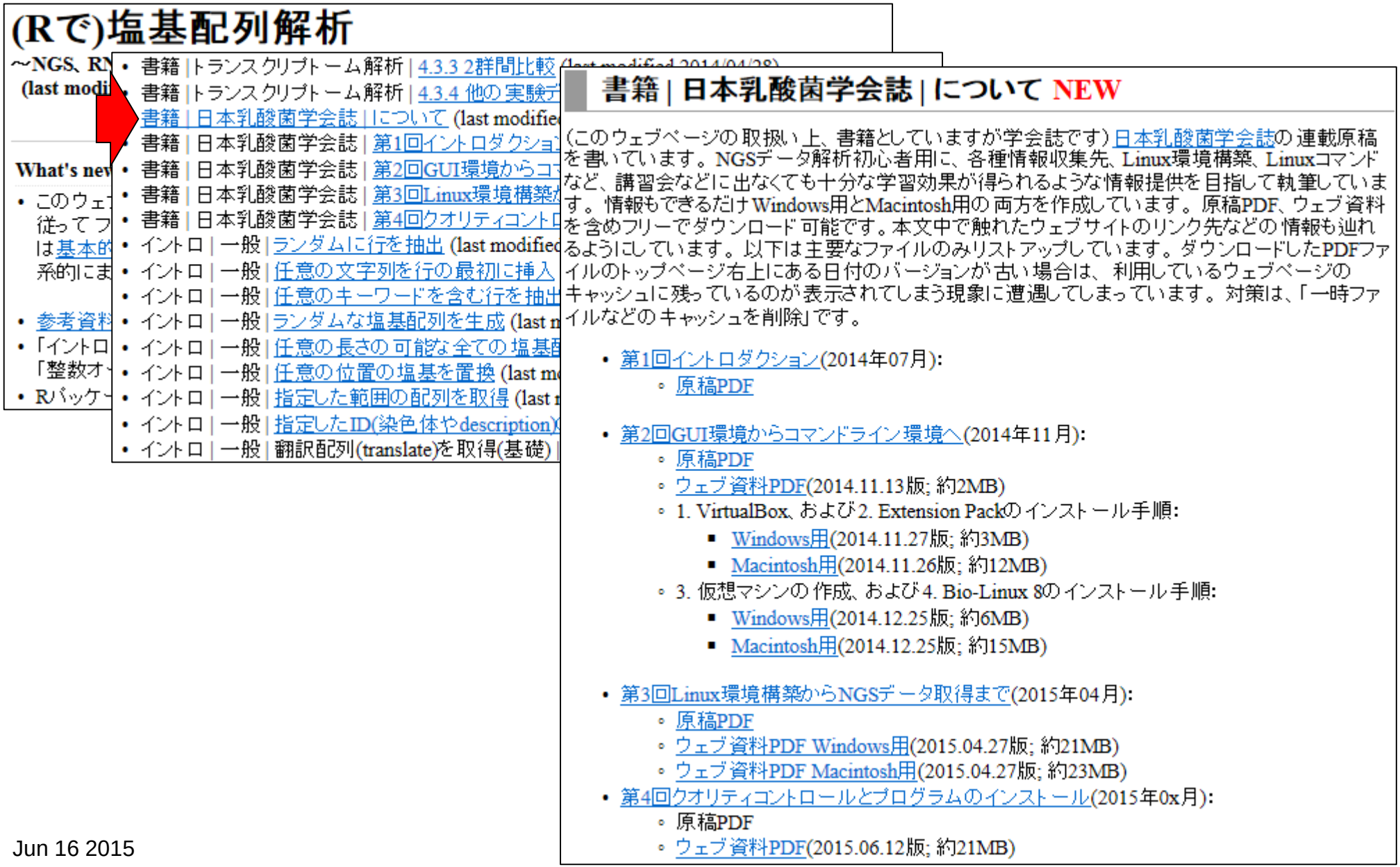
データ解析戦略

クラウドの多くは、Linuxコマンドなどを知らなくてもどうにか解析する方向性を志向。どこを頑張ってどこで手を抜くかなどは、それぞれの与えられている環境やポリシーによっても異なる。

- 解析受託企業に外注
  - 金で解決!、一番手っ取り早い
- クラウド
  - Galaxy、ピタゴラギャラクシー
    - Galaxy Meetupというのも毎月やっているらしい
  - DDBJ Read Annotation Pipeline
  - Illumina BaseSpace
  - BioDevOps by RIKENの二階堂愛先生
    - 2015年6月12日のDDBJingでの講演資料もslideshare上で公開されています
- Linuxコマンドを駆使（旧来型）
  - なるべく自力で解析
  - LinuxコマンドやNGS解析用プログラムのインストールなどを練習し、遺伝研スパコンなどを使いこなす
  - NBDC/東大アグリバイオの「NGSハンズオン講習会」の方向性

Bio-LinuxでNGS解析

自力でLinux環境を構築し、NGS解析を行ってみたいヒトはこちら



データ解析の全体像

クラウドだと何をやっているのかよく分からない、というヒトが一定数存在。NGS解析の全体像をRという手段を使ってできるだけ分かりやすく提供するのが「特論I」の役割。

	マイクロアレイ	RNA-s
公共データ取得	GEO, ArrayExpress	GEO, ArrayExpress, EBI ENA, DDBJ SRA (DRA)
解析対象生物種	配列情報既知（アレイが提供されているもののみ）	モデル・非モデル問わず
生データ	プローブレベル数値データ	塩基配列（数億リード程度、数百塩基長）
		QC (Quality Control)：クオリティチェック、フィルタリング、トリミング
		アセンブリでトランスクリプトーム配列取得（マッピング時のリファレンスとしても利用）
		マッピング(bowtie2, TopHat2など)でSAM/BAMファイル取得
発現行列作成	前処理法(MAS5, RMAなど)適用後に遺伝子発現行列を得る	アノテーションファイルを利用してカウントデータ、配列長補正後のRPKM/FPKM、転写物レベルの発現情報など取得
発現変動遺伝子(DEG)同定	基本Rを利用(limma, SAM, Rank productsなど)	基本Rを利用(cuffdiff2, edgeR, DESeq2, TCCなど)
機能解析	GSEA, GSA, Cytoscapeなど	Rパッケージ SeqGSEAなどを利用。

Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

NGSデータ取得1

SRAdbパッケージを利用すれば、ファイルを1つ1つダウンロードする手間が省けます。但し、GSEから始まるIDは受け付けないので、SRPから始まるID情報を予め入手しておく必要がある。**絶対にやらないで!!**Do NOT perform it.

- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2014/06/25)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2015/02/23) NEW
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu_2013) (last modified 2014/06/26)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/1

① ②

イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu_2013)

SRAdbパッケージを用いてRNA-seq配列を取得するやり方を示します。SRA形式ファイルの場合はNCBIからダウンロードしているようですが、FASTQ形式ファイルの場合はEBIからダウンロードしているようです(2014年6月23日、孫建強氏提供情報)。ここではFASTQファイルをダウンロードするやり方を示します。
「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに移動し以下をコピペ。

1. RNA-seqデータ("SRA000299":Marioni et al., Genome Res., 2008)の実験デザインの全体像やファイルサイズを眺める場合

論文中の記述が
かったうえで下
います:腎臓3.(
(SRX000604)。

```
param <- "S

#必要なパッ
library(SRA

#前処理
#sqlfile <-
sqlfile <-
sra_con <-

#本番(実験デ
hoge <- sra
hoge
```

② ③

3. RNA-seqデータ("SRP017142":Neyret-Kahn et al., Genome Res., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合

論文中の記述からGSE42213を頼りに、RNA-seqデータがGSE42212として収められていることを見出し、その情報からSRP017142にたどり着いています。計6ファイル、合計6Gb程度の容量のファイルがダウンロードされます。東大の有線LANで一時間弱程度かかります。早く終わらせたい場合は、最後のgetFASTQfile関数のオプションを'ftp'から'fasp'に変更すると時間短縮可能です。

```
param <- "SRP017142"                          #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)                                #パッケージの読み込み

#前処理
#sqlfile <- "SRAmetadb.sqlite"                #最新でなくてもよく、手元に予めダウンロードして
sqlfile <- getSRAdbFile()                     #最新のSRAmetadb SQLiteファイルをダウンロード
sra_con <- dbConnect(SQLite(), sqlfile)#おまじない

#前処理(実験デザインの全体像を表示)
```

NGSデータ取得1

原著論文中の記述はGSE42213のみ。重要なのは全体像の理解です。GSE42213は、ChIP-seq(GSE42211)とRNA-seq(GSE42212)をまとめたID。

NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display** | Not logged in | Login

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE42213 | GO

Series GSE42213 | Query DataSets for GSE42213

Status	Public on Jul 24, 2013
Title	SUMO is an Integral and Instructive Compone and Senescence
Organism	Homo sapiens
Experiment type	Genome binding/occupancy profiling by high th Expression profiling by high throughput sequen
Summary	This SuperSeries is composed of the SubSeries
Overall design	Refer to individual Series
Citation(s)	Neyret-Kahn H, Benhamed M, Ye T, Le Gras S e governs coordinated repression of a transcriptio growth and proliferation. *Genome Res* 2013 Oc PMID: 23893515
Submission date	Nov 09, 2012
Last update date	Feb 17, 2015
Contact name	Tao YE
Organization name	IGBMC (CNRS/INSERM/UDS)
Street address	1 rue Laurent Fries
City	Illkirch
ZIP/Postal code	67404
Country	France

Platforms (2)	GPL10999	Illumina Genome Analyzer IIx (Homo sapiens)
	GPL11154	Illumina HiSeq 2000 (Homo sapiens)
Samples (26)	GSM1035423	prolif_input_DNA
More...	GSM1035424	prolif_SUMO1_rep1_ChIPSeq
	GSM1035425	prolif_SUMO1_rep2_ChIPSeq

This SuperSeries is composed of the following SubSeries:

GSE42211 Genome wide occupancy of SUMO machinery in proliferative and Ras-induced senescent human primary fibroblasts

GSE42212 Quantitative analysis of proliferative and Ras-induced senescent human primary fibroblasts transcriptomes

①

Relations

BioProject PRJNA179295

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE42213_RAW.tar	2.8 Gb	(http)(custom)	TAR (of BED, WIG)

NGSデータ取得1

RNA-seqデータのみをまとめたID (GSE42212)のページを眺めると、SRP017142というSRAdbパッケージで入力として利用可能なSRP IDを取得できる。

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE42212 GO

Series GSE42212 Query DataSets for GSE42212

Status	Public on Jul 24, 2013
Title	Quantitative analysis of proliferative and Ras-induced senescent human primary fibroblasts transcriptomes
Organism	Homo sapiens
Experiment type	Expression profiling by high throughpu
Summary	The goals of this study is to analyse tr and to compare them with ChIPseq proteins of the SUMO machinery
Overall design	mRNA profiling of proliferative versus fibroblasts 5 days post-infection
Contributor(s)	Neyret-Kahn H, Benhamed M, Ye T, Le Dasso M, Seeler J, Davidson I, Dejean
Citation(s)	Neyret-Kahn H, Benhamed M, Ye T, Le governs coordinated repression of a tra growth and proliferation. *Genome Res* PMID: 23893515
Submission date	Nov 09, 2012
Last update date	Dec 22, 2014
Contact name	Tao YE
Organization name	IGBMC (CNRS/INSERM/UDS)
Street address	1 rue Laurent Fries
City	Illkirch
ZIP/Postal code	67404
Country	France

Platforms (1) GPL10999 Illumina Genome Analyzer IIx (Homo sapiens)

Samples (6) GSM1035443 Proliferatives_rep1_RNAseq
More... GSM1035444 Proliferatives_rep2_RNAseq
GSM1035445 Proliferatives_rep3_RNAseq

This SubSeries is part of SuperSeries:
GSE42213 SUMO is an Integral and Instructive Component of Chromatin in Cell Growth and Senescence

Relations

BioProject	PRJNA179305
SRA	SRP017142

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE42212_RNASeq_RPKM.txt.gz	11.4 Kb	(ftp)(http)	TXT
SRP/SRP017/SRP017142		(ftp)	SRA Study

Raw data provided as supplementary file

Processed data is available on Series record

NGSデータ取得1

ここではフォルダ中に何もない状態でNGSデータをダウンロードする手順を示す。**絶対にやらないで!!**

- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2014/06/25)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2015/02/23) NEW
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu_2013) (last modified 2014/06/26)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)

イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu_2013)

SRAdbパッケージを用いてRNA-seq配列を取得するやり方を示します。SRA形式ファイルの場合はNCBIからダ

3. RNA-seqデータ("SRP017142":Neyret-Kahn et al., Genome Res., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

論文中の記述からGSE42213を頼りに、RNA-seqデータがGSE42212として収められていることを見出し、その情報からSRP017142にたどり着いています。計6ファイル、合計6Gb程度の容量のファイルがダウンロードされます。東大の有線LANで一時間弱程度かかります。早く終わらせたい場合は、最後のgetFASTQfile関数のオプションを'ftp'から'fasp'に変更すると時間短縮可能です。

```
param <- "SRP017142"                    #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)

#前処理
#sqlfile <- "SRA
sqlfile <- getSR
sra_con <- dbCon

#前処理(実験デザイ
```

R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/mapping_srp017142"
> list.files()
character(0)
>
```

NGSデータ取得1

Windows IEでの全選択は、CTRLとALTキーを押しながらコードの枠内で左クリック。コピペして数分後の状態。SRAmetadb.sqlite.gzというファイルのダウンロードで結構時間がかかる。これはNCBI SRA中の全メタデータ情報を含んでいる。そのおかげで、SRP017142を入力として与えるだけで関連する全サンプルのFASTQファイルの情報を自動的に取り出すことができる。

3. RNA-seqデータ("SRP017142":Nevret-Kahn et al., Genome Res., 2013)のgzip圧縮済みのFAS をダウンロードする場合:

論文中の記述からGSE42213を頼りに、RNA-seqデータがGSE42212として収められていることを見 情報からSRP017142にたどり着いています。計6ファイル、合計6GB程度の容量のファイルがダウ ます。東大の有線LANで一時間弱程度かかります。早く オプションを'ftp'から'fasp'に変更すると時間短縮可能です

```
param <- "SRP017142"                          #取得
#必要なパッケージをロード
library(SRAdb)                                #パ
#前処理
#sqlfile <- "SRAmetadb.sqlite"                #最新
sqlfile <- getSRAdbFile()                     #最新
sra_con <- dbConnect(SQLite(), sqlfile)#お
#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#
hoge                                          #hog
apply(hoge, 2, unique)                        #hog
getFASTQinfo(in_acc=hoge$run)                 #hog
#本番(FASTQファイルのダウンロード)
getFASTQfile(hoge$run, srcType='ftp')  #hog
```

R Console

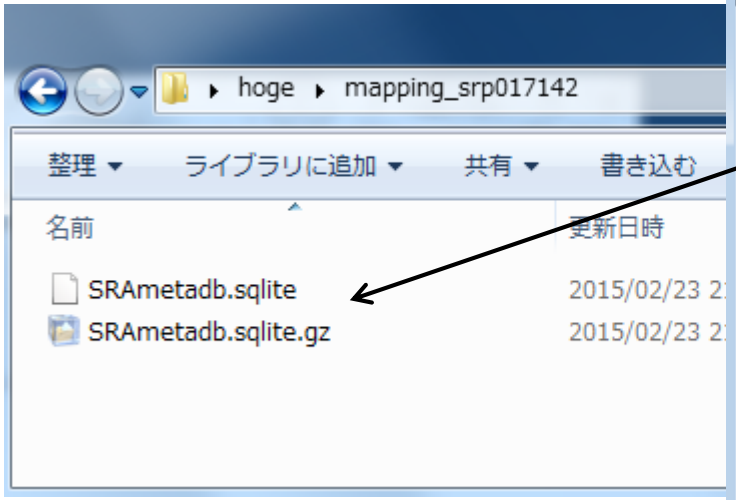
```
> library(SRAdb)
 要求されたパッケージ RSQLite
 要求されたパッケージ DBI をロ
 要求されたパッケージ graph
 次のパッケージを付け加えます: 'graph'
The
 要
 要求されたパッケージ bitops をロード中です
Setting options('download.file.method.GEOquery'='aut$
>
> #前処理
> #sqlfile <- "SRAmetadb.sqlite"          #最新でなく$
> sqlfile <- getSRAdbFile()               #最新のSRAm$
 URL 'http://gbnci.abcc.ncifcrf.gov/backup/SRAmetadb$
Content type 'application/x-gzip' length 776702957 b$
 開かれた URL
```

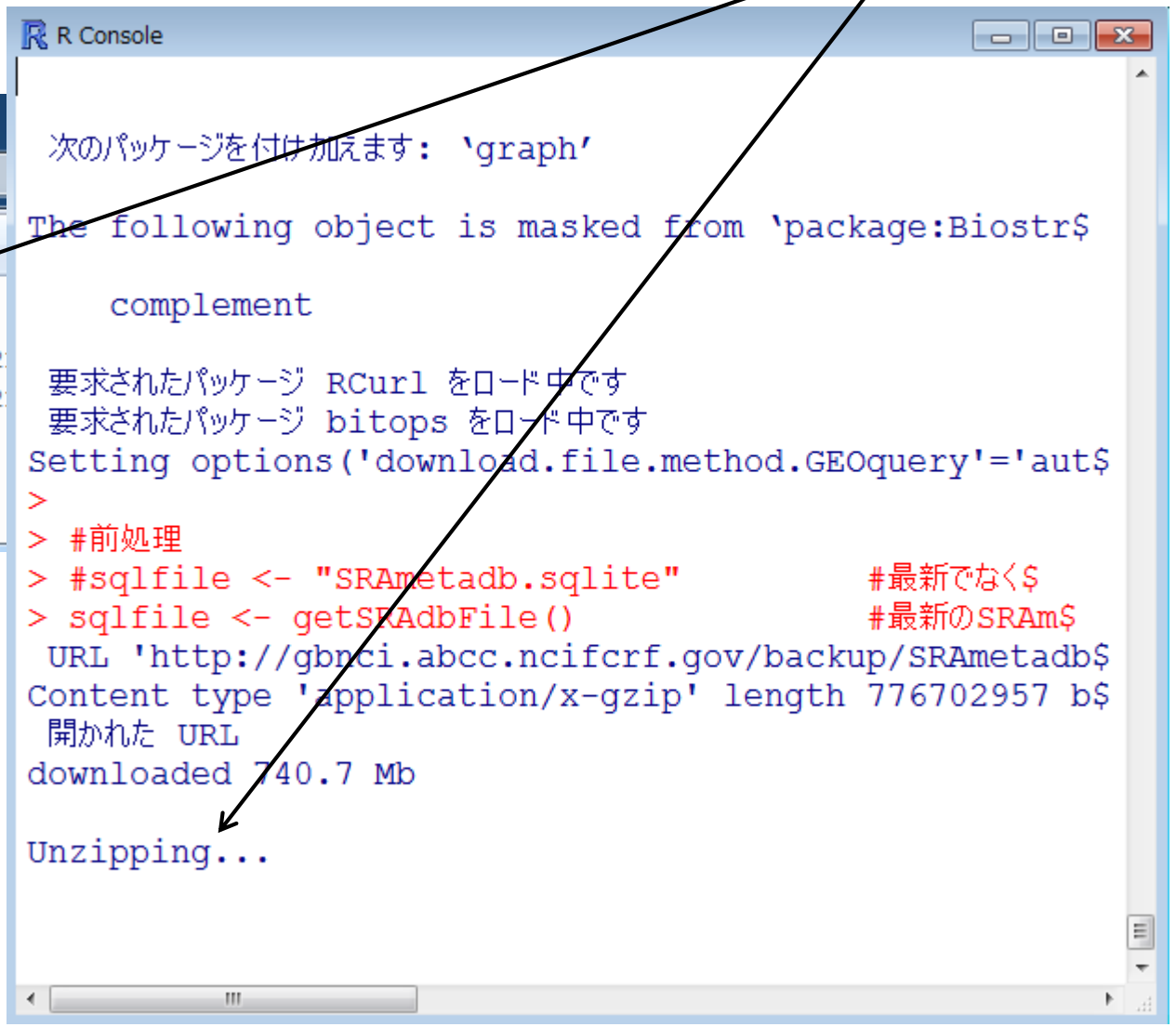
15% downloaded

URL: http://gbnci.abcc.ncifcrf.gov/backup/SRAmetadb.sqlite.gz

NGSデータ取得1

SRAmetadb.sqlite.gzというファイルのダウンロードが終了し、解凍(unzip)しているところ。フォルダ上でも解凍されていることがわかる。





NGSデータ取得1

SRAdbパッケージは、sra形式ファイルの場合はSRAに、fastq形式ファイルの場合はENAにアクセスするようになっているようです。①ftpでSRR*.fastq.gzファイルのURLリストを自動的に取得できていることがわかります。

3. RNA-seqデータ("SRP017142":Nevret-Kahn et al., Genome Res., 2013)のgzip圧縮済みのFAST
をダウンロードする場合:

論文中の記述からGSE42213を頼りに、RNA-seqデータがGSE42212として収められていることを見
情報からSRP017142にたどり着いています。計6ファイル、合計6Gb程度の容量のファイルがダウン
ます。東大の有線LANで一時間弱程度かかります。早く終わらせたい場合は、最後のgetFASTQfi
オプションを'ftp'から'fasp'に変更すると時間短縮可能です。

```
param <- "SRP017142"

#必要なパッケージをロー
library(SRAdb)

#前処理
#sqlfile <- "SRAmetadb
sqlfile <- getSRAdbFil
sra_con <- dbConnect(S

#前処理(実験デザインの全
hoge <- sraConvert(par
hoge
apply(hoge, 2, unique)
getFASTQinfo(in_acc=ho

#本番(FASTQファイルのダ
getFASTQfile(hoge$run,
```

R Console

```
> getFASTQinfo(in_acc=hoge$run)          #hoge$runで指定したSRRから始ま$
        run submission     study    sample experiment fastq_ID
1 SRR616151  SRA061444 SRP017142 SRS375081  SRX204181  1212951
2 SRR616152  SRA061444 SRP017142 SRS375082  SRX204182  1212952
3 SRR616153  SRA061444 SRP017142 SRS375083  SRX204183  1212953
4 SRR616154  SRA061444 SRP017142 SRS375084  SRX204184  1212954
5 SRR616155  SRA061444 SRP017142 SRS375085  SRX204185  1212955
6 SRR616156  SRA061444 SRP017142 SRS375086  SRX204186  1212956
                                                                ftp
1 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616151/SRR616151.fastq.gz
2 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616152/SRR616152.fastq.gz
3 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616153/SRR616153.fastq.gz
4 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616154/SRR616154.fastq.gz
5 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616155/SRR616155.fastq.gz
6 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616156/SRR616156.fastq.gz
                                 md5      bytes
1 79f7663a958458f8fe77e3707f2602d5 1694201684
2 1fd9743365db4c76381a0c5cab018dcd 1657696979
3 b11cd4c2b9e96ff8da2b6e7294461e0c 1452944898
4 8e0a7d2d8b5a096385b07ae76a4dafe8 1025583201
```

①

ダウンロード途中経過

NGSデータ取得1

3. RNA-seqデータ("SRP017142":Nevret-Kahn et al., Genome Res., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

論文中の記述からGSE42213を頼りに、RNA-seqデータがGSE42212として収められていることを見出し、その情報からSRP017142にたどり着いています。計6ファイル、合計6Gb程度の容量のファイルがダウンロードされます。東大の有線LANで一時間弱程度かかります。早く終わらせたい場合は、最後のgetFASTQfile関数のオプションを'ftp'から'fasp'に変更すると時間短縮可能です。

```
param <- "SRP017142"

#必要なパッケージをロー
library(SRAdb)

#前処理
#sqlfile <- "SRAmetadb
sqlfile <- getSRAdbFil
sra_con <- dbConnect(S

#前処理(実験デザインの全
hoge <- sraConvert(par
hoge
apply(hoge, 2, unique)
getFASTQinfo(in_acc=ho

#本番(FASTQファイルのダ
getFASTQfile(hoge$run,
```

R Console

```
2 2014-01-15 05:59:49 2015-02-21 19:29:02
3 2014-01-15 05:52:10 2015-02-21 19:29:02
4 2014-01-15 05:31:20 2015-02-21 19:29:02
5 2014-01-15 05:53:12 2015-02-21 19:29:02
6 2014-01-15 05:48:28 2015-02-21 19:29:02
>
> #本番(FAST
> getFASTQf
Files are sa
'C:/Users/ka

 URL 'ftp://
ftp data connection made, file length 1694201684 bytes
 開かれた URL
downloaded 1615.7 Mb

 URL 'ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616152/SRR616152.fas$
ftp data connection made, file length 1657696979 bytes
 開かれた URL
```

45% downloaded

URL: ... sra.ebi.ac.uk/vol1/fastq/SRR616/SRR616152/SRR616152.fastq.gz

NGSデータ取得1

ダウンロード終了後の状態。
list.files()実行結果からもわかるように、確かに目的のgzip圧縮FASTQファイルが6個存在する。

3. RNA-seqデータ("SRP017142":Neyret-Kahn et al., Genome Res., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

論文中の記述からGSE42213を頼りに、RNA-seqデータがGSE42212として収められていることを見出し、その情報からSRP017142にたどり着いています。計7ファイル、合計6GB程度の容量のファイルがダウンロードされます。東大の有線LANオプションを'ftp'から'fa

```
param <- "SRP017

#必要なパッケージ
library(SRAdb)

#前処理
#sqlfile <- "SRA
sqlfile <- getSR
sra_con <- dbCon

#前処理(実験デザイ
hoge <- sraConve
hoge
apply(hoge, 2, u
getFASTQinfo(in_

#本番(FASTQファイ
getFASTQfile(hog
```

R Console

```
                                  md5       bytes           audit_time
1 79f7663a958458f8fe77e3707f2602d5 1694201684 2014-01-15 06:01:01
2 1fd9743365db4c76381a0c5cab018dcd 1657696979 2014-01-15 05:59:49
3 b11cd4c2b9e96ff8da2b6e7294461e0c 1452944898 2014-01-15 05:52:10
4 8e0a7d2d8b5a096385b07ae76a4dafe8 1025583201 2014-01-15 05:31:20
5 3c3fefcef86d12cdd8769ebb5395e41b 1482959798 2014-01-15 05:53:12
6 2a15030e622a5b939ccd172262dcb2f4 1372463724 2014-01-15 05:48:28
         sradb_updated
1 2015-02-21 19:29:02
2 2015-02-21 19:29:02
3 2015-02-21 19:29:02
4 2015-02-21 19:29:02
5 2015-02-21 19:29:02
6 2015-02-21 19:29:02
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/mapping_SRP017142"
> list.files()
[1] "SRAmetadb.sqlite"   "SRR616151.fastq.gz" "SRR616152.fastq.gz"
[4] "SRR616153.fastq.gz" "SRR616154.fastq.gz" "SRR616155.fastq.gz"
[7] "SRR616156.fastq.gz"
>
```

NGSデータ取得1

gzip圧縮状態でも6個のFASTQファイルだけで8GB以上になっていることがわかる。ここまでで、RNA-seqデータ取得が完了。SRAmetadb.sqliteファイルは後述するやり方で再利用可能。

3. RNA-seqデータ("SRP017142":Nevret-Kahn et al., Genome Res., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

論文中の記述からGSE42213を頼りに、RNA-seqデータがGSE42212として収められていることを見出し情報からSRP017142にたどり着いています。計6ファイル、合計6Gb程度の容量のファイルがダウンロードされます。東大の有線LANで一時間弱程度かかります。早く終わらせたい場合は、最後のgetFASTQfile関数のオプションを'ftp'から'fasp'に変更すると時間短縮可能です。

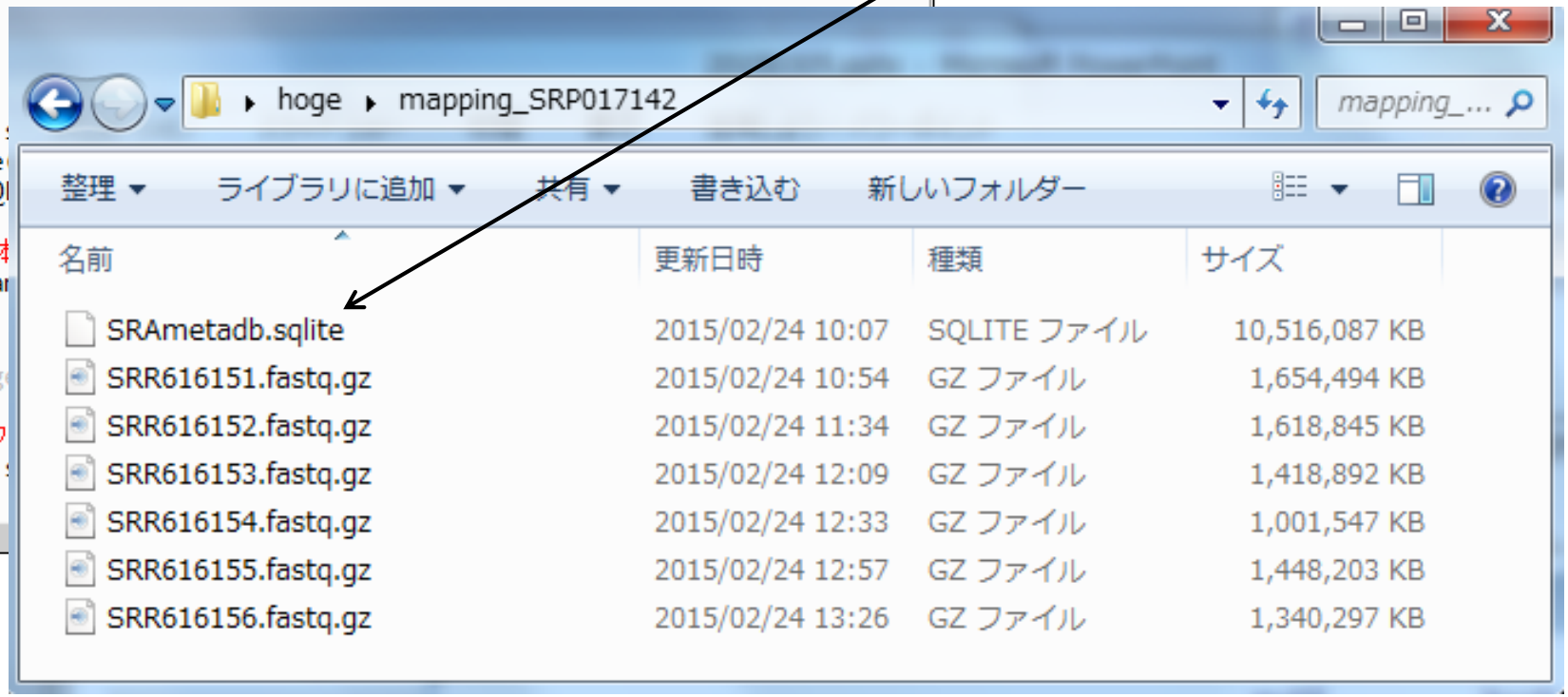
```
param <- "SRP017142"                    #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)

#前処理
#sqlfile <- "SRAmetadb.
sqlfile <- getSRAdbFile
sra_con <- dbConnect(SQ

#前処理(実験デザインの全体
hoge <- sraConvert(para
hoge
apply(hoge, 2, unique)
getFASTQinfo(in_acc=hog

#本番(FASTQファイルのダウ
getFASTQfile(hoge$run,
```



Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

NGSデータ取得2

これも見るだけですが、ファイルサイズが比較的小さいので復習する例題としてはおすすめです。ここでは以前にダウンロードしておいたSRAmetadb.sqliteファイルを再利用するやり方を示す。

- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/11/12)
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) (last modified 2014/06/25)
- イントロ | NGS | 配列取得 | FASTQ or SRA | 公共DBから (last modified 2015/02/23) NEW
- イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu_2013) (last modified 2014/06/26)
- イントロ | NGS | 配列取得 | シミュレーションデータ | について (last modified 2015/01/18)
- イントロ | NGS | 配列取得 | シミュレーションデータ | ランダムな塩基配列の生成から (last modified 2015/01/18)

①

イントロ | NGS | 配列取得 | FASTQ or SRA | SRAdb(Zhu_2013)

SRAdbパッケージを用いてRNA-seq配列を取得するやり方を示します。SRA形式ファイルの場合はNCBIからダウンロードしているようですが、FASTQ形式ファイルの場合はEBIからダウンロードしているようです(2014年6月23日、孫建強氏提供情報)。ここではFASTQファイルをダウンロードするやり方を示します。
「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに移動し以下をコピペ。

②

1. RNA-seqデータ("SRA000299": Marioni et al., Genome Res., 2008)の実験デザインの全体像やファイルサイズを眺める場合:

論文中の記述からGSE19600を頼りに、SRP001538にたどり着いています。したがって、ここで指定するのは"SRP001538"となります。以下を実行して得られるRNA-seqファイルは4つ(SRX000603, SRX000604, SRX000605, SRX000606)。

```
param <- "S

#必要なパッ
library(SRA

#前処理
#sqlfile <-
sqlfile <-
sra_con <-

#本番(実験デ
hoge <- sra
hoge
```

③

7. カイコのsmall RNA-seqデータ("SRP016842": Nie et al., BMC Genomics, 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

論文中の記述からGSE41841を頼りに、SRP016842にたどり着いています。したがって、ここで指定するのは"SRP016842"となります。以下を実行して得られるsmall RNA-seqファイルは1つ(SRR609266.fastq.gz)で、ファイルサイズは400Mb弱、11928428リードであることがわかります。

```
param <- "SRP016842"                          #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)                                #パッケージの読み込み

#前処理
#sqlfile <- "SRAmetadb.sqlite"                #最新でなくてもよく、手元に予めダウンロード
sqlfile <- getSRAdbFile()                     #最新のSRAmetadb SQLiteファイルをダウンロ
sra_con <- dbConnect(SQLite(), sqlfile)#おまじない

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#paramで指定したSRA IDに付随するstudy
hoge                                          #hogeの中身を表示
```

NGSデータ取得2

テンプレートコードをテキストエディタにコピペして、必要な箇所を変更。こうすることで、またgetSRAdbFile関数で改めて取得することなく、手元のSRAmetadb.sqliteファイルを再利用することができる。How to reuse the huge “SRAmetadb.sqlite” file in your local PC.

7. カイコのsmall RNA-seqデータ("SRP016842": Nie et al., BMC Genomics, 2013)のgz
FASTQファイルをダウンロードする場合:

論文中の記述からGSE41841を頼りに、SRP016842にたどり着いています。したがって
のは"SRP016842"となります。以下を実行して得られるsmall RNA-seqファイルは1つ
(SRR609266.fastq.gz)で、ファイルサイズは400Mb弱、11928428リードであることがわか

```
param <- "SRP016842"                    #取得したいSRA IDを指定

#必要なパッケージをロード
library(SRAdb)

#前処理
#sqlfile <- "SRAmetadb.sqlite"
sqlfile <- getSRAdbFile()
sra_con <- dbConnect(SQLite(), sqlfile)

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_co
hoge
```

```
param <- "SRP016842"                    #取得したいSRA IDを指定
↓
#必要なパッケージをロード
library(SRAdb)                          #パッケージの読み込み
↓
#前処理
sqlfile <- "SRAmetadb.sqlite"           #最新でなくてもよく、手元に
#sqlfile <- getSRAdbFile()               #最新のSRAmetadb SQLiteフ
sra_con <- dbConnect(SQLite(), sqlfile)#おまじない
↓
#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#paramで指定したSRA ID
hoge                                    #hogeの中身を表示
apply(hoge, 2, unique)                  #hoge行列の列ごとにユニー
getFASTQinfo(in_acc=hoge$run)           #hoge$runで指定したSRRか
↓
#本番(FASTQファイルのダウンロード)
getFASTQfile(hoge$run, srcType='ftp')   #hoge$runで指定したSRRか
```

NGSデータ取得2

作業ディレクトリはどこでもいいが、当然のことながらSRAmetadb.sqliteファイルが作業ディレクトリにあるという前提。

```
param <- "SRP016842"                    #取得したいSRA IDを指定
↓
#必要なパッケージをロード
library(SRAdb)                          #パッケージの読み込み
↓
#前処理
sqlfile <- "SRAmetadb.sqlite"           #最新でなくてもよく、手元に
#sqlfile <- getSRAdbFile()               #最新のSRAmetadb.sqliteフ
sra_con <- dbConnect(SQLite(), sqlfile)#おま
↓
#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#
hoge                                    #hoge
apply(hoge, 2, unique)                  #hoge
getFASTQinfo(in_acc=hoge$run)           #hoge
↓
#本番(FASTQファイルのダウンロード)
getFASTQfile(hoge$run, srcType='ftp')   #hoge
```

```
R Console

  配布条件の詳細に関しては、'license()' あるいは 'l$

R は多くの貢献者による共同プロジェクトです。
  詳しくは 'contributors()' と入力してください。
  また、R や R のパッケージを出版物で引用する際の形$
  'citation()' と入力してください。

  'demo()' と入力すればデモをみることができます。
  'help()' とすればオンラインヘルプが出ます。
  'help.start()' で HTML ブラウザによるヘルプがみら$
  'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge/SRP016842"
> list.files()
[1] "SRAmetadb.sqlite"
>
```

NGSデータ取得2

途中経過。SRR016842.fastq.gzのダウンロードのところまで一気に進んでいることが分かる。

```
param <- "SRP016842"                    #取得したいSRA IDを指定
↓
#必要なパッケージをロード
library(SRAdb)                          #パッケージの読み込み
↓
#前処理
sqlfile <- "SRAmetadb.sqlite"           #最新
#sqlfile <- getSRAdbFile()              #最
sra_con <- dbConnect(SQLite(), sqlfile)#おま
↓
#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#
hoge                                    #hoge
apply(hoge, 2, unique)                  #hoge
getFASTQinfo(in_acc=hoge$run)           #hoge
↓
#本番(FASTQファイルのダウンロード)
getFASTQfile(hoge$run, srcType='ftp')   #hoge
```

R Console

```
> getFASTQinfo(in_acc=hoge$run)          #hoge$run$
       run submission     study    sample
1 SRR609266  SRA060712 SRP016842 SRS373230
  experiment fastq_ID
1  SRX201604  1209776
                                                $
1 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR609/SRR609$
                                 md5     bytes
1 d1c7189d85d491a58f0de4f991babfeb 383840833
          audit_time       sradb_updated
1 2013-03-10 00:24:17 2015-06-01 23:02:32
```

86% downloaded

URL: ... sra.ebi.ac.uk/vol1/fastq/SRR609/SRR609266/SRR609266.fastq.gz

```
ftp data connection made, file length 383840833 by$
```

NGSデータ取得2

終了後の状態。無事fastq.gzファイルのダウンロードができてるっぽい

```
param <- "SRP016842"                    #取得したいSRA IDを指定
↓
#必要なパッケージをロード
library(SRAdb)                          #パッ
↓
#前処理
sqlfile <- "SRAmetadb.sqlite"           #最新
#sqlfile <- getSRAdbFile()               #最
sra_con <- dbConnect(SQLite(), sqlfile)#おま
↓
#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#
hoge                                    #hoge
apply(hoge, 2, unique)                  #hoge
getFASTQinfo(in_acc=hoge$run)           #hoge
↓
#本番(FASTQファイルのダウンロード)
getFASTQfile(hoge$run, srcType='ftp')   #hoge
```

R Console

```
> #本番(FASTQファイルのダウンロード)
> getFASTQfile(hoge$run, srcType='ftp')  #hoge$run$
Files are saved to:
'C:/Users/kadota/Desktop/hoge/SRP016842'

 URL 'ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR609/SR$
ftp data connection made, file length 383840833 by$
downloaded 366.1 MB

        run submission     study    sample
1 SRR609266  SRA060712 SRP016842 SRS373230
  experiment fastq_ID
1  SRX201604  1209776
                                                  $
1 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR609/SRR609$
                                 md5     bytes
1 d1c7189d85d491a58f0de4f991babfeb 383840833
          audit_time       sradb_updated
1 2013-03-10 00:24:17 2015-06-01 23:02:32
> list.files()
[1] "SRAmetadb.sqlite"   "SRR609266.fastq.gz"
>
```

確認(check)

①FASTQを提供する公共DB側のファイルサイズ情報と②ダウンロード後の手元のファイルサイズを見比べて同じ値になっていることを確認。

7. カイコのsmall RNA-seqデータ("SRP016842": Nie et al., B
FASTQファイルをダウンロードする場合:

論文中の記述からGSE41841を頼りに、SRP016842にたどり
のは"SRP016842"となります。以下を実行して得られるsma
(SRR609266.fastq.gz)で、ファイルサイズは400Mb弱、11928

```
param <- "SRP016842"                          #取得

#必要なパッケージをロード
library(SRAdb)                                #パッ

#前処理
#sqlfile <- "SRAmetadb.sqlite"                #最新
sqlfile <- getSRAdbFile()                     #最新(
sra_con <- dbConnect(SQLite(), sqlfile)#おま

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#p
hoge                                          #hoge
apply(hoge, 2, unique)                        #hoge
getFASTQinfo(in_acc=hoge$run)                 #hoge

#本番(FASTQファイルのダウンロード)
getFASTQfile(hoge$run, srcType='ftp')  #hoge
```

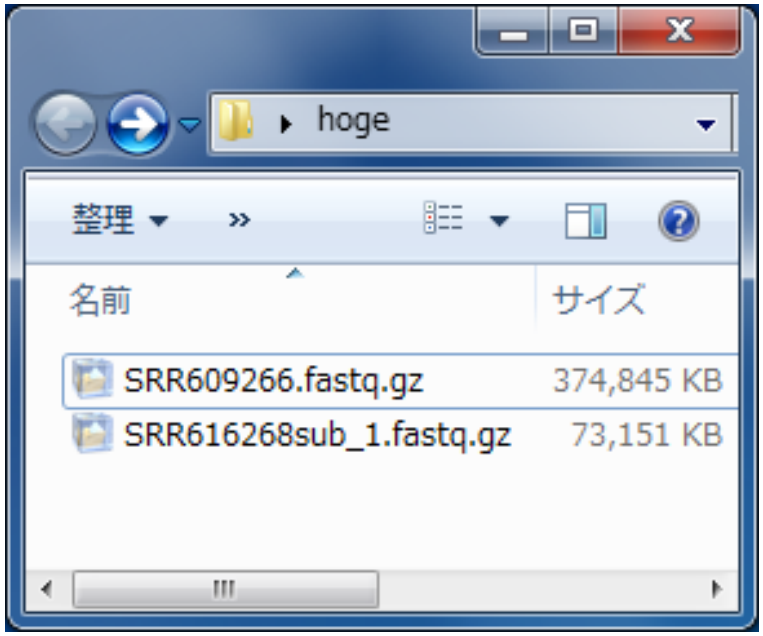
R Console

```
> list.files()
[1] "SRAmetadb.sqlite"    "SRR609266.fastq.gz"
> list.files()[2]
[1] "SRR609266.fastq.gz"
> getFASTQinfo(in_acc=hoge$run)
        run submission     study     sample
1 SRR609266  SRA060712 SRP016842 SRS373230
  experiment fastq_ID
1  SRX201604  1209776
                                                    $
1 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR609/SRR6092$
                                  md5     bytes
1 d1c7189d85d491a58f0de4f991babfeb 383840833 ①
          audit_time       sradb_updated
1 2013-03-10 00:24:17 2015-06-01 23:02:32
> file.info(list.files()[2])
                        size isdir mode
SRR609266.fastq.gz 383840833 FALSE  666
                        ②          mtime
SRR609266.fastq.gz 2015-06-15 17:16:52
                                   ctime
SRR609266.fastq.gz 2015-06-15 17:13:19
                                   atime exe
SRR609266.fastq.gz 2015-06-15 17:13:19  no
>
```

確認(check)

- 真の値
- SRR609266.fastq.gz
  - 383,840,833 bytes
- SRR616268sub_1.fastq.gz
  - 74,906,576 bytes

デスクトップにコピーしたhogeフォルダ中の2つのfastq.gzのファイルサイズを提供側（つまり門田）のものと同一かどうかを確認し、違っていたら別のUSBメモリからコピーし直そう。



R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
[1] "SRR609266.fastq.gz"
[2] "SRR616268sub_1.fastq.gz"
> file.info(list.files())
                          size isdir mode
SRR609266.fastq.gz     383840833 FALSE  666
SRR616268sub_1.fastq.gz 74906576 FALSE  666
                                    mtime
SRR609266.fastq.gz      2015-06-15 17:16:52
SRR616268sub_1.fastq.gz 2015-05-11 16:52:40
                                    ctime
SRR609266.fastq.gz      2015-06-15 17:58:46
SRR616268sub_1.fastq.gz 2015-06-15 18:08:28
                                    atime exe
SRR609266.fastq.gz      2015-06-15 17:58:46  no
SRR616268sub_1.fastq.gz 2015-06-15 18:08:28  no
> |
```

Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

Quality Control (QC)

クオリティーコントロール(Quality Control)の最初のステップがクオリティーチェック(Quality Check)なのかなと最近思い始めています。

- イントロ | ファイル形式の変換 | qseq --> Illumina FASTQ (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | qseq --> Sanger FASTQ (last modified 2013
- 前処理 | クオリティコントロール | について (last modified 2015/05/04)
- 前処理 | クオリティチェック | QuasR(Gaidatzis, 2015) (last modified 2015/06
- 前処理 | クオリティチェック | qrqc (last modified 2014/07/17)
- 前処理 | クオリティチェック | PHREDスコアに変換 (last modified 2013/06/18
- 前処理 | クオリティチェック | 配列長分布を調べる (last modified 2013/06/18
- 前処理 | フィルタリング | PHREDスコアが低い塩基をNに置換 (last modifie
- 前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去 (last mod
- 前処理 | フィルタリング | ACGTのみからなる配列を抽出 (last modified 201
- 前処理 | フィルタリング | ACGT以外のcharacter"-"をNに変換 (last modifie
- 前処理 | フィルタリング | ACGT以外の文字数が閾値以下の配列を抽出 (l
- 前処理 | フィルタリング | 重複のない配列セットを作成 (last modified 2013/
- 前処理 | フィルタリング | 指定した長さ以上の配列を抽出 (last modified 201

前処理 | クオリティコントロール | について

数億～数十億リードからなるNGSデータの全体的な精度チェック、クオリティの低いリードのフィルタリング、リードに含まれるアダプター配列やクオリティの低い配列部分の除去(トリミング)などを実行する様々な方法をリストアップします。Krakenなどアダプター配列除去などが行えるものも含みます。FaQCs (Lo and Chain, 2014)など比較的最近のものはpaired-endリードの処理にデフォルトで対応しています。

R用:

- qrqc: 原著論文なし
- PIQA: Martinez-Alcantara et al., Bioinformatics, 2009
- ShortRead: Morgan et al., Bioinformatics, 2009
- girafe: Toedling et al., Bioinformatics, 2010
- QuasR: Gaidatzis et al., Bioinformatics, 2015

R以外:

- FastQC: 原著論文なし
- FASTX-Toolkit: 原著論文なし
- SolexaQA: Cox et al., BMC Bioinformatics, 2010
- Quake: Kelley et al., Genome Biol., 2010
- NGSQC: Dai et al., BMC Genomics, 2010
- Cutadapt: Martin, M., EMBnet.journal, 2011
- PRINSEQ: Schmieder and Edwards, Bioinformatics, 2011
- ECHO: Kao et al., Genome Res., 2011
- Btrim: Kong Y., Genomics, 2011
- Hammer: Medvedev et al., Bioinformatics, 2011
- ConDeTri: Smeds et al., PLoS One, 2011
- BIGpre: Zhang et al., Genomics Proteomics Bioinformatics, 2011
- NGS QC Toolkit: Patel et al., PLoS One, 2012
- RobiNA: Lohse et al., Nucleic Acids Res., 2012
- SEQuel: Ronen et al., Bioinformatics, 2012
- AdapterRemoval: Lindgreen S., BMC Res Notes, 2012
- Slim-Filter: Golovko et al., BMC Bioinformatics, 2012
- HTQC: Yang et al., BMC Bioinformatics, 2013
- QC-Chain: Zhou et al., PLoS One, 2013
- Kraken: Davis et al., Methods, 2013
- AlienTrimmer: Criscuolo and Brisse, Genomics, 2013
- NextClip: Leggett et al., Bioinformatics, 2014
- QTrim (Roche/454などのlong read用): Shrestha et al., BMC Bioinformatics, 2014
- Trimmomatic: Bolger et al., Bioinformatics, 2014
- Skewer: Jiang et al., BMC Bioinformatics, 2014
- FaQCs: Lo and Chain, BMC Bioinformatics, 2014

Review、ガイドライン、パイプライン系:

- Review: Paszkiewicz et al., Front Genet., 2014

Qual

前処理 | クオリティコントロール | について

数億～数十億リードからなるNGSデータの全体的な精度チェック、クオリティの低いリードのフィルタリング、リードに含まれるアダプター配列やクオリティの低い配列部分の除去（トリミング）などを実行する様々な方法をリストアップします。Krakenなどアダプター配列除去などが行えるものも含みます。FaQCs (Lo and Chain, 2014)など比較的最近のものはpaired-endリードの処理にデフォルトで対応しています。

R用:

- qrqc: 原著論文なし
- PIQA: Martinez-Alcantara et al., Bioinformatics, 2009
- ShortRead: Morgan et al., Bioinformatics, 2009
- girafe: Toedling et al., Bioinformatics, 2010
- QuasR: Gaidatzis et al., Bioinformatics, 2015

R以外:

- FastQC: 原著論文なし
- FASTX-Toolkit: 原著論文なし
- SolexaQA: Cox et al., BMC Bioinformatics, 2010
- Quake: Kelley et al., Genome Biol., 2010
- NGSQC: Dai et al., BMC Genomics, 2010
- Cutadapt: Martin, M., EMBnet.journal, 2011
- PRINSEQ: Schmieder and Edwards, Bioinformatics, 2011
- ECHO: Kao et al., Genome Res., 2011
- Btrim: Kong Y., Genomics, 2011
- Hammer: Medvedev et al., Bioinformatics, 2011
- ConDeTri: Smeds et al., PLoS One, 2011
- BIGpre: Zhang et al., Genomics Proteomics Bioinformatics, 2011
- NGS QC Toolkit: Patel et al., PLoS One, 2012
- RobiNA: Lohse et al., Nucleic Acids Res., 2012
- SEQuel: Ronen et al., Bioinformatics, 2012
- AdapterRemoval: Lindgreen S., BMC Res Notes, 2012
- Slim-Filter: Golovko et al., BMC Bioinformatics, 2012
- HTQC: Yang et al., BMC Bioinformatics, 2013
- QC-Chain: Zhou et al., PLoS One, 2013
- Kraken: Davis et al., Methods, 2013
- AlienTrimmer: Criscuolo and Brisse, Genomics, 2013
- NextClip: Leggett et al., Bioinformatics, 2014
- QTrim (Roche/454などのlong read用): Shrestha et al., BMC Bioinformatics, 2014
- Trimmomatic: Bolger et al., Bioinformatics, 2014
- Skewer: Jiang et al., BMC Bioinformatics, 2014
- FaQCs: Lo and Chain, BMC Bioinformatics, 2014

FASTQ形式ファイルを入力として全体像を眺める作業。FastQCが有名だが、Rパッケージもいくつかある。FastQC is famous.

Quality Control (QC)

実験デザインや使用する機器にもよるが様々な前処理が行われるようです

- 作業内容
  - クオリティチェック（quality check）
  - フィルタリング（filtering）
    - クオリティ値の低い塩基やリードの除去
    - rRNAやtRNAの除去
  - トリミング（trimming）
    - 最初の35塩基のみ利用など
  - 重複除去（de-duplication）
  - コンタミ（contamination）
  - バーコード配列（barcoding）
  - アダプター配列除去（adapter removal）
  - …

前処理 | クオリティコントロール | について

数億～数十億リードからなるNGSデータの全体的な精度チェック、クオリティの低いリードのフィルタリング、リードに含まれるアダプター配列やクオリティの低い配列部分の除去（トリミング）などを実行する様々な方法をリストアップします。Krakenなどアダプター配列除去などが行えるものも含みます。FaQCs (Lo and Chain, 2014)など比較的最近のものはpaired-endリードの処理にデフォルトで対応しています。

R用:

- qrqc: 原著論文なし
- PIQA: Martinez-Alcantara et al., Bioinformatics, 2009
- ShortRead: Morgan et al., Bioinformatics, 2009
- girafe: Toedling et al., Bioinformatics, 2010
- QuasR: Gaidatzis et al., Bioinformatics, 2015

R以外:

- FastQC: 原著論文なし
- FASTX-Toolkit: 原著論文なし
- SolexaQA: Cox et al., BMC Bioinformatics, 2010
- Quake: Kelley et al., Genome Biol., 2010
- NGSQC: Dai et al., BMC Genomics, 2010
- Cutadapt: Martin, M., EMBnet.journal, 2011
- PRINSEQ: Schmieder and Edwards, Bioinformatics, 2011
- ECHO: Kao et al., Genome Res., 2011
- Btrim: Kong Y., Genomics, 2011
- Hammer: Medvedev et al., Bioinformatics, 2011
- ConDeTri: Smeds et al., PLoS One, 2011
- BIGpre: Zhang et al., Genomics Proteomics Bioinformatics, 2011
- NGS QC Toolkit: Patel et al., PLoS One, 2012
- RobiNA: Lohse et al., Nucleic Acids Res., 2012
- SEQuel: Ronen et al., Bioinformatics, 2012
- AdapterRemoval: Lindgreen S., BMC Res Notes, 2012
- Slim-Filter: Golovko et al., BMC Bioinformatics, 2012
- HTQC: Yang et al., BMC Bioinformatics, 2013
- QC-Chain: Zhou et al., PLoS One, 2013
- Kraken: Davis et al., Methods, 2013
- AlienTrimmer: Criscuolo and Brisse, Genomics, 2013
- NextClip: Leggett et al., Bioinformatics, 2014
- QTrim (Roche/454などのlong read用): Shrestha et al., BMC Bioinformatics, 2014
- Trimmomatic: Bolger et al., Bioinformatics, 2014
- Skewer: Jiang et al., BMC Bioinformatics, 2014
- FaQCs: Lo and Chain, BMC Bioinformatics, 2014

Review、ガイドライン、パイプライン系:

- Review: Paszkiewicz et al., Front Genet., 2014

QuasR

このウェブページではQuasRとqrqcパッケージのやり方を示している。QuasRは安定して動くが、情報量がイマイチ。gzip圧縮ファイルに対応しているが、Winでは正常に動作する。Macは2014年受講生は圧縮ファイルに対応していないと報告したが、2015年受講生は普通にfastq.gzを読み込めたとのこと。qrqcは、動作が不安定且つ圧縮ファイルの入力に対応していない。This page shows the usage of two R packages (QuasR and qrqc). However, those have some disadvantages compared to FastQC.

- イントロ | ファイル形式の変換 | qseq --> Illumina FASTQ (l
- イントロ | ファイル形式の変換 | qseq --> Sanger FASTQ (la
- 前処理 | クオリティコントロール | について (last modified 20
- 前処理 | クオリティチェック | QuasR(Gaidatzis 2015) t m
- 前処理 | クオリティチェック | qrqc (last modified 2014/07/17)
- 前処理 | クオリテ
- 前処理 | クオリテ
- 前処理 | フィルタ
- 前処理 | フィルタ
- 前処理 | フィルタ
- 前処理 | フィルタ
- 前処理 | フィルタ
- 前処理 | フィルタ

前処理 | クオリティチェック | QuasR

QuasRパッケージを用いてQCレポート
相当ストイックな出力結果です…。
「ファイル」-「ディレクトリの変更」で解析

1. サンプルデータ7のgzip圧縮FASTQ

SRR037439から得られるFASTQファイ

```
in_f <- "SRR037439.fastq.gz"
out_f <- "hoge1.pdf"

#必要なパッケージをロード
library(QuasR)

#本番
qQCReport(in_f, pdfFilename=
```

2. サンプルデータ7の非圧縮FASTQ形

SRR037439から得られるFASTQファイ

```
in_f <- "SRR037439.fastq"
out_f <- "hoge2.pdf"

#必要なパッケージをロード
library(QuasR)
```

3. FASTQ形

カイコsmall RNA-seqデータ(Nie et al., BMC Genomics, 2013; 約375MB)です。

```
in_f <- "SRR609266.fastq.gz"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.pdf"                  #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(QuasR)                        #パッケージの読み込み

#本番
qQCReport(in_f, pdfFilename=out_f)    #pdfレポートファイルの作成
```

4. FASTQ形式ファイル(SRR616268sub_1.fastq.gz)の場合:

乳酸菌RNA-seqデータSRR616268の最初の100万リード分(約73MB)です。

```
in_f <- "SRR616268sub_1.fastq.gz"     #入力ファイル名を指定してin_fに格納
out_f <- "hoge4.pdf"                  #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(QuasR)                        #パッケージの読み込み

#本番
qQCReport(in_f, pdfFilename=out_f)    #pdfレポートファイルの作成
```

QuasR

3. FASTQ形式ファイル(SRR609266.fastq.gz)の場合:

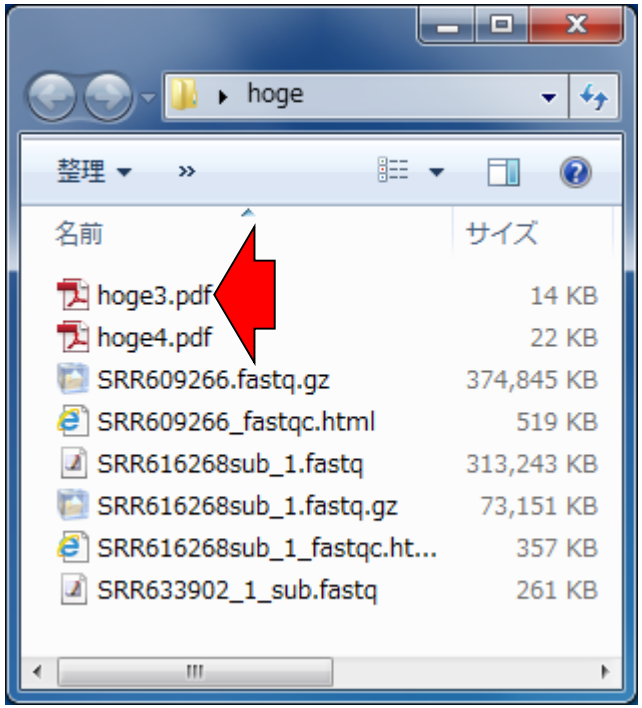
カイコsmall RNA-seqデータ(Nie et al., BMC Genomics, 2013; 約375MB)です。

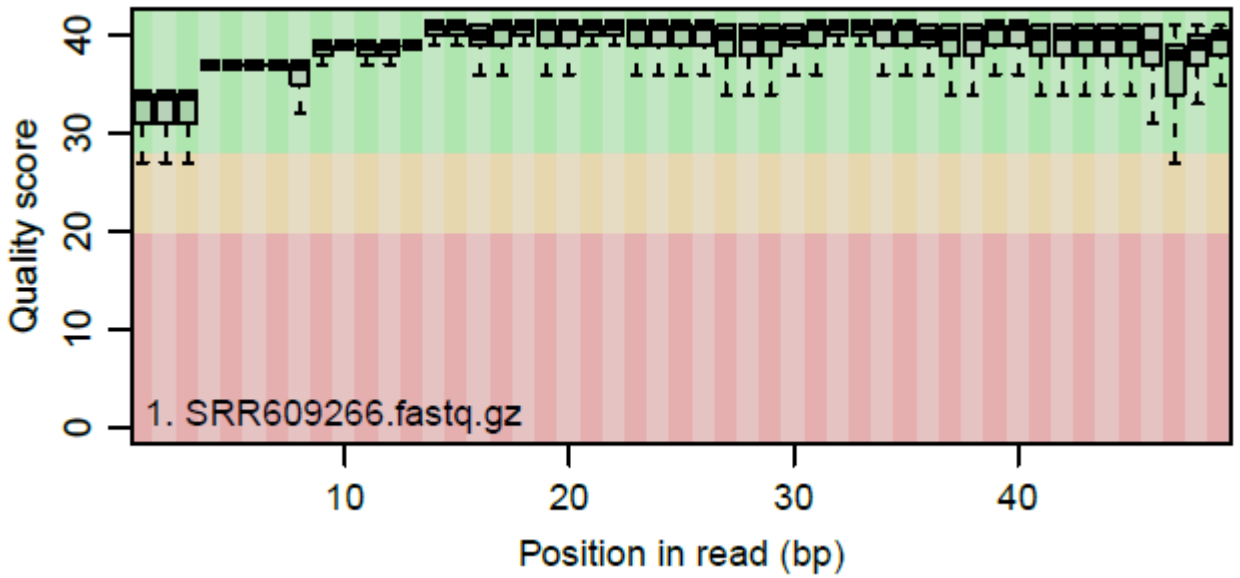
```
in_f <- "SRR609266.fastq.gz"          #入力ファイル名を指定し
out_f <- "hoge3.pdf"                  #出力ファイル名を指定し

#必要なパッケージをロード
library(QuasR)                        #パッケージの読み込み

#本番
qQCReport(in_f, pdfFilename=out_f)    #pdfレポートファイルの作
```

横軸の範囲から、このデータ(SRR609266)の配列長は50 bp弱なのだろうと推測できる。縦軸がquality score。値が大きいほどqualityが高い。一つの目安は20。このデータの場合、ほとんどのリードがどのポジションでもscore > 30はあると判断できる。Horizontal axis indicates the positon in read, implying about 50 bp length for this data. Vertical axis indicates the quality score: the higher score corresponds to accurate base calling.





QuasR

4. FASTQ形式ファイル(SRR616268sub_1.fastq.gz)の場合:

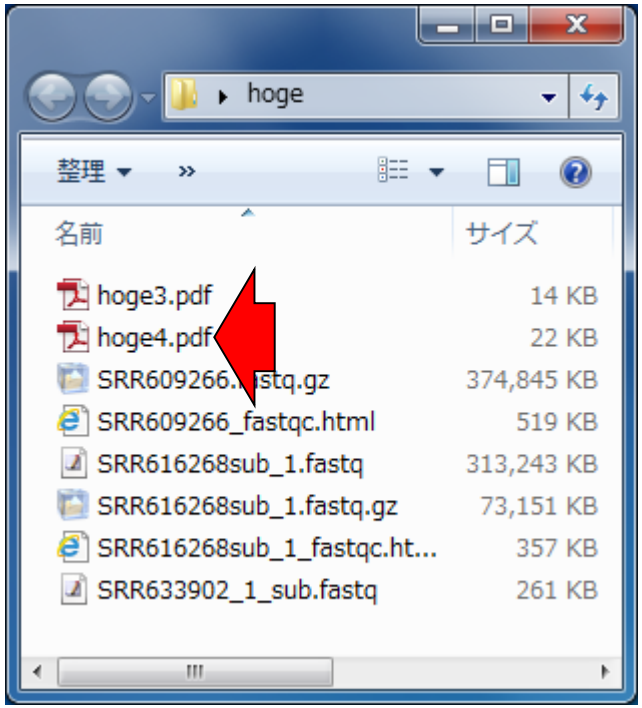
乳酸菌RNA-seqデータSRR616268の最初の100万リード分(約73MB)です。

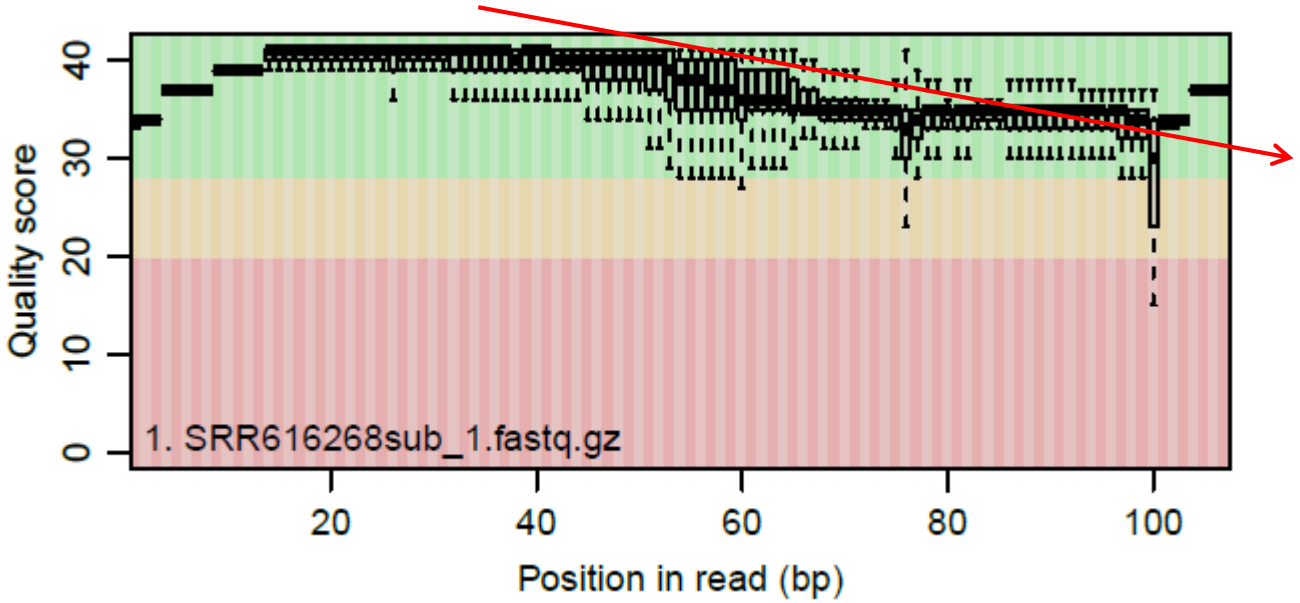
```
in_f <- "SRR616268sub_1.fastq.gz"      #入力ファイル名を指定してin_
out_f <- "hoge4.pdf"                   #出力ファイル名を指定してout

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み

#本番
qQCReport(in_f, pdfFilename=out_f)     #pdfレポートファイルの作成
```

これは乳酸菌RNA-seqデータ。長さが100 bpちょっとであることが分かる。赤矢印で示すように読み進めるにしたがって、徐々にquality scoreが下がる傾向にあることがわかる。これが一般的。This lactobasillus (←乳酸菌) RNA-seq data has about 107 bp in length. As shown in the red arrow, the quality scores tend to gradually decrease. This is general tendency for sequencer.





Contents

- イントロダクション(Introduction)
  - NGSデータ概観（single-end; PacBio; SRP038897） by NCBI SRA (SRA)
  - NGSデータ概観（single-end; PacBio; SRP038897） by DDBJ SRA (DRA)
  - NGSデータ概観（single-end; Illumina; GSE36469） by ENA, DRA, and SRA
  - NGSデータ概観（paired-end; Illumina; GSE42960）
- 公共DBとファイル形式(Public database and file format)
  - 課題1
  - SRA, DRA, ENA。.sraと.fastq。今後の方向性や雑感
  - NGSデータ取得（SRAdbパッケージ）
    - SRP017142
    - SRP016842
- QC(Quality Control)
  - データの全体像を眺めるQuality Check
    - QuasRパッケージ
    - FastQC
    - 課題2, 3

FastQC ver. 0.11.3

FastQC実行結果はhtmlファイルとして用意してあります。自力で実行したいヒトは、乳酸菌学会誌の連載で自習。

- イントロ | ファイル形式の変換 | qseq --> Illumina FASTQ (last modified 2013/06/17)
- イントロ | ファイル形式の変換 | qseq --> Sanger FASTQ (last modified 2013
- 前処理 | クオリティコントロール | について (last modified 2015/05/04)
- 前処理 | クオリティチェック | QuasR(Gaidatzis, 2015) (last modified 2015/06
- 前処理 | クオリティチェック | qrqc (last modified 2014/07/17)
- 前処理 | クオリティチェック | PHREDスコアに変換 (last modified 2013/06/18
- 前処理 | クオリティチェック | 配列長分布を調べる (last modified 2013/06/18
- 前処理 | フィルタリング | PHREDスコアが低い塩基をNに置換 (last modifie
- 前処理 | フィルタリング | PHREDスコアが低い配列(リード)を除去 (last mod
- 前処理 | フィルタリング | ACGTのみからなる配列を抽出 (last modified 201

①

前処理 | クオリティコントロール | について

数億～数十億リードからなるNGSデータの全体的な精度チェック、クオリティの低いリードのフィルタリング、リードに含まれるアダプター配列やクオリティの低い配列部分の除去(トリミング)などを実行する様々な方法をリストアップします。Krakenなどアダプター配列除去などが行えるものも含みます。FaQCs (Lo and Chain, 2014)など比較的最近のものはpaired-endリードの処理にデフォルトで対応しています。

R用:

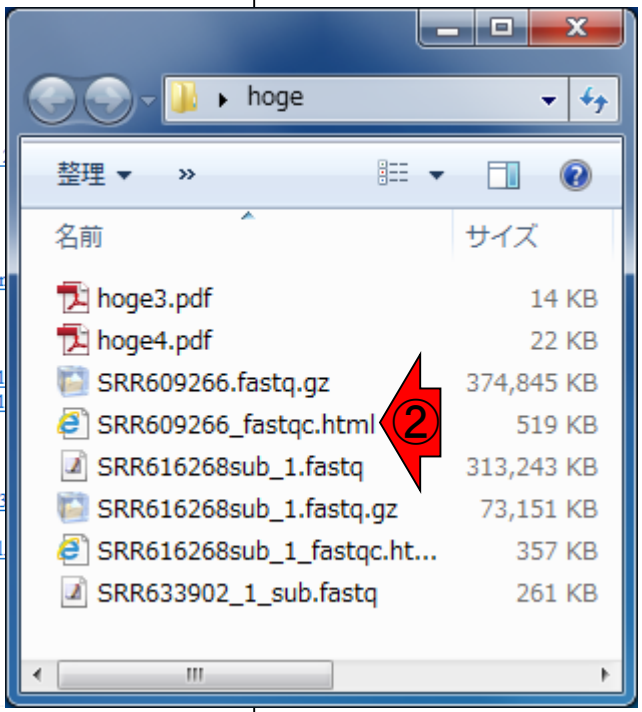
- qrqc: 原著論文なし
- PIQA: Martinez-Alcantara et al., Bioinformatics, 2009
- ShortRead: Morgan et al., Bioinformatics, 2009
- girafe: Toedling et al., Bioinformatics, 2010
- QuasR: Gaidatzis et al., Bioinformatics, 2015

R以外:

- FastQC: 原著論文なし
- FASTX-Toolkit: 原著論文なし
- SolexaQA: Cox et al., BMC Bioinformatics, 2010
- Quake: Kelley et al., Genome Biol., 2010
- NGSQC: Dai et al., BMC Genomics, 2010
- Cutadapt: Martin, M., EMBnet.journal, 2011
- PRINSEQ: Schmieder and Edwards, Bioinformatics,
- ECHO: Kao et al., Genome Res., 2011
- Btrim: Kong Y., Genomics, 2011
- Hammer: Medvedev et al., Bioinformatics, 2011
- ConDeTri: Smeds et al., PLoS One, 2011
- BIGpre: Zhang et al., Genomics Proteomics Bioinfor
- NGS QC Toolkit: Patel et al., PLoS One, 2012
- RobiNA: Lohse et al., Nucleic Acids Res., 2012
- SEQuel: Ronen et al., Bioinformatics, 2012
- AdapterRemoval: Lindgreen S., BMC Res Notes, 201
- Slim-Filter: Golovko et al., BMC Bioinformatics, 201
- HTQC: Yang et al., BMC Bioinformatics, 2013
- QC-Chain: Zhou et al., PLoS One, 2013
- Kraken: Davis et al., Methods, 2013
- AlienTrimmer: Criscuolo and Brisse, Genomics, 2013
- NextClip: Leggett et al., Bioinformatics, 2014
- QTrim (Roche/454などのlong read用): Shrestha et al.
- Trimmomatic: Bolger et al., Bioinformatics, 2014
- Skewer: Jiang et al., BMC Bioinformatics, 2014
- FaQCs: Lo and Chain, BMC Bioinformatics, 2014

Review、ガイドライン、パイプライン系:

- Review: Paszkiewicz et al., Front Genet., 2014



②

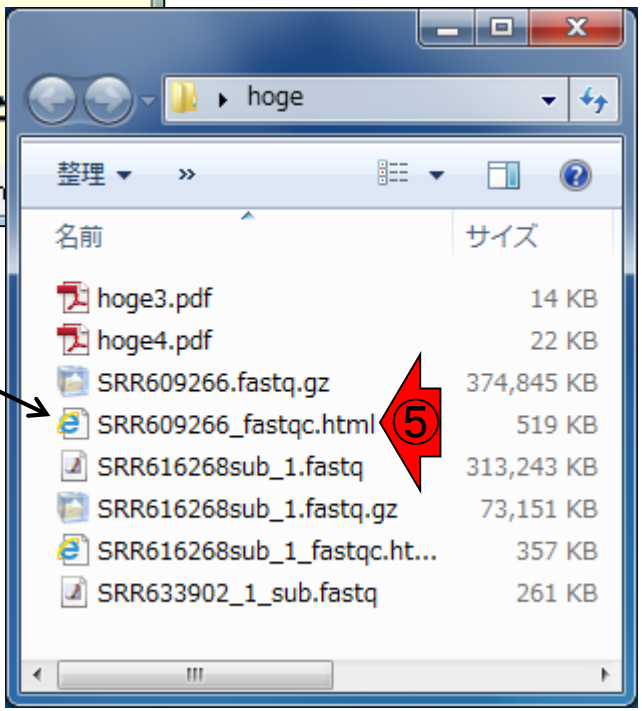
FastQC ver. 0.11.3

①「pwd」=「getwd()」、②「ls SRR609266*」=「list.files(pattern="SRR609266")」。③FastQCプログラムを実行するところ。④再び「ls SRR609266*」とやって、作業ディレクトリ内のSRR609266関連ファイルを表示。FastQC実行結果ファイルが作成されていることが分かる。

BioLinux8 [実行中] - Oracle VM VirtualBox

仮想マシン　ビュー　デバイス　ヘルプ

```
iu@bielinux[~/Desktop/mac_share]
①iu@bielinux[mac_share] pwd                                    [11:44午前]
/home/iu/Desktop/mac_share
②iu@bielinux[mac_share] ls SRR609266*                          [11:44午前]
SRR609266.fastq.gz
③iu@bielinux[mac_share] fastqc2 -q SRR609266.fastq.gz
iu@bielinux[mac_share] ls SRR609266*                           [11:47午前]
SRR609266_fastqc.html  SRR609266.fastq.gz
SRR609266_fastqc.zip                                                    ④
iu@bielinux[mac_share]                                         [11:47午
```



FastQC ver. 0.11.3

①総リード数は11,928,428（約1,200万）。
②リードの長さは49 bp。③GC含量は52%

C:¥Users¥kadota¥Desktop¥hoge¥SRR609266_fastqc.html　SRR609266.fastq.gz Fas...

FastQC Report

Mon 15 Jun 2015
SRR609266.fastq.gz

Summary

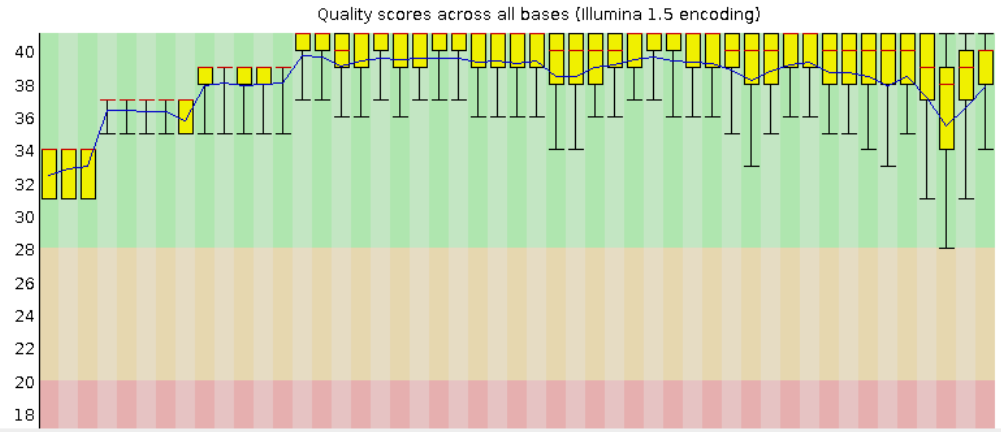
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	SRR609266.fastq.gz
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	11928428
Sequences flagged as poor quality	0
Sequence length	49
%GC	52

①
②
③

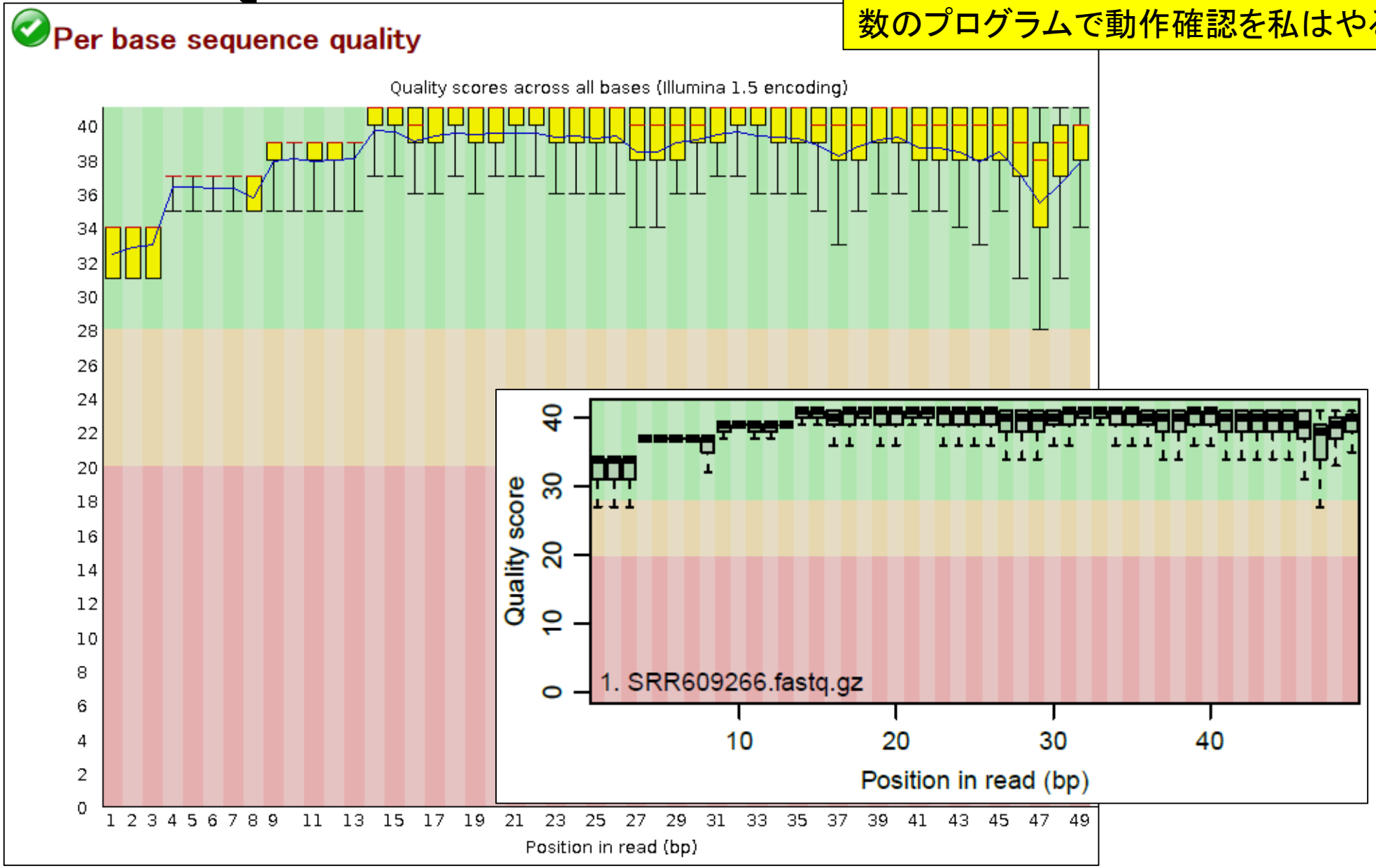
Per base sequence quality



Produced by FastQC (version 0.11.3)

FastQC ver. 0.11.3

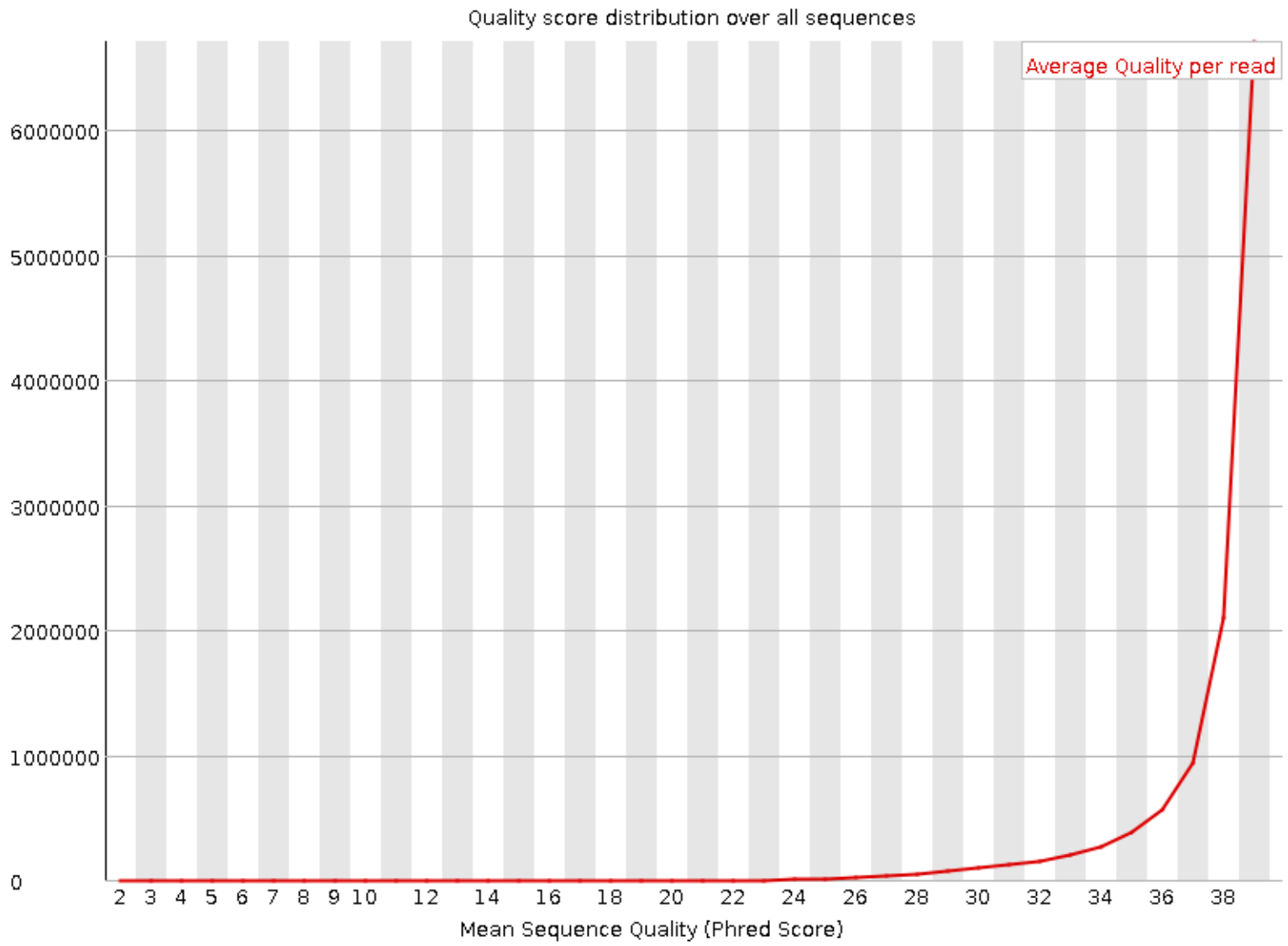
クオリティスコア分布（quality score distribution）。FastQCの結果とQuasRの結果は見た目同じ。このような具合で複数のプログラムで動作確認を私はやる。



FastQC ver. 0.11.3

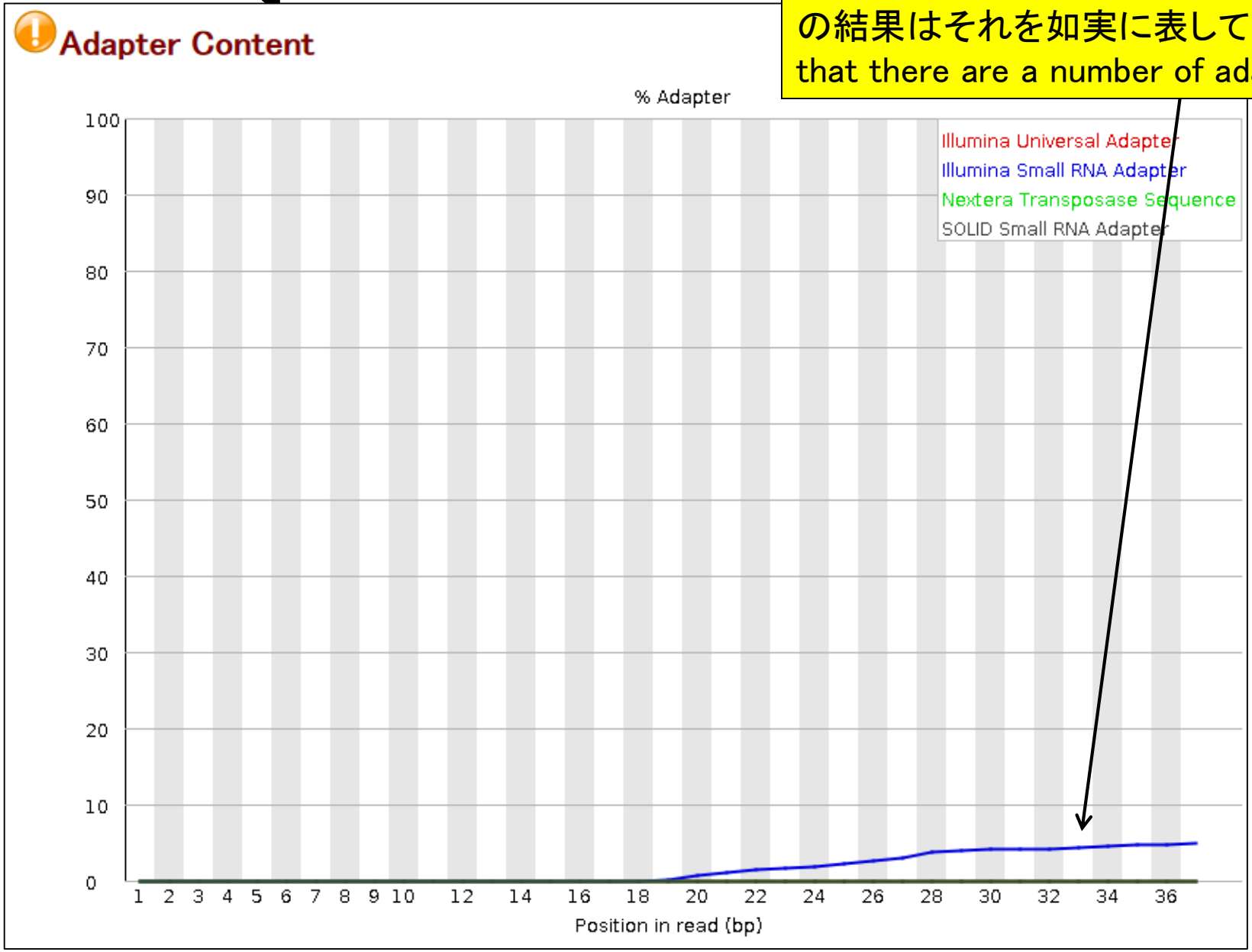
リードごと（per sequence）のクオリティスコア。見方はよく分からないが「average scoreが39のものが最も多いのだろう」くらいは分かる。

Per sequence quality scores



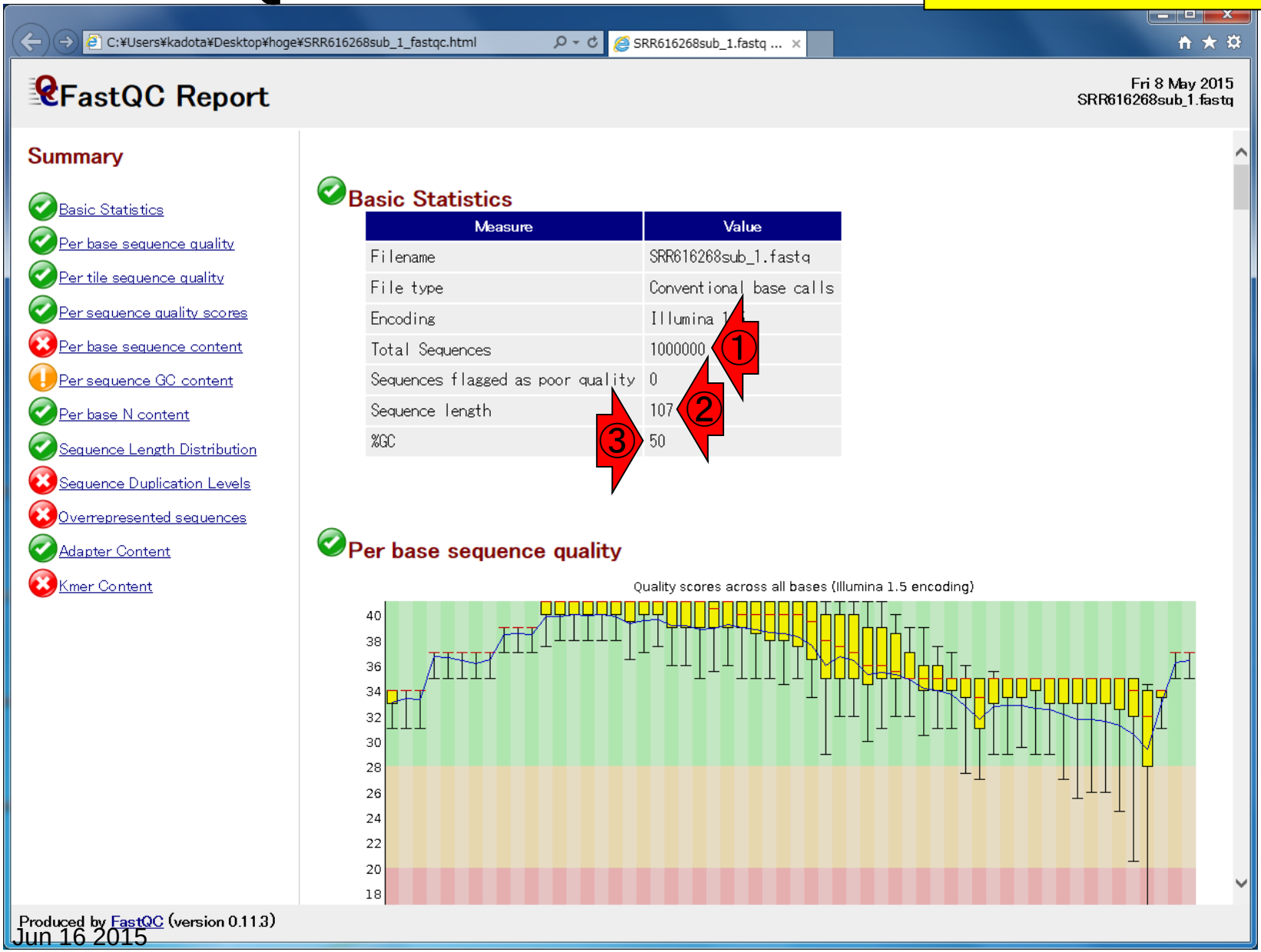
FastQC ver. 0.11.3

これは「カイコsmall RNA-seqデータ」。small RNAは20-40 bp程度の長さなので、リード中に大抵アダプター配列（adapter sequence）が含まれる。この結果はそれを如実に表している。It can be seen that there are a number of adapter sequences.



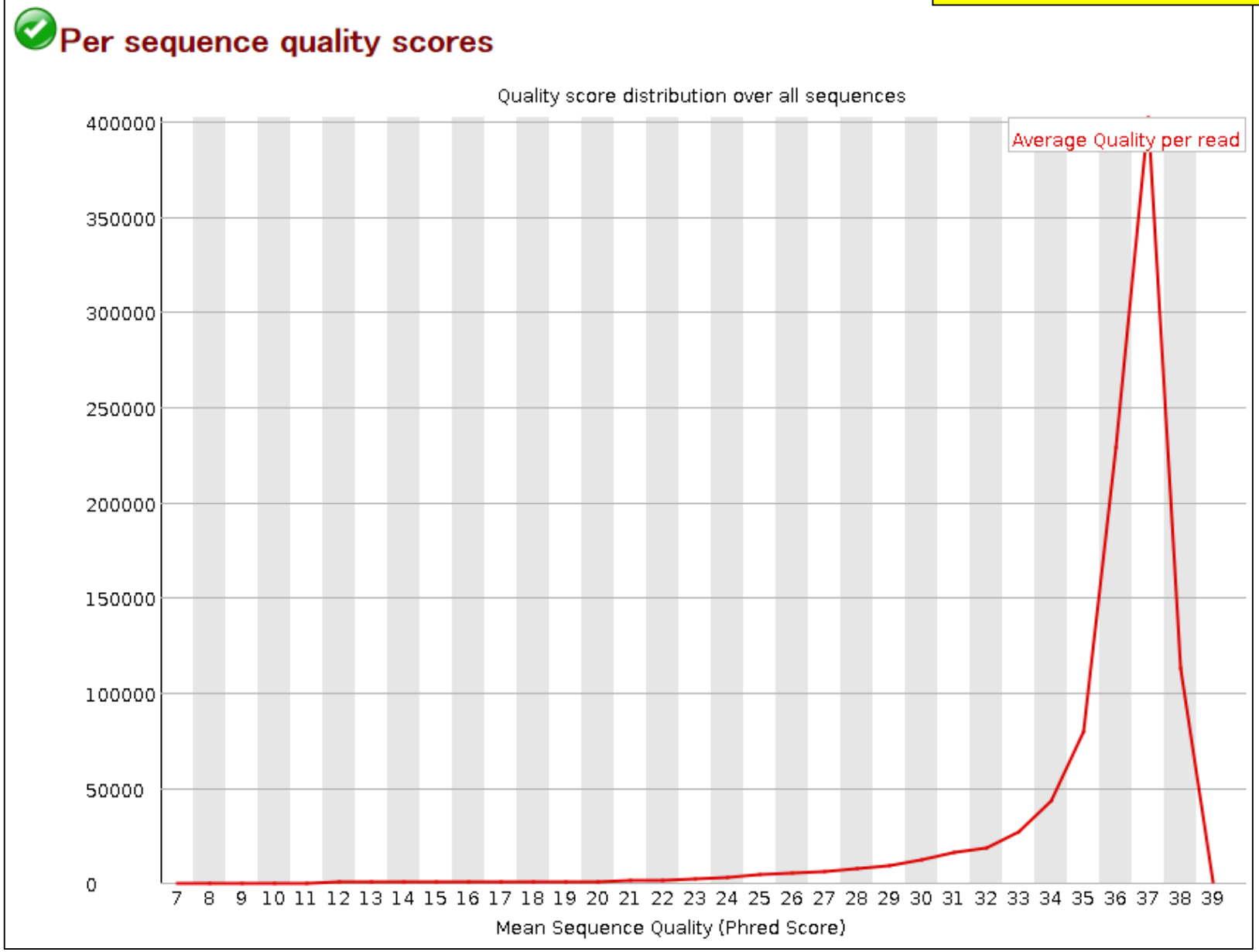
FastQC ver. 0.11.3

課題2。乳酸菌RNA-seqデータ(SRR616268sub_1.fastq)の①総リード数?。②リード長さ?。③GC content?



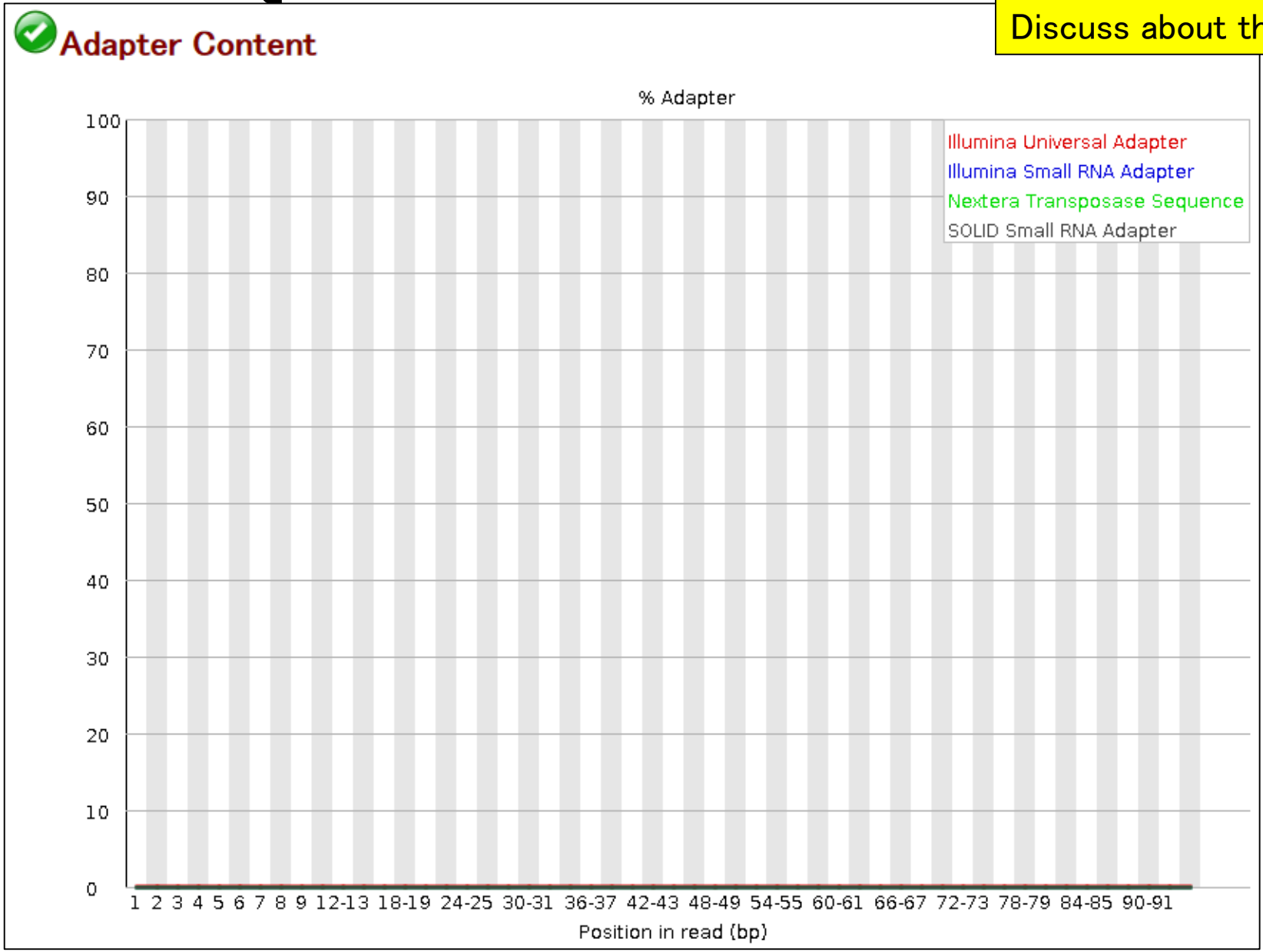
FastQC ver. 0.11.3

課題2。乳酸菌RNA-seqデータ(SRR616268sub_1.fastq)の④average quality score?



FastQC ver. 0.11.3

課題3。乳酸菌RNA-seqデータ(SRR616268sub_1.fastq)中のアダプター配列について簡単に考察せよ。
Discuss about this result.



アグリバイオインフォマティクス

他大学の学生や社会人も受講できる、希少なバイオインフォ教育プログラム

- 人材養成プログラム（科学技術振興調整費: 2004/10-2009/3）
- 教育研究プログラム（特別教育研究経費: 2009/4~2014/3）
- 教育研究プログラム（研究科経費: 2014/4~）

年度	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
修士課程	12	65	73	83	68	72	107	100	121	124	108	
博士課程	3	7	11	13	6	8	12	21	16	19	24	
社会人	5	3	8	4	1	0	11	19	32	26	55	
合計	**20**	**75**	**92**	**100**	**75**	**80**	**130**	**140**	**169**	**169**	**187**	↗
開講科目数	9	15	15	15	15	12	15	15	14	15	13	13
常勤教員数	6	6	7	7	7	3	4	4	3	2	2	2
ポスドク数	>2	>2	>2	>2	>2	1	1	1	1	1	1	0
門田担当コマ数	3	?	?	8	8	5	5	11	13	14	18	20

1科目以上の合格者数