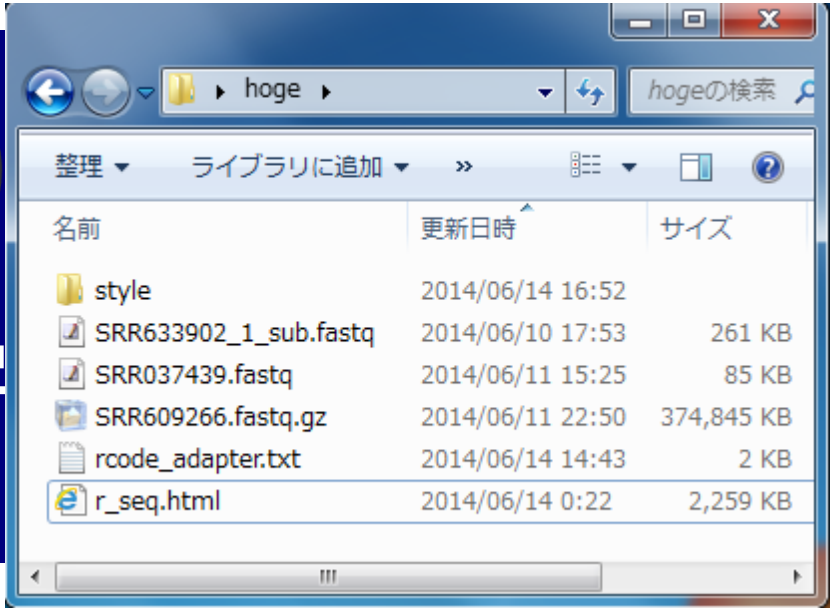
講義室後ろにあるUSBメモリ中のhogeフォルダをデスクトップにコピーしておいてください。



農学生命情報科学特論I 第2回

東京大学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ユニット
門田幸二
kadota@iu.a.u-tokyo.ac.jp

講義予定

- 第1回（2014年6月11日）
  - **西**：NSG概論。現状や展望など。講義のみ
- 第2回（2014年6月18日）
  - **門田**：データベース、データ取得、ファイル形式および変換、前処理
  - 教科書の1.3節周辺
- 第3回（2014年6月25日）
  - **門田**：アセンブル、マッピング、カウント情報取得
  - 教科書の2.3節周辺
- 第4回（2014年7月2日）
  - **門田**：クラスタリング、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

授業の目標・概要

次世代シーケンサ(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シークエンス（RNA-Seq）に基づく（非モデル生物の）トランスクリプトーム解析を題材とした実習を含む講義を行います。

Contents（第2回）

- **イントロダクション(Introduction)**
  - **NGSデータ概観（PacBioとIllumina）**
  - NGSデータベース(DB)、データ形式（FASTQ形式）
  - SRAdbパッケージを用いたデータ取得、エラーへの対処
- 前処理(Pre-processing)
  - qrqcパッケージを用いたQuality Control (QC)
- アダプター配列除去
  - 基本戦略（girafeパッケージ）
  - 昔は正常に動作していたのに…という例（QuasRパッケージ）
  - アダプター除去を含む様々なフィルタリングの組合せ（ShortReadパッケージ）
  - 課題

トランスクリプトームとは

- ある特定の状態の組織や細胞中に存在する全RNA（転写物、transcripts）の総体
- 様々なトランスクリプトーム解析技術
  - マイクロアレイ
    - DNAマイクロアレイ、Affymetrix GeneChip、タイリングアレイなど
  - 配列決定に基づく方法
    - EST、SAGE、CAGE、次世代シーケンサ(RNA-seq)など
  - 電気泳動に基づく方法
    - Differential Display、AFLP、HiCEPなど

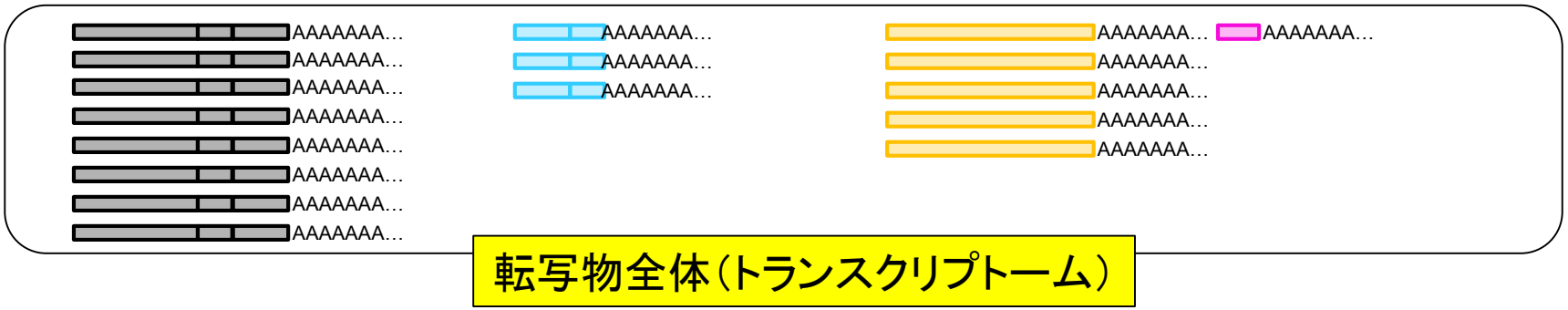
調べたい組織でどの遺伝子がどの程度発現しているのかを一度に観察

トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域

ヒト

・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）

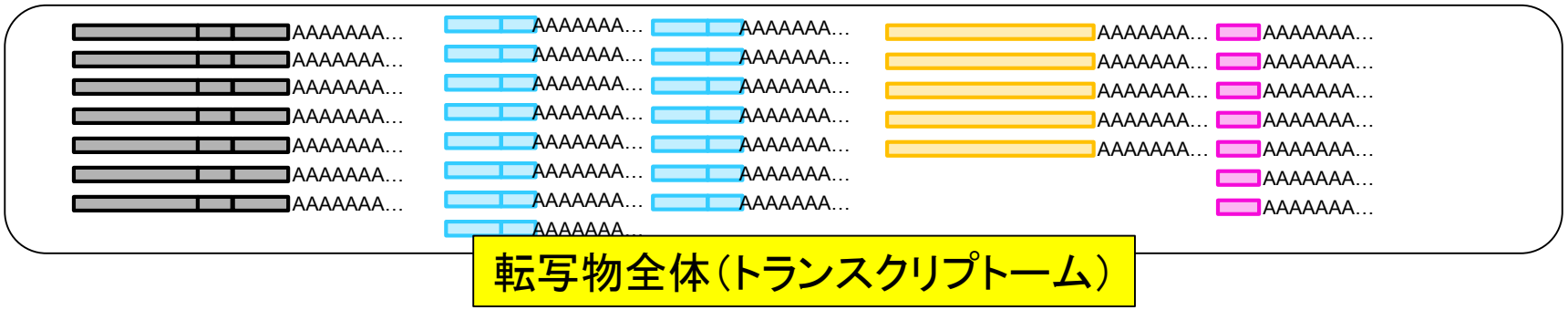


・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域

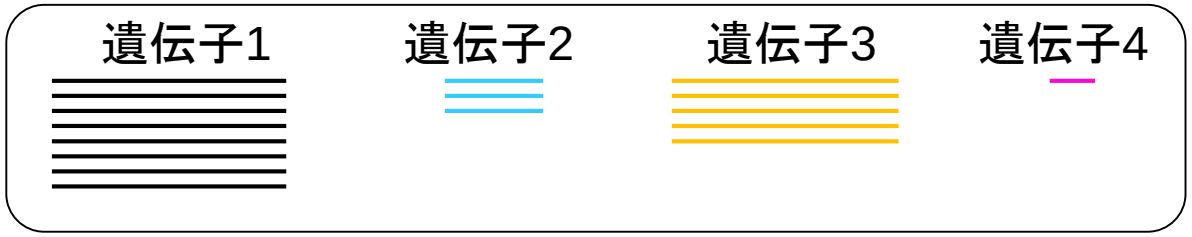
・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）



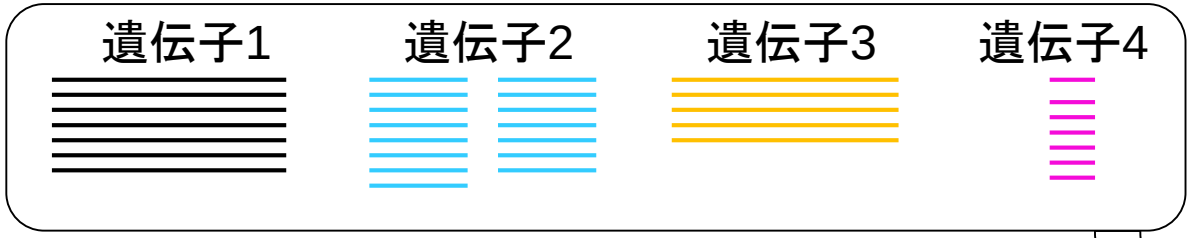
・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

トランスクリプトーム情報を得る手段

- 光刺激前（T1）の目のトランスクリプトーム



- 光刺激後（T2）の目のトランスクリプトーム



これがいわゆる
「遺伝子発現行列」

	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
…	…	…

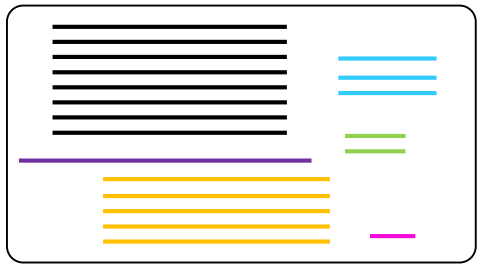
・マイクロアレイ
・RNA-seq

教科書p9の図1−8に示してあるように、実際には「遺伝子 = 転写物」ではない点に注意!

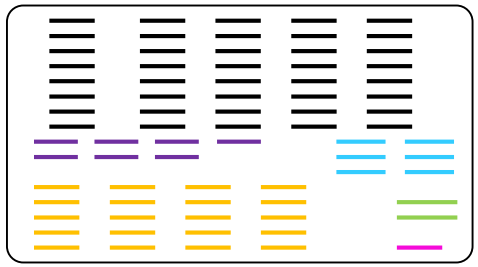
トランスクリプトーム取得（NGS）

- 次世代シーケンサー（Illumina社の場合）

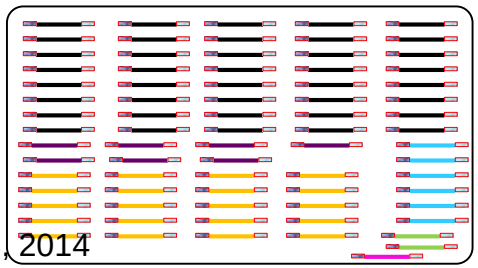
光刺激前（T1）の目のトランスクリプトーム



数百塩基程度
に断片化



アダプター配列
を両末端に付加



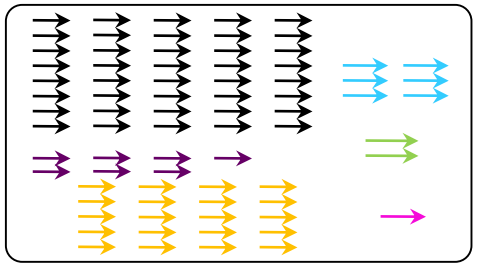
配列決定

・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



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 | BSgenome(last modified 2014/04/25)
- イントロ | 一般 | 配列取得 | プロモーター配列 | GenomicFeatures(Lawrence 2013)(last modified 2014/04/23)
- 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 2014/04/02)
- 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 2013/09/25)
- イントロ | NGS | 様々なプラットフォーム (last modified 2014/06/08) NEW
- イントロ | NGS | qPCRやmicroarrayなどとの比較(last modified 2014/06/05) NEW
- イントロ | NGS | Viewer(last modified 2014/01/29)
- イントロ | NGS | 配列取得 | FA
- イントロ | NGS | 配列取得 | FA
- イントロ | NGS | アノテーション
- イントロ | NGS | アノテーション
- イントロ | NGS | アノテーション
- イントロ | NGS | アノテーション
- イントロ | NGS | アノテーション

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 2500, HiSeq X Ten, ...
- Roche: GS FLX+ System, GS Junior+ System
- Life Technologies: SOLiD
- Life Technologies: Ion PGM System
- Pacific Biosciences: PacBio RS II System
- Dover社など「POLONATOR G.007」
- ...

Pacific Biosciences (PacBio)について:

PacBio RS II Systemは最長で20,000bp以上(平均は4,500 bp)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。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

- 過去のお知らせ (last modified 2014/04/21)
- Rのインストールと起動 (last modified 2014/05/14) NEW
- サンプルデータ (last modified 2014/04/01)
- 書籍 | トランスクリプトームについて (last modified 2014/05/12) NEW
- 書籍 | トランスクリプトーム解析 | 2.3.1 RNA-seqデータ(FASTQファイル) (last modifie

サンプルデータ NEW

1. Illumina/36bp/single-end/human (SRA000299) data (Marioni et al., Genome Res., 200
「Kidney 7 samples vs Liver 7 samples」のRNA-seqの遺伝子発現行列データ(Supplen

25. Illumina HiSeq 2000/200bp?!/single-end/human (SRA062939) data (Chan et al., Hum. Mol. Genet., 2013)
26. Illumina Genome Analyzer II/54bp/single-end/human (SRP017142; GSE42212) data (Neyret-Kahn et al., Genome Res., 2013)
ヒト fibroblastsの2群間比較用データ: 3 proliferative samples vs. 3 Ras samples
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 replic
(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 (xxx) 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 [illegible]RA; 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。

PacBioのロングリードデータも出始めています

30. PacBio/xxx bp/Human (xxx) data (Sharon et al., Nat Biotechnol., 2013)
ヒトの長鎖RNA-seqデータ[illegible]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 SRA) data (Sharon et al., PLoS One, 2014)
32. k-m
(san
50
を1
Cが

• サンプルデータ

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
Abstract	The chicken has long served as an

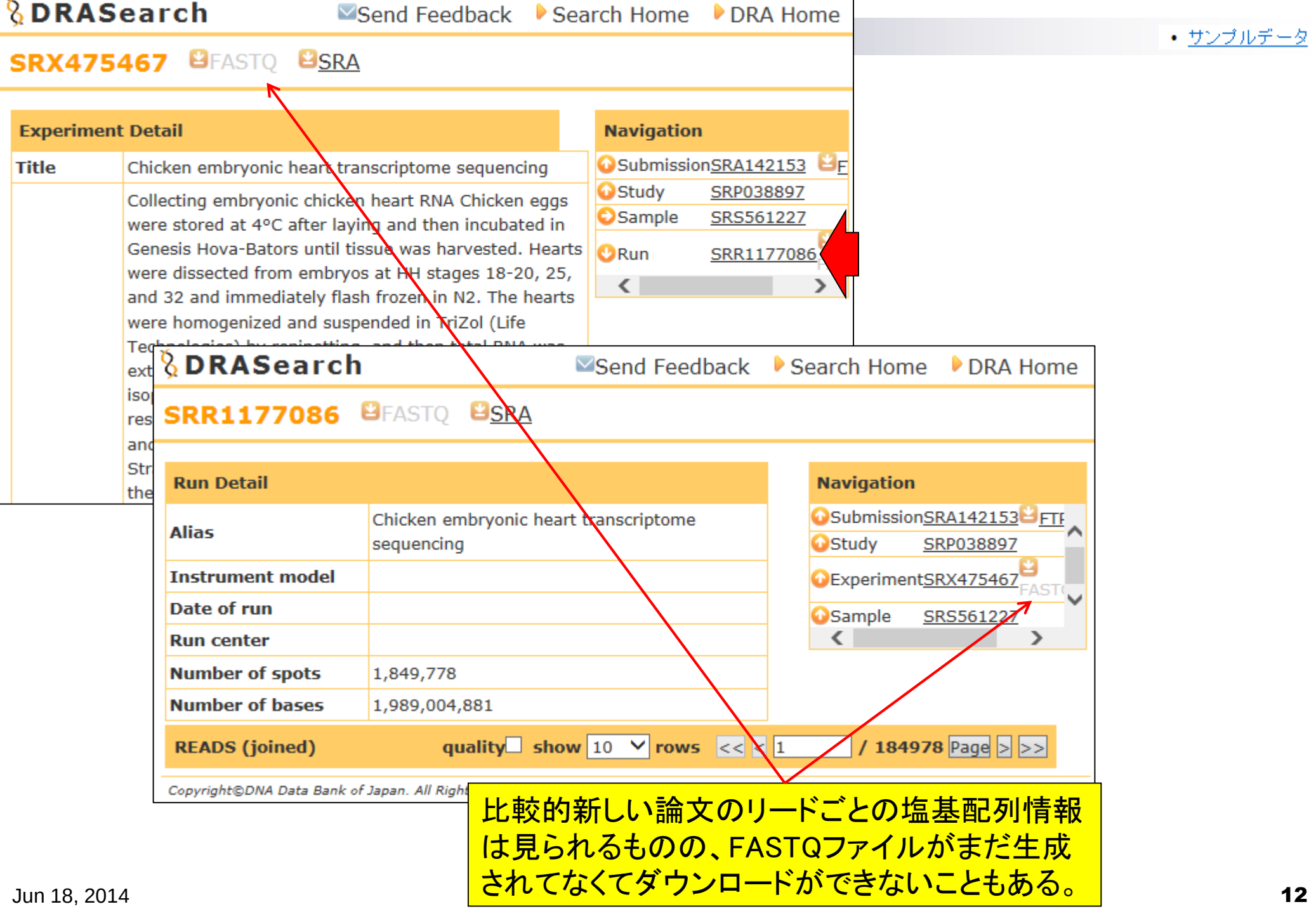
Navigation
Submission SRA142153 F
SRA142158
Experiment SRX475467
Sample SRS561227

DRASearch Send Feedback Search Home DRA Home

SRX475467 FASTQ SRA

Experiment Detail	
Title	Chicken embryonic heart transcriptome sequencing
Description	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 the aqueous phase, precipitated with isopropanol, washed with 75% ethanol and then resuspended into nuclease free water. RNA purification and cDNA synthesis mRNA were purified using the Strategene Absolutely mRNA purification kit: Briefly, the RNA were hybridized to oligo-dT magnetic beads,
Center Name	

Navigation
Submission SRA142153 F
Study SRP038897
Sample SRS561227
Run SRR1177086


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
Study SRP038897
Sample SRS561227
Run SRR1177086
サンプルデータ
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
Study SRP038897
Experiment SRX475467
Sample SRS561227
READS (joined)
quality
show 10 rows
<< < 1 / 184978 Page > >>
Copyright©DNA Data Bank of Japan. All Right
比較的新しい論文のリードごとの塩基配列情報は見られるものの、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。

29. Illumina HiSeq 2000/36bp/single-end/Arabidopsis (SRP011435; GSE36469) data (Huang et al., Developm
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 (xxx) 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 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

DDBJ SRA (DRA)がだめな場合は、NCBI SRAにトライ

NCBI Resources How To

SRA SRA SRP038897

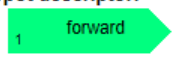
Save search Advanced

Display Settings: Full Send to:

SRX475467: **Chicken embryonic heart transcriptome sequencing**
1 PACBIO_SMRT (PacBio RS) run: 1.8M spots, 2G bases, 486.4Mb downloads

Accession: SRX475467
Experiment design: 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 the aqueous phase, precipitated with isopropanol, washed with 75% ethanol and then resuspended into nuclease free water. RNA purification and cDNA synthesis mRNA were purified using the Strategene Absolutely mRNA purification kit: Briefly, the RNA were hybridized to oligo-dT magnetic beads, separated from solution on a magnetic stand, washed, and then resuspended into the kit's elution buffer. First strand cDNA synthesis was performed using the SMART cDNA kit (Clontech): The first cDNA strand was synthesized from purified poly-A RNA using the SMARTScribe MMLV Reverse Transcriptase (Clontech), the CDS III oligo-dT primer and the SMART IV primer for template switching in order to add a consistent 5' site for LD-PCR amplification using the CDS III primer and the 5' PCR Primer. CDSIII primer: 5'-ATTCTAGAGGCCGAGGCGGCCGACATG-d(T)30N–1N-3' (N = A, G, C, or T; N–1 = A, G, or C) SMART IV oligonucleotide: 5'-AAGCAGTGGTATCAACGCAGAGTGGCCATTACGGCCGGG-3' 5' PCR Primer: 5'-AAGCAGTGGTATCAACGCAGAGT-3' Library preparation and sequencing The cDNA was run on an agarose gel and four separate size ranges were fractionated: 0-1 kb, 1-2 kb, 2-3 kb, and over 3kb. Each size fraction was extracted from the gel and SMRTbell libraries were created using the DNA Template Library Preparation kit (Pacific Biosciences): The cDNA was cleaned using Ampure beads and the ends were repaired. Blunt hairpin 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:



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

#		# of Spots	# of Bases	Size
1.	SRR1177086	1,849,778	2G	486.4Mb

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 [illegible]M spots, 2G bases, 486.4Mb

#		# of Spots	# of Bases	Size
1.	SRR1177086	1,849,778	2G	486.4Mb

• サンプルデータ

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

Chicken embryonic heart transcriptome sequencing (SRR1177086)

Change accession...

Metadata | Reads | Download

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

Quality graph (bigger)

リードごとの塩基配列情報を見る場合はここ

This run has 1 read per spot:

L̄=1075, σ=685.6, 100%

Legend

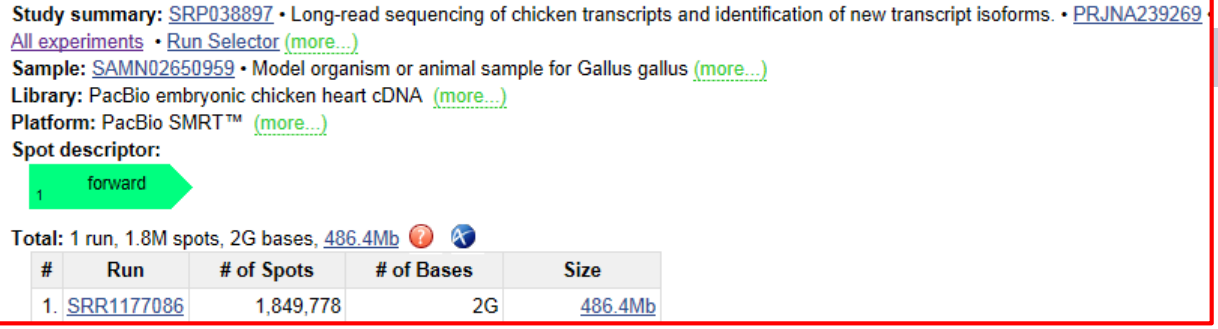
Experiment	Library					
	Name	Platform	Strategy	Source	Selection	Layout
SRX475467	PacBio embryonic chicken heart cDNA	PacBio SMRT™	OTHER	TRANSCRIPTOMIC	cDNA	SINGLE

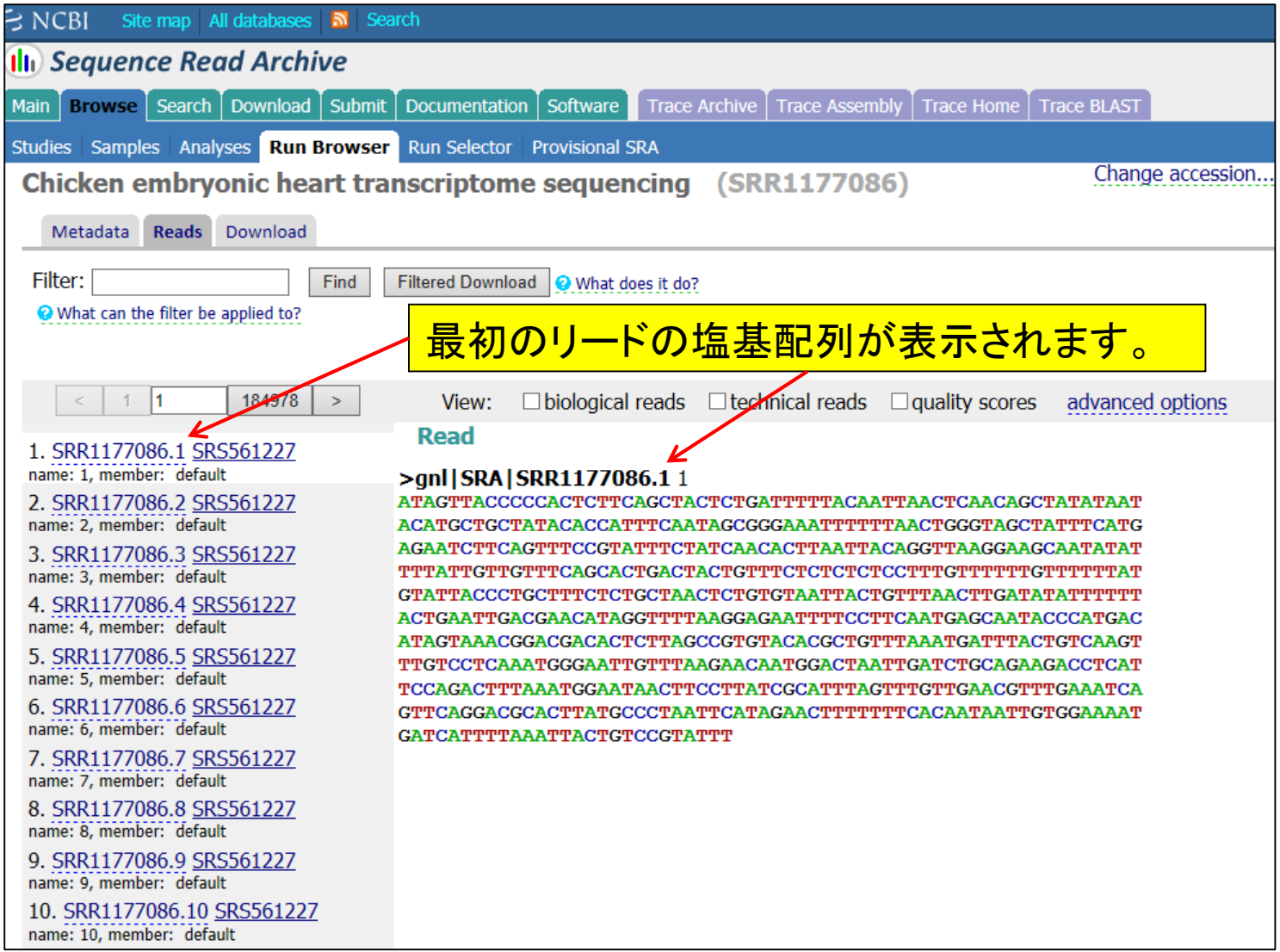
Show design

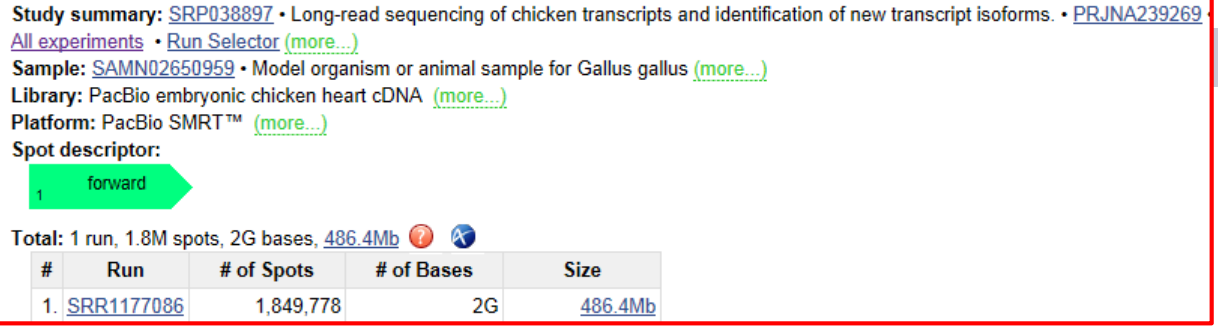
Biosample	Sample Description	Organism	Links
SAMN02650959 (SRS561227)		Gallus gallus	• Model organism or animal sample for Gallus gallus • PRJNA239269 [Gallus gallus]

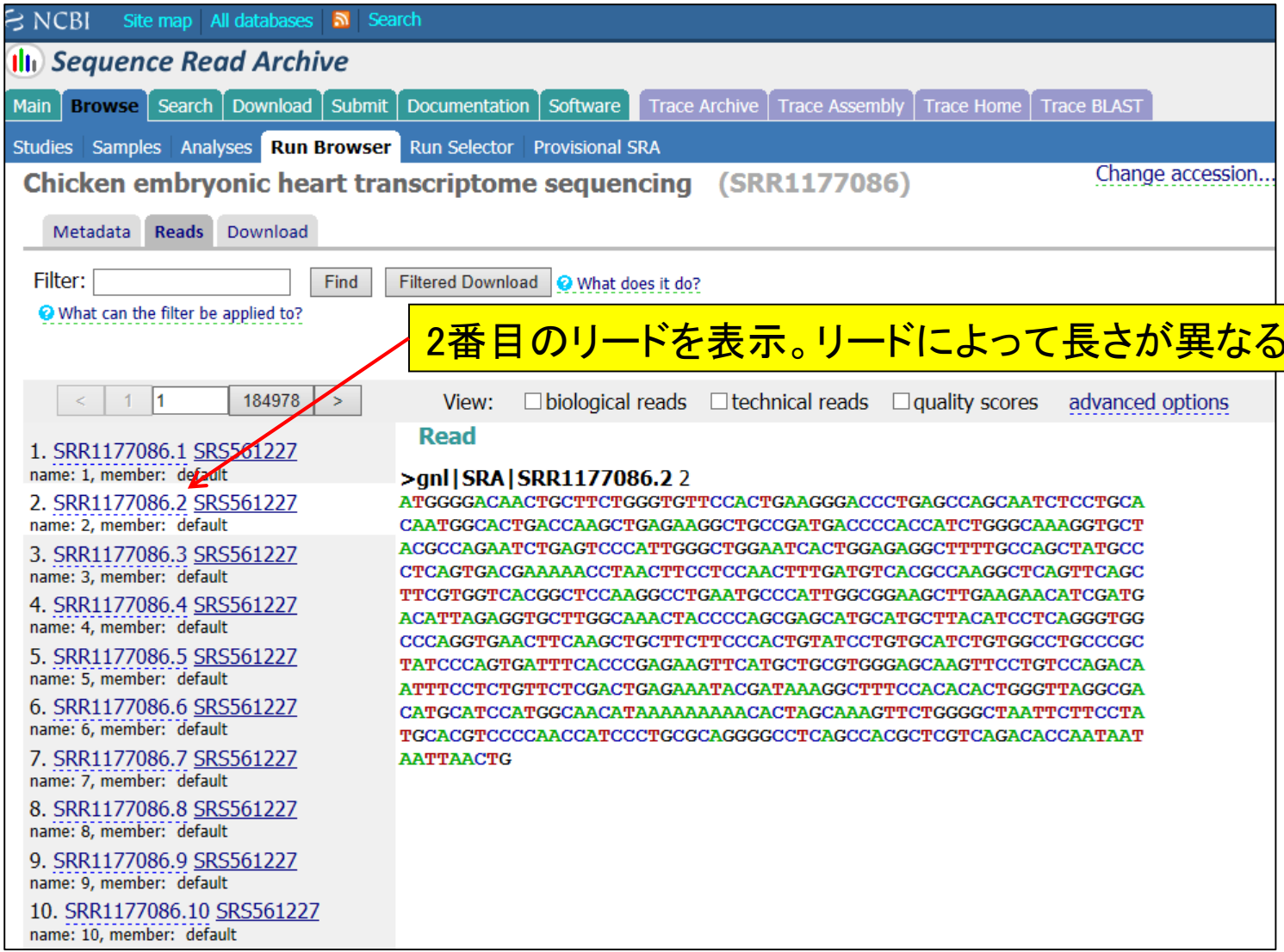
Bioproject	SRA Study	Title
PRJNA239269	SRP038897	Long-read sequencing of chicken transcripts and identification of new transcript isoforms.

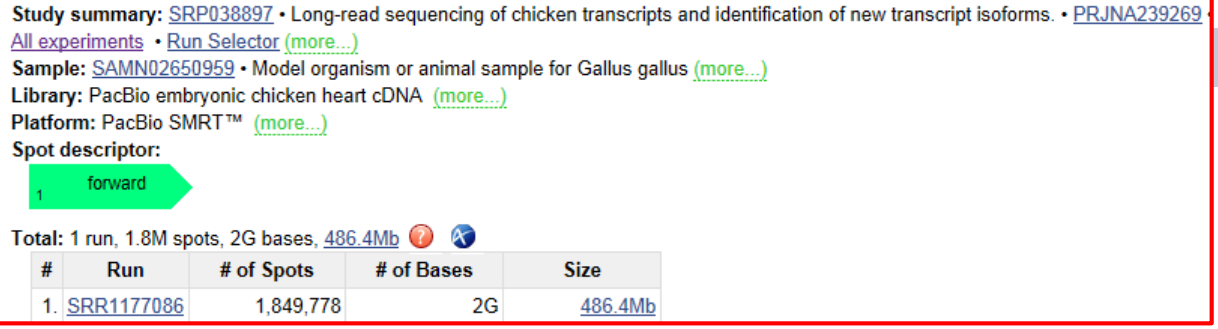
Show abstract

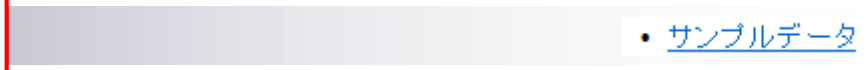

Study summary: SRP038897 • Long-read sequencing of chicken transcripts and identification of new transcript isoforms. • PRJNA239269
All experiments • Run Selector (more...)
Sample: SAMN02650959 • Model organism or animal sample for Gallus gallus (more...)
Library: PacBio embryonic chicken heart cDNA (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
1. SRR1177086 1,849,778 2G 486.4Mb

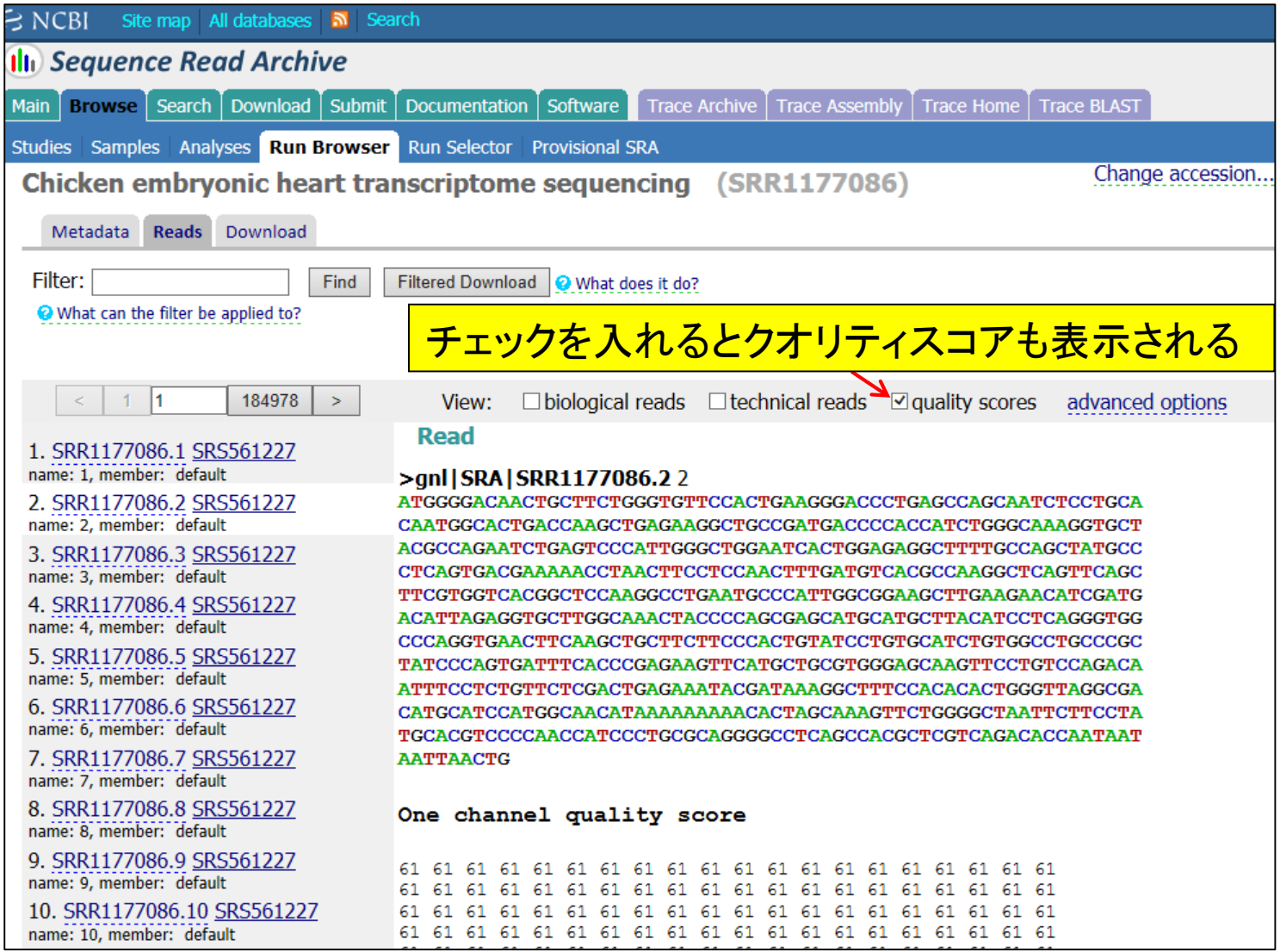

NCBI Site map All databases Search
Sequence Read Archive
Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST
Studies Samples Analyses Run Browser Run Selector Provisional SRA
Chicken embryonic heart transcriptome sequencing (SRR1177086)
Change accession...
Metadata Reads Download
Filter: Find Filtered Download What does it do?
What can the filter be applied to?
最初のリードの塩基配列が表示されます。
< 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
TTTATTGTTGTTTCAGCACTGACTACTGTTTCTCTCTCTCCTTTGTTTTTTGTTTTTTAT
GTATTACCCTGCTTTCTCTGCTAACTCTGTGTAATTACTGTTTAACTTGATATATTTTTT
ACTGAATTGACGAACATAGGTTTTAAGGAGAATTTTCCTTCAATGAGCAATACCCATGAC
ATAGTAAACGGACGACACTCTTAGCCGTGTACACGCTGTTTAAATGATTTACTGTCAAGT
TTGTCCTCAAATGGGAATTGTTTAAGAACAATGGACTAATTGATCTGCAGAAGACCTCAT
TCCAGACTTTAAATGGAATAACTTCCTTATCGCATTTAGTTTGTTGAACGTTTGAAATCA
GTTCAGGACGCACTTATGCCCTAATTCATAGAACTTTTTTTCACAATAATTGTGGAAAAT
GATCATTTTAAATTACTGTCCGTATTT


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
1. SRR1177086 1,849,778 2G 486.4Mb


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
Chicken embryonic heart transcriptome sequencing (SRR1177086)
Change accession...
Metadata Reads Download
Filter: Find Filtered Download What does it do?
What can the filter be applied to?
2番目のリードを表示。リードによって長さが異なる
< 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|SRR1177086.2 2
ATGGGGACAACTGCTTCTGGGTGTTCCACTGAAGGGACCCTGAGCCAGCAATCTCCTGCA
CAATGGCACTGACCAAGCTGAGAAGGCTGCCGATGACCCCACCATCTGGGCAAAGGTGCT
ACGCCAGAATCTGAGTCCCATTGGGCTGGAATCACTGGAGAGGCTTTTGCCAGCTATGCC
CTCAGTGACGAAAAACCTAACTTCCTCCAACTTTGATGTCACGCCAAGGCTCAGTTCAGC
TTCGTGGTCACGGCTCCAAGGCCTGAATGCCCATTGGCGGAAGCTTGAAGAACATCGATG
ACATTAGAGGTGCTTGGCAAACTACCCCAGCGAGCATGCATGCTTACATCCTCAGGGTGG
CCCAGGTGAACTTCAAGCTGCTTCTTCCCACTGTATCCTGTGCATCTGTGGCCTGCCCGC
TATCCCAGTGATTTCACCCGAGAAGTTCATGCTGCGTGGGAGCAAGTTCCTGTCCAGACA
ATTTCCTCTGTTCTCGACTGAGAAATACGATAAAGGCTTTCCACACACTGGGTTAGGCGA
CATGCATCCATGGCAACATAAAAAAAAACACTAGCAAAGTTCTGGGGCTAATTCTTCCTA
TGCACGTCCCCAACCATCCCTGCGCAGGGGCCTCAGCCACGCTCGTCAGACACCAATAAT
AATTAACTG


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
1. SRR1177086 1,849,778 2G 486.4Mb

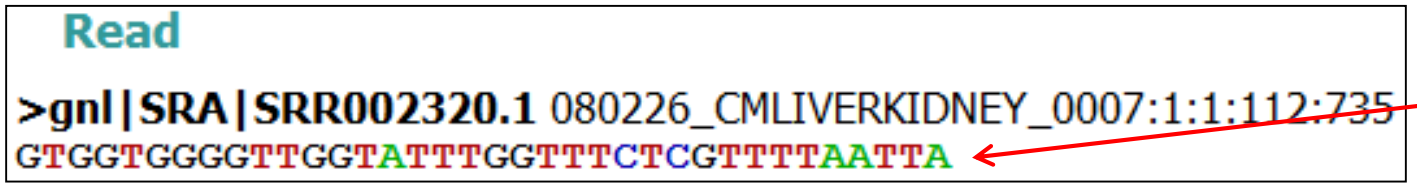

• サンプルデータ


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
Chicken embryonic heart transcriptome sequencing (SRR1177086)
Change accession...
Metadata Reads Download
Filter: Find Filtered Download What does it do?
What can the filter be applied to?
チェックを入れるとクオリティスコアも表示される
< 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|SRR1177086.2 2
ATGGGGACAACTGCTTCTGGGTGTTCCACTGAAGGGACCCTGAGCCAGCAATCTCCTGCA
CAATGGCACTGACCAAGCTGAGAAGGCTGCCGATGACCCCACCATCTGGGCAAAGGTGCT
ACGCCAGAATCTGAGTCCCATTGGGCTGGAATCACTGGAGAGGCTTTTGCCAGCTATGCC
CTCAGTGACGAAAAACCTAACTTCCTCCAACTTTGATGTCACGCCAAGGCTCAGTTCAGC
TTCGTGGTCACGGCTCCAAGGCCTGAATGCCCATTGGCGGAAGCTTGAAGAACATCGATG
ACATTAGAGGTGCTTGGCAAACTACCCCAGCGAGCATGCATGCTTACATCCTCAGGGTGG
CCCAGGTGAACTTCAAGCTGCTTCTTCCCACTGTATCCTGTGCATCTGTGGCCTGCCCGC
TATCCCAGTGATTTCACCCGAGAAGTTCATGCTGCGTGGGAGCAAGTTCCTGTCCAGACA
ATTTCCTCTGTTCTCGACTGAGAAATACGATAAAGGCTTTCCACACACTGGGTTAGGCGA
CATGCATCCATGGCAACATAAAAAAAAACACTAGCAAAGTTCTGGGGCTAATTCTTCCTA
TGCACGTCCCCAACCATCCCTGCGCAGGGGCCTCAGCCACGCTCGTCAGACACCAATAAT
AATTAACTG
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

実際のデータ

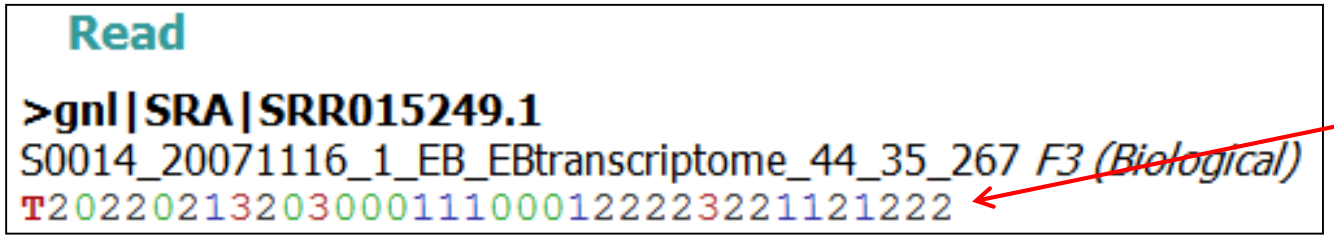
PacBioデータの長さがよく分かります

- Illumina社のGenome Analyzer
  - SRA061145：Marioni et al., *Genome Res.*, **18**: 1509–1517, 2008



サンプルデータ1, 36 bp

- Applied Biosystems社のSOLiD4 System
  - SRA000306：Cloonan et al., *Nat. Methods*, **5**: 613–619, 2008



サンプルデータ5, 25–35 bp

- Illumina社のHiSeq 2000
  - SRA062939：Chan et al., *Hum. Mol. Genet.*, **22**: 2662–2675, 2013

サンプルデータ25, 100 bp

Contents（第2回）

- イントロダクション(Introduction)
  - NGSデータ概観（PacBioとIllumina）
  - **NGSデータベース(DB)、データ形式（FASTQ形式）**
  - SRAdbパッケージを用いたデータ取得、エラーへの対処
- 前処理(Pre-processing)
  - qrqcパッケージを用いたQuality Control (QC)
- アダプター配列除去
  - 基本戦略（girafeパッケージ）
  - 昔は正常に動作していたのに…という例（QuasRパッケージ）
  - アダプター除去を含む様々なフィルタリングの組合せ（ShortReadパッケージ）
  - 課題

NGS用データベース

- 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 2013/09/25)
- イントロ | NGS | 様々なプラットフォーム(last modified 2014/06/08) NEW
- イントロ | NGS | qPCRやmicroarrayなどとの比較(last modified 2014/06/05) NEW
- イントロ | NGS | Viewer(last modified 2014/01/29)
- イントロ | NGS | 配列取得 | FASTQ or SRALite | 公共DBから (last modified 2014/03/27)
- イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu 2013)(last modified 2014/06/08) NEW
- イントロ | NGS | アノテーション情報取得 | について(last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取得 | GFF/GTF形式ファイル(last modified 2014/04/11)

SRA, DRA, ENAが公式?!なNGS data repository

イントロ | NGS | 配列取得 | FASTQ or SRALite | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用すると便利です。

データの形式は基本的にSanger typeのFASTQ形式です。

FASTA形式はリードあたり二行(idの行と配列の行)で表現します。

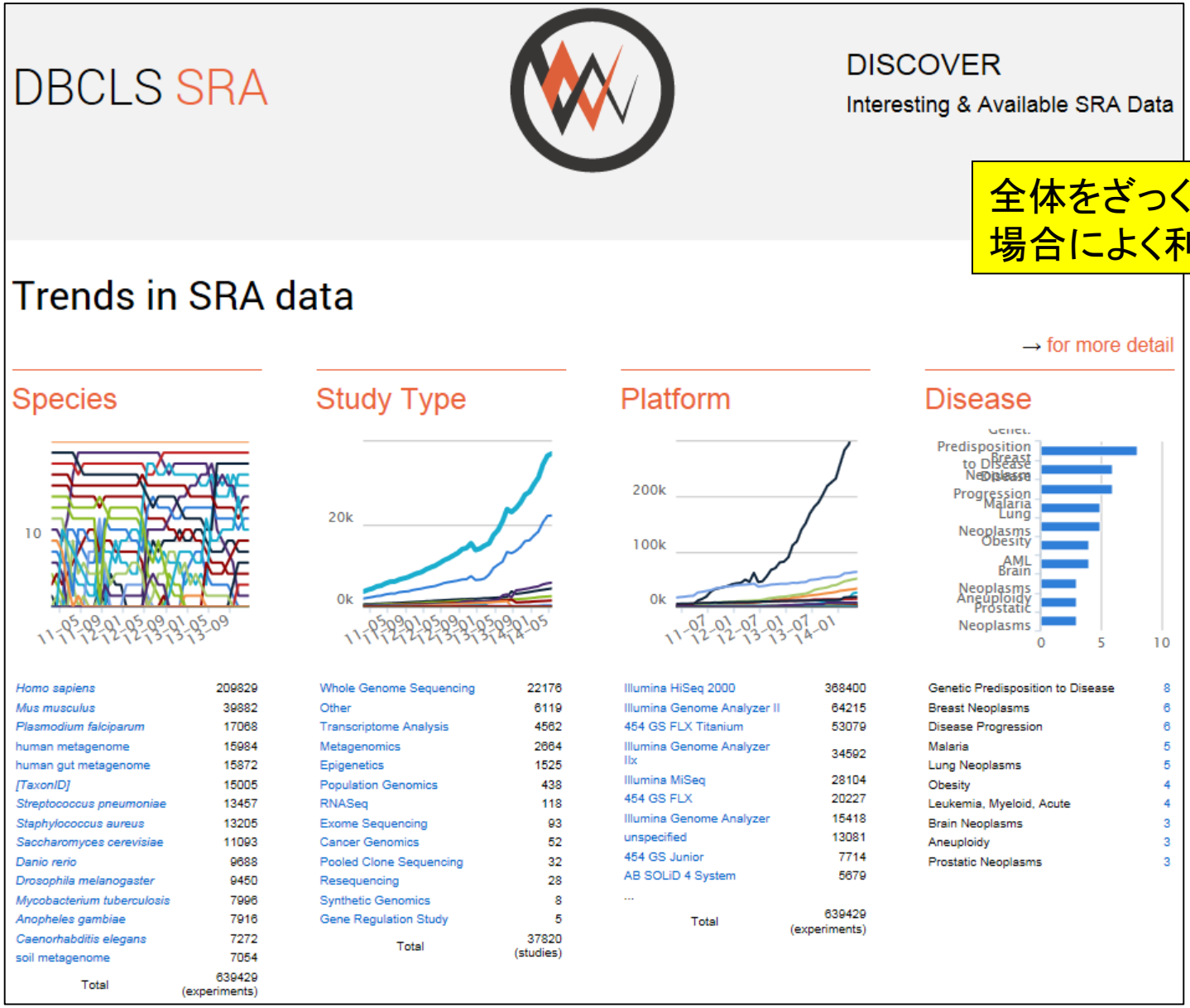
FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。

FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

- NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
- DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
- European Nucleotide Archive (ENA): Leinonen et al., Nucleic Acids Res., 2011
- DBCLS SRA: Nakazato et al., PLoS One, 2013

2次DBのDBCLS SRAも面白いサイトです

DBCLS SRA



全体をざっくりと知りたい
場合によく利用しています

NGSデータ取得は全体像の理解が大事

- Chan et al., *Hum. Mol. Genet.*, **22**: 2662–2675, 2013
  - 論文中に、NGSデータはGene Expression Omnibus (GEO)中にGSE42960で登録したという記載あり

NCBI SRA内で GSE42960で検索

イントロ | NGS | 配列取得 | FASTQ or SRALite | 公共DBから NEW

次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用すると便利です。
データの形式は基本的にSanger typeのFASTQ形式です。
FASTA形式はリードあたり二行(idの行と配
FASTQ形式はリードあたり4行(@から始まる
行)で表現します。
FASTQ形式は、Sangerのものがデファクト
るのはFASTQ-like formatという表現がなさ
2013年頃には、IlluminaデータもBaseSpace
FASTQ形式(つまりSanger typeのデータ)に
は全てはSanger typeのデータになっていた

- NCBI Sequence Read Archive (SRA):
- DDBJ Sequence Read Archive (DRA):
- European Nucleotide Archive (ENA):
- DBCLS SRA: Nakazato et al., PLoS On

NCBI Resources How To Sign in to NCBI

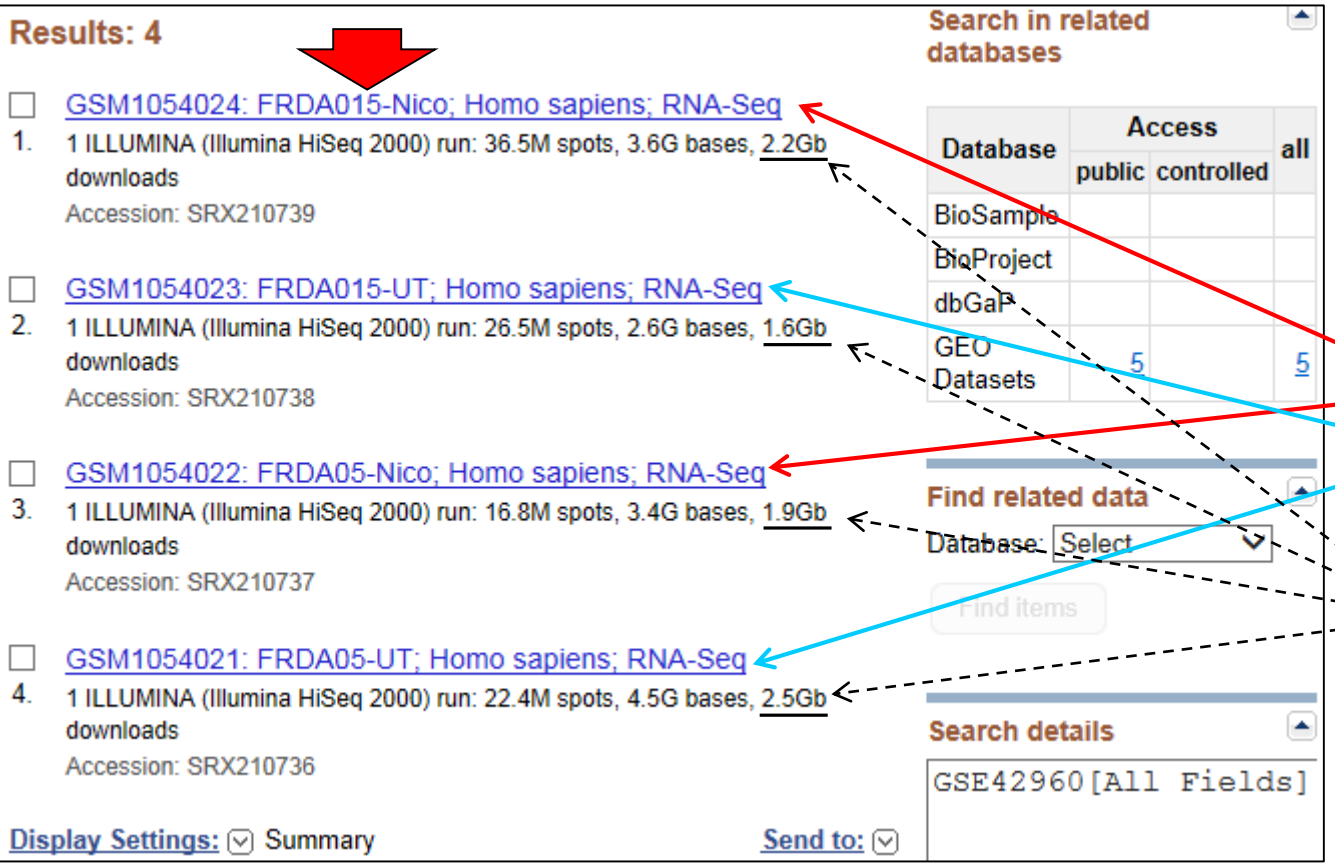
SRA | SRA | GSE42960 | Search | Advanced | Help

SRA

The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Using SRA	Tools	Other Resources
Handbook	BLAST	SRA Home
Download	SRA Run browser	Trace Archive
E-Utilities	Submit to SRA	Trace Assembly
Factsheet	SRA software	GenBank Home

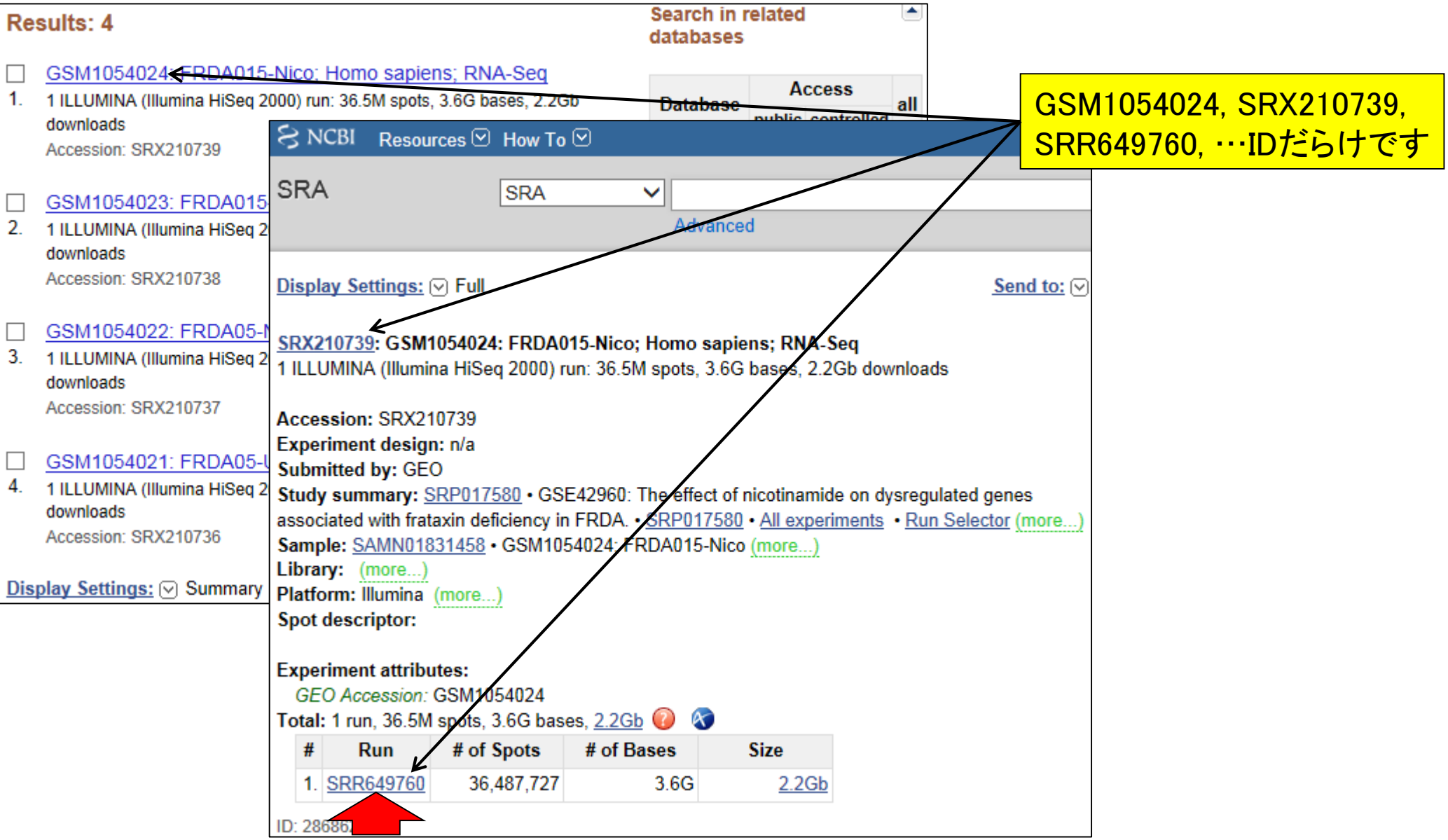
NGSデータ取得は全体像の理解が大事



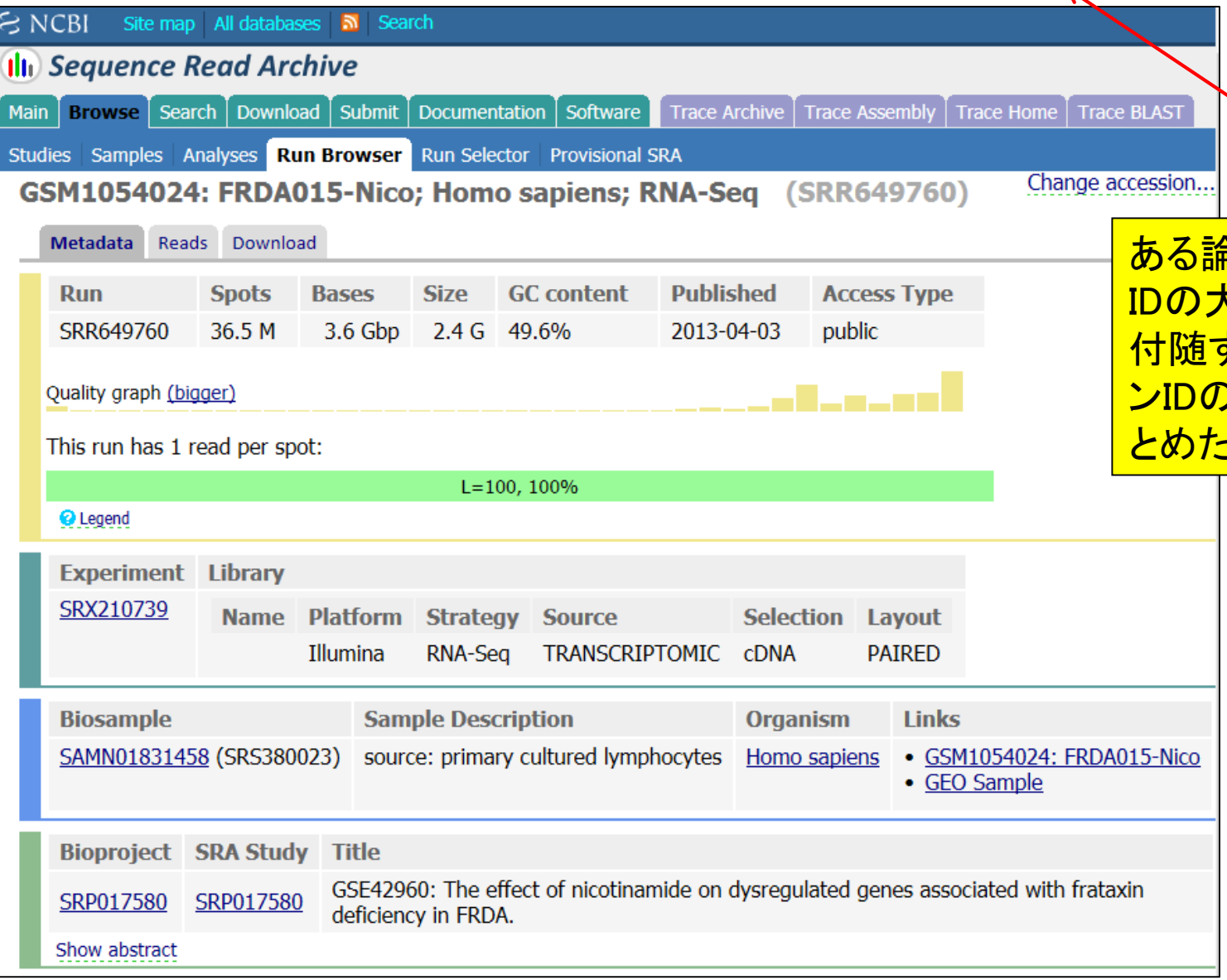
原著論文を読むことでIllumina HiSeq 2000を使っていること、2群間比較用データであることは既知。ニコチンアミド処理群（Nico） 対 未処理群（Untreated; UT）。

生データをダウンロードすると、9GB程度になる。

NGSデータ取得は全体像の理解が大事



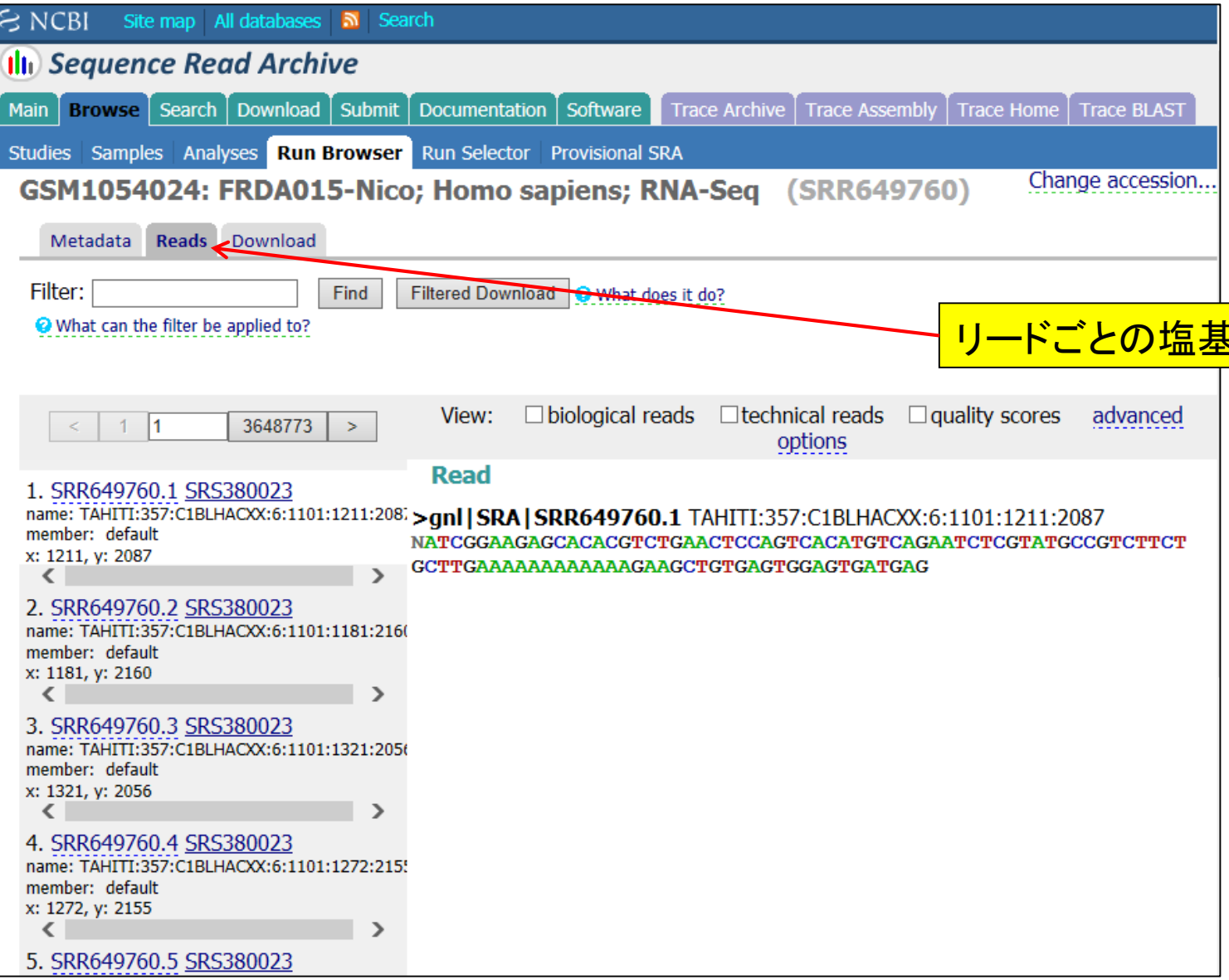
NGSデータ取得は全体像の理解が大事



ある論文のデータ全体を指し示すIDの大元がGSE42960で、それに付随する実験IDのSRX210739、ランIDのSRR649760などの全体をまとめたデータをメタデータという。

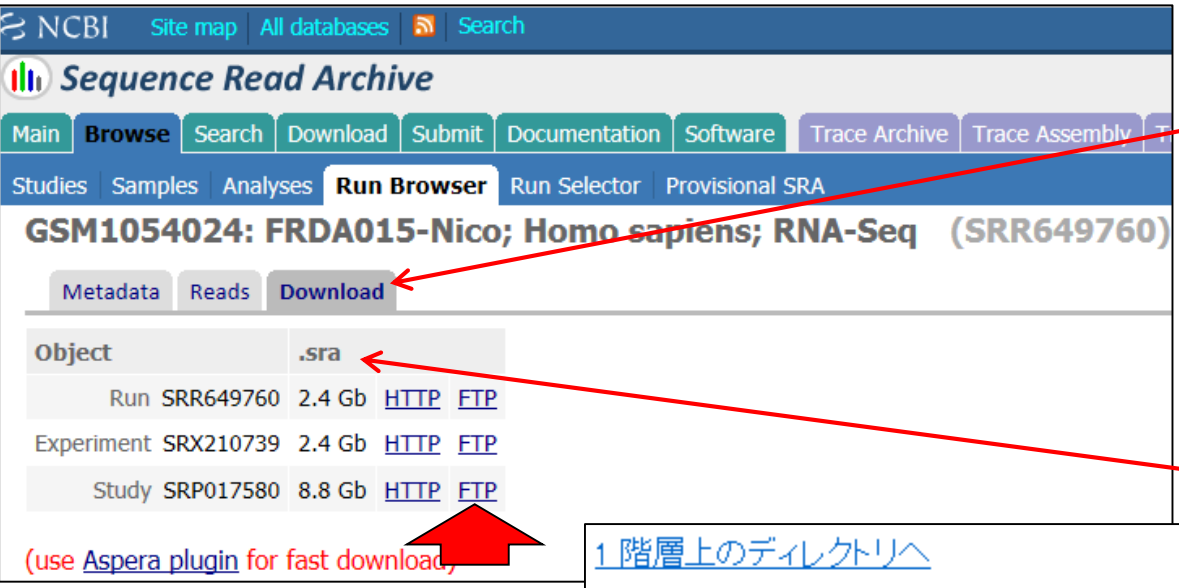
Illuminaの場合はフローセルと呼ばれるスライドグラス程度の大きさのものを用いて、一度に8サンプル分のsequencingが可能。この実験をラン(RUN)という。

NGSデータ取得は全体像の理解が大事



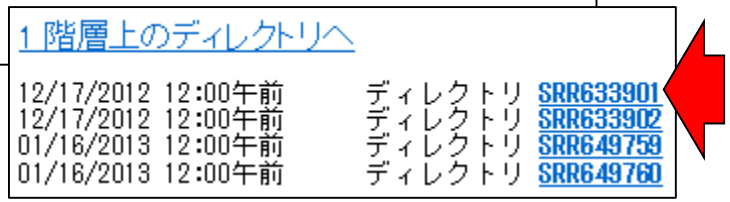
リードごとの塩基配列情報を見る場合はここ

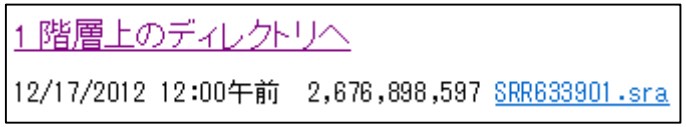
NGSデータ取得は全体像の理解が大事



ダウンロードはここだが…

FASTQ形式ではなくSRA形式ファイルなので非推奨。





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

FASTQ形式ファイルの利用が基本

FASTA形式とFASTQ形式

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

```
>SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

- FASTQ形式
  - 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

公共DBからデータを取得する場合

- ENA Sequence Read Archive (ERA; 欧)
  - FASTQ形式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
- DDBJ Sequence Read Archive (DRA; 日)
  - FASTQ形式とSRA-Lite形式でダウンロード可能

ENAを概観しながら…
・サンプル数と得られるファイル数の違いなどを認識
・論文中の情報を頼りにSRAまたはSRP IDを入手する手続き
・メタデータ(全体像)情報を把握

イントロ | NGS | 配列取得 | FASTQ or SRALite | 公共DBから NEW

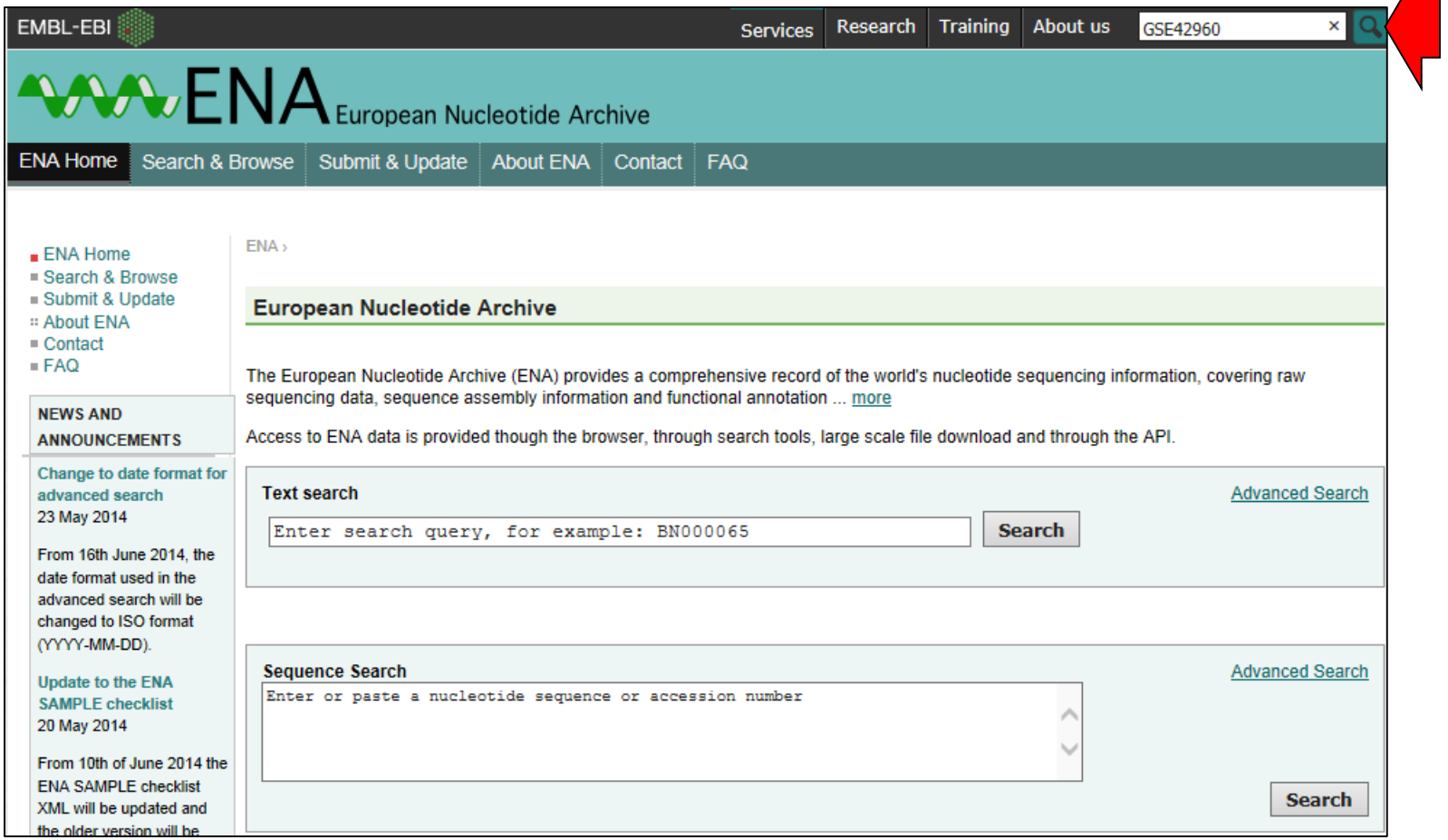
次世代シーケンサ(NGS)から得られる塩基配列データを公共データベースから取得する際には以下を利用すると便利です。
データの形式は基本的にSanger typeのFASTQ形式です。
FASTA形式はリードあたり二行(idの行と配列の行)で表現します。
FASTQ形式はリードあたり4行(@から始まるidの行と配列の行、および+から始まるidの行とbase callの際のqualityの行)で表現します。
FASTQ形式は、Sangerのものがデファクトスタンダード(業界標準)です。かつてIlluminaのプラットフォームから得られるのはFASTQ-like formatという表現がなされていたようです(Cock et al., Nucleic Acids Res., 2010)。しかし少なくとも2013年頃には、IlluminaデータもBaseSpaceやCASAVA1.8のconfigureBclToFastq.plなどを用いることで業界標準のFASTQ形式(つまりSanger typeのデータ)に切り替えられるようですし、NCBI SRAなどの公共DBから取得するデータは全てはSanger typeのデータになっていたと思います(Kibukawa E., テクニカルサポートウェビナー, 2013)。

- NCBI Sequence Read Archive (SRA): NCBI Resource Coordinators., Nucleic Acids Res., 2014
- DDBJ Sequence Read Archive (DRA): Kodama et al., Nucleic Acids Res., 2012
- European Nucleotide Archive (ENA): Leinonen et al., Nucleic Acids Res., 2011
- DBCLS SRA: Nakazato et al., PLoS One, 2013

NGSデータ取得は全体像の理解が大事

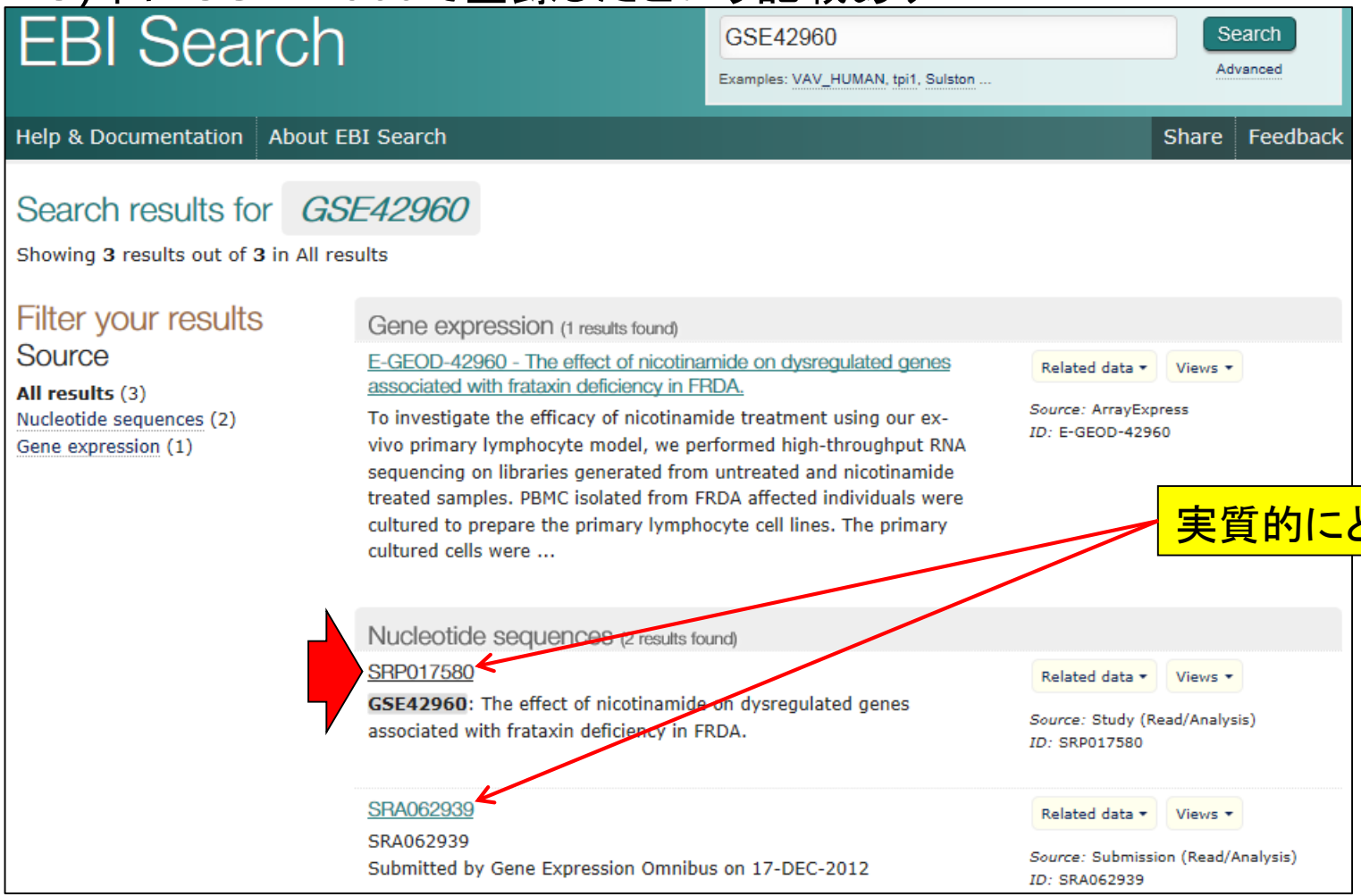
- Chan et al., *Hum. Mol. Genet.*, **22**: 2662–2675, 2013
  - 論文中に、NGSデータはGene Expression Omnibus (GEO)中にGSE42960で登録したという記載あり

ENAでGSE42960で検索



NGSデータ取得は全体像の理解が大事

- Chan et al., *Hum. Mol. Genet.*, **22**: 2662–2675, 2013
  - 論文中に、NGSデータはGene Expression Omnibus (GEO)中にGSE42960で登録したという記載あり



NGSデータ取得は全体像の理解が大事

- Chan et al., *Hum. Mol. Genet.*, **22**: 2662–2675, 2013

計6個のFASTQ形式ファイルになるようだ

ENAは全体像をつかみやすい

Navigation | Read Files | Attributes

▶ Download files

View: TEXT

Download: TEXT

Select columns

Showing results 1 - 4 of 4 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	CoL tax ID	CoL scientific name
SRP017580	SRP017580	SAMN01831455	SRS380020	SRX210736	SRR633901	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			6850099	Homo sapiens Linnaeus, 1758
SRP017580	SRP017580	SAMN01831456	SRS380021	SRX210737	SRR633902	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			6850099	Homo sapiens Linnaeus, 1758
SRP017580	SRP017580	SAMN01831457	SRS380022	SRX210738	SRR649759	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1			6850099	Homo sapiens Linnaeus, 1758
SRP017580	SRP017580	SAMN01831458	SRS380023	SRX210739	SRR649760	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1			6850099	Homo sapiens Linnaeus, 1758

For Aspera download, please download and install Aspera Connect

NGSデータ取得は全体像の理解が大事

- Chan et al., *Hum. Mol. Genet.*, **22**: 2662–2675, 2013

Navigation | Read Files | Attributes

This table contains the files for submission SRA062939

▶ Download files

View: TEXT

Select columns

Showing results 1 - 4 of 4 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	CoL tax ID	CoL scientific name
SRP017580	SRP017580	SAMN01831455	SRS380020	SRX210736	SRR633901	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			6850099	Homo sapiens Linnaeus, 1758
SRP017580	SRP017580	SAMN01831456	SRS380021	SRX210737	SRR633902	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			6850099	Homo sapiens Linnaeus, 1758
SRP017580	SRP017580	SAMN01831457	SRS380022	SRX210738	SRR649759	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1			6850099	Homo sapiens Linnaeus, 1758
SRP017580	SRP017580	SAMN01831458	SRS380023	SRX210739	SRR649760	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1	File 1			6850099	Homo sapiens Linnaeus, 1758

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR633/SRR633901/SRR633901_1.fastq.gz

SRA, DRA, ENAどれでもよいが、論文から得られるGSE IDを頼りに、「SRA IDまたはSRP ID」情報を入手（**SRA062939**または**SRP017580**）するとともに、メタデータ情報を把握すべし

ダウンロードして得られるのは、Run IDから始まるSRR…という名前のファイル

Contents（第2回）

- イントロダクション(Introduction)
  - NGSデータ概観（PacBioとIllumina）
  - NGSデータベース(DB)、データ形式（FASTQ形式）
  - **SRAdbパッケージを用いたデータ取得、エラーへの対処**
- 前処理(Pre-processing)
  - qrqcパッケージを用いたQuality Control (QC)
- アダプター配列除去
  - 基本戦略（girafeパッケージ）
  - 昔は正常に動作していたのに…という例（QuasRパッケージ）
  - アダプター除去を含む様々なフィルタリングの組合せ（ShortReadパッケージ）
  - 課題

RでNGSデータ取得

SRAdb (Zhu et al., *BMC Bioinformatics*, **14**: 19, 2013)

*SRAdb*パッケージを利用して、R経由でNCBI SRAからFASTQファイル群をダウンロード可能。**数時間かかるのでやらないで!**

- イントロ | NGS | qPCRやmicroarrayなどとの比較 (last modified 2014/06/05) NEW
- イントロ | NGS | Viewer (last modified 2014/01/29)
- イントロ | NGS | 配列取得 | FASTQ or SRALite | 公共DBから (last modified 2014/03/27)
- イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu_2013) (last modified 2014/06/08) NEW
- イントロ | NGS | アノテーション情報取得 | について (last modified 2014/03/26)
- イントロ | NGS | アノテーション情報取得 | GFF/GTF形式ファイル (last modified 2014/04/11)
- イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu_2013) NEW

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。SRALite形式などでも取得可能なようですが、ここではFASTQ

6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb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x時間程度かかります。

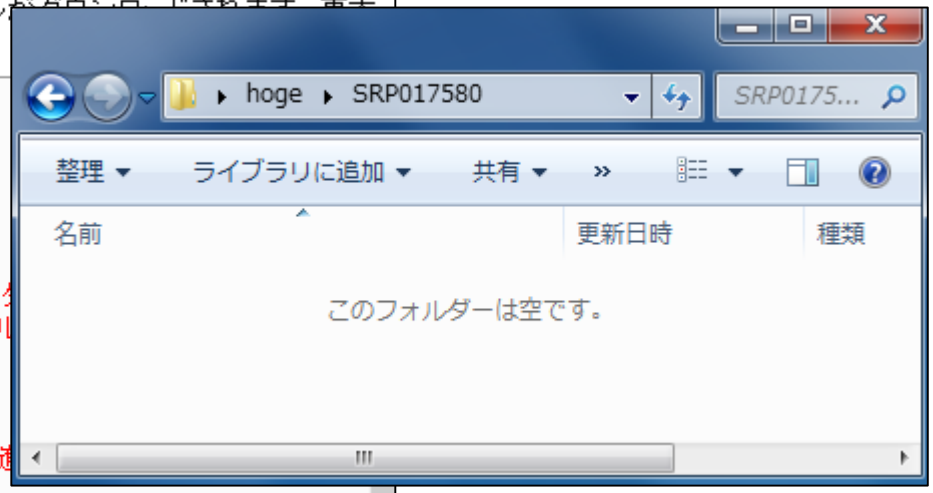
```
param <- "SRP017580"                          #取得したいSRA IDを指定

#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ある"SRAm
sqlfile <- getSRAdbFile()                     #最新のSRAmetadb SQLiteファイルをダウンロードして解凍(
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 (SRP...), sa
hoge                                          #hogeの中身を表示
apply(hoge, 2, unique)                        #hoge行列の列ごとにユニークな文字列を表示させている。

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)                   #「hoge$run」で指定したSRRから始まるIDのFASTQファイルサ
k                                             #kの中身を表示
hoge2 <- cbind(k$library.name,                #ライブラリ名と、
               k$run.read.count,              #総リード数と、
               k$file.name,                   #ファイル名と、
               k$file.size)                   #ファイルサイズ、の順番で列方向で結合した結果をhoge2に格
```

6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

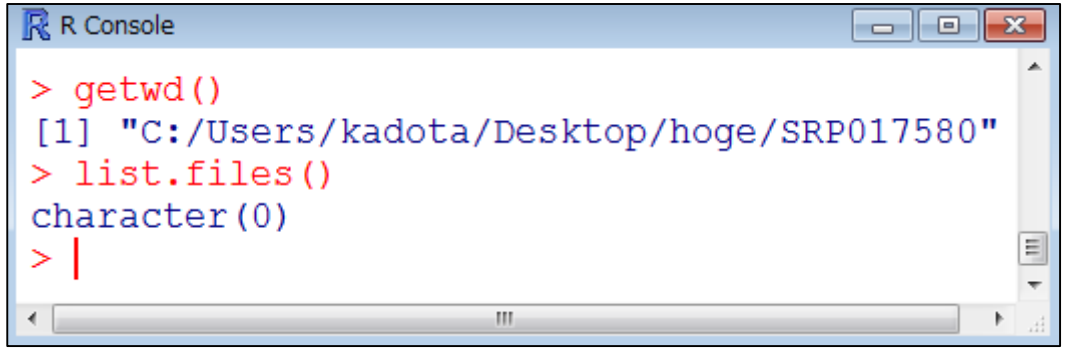
取得 | FASTQ or SRALite | SRAdb(Zhu 2013)

Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb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x時間程度かかります。

```
param <- "SRP017580"                        #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
sqlfile <- getSRAdbFile()                   #最新のSRAmetadb SQLiteファイル
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
hoge                                        #hogeの中身を表示
apply(hoge, 2, unique)                      #hoge行列の列ごとにユニークな文字列を表示させている。

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)                 #「hoge$run」で指定したSRRから始まるIDのFASTQファイルサ
k                                           #kの中身を表示
hoge2 <- cbind(k$library.name,              #ライブラリ名と、
               k$run.read.count,            #総リード数と、
               k$file.name,                 #ファイル名と、
               k$file.size)                 #ファイルサイズ、の順番で列方向で結合した結果をhoge2に格
```



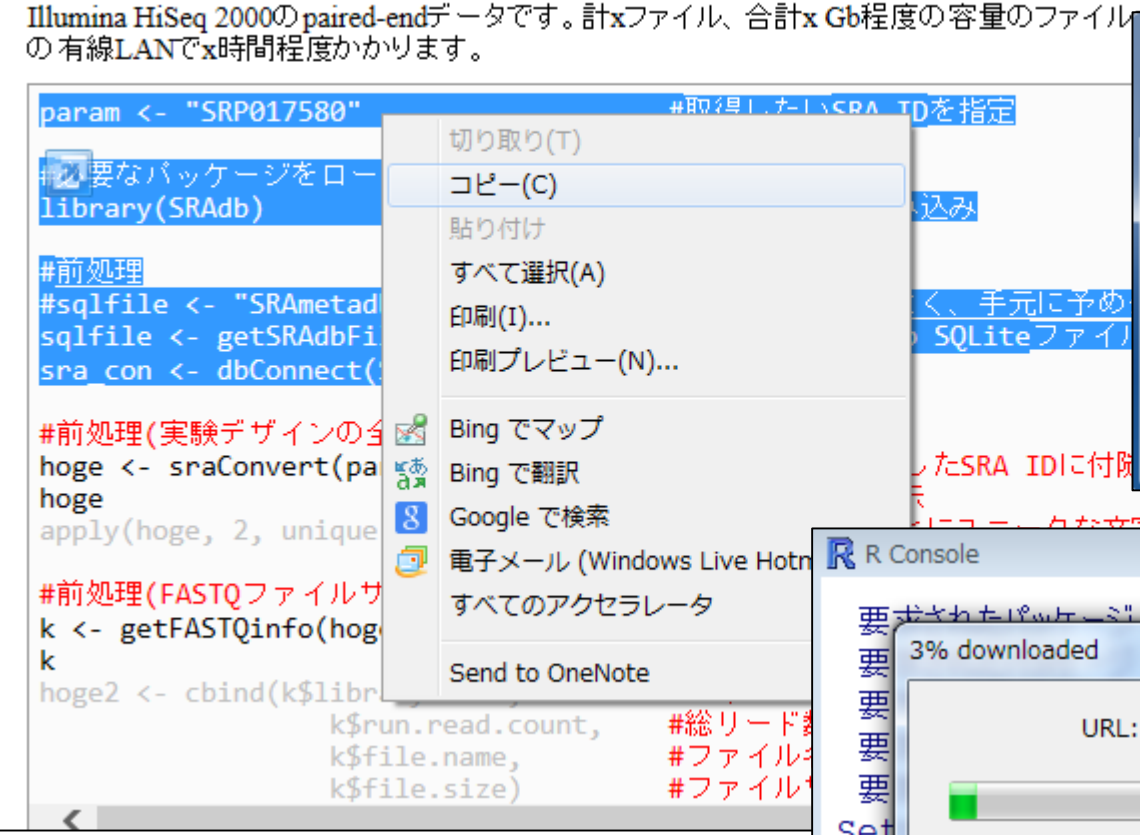
SRAまたはSRP IDを与えることでコピペでFASTQファイルをダウンロード可能です

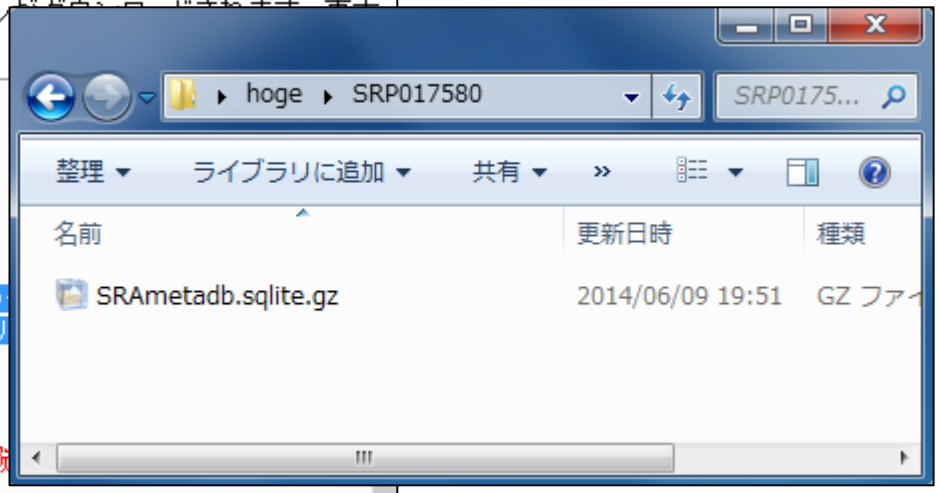


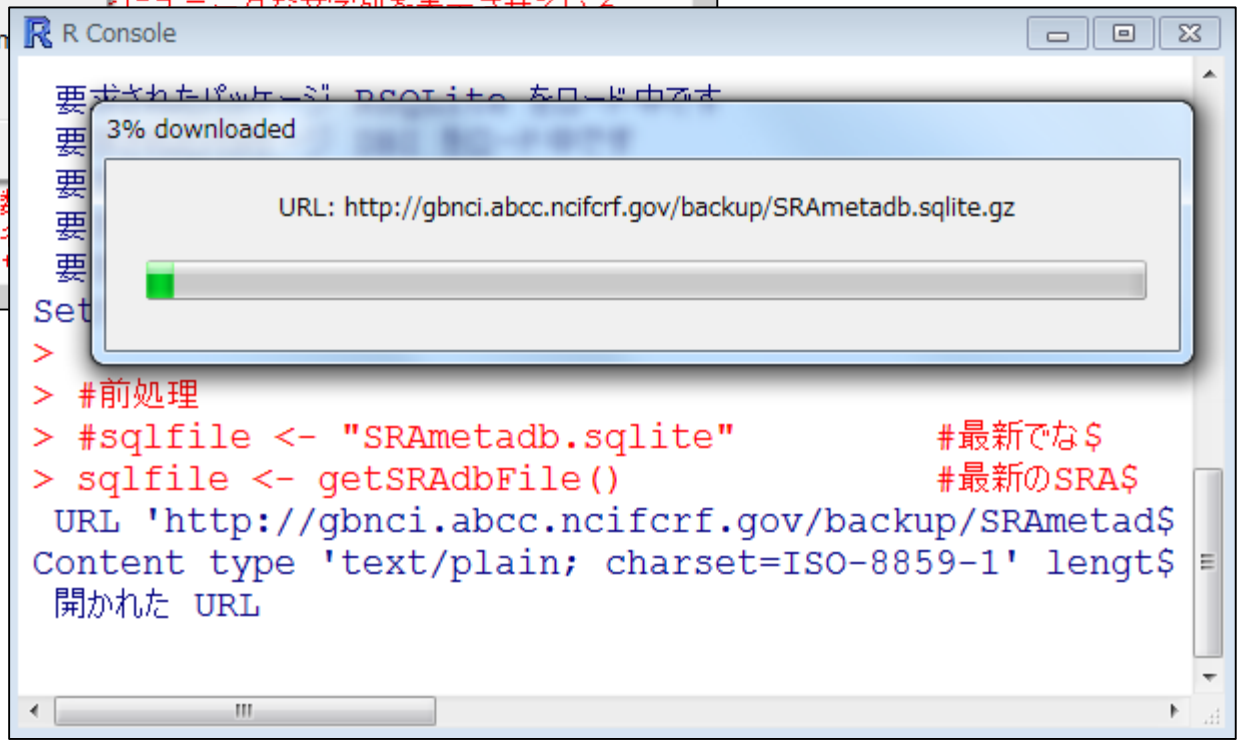
6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb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x時間程度かかります。

```
param <- "SRP017580"                    #取得したいSRA IDを指定

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

#前処理
#sqlfile <- "SRAmetad                   く、手元に予め
sqlfile <- getSRAdbFi                   SQLiteファイル
sra_con <- dbConnect(

#前処理(実験デザインの
hoge <- sraConvert(pa                   したSRA IDに付
hoge
apply(hoge, 2, unique

#前処理(FASTQファイルサ
k <- getFASTQinfo(hog
k
hoge2 <- cbind(k$libr
                k$run.read.count,       #総リード
                k$file.name,            #ファイル
                k$file.size)            #ファイル
```





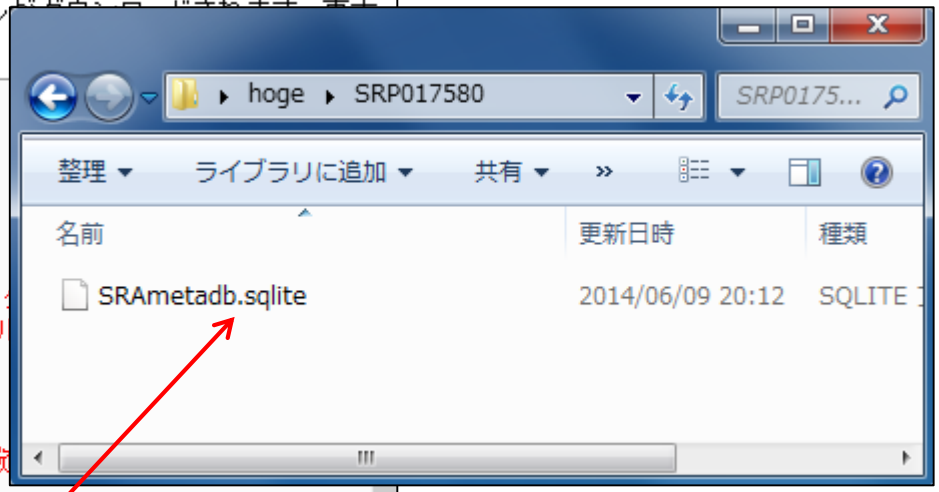


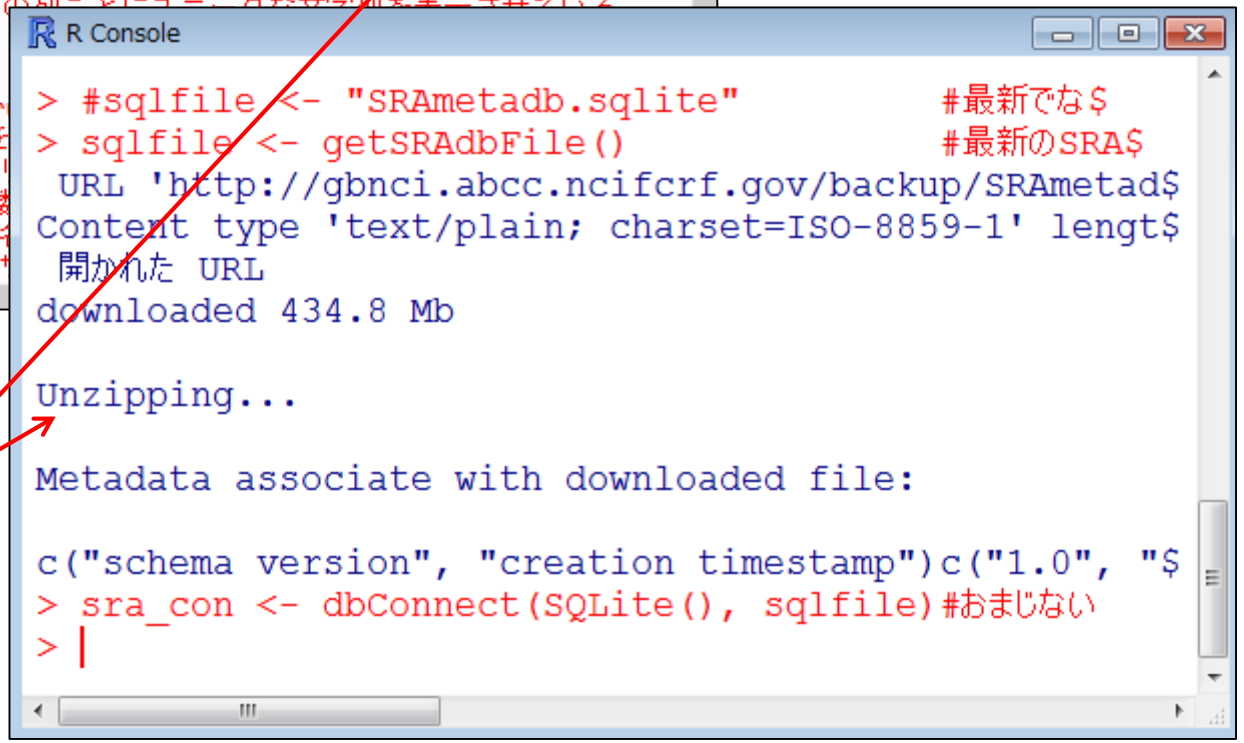
ここまでのコマンドで、NCBI SRA中の全てのメタデータ情報を含んだSRAmetadb.sqliteのgzip圧縮ファイルがダウンロードされる。東大有線LANで20分程度。

6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb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x時間程度かかります。

```
param <- "SRP017580"                          #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
sqlfile <- getSRAdbFile()                     #最新のSRAmetadb SQLiteファイル
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
hoge                                          #hogeの中身を表示
apply(hoge, 2, unique)                        #hoge行列の列ごとにユニークな文字列を表示させている

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)                   # 「hoge$r
k                                             #kの中身を
hoge2 <- cbind(k$library.name,                #ライブラリ
               k$run.read.count,              #総リード数
               k$file.name,                   #ファイル名
               k$file.size)                   #ファイルサ
```



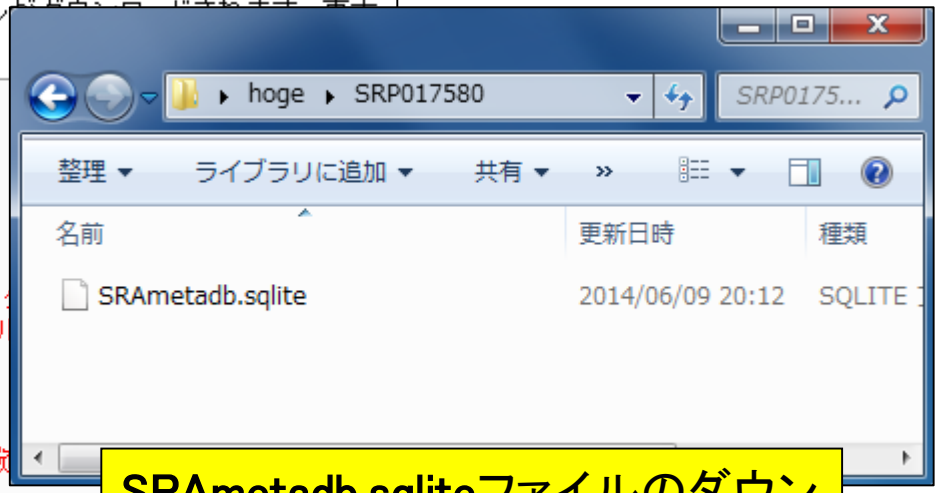


ダウンロード後は自動で解凍。
解凍後のファイルは6GB程度

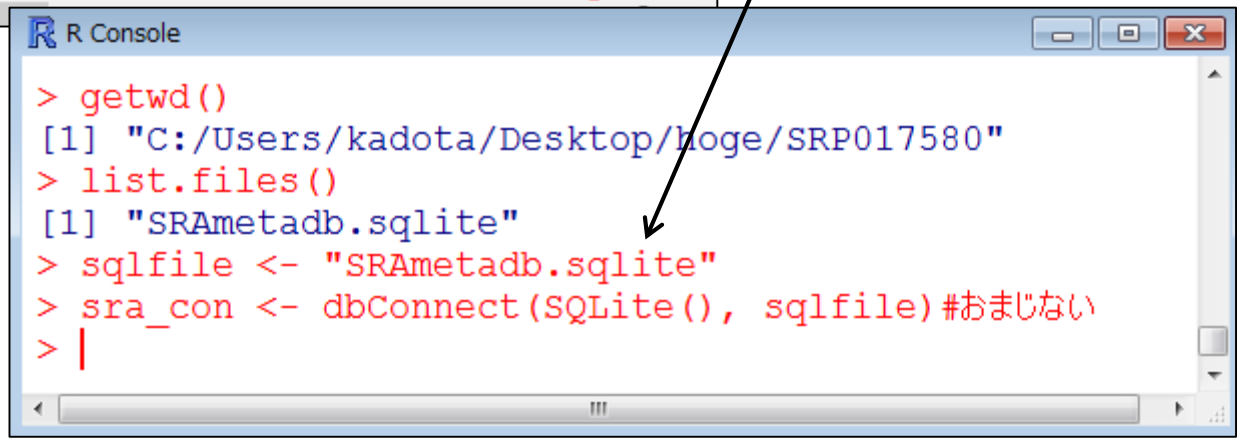
6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb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x時間程度かかります。

```
param <- "SRP017580"                    #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
sqlfile <- getSRAdbFile()               #最新のSRAmetadb SQLiteファイル
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
hoge                                    #hogeの中身を表示
apply(hoge, 2, unique)                  #hoge行列の列ごとにユニークな文字列を

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)             #「hoge$run」で指定したSRRから始まるI
k                                       #kの中身を表示
hoge2 <- cbind(k$library.name,          #ライブラリ名と、
               k$run.read.count,        #総リード数と、
               k$file.name,             #ファイル名と、
               k$file.size)             #ファイルサイズ、の順番で列方向で結合した結果をhoge2に格
```



SRAmetadb.sqliteファイルのダウンロードは結構大変だが、一度ダウンロードしておけば格納されているSRAメタデータ情報は利用可能



6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb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x時間程度かかります。

```
#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ある"SRAm
sqlfile <- getSRAdbFile()               #最新のS
sra_con <- dbConnect(SQLite(), sqlfile)#おまじな

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#para
hoge                                    #hogeの中
apply(hoge, 2, unique)                  #hoge行列

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)             # 「hoge$
k                                       #kの中身
hoge2 <- cbind(k$library.name,          #ライブラ
               k$run.read.count,        #総リード
               k$file.name,             #ファイル
               k$file.size)             #ファイル
hoge2                                   #hoge2の

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

R Console

```
> hoge <- sraConvert(param, sra_con=sra_con)#paramで指$
> hoge                                      #hogeの中身を$
      study submission    sample experiment        run
1 SRP017580  SRA062939 SRS380021  SRX210737 SRR633902
2 SRP017580  SRA062939 SRS380023  SRX210739 SRR649760
3 SRP017580  SRA062939 SRS380022  SRX210738 SRR649759
4 SRP017580  SRA062939 SRS380020  SRX210736 SRR633901
> apply(hoge, 2, unique)                    #hoge行列の列$
$study
[1] "SRP017580"

$submission
[1] "SRA062939"

$sample
[1] "SRS380021" "SRS380023" "SRS380022" "SRS380020"

$experiment
[1] "SRX210737" "SRX210739" "SRX210738" "SRX210736"

$run
[1] "SRR633902" "SRR649760" "SRR649759" "SRR633901"

>
```

ENAでみられるメタデータ情報がR Console画面上でも見られます

6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb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x時間程度かかります。

```
#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ある"SRAm
sqlfile <- getSRAdbFile()               #最新のS
sra_con <- dbConnect(SQLite(), sqlfile)#おまじな

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#para
hoge                                    #hogeの中
apply(hoge, 2, unique)                  #hoge行列

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)             # 「hoge$
k                                       #kの中身
hoge2 <- cbind(k$library.name,          #ライブラ
               k$run.read.count,        #総リード
               k$file.name,             #ファイル
               k$file.size)             #ファイル
hoge2

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

R Console

```
> hoge <- sraConvert(param, sra_con=sra_con)#paramで指$
> hoge                                      #hogeの中身を$
      study submission     sample experiment        run
1 SRP017580  SRA062939 SRS380021  SRX210737 SRR633902
2 SRP017580  SRA062939 SRS380023  SRX210739 SRR649760
3 SRP017580  SRA062939 SRS380022  SRX210738 SRR649759
4 SRP017580  SRA062939 SRS380020  SRX210736 SRR633901
> apply(hoge, 2, unique)                    #hoge行列の列$
$study
```

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

ENAでみられるメタデータ情報が
R Console画面上でも見られます

①ライブラリ名、②総リード数、
③ファイル名、④ファイルサイズ

```
#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
hoge                                      #hogeの中身を表示
apply(hoge, 2, unique)                    #hoge行列の列ごとにユニークな文字列を

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)               # 「hoge$run」で指定したSRRから始まる
k                                         #kの中身を表示
hoge2 <- cbind(k$library.name,            #ライブラリ名
               k$run.read.count,          #総リード数
               k$file.name,               #ファイル名
               k$file.size)               #ファイルサ
hoge2                                     #hoge2の中

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

R Console

```
6 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR633/SRR633901$
> hoge2 <- cbind(k$library.name,          #ライブラリ$ ①
+                k$run.read.count,        #総リード数$ ②
+                k$file.name,             #ファイル名$③
+                k$file.size)             #ファイルサ$  ④
> hoge2                                  #hoge2の中身$
     [,1]   [,2]       [,3]                      [,4]
[1,] "null" "16817117" "SRR633902_1.fastq.gz" "1Gb"
[2,] "null" "16817117" "SRR633902_2.fastq.gz" "1Gb"
[3,] "null" "36487727" "SRR649760.fastq.gz"   "2Gb"
[4,] "null" "26457253" "SRR649759.fastq.gz"   "2Gb"
[5,] "null" "22419833" "SRR633901_1.fastq.gz" "1Gb"
[6,] "null" "22419833" "SRR633901_2.fastq.gz" "1Gb"
>
```

① ② ③ ④

```
#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
hoge                                      #hogeの中身を表示
apply(hoge, 2, unique)                    #hoge行列の列ごとにユニークな文字列を

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)               # 「hoge$
k                                         #kの中身
hoge2 <- cbind(k$library.name,            #ライブラ
               k$run.read.count,          #総リード
               k$file.name,               #ファイル
               k$file.size)               #ファイル
hoge2                                     #hoge2の

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

R Console

```
> hoge$run
[1] "SRR633902" "SRR649760" "SRR649759" "SRR633901"
> dim(k)
[1]  6 19
> colnames(k)
 [1] "study"                "sample"
 [3] "experiment"           "run"
 [5] "analysis"             "organism"
 [7] "instrument.platform"  "instrument.model"
 [9] "library.name"         "library.layout"
[11] "library.strategy"     "library.source"
[13] "library.selection"    "run.read.count"
[15] "run.base.count"       "file.name"
[17] "file.size"            "md5"
[19] "ftp"
> k$sample
[1] "SAMN01831456" "SAMN01831456" "SAMN01831458"
[4] "SAMN01831457" "SAMN01831455" "SAMN01831455"
> k[,2]
[1] "SAMN01831456" "SAMN01831456" "SAMN01831458"
[4] "SAMN01831457" "SAMN01831455" "SAMN01831455"
>
```

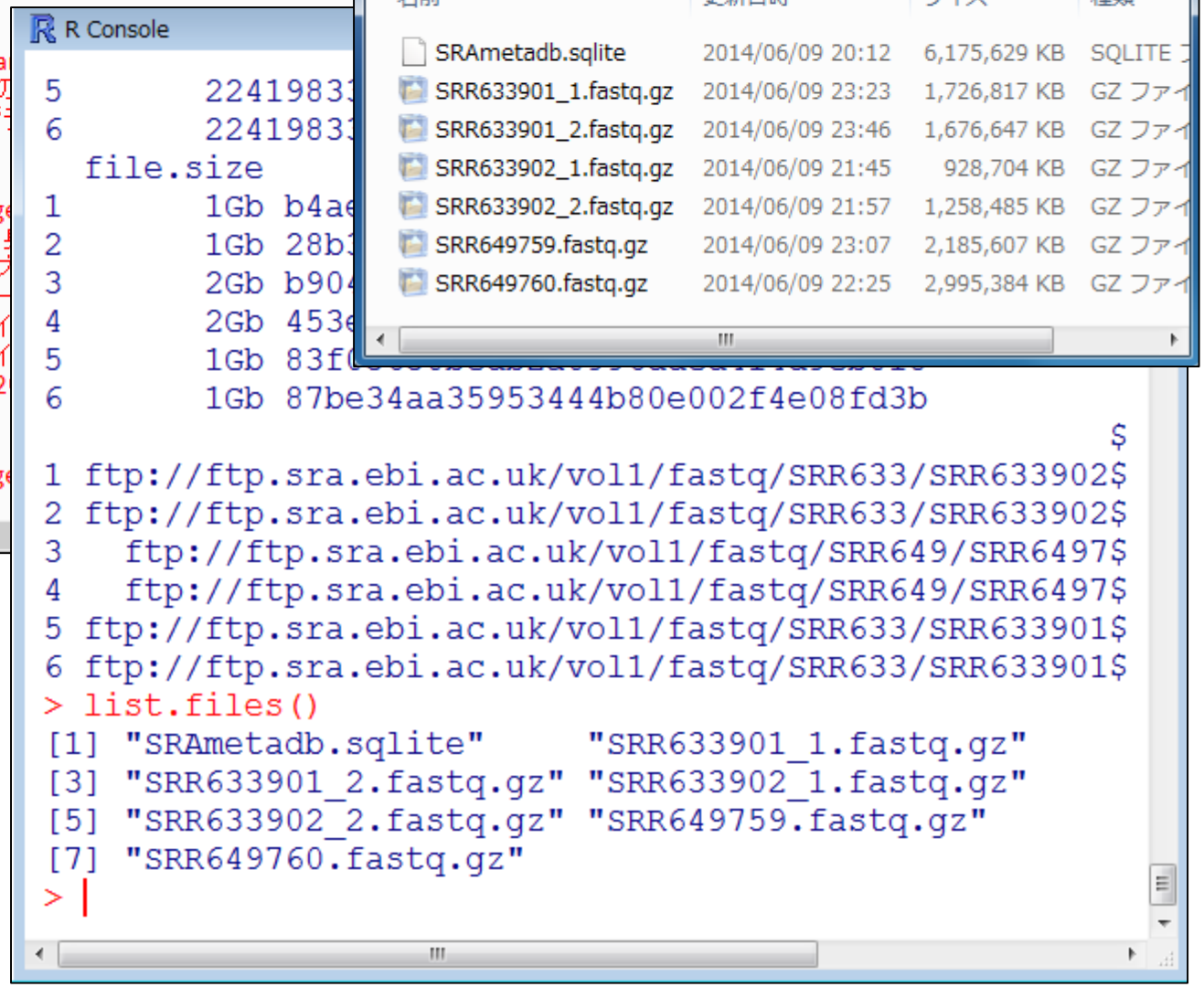
kオブジェクトは6行×19列からなる

①ライブラリ名、②総リード数、
③ファイル名、④ファイルサイズ

6. ヒトのRNA-seqデータ("SRP017580": Chan et al., Hum. Mol. Genet., 2013)のgzip圧縮済みのFASTQファイルをダウンロードする場合:

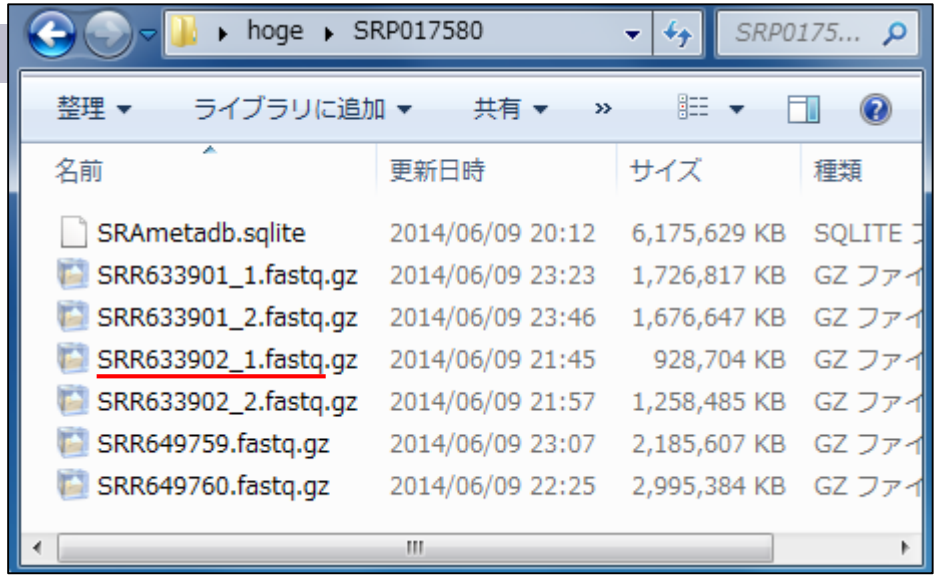
Illumina HiSeq 2000のpaired-endデータです。計xファイル、合計x Gb程度の容量のファイルが
の有線LANでx時間程度かかります。

```
#前処理
#sqlfile <- "SRAmetadb.sqlite"          #最新でなくてもよく、手元に予め
sqlfile <- getSRAdbFile()               #最新のSRAmetadb SQLiteファイル
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)#pa
hoge                                    #hoge
apply(hoge, 2, unique)                  #hoge

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(hoge$run)             # 「hoge
k                                       #kの中
hoge2 <- cbind(k$library.name,          #ライ
               k$run.read.count,        #総リ
               k$file.name,             #ファイ
               k$file.size)             #ファイ
hoge2                                   #hoge2

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



計6個のgzip圧縮FASTQファイルが得られます

FASTQファイル実例

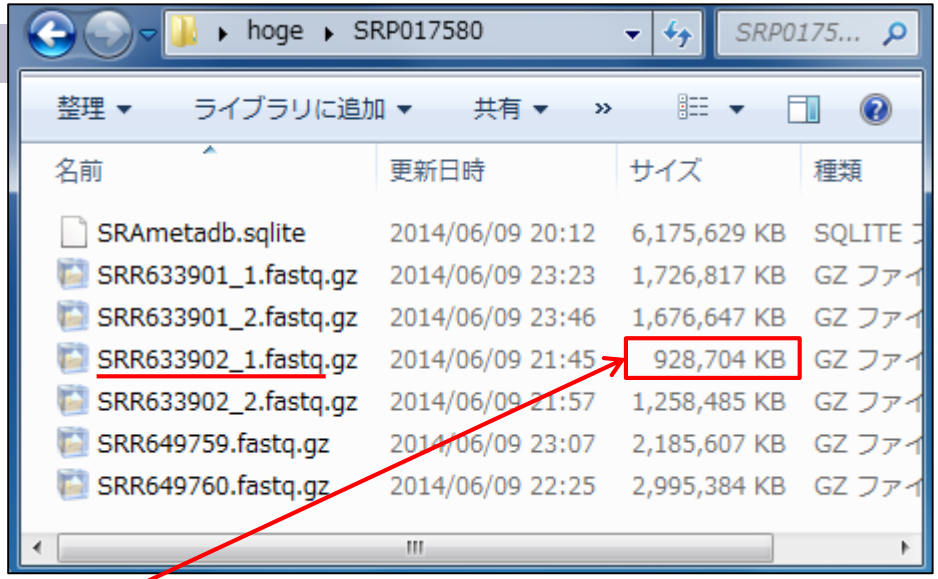


- gzip圧縮ファイル解凍後のSRR633902_1.fastq をテキストエディタで表示

```
@SRR633902.1 HWI-ST674:192:COVP8ACXX:8:1101:1603:2162/1
CTAAACTCCAGCCTGGGCGACAGAGCAAGACTCTACTTGCACTAGTCTCTTAACAAAAGGAGGCAGCTCAGTCATTACATTTTAGTGTTGCATAAAACCT
+
CCCFFFFFGGHHHIJJIGIIIJJJJIIJJIJJJIIIJIJIJGIIJGIGGJJIGGHGGIIJIHHHFFFFEEEEEEDDDCDEFEDDDDCDEDDDDDC:CC?C
@SRR633902.2 HWI-ST674:192:COVP8ACXX:8:1101:1743:2217/1
GCAGTGCACAGATAACATCAGGGTAGACTTGACTGGAGAAAACCAAATTCTGCGCTTGCTCCTGTGTGCCCCCATCCAGCTGTGCATGCACACACAGGAC
+
@@@FFFFFHHHHHGGIGJGJJIJFHIFHIGIJJJJGEIGGGIEGHGIIGJJIGIIJJIDHIIIGIHHHGHHFFA@CAACCCDDDDDDDCDD>B<BBDD?<
@SRR633902.3 HWI-ST674:192:COVP8ACXX:8:1101:1544:2220/1
GTGCTGATCTACACAAAGAAATTCGAGACACTTACTATCAACTTGTTCTGTTTTTGGTCAAAGCAGTTAAAGGATTTAGTAGCCTAAATGACAGGTCCTT
+
CCCFFFFFHHHHHJJJJJHIJJJJJJJIIJJJJJJJJJIFHIJJIIJIJHGHJJJJJJJIIIIJGIGGIHIIGFHHEHFFFFFECECCEEDCCDDCCCDC
@SRR633902.4 HWI-ST674:192:COVP8ACXX:8:1101:1971:2190/1
TTTTTTTTTAACTGATAGATGGTGCAGCATGTCTACATGGTTGTTTGTTGCTAAACTTTATATAATGTGTGGTTTCAATTCAGCTTGAAAAATAATCTCA
+
@@@DDDDDHHHHHF@@@F<DFH?FE@GEBGEHEGGGIIIIFFFIIIIIIIIIIHHFHHH@DBBCBDE@CBA?CBBCCCEE@>ACCCCBCCCB1(:@>CC:
@SRR633902.5 HWI-ST674:192:COVP8ACXX:8:1101:2360:2143/1
```

4行で1つのリード情報を表すのがFASTQ形式です

FASTQファイル実例



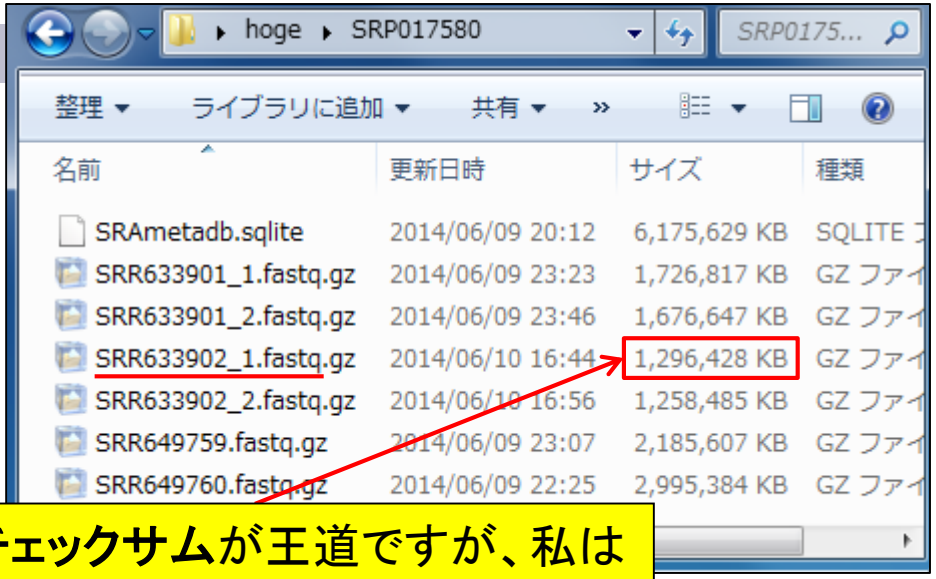
- gzip圧縮ファイル解凍後のSRR633902_1.fastq をテキストエディタで表示

解凍時に「ファイルが壊れています」などというメッセージが出たら、ダウンロードに失敗していると解釈すべし

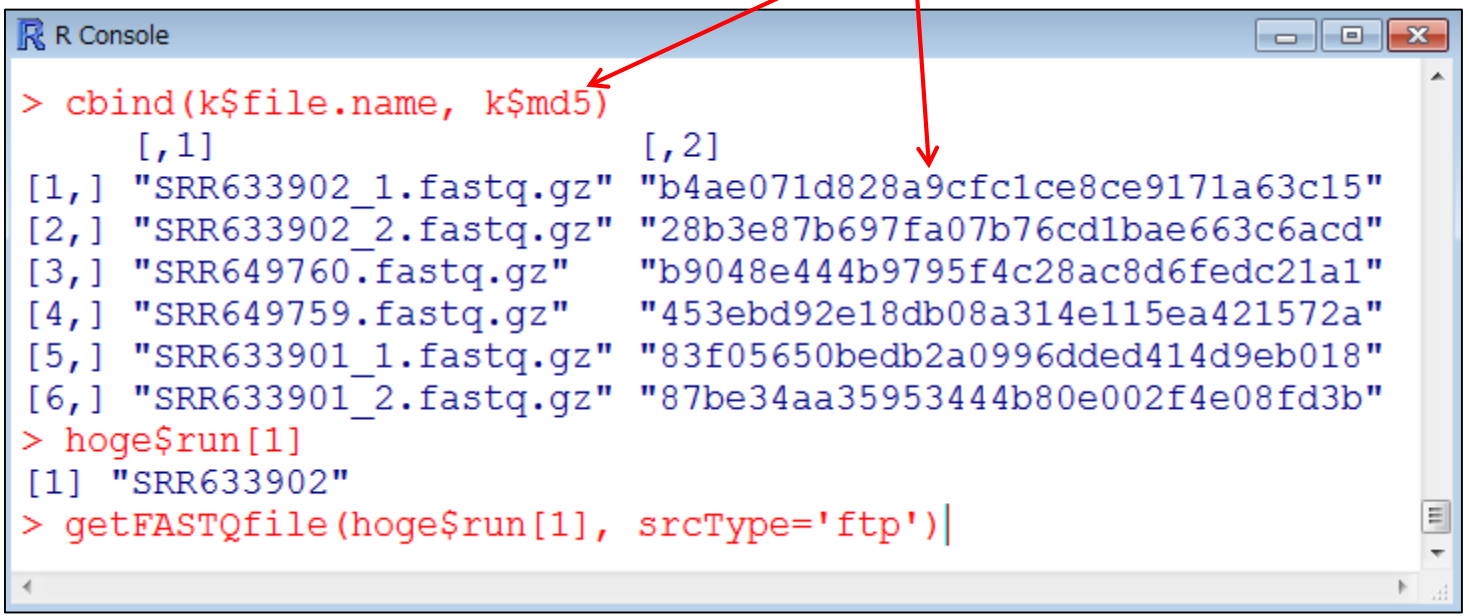
```
@SRR633902.1 HWI-ST674:192:COVP8ACXX:8:1101:1
CTAAACTCCAGCCTGGGCGACAGAGCAAGACTCTACTTGCACTAG                                GCATAAAACCT
+
CCCFFFFFGGHHHIJJIGIIIJJJJIIJJIJJJIIIJIJIJGIIJ                               DDDDDC:CC?C
@SRR633902.2 HWI-ST674:192:COVP8ACXX:8:1101:1743:2217/1
GCAGTGCACAGATAACATCAGGGTAGACTTGACTGGAGAAAACCAAATTCTGCGCTTGCTCCTGTGTGCCCCCATCCAGCTGTGCATGCACACACAGGAC
+
@@@FFFFFHHHHHGGIGJGJJIJFHIFHIGIJJJJGEIGGGIEGHGIIGJJIGIIJJIDHIIIGIHHHGHHFFA@CAACCCDDDDDDDCDD>B<BBDD?<
@SRR633902.3 HWI-ST674:192:COVP8ACXX:8:1101:1544:2220/1
GTGCTGATCTACACAAAGAAATTCGAGACACTTACTATCAACTTGTTCTGTTTTTGGTCAAAGCAGTTAAAGGATTTAGTAGCCTAAATGACAGGTCCTT
+
CCCFFFFFHHHHHJJJJJHIJJJJJJJIIJJJJJJJJJIFHIJJIIJIJHGHJJJJJJJIIIIJGIGGIHIIGFHHEHFFFFFECECCEEDCCDDCCCDC
@SRR633902.4 HWI-ST674:192:COVP8ACXX:8:1101:1971:2190/1
TTTTTTTTTAACTGATAGATGGTGCAGCATGTCTACATGGTTGTTTGTTGCTAAACTTTATATAATGTGTGGTTTCAATTCAGCTTGAAAAATAATCTCA
+
@@@DDDDDHHHHHF@@@F<DFH?FE@GEBGEHEGGGIIIIFFFIIIIIIIIIIHHFHHH@DBBCBDE@CBA?CBBCCCEE@>ACCCCBCCCB1(:@>CC:
@SRR633902.5 HWI-ST674:192:COVP8ACXX:8:1101:2360:2143/1
NTCAGGTCTGCTTTTCCAATTGGTCCTGCTTGTGAGAATGTGGAGGAATCTCTCAGCCCCAGCTCTCTTCAGCCAAGGAATCAAACATGTATTTTTTTCT
```

FASTQファイル実例

- gzip圧縮ファイル解凍後のSRR633902_1.fastq をテキストエディタで表示



MD5チェックサムが王道ですが、私はペアエンドリードの場合は、もう片方のファイルサイズと比較して判断します。基本思考停止して再ダウンロード



Contents（第2回）

- イントロダクション(Introduction)
  - NGSデータ概観（PacBioとIllumina）
  - NGSデータベース(DB)、データ形式（FASTQ形式）
  - SRAdbパッケージを用いたデータ取得、エラーへの対処
- **前処理(Pre-processing)**
  - **qrqcパッケージを用いたQuality Control (QC)**
- アダプター配列除去
  - 基本戦略（girafeパッケージ）
  - 昔は正常に動作していたのに…という例（QuasRパッケージ）
  - アダプター除去を含む様々なフィルタリングの組合せ（ShortReadパッケージ）
  - 課題

Quality Control

- イントロ | ファイル形式の変換 | qseq --> FASTA (last modified 2013/06/17)
- 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 2013/08/19)
- 前処理 | クオリティチェック | について (last modified 2014/06/10) NEW
- 前処理 | クオリティチェック | qrqc (last modified 2014/06/10) NEW
- 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 modified 2014/03/03
- 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 modified 2014/03
- 前処理 | フィルタリング | ACGTのみからなる配列を抽出 (last modified 2013/06/18)

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。

前処理 | クオリティチェック | について NEW

Quality Control (QC)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。

Rパッケージ:

- 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

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
- 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
- SEQuel: Ronen et al., Bioinformatics, 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

Review:

- Paszkiewicz et al., Front Genet., 2014

qrqcパッケージで全体像を眺める

- イントロ | ファイル形式の変換 | qseq --> FASTA (last modified 2013/06/17)
- 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 2013/08/19)
- 前処理 | クオリティチェック | について (last modified 2014/06/10) NEW
- 前処理 | クオリティチェック | qrqc (last modified 2014/06/10) NEW
- 前処理 | クオリティチェック | PHRED スコアに変換 (last modified 2013/06/18)

qrqcはシンプルだが、若干動作が不安定な印象。使われた実績はあり

前処理 | クオリティチェック | qrqc NEW

FastQCのR版のようなものです。Sanger FASTQ形式ファイルを読み込んで、positionごとの「クオリティスコア(quality score)」、「どんな塩基が使われているのか(base frequency and base proportion)」、「リード長の分布」、「GC含量」、「htmlレポート」などを出力してくれます。

2. サンプルデータ25のFASTQ形式ファイル(SRR633902_1_sub.fastq)の場合:

SRR633902から得られるFASTQファイルの最初の2,000行分を抽出したヒトデータです (Chan et al., Hum. Mol. Genet., 2013)。下記を実行すると「SRR633902_1_sub」という名前のフォルダが作成されます。中にあるreport.htmlをダブルクリックするとhtmlレポートを見ることができます。「Error : Insufficient values in manual scale. 2 needed but only 1 provided.」というエラーが出ることは確認済みですが、htmlファイルは作成されます。

```
in_f <- "SRR633902_1_sub.fastq"          #入力ファイル名を指定してin_fに格納

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

#入力ファイルの読み込み
fastq <- readSeqFile(in_f)               #in_fで指定したファイルの読み込み

#本番
makeReport(fastq)                        #htmlレポートの作成
```

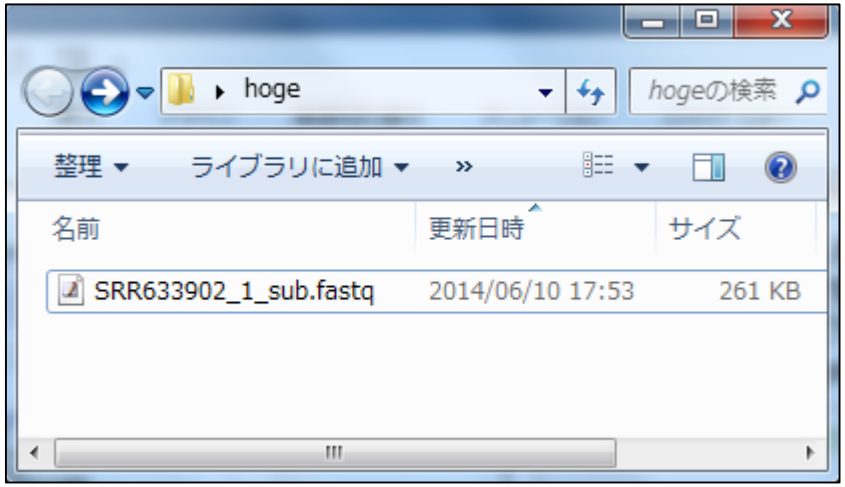
- qrqc:原著論文はまだ?!
- qrqc利用論文:Loman et al., Nat Biotechnol., 2012

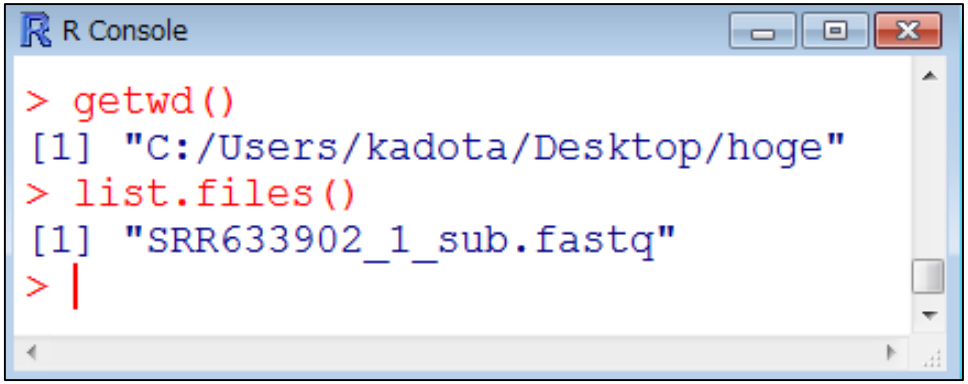
2. サンプルデータ25のFASTQ形式ファイル(SRR633902_1_sub.fastq)の場合:

SRR633902から得られるFASTQファイルの最初の2,000行分を抽出したヒトデータです (Chan et al., Hum. Mol. Genet., 2013)。下記を実行すると「SRR633902_1_sub」という名前のフォルダが作成されます。中にあるreport.htmlをダブルクリックするとhtmlレポートを見ることができます。「Error : Insufficient values in manual scale. 2 needed but only 1 provided.」というエラーが出ることは確認済みですが、htmlファイルは作成されます。

```
in_f <- "SRR633902_1_sub.fastq"          #入力ファイル名を指定してin_fに格納

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

#入力ファイルの読み込み
fastq <- readSeqFile(in_f)               #in_fで指定したファイルの読み込み

#本番
makeReport(fastq)                        #htmlレポートの作成
```

hogeフォルダ中にあり。通常利用時は、(Windowsのヒトは)右クリックでファイルの保存。





2. サンプルデータ25のFASTQ形式ファイル(SRR633902_1_sub.fastq)の場合:

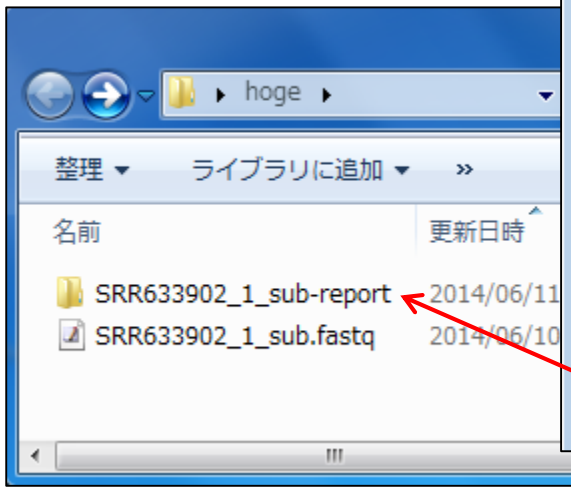
SRR633902から得られるFASTQファイルの最初の2,000行分を抽出したヒトデータです (Chan et al., Hum. Mol. Genet., 2013)。下記を実行すると「SRR633902_1_sub」という名前のフォルダが作成されます。中にあるreport.htmlをダブルクリックするとhtmlレポートを見ることができます。「Error : Insufficient values in manual scale. 2 needed but only 1 provided.」というエラーが出ることは確認済みですが、htmlファイルは作成されます。

```
in_f <- "SRR633902_1_sub.fastq"                    fに格納

#必要なパッケージをロード
library(qrqc)

#入力ファイルの読み込み
fastq <- readSeqFile(in_f)

#本番
makeReport(fastq)
```

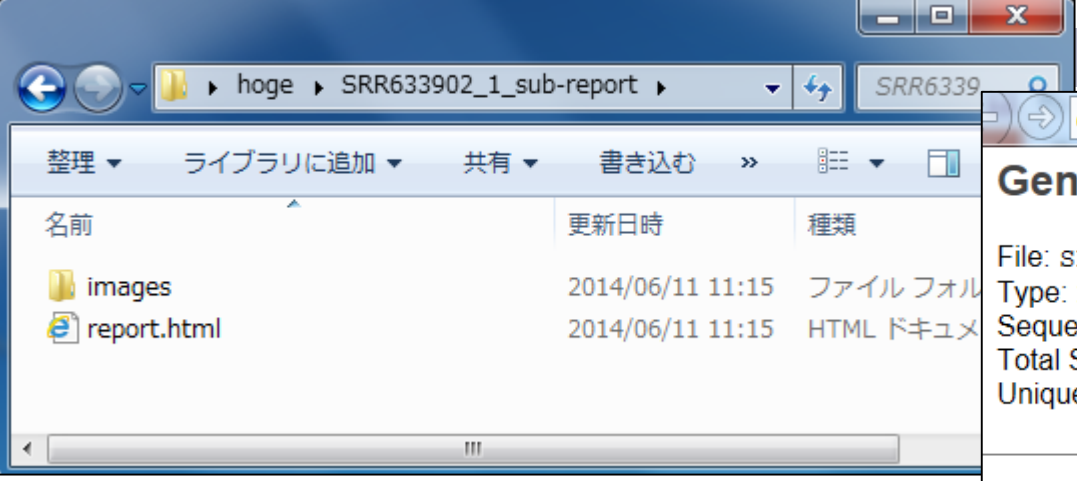
切り取り(T)
コピー(C)

R Console

```
> #入力ファイルの読み込み
> fastq <- readSeqFile(in_f)                    #in_fで指定したファイ$
>
> #本番
> makeReport(fastq)                             #htmlレポートの作成
Error : .onAttach failed in attachNamespace() for 'mgcv', detai$
  call: formatDL(nm, txt, indent = max(nchar(nm, "w")) + 3)
  error: incorrect values of 'indent' and 'width'
Error : mgcv package required for this functionality.  Please i$
Error : Insufficient values in manual scale. 2 needed but only $
Report written to directory './SRR633902_1_sub-report'.
> list.files()
[1] "SRR633902_1_sub-report"
[2] "SRR633902_1_sub.fastq"
>
```



R colsole画面上でペーストすると、エラーは出るが
SRR633902_1_sub-reportというフォルダは作成される



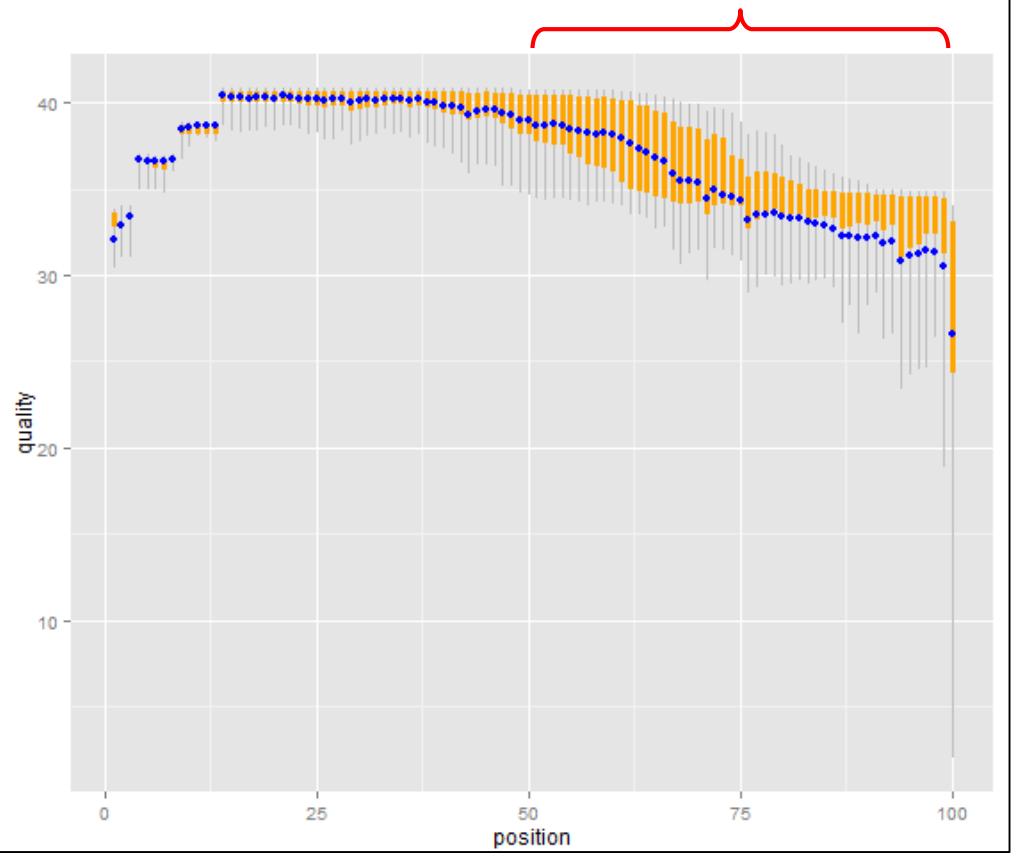
縦軸：Phred quality score、横軸：塩基のポジション。50 bpあたりから100 bpにかけてクオリティ値が下がってますが、これが一般的な傾向です

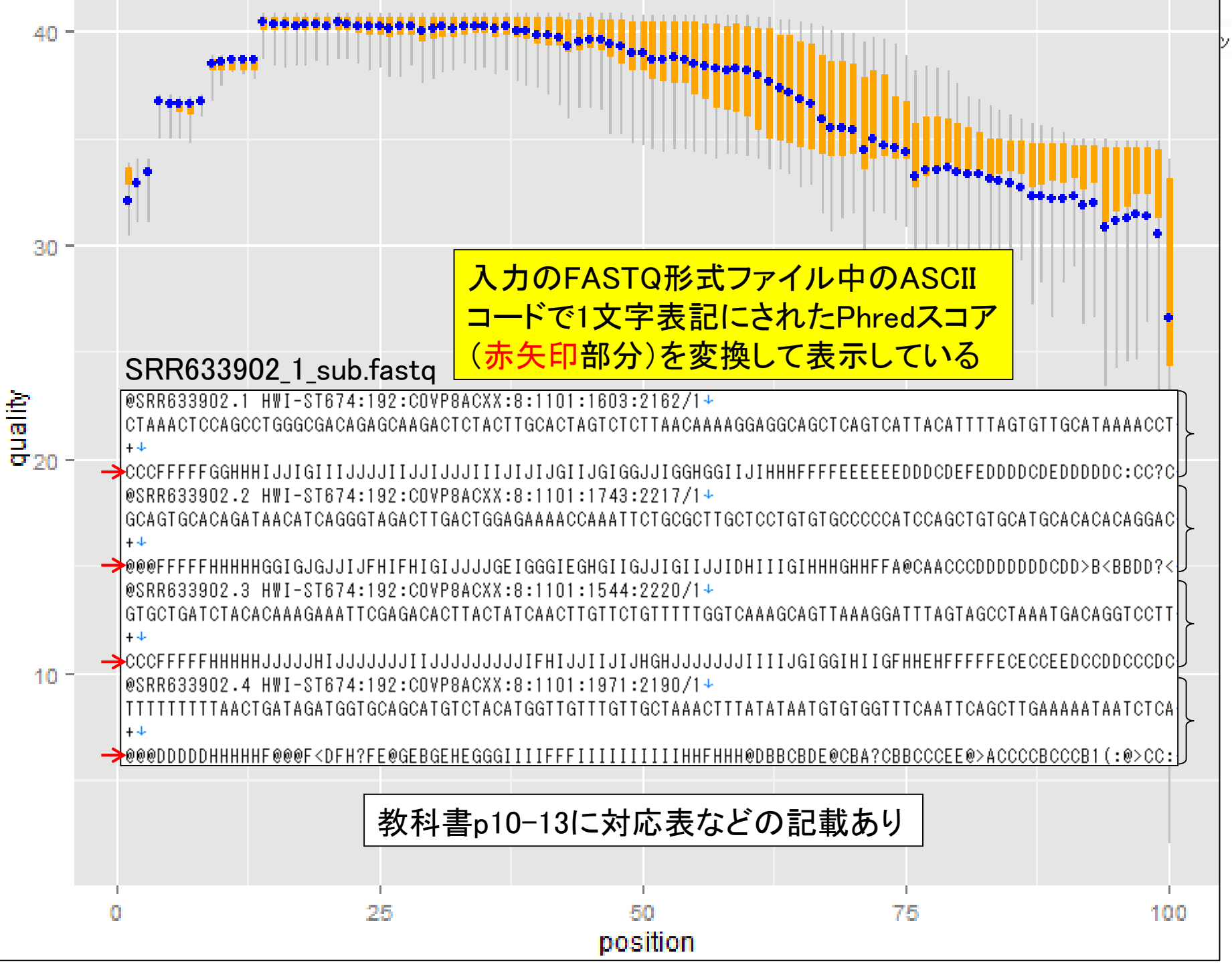
C:¥Users¥kadota¥Desktop¥hoge¥SRR633902_1_sub-report¥report.html

General Information

File: SRR633902_1_sub.fastq
Type: FASTQ
Sequence Length Range: 100 to 100
Total Sequences: 1000
Unique Sequences: 906

Quality by Position





入力のFASTQ形式ファイル中のASCIIコードで1文字表記にされたPhredスコア（赤矢印部分）を変換して表示している

SRR633902_1_sub.fastq

```
@SRR633902.1 HWI-ST674:192:COVP8ACXX:8:1101:1603:2162/1
CTAAACTCCAGCCTGGGCGACAGAGCAAGACTCTACTTGCACTAGTCTCTTAACAAAAGGAGGCAGCTCAGTCATTACATTTTAGTGTTGCATAAAACCT
+
CCCFFFFFGGHHHIJJIGIIIJJJJIIJJIJJJIIIJIJIJGIIJGIGGJJIGGHGGIIJIHHHFFFFEEEEEEDDDCDEFEDDDDCDEDDDDDC:CC?C
@SRR633902.2 HWI-ST674:192:COVP8ACXX:8:1101:1743:2217/1
GCAGTGCACAGATAACATCAGGGTAGACTTGACTGGAGAAAACCAAATTCTGCGCTTGCTCCTGTGTGCCCCCATCCAGCTGTGCATGCACACACAGGAC
+
@@@FFFFFHHHHHGGIGJGJJIJFHIFHIGIJJJJGEIGGGIEGHGIIGJJIGIIJJIDHIIIGIHHHGHHFFA@CAACCCDDDDDDDCDD>B<BBDD?<
@SRR633902.3 HWI-ST674:192:COVP8ACXX:8:1101:1544:2220/1
GTGCTGATCTACACAAAGAAATTCGAGACACTTACTATCAACTTGTTCTGTTTTTGGTCAAAGCAGTTAAAGGATTTAGTAGCCTAAATGACAGGTCCTT
+
CCCFFFFFHHHHHJJJJJHIJJJJJJJIIJJJJJJJJJIFHIJJIIJIJHGHJJJJJJJIIIIJGIGGIHIIGFHHEHFFFFFECECCEEDCCDDCCCDC
@SRR633902.4 HWI-ST674:192:COVP8ACXX:8:1101:1971:2190/1
TTTTTTTTTAACTGATAGATGGTGCAGCATGTCTACATGGTTGTTTGTTGCTAAACTTTATATAATGTGTGGTTTCAATTCAGCTTGAAAAATAATCTCA
+
@@@DDDDDHHHHHF@@@F<DFH?FE@GEBGEHEGGGIIIIFFFIIIIIIIIIIHHFHHH@DBBCBDE@CBA?CBBCCCEE@>ACCCCBCCCB1(:@>CC:
```

教科書p10-13に対応表などの記載あり

Quality Control

- 作業内容
  - 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）
  - …

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ます

前処理 | クオリティチェック | について NEW

Quality Control (QC)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。

Rパッケージ:

- 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

R以外:

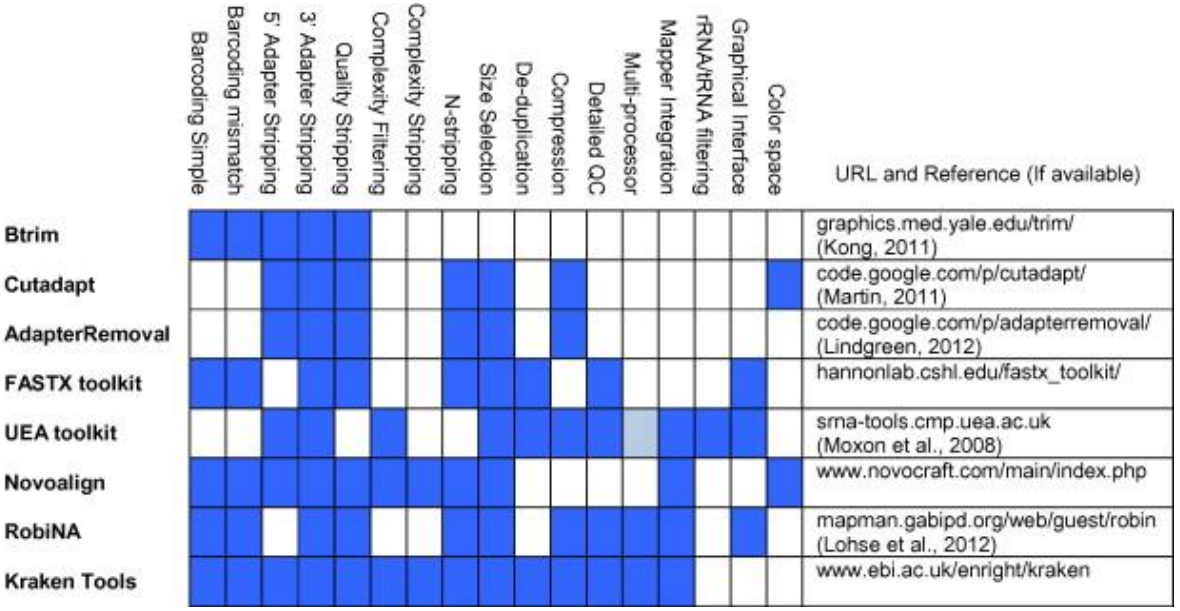
- FastQC: 原著論文なし
- FASTX-ToolKit: 原著論文なし
- SolexaQA: Cox et al., BMC Bioinformatics, 2010
- Quake: Kelley et al., Genome Biol., 2010
- NGSQC: Dai et al., BMC Genomics, 2010
- Cutadapt: Martin, M., EMBnet.journal, 2011
- PRINSEQ: Schmieder and Edwards, Bioinformatics, 2011
- ECHO: Kao et al., Genome Res., 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

Review:

- Paszkiewicz et al., Front Genet., 2014

Kraken

- LinuxとMac用のみ



	Barcoding Simple	Barcoding mismatch	5' Adapter Stripping	3' Adapter Stripping	Quality Stripping	Complexity Filtering	Complexity Stripping	N-stripping	Size Selection	De-duplication	Compression	Detailed QC	Multi-processor	Mapper Integration	rRNA/tRNA filtering	Graphical Interface	Color space	URL and Reference (If available)
Btrim	■	■	■	■	■													graphics.med.yale.edu/trim/ (Kong, 2011)
Cutadapt			■	■	■			■	■		■						■	code.google.com/p/cutadapt/ (Martin, 2011)
AdapterRemoval			■	■	■			■	■		■							code.google.com/p/adapterremoval/ (Lindgreen, 2012)
FASTX toolkit	■	■		■	■			■	■	■		■				■		hannonlab.cshl.edu/fastx_toolkit/
UEA toolkit			■	■		■			■	■	■	■		■	■	■		sma-tools.cmp.uea.ac.uk (Moxon et al., 2008)
Novoalign	■	■	■	■	■	■	■	■	■					■			■	www.novocraft.com/main/index.php
RobiNA	■	■		■	■			■	■		■	■	■	■				mapman.gabipd.org/web/guest/robin (Lohse et al., 2012)
Kraken Tools	■	■	■	■	■	■	■	■	■	■	■	■	■	■				www.ebi.ac.uk/enright/kraken

Table 1

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, 201
- Kraken: Davis et al., Methods, 2013

Review:

- Paszkiewicz et al., Front Genet., 2014

Table 1を見るといろいろできるように見えるが…

Review論文だと…

	HTQC	FastQC	SolexaQA	NGS QC Toolkit	Kraten	QC-Chain
Language	C++	Java	Perl	Perl	C	C++
Q score boxplot	■	■				
Tile based Q scores	■		■			
Duplication removal					■	■
Filtering	■			■	■	■
Trimming	■		■	■	■	■
Adaptor detection		■	■	■	■	■
Contamination detection				■	■	■

Comparison of feature of software packages for quality control of Illumina read data. Shaded areas indicate that the feature is present.

Table 1

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

Review:

- Paszkiewicz et al., Front Genet., 2014

評価項目によって印象は変わりますね…

Quality Control

- 作業内容
  - 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）
  - …

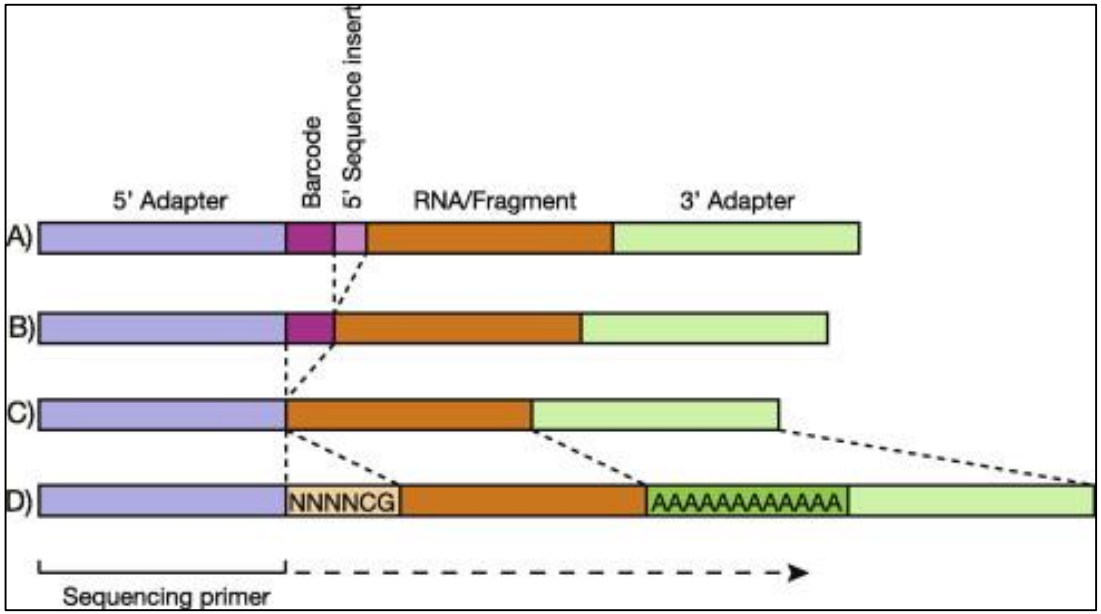


Fig. 2

特にアダプター配列除去(adapter removal)は small RNA sequencing (sRNA-seq)の場合に、マップ率に多大な影響を及ぼします

アダプター配列除去

girafeパッケージのデフォルト設定はイマイチですが、感覚をつかむ上では便利なのでそれを利用して説明します

- 前処理 | トリミング | ポリA配列除去 | ShortRead(Morgan_2009)(last mod…d 2014/06/11) NEW
- 前処理 | トリミング | アダプター配列除去(基礎) | girafe(Toedling_2010)(…modified 2014/06/11)
- 前処理 | トリミング | アダプター配列除去(基礎) | ShortRead(Morgan_200…(last modified 2014/06/1
- 前処理 | トリミング | アダプター配列除去(応用) | QuasR(Lerch_20XX)(last modified 2014/06/11) N
- 前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan_2009)(last modified 2014/06/13)推奨
- 前処理 | トリミング | …
- アセンブル | について
- アセンブル | ゲノム用
- アセンブル | トランス
- アセンブル | ゲノム既

前処理 | トリミング | アダプター配列除去(基礎) | girafe(Toedling_2010) NEW

girafeパッケージを用いたアダプター配列除去を行うやり方を示します。アダプター配列除去を行うやり方を示します。
注意点1としては、実際に塩基配列長が短くなっていてもdescription行の記述(特に配列長情報の記述)は変わりませんので、「なんかおかしい」と気にしなくて大丈夫です。
注意点2としては、例えば、アダプター配列の5'側が「CATCG...」となっているにも関わらずなぜ二番目の配列の3'側「...CATAG」の最後の5塩基がトリムされているのだろう？と疑問に思われる方がいらっしゃるかもしれませんが、これは。R Console上で「?trimAdapter」と打ち込んでデフォルトのオプションを眺めることで理由がわかります。つまり、アラインメントスコア計算時に、この関数はデフォルトで一致に1点、不一致に-1点を与えて一塩基づつオーバーラップの度合いを上げていく、という操作をしているからです。したがって、もし完全一致のみに限定したい場合は、trimAdapter関数のところで、不一致に対して大幅に減点するようなパラメータを与えればいいんです。例えば「mismatch.score = -1000」とか。。。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです(Bullard et al., BMC Bioinformatics, 2010)。
デフォルトのパラメータ(一致に1点、不一致に-1点)でトリムし、FASTQ形式で保存するやり方です。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。

```
in_f <- "SRR037439.fastq"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fastq"                       #出力ファイル名を指定してout_fに格納
param_adapter <- "CATCGATCCTGCAGGCTAGAGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG"#アダプター配

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

#入力ファイルの読み込み
fastq <- readFastq(in_f)                     #in_fで指定したファイルの読み込み
sread(fastq)                                 #配列情報を表示
```

アダプター配列除去のイメージ

- アダプター配列：CATCGATCCTGCAGGCTAGAGACAGAT…
- FASTQ形式ファイル：SRR037439.fastq

```
@SRR037439.1 HWI-E4_6_30ACL:2:1:0:176 length=35
NNNNNNNNNNNNNNCTACCCCCCCAGCCGCCGCA
+SRR037439.1 HWI-E4_6_30ACL:2:1:0:176 length=35
!!!!!!!!!!!!!!!"""""""""""""""""""#"""
@SRR037439.2 HWI-E4_6_30ACL:2:1:0:252 length=35
NNNNNNNNNNNNNNAGACAGTTGATTTAGCATAG
+SRR037439.2 HWI-E4_6_30ACL:2:1:0:252 length=35
!!!!!!!!!!!!!!!"+"""""""""""""""""""&
@SRR037439.3 HWI-E4_6_30ACL:2:1:0:1152 length=35
NNNNNNNNNNNNNNGGGTGGGGCGTTTGTTCTTG
+SRR037439.3 HWI-E4_6_30ACL:2:1:0:1152 length=35
!!!!!!!!!!!!!!!/5$$&""""%""""""""""
@SRR037439.4 HWI-E4_6_30ACL:2:1:0:1349 length=35
NNNNNNNNNNNNNNCCCCGCCCCGCCCCTCCCTC
+SRR037439.4 HWI-E4_6_30ACL:2:1:0:1349 length=35
@SRR037439.5 HWI-E4_6_30ACL:2:1:0:1669 length=35
NNNNNNNNNNNNNNGGTCGCCCCCGACGCCCACA
+SRR037439.5 HWI-E4_6_30ACL:2:1:0:1669 length=35
!!!!!!!!!!!!!!!",""""#""""""""&""""&"
```

3'側の2塩基が除去される

3'側の5塩基が除去される

3'側の0塩基が除去される

3'側の4塩基が除去される

アラインメント時に指定するパラメータ(一致に何点、不一致に…)次第で結果が変わる

3'側の2塩基が除去される

アダプター配列除去のイメージ

- アダプター配列：CATCGATCCTGCAGGCTAGAGACAGAT…
- FASTQ形式ファイル：SRR037439.fastq

```
@SRR037439.1 HWI-E4_6_30ACL:2:1:0:176 length=35
NNNNNNNNNNNNNNCTACCCCCCCAGCCGCCGCA
+SRR037439.1 HWI-E4_6_30ACL:2:1:0:176 length=35
!!!!!!!!!!!!!!!"""""""""""""""""""#"""
@SRR037439.2 HWI-E4_6_30ACL:2:1:0:252 length=35
NNNNNNNNNNNNNNNAGACAGTTGATTTAGCATAG
+SRR037439.2 HWI-E4_6_30ACL:2:1:0:252 length=35
!!!!!!!!!!!!!!!"+"""""""""""""""""""&
@SRR037439.3 HWI-E4_6_30ACL:2:1:0:1152 length=35
NNNNNNNNNNNNNNNGGGTGGGGCGTTTGTTCTTG
+SRR037439.3 HWI-E4_6_30ACL:2:1:0:1152 length=35
!!!!!!!!!!!!!!!/5$$&""""%""""""""""
@SRR037439.4 HWI-E4_6_30ACL:2:1:0:1349 length=35
NNNNNNNNNNNNNNNCCCCGCCCCGCCCCTCCCTC
+SRR037439.4 HWI-E4_6_30ACL:2:1:0:1349 length=35
!!!!!!!
@SRR037
NNNNNNN
+SRR037439.5 HWI-E4_6_30ACL:2:1:0:1669 length=35
!!!!!!!!!!!!!!!",""""#""""""""&""""&"
```

この結果は一致に+1点、不一致に-1点を与えた場合です。具体的にどういう計算をしているのだろう?

3'側の2塩基が除去される

3'側の5塩基が除去される

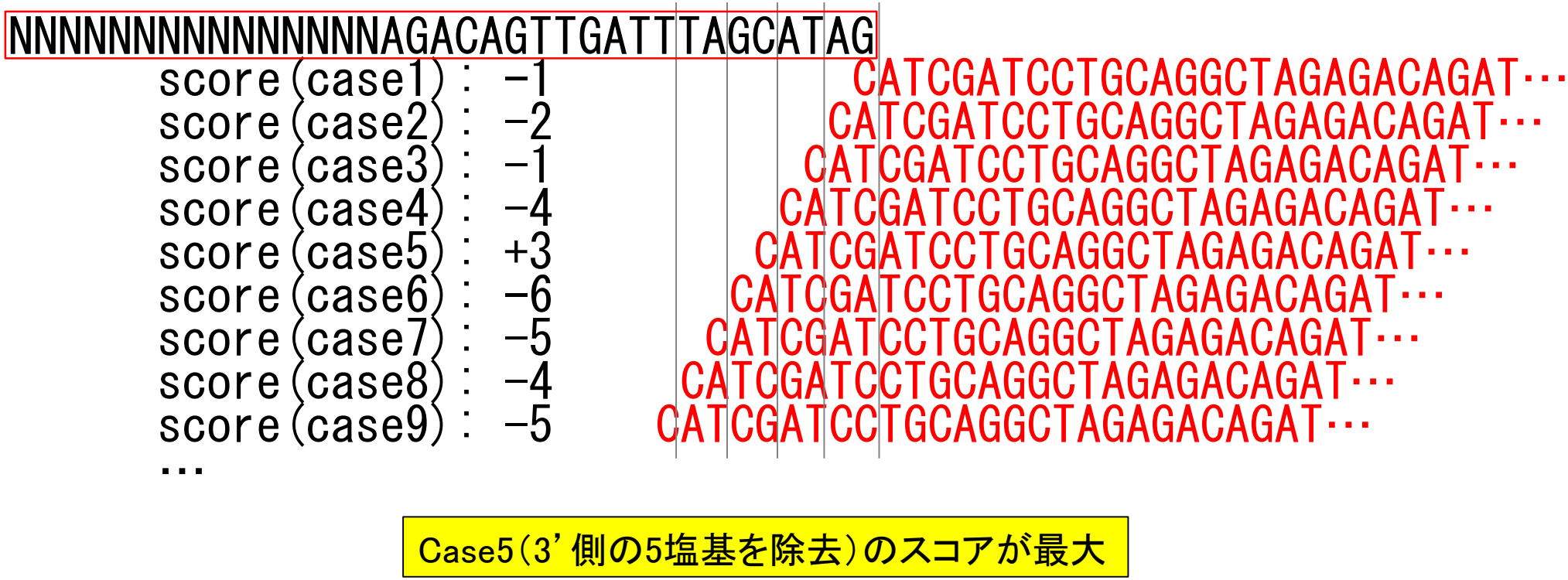
3'側の0塩基が除去される

3'側の4塩基が除去される

3'側の2塩基が除去される

アダプター配列除去のイメージ

- 一塩基づつずらしたアラインメントのoverlapの範囲で一致(+1), 不一致(-1)の総和を計算し、最も得点の高かったものを採用している



Case5(3'側の5塩基を除去)のスコアが最大

アダプター配列除去

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの 最初の2,000行分を抽出したMAQC2 brainデータです (Bullard et al., BMC Bioinformatics, 2010)。
デフォルトのパラメータ(一致に1点、不一致に-1点)でトリムし、FASTQ形式で保存するやり方です。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。

```
in_f <- "SRR037439.fastq"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fastq"                       #出力ファイル名を
param_adapter <- "CATCGATCCTGCAGGCTAGAGACAGATCGGAAGAGC

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

#入力ファイルの読み込み
fastq <- readFastq(in_f)                     #in_fで指定したファイルの読み込み
sread(fastq)                                 #配列情報を表示
table(width(fastq))                          #配列長ごとの出現頻度情報を表示

#本番
fastq <- trimAdapter(fastq, param_adapter)#trimAdapter関数を用いてアダプター配列除去した結果
sread(fastq)                                 #配列情報を表示
table(width(fastq))                          #配列長ごとの出現頻度情報を表示

#ファイルに保存
writeFastq(fastq, out_f, compress=F)    #fastqの中身を指定したファイル名で保存
```

girafeパッケージのデフォルト設定はイマイチですが、感覚をつかむ上では便利なのでそれを利用して説明します

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです(Bullard et al., BMC Bioinformatics, 2010)。デフォルトのパラメータ(一致に1点、不一致に-1点)でトリムし、FASTQ形式で保存するやり方です。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。

```
in_f <- "SRR037439.fastq"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fastq"                       #出力ファイル名を指定してout_fに格納
param_adapter <- "CATCGATCCTGCAGGCTAGAGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG"#アダプター配列を指定

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

#入力ファイルの読み込み
fastq <- readFastq(in_f)                     #in_fで指定
sread(fastq)                                 #配列情報を
table(width(fastq))                          #配列長ごと

#本番
fastq <- trimAdapter(fastq, param_adapter)#trimAd
sread(fastq)                                 #配列情報を
table(width(fastq))                          #配列長ごと

#ファイルに保存
writeFastq(fastq, out_f, compress=F)    #fastqの中
```

R Console

```
> #入力ファイルの読み込み
> fastq <- readFastq(in_f)                    #in_fで$
> sread(fastq)                                #配列情$
  A DNAStringSet instance of length 500
      width seq
  [1]    35 NNNNNNNNNNNNNNNCTACCCCCCCAGCCGCCGCA
  [2]    35 NNNNNNNNNNNNNNNAGACAGTTGATTTAGCATAG
  [3]    35 NNNNNNNNNNNNNNNGGGTGGGGCGTTTGTTCTTG
  [4]    35 NNNNNNNNNNNNNNNCCCCGCCCCGCCCCTCCCTC
  [5]    35 NNNNNNNNNNNNNNNGGTCGCCCCCGACGCCCACA
  ...   ... ...
[496]    35 CTGGACGGCCCCCCCCCCACACACCCACCCCCCCC
[497]    35 TGTCACTTGTGCTTTGCTCTTGTCCCACGGGGCTT
[498]    35 CCGCCCTTTTTCCAGAAATTTTCCGCACAAAAAAA
[499]    35 CGTTCTTGTTGCCCCCGGGGCGGGGGGGAAAAACC
[500]    35 GGAGCCTCCCCCCCCCCCCAAGGGGGGGGGGGGGC
> table(width(fastq))                         #配列長$

 35
500
>
```

FASTQファイル読み込み後のリード塩基配列情報はsread関数で抽出可能。table関数を用いて配列長分布を調べている。この場合35 bpのものが500個あったということ

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです(Bullard et al., BMC Bioinformatics, 2010)。デフォルトのパラメータ(一致に1点、不一致に-1点)でトリムし、FASTQ形式で保存するやり方です。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。

```
in_f <- "SRR037439.fastq"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fastq"                       #出力ファイル名を指定してout_fに格納
param_adapter <- "CATCGATCCTGCAGGCTAGAGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG"#アダプター配列を指定

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

#入力ファイルの読み込み
fastq <- readFastq(in_f)                     #in_fで指定
sread(fastq)                                 #配列情報を
table(width(fastq))                          #配列長ごと

#本番
fastq <- trimAdapter(fastq, param_adapter)#trimAd
sread(fastq)                                 #配列情報を
table(width(fastq))                          #配列長ごと
```

```
@SRR037439.1 HWI-E4_6_30ACL:2:1:0:176
NNNNNNNNNNNNNNCTACCCCCCCAGCCGCCGCA
+SRR037439.1 HWI-E4_6_30ACL:2:1:0:176
!!!!!!!!!!!!!!!"""""""""""""""""#"""
@SRR037439.2 HWI-E4_6_30ACL:2:1:0:252
NNNNNNNNNNNNNNAGACAGTTGATTTAGCATAG
+SRR037439.2 HWI-E4_6_30ACL:2:1:0:252
!!!!!!!!!!!!!!!"+"""""""""""""""""""&
@SRR037439.3 HWI-E4_6_30ACL:2:1:0:1152
NNNNNNNNNNNNNNGGGTGGGGCGTTTGTTCTTG
+SRR037439.3 HWI-E4_6_30ACL:2:1:0:1152
!!!!!!!!!!!!!!!/5$$&""""%""""""""""
```

R Console

```
> #入力ファイルの読み込み
> fastq <- readFastq(in_f)                    #in_fで$
> sread(fastq)                                #配列情$
  A DNAStringSet instance of length 500
      width seq
  [1]    35 NNNNNNNNNNNNNNCTACCCCCCCAGCCGCCGCA
  [2]    35 NNNNNNNNNNNNNNAGACAGTTGATTTAGCATAG
  [3]    35 NNNNNNNNNNNNNNGGGTGGGGCGTTTGTTCTTG
  [4]    35 NNNNNNNNNNNNNNCCCCGCCCCGCCCCTCCCTC
  [5]    35 NNNNNNNNNNNNNNGGTCGCCCCCGACGCCCACA
  ...   ... ...
[496]    35 CTGGACGGCCCCCCCCCCACACACCCACCCCCCCC
[497]    35 TGTCACTTGTGCTTTGCTCTTGTCCCACGGGGCTT
[498]    35 CCGCCCTTTTTCCAGAAATTTTCCGCACAAAAAAA
[499]    35 CGTTCTTGTTGCCCCCGGGGCGGGGGGGAAAAACC
[500]    35 GGAGCCTCCCCCCCCCCCCAAGGGGGGGGGGGGGC
> table(width(fastq))                         #配列長$

 35
500
>
```

入力ファイルをちゃんと読み込めていることがわかります

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです(Bullard et al., BMC Bioinformatics, 2010)。デフォルトのパラメータ(一致に1点、不一致に-1点)でトリムし、FASTQ形式で保存するやり方です。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。

```
in_f <- "SRR037439.fastq"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fastq"                       #出力ファイル名を指定してout_fに格納
param_adapter <- "CATCGATCCTGCAGGCTAGAGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG"#アダプター配列を指定

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

#入力ファイルの読み込み
fastq <- readFastq(in_f)                     #in_fで指定
sread(fastq)                                 #配列情報を
table(width(fastq))                          #配列長ごと

#本番
fastq <- trimAdapter(fastq, param_adapter)#trimAd
sread(fastq)                                 #配列情報を
table(width(fastq))                          #配列長ごと

#ファイルに保存
writeFastq(fastq, out_f, compress=F)         #fastqの中
```

R Console

```
> #本番
> fastq <- trimAdapter(fastq, param1)     #trimAd$
> sread(fastq)                                #配列情$
A DNAStringSet instance of length 500
     width seq
[1]     33 NNNNNNNNNNNNNNNCTACCCCCCCAGCCGCCG
[2]     30 NNNNNNNNNNNNNNNAGACAGTTGATTTAG
[3]     35 NNNNNNNNNNNNNNNGGGTGGGGCGTTTGTTCTTG
[4]     31 NNNNNNNNNNNNNNNCCCCGCCCCGCCCCTC
[5]     33 NNNNNNNNNNNNNNNGGTCGCCCCCGACGCCCA
...    ... ...
[496]   35 CTGGACGGCCCCCCCCCCACACACCCACCCCCCCC
[497]   35 TGTCACTTGTGCTTTGCTCTTGTCCCACGGGGCTT
[498]   35 CCGCCCTTTTTCCAGAAATTTTCCGCACAAAAAAA
[499]   35 CGTTCTTGTTGCCCCCGGGGCGGGGGGGAAAAACC
[500]   35 GGAGCCTCCCCCCCCCCCCAAGGGGGGGGGGGGGC
> table(width(fastq))                         #配列長$

 19  20  23  25  27  29  30  31  32  33  35
  3   1   4   1   4   2   4   8   1   8 464
>
```

アダプター配列除去後のリード塩基配列情報はsread関数で抽出可能。table関数を用いて配列長分布を調べている。この場合19 bpのものが3個など…。

1. サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の場合:

SRR037439から得られるFASTQファイルの最初の2,000行分を抽出したMAQC2 brainデータです(Bullard et al., BMC Bioinformatics, 2010)。デフォルトのパラメータ(一致に1点、不一致に-1点)でトリムし、FASTQ形式で保存するやり方です。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。

```
in_f <- "SRR037439.fastq"                #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.fastq"                   #出力ファイル名を指定してout_fに格納
param_adapter <- "CATCGATCCTGCAGGCTAGAGACAGATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG"#アダプター配列を指定

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

#入力ファイルの読み込み
fastq <- readFastq(in_f)                 #in_fで指定
sread(fastq)                             #配列情報を
table(width(fastq))                      #配列長ごと

#本番
fastq <- trimAdapter(fastq, param_adapter)#trimAd
sread(fastq)                             #配列情報を
table(width(fastq))                      #配列長ごと

#ファイルに保存
writeFastq(fastq, out_f, compress=F)     #fastqの中
```

R Console

```
> #本番
> fastq <- trimAdapter(fastq, param1)     #trimAd$
> sread(fastq)                            #配列情$
  A DNAStringSet instance of length 500
      width seq
  [1]    33 NNNNNNNNNNNNNNNCTACCCCCCCAGCCGCCG
  [2]    30 NNNNNNNNNNNNNNNAGACAGTTGATTTAG
  [3]    35 NNNNNNNNNNNNNNNGGGTGGGGCGTTTGTTCTTG
  [4]    31 NNNNNNNNNNNNNNNCCCCGCCCCGCCCCTC
  [5]    33 NNNNNNNNNNNNNNNGGTCGCCCCCGACGCCCA
  ...   ... ...
[496]    35 CTGGACGGCCCCCCCCCCACACACCCACCCCCCCC
[497]    35 TGTCACTTGTGCTTTGCTCTTGTCCCACGGGGCTT
[498]    35 CCGCCCTTTTTCCAGAAATTTTCCGCACAAAAAAA
[499]    35 CGTTCTTGTTGCCCCCGGGGCGGGGGGGAAAAACC
[500]    35 GGAGCCTCCCCCCCCCCCCAAGGGGGGGGGGGGGC
> table(width(fastq))                     #配列長$

 19  20  23  25  27  29  30  31  32  33  35
  3   1   4   1   4   2   4   8   1   8 464
>
```

table関数を用いて配列長分布を調べている。トリム前は500リードの配列長はすべて35bpだったが、トリム後に19bp長になっているものが3つ存在する。それを調べる

特定の条件を満たすリードを調べる

R Console

```
> table(width(fastq))                          #配列長ごとの出現頻$

 19  20  23  25  27  29  30  31  32  33  35
  3   1   4   1   4   2   4   8   1   8 464
> obj <- as.logical(width(fastq) == 19)
> sum(obj)
[1] 3
> sread(fastq[obj])
  A DNAStringSet instance of length 3
    width seq
[1]    19 CNNNNNNNNNNNNNTGTGT
[2]    19 CTNNNNNNNNNNTGGGGGG
[3]    19 CAAGGACCCTGGGTTCCCA
> id(fastq[obj])
  A BStringSet instance of length 3
    width seq
[1]    47 SRR037439.23 HWI-E4_6_30ACL:2:1:0:181 length=35
[2]    49 SRR037439.130 HWI-E4_6_30ACL:2:1:2:1099 length=35
[3]    49 SRR037439.199 HWI-E4_6_30ACL:2:1:4:1799 length=35
>
```

配列長が19 bpのものの位置情報を取得し、その数を確認

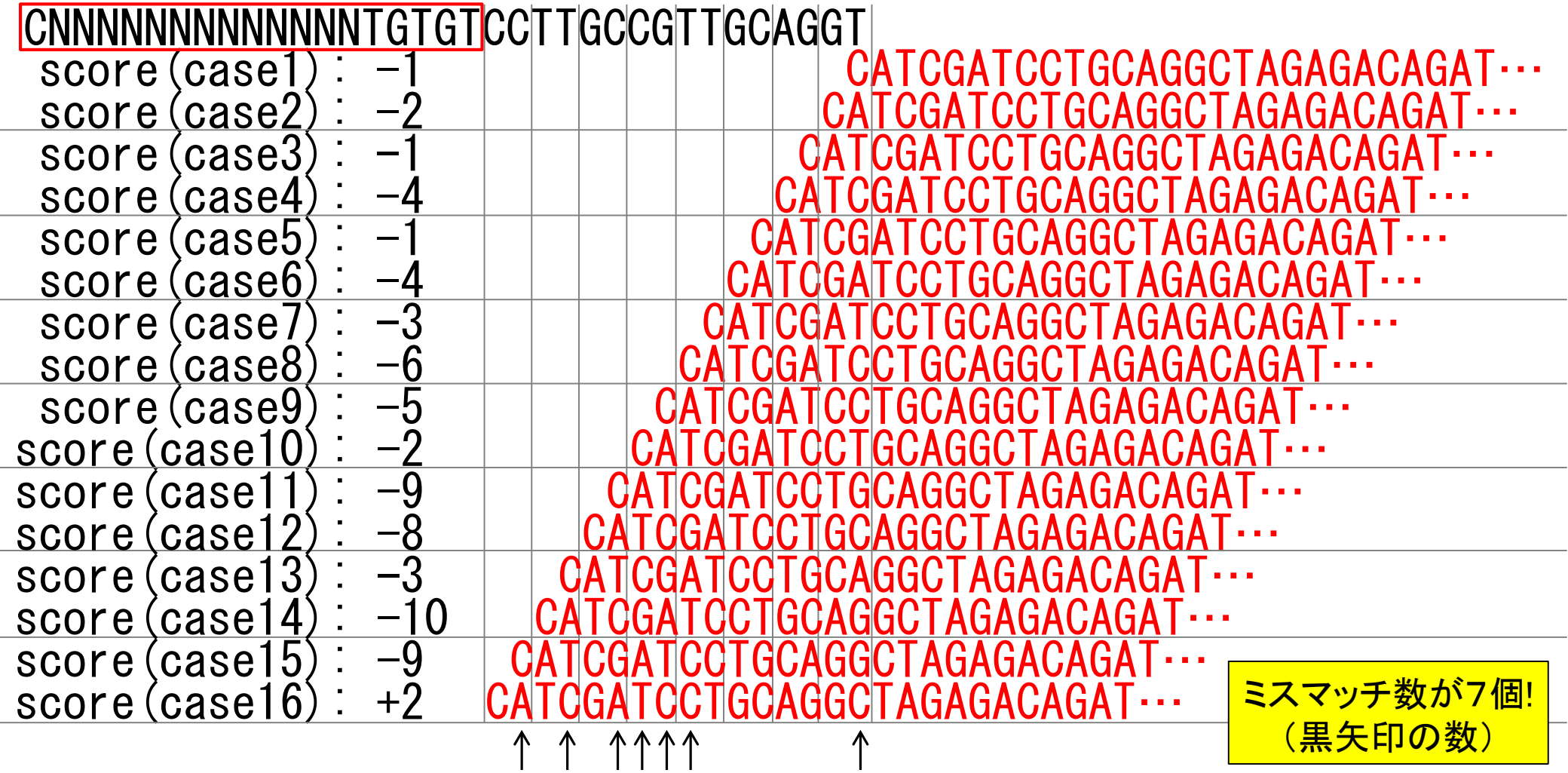
objがTRUEとなる要素のみに対して、塩基配列とdescription情報を表示

最も多くアダプター配列を含むリード IDを特定できた

アダプター配列除去アルゴリズムの詳細を知ることで
girafeのデフォルトパラメータがイマイチであることを知る

アダプター配列除去のイメージ

- 一塩基づつずらしたアラインメントのoverlapの範囲で一致(+1), 不一致(−1)の総和を計算し、最も得点の高かったものを採用している



「?関数名」で詳細な使用法を学ぶ

R Console

```
> ?trimAdapter
starting httpd help server ... done
>
```

Remove 3' adapter contamination

Description

一致に+1、不一致に-1を与えているところ

Function to remove 3' adapter contamination from reads

Usage

```
trimAdapter(fq, adapter, match.score = 1, mismatch.score = -1,
            score.threshold = 2)
```

一致に+1、不一致に-1を与えて、2塩基以上のオーバーラップがないとトリムしない設定にしているのがデフォルト

Arguments

fq	Object of class ShortReadQ; the reads with possible adapter contamination.
adapter	object of class DNAString or class character; the sequence of the 3' adapter which could give rise to the 3' contamination. If of class character, it is converted to a DNAString inside the function.
match.score	numeric; alignment score for matching bases
mismatch.score	numeric; alignment score for mismatches
score.threshold	numeric; minimum total alignment score required for an overlap match between the 3' end of the read and the 5' end of the adapter sequence.

Contents（第2回）

- イントロダクション(Introduction)
  - NGSデータ概観（PacBioとIllumina）
  - NGSデータベース(DB)、データ形式（FASTQ形式）
  - SRAdbパッケージを用いたデータ取得、エラーへの対処
- 前処理(Pre-processing)
  - qrqcパッケージを用いたQuality Control (QC)
- アダプター配列除去
  - 基本戦略（girafeパッケージ）
  - **昔は正常に動作していたのに…という例（QuasRパッケージ）**
  - アダプター除去を含む様々なフィルタリングの組合せ（ShortReadパッケージ）
  - 課題

- 前処理 | トリミング | アダプター配列除去(基礎) | girafe(Toedling 2010)(last modified 2014/06/11) NEW
- 前処理 | トリミング | アダプター配列除去(基礎) | ShortRead(Morgan 200)(last modified 201
- 前処理 | トリミング | アダプター配列除去(応用) | QuasR(Lerch 20XX)(last modified 2014/06/11) NEW
- 前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan 200)(last modified 2014/06/13)推奨
- 前処理 | トリミング | 指定した末端塩基数だけ除去(last modified 2013/06/15)
- アセンブル | について(last modified 2011/07/26)
- アセンブル | ゲノム用(last modified 2014/06/10) NEW
- アセ
- アセ
- マッ

前処理 | トリミング | アダプター配列除去(応用) | QuasR(Lerch_20XX) N

前処理 | トリミング | アダプター配列除去(応用) | QuasR(Lerch 20XX) | QuasR(Lerch 20XX)

2013年11月ごろはうまくいっていましたが、2014年6月に試すとエラーが出ていました、という例

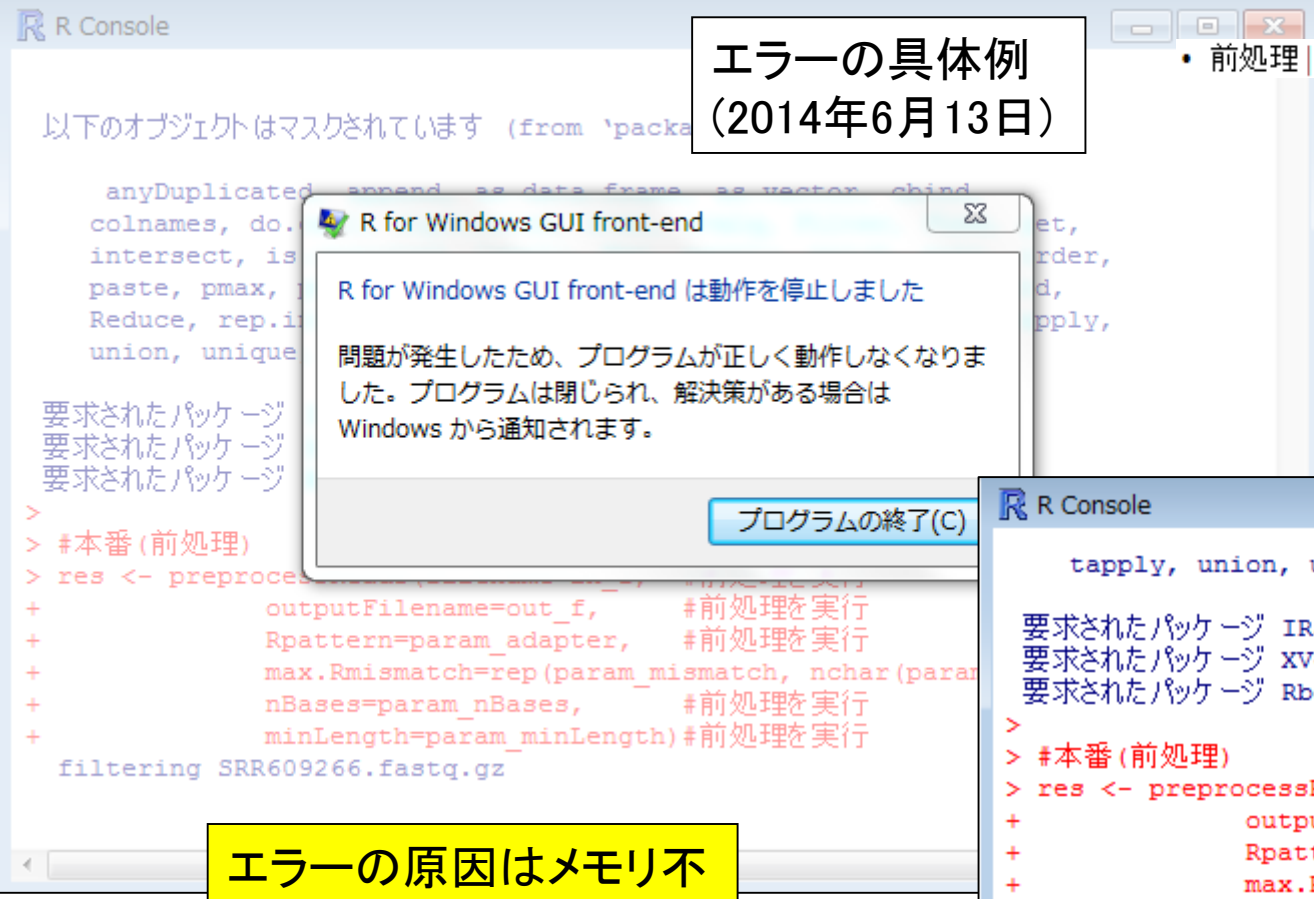
QuasRパッケージを用いたアダプター配列除去とそれに付随する様々な組み合わせのやり方を示します。2013年11月ごろ(R ver. 3.0.2 and Bioconductor 2.13)ではうまくいっていましたが、2014年6月11日(R ver. 3.1.0 and Bioconductor 2.14)で試すと、エラーが発生してR Guiが終了してしまうことを確認済みです。圧縮ファイルのままでアダプター配列除去できるので非常にありがたかったのですが....。誰か対処法が分かったかたはお知らせいただければ幸いです。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. gzip圧縮状態のFASTQ形式ファイル(SRR609266.fastq.gz)の場合:

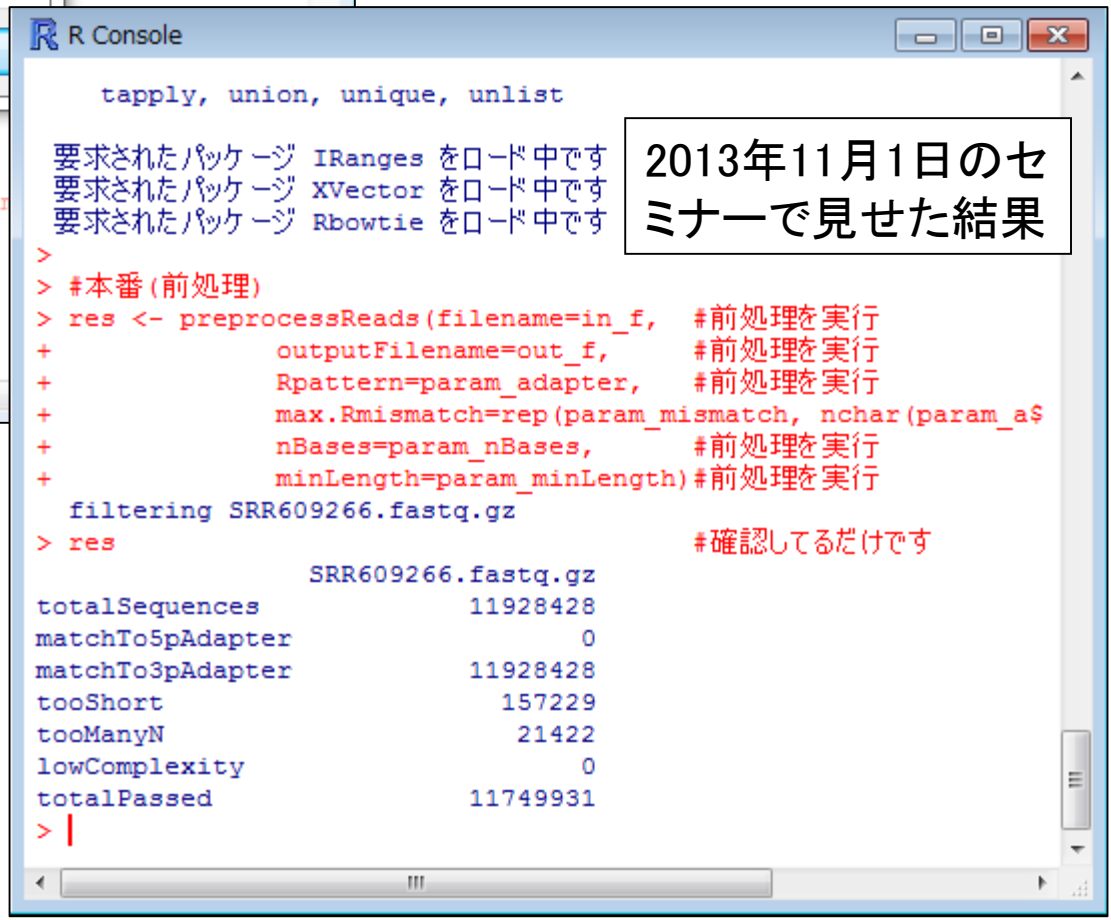
small RNA-seqデータ(ファイルサイズは400Mb弱、11928428リード)です。原著論文(Nie et al., BMC Genomics, 2013)中の記述から GSE41841を頼りに、SRP016842にたどりつき、イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu 2013)の7を実行して得られたものが入力ファイルです。原著論文中では、アダプター配列やクオリティの低いリードを除去したのち、ゲノムにマッピングしたと書いてあります。アダプター配列情報はどこにも書かれていませんでしたが、Table S2中のアダプター配列除去後の最も短いリードが18 nt (例:"GCAGTCGTGGCCGAGCGG")であり、「この18 nt」と「この配列を含む生リード配列の差分」がアダプター配列ということになります。詳細な情報は書かれていませんでしたが、おそらくアダプター配列は"TGGAATTCTCGGGTGCCAAGGAACTCCAGTC..."という感じだろうと推測できます。ここでは、1塩基ミスマッチまで許容して(推定)アダプター配列除去を行ったのち、"ACGT"のみからなる配列(許容するN数が0)で、配列長が18nt以上のものをフィルタリングして出力しています。

```
in_f <- "SRR609266.fastq.gz"            #入力ファイル名を指定してin_fに格納(RNA-seqファイル)
out_f <- "SRR609266_p.fastq.gz"         #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 1                     #許容するミスマッチ数を指定
param_nBases <- 0                       #許容するNの数を指定
param_minLength <- 18                   #アダプター配列除去後の許容する最低配列長を指定

#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(前処理)
res <- preprocessReads(filename=in_f,   #前処理を実行
             outputFilename=out_f,      #前処理を実行
             Rpattern=param_adapter,    #前処理を実行
             max.Rmismatch=rep(param_mismatch, nchar(param_adapter)),#前処理を実行
             nBases=param_nBases,       #前処理を実行
             minLength=param_minLength)#前処理を実行
res                                     #確認してるだけです
```

エラーの具体例
(2014年6月13日)



エラーの原因はメモリ不足だそうです by 孫堅強氏(2014年6月19日)



2013年11月1日のセミナーで見せた結果

アダプター配列除去（推奨のやり方）

- 前処理 | トリミング | アダプター配列除去(基礎) | girafe(Toedling_2010)(last modified 2014/06/11) NEW
- 前処理 | トリミング | アダプター配列除去(基礎) | ShortRead(Morgan_2009)(last modified 2014/06/11)
- 前処理 | トリミング | アダプター配列除去(応用) | QuasR(Lerch_20XX)(last modified 2014/06/11) NEW
- 前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan_2009)(last modified 2014/06/13)
- 前処理 | トリミング | 指定した末端塩基数だけ除去(last modified 2013/06/15)
- アセンブル | に
- アセンブル | ゲノ
- アセンブル | トラ
- アセンブル | ゲノ
- マッピング | 備忘

hogeフォルダ中のファイルを解凍すれば実行できますが見るだけにして

前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan_2009) NEW

ShortReadパッケージを用いたアダプター配列除去とそれに付随する様々な組み合わせのやり方を示します。

「ファイル」ー「デ

1. FASTA形式ファ

アダプター配列既
列(param_nBases

```
in_f <- "samp
out_f <- "hog
param_adapter
param_mismatc
param_nBases
param_range <

#必要なパッケー
library(Short

#入力ファイルの
fasta <- read
fasta

#本番1(アダプタ
fasta <- trim
```

3. FASTQ形式ファイル(SRR609266.fastq)の場合:

small RNA-seqデータ(ファイルサイズは1.8GB、圧縮後で400Mb弱、11928428リード)です。原著論文(Nie et al., BMC Genomics, 2013)中の記述から GSE41841を頼りに、SRP016842にたどりつき、イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu_2013)の7を実行して得られたものが入力ファイルです。

原著論文中では、アダプター配列やクオリティの低いリードを除去したのち、ゲノムにマッピングしたと書いてあります。アダプター配列情報はどこにも書かれていませんでしたが、Table S2中のアダプター配列除去後の最も短いリードが18 nt (例:"GCAGTCGTGGCCGAGCGG")であり、「この18 nt」と「この配列を含む生リード配列の差分」がアダプター配列ということになります。詳細な情報は書かれていませんでしたが、おそらくアダプター配列は"TGGAATTCTCGGGTGCCAAGGAACTCCAGTC..."という感じだろうと推測できます。

アダプター配列既知("TGGAATTCTCGGGTGCCAAGGAACTCCAGTC")で、許容するミスマッチ数が2、ACGTのみからなる配列(param_nBases <- 0)、配列長の範囲指定(20:30)の組み合わせです。

```
in_f <- "SRR609266.fastq"                #入力ファイル名を指定してin_fに格納(RNA-seqファイル)
out_f <- "hoge3.fasta"                   #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 2                      #許容するミスマッチ数を指定
param_nBases <- 0                        #許容するACGT以外の文字数(実質的にはNの許容数に相当)
param_range <- 20:30                     #配列長の範囲を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fastq")#in_fで指定したファイルの読み込み
fasta                                    #確認してるだけです

#本番1(アダプター配列除去)
```

3. FASTQ形式ファイル(SRR609266.fastq)の場合:

small RNA-seqデータ(ファイルサイズは1.8GB、圧縮後で400Mb弱、11928428リード)です。• 前処理|トリミング|アダプター配列除去(応用)|ShortRead(Morgan 2009) Genomics, 2013)中の記述から GSE41841を頼りに、SRP016842にたどりつき、イントロ|NGS|配列取得|FASTQ or SRALite|SRAdb(Zhu 2013)の7を実行して得られたものが入力ファイルです。

原著論文中では、アダプター配列やクオリティの低いリードを除去したのち、ゲノムにマッピングしたと書いてあります。アダプター配列情報はどこにも書かれていませんでしたが、Table S2中のアダプター配列除去後の最も短いリードが18 nt (例:"GCAGTCGTGGCCGAGCGG")であり、「この18 nt」と「この配列を含む生リード配列の差分」がアダプター配列ということになります。詳細な情報は書かれていませんでしたが、おそらくアダプター配列は"TGGAATTCTCGGGTGCCAAGGAACTCCAGTC..."という感じだろうと推測できます。

アダプター配列既知("TGGAATTCTCGGGTGCCAAGGAACTCCAGTC")で、許容するミスマッチ数が2、ACGTのみからなる配列(param_nBases <- 0)、配列長の範囲指定(20:30)の組み合わせです。

```
in_f <- "SRR609266.fastq"                #入力ファイル名を指定してin_fに格納(RNA-seqファイル)
out_f <- "hoge3.fasta"                   #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 2                      #許容するミスマッチ数を指定
param_nBases <- 0                        #許容するACGT以外の文字数(実質的にはNの許容数に相当)
param_range <- 20:30                     #配

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="f
fasta                                    #確

#本番1(アダプター配列除去)
fasta <- trimLRPatterns(Rpattern=param_ad
            subject=fasta,               #T
            max.Rmismatch=rep(param_misma
fasta                                    #確

#本番2(ACGTのみの配列を抽出)
hoge <- rowSums(alphabetFrequency(DNAStri
```

R Console

```
> #入力ファイルの読み込み
> fasta <- readDNAStringSet(in_f, format="fastq")#in_f$
> fasta                                          #確認してるだ$
  A DNAStringSet instance of length 11928428
            width seq                    names
        [1]    49 TCTNCGGT...CGGGTGC SRR609266.1 FCC0N...
        [2]    49 TATGTCTA...CCAAGGA SRR609266.2 FCC0N...
        [3]    49 TCTTCGGT...CGGGTGC SRR609266.3 FCC0N...
        [4]    49 TCTTCGGT...CGGGTGC SRR609266.4 FCC0N...
        [5]    49 TCTTCGGT...CGGGTGC SRR609266.5 FCC0N...
        ...   ... ...
[11928424]    49 NATTATGA...GAATTCT SRR609266.1192842...
[11928425]    49 NCTTCGGT...GGGTGCC SRR609266.1192842...
[11928426]    49 NCAGTTAG...CCAAGGA SRR609266.1192842...
[11928427]    49 NCTTCGGT...CGGGTGC SRR609266.1192842...
[11928428]    49 NCTTCGGT...CGGGTGC SRR609266.1192842...
>
```

全部で11,928,428リード。
配列長は49 bp

```
#本番1(アダプター配列除去)
fasta <- trimLRPatterns(Rpattern=param_adapter,#アダプター配列除去を行った結果をfastaに格納
            subject=fasta,              #アダプター配列除去を行った結果をfastaに格納
            max.Rmismatch=rep(param_mismatch, nchar(param_adapter)))#アダプ
fasta                                   #確認してるだけです

#本番2(ACGTのみの配列を抽出)
hoge <- rowSums(alphabetFr
```

```
R Console

> fasta                                          #確認してるだけです
  A DNAStringSet instance of length 11928428
            width seq                                    names
        [1]    33 TCTNCGGTAGTAT...GTATCCCCGCCT SRR609266.1 FCC0N...
        [2]    27 TATGTCTAAGGAGAATTCAAAAAAGAG  SRR609266.2 FCC0N...
        [3]    33 TCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.3 FCC0N...
        [4]    33 TCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.4 FCC0N...
        [5]    33 TCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.5 FCC0N...
        ...   ... ...
[11928424]    40 NATTATGATGACA...CACTGATTGCAT SRR609266.1192842...
[11928425]    32 NCTTCGGTAGTAT...AGTATCCCCGCC SRR609266.1192842...
[11928426]    27 NCAGTTAGTGGAGGAGTATTCTGGCGT  SRR609266.1192842...
[11928427]    33 NCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.1192842...
[11928428]    33 NCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.1192842...
> table(width(fasta))

   18    19    20    21    22    23    24    25
55342 104978 517762 266238 298548 166321 138027 173525
   26    27    28    29    30    31    32    33
196544 339400 337122 131834  58820  69083 1543235 6517163
   34    35    36    37    38    39    40    41
       783 73928 99457 63429 71625 66676 62746
   43    44    45    46    47
175 37820 46158 145176 63998
>
```

総リード数は不変だが、アダプター配列除去によって配列長にバリエーションができたことがわかる。

最短の18 bpのものが55,342リード、
最長の47 bpのものが63,998リード。

```
#本番2(ACGTのみの配列を抽出)
hoge <- rowSums(alphabetFrequency(DNAStringSet(fasta))[,1:4])#ACGTの総数をカウントした結果
obj <- (width(fasta) - hoge) <= param_nBases#条件を満たすかどうかを判定した結果をobjに格納
fasta <- fasta[obj]                          #objがTRUEとなる要素のみ抽出した結果をfastaに格納
fasta                                        #確認してるだけです

#本番3(指定した長さの範囲の配
obj <- (width(fasta) >= mi
fasta <- fasta[obj]
fasta

#ファイルに保存
writeXStringSet(fasta, fil
```

Nを含むリードがちゃんと消えていることがわかる

R Console

```
> fasta                                          #確認してるだけです
  A DNAStringSet instance of length 11907289
            width seq                                    names
        [1]    27 TATGTCTAAGGAGAATTCAAAAAAGAG  SRR609266.2 FCC0N...
        [2]    33 TCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.3 FCC0N...
        [3]    33 TCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.4 FCC0N...
        [4]    33 TCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.5 FCC0N...
        [5]    22 CCTTGCAAACTGTAAACTCATT        SRR609266.6 FCC0N...
        ...   ... ...
[11907285]    32 GCCGTGATCGTCT...AGGACCCTACGT SRR609266.1192837...
[11907286]    44 TTCTTCACAATTC...GTGCCCAGGAAC SRR609266.1192837...
[11907287]    33 AAGGGAGATATGG...GCGAAGAGCACC SRR609266.1192837...
[11907288]    33 TGCCGTGATCGTC...AGGACCCTACGT SRR609266.1192837...
[11907289]    33 TCTTCGGTAGTAT...GTATCCCCGCCT SRR609266.1192837...
> table(width(fasta))

    18     19     20     21     22     23     24     25
 55264 104821 517002 265848 298040 166022 137801 173238
    26     27     28     29     30     31     32     33
196247 338821 336547 131613  58713  68977 1540653 6505885
    34     35     36     37     38     39     40     41
        669  73800  99268  63313  71490  66552  62624
    43     44     45     46     47
        104  37715  45908 144393  63763
>
```

最短の18 bpのものが55,264リード、
最長の47 bpのものが63,763リード。

```
#本番3(指定した長さの範囲の配列を抽出)
obj <- (width(fasta) >= min(param_range)) & (width(fasta) <= max(param_range))#条件を満た
fasta <- fasta[obj]                        #objがTRUEとなる要素のみ抽出した結果をfastaに格納
fasta                                      #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, fil
```

R Console

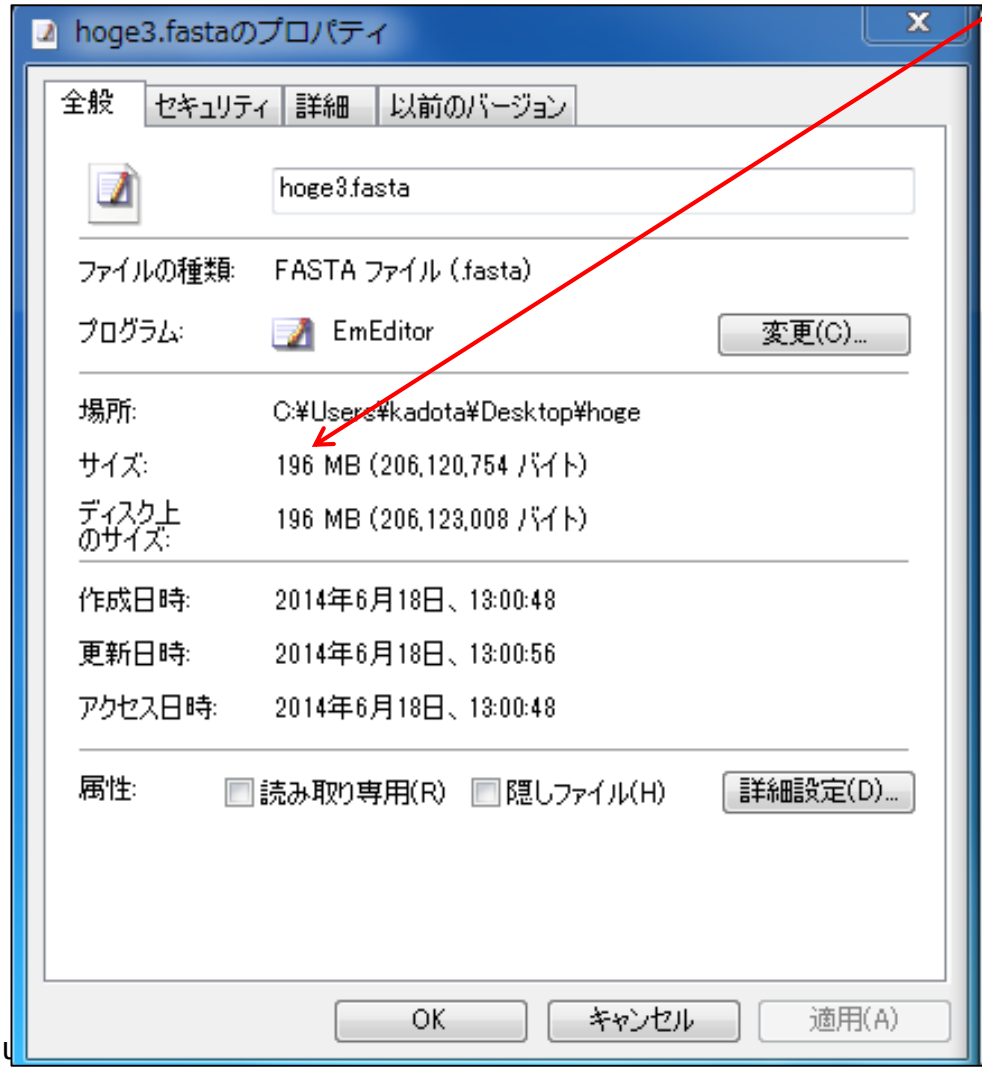
```
> fasta
  A DNAStringSet instance of length 2619892
            width seq                                   names
        [1]    27 TATGTCTAAGGAGAATTCAAAAAAGAG           SRR609266.2 FCC0N...
        [2]    22 CCTTGCAAACTGTAAACTCATT                SRR609266.6 FCC0N...
        [3]    28 TAACGGCAATAAACTCATCATCCAATGC           SRR609266.8 FCC0N...
        [4]    25 TAGAGATGCCTACTGACTAGAACTG              SRR609266.19 FCC0...
        [5]    20 GAAATATTAGGAGGGTTAGG                   SRR609266.22 FCC0...
        ...   ... ...
[2619888]    24 TATATGGTACTGACAATGATTGAT                SRR609266.1192832...
[2619889]    28 TATGTTCGAGTCCTTGTGGAATAATACT            SRR609266.1192833...
[2619890]    28 AAGCTGACATCTGTAGCACTGCCCGGGA            SRR609266.1192834...
[2619891]    26 TAACTCATTTTCGAACACCCAAACTC              SRR609266.1192834...
[2619892]    28 TATTGAACCTAAACAACTATGCACCTGA            SRR609266.1192834...
> table(width(fasta))

    20     21     22     23     24     25     26     27     28
517002 265848 298040 166022 137801 173238 196247 338821 336547
    29     30
131613  58713
> param_range
 [1] 20 21 22 23 24 25 26 27 28 29 30
> min(param_range)
[1] 20
>
```

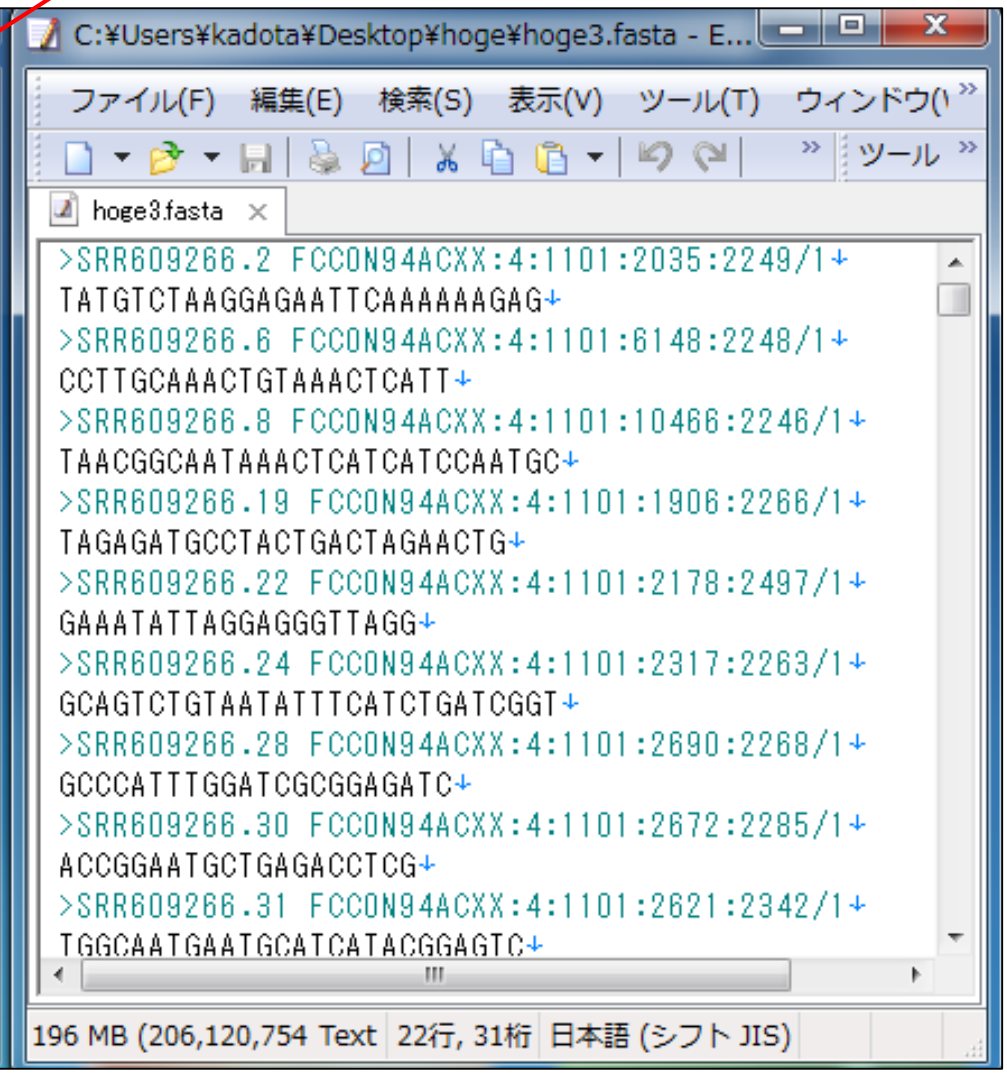
配列長の範囲を20–30 bpに限定すると2,619,892リードに減ることがわかる

最短の20 bpのものが517,002リード、最長の30 bpのものが58,713リード。

```
#本番3(指定した長さの範囲の配列を抽出)
obj <- (width(fasta) >= min(param_range)) & (width(fasta) <= max(param_range))#条件を満た
fasta <- fasta[obj]                          #objがTRUEとなる要素のみ抽出した結
fasta                                        #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50)#fastaの中身
```

出力はFASTA形式にしている。アダプター配列や各種前処理後は、クオリティスコア情報はいらないだろう、という思想。主なメリットはファイルサイズ





アダプター配列除去（推奨のやり方）

前処理 | トリミング | アダプター配列除去(応用) | ShortRead(Morgan_2009) NEW

ShortReadパッケージを用いたアダプター配列除去とそれに付随する様々な組み合わせのやり方を示します。

「ファイル」ー「デ

1. FASTA形式ファ

アダプター配列既
列(param_nBases

```
in_f <- "samp
out_f <- "hog
param_adapter
param_mismatc
param_nBases
param_range <

#必要なパッケー
library(Short

#入力ファイルの
fasta <- read
fasta

#本番1(アダプタ
fasta <- trim
```

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

small RNA-seqデータ(400Mb弱、11928428リード)です。圧縮ファイルもreadDNAStringSet関数で通常手順で読み込めます。原著論文(Nie et al., BMC Genomics, 2013)中の記述から GSE41841を
口 | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu 2013)の7を実行して得
原著論文中では、アダプター配列やクオリティの低いリードを除去したのち、ゲ
アダプター配列情報はどこにも書かれていませんでしたが、Table S2中のアダ
nt (例:"GCAGTCGTGGCCGAGCGG")であり、「この18 nt」と「この配列を含む
ということになります。詳細な情報は書かれていませんでしたが、おそらくアダプター配列
は"TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"という感じだろうと推測できます。
アダプター配列既知("TGGAATTCTCGGGTGCCAAGGAACTCCAGTC")で、許容するミスマッチ数が2、ACGTのみからなる配列(param_nBases <- 0)、配列長の範囲指定(20:30)の組み合わせです。

```
in_f <- "SRR609266.fastq.gz"          #入力ファイル名を指定してin_fに格納(RNA-seqファイル)
out_f <- "hoge4.fasta.gz"             #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
para
para
para

#必要
libr

#入力
fasta
```

```
#本番2(ACGTのみの配列を抽出)
hoge <- rowSums(alphabetFrequency(DNAStringSet(fasta))[,1:4])#ACGTの総数をカウントした結果
obj <- (width(fasta) - hoge) <= param_nBases#条件を満たすかどうかを判定した結果をobjに格納
fasta <- fasta[obj]                   #objがTRUEとなる要素のみ抽出した結果をfastaに格納
fasta                                 #確認してるだけです

#本番3(指定した長さの範囲の配列を抽出)
obj <- (width(fasta) >= min(param_range)) & (width(fasta) <= max(param_range))#条件を満た
fasta <- fasta[obj]                   #objがTRUEとなる要素のみ抽出した結果をfastaに格納
fasta                                 #確認してるだけです

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50, compress=T)#fastaの中身を指定
```

readDNAStringSet関数はgzip圧縮ファイルも読み込み可能。gzip圧縮ファイルとして保存することも可能

Tips

4. FASTQ形式ファイル(SRR609266.fastq.gz)

small RNA-seqデータ(400Mb弱、11928428リ
ます。原著論文(Nie et al., BMC Genomics,
口 | NGS | 配列取得 | FASTQ or SRALite | S
原著論文中では、アダプター配列やクオリテ
アダプター配列情報はどこにも書かれていま
nt (例:"GCAGTCGTGGCCGAGCGG")であ
ということになります。詳細な情報は書かれて
は"TGGAATTCTCGGGTGCCAAGGAACT
アダプター配列既知("TGGAATTCTCGGGT
からなる配列(param_nBases <- 0)、配列長の

```
in_f <- "SRR609266.fastq.gz"
out_f <- "hoge4.fasta.gz"
param_adapter <- "TGGAATTCTCGGGTG
param_mismatch <- 2
param_nBases <- 0
param_range <- 20:30

#必要なパッケージをロード
library(ShortRead)

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, f
```

rcode_adapter.txt

```
in_f <- "SRR609266.fastq.gz"          #入力ファイル名を指定してin_fに格納(RNA
out_f <- "hoge4.fasta.gz"             #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 2                   #許容するミスマッチ数を指定
#param_nBases <- 0                     #許容するACGT以外の文字数(実質的にはNの
param_range <- 20:30                  #配列長の範囲を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format="fastq")
fasta                                 #確認し

#本番1(アダプター配列除去)
fasta <- trimLRPatterns(Rpattern=param_adapter,#アダプター配列除去を行った結果
            subject=fasta,            #アダプター配列除去を行った結果をfastaに
            max.Rmismatch=rep(param_mismatch, nchar(param_adapter)))#アダプター
fasta                                 #確認してるだけです

#本番2(ACGTのみの配列を抽出)
#hoge <- rowSums(alphabetFrequency(DNAStringSet(fasta))[,1:4])#ACGTの総数をカウ
#obj <- (width(fasta) - hoge) <= param_nBases#条件を満たすかどうかを判定した結
#fasta <- fasta[obj]                   #objがTRUEとなる要素のみ抽出した結果を
#fasta                                 #確認してるだけです

#本番3(指定した長さの範囲の配列を抽出)
obj <- (width(fasta) >= min(param_range)) & (width(fasta) <= max(param_range))
fasta <- fasta[obj]                   #objがTRUEとなる要素のみ抽出した結果をf
fasta                                 #確認してるだけです
table(width(fasta))                   #配列長ごとの出現頻度を表示

#ファイルに保存
writeXStringSet(fasta, file=out_f, format="fasta", width=50, compress=T)#fasta
```

Nを一つでも含むリードの除去を行うステップを省く場合。(#を左端に入れれば、そのコマンドは実行されない)

課題1と2

- 原著論文中のアダプター配列除去後の配列長分布は右表のとおりであった。

Nie et al., *BMC Genomics*, **14**: 661, 2013

Table S1

Length	Total
18nt	55702
19nt	106074
20nt	521188
21nt	267851
22nt	300121
23nt	167156
24nt	138656
25nt	174267
26nt	197271
27nt	339876
28nt	337492
29nt	132016
30nt	58901
31nt	69938
32nt	1551638
33nt	6529060
34nt	141067
35nt	72969
36nt	73926
37nt	99344
38nt	63419
39nt	71469
40nt	66514
41nt	61526
42nt	37604
43nt	27361
44nt	29035
all	11691441

rcode_adapter.txt(の一部)

```
in_f <- "SRR609266.fastq.gz"              #入力ファイル名を指定してin_fに格納(RNA
out_f <- "hoge4.fasta.gz"                 #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 2                       #許容するミスマッチ数を指定
#param_nBases <- 0                         #許容するACGT以外の文字数(実質的にはNの
param_range <- 20:30                      #配列長の範囲を指定

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

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, form
fasta

#本番1(アダプター配列除去)
```

1. 右表と同じように18-44塩基の範囲内にあるsmall RNAリードのみを抽出するためにはどこをどう変更すればよいか示せ。

2. 指定した範囲に含まれる総リード数を示せ。

課題3

Nie et al., *BMC Genomics*, **14**: 661, 2013

- 右表に示されているように、原著論文中のアダプター配列除去を含むフィルタリング後の総リード数は11,691,441個であった。以下に様々な条件で得られた総リード数を示す。

条件1-1(許容するミスマッチ数=1; Nを含んでもよい) : 11,619,415個

```
param_mismatch <- 1          #許容するミスマッチ数を指定
#param_nBases <- 0           #許容するACGT以外の文字数(実質的にはNの許容数に相当)を指定
param_range <- 18:44         #配列長の範囲を指定
```

条件1-2(許容するミスマッチ数=1; Nを全く含まない) : 11,599,894個

```
param_mismatch <- 1          #許容するミスマッチ数を指定
param_nBases <- 0            #許容するACGT以外の文字数(実質的にはNの許容数に相当)を指定
param_range <- 18:44         #配列長の範囲を指定
```

条件2-1(許容するミスマッチ数=0; Nを含んでもよい) : 11,357,039個

```
param_mismatch <- 0          #許容するミスマッチ数を指定
#param_nBases <- 0           #許容するACGT以外の文字数(実質的にはNの許容数に相当)を指定
param_range <- 18:44         #配列長の範囲を指定
```

条件2-2(許容するミスマッチ数=0; Nを全く含まない) : 11,338,479個

```
param_mismatch <- 0          #許容するミスマッチ数を指定
param_nBases <- 0            #許容するACGT以外の文字数(実質的にはNの許容数に相当)を指定
param_range <- 18:44         #配列長の範囲を指定
```

自分でもいくつか試し、結果を簡単に考察せよ。原著論文も明確に条件を記述しているわけではないので細かな違いは気にしなくてよい。

Table S1

Length	Total
18nt	55702
19nt	106074
20nt	521188
21nt	267851
22nt	300121
23nt	167156
24nt	138656
25nt	174267
26nt	197271
27nt	339876
28nt	337492
29nt	132016
30nt	58901
31nt	69938
32nt	1551638
33nt	6529060
34nt	141067
35nt	72969
36nt	73926
37nt	99344
38nt	63419
39nt	71469
40nt	66514
41nt	61526
42nt	37604
43nt	27361
44nt	29035
all	11691441

課題遂行時に何人か遭遇したエラーの解説

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

small RNA-seqデータ(400Mb弱、11928428リード)です。圧縮ファイルもreadDNAStringSet関数で通常手順で読み込めます。原著論文(Nie et al., BMC Genomics, 2013)中の記述から GSE41841を頼りに、SRP016842にたどりつき、イントロ | NGS | 配列取得 | FASTQ or SRALite | SRAdb(Zhu 2013)の7を実行して得られたものが入力ファ
原著論文中では、アダプター配列やクオリティの低いリードを除去したのち、ゲノムにマッピングしたと
アダプター配列情報はどこにも書かれていませんでしたが、Table S2中のアダプター配列除去後の最
nt (例:"GCAGTCGTGGCCGAGCGG")であり、「この18 nt」と「この配列を含む生リード配列の差分」
ということになります。詳細な情報は書かれていませんでしたが、おそらくアダプター配列
は"TGGAATTCTCGGGTGCCAAGGAACTCCAGTC..."という感じだろうと推測できます。
アダプター配列既知("TGGAATTCTCGGGTGCCAAGGAACTCCAGTC")で、許容するミスマッチ数が2、ACGTのみからなる配列(param_nBases <- 0)、配列長の範囲指定(20:30)の組み合わせです。

```
in_f <- "SRR609266.fastq.gz"              #入力ファイル名を指定してin_fに格納(RNA-seqファイル)
out_f <- "hoge4.fasta.gz"                 #出力ファイル名を指定してout_fに格納
param_adapter <- "TGGAATTCTCGGGTGCCAAGGAACTCCAGTC"#アダプター配列を指定
param_mismatch <- 2                       #許容するミスマッチ数を指定
param_nBases <- 0                         #許容するACGT以外の文字数(実質的にはNの許容数に相当)を
param_range <- 20:30                      #配列長の範囲を指定

#必要なパッケージをロード
library(ShortRead)

#入力ファイルの読み込み
fasta <- readDNAStringSet(in_f, format
```

Macだとメモリ云々の問題に関わらず、gzファイルのままでは読み込めないようです by 受講生

何人かの方が、作業ディレクトリの変更も正しく行い、SRR609266.fastq.gzファイルもhogeフォルダ中に存在するにも関わらず、入力ファイル読み込み時にエラーに遭遇しました。この理由は2つ考えられます。1つめは、USBメモリにコピーする際に正しくコピーできていなかった可能性、そして2つめはUSBメモリ中のSRR609266.fastq.gzファイル段階では正しいものであったが、各自のPCにコピーする際に正しくコピーできなかった可能性です。講義中に述べたMD5チェックサム(MD5 check sum)でファイルの同一性を確認するのは重要…ですね。