

次世代シーケンサーデータの解析手法 第7回ロングリードアセンブリ: ウェブ資料

谷澤靖洋、神沼英里*、中村保一、遠野雅徳、大崎 研、
清水謙多郎、門田 幸二*

*東京大学・大学院農学生命科学研究科

kadota@bi.a.u-tokyo.ac.jp

<http://www.iu.a.u-tokyo.ac.jp/~kadota/>

W1-1 : PacBio RS II

YouTubeで見られるPacBioのSMART Cell sequencingの原理説明番組

The screenshot shows a YouTube video player for the video 'Introduction to SMRT Sequencing' by Pacific Biosciences. The video thumbnail displays a SMRT Cell. The video title is 'Introduction to SMRT Sequencing' and it has 41,211 views. The channel name is 'Pacific Biosciences' with 713 subscribers. The video player interface includes a search bar, a language selection dropdown set to Japanese, and a list of recommended videos on the right side.

YouTube JP

言語を選択してください。

YouTube を日本語でご覧いただいています。この設定は下で変更できます。

詳細
View this message in English

次の動画 自動再生

- Single Molecule Real Time Sequencing - Pacific Biosciences**
haemse
視聴回数 82,805 回
4:05
- Next-Generation Sequencing Technologies - Elaine Mardis**
GenomeTV
視聴回数 67,502 回
1:34:36
- 1) Next Generation Sequencing (NGS) - An Introduction**
Applied Biological Materials - abm
視聴回数 64,540 回
9:30
- PacBio AGBT 2015 Live Workshop**
PacificBiosciences
視聴回数 2,922 回
2:04:50
- Reveal Hidden Variation—C1™ Single-Cell DNA Sequencing**
fluidigm
視聴回数 2,134 回
58:27
- 1st, 2nd, and 3rd Generation Genome Sequencing**
ImGenTechWPI
視聴回数 51,325 回
14:40

W2-1: PacBioデータ

DRR024500は(連載第6回のW2-2作成当時
は見られたが)2016年3月頃から見られない

DRASearch [Send Feedback](#) [Search Home](#) [DRA Home](#)

DRR024500 [FASTQ](#) [SRA](#)

Run Detail	
Alias	
Instrument model	
Date of run	
Run center	
Number of spots	
Number of bases	

Navigation

READS (joined) quality show 10 rows << < 1 / 0 Page > >>

Website policy | © DNA Data Bank of Japan

W2-2: 乳酸菌データ

原著論文(PMID: 25879859)のPubMedの①Full textリンク先で全文を見られる。②Availability of supporting dataという項目をよく眺めると、NGS生データがDDBJ Sequence Read Archive (DDBJ SRA; 略してDRA)にDRR024500とDRR024501というIDで登録されていることがわかる。③Illumina MiSeqデータのDRR024501を頼りに調べていく

Abstract

BMC Genomics. 2015 Mar 25;16:240. doi: 10.1186/s12864-015-1435-2.

Complete genome sequence and analysis of *Lactobacillus hokkaidonensis* LOOC260(T), a psychrotrophic lactic acid bacterium isolated from silage.

Tanizawa Y^{1,2}, Tohno M³, Kaminuma E⁴, Nakamura Y⁵, Arita M^{6,7}.

Author information

Abstract

BACKGROUND: *Lactobacillus hokkaidonensis* is an obligate heterofermentative lactic acid bacterium, which is isolated from Timothy grass silage in Hokkaido, a subarctic region of Japan. This bacterium is expected to be useful as a silage starter culture in cold regions because of its remarkable psychrotolerance; it can grow at temperatures as low as 4°C. To elucidate its genetic background, particularly in relation to the source of psychrotolerance, we constructed the complete genome sequence of *L. hokkaidonensis* LOOC260(T) using PacBio single-molecule real-time sequencing technology.

RESULTS: The genome of LOOC260(T) comprises one circular chromosome (2.28 Mbp) and two circular plasmids: pLOOC260-1 (81.6 kbp) and pLOOC260-2 (41.0 kbp). We identified diverse genetic elements, such as prophages, integrated and conjugative elements, and conjugative elements which may reflect adaptation to plant-associated niches. Comparative genome analysis also revealed unique genomic features, such as genes involved in pentose assimilation and NADPH generation.

CONCLUSIONS: This is the first complete genome in the *L. vaccinostercus* group, which is characterized, so the genomic information obtained in this study provides insight into the evolution of this group. We also found several factors that may contribute to the ability of *L. hokkaidonensis* to grow at cold temperatures. The results of this study will facilitate future research for the cold-tolerance mechanism of *L. hokkaidonensis*.

PMID: 25879859 [PubMed - in process] PMCID: PMC4377027 Free PMC Article

Full text links

Read free full text at BioMed Central

PMC Full text

Save items

Add to Favorites

Similar articles

Availability of supporting data

The complete genome sequence of *L. hokkaidonensis* LOOC260^T and its annotations were deposited at DDBJ/ENA/GenBank under accession numbers AP014680 (chromosome), AP014681 (plasmid pLOOC260-1), and AP014682 (plasmid pLOOC260-2). All of the sequencing data were deposited in the DDBJ Sequence Read Archive under accession numbers DRR024500 and DRR024501. The phylogenetic tree and associated data matrix for in Additional file 1: Figure S2 are available in TreeBASE database (Accession URL: <http://purl.org/phylo/treebase/phyloWS/study/TB2:S17206>).

Related information

Tanizawa et al., *BMC Genomics*, 16: 240, 2015

W2-3: 乳酸菌データ

①DRR024501の上位階層である②DRP002401をクリックすると、PacBioの新しいDRR IDに辿れる

The screenshot shows the DRA Search interface. At the top, the URL is <https://trace.ddbj.nig.ac.jp/DRAsearch/run?acc=DRR024501>. The page title is "DRASearch" and it includes links for "Send Feedback", "Search Home", and "DRA Home".

The main content area displays "DRR024501" with download options for "FASTQ" and "SRA". Below this is a "Run Detail" table:

Run Detail	
Alias	DRR024501
Instrument model	
Date of run	
Run center	
Number of spots	2,971,310
Number of bases	1,491,597,620

To the right of the table is a "Navigation" menu with the following items:

- Submission [DRA002643](#) (FASTQ, SRA)
- Study [DRP002401](#) (highlighted with a red arrow and circled '2')
- Experiment [DRX022186](#) (FASTQ, SRA)

Below the navigation menu is a "READS (joined)" section with a "quality" checkbox, a "show" dropdown set to "10", and "rows" navigation. The first read is shown as:

```
>DRR024501.1
ATGNATCGAAACAGTATTTACAAGATTTGCATACTGAAATTGAAGCTGATCAACACGAAACCATTCCAGCCGGCAAGGGT
AATCTAAACCAACCCATTAGCTGTTATTGAAGCTTTGCAGCAACGAGTTGATGATAAAATGACCGTTTCGGTTGATGTGGG
GAGCCATTATATTTGGATGGCCCGGCACCTCCGAAGTTATGAGCCTCGCCATTTATTGTTTAGTAATGGGATGCAGACGC
TTGGAGTGGCGATGAACCGTATTAAGGCCTAAACGAACGGCTGTCTCCAGTTCCTGTCCAGTAAATAAGAATCCGGCATC
CCCAGAAACAGAGACTGATTTAGCATTGGGCCGAACCTAACCGAGCCGAAATTGACCAAGGTAGCGCCACTCCAAGCGTCT
GCATCCCATTACTAAACAATAAATGGCGAGGCTCATAACTTCGGAAGTGCCGGGCCATCCAAATATAATGGCTCCCCACA
```

W2-3: 乳酸菌データ

①DRP002401。②DRX022185から、③新しいPacBioのDRR ID (DRR054113-054116)に辿れる

DRASearch [Send Feedback](#) [Search Home](#) [DRA Home](#)

DRP002401

Study Detail

Title	
Study Type	
Abstract	
Description	
Center Name	

Navigation

- Submission [DRA002643](#) [FTP](#)
- Experiment [DRX022185](#) [FASTQ](#) [SRA](#)
- DRX022186 [FASTQ](#) [SRA](#)

DRASearch [Send Feedback](#) [Search Home](#) [DRA Home](#)

DRX022185

[FASTQ](#) [SRA](#)

Experiment Detail

Title	Whole genome sequencing of Lactobacillus hokkaidonensis: Lactobacillus hokkaidonensis LOOC260
Design Description	
Organism	

Library Description

Name	LH_LOOC260_lib1
Strategy	WGS
Source	GENOMIC

Navigation

- Submission [DRA002643](#) [FTP](#)
- Study [DRP002401](#)
- Sample [DRS016698](#)
- Run
 - [DRR054113](#) [FASTQ](#) [SRA](#)
 - [DRR054114](#) [FASTQ](#) [SRA](#)
 - [DRR054115](#) [FASTQ](#) [SRA](#)
 - [DRR054116](#) [FASTQ](#) [SRA](#)

W2-4: 乳酸菌データ

①DRR024501の上位階層である②DRA002643をクリックすると、PacBioの新しいDRR IDに辿れる

The screenshot shows the DRA Search interface. At the top, the URL is <https://trace.ddbj.nig.ac.jp/DRASearch/run?acc=DRR024501>. The page title is "DRR024501" with links for FASTQ and SRA. A red arrow labeled "1" points to the "Run Detail" table.

Run Detail	
Alias	DRR024501
Instrument model	
Date of run	
Run center	
Number of spots	2,971,310
Number of bases	1,491,597,620

On the right, the "Navigation" menu is visible, with a red arrow labeled "2" pointing to the "Submission DRA002643" link.

Below the navigation menu, the "READS (joined)" section shows a sequence viewer for run DRR024501.1. The sequence is displayed in a color-coded format (ATCG) and includes a "quality" checkbox, a "show" dropdown set to 10, and a "rows" dropdown set to 1. The page number is 1 / 297131.

```
>DRR024501.1
ATGNATCGAAACAGTATTTACAAGATTTGCATACTGAAATTGAAGCTGATCAACACGAAACCATTCCAGCCGGCAAGGGT
AATCTAAACCACCCATTAGCTGTTATTGAAGCTTTGCAGCAACGAGTTGATGATAAAATGACCGTTTCGGTTGATGTGGG
GAGCCATTATATTTGGATGGCCCGGCACCTCCGAAGTTATGAGCCTCGCCATTTATTGTTTAGTAATGGGATGCAGACGC
TTGGAGTGGCGATGAACCGTATTAAGGCCTAAACGAACGGCTGTCTCCAGTTCCTTGTCCAGTAAATAAGAATCCGGCATC
CCCAGAAACAGAGACTGATTTAGCATTGGGCCGAACCTAACCGAGCCGAAATTGACCAAGGTAGCGCCACTCCAAGCGTCT
GCATCCCATTACTAAACAATAAATGGCGAGGCTCATAACTTCGGAAGTGCCGGGCCATCCAAATATAATGGCTCCCCACA
```

W2-4: 乳酸菌データ

①DRA002643でも、②新しいPacBioのDRR ID (DRR054113-054116)に辿れる

The screenshot shows the DRA Search interface. The main heading is "DRA002643" with an FTP icon. Below it is a "Submission Detail" table. To the right is a "Navigation" menu. A red arrow labeled "1" points to the "DRA002643" heading. A red arrow labeled "2" points to a red-bordered box in the "Run" section of the navigation menu, which contains the DRR IDs DRR054113, DRR054114, DRR054115, and DRR054116, each with "FASTQ" and "SRA" links.

Submission Detail	
Alias	DRA002643
Submission ID	
Submission Date	2014-11-07
Center Name	NILGS
Lab Name	Animal Feeding and Management Research Division

Navigation	
Study	DRP002401
Experiment	DRX022185 FASTQ SRA
	DRX022186 FASTQ SRA
Sample	DRS016698
Run	DRR024501 FASTQ SRA
	DRR054113 FASTQ SRA
	DRR054114 FASTQ SRA
	DRR054115 FASTQ SRA
	DRR054116 FASTQ SRA

W2-5: PacBio概観

①各DRR IDをクリックして、PacBioデータを眺める

The screenshot shows the DRA Search web interface. The browser address bar displays the URL: <https://trace.ddbj.nig.ac.jp/DRAsearch/submission?acc=DRA002643>. The page title is "DRA Search" and the submission ID is "DRA002643".

Submission Detail

Alias	DRA002643
Submission ID	
Submission Date	2014-11-07
Center Name	NILGS
Lab Name	Animal Feeding and Management Research Division

Navigation

- Study: [DRP002401](#)
- Experiment: [DRX022185](#) (FASTQ, SRA)
- [DRX022186](#) (FASTQ, SRA)
- Sample: [DRS016698](#)
- Run: [DRR024501](#) (FASTQ, SRA)
- [DRR054113](#) (FASTQ, SRA)
- [DRR054114](#) (FASTQ, SRA)
- [DRR054115](#) (FASTQ, SRA)
- [DRR054116](#) (FASTQ, SRA)

A red box highlights the DRR054113 ID in the Run section, and a red arrow points to it from below.

W2-5: DRR054113

DRASearch [Send Feedback](#) [Search Home](#) [DRA Home](#)

DRR054113 [FASTQ](#) [SRA](#)

Run Detail	
Alias	DRR054113
Instrument model	
Date of run	
Run center	
Number of spots	163,482
Number of bases	360,244,590

Navigation

- Submission [DRA002643](#) [FTP](#)
- Study [DRP002401](#)
- Experiment [DRX022185](#) [FASTQ](#) [SF](#)

READS (joined) quality show 10 rows << < 1 / 16349 Page > >>

```
>DRR054113.1
CCTATGCTGTCAGCATTGATTGCTAGTTGATGGTTCTATATTTACGTATCACATTGAGATATATCGCTCATCAGCTTCT
GCTACTAGTCTTGAGTCTGCCTGACTGATTGATTGATCATGCGTGGATTCTGATGATACTTTATGCATTATACGAGTTAC
GTACGGCAGGCATGTAGTACTGGTGCATGACCTCATGAGCTAGCATTGAGTTATCGTGATCCATAACTGGATCAGTACTT
GCAGGTCATGATACCAGGTCGATTTCAGTATTCGATGTCTAGACTTAGCTGACATAGCAGATTGATCTTCTTGATTACAGG
CGATATCGCACTGCGTCATACGATTCACAGTCA

>DRR054113.2
TCATATACTCGGCACAATGTGTGTCGATCGTAAAGGGATGTCATTGTGTAGTATTGTATTCTATATGTCGAGCATCAGCG
TTCTACTGCTGAGATGATATATTCTGAGTATTATGGTTATGTATTTTACGTGAACCTGGATTATGTCGTGGACGGACGT
TGTACGGATTTCTAACTGTTAGTATCGAGCATTGATCGTTCGATGGATTGATAGTGCTTCCGTTGAGTCGTAATGATTGTT
CAGTTAGTCGATCAGTTTCGTGGATCAGTGATTTTGTAGGCGAAGATTATGAATCTTTACGATCCTTATGGCTGAGTTGAA
TGATCTGCTGCTAGTGTCTGTATGTTTCGTATGATCACATGACGATACGTGATATTTATTATTGTCTACGCATCGATTGAG
```

W2-5: DRR054114

DRASearch [Send Feedback](#) [Search Home](#) [DRA Home](#)

DRR054114 [FASTQ](#) [SRA](#)

Run Detail	
Alias	DRR054114
Instrument model	
Date of run	
Run center	
Number of spots	163,482
Number of bases	353,390,616

Navigation

- Submission [DRA002643](#) [FTP](#)
- Study [DRP002401](#)
- Experiment [DRX022185](#) [FASTQ](#) [SF](#)

READS (joined) quality show 10 rows << < 1 / 16349 Page > >>

```
>DRR054114.1
GTAGATACGTATGGGTGCGAGACGCTATCATTGAGCTGTCATGGCGTTTCAGGTGGTTGCATCAGCAATTAGGTGTGTCTT
GGATTTTGAAGAGAGTTTCATTTCGCTGGAGTCTCGACGAGAGCCGGCGACTCTTATTAGTACGTCAGGACGTGATATGATC
ACATTACGCAGTACTGAGCGATATGTTTATCACAGTCGAGTTCTGTTTATGTTATGTTGAGTAGTTATTGTAGTGCGCTT
CATTTCGCTCTAACGGATGGACAGCTATTGGCTTCGGAATTTTATGTCAGTTCCACATTTACGTAGTTTGGAAATTTTCGTG
GTGTTTA

>DRR054114.2
GTTTCGAAAGCATCTCCACACACGTTACCTCTGGTCACTCCTTCGCGGTACGCCGAACCCCTCTCCACAGCCACCAGAAGCA
CACGTGTGCGGTAGCAAGCCCACCAGAGCGCGCCCTCCGATACGCCACAGAAGAGCCCCCTAGTTTCTTCTTCGGCAGTAC
CACTATCCTCTCACAGAGGACTCCTCCACCACCGCCACTGCTCTCAAGACCCCTAGTCCCCTGCGGCTCGGTCCACTCCT
AGCCCAATATCCAACGGCCTTCGCTGCTCGCGCCACACACCCCTTACCATCGCCGCCACGGAACAAGACGATCCTAGCCCT
GCCTGGTTCGTGCGCGTCAACAGAACACCTCGTCCGCCACGTACGCCGCACTCCTCAGCCAGCACCCTGGAACCCCGCC
```

W2-5: DRR054115

DRASearch [Send Feedback](#) [Search Home](#) [DRA Home](#)

DRR054115 [FASTQ](#) [SRA](#)

Run Detail	
Alias	DRR054115
Instrument model	
Date of run	
Run center	
Number of spots	163,482
Number of bases	376,482,867

Navigation

- Submission [DRA002643](#) [FTP](#)
- Study [DRP002401](#)
- Experiment [DRX022185](#) [FASTQ](#) [SF](#)

READS (joined) quality show 10 rows << < 1 / 16349 Page > >>

```
>DRR054115.1
ATCGGATTATTAGTATCGATGCCAGTTGACATAGTCGTTGTTGTGACGTGCTTAGGGGGTCAGGGGATGTATGAATGAAG
AGGAGATCGCATCGACATGTGTCCGACTAGATCCGAGCATGGTGATTATCGAGTCCTTGTATATCGAACGACGTGTAG
CATATAGTATGCCTAATATTATTATCAGCTAATTATTGTGCGATTATACGGTATGTCACTCAGCCATGATTGTCATCATC
AGTCTCTATCGTGGTATATTCATTCCGATCAGCATGTATGATGATATACGTAGCTGTCATAGTAGATAGTATGTCATTGC
ATACGTGAACGACTGATTAGCAG

>DRR054115.2
TGTCGATACAGGTATAGTCATAGCATTGATTTTGTAGTACGACGAAGACGTGGATACGGTGCATCCTGACTTTCTCGTATT
TACGTTTATATAGTTGATTTTGGATGCTGTATATGATGATGCCTCCTGACTAATATCAGCACTGCTGAGGTCGTGATATT
AAGTACTACATGTGATGCTGTATGCACAGTTGTCTGTTATGCGATTATGATGAGGATAGTCGCTTGGTTATGATATTATC
TGTGGTTCAGTGGCATTCTTTTATTACTTGATCCTGTCCGGAGGACGTGATTTTCTTCGTCCTCGCAGTCCATGTATAT
GAGCATAGTGGACAGGATTCGTTCTTTGACACGCTGCTTGTAGCGTGGTGGAGCTCTTGATGTTTCAGGATCGAGATTCCA
```

W2-5: DRR054116

① DRR054116中の総リード数は163,482
。4 SMRT Cells由来の全DRR IDsのリード数が同じになっている

DRASearch [Send Feedback](#) [Search Home](#) [DRA Home](#)

DRR054116 [FASTQ](#) [SRA](#)

Run Detail	
Alias	DRR054116
Instrument model	
Date of run	
Run center	
Number of spots	163,482
Number of bases	532,802,277

Navigation

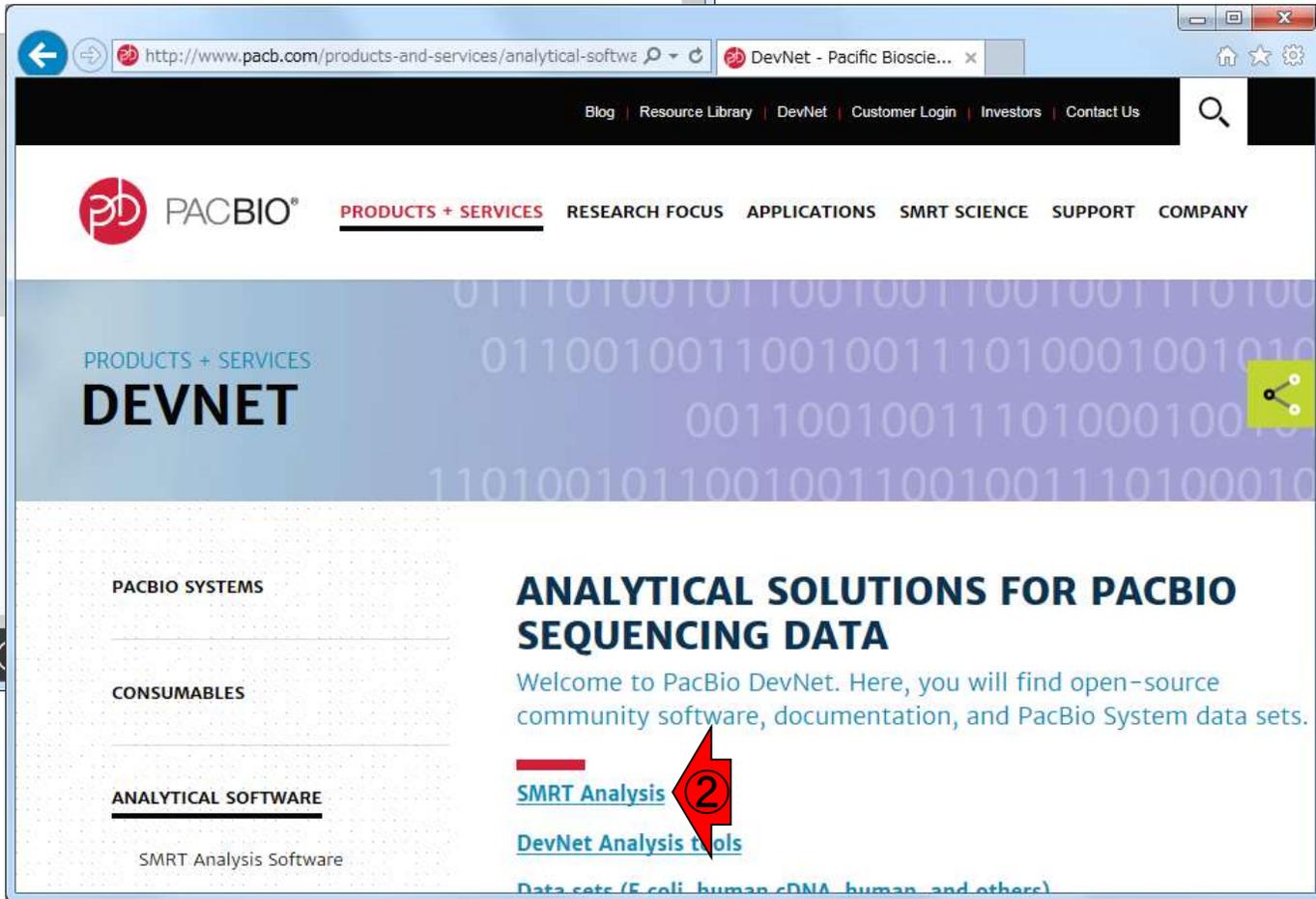
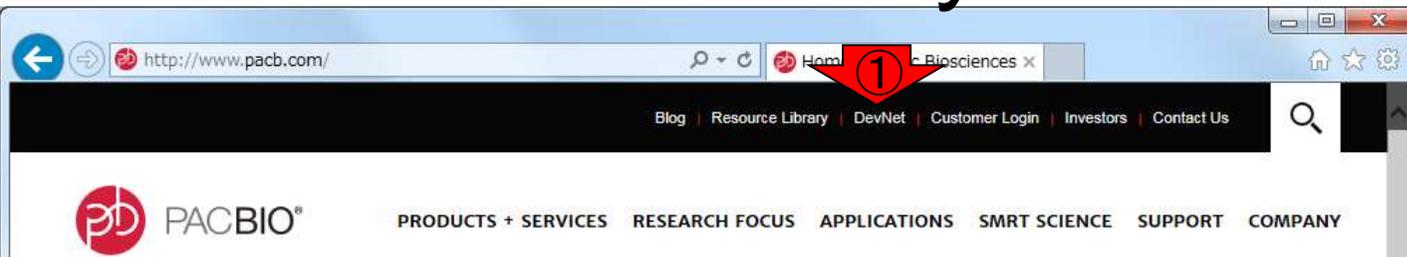
- Submission [DRA002643](#) [FTP](#)
- Study [DRP002401](#)
- Experiment [DRX022185](#) [FASTQ](#) [SF](#)

READS (joined) quality show 10 rows << < 1 / 16349 Page > >>

```
>DRR054116.1
AGTATGAGTAATGCTGCTAGGTTACGTATATGTGACGACTGAATTTTCTCGCCGTGCTAGTTTAGTAATGATCATAkata
TGTTGGACGTGTTGATGGTACTGACGGCTTCTGCAAGCGAGTCATGGATAGTGACTGTAATAGTATTATGTTATCGGATG
GAGCACGTACGTGTGTGACGTAGATGACTTGTTCGATGATTATTTCTCATGCGCAGTGTGTTAGTTCTCATATAGTTGCTAC
TGTGGCTGTTTTGCACATGCAGGAGAGTTCTGACGGGGATAATCGACTATGTGGGTGATGGACATCTACTGACGCGCATC
TTATTATGGTTCTGATGCCTCTTTATAAAAATTTGTCA

>DRR054116.2
CATGCTGACGTTTTTCGAGGATGGGAGGACACGGATGGTTGATCGTACTGACAGTATCTTCAAGTGCATTGTTTGCAGTTC
GGTGGTGATCGATCGTGATTGATTCTTTGAGTAGATTATTGTTAGGCAGTAGTGTGACCTGTGATTGTAGATAGTACAG
GTCTACTAGTTGATGAAAGGTCATGTGACTGAGGATTGGAAGTAGTTCTATCTTAGAGTGTTTCGATTTCGTTAGAGGTTAG
GAGATGTCCGAGTGGATGTACTAGAATGCTGATCAGTGTACGTGTGATGATCTGTATTAGTTCTCGTACTTACGCATCG
GCTCGATGATGTGATGAATGACCAGTCAGACTATGTTCTTAGATTGTGTATCGACGACGGGTCGACACCAGTTGCGTGTA
```

W2-6: SMRT Analysis



W2-6: SMRT Analysis

①ページ下部に移動し、②の手順を参考にインストールを行う。③が手順書の中に書かれているwgetでダウンロードするもの

The screenshot shows the PacBio website's 'SMRT Analysis Software' page. The page layout includes a top navigation bar with links for 'Blog', 'Resource Library', 'DevNet', 'Customer Login', 'Investors', and 'Contact Us'. Below this is a secondary navigation bar with 'PRODUCTS + SERVICES', 'RESEARCH FOCUS', 'APPLICATIONS', 'SMRT SCIENCE', 'SUPPORT', and 'COMPANY'. The main content area features a green 'CURRENT' banner with a 'Visit our blog >>' link. A 'FEATURED RESOURCE' section highlights a registration event for the SMRT Informatics Developers Conference in San Diego. The 'Additional resources' section contains several links: 'PacBio software downloads' (marked with a red arrow 3), 'SMRT Analysis release notes', 'SMRT Analysis system requirements', 'SMRT Analysis installation' (marked with a red arrow 2), and 'SMRT Analysis web services API'. A red arrow 1 points to the bottom right corner of the page, indicating the location for scrolling to find the wget command mentioned in the text.

W2-7: bax.h5ファイル

①1セル分のみでも(747MB + 766MB + 901MB) = 2,414MB (約2.4GB)。実際のダウンロードはW7-1で行う

PacBioのファイル形式とデータ解析の概要

- W1-1: PacificBiosciencesのYouTubeサイト
 - [Introduction to SMRT Sequencing](#)
 - [Single Molecule Real Time Sequencing](#)
- W2-1: PacBioデータ(原著論文中のDRR IDだが削除されている)
 - [DRR024500](#): [Tanizawa et al., BMC Genomics, 2015](#)
- W2-3: [DRR024501](#) -> [DRP002401](#) -> [DRX022185](#)
- W2-4: [DRR024501](#) -> [DRA002643](#)
- W2-5: PacBioデータ概観
 - [DRR054113](#)
 - [DRR054114](#)
 - [DRR054115](#)
 - [DRR054116](#)
- W2-6: SMRT Portal(PacBio提供のHGAPを含む解析ソフトウェア群)の場所
 - [PacBio](#) -> [DevNet](#) -> [SMRT Analysis](#)
 - SMRT Analysis 2.3までは、HGAPを実行するためには**bax.h5**ファイルが必須。
 - SMRT Analysis 3.0からは、BAMファイルが入力フォーマットになる。但しここでのBAMファイルは、マッピングデータではなく、シーケンス生データ。
 - PacBio RSIIの後継機である [Sequel](#) の出力ファイル形式はBAM。
 - PacBioのファイル形式の説明については [こちら](http://pacbiofileformats.readthedocs.io/) (<http://pacbiofileformats.readthedocs.io/> (3.0))。
- W2-7: [DRR054113](#)のbax.h5ファイル(下記3ファイル合わせてDRR054113に相当)
 - [m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5](#) (747 MB; 784,301,199 bytes)
 - [m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5](#) (766 MB; 803,938,042 bytes)
 - [m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5](#) (901 MB; 945,597,712 bytes)

①

W3-1: FASTQダウンロード

①DRR054113、②FASTQ、③bzip2
 圧縮FASTQファイルをダウンロード。右クリックで「ショートカットのコピー」などでURL情報を取得(第4回W9-2やW18-1)してwgetしてもよいし、共有フォルダ経由でBio-Linux上に置いてもよい。

DRASearch Send Feedback

DRR054113

Run Detail	
Alias	DRR054113
Instrument model	
Date of run	
Run center	
Number of spots	163,482
Number of bases	360,244,590

Navigation	
Submission	DRA002643 <input type="button" value="FTP"/>
Study	DRP002401
Experiment	DRX022185 <input type="button" value="FASTQ"/>

READS (joined)	qual	01/27/2016 02:32午後	2,392,259	DRR054113.fastq.bz2
>DRR054113.1		01/27/2016 02:32午後	2,901,836	DRR054114.fastq.bz2
CCTATGCTGTCAGCATTGATTGCTAGTTGAT		01/27/2016 02:33午後	3,858,646	DRR054115.fastq.bz2
GCTACTAGTCTTGAGTCTGCCTGACTGATTGA		01/27/2016 02:33午後	4,455,155	DRR054116.fastq.bz2
GTACGGCAGCATGTAGTACTGGTGCATGACC				
GCAGGTCATGATACCAGGTCGATTTCAGTATTCGATGTCTAGACTTAGCTGACATAGCAGATTGATCTTCTTGATTACAGG				
CGATATCGCACTGCGTCATACGATTCACAGTCA				
>DRR054113.2				
TCATATACTCGGCACAATGTGTGTCGATCGTAAAGGGATGTCATTGTGTAGTATTGTATTCTATATGTCGAGCATCAGCG				
TTCTACTGCTGAGATGATATATTCTGAGTATTATGGTTATGTATTTTACGCTGAACCTGGATTATGTCGTGGACGGACGT				
TGTACGGATTTCTAACTGTTAGTATCGAGCATTGATCGTCCGATGGATTGATAGTGCTTCCGTTGAGTCGTAATGATTGTT				
CAGTTAGTCGATCAGTTCGTGGATCAGTGATTTTGTAGGCGAAGATTATGAATCTTTACGATCCTTATGGCTGAGTTGAA				
TGATCTGCTGCTAGTGTCTGTATGTTTCGTATGATCACATGACGATACGTCGATATTTATTATTGTCTACGCATCGATTGAG				

W3-1: FASTQダウンロード

作業ディレクトリはどこでもよいが、ここでは①~/Documents/DRR054113で②wgetしている。③ファイルサイズは数MB程度なので、ダウンロード自体はほぼ一瞬で終わる。

```
iu@bielinux[iu] cd ~/Documents [12:32午後]
iu@bielinux[Documents] pwd [12:32午後]
/home/iu/Documents
iu@bielinux[Documents] mkdir DRR054113 [12:32午後]
iu@bielinux[Documents] cd DRR054113 [12:32午後]
iu@bielinux[DRR054113] pwd [12:32午後]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] wget -cq ftp://ftp.ddbj.nig.ac.jp/ddbj_d [12:32午後]
atabase/dra/fastq/DRA002/DRA002643/DRX022185/DRR054113.fastq.bz [12:32午後]
2
iu@bielinux[DRR054113] ls -l [12:32午後]
total 2340
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
iu@bielinux[DRR054113] ls -lh [12:33午後]
total 2.3M
-rw-rw-r-- 1 iu iu 2.3M 3月 22 12:32 DRR054113.fastq.bz2
iu@bielinux[DRR054113] █ [12:33午後]
```

W3-2: FastQC

①FastQC ver. 0.11.4は、第4回W9-2でインストールし、fastqc2というコマンドでパスを通してている。②FastQCを実行し、共有フォルダ(~/Desktop/mac_share)に保存している。③ファイルの確認。

```
iu@bielinux[DRR054113] pwd [12:35午後]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [12:35午後]
total 2340
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
① iu@bielinux[DRR054113] fastqc2 -v [12:35午後]
FastQC v0.11.4
② iu@bielinux[DRR054113] fastqc2 -q DRR054113.fastq.bz2 --outdir=
/home/iu/Desktop/mac_share
③ iu@bielinux[DRR054113] ls -l ~/Desktop/mac_share [12:35午後]
total 751
-rwxrwxrwx 1 iu iu 392406 3月 22 12:35 DRR054113_fastqc.html
-rwxrwxrwx 1 iu iu 375781 3月 22 12:35 DRR054113_fastqc.zip
iu@bielinux[DRR054113] [12:35午後]
```

W3-3: 結果を眺める

①共有フォルダに保存することで、使いなれた
ホストOS(この場合Windows)上で②FastQC
実行結果ファイルを眺めることができる。

The screenshot shows a Windows desktop environment. On the left, a 'share' folder is visible in the taskbar, with a red arrow labeled '1' pointing to it. Below it, a file explorer window shows the contents of the 'share' folder, including 'DRR054113_fastqc.html' and 'DRR054113_fastqc.zip', with a red arrow labeled '2' pointing to the HTML file. On the right, a web browser window displays the 'FastQC Report' for the file 'DRR054113.fastq.bz2'. The report includes a 'Summary' section with a list of metrics and their status, and a 'Basic Statistics' table.

Summary

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

Basic Statistics

Measure	Value
Filename	DRR054113.fastq.bz2
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	915
Sequences flagged as poor quality	0
Sequence length	923-8076
%GC	38

Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by FastQC (version 0.11.4)

W3-3: 結果を眺める

FastQC実行結果の解説は第4回W8とW17、および第6回W4にもあり。①入力ファイル。②リード数は915、③配列長は923-8076 bpの範囲であることがわかる。

FastQC Report

Summary

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

✓ Basic Statistics

Measure	Value
Filename	DRR054113.fastq.bz2 ①
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	915 ②
Sequences flagged as poor quality	0
Sequence length	923-8076 ③
%GC	38

✗ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)



W3-3: 結果を眺める

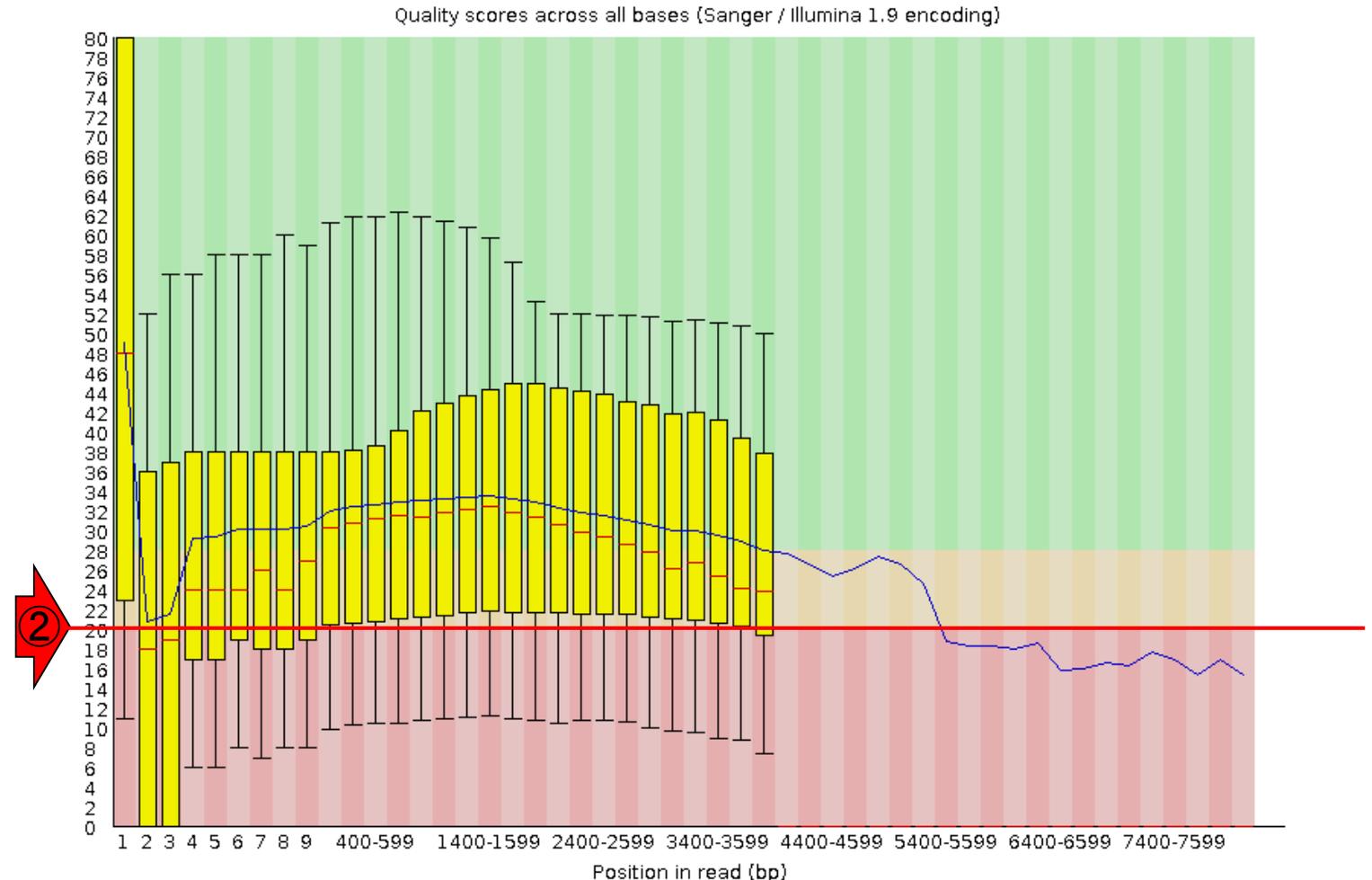
①Per base sequence quality. この図の縦軸はクオリティスコア。②赤線のスコア20を超えているかどうか1つの目安。Illumina HiSeq2000 (第4回のW8)やMiSeq (第6回のW4)とは傾向が異なる。

FastQC Report

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality **①**
- ✗ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✗ Per base sequence quality



W3-3: 結果を眺める

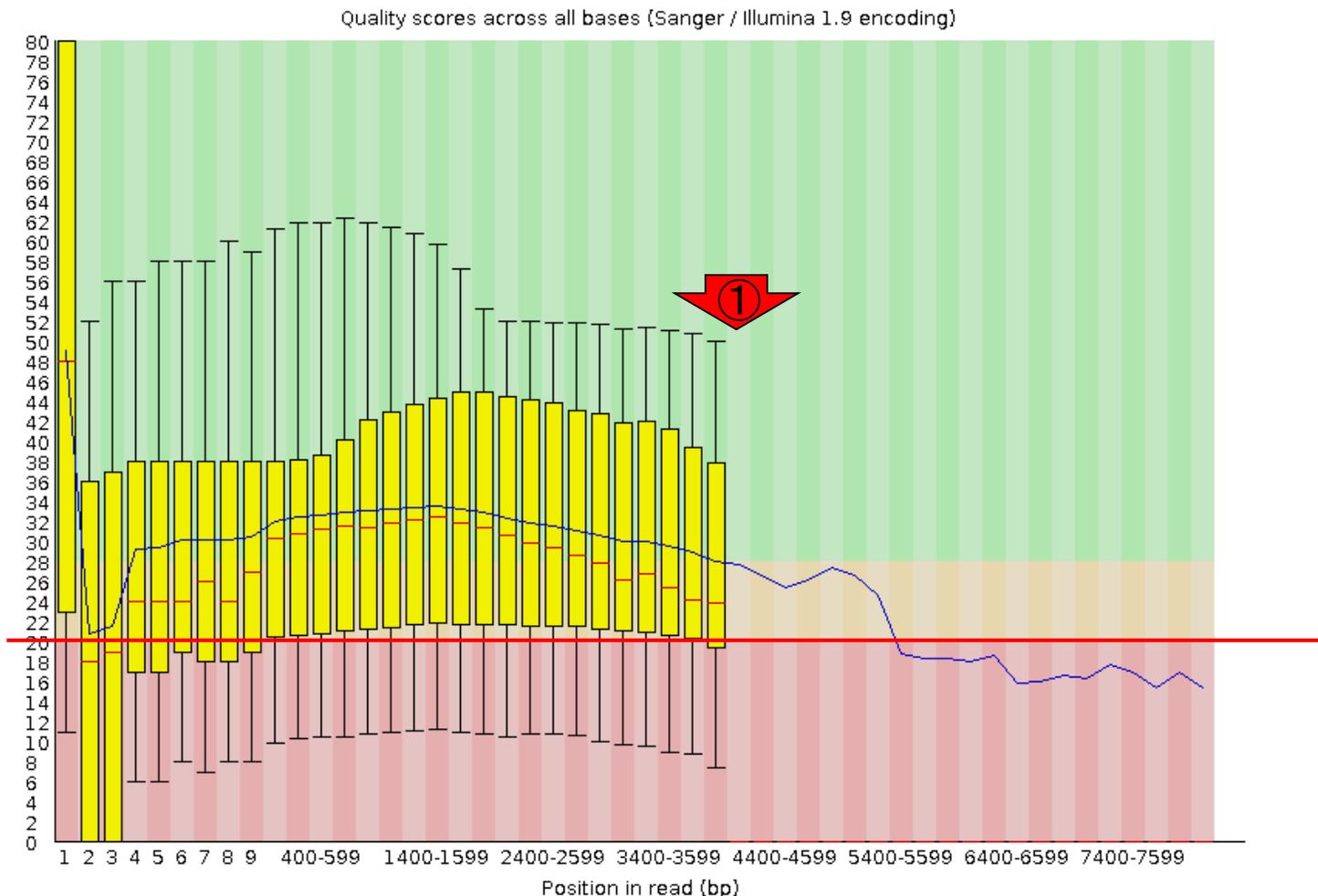
①横軸のリードポジションが4000 bpあたりで黄色の縦棒がなくなっているのは、4000 bp以上のリードが少数だからだと思われる。それは②の配列長分布(Sequence Length Distribution)で確認できる

FastQC Report

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution ②
- ✓ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✗ Per base sequence quality



W3-4: 配列長分布

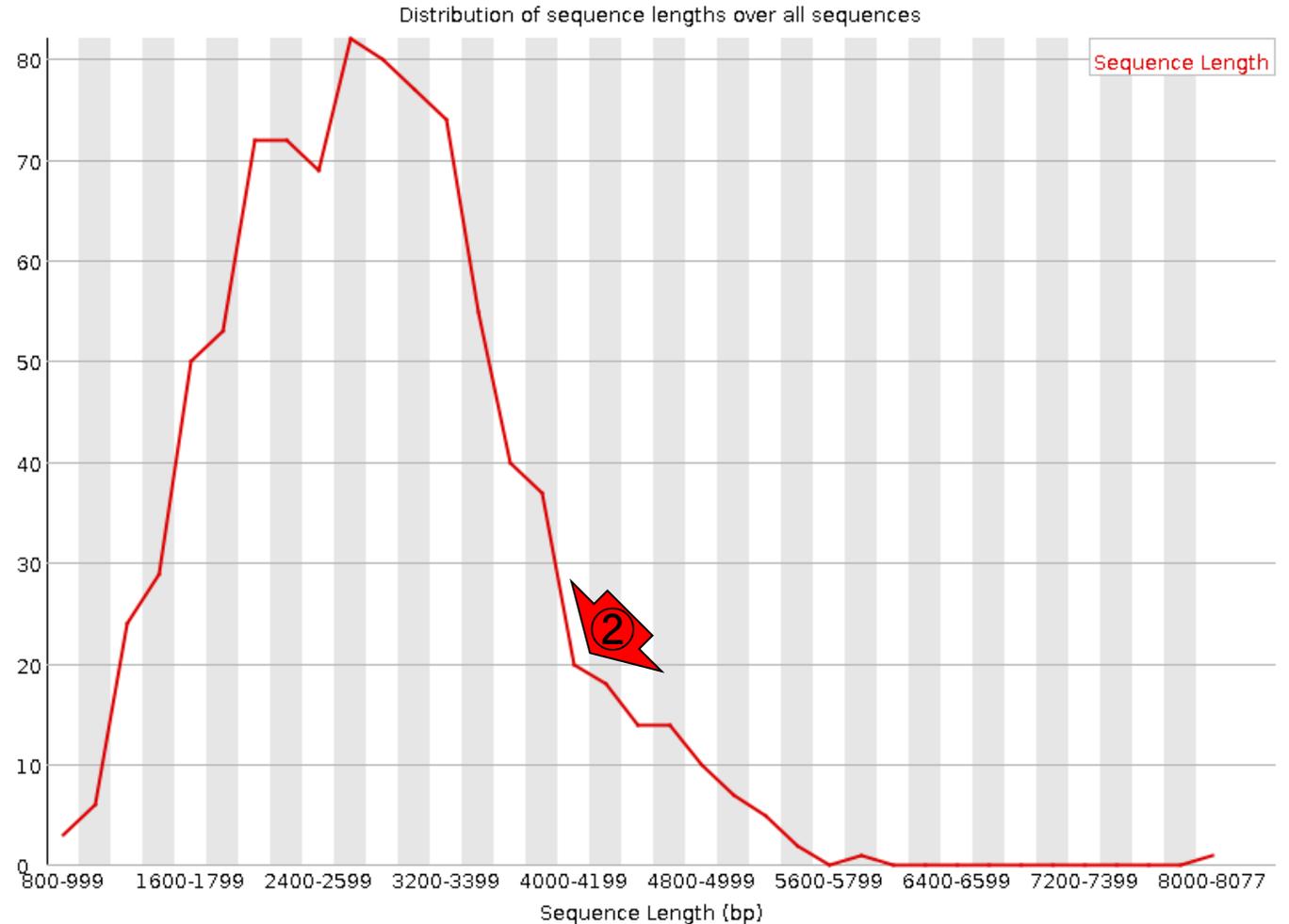
①配列長分布(Sequence Length Distribution)。②このあたりで黄色の縦棒がなくなっているのので、おそらく20リードが黄色の縦棒の有無の閾値なのだろう

FastQC Report

Summary

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

! Sequence Length Distribution



W3-5: bzip2 → gz

①元のbzip2ファイルを残したままgzipファイルを作成。bzip2の-dは解凍オプション、-cは解凍結果を標準出力させるオプション。bzip2 -dc使用例は、第3回W22-2、第6回W3にもあり。②bzip2とgzipの圧縮効率に関しては第3回W13にもあり

```
File Edit View Search Terminal Help
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l
total 2340
-rw-rw-r-- 1 iu iu 2392259  3月 22 12:32 DRR054113.fastq.bz2
iu@bielinux[DRR054113] bzip2 -dc DRR054113.fastq.bz2 | gzip > D
RR054113.fastq.gz
iu@bielinux[DRR054113] ls -l
total 5000
-rw-rw-r-- 1 iu iu 2392259  3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482  3月 22 12:38 DRR054113.fastq.gz
iu@bielinux[DRR054113]
```

[12:38午後]

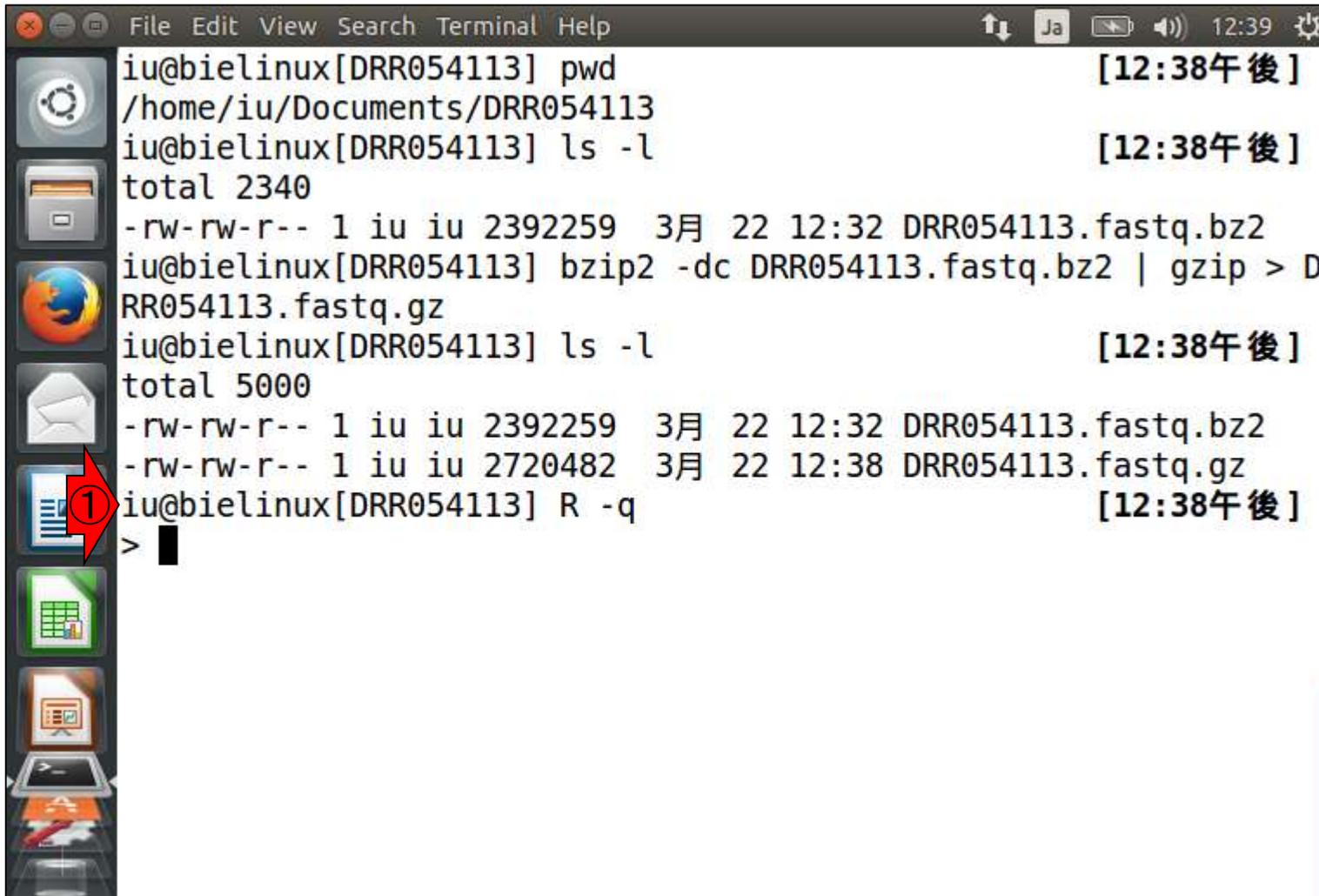
[12:38午後]

[12:38午後]



Rの起動の基本は、Rのみでよい(第5回W7)。①-qオプションをつけてメッセージ表示を省略している(第5回W9-7)。

W3-6: R起動



A terminal window titled "Terminal" with a menu bar (File, Edit, View, Search, Terminal, Help) and system icons (language: Ja, battery, volume, time: 12:39). The terminal shows the following commands and outputs:

```
iu@bielinux[DRR054113] pwd [12:38午後]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [12:38午後]
total 2340
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
iu@bielinux[DRR054113] bzip2 -dc DRR054113.fastq.bz2 | gzip > D
RR054113.fastq.gz
iu@bielinux[DRR054113] ls -l [12:38午後]
total 5000
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
iu@bielinux[DRR054113] R -q [12:38午後]
> █
```

A red arrow with the number "1" points to the command `R -q`.

①入力はgzip圧縮ファイル(DRR054113.fastq.gz)
、②出力はhoge3.txt。

W3-6: 入出力

```
iu@bielinux[DRR054113] pwd [12:38午後]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [12:38午後]
total 2340
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
iu@bielinux[DRR054113] bzip2 -dc DRR054113.fastq.bz2 | gzip > D
RR054113.fastq.gz
iu@bielinux[DRR054113] ls -l [12:38午後]
total 5000
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
iu@bielinux[DRR054113] R -q [12:38午後]
> in_f <- "DRR054113.fastq.gz" #入力ファイル名を指定
してin_fに格納
> out_f <- "hoge3.txt" #出力ファイル名を指定
してout_fに格納
>
```

- W3-6: Rで配列長の具体的な数値情報を取得
「前処理 | クオリティチェック | [配列長分布を調べる](#)」の例題3と基本的に同じです。

```
pwd
R -q

in_f <- "DRR054113.fastq.gz" #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt" #出力ファイル名を指定してout_fに格納
library(ShortRead) #パッケージの読み込み
fastq <- readFastq(in_f) #in_fで指定したファイルの読み込み
out <- table(width(fastq)) #長さごとの出現頻度情報を得た結果をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=F, col.names=F)
q(save="no")
```

①赤枠をコピーで実行。②特にエラーメッセージも出ずに、無事Linuxコマンド入力待ち状態になっていることがわかる

W3-6: コピペ後

```
File Edit View Search Terminal Help
Loading required package: BiocParallel
Loading required package: Biostrings
Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' from package 'base' in
package 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
Loading required package: Rsamtools
Loading required package: GenomeInfoDb
Loading required package: GenomicRanges
Loading required package: GenomicAlignments
> fastq <- readFastq(in_f) #in_fで指定したファイルの読み込み
> out <- table(width(fastq)) #長さごとの出現頻度情報を得た結果をoutに格納
> write.table(out, out_f, sep="\t", append=F, quote=F, row.names=F, col.names=F) #outの中身をout_fに格納
>
> q(save="no")
iu@bielinux[DRR054113] █
```

• W3-6: Rで配列長の具体的な数値情報を取得
「前処理 | クオリティチェック | [配列長分布を調べる](#)」の例題3と基本的に同じです。

```
pwd
R -q

in_f <- "DRR054113.fastq.gz" #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.txt" #出力ファイル名を指定してout_fに格納
library(ShortRead) #パッケージの読み込み
fastq <- readFastq(in_f) #in_fで指定したファイルの読み込み
out <- table(width(fastq)) #長さごとの出現頻度情報を得た結果をoutに格納
write.table(out, out_f, sep="\t", append=F, quote=F, row.names=F, col.names=F)
q(save="no")
```



W3-7: 結果の確認

①確かに出力ファイルとして指定したhoge3.txtが作成されている。②最初の5行分と③最後の5行分を表示。FastQC実行結果(W3-3)で見られた配列長の範囲(923-8076 bp)と同じである。

```
iu@bielinux[DRR054113] pwd [12:41午後]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [12:41午後]
total 5008
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 5541 3月 22 12:40 hoge3.txt
iu@bielinux[DRR054113] head -n5 hoge3.txt [12:41午後]
923 1
951 1
973 1
1079 1
1100 1
iu@bielinux[DRR054113] tail -n5 hoge3.txt [12:41午後]
5393 1
5411 1
5597 1
5850 1
8076 1
iu@bielinux[DRR054113] [12:41午後]
```



W3-7: 結果の確認

①2列目部分の数値は出現回数。ここで見えているものは全て1になっているが、例えば923 bpの長さのリードは1つしかなかった、という風に解釈する。

```
iu@bielinux[DRR054113] pwd [12:41午後]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [12:41午後]
total 5008
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 5541 3月 22 12:40 hoge3.txt
iu@bielinux[DRR054113] head -n5 hoge3.txt [12:41午後]
923 1
951 1
973 1
1079 1
1100 1
iu@bielinux[DRR054113] tail -n5 hoge3.txt [12:41午後]
5393 1
5411 1
5597 1
5850 1
8076 1
iu@bielinux[DRR054113] [12:41午後]
```



923	1
951	1
973	1
1079	1
1100	1

5393	1
5411	1
5597	1
5850	1
8076	1

W3-7: 結果の確認

①hoge3.txtの行数は792。総リード数が915個、配列長の範囲が923-8076 bpなので、ほとんどの配列長のものが1回しか出現しないという結果(792/915)は妥当

```
File Edit View Search Terminal Help
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [12:41午後]
total 5008
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 5541 3月 22 12:40 hoge3.txt
iu@bielinux[DRR054113] head -n5 hoge3.txt [12:41午後]
923 1
951 1
973 1
1079 1
1100 1
iu@bielinux[DRR054113] tail -n5 hoge3.txt [12:41午後]
5393 1
5411 1
5597 1
5850 1
8076 1
① iu@bielinux[DRR054113] wc hoge3.txt [12:41午後]
792 1584 5541 hoge3.txt
iu@bielinux[DRR054113] [12:42午後]
```

W3-4で5000 bp以上のリード数は20個以下だという当たりをつけているので、適当に「tail -n19 ...」などで調べる。①ここでは最後の17行分を表示させ、②16リードが5000 bp以上だと確認

W3-7: 結果の確

```
iu@bielinux[DRR054113] tail -n17 hoge3.txt [12:43午後]
4995 1
5005 1
5019 1
5059 1
5062 1
5075 1
5109 1
5161 1
5213 1
5274 1
5304 1
5353 1
5393 1
5411 1
5597 1
5850 1
8076 1
iu@bielinux[DRR054113] [12:43午後]
```

W3-8: sraダウンロード

①DRR054113、②FASTQ、③sraファイルをダウンロード。約1.4GBあるので、エアーハンズオン(やったつもり)でもよい。

DRASearch Send Feedback [Search Home](#) [DRA Home](#)

DRR054113 [FASTQ](#) [SRA](#)

Run Detail

Alias	DRR054113
Instrument model	
Date of run	
Run center	
Number of spots	163,482
Number of bases	360,244,590

Navigation

- Submission [DRA002643](#) [FTP](#)
- Study [DRP002401](#)
- Experiment [DRX022185](#) [FASTQ](#) [SF](#)

READS (joined) quality show 10 rows << < 1 / 16349 Page > >>

```
>DRR054113.1
CCTATGCTGTCAGCATTGATTGCTAGTTGATGGTTCTATATTTACGTATCACATTGAGATATATCGCTCATCAGCTTCT
GCTACTAGTCTTGAGTCTGCCTGACTGATTGATTGATCATGCGTGGATTCTGATGATACTTTATGCATTATACGAGTTAC
GTACGGCAGCATGTAGTACTGGTGCATGACCTCATGAGCTAGCATTGAGTTATCGTGATCCATAACTGGATCAGTACTT
GCAGGTCATGATACCAGGTCGATTTCAGTATTCGATGTCTAGACTTAGCTGACATAGCAGATTGATCTTCTTGATTACAGG
CGATATCGCACTGCGTCATACGATTCACAGTCA

>DRR054113.2
TCATATACTCGGCACAATGTGTGTCGATCGTAAAGGGATGTCATTGTGTAGTATTGATTCTATATGTCGAGCATCAGCG
TTCTACTGCTGAGATGATATATTCTGAGTATTATGGTTATGTATTTTACGTGAACCTGGATTATGTCGTGGACGGACGT
TGTACGGATTTCTAACTGTTAGTATCGAGCATTGATCGTCCGATGGATTGATAGTGCTTCCGTTGAGTCGTAATGATTGTT
CAGTTAGTCGATCAGTTCGTGGATCAGTGATTTTGTAGGCGAAGATTATGAATCTTTACGATCCTTATGGCTGAGTTGAA
TGATCTGCTGCTAGTGTCTGTATGTTTCGATGATCACATGACGATACGTGATATTTATTATTGTCTACGCATCGATTGAG
```

01/27/2016 02:34午後 1,418,046,334 [DRR054113.sra](#)

W3-8: sraダウンロード

- ①wget。東大有線LAN環境では約2分。
- ②DRA上のファイルサイズと同じで安心。

```
File Edit View Search Terminal Help 15:30
iu@bielinux[DRR054113] pwd [ 3:21午後 ]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [ 3:21午後 ]
total 5008
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 5541 3月 22 12:40 hoge3.txt
① iu@bielinux[DRR054113] wget -cq ftp://ftp.ddbj.nig.ac.jp/ddbj_d
atabase/dra/sra/ByExp/sra/DRX/DRX022/DRX022185/DRR054113/DRR054
113.sra
iu@bielinux[DRR054113] ls -l [ 3:23午後 ]
total 1389824
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra ②
-rw-rw-r-- 1 iu iu 5541 3月 22 12:40 hoge3.txt
iu@bielinux[DRR054113] █ [ 3:23午後 ]
```

W3-8: ファイルサイズ

01/27/2016	02:32午後	2,392,259	DRR054113.fastq.bz2
01/27/2016	02:32午後	2,901,836	DRR054114.fastq.bz2
01/27/2016	02:33午後	3,858,646	DRR054115.fastq.bz2
01/27/2016	02:33午後	4,455,155	DRR054116.fastq.bz2



01/27/2016	02:34午後	1,418,046,334	DRR054113.sra
01/27/2016	02:34午後	1,395,142,624	DRR054114.sra
01/27/2016	02:34午後	1,487,687,328	DRR054115.sra
01/27/2016	02:35午後	2,031,098,626	DRR054116.sra



W3-9:リード数の違い

DRA で公開されている fastq のリード数が生データのそれよりも少ないのは何故でしょうか?

DRA では NCBI SRA Toolkit に含まれている fastq-dump を使い、以下のオプションで生データである SRA ファイルから fastq ファイルを作成しています。

```
fastq-dump -M 25 -E --skip-technical --split-3 -W <SRA file>
```

- -M 25: 25 塩基以上の配列のみを含める。デフォルトは 25。
- -E: リードの開始、もしくは終わりに 10 以上の N が存在しない
- --skip-technical: technical read を除き biological read のみを出力
- --split-3: ペアリードで最初と二番目の biological read をそれぞれ *_1.fastq と *_2.fastq として出力する。一つしか biological read が存在しない場合、*.fastq として出力する。
- -W: 指定されていた場合、left と right を clip する

上記の出力条件でリードがフィルタリング、トリミングされるため、一般的に fastq のリード数は SRA ファイルのそれよりも少なくなっています。フィルタリング、トリミングされていない fastq ファイルを得るには以下のコマンドで fastq を生成します。

```
fastq-dump -M 1 --split-3 <SRA file>
```

作成日: 2013年10月8日; 最終更新日: 2014年6月6日

<http://trace.ddbj.nig.ac.jp/dra/faq.html#read-number-fastq>

W4-1 : SRA Toolkit

①最新版はver. 2.5.7(2016年3月22日現在)。②プログラムはOSの種類ごとに用意されている。Bio-Linuxの実体はUbuntuなので③ここ。話についてこれないヒトは、連載第1回を復習。④documentationをクリック

The screenshot shows the NCBI SRA Toolkit download page. The browser address bar shows the URL: <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>. The page title is "Sequence Read Archive". The navigation menu includes "Main", "Browse", "Search", "Download", "Submit", "Documentation", "Software", "Trace Archive", "Trace Assembly", "Trace Home", and "Trace BLAST". The "Download" tab is active, showing "Toolkit Documentation" and "XML Schema". The main heading is "SRA Toolkit". Below the heading, there is a link for "For Toolkit documentation [click here](#)". A list of download options is provided, with annotations: ① points to the first item, ② points to the OS-specific links, ③ points to the Ubuntu link, and ④ points to the documentation link. The list includes:

1. NCBI SRA Toolkit latest release (December 23 2015, version 2.5.7 release) compiled binaries and [md5 checksums](#)*:
 - [CentOS Linux 64 bit architecture](#)
 - [Ubuntu Linux 64 bit architecture](#)
 - [MacOS 64 bit architecture](#)
 - [MS Windows 64 bit architecture](#)
 - [vdb-view Windows Installer](#) is a spreadsheet-like browser for viewing SRA and vdb objects - Windows only
2. NCBI Decryption Tools latest release binaries and [md5 checksums](#)*:
 - [CentOS Linux 64 bit architecture](#)
 - [CentOS Linux 32 bit architecture](#)
 - [Ubuntu Linux 64 bit architecture](#)
 - [Ubuntu Linux 32 bit architecture](#)
 - [MacOS 64 bit architecture](#)
 - [MacOS 32 bit architecture](#)
 - [MS Windows 64 bit architecture](#)
 - [MS Windows 32 bit architecture](#)
3. Latest Source Code:
 1. [NGS Software Development Kit](#) – November 24 2015, version 1.2.3 release
 2. [NCBI VDB Software Development Kit](#) – December 23 2015, version 2.5.7 release
 3. [NCBI SRA Toolkit](#) – December 23 2015, version 2.5.7 release

W4-1 : SRA Toolkit

①目的のfastq-dumpプログラムは、よく使われるツール群(Frequently Used Tools)の最初に位置する。fastq-dumpを利用したいがために、SRA Toolkitをインストールするヒトがほとんどであろう。基本的には②を参考にインストール、クリック

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc

NCBI Site map All databases Search

Sequence Read Archive

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

Download **Toolkit Documentation** XML Schema

SRA Toolkit Documentation

[SRA Toolkit Installation and Configuration Guide](#) ②

[Protected Data Usage Guide](#)

Frequently Used Tools:

[fastq-dump: Convert SRA data into fastq format](#) ①

[prefetch](#): Allows command-line downloading of SRA, dbGaP, and ADSP data

[sam-dump](#): Convert SRA data to sam format

[sra-pileup](#): Generate pileup statistics on aligned SRA data

[vdb-config](#): Display and modify VDB configuration information

[vdb-decrypt](#): Decrypt non-SRA dbGaP data ("phenotype data")

Additional Tools:

[abi-dump](#): Convert SRA data into ABI format (csfasta / qual)

[illumina-dump](#): Convert SRA data into Illumina native formats (qseq, etc.)

[sff-dump](#): Convert SRA data to sff format

[sra-stat](#): Generate statistics about SRA data (quality distribution, etc.)

[vdb-dump](#): Output the native VDB format of SRA data.

[vdb-encrypt](#): Encrypt non-SRA dbGaP data ("phenotype data")

日本乳酸菌学会誌の連載第7回

W4-1 : SRA Toolkit

SRA Toolkit Installation and Configuration

①wgetでtar.gzをダウンロードし、②解凍するのが基本だが…折角なので第4回W4-5、W13-5、W14、W15で紹介した「sudo apt-get install **ソフトウェア名**」でSRA Toolkitのインストールを行うやり方を伝授

Table of Contents

1. [Downloading and installing the SRA Toolkit](#)
2. [Testing the Toolkit configuration](#)
3. [Configuring the Toolkit](#)
4. [Links and help documents](#)

Contact: sra-tools@ncbi.nlm.nih.gov

The following guide will outline the download, installation, and configuration of the SRA Toolkit. Detailed information regarding the usage of individual tools in the SRA Toolkit can be found on the tool-specific documentation pages.

The NCBI SRA Toolkit enables reading ("dumping") of sequencing files from the SRA database and writing ("loading") files into the .sra format (Note that this is not required for submission). The Toolkit source code is provided in the form of the [SRA SDK](#), and may be compiled with GCC. However, pre-built software executables are available for Linux, Windows, and Mac OS X, and we highly recommend using these pre-built executables whenever possible.

[Downloading and installing the SRA Toolkit](#)

Download the Toolkit from the SRA website

1. If you are using a web browser, the following page contains download links to the most current version of the toolkit for each of the supported platforms: SRA Toolkit download page: [//www.ncbi.nlm.nih.gov/Traces/sra/?view=software](http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software)
2. If you are instead working from a command line interface, you may use FTP or wget to obtain the software from the following directory: "[//ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current](ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current)". Example:

```
wget "//ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-centos\_linux64.tar.gz"
```

Unpack the Toolkit:

1. For Linux, use tar:

```
tar -xzf sratoolkit.current-centos\_linux64.tar.gz
```

W4-2: apt-cache

目的:「sudo apt-get install ソフトウェア名」のソフトウェア名のところで指定する名前を知りたい! やり方:「apt-cache -n search キーワード」で任意のキーワードを含むソフトウェア名をリストアップする。ここでは、①SRAを含むソフトウェア名をリストアップ。②欲しいソフトウェア名は、sra-toolkit。

```
File Edit View Search Terminal Help
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
① iu@bielinux[DRR054113] apt-cache -n search SRA [ 8:16午後 ]
libratom-0-0 - library for serialising LV2 atoms to/from Turtle
libratom-dev - library for serialising LV2 atoms to/from Turtle
- development files
libratom-doc - library for serialising LV2 atoms to/from Turtle
- documentation
sra-toolkit - utilities for the NCBI Sequence Read Archive ②
sra-toolkit-libs-dev - Development files for the NCBI SRA Toolkit's libraries
sra-toolkit-libs0 - Libraries for the SRA Toolkit
iu@bielinux[DRR054113] [ 8:16午後 ]
```

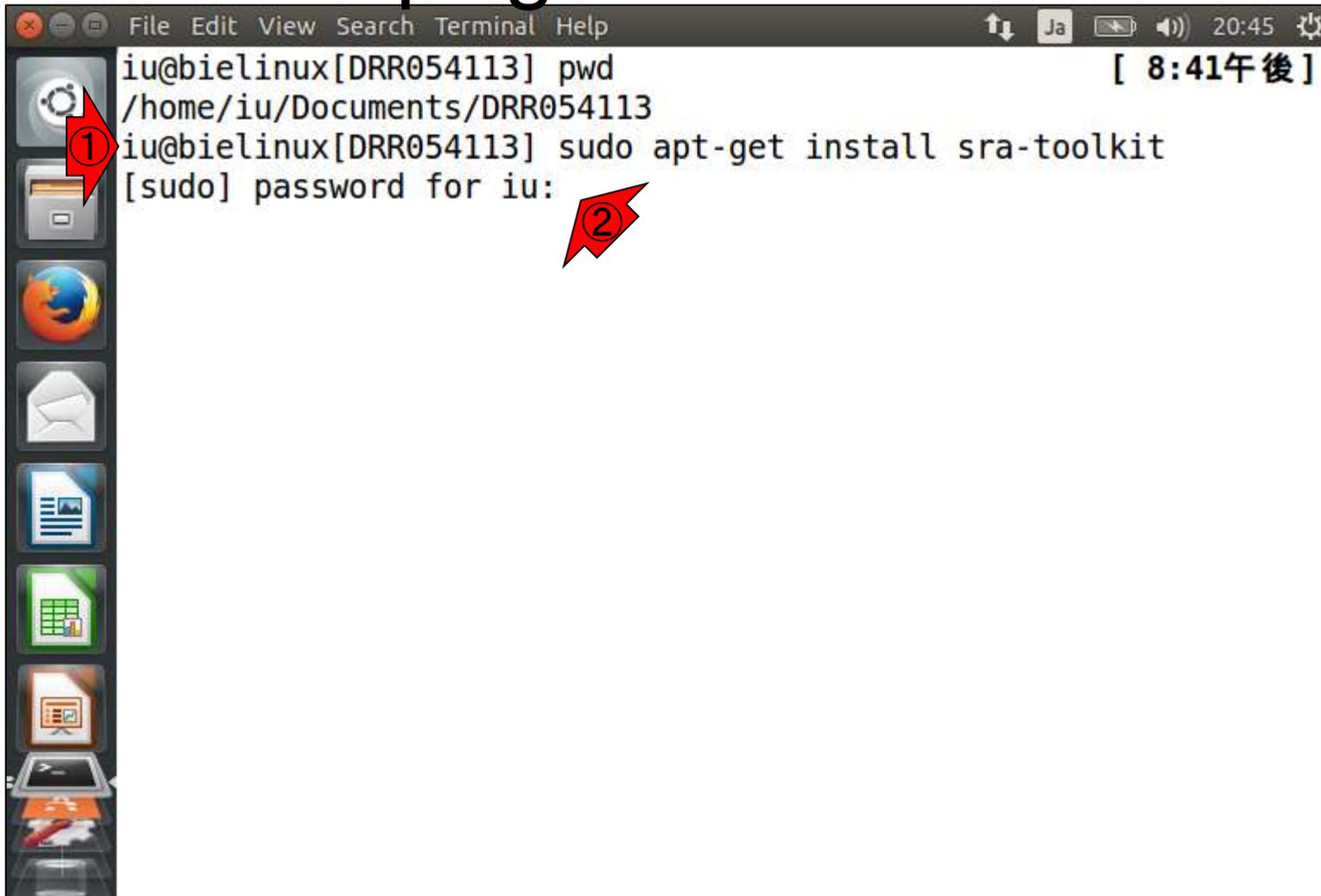
W4-2: apt-cache

おまけ。①「apt-cache -n search SRA」実行結果として、小文字のsraを含むソフトウェア名もリストアップされたことから、**キーワード**部分は、大文字でも小文字でもどちらでもいいのだろうと学習する。また、②「| wc」を追加することで、ソフトウェア名が6個だったと認識

```
File Edit View Search Terminal Help
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
① iu@bielinux[DRR054113] apt-cache -n search SRA [ 8:16午後 ]
libratorom-0-0 - library for serialising LV2 atoms to/from Turtle
libratorom-dev - library for serialising LV2 atoms to/from Turtle
- development files
libratorom-doc - library for serialising LV2 atoms to/from Turtle
- documentation
sra-toolkit - utilities for the NCBI Sequence Read Archive
sra-toolkit-libs-dev - Development files for the NCBI SRA Toolkit's
libraries
sra-toolkit-libs0 - Libraries for the SRA Toolkit
② iu@bielinux[DRR054113] apt-cache -n search SRA | wc [ 8:16午後 ]
      6      58     418
iu@bielinux[DRR054113] [ 8:29午後 ]
```

W4-3: apt-get

①「sudo apt-get install sra-toolkit」。rootのパスワードを聞かれたら打ち込む(推奨手順通りだとpass1409)



```
iu@bielinux[DRR054113] pwd [ 8:41午後 ]  
/home/iu/Documents/DRR054113  
① iu@bielinux[DRR054113] sudo apt-get install sra-toolkit  
[sudo] password for iu: ②
```

W4-3: apt-get

①Do you want to continue?と聞かれるので、y。イチイチ聞かれたくない場合は、「sudo apt-get -y install sra-toolkit」と-yオプションをつけておけばよい(第4回W15-1)

```
iu@bielinux[~/Documents/DRR054113] 20:48
Reading state information... Done
The following packages were automatically installed and are no longer required:
linux-headers-3.13.0-55 linux-headers-3.13.0-55-generic
linux-headers-3.13.0-68 linux-headers-3.13.0-68-generic
linux-headers-3.13.0-71 linux-headers-3.13.0-71-generic
linux-image-3.13.0-55-generic linux-image-3.13.0-68-generic
linux-image-3.13.0-71-generic linux-image-extra-3.13.0-55-generic
linux-image-extra-3.13.0-68-generic linux-image-extra-3.13.0-71-generic
Use 'apt-get autoremove' to remove them.
The following extra packages will be installed:
sra-toolkit-libs0
The following NEW packages will be installed:
sra-toolkit sra-toolkit-libs0
0 upgraded, 2 newly installed, 0 to remove and 138 not upgraded.
Need to get 2,311 kB of archives.
After this operation, 6,065 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
```



W4-3: apt-get

①インストール完了後の状態。2016年06月19日に気づいたこととして、赤下線部分とこの後のスライドを眺めればわかるが、バージョン番号がかなり古い。(既に校正も終わっている)2016年7月公開予定の原稿PDFではapt-getの手順を推奨しているが、“最新版”をインストールしたい場合はwgetでやりましょう



```
used.
Do you want to continue? [Y/n] y
Get:1 http://jp.archive.ubuntu.com/ubuntu/ sra-toolkit-libs0 amd64 2.1.7a-1ubuntu2 [924 kB]
Get:2 http://jp.archive.ubuntu.com/ubuntu/ trusty/universe sra-toolkit amd64 2.1.7a-1ubuntu2 [1,387 kB]
Fetched 2,311 kB in 0s (5,838 kB/s)
Selecting previously unselected package sra-toolkit-libs0.
(Reading database ... 439956 files and directories currently installed.)
Preparing to unpack .../sra-toolkit-libs0_2.1.7a-1ubuntu2_amd64.deb ...
Unpacking sra-toolkit-libs0 (2.1.7a-1ubuntu2) ...
Selecting previously unselected package sra-toolkit.
Preparing to unpack .../sra-toolkit_2.1.7a-1ubuntu2_amd64.deb ...
Unpacking sra-toolkit (2.1.7a-1ubuntu2) ...
Setting up sra-toolkit-libs0 (2.1.7a-1ubuntu2) ...
Setting up sra-toolkit (2.1.7a-1ubuntu2) ...
Processing triggers for libc-bin (2.19-0ubuntu6.7) ...
iu@bielinux[DRR054113] [ 8:48午後 ]
```



W4-4: 確認

①インストール作業は「~/Documents/DRR054113」で行ったが、「sudo apt-get install **ソフトウェア名**」でやる場合は、基本的にどの作業ディレクトリ上でもよい。②インストール後は、fastq-dumpを使用可能。③パスも既に通されている。

```
File Edit View Search Terminal Help
① iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
② iu@bielinux[DRR054113] fastq-dump [12:31午後]

Usage:
fastq-dump [options] [ -A ] <accession>
fastq-dump [options] <path [path...]>

Use option --help for more information

fastq-dump : 2.1.7

③ iu@bielinux[DRR054113] where fastq-dump [12:31午後]
/usr/bin/fastq-dump
iu@bielinux[DRR054113] [12:31午後]
```

W4-5: パスが通っている

SRA Toolkit Installation and Configuration Guide

Table of Contents

1. [Downloading and installing the SRA Toolkit](#)
2. [Testing the Toolkit configuration](#)
3. [Configuring the Toolkit](#)
4. [Links and help documents](#)

Contact: sra-tools@ncbi.nlm.nih.gov

The following guide will outline the steps for the installation and usage of individual tools in the SRA Toolkit.

The NCBI SRA Toolkit can be used to process data in the .sra format (Note that the .sra format may be compiled with GC content). We recommend using these procedures.

Downloading and installing the SRA Toolkit

Download the Toolkit from the following links:

1. If you are using a web browser, click on the supported platform link.
2. If you are instead using a command-line interface, use the following command to download the Toolkit to the directory: `"/ftp-trace.ncbi.nlm.nih.gov/sra/sdk/centos64/".`

```
wget "http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/centos64/centos64_linux64.tar.gz"
```

Unpack the Toolkit:

1. For Linux, use tar:

```
tar -xzf sratoolkit.current-centos_linux64.tar.gz
```



「`sudo apt-get install sra-toolkit`」で無事インストール完了したあとは、W4-4で示したようにパスを通し終わった状態。それゆえ①SRA Toolkit Installation and Configuration Guide中の、②パスに関する注意書きは、気にしなくてもよい。

Unpack the Toolkit:

1. For Linux, use tar:

```
tar -xzf sratoolkit.current-centos_linux64.tar.gz
```

2. For Mac OS X, double-click on the .tar.gz file and the Archive Utility will unpack it. Alternatively, command-line tar will also work (see Linux example, above).
3. For Windows, either use an archiving and compression utility (e.g., Winzip, 7-Zip, etc.), or simply double-click on the .zip file and drag the 'sratoolkit...' folder to the preferred install location.

Note: For most users, the Toolkit functions (fastq-dump, sam-dump, etc.) will not be located in their [PATH environmental variable](#). This may require providing directory information about the location of the Toolkit. See the below examples for how 'fastq-dump' would be called in different circumstances:

- `~/[user_name]/sra-toolkit/fastq-dump`
YES: The Toolkit "bin" directory has been placed in the user-specified directory "sra-toolkit"
- `./fastq-dump`
YES: The Toolkit components are in the current working directory
- `fastq-dump`
NO: If the toolkit location is not specified in your \$PATH variable, then the OS cannot locate the fastq-dump program, even if it is in the current directory. NOTE: Windows users should be able to enter only "fastq-dump.exe" if you have navigated to the Toolkit "bin" directory.

W5-1: おさらいと準備

①W3-7で作成したhoge3.txtを削除。②
(fastq-dump実行時に上書きされるのを防ぐ
ため)ファイル名変更。③DRAから直接ダウ
ンロードしたFASTQファイルの行数が3660
であることを確認(リード数は3660/4 = 915)

```
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l
total 1389824
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra
-rw-rw-r-- 1 iu iu 5541 3月 22 12:40 hoge3.txt
iu@bielinux[DRR054113] rm -f hoge3.txt
iu@bielinux[DRR054113] mv DRR054113.fastq.bz2 DRR054113_DRA.fastq.bz2
iu@bielinux[DRR054113] mv DRR054113.fastq.gz DRR054113_DRA.fastq.gz
iu@bielinux[DRR054113] bzip2 -dc DRR054113_DRA.fastq.bz2 | wc
3660 5490 5264936
iu@bielinux[DRR054113] gzip -dc DRR054113_DRA.fastq.gz | wc
3660 5490 5264936
iu@bielinux[DRR054113]
```



W5-2: DRAの手順で...

①DRAのFASTQファイル作成手順通りに fastq-dump を実行してみる。入力は PacBio データなので single-end 扱いのはずだが、結論としては②のオプションによって、計3つのファイルが新規作成される

DRA で公開されている fastq のリード数が生データのそれよりも少ないのは何故でしょうか?

DRA では NCBI SRA Toolkit に含まれている fastq-dump を使い、以下のオプションで生データである SRA ファイルから fastq ファイルを作成しています。

①

```
fastq-dump -M 25 -E --skip-technical --split-3 -W <SRA file>
```

- -M 25: 25 塩基以上の配列のみを含める。デフォルトは 25。
- -E: リードの開始、もしくは終わりに 10 以上の N が存在しない
- --skip-technical: technical read を除き biological read のみを出力
- --split-3: ペアリードで最初と二番目の biological read をそれぞれ *_1.fastq と *_2.fastq として出力する。一つしか biological read が存在しない場合、*.fastq として出力する。
- -W: 指定されていた場合、left と right を clip する

②

上記の出力条件でリードがフィルタリング、トリミングされるため、一般的に fastq のリード数は SRA ファイルのそれよりも少なくなっています。フィルタリング、トリミングされていない fastq ファイルを得るには以下のコマンドで fastq を生成します。

```
fastq-dump -M 1 --split-3 <SRA file>
```

作成日: 2013年10月8日; 最終更新日: 2014年6月6日

<http://trace.ddbj.nig.ac.jp/dra/faq.html#read-number-fastq>

W5-2: DRAの手順で...

①fastq-dumpを実行。約2分。実行後は赤枠で示す3つのファイルが作成される。②がpaired-endリードで、③がpairのないリード、ということになるが、PacBioデータでなぜこんな結果になるのかは意味不明。③のファイルサイズが②より大きいのは妥当

```
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l
total 1389816
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
① iu@bielinux[DRR054113] fastq-dump -M 25 -E --skip-technical --split
-3 -W ./DRR054113.sra
Written 41793 spots for ./DRR054113.sra
Written 41793 spots total
iu@bielinux[DRR054113] ls -l
total 1732152
-rw-rw-r-- 1 iu iu    44401670  3月 23 16:48 DRR054113_1.fastq
-rw-rw-r-- 1 iu iu    23673718  3月 23 16:48 DRR054113_2.fastq
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 282463422  3月 23 16:48 DRR054113.fastq
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] █
```

[4:48午後]

[4:48午後]



Paired-endファイルのリードIDを眺めている。最初は赤下線のDRR054113.753というID。長さも異なっていて、なんだかよくわからないがうまく分割されているようだ

W5-3: 確認

```
File Edit View Search Terminal Help
iu@bielinux[DRR054113] pwd [ 4:49午後 ]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l *.fastq [ 4:49午後 ]
-rw-rw-r-- 1 iu iu 44401670 3月 23 16:48 DRR054113_1.fastq
-rw-rw-r-- 1 iu iu 23673718 3月 23 16:48 DRR054113_2.fastq
-rw-rw-r-- 1 iu iu 282463422 3月 23 16:48 DRR054113.fastq
① iu@bielinux[DRR054113] grep "@" DRR054113_1.fastq | head -n 4
@DRR054113.753 length=3654
@DRR054113.1452 length=2489
@DRR054113.2205 length=10147
@DRR054113.2581 length=8802
② iu@bielinux[DRR054113] grep "@" DRR054113_2.fastq | head -n 4
@DRR054113.753 length=701
@DRR054113.1452 length=2421
@DRR054113.2205 length=2389
@DRR054113.2581 length=5401
iu@bielinux[DRR054113] [ 4:49午後 ]
```

W5-3: 確認

③pairのないリードファイルについて、最初の14個分のリードIDを表示。④IDのシリアル番号で、744の次は763となっている。②のpairedのほうのシリアル番号とは、重なりはなさそう

```
File Edit View Search Terminal Help 16:51
iu@bielinux[DRR054113] grep "@" DRR054113_2.fastq | head -n 4
@DRR054113.753 length=701
@DRR054113.1452 length=2421
@DRR054113.2205 length=2389
@DRR054113.2581 length=5401
iu@bielinux[DRR054113] grep "@" DRR054113.fastq | head -n 14
@DRR054113.111 length=86
@DRR054113.227 length=11592
@DRR054113.298 length=13611
@DRR054113.392 length=1346
@DRR054113.393 length=478
@DRR054113.488 length=484
@DRR054113.643 length=7134
@DRR054113.669 length=6653
@DRR054113.692 length=6385
@DRR054113.733 length=1184
@DRR054113.744 length=3232
@DRR054113.763 length=99
@DRR054113.764 length=2546
@DRR054113.783 length=3282
iu@bielinux[DRR054113]
```

[4:50午後]

W5-3: 確認

①pairedのほうの行数は22,532。リード数は $22,532/4 = 5,633$ 個。②unpairedのほうの行数は144,640。リード数は $144,640/4 = 36,160$ 個。どう転んでもDRAから直接ダウンロードするFASTQファイルのリード数(=915個)よりも多いが、理由は不明

```
iu@bielinux[DRR054113] pwd [ 4:52午後 ]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l *.fastq [ 4:52午後 ]
-rw-rw-r-- 1 iu iu 44401670 3月 23 16:48 DRR054113_1.fastq
-rw-rw-r-- 1 iu iu 23673718 3月 23 16:48 DRR054113_2.fastq
-rw-rw-r-- 1 iu iu 282463422 3月 23 16:48 DRR054113.fastq
iu@bielinux[DRR054113] wc *.fastq [ 4:52午後 ]
 22532   33798 44401670 DRR054113_1.fastq
 22532   33798 23673718 DRR054113_2.fastq
144640  216960 282463422 DRR054113.fastq
189704  284556 350538810 total
iu@bielinux[DRR054113] █ [ 4:52午後 ]
```



W5-4: DRAのFAQ再訪

①この記述内容は最終更新が2014年6月。しかもIlluminaデータを想定(していると言える根拠は、この質問は門田が投げたものだから)。PacBioデータに対してこのオプションを使っているかどうかはそもそも不明だし、最終更新日時も古いので参考程度にしておいたほうがいいのかもかもしれない

DRA で公開されている fastq のリード数が生データのそれよりも少ないのは何故でしょうか?

DRA では NCBI SRA Toolkit に含まれている fastq-dump を使い、以下のオプションで生データである SRA ファイルから fastq ファイルを作成しています。

```
fastq-dump -M 25 -E --skip-technical --split-3 -W <SRA file>
```

- -M 25: 25 塩基以上の配列のみを含める。デフォルトは 25。
- -E: リードの開始、もしくは終わりに 10 以上の N が存在しない
- --skip-technical: technical read を除き biological read のみを出力
- --split-3: ペアリードで最初と二番目の biological read をそれぞれ *_1.fastq と *_2.fastq として出力する。一つしか biological read が存在しない場合、*.fastq として出力する。
- -W: 指定されていた場合、left と right を clip する

上記の出力条件でリードがフィルタリング、トリミングされるため、一般的に fastq のリード数は SRA ファイルのそれよりも少なくなっています。フィルタリング、トリミングされていない fastq ファイルを得るには以下のコマンドで fastq を生成します。

```
fastq-dump -M 1 --split-3 <SRA file>
```

作成日: 2013年10月8日; 最終更新日: 2014年6月6日



<http://trace.ddbj.nig.ac.jp/dra/faq.html#read-number-fastq>

W5-5: --split-3抜きで...

①W5-2で作成した3つの.fastqファイルを削除。②DRAのFASTQ作成手順から--split-3オプションのみ外して再度実行。約2分。③1つのFASTQファイルのみ作成された。

```
iu@bielinux[DRR054113] pwd [ 5:12午後 ]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls [ 5:12午後 ]
DRR054113_1.fastq DRR054113_DRA.fastq.bz2 DRR054113.fastq
DRR054113_2.fastq DRR054113_DRA.fastq.gz DRR054113.sra
① iu@bielinux[DRR054113] rm -f *.fastq [ 5:12午後 ]
iu@bielinux[DRR054113] ls -l [ 5:12午後 ]
total 1389816
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra
② iu@bielinux[DRR054113] fastq-dump -M 25 -E --skip-technical -W ./DR
R054113.sra
Written 42038 spots for ./DRR054113.sra
Written 42038 spots total
iu@bielinux[DRR054113] ls -l [ 5:12午後 ] ←
total 1744120
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 362806102 3月 23 17:12 DRR054113.fastq ③
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] [ 5:12午後 ]
```

W5-5: --split-3抜きで...

①行数は168,152。リード数は168,152/4 = 42,038個。一方、DRAのウェブ上で見られる数値は163,482リード(W2-5)なので、②赤下線のオプションの効果でトリミングやフィルタリングがかかっているのであろう。

```
iu@bielinux[DRR054113] ls
DRR054113_1.fastq  DRR054113_DRA.fastq.bz2  DRR054113.fastq
DRR054113_2.fastq  DRR054113_DRA.fastq.gz  DRR054113.sra
iu@bielinux[DRR054113] rm -f *.fastq [ 5:12午後]
iu@bielinux[DRR054113] ls -l [ 5:12午後]
total 1389816
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] fastq-dump -M 25 -E --skip-technical -W ./DR
R054113.sra
Written 42038 spots for ./DRR054113.sra
Written 42038 spots total
iu@bielinux[DRR054113] ls -l [ 5:12午後]
total 1744120
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 362806102 3月 23 17:12 DRR054113.fastq
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] wc DRR054113.fastq [ 5:12午後]
168152 252228 362806102 DRR054113.fastq
iu@bielinux[DRR054113] [ 5:13午後]
```



W5-6: デフォルトで実行

①W5-5で作成した.fastqファイルを削除。②オプション無指定のデフォルトで再度実行。約2分。③1つのFASTQファイルのみ作成された

```
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls
DRR054113_DRA.fastq.bz2  DRR054113.fastq
DRR054113_DRA.fastq.gz  DRR054113.sra
iu@bielinux[DRR054113] rm -f *.fastq
iu@bielinux[DRR054113] ls -l
total 1389816
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] fastq-dump ./DRR054113.sra
Written 163380 spots for ./DRR054113.sra
Written 163380 spots total
iu@bielinux[DRR054113] ls -l
total 2102944
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu   730238332  3月 23 17:15 DRR054113.fastq
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113]
```



W5-6: デフォルトで実行

①行数は653,520。リード数は653,520/4 = 163,380個。このリード数情報は、②を眺めるのでもわかる。③赤下線のオプションなし効果でトリミングやフィルタリングがほとんどかかっていないため、DRAのウェブ上で見られる数値(163,482リード; W2-5)とほぼ同じになっていることがわかる。

```
File Edit View Search Terminal Help
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls
DRR054113_DRA.fastq.bz2  DRR054113.fastq
DRR054113_DRA.fastq.gz  DRR054113.sra
iu@bielinux[DRR054113] rm -f *.fastq
iu@bielinux[DRR054113] ls -l
total 1389816
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] fastq-dump ./DRR054113.sra
Written 163380 spots for ./DRR054113.sra
Written 163380 spots total
iu@bielinux[DRR054113] ls -l
total 2102944
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu  730238332  3月 23 17:15 DRR054113.fastq
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] wc DRR054113.fastq
653520    980280 730238332 DRR054113.fastq
iu@bielinux[DRR054113] █
```

[5:14午後]

[5:14午後]

[5:14午後]

[5:15午後]

[5:15午後]

[5:19午後]

W5-7: --gzip

① --gzip オプションをつけると、② 出力ファイルが gzip 圧縮された状態になるのでおススメ。-M 1 オプションは、1塩基以上の長さの配列を出力せよ、という意味です。③ リード数は何も指定していないときと同じ163,380個なので無指定のときと同じ結果になることを確認しただけになります。ここでは示しませんが、--bzip2 オプションもあります。

```
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls
DRR054113_DRA.fastq.bz2  DRR054113.fastq
DRR054113_DRA.fastq.gz  DRR054113.sra
iu@bielinux[DRR054113] rm -f *.fastq
iu@bielinux[DRR054113] ls -l
total 1389816
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
① iu@bielinux[DRR054113] fastq-dump -M 1 --gzip ./DRR054113.sra
Written 163380 spots for ./DRR054113.sra
Written 163380 spots total
iu@bielinux[DRR054113] ls -l
total 1690224
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 307610739  3月 23 17:35 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] █
```

[5:33午後]

[5:33午後]

[5:35午後]

[5:35午後]

W5-8: 入力ファイル

これまで特に述べてこなかったが、①入力ファイル名の前に./を付け忘れないようにしましょう。これは実質的にfastq-dump特有の指定法。普通のプログラムは、作業ディレクトリ中のファイルを自動で見に行ってくれるので./をつけなくてもよい。しかしfastq-dumpの場合は、「このディレクトリ上にある」を意味する「./」をつけないと動作しない。

```
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls
DRR054113_DRA.fastq.bz2  DRR054113.fastq
DRR054113_DRA.fastq.gz  DRR054113.sra
iu@bielinux[DRR054113] rm -f *.fastq
iu@bielinux[DRR054113] ls -l
total 1389816
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] fastq-dump -M 1 --gzip ./DRR054113.sra
Written 163380 spots for ./DRR054113.sra
Written 163380 spots total
iu@bielinux[DRR054113] ls -l
total 1690224
-rw-rw-r-- 1 iu iu    2392259  3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu    2720482  3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 307610739  3月 23 17:35 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334  3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] █
```

[5:33午後]
[5:33午後]
[5:35午後]
[5:35午後]



W6-1 : FastQC

163,380リードからなるDRR054113.fastq.gzを入力としてFastQC (ver. 0.11.4)を実行。W3-2と違って出力先を指定していないので、結果ファイルはカレントディレクトリ上に作成される

```
iu@bielinux[DRR054113] pwd [12:48午後]
/home/iu/Documents/DRR054113
iu@bielinux[DRR054113] ls -l [12:48午後]
total 1690224
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 307610739 3月 23 17:35 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] fastqc2 -v [12:48午後]
FastQC v0.11.4
iu@bielinux[DRR054113] fastqc2 -q DRR054113.fastq.gz [12:48午後]
iu@bielinux[DRR054113] ls -l [12:49午後]
total 1690836
-rw-rw-r-- 1 iu iu 2392259 3月 22 12:32 DRR054113_DRA.fastq.bz2
-rw-rw-r-- 1 iu iu 2720482 3月 22 12:38 DRR054113_DRA.fastq.gz
-rw-rw-r-- 1 iu iu 296994 3月 24 12:49 DRR054113_fastqc.html
-rw-rw-r-- 1 iu iu 325042 3月 24 12:49 DRR054113_fastqc.zip
-rw-rw-r-- 1 iu iu 307610739 3月 23 17:35 DRR054113.fastq.gz
-rw-rw-r-- 1 iu iu 1418046334 3月 22 15:23 DRR054113.sra
iu@bielinux[DRR054113] [12:49午後]
```



W6-2: 改名して移動

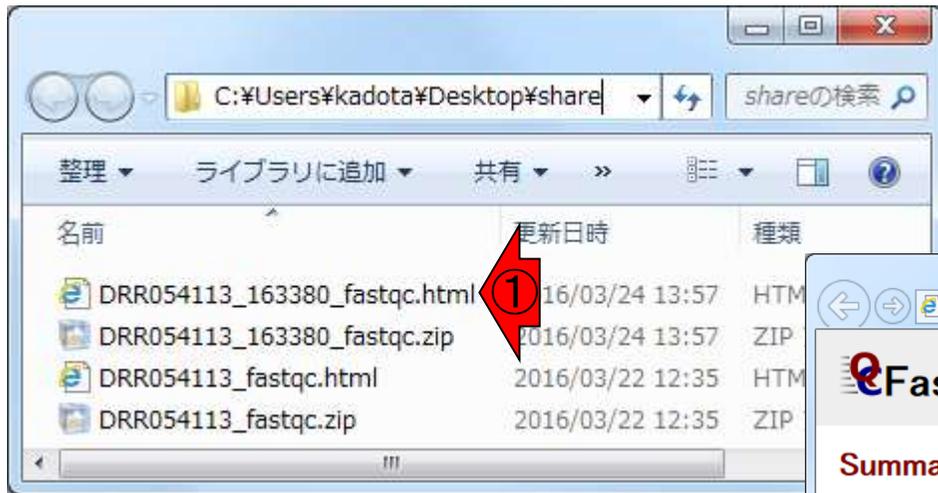
①163,380リードからなるDRR054113.fastq.gz
を入力としてFastQC (ver. 0.11.4)を実行した
結果ファイルと同じ名前のものが②共有フォル
ダ内に存在する。これはW3-2で作成。そ
のため、③mvコマンドで共有フォルダに移動
させる際に④_163380を追加して改名している

```
iu@bielinux[DRR054113] pwd
/home/iu/Documents/DRR054113
① iu@bielinux[DRR054113] ls -l *fastqc*
-rw-rw-r-- 1 iu iu 296994  3月 24 12:49 DRR054113_fastqc.html
-rw-rw-r-- 1 iu iu 325042  3月 24 12:49 DRR054113_fastqc.zip
② iu@bielinux[DRR054113] ls -l ~/Desktop/mac_share [ 1:57午後 ]
total 751
-rwxrwxrwx 1 iu iu 392406  3月 22 12:35 DRR054113_fastqc.html
-rwxrwxrwx 1 iu iu 375781  3月 22 12:35 DRR054113_fastqc.zip
iu@bielinux[DRR054113] mv DRR054113_fastqc.html ~/Desktop/mac_share
/DRR054113_163380_fastqc.html
iu@bielinux[DRR054113] mv DRR054113_fastqc.zip ~/Desktop/mac_share/
DRR054113_163380_fastqc.zip
③
iu@bielinux[DRR054113] ls -l ~/Desktop/mac_share [ 1:57午後 ]
total 1359
-rwxrwxrwx 1 iu iu 296994  3月 24 13:57 DRR054113_163380_fastqc.htm
l
-rwxrwxrwx 1 iu iu 325042  3月 24 13:57 DRR054113_163380_fastqc.zip
-rwxrwxrwx 1 iu iu 392406  3月 22 12:35 DRR054113_fastqc.html
-rwxrwxrwx 1 iu iu 375781  3月 22 12:35 DRR054113_fastqc.zip
iu@bielinux[DRR054113] █ [ 1:57午後 ]
④
```



①ホストOS(ここではWindows)上でFastQC実行結果ファイルを眺める

W6-3: 結果を眺める



The screenshot shows a web browser displaying a FastQC Report. The report title is 'FastQC Report' and the date is 'Thu 24 Mar 2016'. The report is for the file 'DRR054113.fastq.gz'.

Summary

- Basic Statistics (Green checkmark)
- Per base sequence quality (Red X)
- Per sequence quality scores (Red X)
- Per base sequence content (Yellow exclamation mark)
- Per sequence GC content (Green checkmark)
- Per base N content (Green checkmark)
- Sequence Length Distribution (Yellow exclamation mark)
- Sequence Duplication Levels (Green checkmark)
- Overrepresented sequences (Green checkmark)
- Adapter Content (Green checkmark)
- Kmer Content (Red X)

Basic Statistics

Measure	Value
Filename	DRR054113.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	163380
Sequences flagged as poor quality	0
Sequence length	116-28874
%GC	44

Per base sequence quality

Quality scores across all bases (5)

Produced by [FastQC](#) (version 0.11.4)

W6-3: 結果を眺める

①入力ファイル。②リード数は163,380、③配列長は116-28874 bpの範囲であることがわかる

FastQC Report Thu 24 Mar 2016
DRR054113.fastq.gz

Summary

- ✓ [Basic Statistics](#)
- ✗ [Per base sequence quality](#)
- ✗ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ! [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)

✓ Basic Statistics

Measure	Value
Filename	DRR054113.fastq.gz ①
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	163380 ②
Sequences flagged as poor quality	0
Sequence length	116-28874 ③
%GC	44

W6-3: 結果を眺める

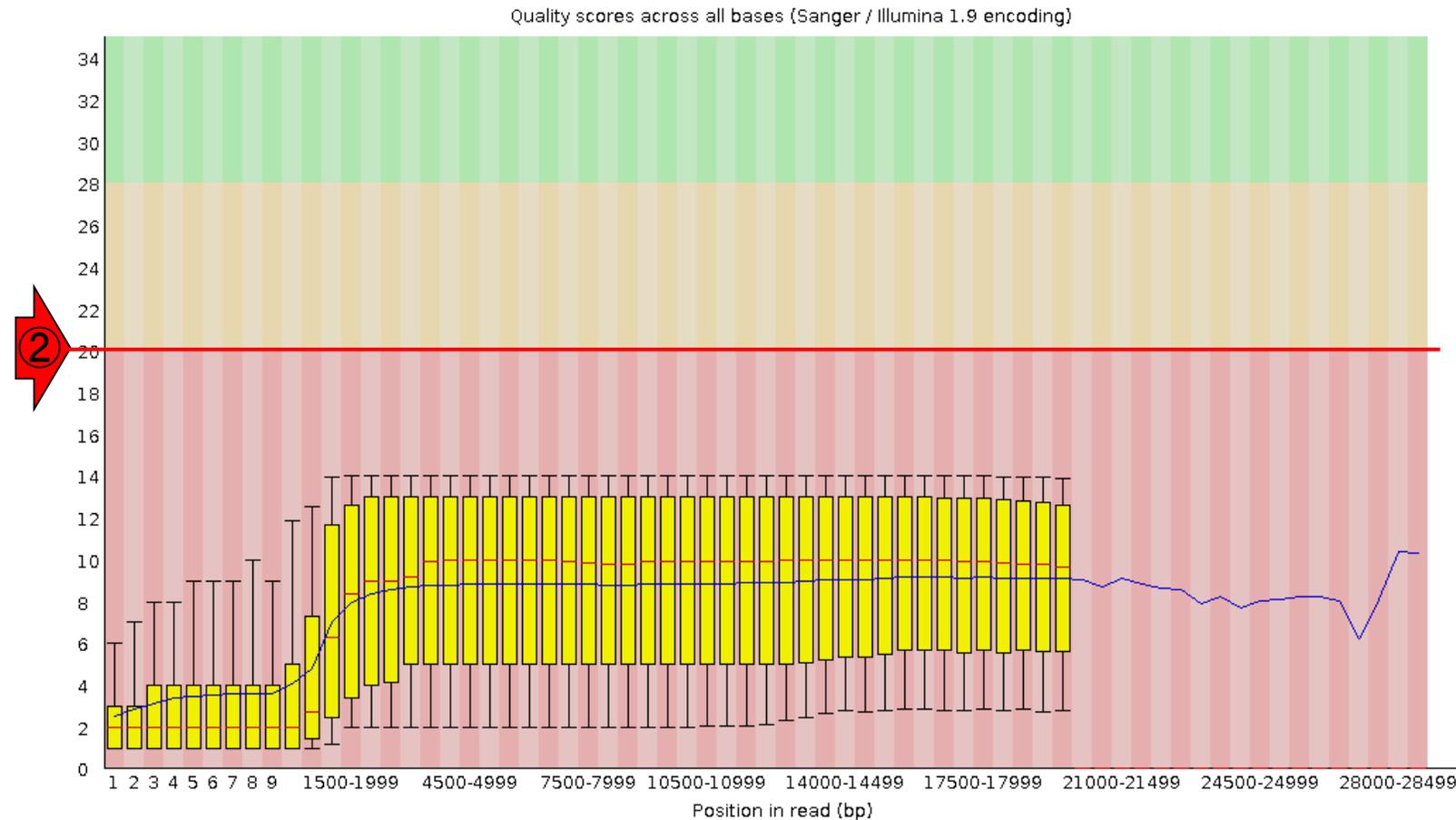
① Per base sequence quality. この図の縦軸はクオリティスコア。② 赤線のスコア20を基準としてみると、DRAから直接ダウンロードした915リードからなるFASTQファイルのFastQC実行結果(W3-3)と比べて、明らかに低くなっていることがわかる。

FastQC Report

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality ①
- ✗ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

✗ Per base sequence quality



W6-4: Rで計算

塩基配列決定精度(エラー率)からクオリティスコアをRで計算する。①エラー率13%のときは、スコア8.86となるので、実際のスペック通りで安心

```
iu@bielinux[DRR054113] R -q [ 4:32午後]
> -10*log10(0.01) #エラー率1%のときのクオリティスコア
[1] 20
> -10*log10(0.10) #エラー率10%のときのクオリティスコア
[1] 10
① > -10*log10(0.13) #エラー率13%のときのクオリティスコア
[1] 8.860566
> q(save="no")
iu@bielinux[DRR054113] [ 4:33午後]
```



W6-5: 配列長分布

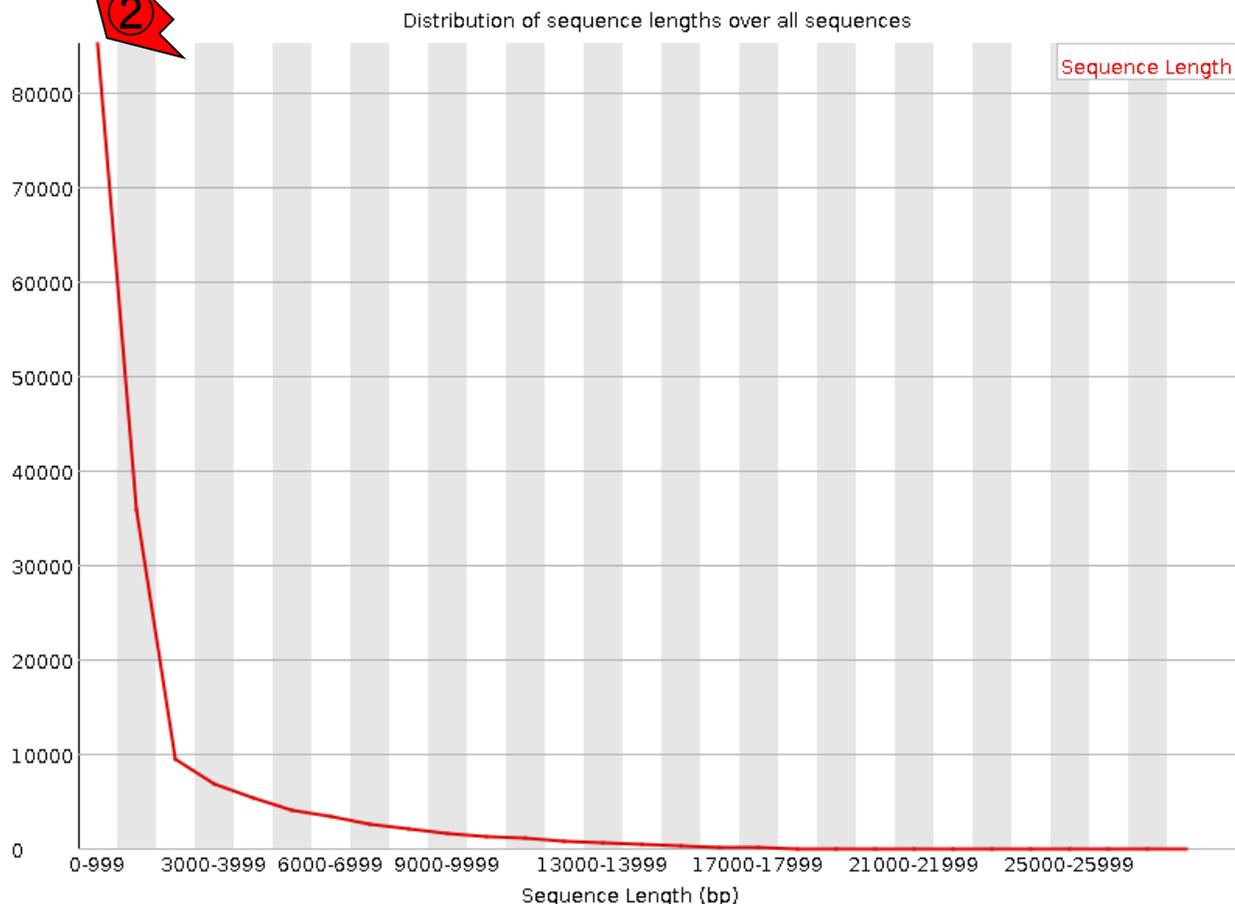
①配列長分布(Sequence Length Distribution)。短いものが多いため評価しづらいが、平均すると数千bp程度の長さを読めているのだろう。②163,380リードの半分以上が1,000 bp未満だと判断

FastQC Report

Summary

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

! Sequence Length Distribution



Rを起動し、①赤枠内をコピペ。Rでも②配列長の最短と最長、および③全リード数を得ることができる

W6-6: Rで計算

```
File Edit View Search Terminal Help
t,
rownames, sapply, setdiff, sort, table, tapply, union, unique
unlist, unsplit

Loading required package:
Loading required package:
Loading required package:
Loading required package:
Creating a generic function for 'S4Vectors'
Loading required package:
Loading required package: GenomicAlignments
> fastq <- readFastq(in_f)
> range(width(fastq))
[1] 116 28874
> length(width(fastq))
[1] 163380
>
```

```
• W6-6: Rで計算

pwd
ls -l
R -q

in_f <- "DRR054113.fastq.gz" #入力ファイル名を指定してin_fに格納
library(ShortRead) #パッケージの読み込み
fastq <- readFastq(in_f) #ファイルの読み込み
range(width(fastq)) #配列長の最短と最長を表示
length(width(fastq)) #全リード数を表示

sum(width(fastq) < 1000) #1000 bp未満のリード数を表示
sum(width(fastq) < 1000)/length(width(fastq)) #1000 bp未満のリード数の割合を表示

sum(width(fastq) > 10000) #10000 bpより長いリード数を表示
sum(width(fastq) > 10000)/length(width(fastq)) #10000 bpより長いリード数の割合を表示

q(save="no")
```



①赤枠内をコピペ。②1000 bp未満のリード数は85,134個で、③その割合は52.1%

W6-6: Rで計算

```
File Edit View Search Terminal Help
Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for method 'S4Vectors'
Loading required package:
> fastq <- readFastq(in_f)
> range(width(fastq))
[1] 116 28874
> length(width(fastq))
[1] 163380
> sum(width(fastq) < 1000)
[1] 85134
> sum(width(fastq) < 1000)/length(width(fastq))#1000 bp未満のリード
数の割合を表示
[1] 0.5210797
>
```

W6-6: Rで計算

```
pwd
ls -l
R -q

in_f <- "DRR054113.fastq.gz"
library(ShortRead)
fastq <- readFastq(in_f)
range(width(fastq))
length(width(fastq))

sum(width(fastq) < 1000)
sum(width(fastq) < 1000)/length(width(fastq))#1000 bp未満のリード数の割合を表示

sum(width(fastq) > 10000)
sum(width(fastq) > 10000)/length(width(fastq))#10000 bpより長いリード数の割合を表示

q(save="no")
```

#入力ファイル名を指定してin_fに格納
#パッケージの読み込み
#ファイルの読み込み
#配列長の最短と最長を表示
#全リード数を表示
#1000 bp未満のリード数を表示
#1000 bp未満のリード数の割合を表示
#10000 bpより長いリード数を表示
#10000 bpより長いリード数の割合を表示



① `sum(width(fastq) < 1000)`
`sum(width(fastq) < 1000)/length(width(fastq))#1000 bp未満のリード数の割合を表示`

② `sum(width(fastq) < 1000)` #1000 bp未満のリード数を表示

③ `sum(width(fastq) < 1000)/length(width(fastq))#1000 bp未満のリード数の割合を表示`

①赤枠内をコピペ。②10000 bpより長いリード数は5,985個で、③その割合は3.66%

W6-6: Rで計算

```
File Edit View Search Terminal Help
Loading required package: Rsamtools
Loading required package: GenomeInfoDb
Loading required package:
Loading required package:
> fastq <- readFastq(in_f)
> range(width(fastq))
[1] 116 28874
> length(width(fastq))
[1] 163380
> sum(width(fastq) < 1000)
示
[1] 85134
> sum(width(fastq) < 1000)
数の割合を表示
[1] 0.5210797
> sum(width(fastq) > 10000)
を表示
[1] 5985
> sum(width(fastq) > 10000)/length(width(fastq))#10000 bpより長いリ
ード数の割合を表示
[1] 0.03663239
>
```

• W6-6: Rで計算

```
pwd
ls -l
R -q

in_f <- "DRR054113.fastq.gz" #入力ファイル名を指定してin_fに格納
library(ShortRead) #パッケージの読み込み
fastq <- readFastq(in_f) #ファイルの読み込み
range(width(fastq)) #配列長の最短と最長を表示
length(width(fastq)) #全リード数を表示

sum(width(fastq) < 1000) #1000 bp未満のリード数を表示
sum(width(fastq) < 1000)/length(width(fastq))#1000 bp未満のリード数の割合を表示

sum(width(fastq) > 10000) #10000 bpより長いリード数を表示
sum(width(fastq) > 10000)/length(width(fastq))#10000 bpより長いリード数の割合を表示

q(save="no")
```

#10000 bpより長いリード数

W6-6: Rで計算

```

File Edit View Search Terminal Help
Loading required package: GenomeInfoDb
Loading required package: GenomicRanges
Loading required package:
> fastq <- readFastq(in_f)
> range(width(fastq))
[1] 116 28874
> length(width(fastq))
[1] 163380
> sum(width(fastq) < 1000)
示
[1] 85134
> sum(width(fastq) < 1000)
数の割合を表示
[1] 0.5210797
> sum(width(fastq) > 1000)
を表示
[1] 5985
> sum(width(fastq) > 10000)/length(width(fastq))#10000 bpより長いリ
ード数の割合を表示
[1] 0.03663239
> q(save="no")
iu@bielinux[DRR054113]

```

• W6-6: Rで計算

```

pwd
ls -l
R -q

in_f <- "DRR054113.fastq.gz" #入力ファイル名を指定してin_fに格納
library(ShortRead) #パッケージの読み込み
fastq <- readFastq(in_f) #ファイルの読み込み
range(width(fastq)) #配列長の最短と最長を表示
length(width(fastq)) #全リード数を表示

sum(width(fastq) < 1000) #1000 bp未満のリード数を表示
sum(width(fastq) < 1000)/length(width(fastq))#1000 bp未満のリード数の割合を表示

sum(width(fastq) > 10000) #10000 bpより長いリード数を表示
sum(width(fastq) > 10000)/length(width(fastq))#10000 bpより長いリード数の割合を表示

q(save="no")

```

[8:18午後]

W6-7: SMRTbell[®]

The image shows a YouTube video player interface. At the top left is the YouTube logo with 'JP' next to it. The video content is a diagram of the SMRTbell complex, which is a circular DNA structure. A polymerase is shown at one end of the circle, and a primer is at the other. A poly-A tail connects the two ends. Labels with arrows point to 'Polymerase', 'poly-A tail', and 'Primer'. Below the diagram, the text 'SMRTbell™ complex' is displayed. The video player controls at the bottom show a progress bar at 0:15 / 1:03, along with play, next, volume, and settings icons. Below the video player, the video title 'Using MagBeads to load SMRTbells in the PacBio RS II' is visible, followed by the Pacific Biosciences logo and channel name. A red button indicates 'チャンネル登録' (Subscribe) with 726 subscribers, and the view count is 4,401.

W6-8:リードのフラグ

- Productivity0 (P0)
 - シグナルが検出限界未満のため、リードとして出力されなかったもの。その後の解析に利用されない。
- Productivity1 (P1)
 - 一分子DNAのリードデータらしき配列データ(Read Quality=75以上; RQ75)で出力されたもの。その後の解析に利用される。
- Productivity2 (P2)
 - P0およびP1以外の全て。ノイズが大きかったり、一分子として認識されなかったもの。その後の解析に利用されない。
- 一般にアプライするDNA濃度によって…
 - 濃度が低いとP0の割合が増え、高すぎるとP2の割合が増える
 - 適度な濃度にするとP1の割合は、全リードの30-40%になる。解析に使えるリード数は、例えばPacBio RSIIの場合、上限の150,292リードの約30-40%ということで、約5-6万になる

W6-8: フィルタリング

①原著論文では生リード数に関する言及はないが、P1は21.6%–32.6%で、P1リードの合計は155,039個。「RQ = 80, リード長 = 500」でリードのフィルタリング、およびアダプター除去後のサブリード数は②163,376個であった。入力配列数の約1/4

生リード数: $150,292 \times 4 \text{セル} = 601,168 \text{個}$

⇩ RQ = 80, Length = 500
でリードのフィルタリング

⇩ アダプター除去(サブリード
の作成)

サブリード数: 163,376個



① Genome sequencing and *de novo* assembly

The cells of *L. hokkaidonensis* LOOC260^T were cultured in MRS (de Man, Rogosa, and Sharpe) broth (Difco) and were harvested in the mid-logarithmic phase. The genomic DNA was extracted and purified using Qiagen Genomic-tip 500/G and Qiagen Genomic DNA Buffer Set with lysozyme (Sigma) and proteinase K (Qiagen) according to the manufacturer's instruction. PacBio SMRT whole-genome sequencing was performed using a PacBio RSII sequencer with P4-C2 chemistry. Four SMRT cells were used for sequencing, thereby yielding 163,376 adapter-trimmed reads (subreads) with an average read length of approximately 4 kbp, which corresponded to approximately 250-fold coverage. *De novo* assembly was conducted using the HGAP method based on the SMRT Analysis package 2.0, which yielded seven contigs. Independent genome sequencing using the 250-bp paired-end Illumina MiSeq system generated 5,942,620 reads, which were assembled into contigs using Platanus assembler ver 1.2 with the default settings [40]. The initial contigs

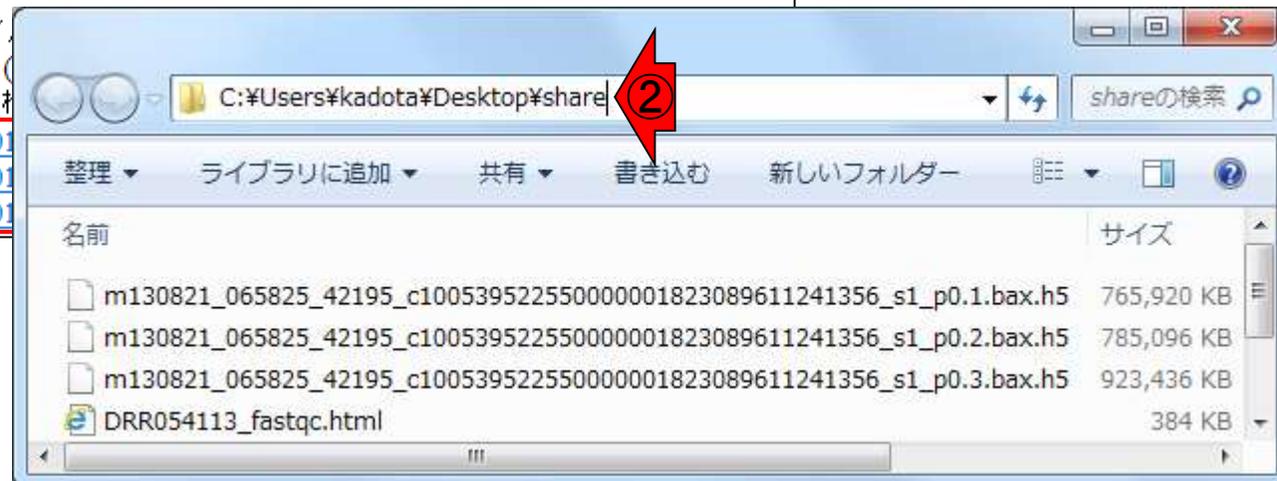


W7-1: bax.h5ファイル準備

①DDBJ Pipelineにアップロードしたいbax.h5ファイルを手元のPCにダウンロードしておく。②ここではDesktop上のshareフォルダにダウンロード。ファイルサイズは約2.4GBに達するため、それなりに時間はかかるだろう

PacBioのファイル形式とデータ解析の概要

- W1-1: PacificBiosciencesのYouTubeサイト
 - [Introduction to SMRT Sequencing](#)
 - [Single Molecule Real Time Sequencing](#)
- W2-1: PacBioデータ(原著論文中のDRR IDだが削除されている)
 - [DRR024500: Tanizawa et al., BMC Genomics, 2015](#)
- W2-3: [DRR024501](#) -> [DRP002401](#) -> [DRX022185](#)
- W2-4: [DRR024501](#) -> [DRA002643](#)
- W2-5: PacBioデータ概観
 - [DRR054113](#)
 - [DRR054114](#)
 - [DRR054115](#)
 - [DRR054116](#)
- W2-6: SMRT Portal(PacBio提供のHGAPを含む解析ソフトウェア群)の場所
 - [PacBio](#) -> [DevNet](#) -> [SMRT Analysis](#)
 - SMRT Analysis 2.3までは、HGAPを実行するためにはbax.h5ファイルが必須。
 - SMRT Analysis 3.0からは、BAMファイルが入力フォーマットになる。但しここでのBAMファイルは、マッピングデータではなく、シーケンス生データ。
 - PacBio RSIIの後継機である [Sequel](#) の出力ファイル
 - PacBioのファイル形式の説明については [こちら](#) (英語)
- W2-7: [DRR054113](#)のbax.h5ファイル(下記3ファイル合計約2.4GB)
 - [m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5](#)
 - [m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5](#)
 - [m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5](#)

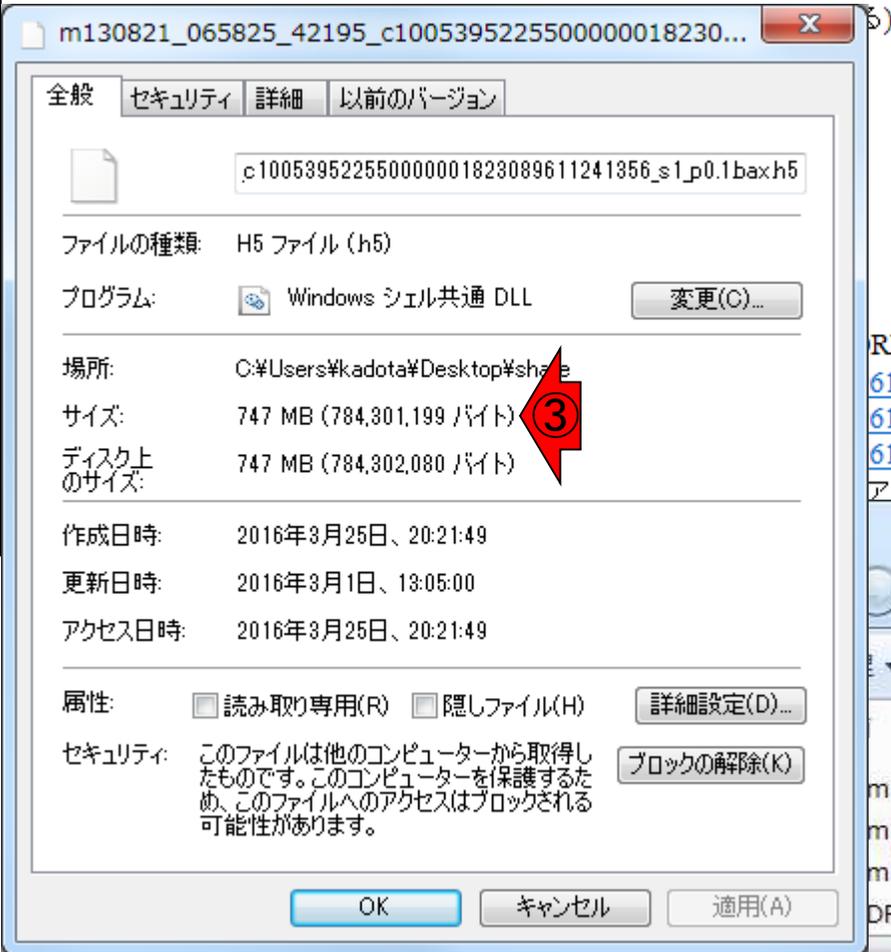


W7-1: bax.h5ファイル準備

赤下線部分の数値の微妙な違いは気にしなくてもよい。どうしても気になるヒトは、②確認したいファイル上で右クリックし、「プロパティ」。③サイズのところを眺めて安心してください

PacBioのファイル形式とデータ解析の概要

- W1-1: PacificBiosciencesのYouTubeサイト
 - [Introduction to SMRT Sequencing](#)
 - [Single Molecule Real Time Sequencing](#)



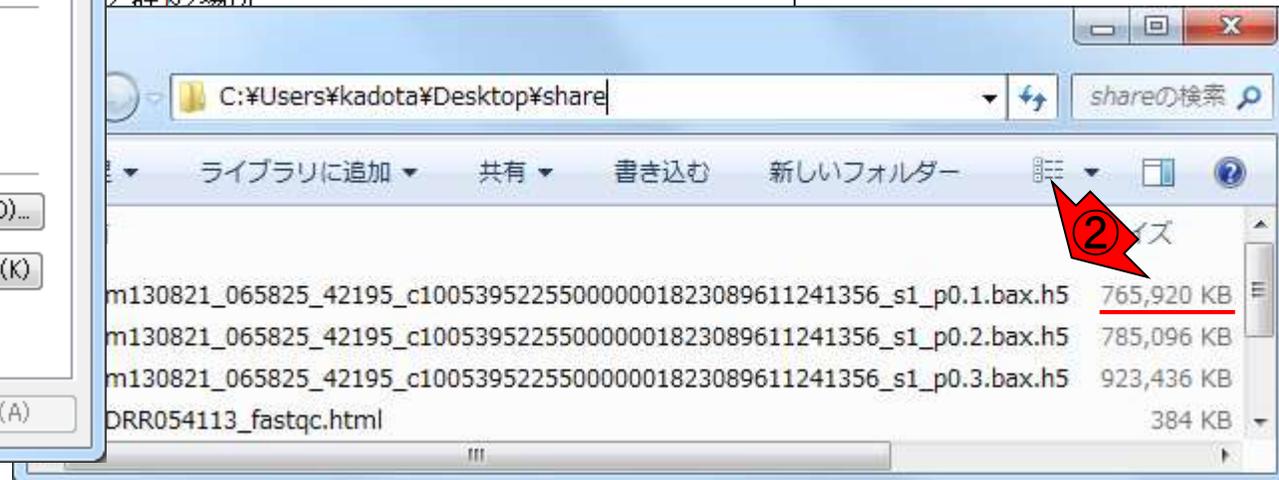
RR054113に相当)

[611241356_s1_p0.1.bax.h5](#) (747 MB; 784,301,199 bytes)

[611241356_s1_p0.2.bax.h5](#) (766 MB; 803,938,042 bytes)

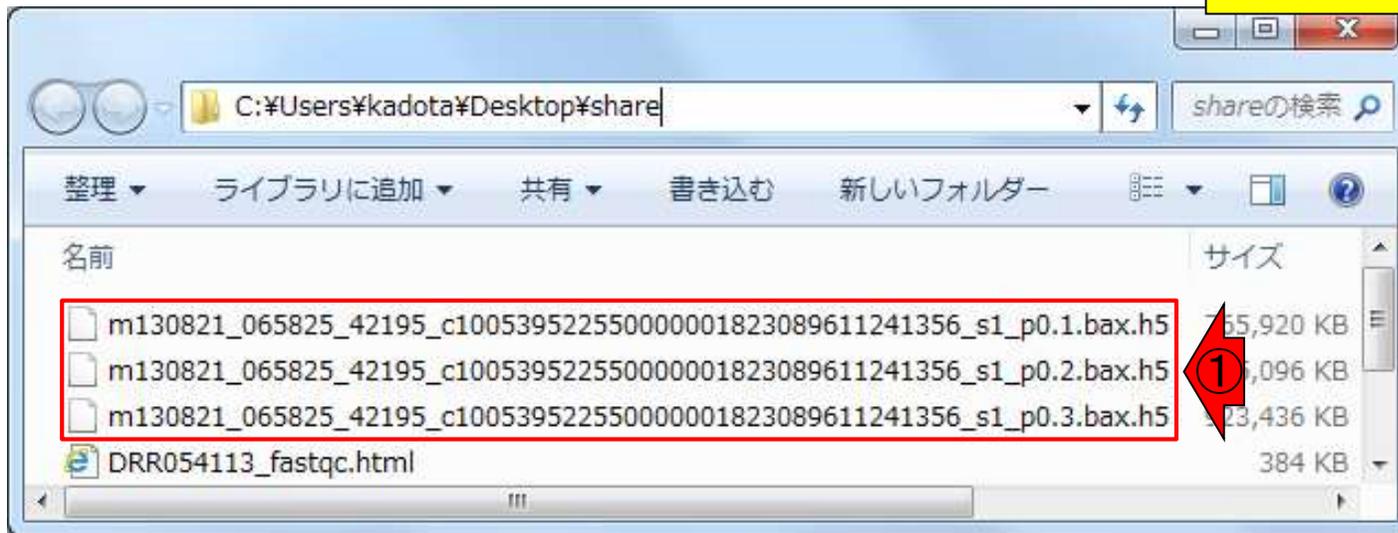
[611241356_s1_p0.3.bax.h5](#) (901 MB; 945,597,712 bytes)

ア群)の場所



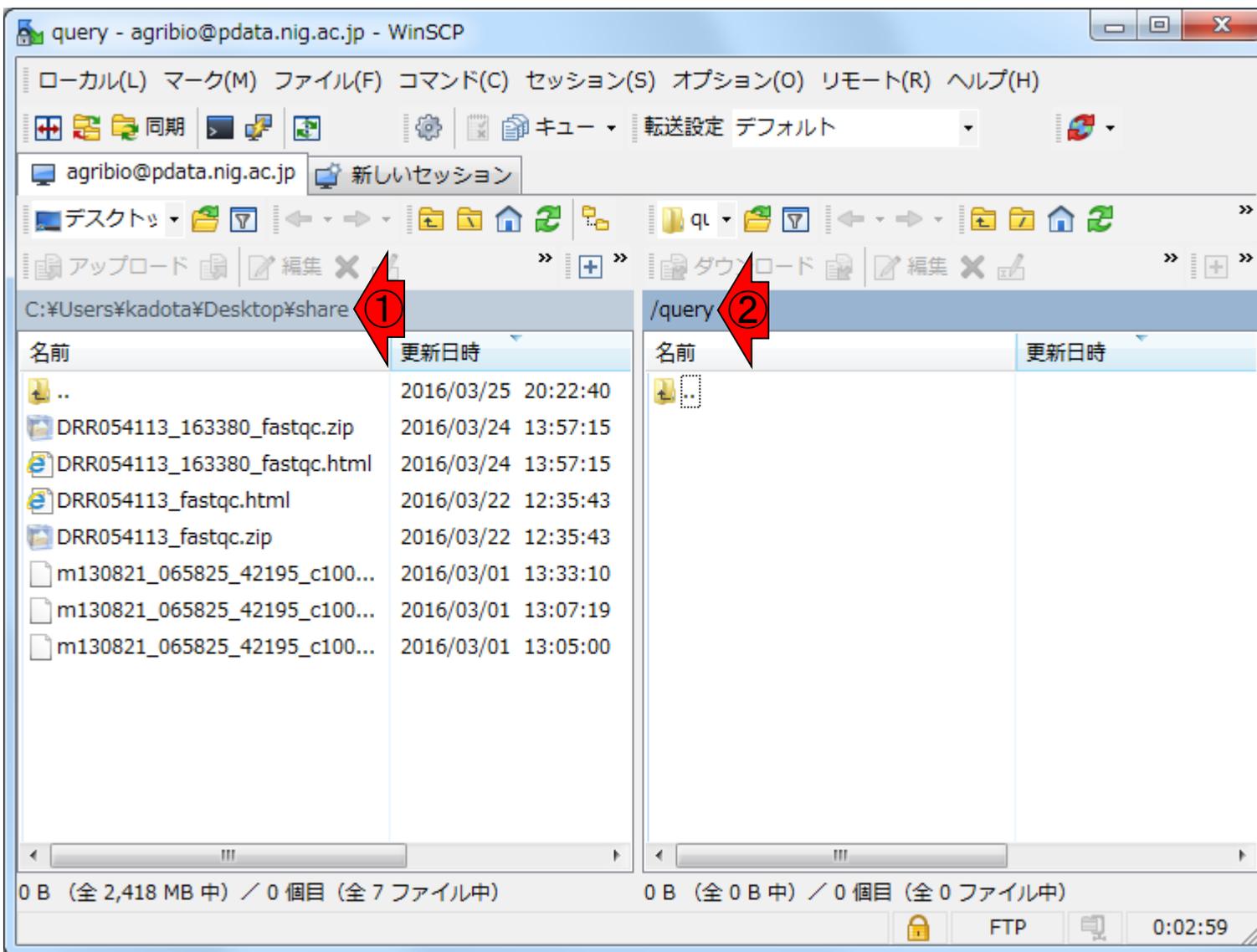
W7-2: アップロード

第6回W14を参考にしてpdata.nig.ac.jp
にログインし、①3つのbax.h5ファイルを
DDBJ Pipelineにアップロードする



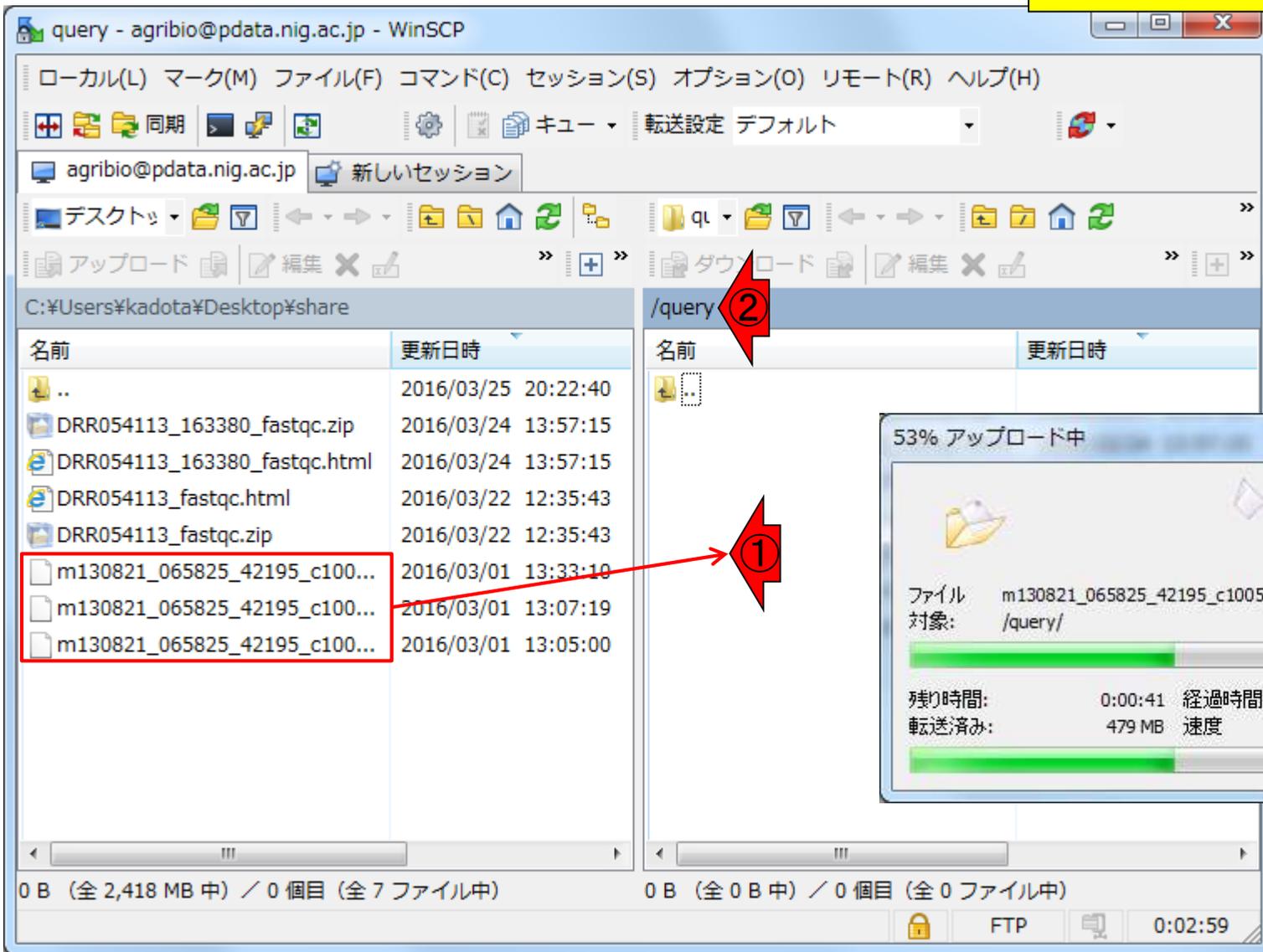
W7-2: アップロード

WinSCP上でログインし、①「デスクトップ - share」、②queryフォルダに移動した状態。



W7-2: アップロード

①アップロードしたい3つのファイルを
②queryフォルダにドラッグ&ドロップ。
アップロードもそれなりに時間がかかる



W7-3: ログイン

① ユーザIDとパスワードを打ち込んで、DDBJ Pipelineに② Login

https://p.ddbj.nig.ac.jp/pipeline/Login.do

DDBJ Read Annotation Pipeline

English Japanese

DDBJ Read Annotation Pipelineは、次世代シーケンサ配列のクラウド型データ解析プラットフォームです。

LOGIN 新規アカウント作成 ゲストとしてログイン

User ID: agriblio

Password: [masked]

Login

動作中JOBの確認

PipelineのIDをお持ちでない場合、[ゲストとしてログインすることができます。](#)

マニュアルおよびチュートリアル

- 日本語チュートリアル (FAQ)
- 英語マニュアル
- DBCLS 総合TV チュートリアル1 - 今日からはじめるDDBJ Read Annotation Pipeline
- DBCLS 総合TV チュートリアル2 - DDBJ Read Annotation Pipelineによるde novo Assembly解析
- チュートリアル: FTPでファイルをアップロードし、DDBJ Pipelineへ登録する方法
- チュートリアル: DDBJ PipelineでHGAP法でPacBioリードのアセンブリを行う方法

塩基配列・解析結果の登録

- DRA: NGS出力データの登録
- DDBJ-INSDC: アノテーション済の塩基配列データの登録

Citation

- Nagasaki, H. et al., "DDBJ Read Annotation Pipeline: A

Pipelineフローチャート

ユーザー

File Upload

DRS Start

Reads, metadata

DDBJ Read Archive

DDBJ

Reads, metadata

map positions

metadata, annotation

Manual curation

WGS/CON

基礎処理部

Mapping

de novo assembly

解析目的別ワークフロー

ゲノム解析

RNA-seq

ChIP-seq

Pipeline

Tweets by @pipeline_info

pipeline @pipeline_info

Several jobs might have failed due to the shutdown. Please check your job results carefully.

W7-5: 登録

The screenshot shows the DDBJ pipeline web interface. The main navigation bar includes steps: Select Query Files (highlighted), Select Tools, Set QuerySet, Set GenomeSet, Set Map Options, and Confirmation. The current step is 'Selecting Query Files', which has sub-options: FTP upload (highlighted), Private DRA entry, Import public DRA, Preprocessing, and HTTP upload. Below these options is a text area for 'List of your uploaded files by FTP client.' with a link '[Add new files]' (highlighted with a red arrow labeled '2'). A table below shows one file upload:

	Filename	Description	Layout	Instrument model	File size
<input type="checkbox"/>	QC.1.trimmed.fastq.gz (more 1 files)	L.hokkaidonensis_MiSeq_denovo	paired	ILLUMINA	120.0 MB

Buttons for 'Select All', 'Clear All', 'DELETE', and 'NEXT' are also visible.

W7-5: 登録

http://p.ddbj.nig.ac.jp/pipeline/RegistQueryDeleteFiles.do

DDBJ Read Annotation P... x

DDBJ
DNA Data Bank of Japan

ACCOUNT
login ID [agribio]
Logout
Change password

ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping /
de novo Assembly
step-2
Workflow
Genome (SNP/Short
Indel)
RNA-seq (Tag count)
ChIP-seq
JOB STATUS
step1.
Preprocessing
step1.
Mapping
step1.
de novo Assembly
step2-All status
HELP
HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation

Registration of fastq/fastq files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please upload query files.

To use your fasta or fastq files as pipeline query, you need to upload files to our server via FTP or HTTP. FTP uploading works faster than HTTP uploading. Therefore we recommend using FTP rather than HTTP

By FTP (Recommended)

FTP Configuration.

Server : Port	pdata.nig.ac.jp:21
Security	SSL Explicit encryption
User ID/password	Your Pipeline login ID/password If you can't login via FTP, retry after changing password .

[FTP setting manual \(English\)](#)
[FTP setting manual \(Japanese\)](#)

Recommended FTP client softwares.

Windows	FFFTP WinSCP
Mac OS X	Cyberduck
LinuxOS	FileZilla

For security our FTP server utilizes FTP over SSL protocol (FTPS). Other FTP client softwares can be used if they support FTPS.

NOTICE

- Uploaded files **cannot be seen** from other Pipeline users.
- when you connected to FTP server, there are two directories, "query" and "galaxy".
Please upload into "query" directory. If you uploaded to same level as the "query" directory, the file cannot be used in DDBJ Pipeline.
- The uploaded files will be displayed in the list below after a few minutes. (It takes 2-5 min per 1GB)
When uploading is completed, files are transferred to Pipeline data directory from FTP server.
So files seem to be removed, but it is normal operation.
- Please ensure that uploading files have appropriate file extensions.
eg. In the case of Bzip2 files, please add the ".bz2" extension.

Supported file type



W7-5: 登録

①W7-2でアップロードしたファイルが見えていないはず。②Next STEP。③もし見えていなければリロード、またはアップロードのやり直し



http://p.ddbj.nig.ac.jp/pipeline/RegistQueryDeleteFiles.do

Mapping
step1.
de novo Assembly
step2-All status

HELP
HELP [?](#)
TUTORIAL
Contact Us.
DDBJ Read Annotation Pipeline.
Development Team.

- Uploaded files **cannot be seen** from other Pipeline users.
- when you connected to FTP server, there are two directories, "query" and "galaxy".
Please upload into "query" directory. If you uploaded to same level as the "query" directory, the file cannot be used in DDBJ Pipeline.
- The uploaded files will be displayed in the list below after a few minutes. (It takes 2-5 min per 1GB)
When uploading is completed, files are transferred to Pipeline data directory from FTP server.
So files seem to be removed, but it is normal operation.
- Please ensure that uploading files have appropriate file extensions.
eg. In the case of Bzip2 files, please add the ".bz2" extension.

Supported file type

Filetype	Extension
Plain text	.fasta, .fq .fastq, .fa etc...
Gzip	.gz
Bzip2	.bz2

Bzip2 is recommended, because save disk space usage and transfer.

By HTTP (slower)

If you can't use FTP uploading, click "Browse and Upload" button and select FASTA/FASTQ files to be uploaded.

[Browse and Upload](#) [Delete Files](#)

	filename	type	size	timestamp
<input type="checkbox"/>	QC.1.trimmed.fastq.gz	fastq	57.0 MB	2016-01-15 04:19:05
<input type="checkbox"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:18:57
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 08:56:17
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:53:41
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.3.bax.h5	invalid	945.5 MB	2016-03-25 08:51:51

Go to the next page after uploading files.

[Next STEP >](#)



W7-5: 登録

2. Select a FASTA/FASTQ file.

If you are select Paired-end, please specify

	filename	type	size	timestamp
<input checked="" type="radio"/>	Not select			
<input type="radio"/>	QC.1.trimmed.fastq.gz	fastq	57.0 MB	2016-01-15 04:19:05
<input type="radio"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:18:57
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 08:56:17
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:53:41

Go to the next page after you select a file.

Next STEP >



W7-6: PacBioはsingle-end

第6回W14-5ではpaired-endと指定したが、PacBioの場合は①デフォルトのsingle-endとして取り扱う。②1つめのbax.h5ファイルを選択して、③Next STEP。

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please specify read layout to uploaded files.

1. Select a read layout:

Read layout : ①

2. Select a FASTA/FASTQ file:

If you are select Paired-end, please specify

	filename	type	size	timestamp
<input type="radio"/>	Not select			
<input type="radio"/>	QC.1.trimmed.fastq.gz	fastq	57.0 MB	2016-01-15 04:19:05
<input type="radio"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:18:57
<input checked="" type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 08:56:17
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:53:41

Go to the next page after you select a file.

③ Next STEP >

W7-6: PacBioはsingle-end

①上部に移動。②の部分を変更および記載し、③SUBMIT、④OK

Registration of fastq/fastq files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please specify instrument model.

SelectedFile 1	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5
SelectedFile 2	Not select
Read layout	Single-end
Instrument model	PacBio
(Required) Study title	L.hokkaidonensis.PacBio1

NOTICE: After confirming your entries, push the SUBMIT button to register uploaded files.

SUBMIT

Web ページからのメッセージ

Are you sure you want to submit?

OK キャンセル

W7-7: 2つめのbax.h5

こんな感じに見えます。2つめのbax.h5
ファイルを登録すべく、①Add new files

The screenshot shows the DDBJ pipeline web interface. The main navigation bar includes steps: Select Query Files (highlighted), Select Tools, Set QuerySet, Set GenomeSet, Set Map Options, and Confirmation. The current step is 'Selecting Query Files', which has sub-steps: FTP upload (highlighted), Private DRA entry, Import public DRA, Preprocessing, and HTTP upload. Below the sub-steps, there is a section titled 'List of your uploaded files by FTP client. [Add new files]'. A red arrow with the number '1' points to the '[Add new files]' link. Below this section is a table of uploaded files.

	Filename	Description	Layout	Instru mo
<input type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	L.hokkaidonensis.PacBio1	single	PacBio
<input type="checkbox"/>	QC.1.trimmed.fastq.gz (more 1 files)	L.hokkaidonensis_MiSeq_denovo	paired	ILLUMI

W7-7: 2つめのbax.h5

The screenshot shows the DDBJ Read Annotation Pipeline web interface. The browser address bar displays `http://p.ddbj.nig.ac.jp/pipeline/RegistQuery.do`. The page content includes a navigation sidebar on the left with links for Mapping, step1, de novo Assembly, step2-All status, HELP, and TUTORIAL. The main content area contains instructions for file uploads, a table of supported file types, and a table of uploaded files. A red arrow labeled '1' points to the third row of the file table, and another red arrow labeled '2' points to the 'Next STEP >' button at the bottom right.

Mapping

step1.
de novo Assembly
step2-All status

HELP
HELP [?](#)
TUTORIAL

Contact Us.
DDBJ Read Annotation Pipeline.
Development Team.

- Uploaded files **cannot be seen** from other Pipeline users.
- when you connected to FTP server, there are two directories, "query" and "galaxy".
Please upload into "query" directory. If you uploaded to same level as the "query" directory, the file cannot be used in DDBJ Pipeline.
- The uploaded files will be displayed in the list below after a few minutes. (It takes 2-5 min per 1GB)
When uploading is completed, files are transferred to Pipeline data directory from FTP server.
So files seem to be removed, but it is normal operation.
- Please ensure that uploading files have appropriate file extensions.
eg. In the case of Bzip2 files, please add the ".bz2" extension.

Supported file type

Filetype	Extension
Plain text	.fasta, .fq .fastq, .fa etc...
Gzip	.gz
Bzip2	.bz2

Bzip2 is recommended, because save disk space usage and transfer.

By HTTP (slower)

If you can't use FTP uploading, click "Browse and Upload" button and select FASTA/FASTQ files to be uploaded.

[Browse and Upload](#) [Delete Files](#)

	filename	type	size	timestamp
<input type="checkbox"/>	QC.1.trimmed.fastq.gz	fastq	57.0 MB	2016-01-15 04:19:05
<input type="checkbox"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:18:57
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 08:56:17
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:53:41
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.3.bax.h5	invalid	945.5 MB	2016-03-25 08:51:51

Go to the next page after uploading files.

[Next STEP >](#)

W7-7: 2つめのbax.h5

①上部に移動し、②single-endになっていることを確認して、③2つめのbax.h5ファイルにチェックを入れて、④Next STEP

Registration of fastq/fasta files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please specify read layout to uploaded files.

1. Select a read layout:

Read layout :

2. Select a FASTA/FASTQ file:

If you are select Paired-end, please specify

	filename	type	size	timestamp
<input type="radio"/>	Not select			
<input type="radio"/>	QC.1.trimmed.fastq.gz	fastq	57.0 MB	2016-01-15 04:19:05
<input type="radio"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:18:57
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 08:56:17
<input checked="" type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:53:41

Go to the next page after you select a file.

Next STEP >

W7-7: 2つめのbax.h5

①上部に移動。②の部分を変更および記載し、③SUBMIT、④OK。
⑤念のために、ここは2にしています

Registration of fastq/fastq files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please specify instrument model.

SelectedFile 1	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5
SelectedFile 2	Not select
Read layout	Single-end
Instrument model	PacBio
(Required) Study title	L.hokkaidonensis.PacBio2

NOTICE: After confirming your entries, push the SUBMIT button to register uploaded files.

SUBMIT

Web ページからのメッセージ

Are you sure you want to submit?

OK キャンセル

W7-8: 3つめのbax.h5

こんな感じに見えます。3つめのbax.h5
ファイルを登録すべく、①Add new files

The screenshot shows the DDBJ pipeline web interface. The main navigation bar includes: Select Query Files (highlighted), Select Tools, Set QuerySet, Set GenomeSet, Set Map Options, and Confirmation. The current step is 'Selecting Query Files'. The left sidebar contains sections for ACCOUNT (login ID, Logout, Change password), ANALYSIS (Data setup, DRA Start, FTP upload, HTTP upload, DRA Import, Preprocessing Start), step-1 (Preprocessing, Mapping / de novo Assembly), step-2 (Workflow: Genome (SNP/Short Indel), RNA-seq (Tag count), ChIP-seq), JOB STATUS (step1. Preprocessing, Mapping, de novo Assembly), and HELP (HELP, TUTORIAL, Contact Us).

The 'Selecting Query Files' section has a 'NEXT' button and tabs for: FTP upload (highlighted), Private DRA entry, Import public DRA, Preprocessing, and HTTP upload. Below the tabs is a text input field with the placeholder 'List of your uploaded files by FTP client. [Add new files]' and a red arrow pointing to the '[Add new files]' link, labeled with a circled '1'. There are 'Select All' and 'Clear All' buttons below the input field.

	Filename	Description	Layout	Instru mo
<input type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	L.hokkaidonensis.PacBio2	single	PacBio
<input type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	L.hokkaidonensis.PacBio1	single	PacBio
<input type="checkbox"/>	QC.1.trimmed.fastq.gz (more 1 files)	L.hokkaidonensis_MiSeq_denovo	paired	ILLUMI

At the bottom of the table area are 'DELETE' and 'NEXT' buttons.

W7-8: 3つめのbax.h5

The screenshot shows the DDBJ Read Annotation Pipeline web interface. The browser address bar displays `http://p.ddbj.nig.ac.jp/pipeline/RegistQuery.do`. The page content includes a navigation sidebar on the left with links for Mapping, step1, de novo Assembly, step2-All status, HELP, and TUTORIAL. The main content area contains instructions for file uploads, a table of supported file types, and a table of uploaded files. A red arrow labeled '1' points to the third row of the file table, and another red arrow labeled '2' points to the 'Next STEP >' button at the bottom right.

Mapping

step1.
de novo Assembly
step2-All status

HELP
HELP [?](#)
TUTORIAL

Contact Us.
DDBJ Read Annotation Pipeline.
Development Team.

- Uploaded files **cannot be seen** from other Pipeline users.
- when you connected to FTP server, there are two directories, "query" and "galaxy".
Please upload into "query" directory. If you uploaded to same level as the "query" directory, the file cannot be used in DDBJ Pipeline.
- The uploaded files will be displayed in the list below after a few minutes. (It takes 2-5 min per 1GB)
When uploading is completed, files are transferred to Pipeline data directory from FTP server.
So files seem to be removed, but it is normal operation.
- Please ensure that uploading files have appropriate file extensions.
eg. In the case of Bzip2 files, please add the ".bz2" extension.

Supported file type

Filetype	Extension
Plain text	.fasta, .fq .fastq, .fa etc...
Gzip	.gz
Bzip2	.bz2

Bzip2 is recommended, because save disk space usage and transfer.

By HTTP (slower)

If you can't use FTP uploading, click "Browse and Upload" button and select FASTA/FASTQ files to be uploaded.

[Browse and Upload](#) [Delete Files](#)

	filename	type	size	timestamp
<input type="checkbox"/>	QC.1.trimmed.fastq.gz	fastq	57.0 MB	2016-01-15 04:19:05
<input type="checkbox"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:18:57
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 08:56:17
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:53:41
<input type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.3.bax.h5	invalid	945.5 MB	2016-03-25 08:51:51

Go to the next page after uploading files.

[Next STEP >](#)

W7-8: 3つめのbax.h5

Registration of fastq/fastq files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please specify read layout to uploaded files.

1. Select a read layout:

Read layout :

2. Select a FASTA/FASTQ file:

If you are select Paired-end, please specify

	filename	type	size	timestamp
<input type="radio"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:19:05
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 04:18:57
<input checked="" type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:56:17
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5	invalid	945.5 MB	2016-03-25 08:53:41
<input type="radio"/>				2016-03-25 08:51:51

Go to the next page after you select a file.

Next STEP >

W7-8: 3つめのbax.h5

①single-endになっていることを確認して、②3つめのbax.h5ファイルにチェックを入れて、③Next STEP

Registration of fastq/fastq files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please specify read layout to uploaded files.

1. Select a read layout:
Read layout : ①

2. Select a FASTA/FASTQ file:

If you are select Paired-end, please specify

	filename	type	size	timestamp
<input type="radio"/>	QC.2.trimmed.fastq.gz	fastq	62.9 MB	2016-01-15 04:19:05
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	invalid	784.3 MB	2016-03-25 04:18:57
<input type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	invalid	803.9 MB	2016-03-25 08:56:17
<input checked="" type="radio"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5	invalid	945.5 MB	2016-03-25 08:53:41
				2016-03-25 08:51:51

Go to the next page after you select a file.

Next STEP > ③

W7-8: 3つめのbax.h5

①上部に移動。②の部分を変更に適切に変更および記載し、③SUBMIT、④OK。
⑤念のために、ここは3にしています

Registration of fastq/fastq files

1. Upload FASTA/FASTQ files 2. Select FASTA/FASTQ files 3. Registration

Please specify instrument model.

SelectedFile 1	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5
SelectedFile 2	Not select

Read layout: Single-end

Instrument model: PacBio

(Required) Study title: L.hokkaidonensis.PacBio3

NOTICE: After confirming your entries, push the SUBMIT button to register uploaded files.

SUBMIT

Web ページからのメッセージ

Are you sure you want to submit?

OK キャンセル

W7-9: 登録完了後

2016年3月28日現在、bax.h5ファイルを一つ一つ登録していく必要があり、非常に面倒。近い将来改善されていくであろうが、HGAPを実行できるだけでもありがたい

The screenshot shows the DDBJ Pipeline web interface. The main heading is "Selecting Query Files". Below it, there are tabs for "FTP upload", "Private DRA entry", "Import public DRA", "Preprocessing", and "HTTP upload". A table lists the uploaded files:

	Filename	Description	Layout	Instrument model	File size
<input type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5	L.hokkaidonensis.PacBio3	single	PacBio	945.5 MB
<input type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	L.hokkaidonensis.PacBio2	single	PacBio	803.9 MB
<input type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	L.hokkaidonensis.PacBio1	single	PacBio	784.3 MB
<input type="checkbox"/>	QC.1.trimmed.fastq.gz (more 1 files)	L.hokkaidonensis_MiSeq_denovo	paired	ILLUMINA	120.0 MB

Buttons for "Select All", "Clear All", "DELETE", and "NEXT" are visible. The left sidebar contains navigation options for ACCOUNT, ANALYSIS, and JOB STATUS.

W8-1: HGAP実行

①HGAPの入力として用いる3つのbax.h5ファイルを選択し、②NEXT

The screenshot shows the DDBJ pipeline web interface. The main heading is "Selecting Query Files". The interface includes a navigation menu on the left with sections for ACCOUNT, ANALYSIS, JOB STATUS, and HELP. The main content area shows a progress bar with steps: Select Query Files, Select Tools, Set QuerySet, Set GenomeSet, Set Map Options, and Confirmation. Below the progress bar, there are tabs for "FTP upload", "Private DRA entry", "Import public DRA", "Preprocessing", and "HTTP upload". A table lists uploaded files with columns for Filename, Description, Layout, Instrument model, and File size. Three files are selected (checked):

	Filename	Description	Layout	Instrument model	File size
<input checked="" type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.3.bax.h5	L.hokkaidonensis.PacBio3	single	PacBio	945.5 MB
<input checked="" type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.2.bax.h5	L.hokkaidonensis.PacBio2	single	PacBio	803.9 MB
<input checked="" type="checkbox"/>	m130821_065825_42195_c100539522550000001823089611241356_s1_p0.1.bax.h5	L.hokkaidonensis.PacBio1	single	PacBio	784.3 MB
<input type="checkbox"/>	QC.1.trimmed.fastq.gz (more 1 files)	L.hokkaidonensis_MiSeq_denovo	paired	ILLUMINA	120.0 MB

At the bottom right, there are "DELETE" and "NEXT" buttons. A red arrow labeled "2" points to the "NEXT" button. A red arrow labeled "1" points to the "FTP upload" tab.

W8-1: HGAP実行

①下部に移動、②de novo Assembly、③HGAP、④NEXT

Preprocessing Start

step-1
Preprocessing

Mapping / de novo Assembly

step-2

Workflow

Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS

step1. Preprocessing

step1. Mapping

step1. de novo Assembly

step2-All status

HELP

HELP [?](#)

TUTORIAL

Contact Us.
DDBJ Read Annotation Pipeline.
Development Team.

Tool	Help	Version	Base space	Color space	Paired-end	MSS (WGS)	Comment
<input type="checkbox"/> BLAT	?	34	✓		✓		Single-end analysis only
<input type="checkbox"/> bwa	?	0.6.1	✓	✓	✓	✓	
<input type="checkbox"/> Bowtie	?	0.12.7	✓	✓	✓	✓	
<input type="checkbox"/> TopHat	?	1.0.11	✓	✓	✓	✓	
<input type="checkbox"/> Bowtie2	?	2.2.6	✓	✓	✓	✓	For reads longer than about 50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie1.
<input type="checkbox"/> TopHat2	?	2.1.0	✓	✓	✓	✓	

de novo Assembly
Total limit = 22 Gbp

Tool	Help	Version	Base space	Color space	Paired-end	MSS (WGS)	Comment
<input type="checkbox"/> SOAPdenovo	?	2.04-r240	✓		✓		
<input type="checkbox"/> ABySS	?	1.3.2	✓		✓		Maximum K-mer value is 64.
<input type="checkbox"/> Velvet	?	1.2.10	✓		✓	✓	We severe recommend when performing Velvet, total length of those reads is up to 22G bp.Maximum K-mer value is 64.
<input type="checkbox"/> Trinity	?	2.1.1	✓		✓		RNA-Seq De novo Assembly
<input type="checkbox"/> Platanus	?	1.2.2	✓		✓		
<input checked="" type="checkbox"/> HGAP	?	Protocol3 (v 2.2.0)					HGAP Pipeline for PacBio Sequence based on SMRT Analysis v2.2.0. For bax.h5 file only. (Beta version)

Mapping Contigs by de novo Assemble to Reference Sequences.
The contigs will be aligned to reference genome.

Tool	Comment
<input checked="" type="radio"/> BLAT	Single-end analysis only

BACK NEXT

W8-1: HGAP実行

①解析したい3つのファイルにチェックを入れて、②confirm

http://p.ddbj.nig.ac.jp/pipeline/SelectQuery.do

Generating Query Sets f... x

Select Query Files → Select Tools → **Set QuerySet** → Set Ass. Options → Confirmation → Running Status

Generating Query Sets from Query Read Files

RESET BACK NEXT

Single analysis
Layout of single sequence.

5' 3'

Linker(1) Target Linker(2)

	Run ACCESSION	Read length	Quality Score
<input checked="" type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.3.bax.h5	bp	
<input checked="" type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.2.bax.h5	bp	
<input checked="" type="checkbox"/>	m130821_065825_42195_c10053952255000001823089611241356_s1_p0.1.bax.h5	bp	

confirm

RESET BACK NEXT

QUERY SET

ACCOUNT
login ID [agribio]
Logout
Change password

ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping / de novo Assembly
step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS
step1. Preprocessing
step1. Mapping
step1. de novo Assembly
step2-All status

HELP
HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation

W8-1: HGAP実行

The screenshot shows the DDBJ pipeline web interface. The browser address bar displays `http://p.ddbj.nig.ac.jp/pipeline/Confirm1ToSelectQuery.do`. The navigation bar includes steps: Select Query Files, Select Tools, Set QuerySet (highlighted), Set Ass. Options, Confirmation, and Running Status.

ACCOUNT

- login ID [agribio]
- Logout
- Change password

ANALYSIS

Data setup

- DRA Start
- FTP upload
- HTTP upload
- DRA Import
- Preprocessing Start

step-1

- Preprocessing
- Mapping / *de novo* Assembly

step-2

Workflow

- Genome (SNP/Short Indel)
- RNA-seq (Tag count)
- ChIP-seq

JOB STATUS

- step1. **Preprocessing**
- step1. **Mapping**
- step1. *de novo* Assembly
- step2-All status

HELP

- HELP
- TUTORIAL
- Contact Us.
- DDBJ Read Annotation

Generating Query Sets from Query Read Files

RESET BACK NEXT

Single analysis

Layout of single sequence.

5' 3'

Linker(1) Target Linker(2)

Run ACCESSION Read length Quality Score

confirm

QUERY SET

Query set1

PairedOrientation	RunAccession	RunAlias	RowLength	Quality Score1	Quality Score2
single	21144	L.hokkaidonensis.PacBio3			
single	21143	L.hokkaidonensis.PacBio2			
single	21142	L.hokkaidonensis.PacBio1			

RESET BACK NEXT

①

W8-2: 2つのパラメータ

http://p.ddbj.nig.ac.jp/pipeline/SettingAssembly.do

Setting for De Novo Ass...

Select Query Files → Select Tools → Set QuerySet → **Set Ass. Options** → Confirmation → Running Status

Setting for De Novo Assembly

BACK NEXT

hgap

Set optional parameters for HGAP pipeline

Select UGE-node to run :

month_fat (32 CPUs and 320GB memory)

month_medium (32 CPUs and 256GB memory)

Same results will be generated with either option.
You can check the CPU and memory usage at [NIG-SC Website](#).

1: The approximate genome size, in base pairs. (Must be a value between 1 and 150000000)

GenomeSize =

2: The minimum length of reads (in base pairs) to use as seeds for pre-assembly.

Minimum Seed Length :

Automatic Estimation

If the coverage exceeds 30X, the Minimum Seed Read Length that results in at least 30X coverage by the longest subreads will be calculated automatically. If the coverage is less than 30X, the user-specified value will be used.

Use Manually Specified Value (regardless of the coverage)

BACK NEXT

W8-2: パラメータの解説

The screenshot shows a GitHub Wiki page for 'HGAP in SMRT Analysis'. The page content is as follows:

HGAP in SMRT Analysis
lhon edited this page on Jun 10 2014 · 2 revisions

This page contains information about the current release of HGAP.

There have been multiple iterations of the HGAP implementation in SMRT Analysis, with performance improvements added to each iteration. In SMRT Analysis v2.0, significant improvements were introduced, significantly speeding up HGAP execution. In most cases, HGAP.3 makes it the preferred protocol. In production environments, we recommend using the latest version of SMRT Analysis to ensure the best performance with HGAP.

SMRT Analysis v2.1 has a new implementation of HGAP that speeds up the process. This is found in the `RS_HGAP_Assembly.2` protocol. `RS_HGAP_Assembly.2` is used in SMRT Analysis v2.2.0 and later versions.

SMRT Analysis v2.2 contains a further improvement to HGAP, in which the assembly stage is sped up, this new protocol is named `RS_HGAP_Assembly.3`. The `RS_HGAP_Assembly.2` and `RS_HGAP_Assembly.3` versions 2 and 3 is largely the same.

Important parameters

- 1. Genome Size**
To accurately determine the Minimum Seed Read Length and the coverage of trimmed preassembled reads going into the assembly step, it is important to adjust the target genome size as accurately as possible.
- 2. Automatic Minimum Seed Read Length calculation**
The Minimum Seed Read Length that results in at least 30X target genome coverage by the longest subreads is being calculated automatically (the default option). To use the user-selected Minimum Seed Read Length, the default option has to be **deselected**. If less than 30X coverage is being used for the HGAP process, the algorithm will use the user-selected Minimum Seed Length (6kb default), so lowering the default setting to 500bp is required to allow all-vs-all PreAssembly at lower than 30X coverage.

Genome Size

At the moment, HGAP in SMRT Analysis supports genomes up to 130 MB; further improvements to scaling the workflow will enable support for larger genomes.

Older versions of SMRT Analysis may have lower genome size limits. SMRT Analysis 2.0 was limited to a 10 Mb genome size. We do not recommend using older versions of SMRT Analysis since they can have significant performance limitations; please upgrade if possible.

Usage notes

For microbial assemblies we have seen improved assembly results using the latest workflows

W8-3: HGAP実行

①ゲノムサイズは、乳酸菌の平均的なゲノムサイズである2.5MB。②Minimum Seed Lengthは、Automatic Estimation (デフォルト)を指定して③NEXT

http://p.ddbj.nig.ac.jp/pipeline/SettingAssembly.do

Setting for De Novo Assembly

Select Query Files → Select Tools → Set QuerySet → **Set Ass. Options** → Confirmation → Running Status

ACCOUNT

login ID [agribio]
Logout
Change password

ANALYSIS

Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start

step-1
Preprocessing
Mapping / de novo Assembly

step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS

step1. Preprocessing
step1. Mapping
step1. **de novo Assembly**
step2-All status

HELP
HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation

Setting for De Novo Assembly

BACK NEXT

hgap

Set optional parameters for HGAP pipeline

Select UGE-node to run :

month_fat (32 CPUs and 320GB memory)
 month_medium (32 CPUs and 256GB memory)

Same results will be generated with either option.
You can check the CPU and memory usage at [NIG-SC Website](#).

1 : The approximate genome size, in base pairs.(Must be a value between 1 and 150000000)

GenomeSize = x

2 : The minimum length of reads (in base pairs) to use as seeds for pre-assembly.

Minimum Seed Length :

Automatic Estimation
If the coverage exceeds 30X, the Minimum Seed Read Length that results in at least 30X coverage by the longest subreads will be calculated automatically. If the coverage is less than 30X, the user-specified value will be used.

Use Manually Specified Value (regardless of the coverage)

BACK NEXT

W8-3: HGAP実行

http://p.ddbj.nig.ac.jp/pipeline/Confirm.do

Run Confirmation

Select Query Files → Select Tools → Set QuerySet → Set Ass. Options → **Confirmation** → Running Status

ACCOUNT
login ID [agribio]
Logout
Change password

ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start
step-1
Preprocessing
Mapping / de novo Assembly
step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq
JOB STATUS
step1. Preprocessing
step1. Mapping
step1. de novo Assembly
step2-All status
HELP
HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation

Run Confirmation

BACK RUN

Destination of mail
When the request is completed, the system sends an email to this address.
kadota@bi.a.u-tokyo.ac.jp * Required
Result files will be deleted 60 days after submission.

Assembly [hgap]

Query sets

PairedOrientation	RunAccession	RunAlias	RowLength	QualityScore1	QualityScore2
single	21144	L.hokkaidonensis.PacBio3			
single	21143	L.hokkaidonensis.PacBio2			
single	21142	L.hokkaidonensis.PacBio1			

Assembly commands
hgap

Set optional parameters for HGAP pipeline

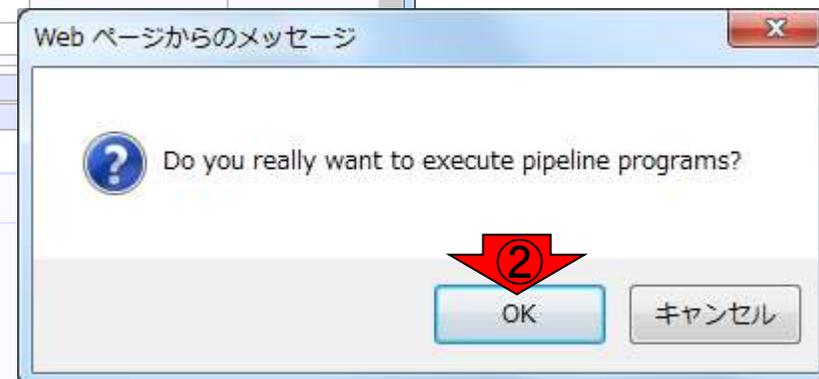
Select UGE-node to run :

month_fat (32 CPUs and 320GB memory)
 month_medium (32 CPUs and 256GB memory)

Same results will be generated with either option.
You can check the CPU and memory usage at [NIG-SC Website](#).

1 : The approximate genome size, in base pairs.(Must be a value between 1 and 150000000)

GenomeSize = 2500000



W8-3: HGAP実行

The screenshot shows a web browser window with the URL <http://p.ddbj.nig.ac.jp/pipeline/ConfirmRun.do>. The page features a navigation breadcrumb at the top: [Select Query Files](#) → [Select Tools](#) → [Set QuerySet](#) → [Set Ass. Options](#) → [Confirmation](#) → [Running Status](#). A prominent blue message box in the center states "The reservation was completed." Below this message are two buttons: "STATUS" and "NEXT JOB". A red arrow with the number "1" points to the "STATUS" button. The left sidebar contains several menu sections: "ACCOUNT" (login ID [agribio], Logout, Change password), "ANALYSIS" (Data setup, DRA Start, FTP upload, HTTP upload, DRA Import, Preprocessing Start, step-1, step-2, Workflow), "JOB STATUS" (step1. Preprocessing, step1. Mapping, step1. de novo Assembly, step2-All status), and "HELP" (HELP, TUTORIAL, Contact Us.).

W8-4: Status

実行中(Running)…。①登録作業(W7-8)の最後のほうで記載したStudy titleはここで見られる。自分が他のヒトのStudy titleを見られるように、他のヒトも自分のStudy titleを見ることができるので、気をつけよう

ACCOUNT
login ID [agribio]
Logout
Change password

ANALYSIS
Data setup
DRA Start
FTP upload
HTTP upload
DRA Import
Preprocessing Start

step-1
Preprocessing
Mapping / de novo Assembly

step-2
Workflow
Genome (SNP/Short Indel)
RNA-seq (Tag count)
ChIP-seq

JOB STATUS
step1. Preprocessing
step1. Mapping
step1. de novo Assembly
step2-All status

HELP
HELP
TUTORIAL
Contact Us.
DDBJ Read Annotation Pipeline

Select Query Files → Select Tools → Set QuerySet → Set Ass. Options → Confirmation → Running Status

Status - de novo Assembly

Mapping Job | **de novo Assembly Job** | Preprocessing Job

Order

Sort by : ID ▼ Descending ▼ Show Only Your Own Job Reload

Delete * page 1 ▼ NEXT >

	ID	UserID	Submission accession	P/S	Status	Tool	Read #	Read length	Assembly detail	Mapping detail	Start time	End time	Elapsed time
<input type="checkbox"/>	21965	agribio	L.hokkaidonensi L.hokkaidonensi L.hokkaidonensi	S	running	HGAP		---	View		2016-03-28 20:14:25	---	
<input type="checkbox"/>	21953	---	---	P	error	Trinity		---			---	---	
<input type="checkbox"/>	21952	---	---	P	error	Trinity		---			---	---	
<input type="checkbox"/>	21951	---	---	P	error	Trinity		---			---	---	
<input type="checkbox"/>	21950	---	---	P	error	Trinity		---			---	---	

Top of page

W9-1: 計算終了

①このときは約23時間後の19:13に、②DDBJ Pipelineから計算終了メールが届いた。計算結果を眺めるべく、③DDBJ Pipelineにログイン

2016/03/29 (火) 19:13

pipeline_team@g.nig.ac.jp

Job finished : DDBJ Read Annotation Pipeline

宛先 kadota@bi.a.u-tokyo.ac.jp

C C pipeline_report@g.nig.ac.jp

 このメッセージから余分な改行を削除しました。

Dear agribio,

Your request to DDBJ pipeline service has finished.
Please visit the web site to obtain analytical results.

Request ID: 21965

URL: <https://p.ddbj.nig.ac.jp/>

If you have troubles in this service, please write to pipeline_dev@ddbj.nig.ac.jp Thank you for trying our analytical service.

Regards,

DDBJ

W9-2: 結果を眺める

① *de novo* Assembly、② Job ID番号(21965)を頼りにすれば、このページに辿り着ける。
③ 赤枠部分を見ると、④ コンティグ数は4つ

The screenshot shows the DDBJ pipeline detail view for Job ID 21965. The page is titled "Detail view" and includes a navigation bar with steps: Select Query Files, Select Tools, Set QuerySet, Set Ass. Options, Confirmation, and Running Status. The left sidebar contains sections for ACCOUNT, ANALYSIS, and JOB STATUS. The main content area is divided into several sections:

- Job info:** ID: 21965, Tool (Version): HGAP (Protocol3(v 2.2.0)).
- RunAccession or Filename:** A table listing three files with their respective download links.
- Download modified queries:** A message stating that the modified query file does not exist due to reasons like expiration or queueing.
- Download wgs file:** A link to download the WGS fasta file (2.4 MB).
- Assembly statistics:** A table showing 4 contigs with their total, maximum, minimum, and N50 sizes.
- Time:** A table showing the wait time, start time, and end time of the job.
- Command:** A table showing the command used to run the pipeline.

Contig # : 4
Total contig size : 2,433,614
Maximum contig size : 2,289,497
Minimum contig size : 11,372
N50 contig size : 2,289,497

Contig # : 4
Total contig size : 2,433,614
Maximum contig size : 2,289,497
Minimum contig size : 11,372
N50 contig size : 2,289,497

W9-2: 結果を眺める

① Total contig sizeは乳酸菌の一般的なゲノムサイズと近く、妥当。② Maximum contig sizeのものが全体の9割以上を占めていることから、これが乳酸菌の染色体ゲノムなのだろうと妄想する

The screenshot shows the DDBJ pipeline detail view for job ID 21965. The job was run using HGAP (Protocol3 v 2.2.0). The assembly statistics are as follows:

Contig #	Total contig size	Maximum contig size	Minimum contig size	N50 contig size
4	2,433,614	2,289,497	11,372	2,289,497

The command used for the run is: `run HGAP through smrtpipe.py : GenomeSize=2500000,minSeedLength=6000`. The run started on 2016-03-28 at 20:14:25 and ended on 2016-03-29 at 19:13:20.

Contig # : 4
Total contig size : 2,433,614
Maximum contig size : 2,289,497
Minimum contig size : 11,372
N50 contig size : 2,289,497



W9-3:ダウンロード

- ①result.zipというzip圧縮ファイルを共有フォルダ(ホストOS側はDesktop/share)にダウンロード。
- ②MD5については、連載第3回W12で説明した。
- ③リンク先のMD5チェックサム値

The screenshot shows the DDBJ pipeline detail view for job ID 21965. The job info section indicates the tool used is HGAP (Protocol3(v 2.2.0)). Below this, a table lists commands and their results. A red box highlights a specific command and its result, with arrows pointing to the download link and MD5 value. A callout box shows the MD5 value and the filename 'result.zip'.

Command	Start time	End time	Log1	Log2	Result	MD5
run HGAP through smrtpipe.py : GenomeSize=2500000,minSeedLength=6000	2016-03-28 20:14:25	2016-03-29 19:12:37	View		Download(13.1 MB)	MD5

5cf4ed21fd476edce6625afaec577815 result.zip

W9-3: ダウンロード

ここでは、①ゲストOSの共有フォルダ上で②wgetを用いて(門田のサイトから)取得しているが、result.zipが共有フォルダにダウンロードされていれば手段はなんでもよい。③md5sumコマンド実行結果。④ウェブサイト上の数値と合っていることを確認(ダウンロード時にファイルが壊れていないことを意味する)

```
File Edit View Search Terminal Help
① iu@bielinux[iu] cd ~/Desktop/mac_share
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
② iu@bielinux[mac_share] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kado
ta/book/DRR054113/result.zip
iu@bielinux[mac_share] ls -l result*           [ 8:15午後]
-rwxrwxrwx 1 iu iu 13128969 3月 29 19:57 result.zip
③ iu@bielinux[mac_share] md5sum result.zip     [ 8:15午後]
5cf4ed21fd476edce6625afaec577815 result.zip
iu@bielinux[mac_share] █                       [ 8:15午後]

5cf4ed21fd476edce6625afaec577815 result.zip ④
```

W9-4: 解凍して概観

①result.zipを解凍して、②中身を確認。
③計4つのファイルがある。④欲しい最終結果ファイルはpolished_assembly.fasta

```
iu@bielinux[mac_share] pwd [11:30午前]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l result* [11:30午前]
-rwxrwxrwx 1 iu iu 13128969 3月 29 19:57 result.zip
iu@bielinux[mac_share] unzip result.zip [11:30午前]
Archive: result.zip
  creating: result/
  inflating: result/corrected.fastq
  inflating: result/smrtpipe.log
  inflating: result/polished_assembly.fastq
  inflating: result/polished_assembly.fasta
iu@bielinux[mac_share] ls -l result* [11:30午前]
-rwxrwxrwx 1 iu iu 13128969 3月 29 19:57 result.zip

result:
total 75356
-rwxrwxrwx 1 iu iu 69756928 3月 29 19:12 corrected.fastq
-rwxrwxrwx 1 iu iu 2474245 3月 29 19:12 polished_assembly.fasta
-rwxrwxrwx 1 iu iu 4867312 3月 29 19:12 polished_assembly.fastq
-rwxrwxrwx 1 iu iu 64556 3月 29 19:12 smrtpipe.log
iu@bielinux[mac_share] [11:30午前]
```



W9-5: コンティグ数

①resultディレクトリに移動し、②コンティグ数を表示。FASTA形式ファイルなので">"を含む行数がコンティグ数に相当する。③description行を表示。こんな感じの記述内容か~と思うだけ。④行数は、40,567行

```
iu@bielinux[mac_share] pwd [12:25午後]
/home/iu/Desktop/mac_share
① iu@bielinux[mac_share] cd result [12:26午後]
iu@bielinux[result] pwd [12:26午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l [12:26午後]
total 75356
-rwxrwxrwx 1 iu iu 69756928 3月 29 19:12 corrected.fastq
-rwxrwxrwx 1 iu iu 2474245 3月 29 19:12 polished_assembly.fasta
-rwxrwxrwx 1 iu iu 4867312 3月 29 19:12 polished_assembly.fastq
-rwxrwxrwx 1 iu iu 64556 3月 29 19:12 smrtpipe.log
② iu@bielinux[result] grep -c ">" polished_assembly.fasta
4
③ iu@bielinux[result] grep ">" polished_assembly.fasta [12:26午後]
>unitig_0|quiver
>unitig_2|quiver
>unitig_3|quiver
>unitig_1|quiver
④ iu@bielinux[result] wc polished_assembly.fasta [12:26午後]
 40567 40567 2474245 polished_assembly.fasta
iu@bielinux[result] █ [12:35午後]
```

W9-6: Rで配列長

Rで配列長情報を把握。これはどのコンティグが最も長いものかなどの全体像を把握するのが目的。このファイル(polished_assembly.fasta)の場合は、①長い順にソートされていることがわかる

```
File Edit View Search Terminal Help
anyDuplicated, append, as.data.frame, as.vector, cbind, colname
s,
do.call, duplicated, eval, evalq, Filter, Find, get, intersect,
is.unsorted, lapply, Map, mapply, match, mget, order, paste, pm
ax,
pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.in
t,
rownames, sapply, setdiff, sort,
unlist, unsplit

Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' in package 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
> fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したフ
ファイルの読み込み
> width(fasta)
[1] 2289497 86892 45853 11372
> q(save="no")
iu@bielinux[result] █
```

• W9-6: Rで配列長
最終結果ファイル(polished_assembly.fasta)の配列の並びを把握すべく、配列長を表示。

```
pwd
ls -l
R -q

in_f <- "polished_assembly.fasta" #入力ファイル名を指定してin_f
library(Biostrings) #パッケージの読み込み
fasta <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファ
width(fasta) #配列長を表示

q(save="no")
```

#配列長を表示

[1:02午後]



W10-1: ファイル分割

Rでmulti-FASTAファイル(polished_assembly.fasta)を読み込んで、コンティグごとにsequence1_R.fa, sequence2_R.faなどのファイル名で保存するコード

- W10-1: multi-FASTAファイルの分割(Rの場合)
(配列の長い順にソートして) description部分をsequence1, sequence2などと変更し、それを分割後のファイル名として利用するやり方です。

```
R -q

in_f <- "polished_assembly.fasta" #入力ファイル名を指定してin_fに格納
library(Biostrings) #パッケージの読み込み
fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

#コンティグを長い順にソート
order(width(fasta)) #配列長の短い順にコンティグごとの順位を表示
order(width(fasta), decreasing=T) #配列長の長い順にコンティグごとの順位を表示
fasta <- fasta[order(width(fasta), decreasing=T)]#長い順に配列をソート

#description部分を変更
names(fasta) #変更前
names(fasta) <- paste("sequence", 1:length(fasta), sep="")#変更後の文字列を作成して代入
names(fasta) #変更後

#分割本番
for(i in 1:length(fasta)){ #コンティグ数分だけループを回す
  out_f <- paste(names(fasta)[i], "_R", ".fa", sep="")#出力ファイル名を作成
  writeXStringSet(fasta[i], file=out_f, format="fasta", width=50)#配列ごとに保存
}

q(save="no")
```

W10-1: ファイル分割1

①コピー実行後にls。②うまく作成できているようだ。③ファイルサイズもコンテイングとの塩基数(W9-6)と類似しており妥当

- W10-1: multi-FASTAファイルの分割(Rの場合)

(配列の長い順にソートして) description部分を変更し、それを分割後のファイル名として利用するやり方です。

```
R -q
in_f <- "polished_assembly.fasta"
library(Biostrings)
fasta <- readDNASTringSet(in_f)

#コンテイングを長い順にソート
order(width(fasta))
order(width(fasta),
fasta <- fasta[order(width(fasta))]

#description部分を変更
names(fasta)
names(fasta) <- paste0("sequence", 1:length(fasta), ".fa")
names(fasta)

#分割本番
for(i in 1:length(fasta)){
  out_f <- paste0(names(fasta)[i], "_R", ".fa", sep="")
  writeXStringSet(fasta[i], file=out_f, format="fasta", width=50)
}

q(save="no")
```

```
File Edit View Search Terminal Help
> for(i in 1:length(fasta)){ #コンテイング数分だけループ
を回す
+   out_f <- paste(names(fasta)[i], "_R", ".fa", sep="") #出力ファイル名を作成
+   writeXStringSet(fasta[i], file=out_f, format="fasta", width=50) #配列ごとに保存
+ }
>
> q(save="no")
iu@bielinux[result] pwd [ 4:18午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l [ 4:18午後]
total 77781
-rwxrwxrwx 1 iu iu 69756928 3月 29 19:12 corrected.fastq
-rwxrwxrwx 1 iu iu 2474245 3月 29 19:12 polished_assembly.fasta
-rwxrwxrwx 1 iu iu 4867312 3月 29 19:12 polished_assembly.fastq
-rwxrwxrwx 1 iu iu 2335298 3月 30 16:18 sequence1_R.fa
-rwxrwxrwx 1 iu iu 88641 3月 30 16:18 sequence2_R.fa
-rwxrwxrwx 1 iu iu 46782 3月 30 16:18 sequence3_R.fa
-rwxrwxrwx 1 iu iu 11611 3月 30 16:18 sequence4_R.fa
-rwxrwxrwx 1 iu iu 69756 3月 29 19:12 smrtpipe.log
iu@bielinux[result]
```



W10-1: ファイル分割1

①grepでdescription行部分を表示させ、イメージ通りになっていることを確認。②各配列の行数は、45,791、1,739、919、229行。この理由は

- W10-1: multi-FASTAファイルの分割(Rの場合)

(配列の長い順にソートして) description部分をsequence1, sequence2などと変更し、それを分割後のファイル名として利用するやり方です。

```
R -q

in_f <- "polished_assembly.fasta"
library(Biostrings)
fasta <- readDNASTrings(in_f)

#コンテイングを長い順にソート
order(width(fasta))
order(width(fasta),
fasta <- fasta[order(width(fasta),
names(fasta))

#description部分を変更
names(fasta) <- paste0("sequence", 1:length(fasta), ".fa")
names(fasta)

#分割本番
for(i in 1:length(fasta)) {
  out_f <- paste0("sequence", i, ".fa")
  writeXStringSet(fasta[[i]], out_f)
}

q(save="no")
```

The terminal window shows the following sequence of commands and outputs:

```
iu@bielinux[result] ls -l
total 77781
-rwxrwxrwx 1 iu iu 69756928 3月 29 19:12 corrected.fastq
-rwxrwxrwx 1 iu iu 2474245 3月 29 19:12 polished_assembly.fasta
-rwxrwxrwx 1 iu iu 4867312 3月 29 19:12 polished_assembly.fastq
-rwxrwxrwx 1 iu iu 2335298 3月 30 16:18 sequence1_R.fa
-rwxrwxrwx 1 iu iu 88641 3月 30 16:18 sequence2_R.fa
-rwxrwxrwx 1 iu iu 46782 3月 30 16:18 sequence3_R.fa
-rwxrwxrwx 1 iu iu 11611 3月 30 16:18 sequence4_R.fa
-rwxrwxrwx 1 iu iu 64556 3月 29 19:12 smrtpipe.log

iu@bielinux[result] grep ">" *_R.fa
sequence1_R.fa:>sequence1
sequence2_R.fa:>sequence2
sequence3_R.fa:>sequence3
sequence4_R.fa:>sequence4

iu@bielinux[result] wc *_R.fa
45791 45791 2335298 sequence1_R.fa
1739 1739 88641 sequence2_R.fa
919 919 46782 sequence3_R.fa
229 229 11611 sequence4_R.fa
48678 48678 2482332 total

iu@bielinux[result]
```

Annotations in the image include a red arrow labeled '1' pointing to the grep command and another red arrow labeled '2' pointing to the wc command. The terminal window title is 'File Edit View Search Terminal Help' and the system tray shows 'Ja', '16:41', and a settings icon.

W10-1: ファイル分割1

- W10-1: multi-FASTAファイルの分割(Rの場合)
(配列の長い順にソートして) description部分を `sequence1`, `sequence2` などと変更し、それを分割後のファイル名として利用するやり方です。

```
R -q

in_f <- "polished_assembly.fasta"      #入力ファイル名を指定してin_fに格納
library(Biostrings)                    #パッケージの読み込み
fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み

#コンティグを長い順にソート
order(width(fasta))                    #配列長の短い順にコンティグごとの順位を表示
order(width(fasta), decreasing=T)      #配列長の長い順にコンティグごとの順位を表示
fasta <- fasta[order(width(fasta), decreasing=T)]#長い順に配列をソート

#description部分を変更
names(fasta)                            #変更前
names(fasta) <- paste("sequence", 1:length(fasta), sep="")#変更後の文字列を作成して代入
names(fasta)                             #変更後

#分割本番
for(i in 1:length(fasta)){              #コンティグ数分だけループを回す
  out_f <- paste(names(fasta)[i], "_R", ".fa", sep="")#出力ファイル名を作成
  writeXStringSet(fasta[i], file=out_f, format="fasta", width=50)#配列ごとに保存
}

q(save="no")
```



W10-2: ファイル分割2

自作プログラム(fastaLengthFilter.py; 第6回のW12)とLinuxコマンドを組み合わせたやり方。①パスが通っていることを確認、②実行、③description行を表示、④行数は8

```
iu@bielinux[result] pwd [ 5:41午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls [ 5:41午後 ]
corrected.fastq          sequence1_R.fa  sequence4_R.fa
polished_assembly.fasta sequence2_R.fa  smrtpipe.log
polished_assembly.fastq sequence3_R.fa
① iu@bielinux[result] where fastaLengthFilter.py [ 5:41午後 ]
/home/iu/bin/fastaLengthFilter.py
/home/iu/bin/fastaLengthFilter.py
② iu@bielinux[result] fastaLengthFilter.py polished_assembly.fasta 0
> LH_hgap.fa
iu@bielinux[result] ls -l LH* [ 5:41午後 ]
-rwxrwxrwx 1 iu iu 2433662 3月 30 2016 LH_hgap.fa
③ iu@bielinux[result] grep ">" LH_hgap.fa [ 5:41午後 ]
>sequence1
>sequence2
>sequence3
>sequence4
④ iu@bielinux[result] wc LH_hgap.fa [ 5:41午後 ]
      8      8 2433662 LH_hgap.fa
iu@bielinux[result] █ [ 5:41午後 ]
```

W10-3: ファイル分割2

sequence1は、最初の2行分に相当する。それゆえ、①headコマンドで最初の2行のみ抽出した結果をsequence1.faというファイル名で保存している。それ以外はただの確認

```
iu@bielinux[result] pwd [11:27午前]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l LH* [11:27午前]
-rwxrwxrwx 1 iu iu 2433662 3月 30 17:41 LH_hgap.fa
iu@bielinux[result] head -n 2 LH_hgap.fa > sequence1.fa [11:27午前]
iu@bielinux[result] ls -l sequence1.fa [11:27午前]
-rwxrwxrwx 1 iu iu 2289509 4月 1 2016 sequence1.fa
iu@bielinux[result] grep ">" sequence1.fa [11:27午前]
>sequence1
iu@bielinux[result] wc sequence1.fa [11:27午前]
 2      2 2289509 sequence1.fa
iu@bielinux[result] █ [11:27午前]
```



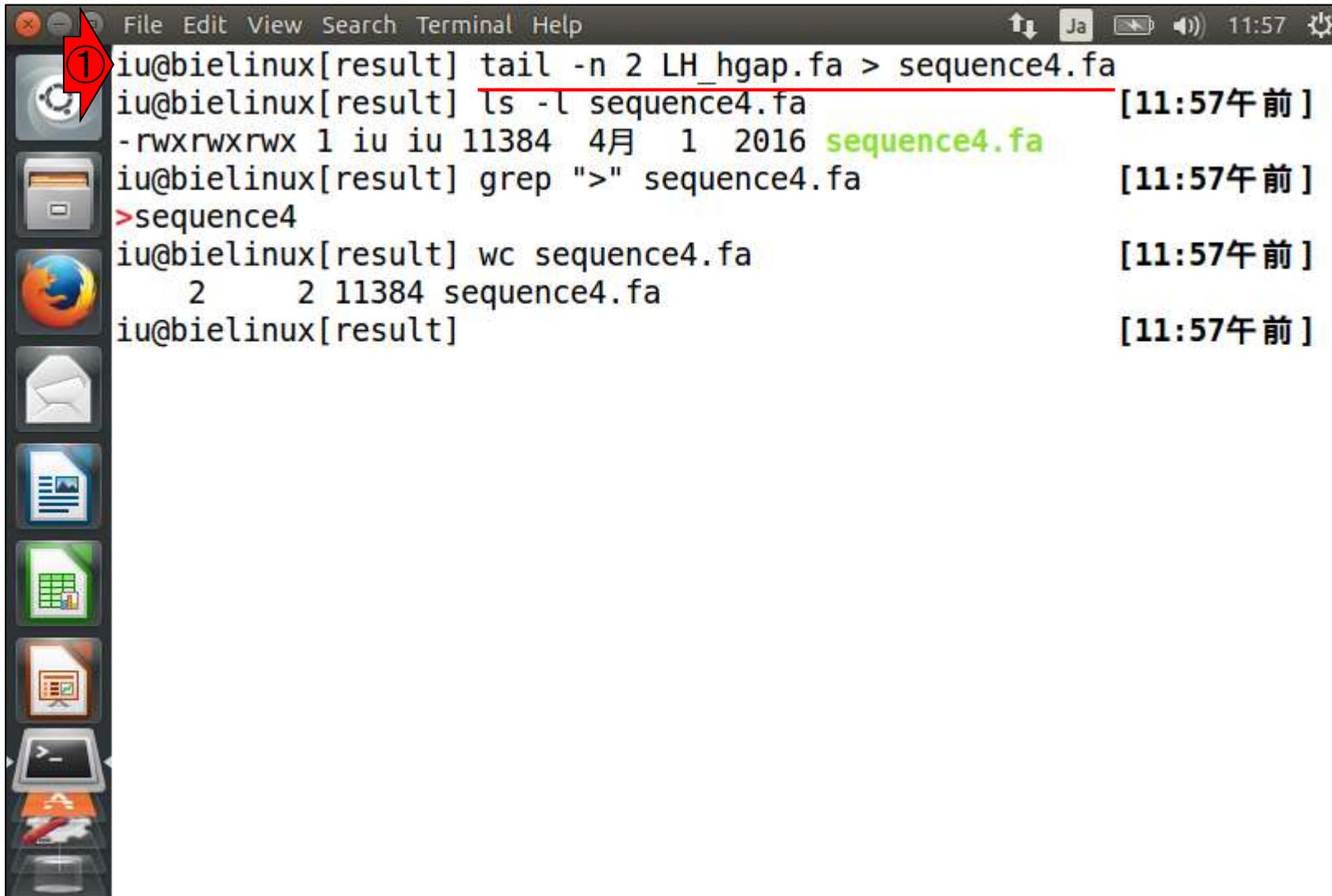
①

W10-4: ファイル分割2

①sequence2と②sequence3は、headとtailを組み合わせて目的の配列のみ抽出。連載第3回のW19-3にもあり

```
iu@bielinux[result] pwd [11:28午前]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l LH* [11:56午前]
-rwxrwxrwx 1 iu iu 2433662 3月 30 17:41 LH_hgap.fa
① iu@bielinux[result] head -n 4 LH_hgap.fa | tail -n 2 > sequence2.fa
iu@bielinux[result] ls -l sequence2.fa [11:56午前]
-rwxrwxrwx 1 iu iu 86904 4月 1 2016 sequence2.fa
iu@bielinux[result] grep ">" sequence2.fa [11:56午前]
>sequence2
iu@bielinux[result] wc sequence2.fa [11:56午前]
 2      2 86904 sequence2.fa
iu@bielinux[result] [11:56午前]
② iu@bielinux[result] head -n 6 LH_hgap.fa | tail -n 2 > sequence3.fa
iu@bielinux[result] ls -l sequence3.fa [11:56午前]
-rwxrwxrwx 1 iu iu 45865 4月 1 2016 sequence3.fa
iu@bielinux[result] grep ">" sequence3.fa [11:56午前]
>sequence3
iu@bielinux[result] wc sequence3.fa [11:56午前]
 2      2 45865 sequence3.fa
iu@bielinux[result] [11:56午前]
```

W10-4: ファイル分割2



```
iu@bielinux[result] tail -n 2 LH_hgap.fa > sequence4.fa
iu@bielinux[result] ls -l sequence4.fa [11:57午前]
-rwxrwxrwx 1 iu iu 11384 4月 1 2016 sequence4.fa
iu@bielinux[result] grep ">" sequence4.fa [11:57午前]
>sequence4
iu@bielinux[result] wc sequence4.fa [11:57午前]
  2    2 11384 sequence4.fa
iu@bielinux[result] [11:57午前]
```

W10-5: 確認

①Rで分割した結果(W10-1)、および②自作プログラムとLinuxコマンドの組み合わせで分割した結果(W10-2からW10-4)。①のほうが改行コードが余分に入っている分だけファイルサイズが大きくなっている。②のほうは、実際の塩基数(2289497, 86892, 45853, and 11372 bp; W9-6)とほぼ同じ

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence* R.fa
-rwxrwxrwx 1 iu iu 2335298 4月 1 12:00 sequence1_R.fa
-rwxrwxrwx 1 iu iu 88641 4月 1 12:00 sequence2_R.fa
-rwxrwxrwx 1 iu iu 46782 4月 1 12:00 sequence3_R.fa
-rwxrwxrwx 1 iu iu 11611 4月 1 12:00 sequence4_R.fa
iu@bielinux[result] ls -l sequence[0-9].fa [12:01午後]
-rwxrwxrwx 1 iu iu 2289509 4月 1 11:27 sequence1.fa
-rwxrwxrwx 1 iu iu 86904 4月 1 11:56 sequence2.fa
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 11384 4月 1 11:57 sequence4.fa
iu@bielinux[result] [12:01午後]
```



-rwxrwxrwx	1	iu	iu	2335298	4月	1	12:00	sequence1_R.fa
-rwxrwxrwx	1	iu	iu	88641	4月	1	12:00	sequence2_R.fa
-rwxrwxrwx	1	iu	iu	46782	4月	1	12:00	sequence3_R.fa
-rwxrwxrwx	1	iu	iu	11611	4月	1	12:00	sequence4_R.fa

-rwxrwxrwx	1	iu	iu	2289509	4月	1	11:27	sequence1.fa
-rwxrwxrwx	1	iu	iu	86904	4月	1	11:56	sequence2.fa
-rwxrwxrwx	1	iu	iu	45865	4月	1	11:56	sequence3.fa
-rwxrwxrwx	1	iu	iu	11384	4月	1	11:57	sequence4.fa

W10-5: 削除

①Rで作成したほうを削除。中身は同じだが、50 bpごとに改行が入っていることを想定しない操作も後に行うため、言われるがまま*_R.faのほうを削除

```
iu@bielinux[result] pwd [12:01午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence*_R.fa [12:01午後]
-rwxrwxrwx 1 iu iu 2335298 4月 1 12:00 sequence1_R.fa
-rwxrwxrwx 1 iu iu 88641 4月 1 12:00 sequence2_R.fa
-rwxrwxrwx 1 iu iu 46782 4月 1 12:00 sequence3_R.fa
-rwxrwxrwx 1 iu iu 11611 4月 1 12:00 sequence4_R.fa
iu@bielinux[result] ls -l sequence[0-9].fa [12:01午後]
-rwxrwxrwx 1 iu iu 2289509 4月 1 11:27 sequence1.fa
-rwxrwxrwx 1 iu iu 86904 4月 1 11:56 sequence2.fa
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 11384 4月 1 11:57 sequence4.fa
① iu@bielinux[result] rm -f sequence*_R.fa [12:01午後]
iu@bielinux[result] ls -l sequence* [12:02午後]
-rwxrwxrwx 1 iu iu 2289509 4月 1 11:27 sequence1.fa
-rwxrwxrwx 1 iu iu 86904 4月 1 11:56 sequence2.fa
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 11384 4月 1 11:57 sequence4.fa
iu@bielinux[result] [12:02午後]
```

W11-1: FASTQ

FASTQファイルも分割。①fastaLengthFilter.pyは、入力としてFASTQファイルを想定していない。実行してもエラーは出ないので一見うまくできたのではないかと思うかもしれない。しかし、FASTQファイルはクオリティスコア情報を含むためFASTAファイルの約2倍のファイルサイズになるはず、という視点で見ると明らかにおかしな出力ファイルになっていることがわかる。②中身を見るまでもなく削除

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l
total 80110
-rwxrwxrwx 1 iu iu 69756928 3月 29 19:12 corrected.fastq
-rwxrwxrwx 1 iu iu 2433662 3月 30 17:41 LH_hgap.fa
-rwxrwxrwx 1 iu iu 2474245 3月 29 19:12 polished_assembly.fasta
-rwxrwxrwx 1 iu iu 4867312 3月 29 19:12 polished_assembly.fastq
-rwxrwxrwx 1 iu iu 2289509 4月 1 11:27 sequence1.fa
-rwxrwxrwx 1 iu iu 86904 4月 1 11:56 sequence2.fa
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 11384 4月 1 11:57 sequence4.fa
-rwxrwxrwx 1 iu iu 64556 3月 29 19:12 smrtpipe.log
① iu@bielinux[result] fastaLengthFilter.py polished_assembly.fastq 0
> LH_hgap.fq
iu@bielinux[result] ls -l LH* [12:13午後]
-rwxrwxrwx 1 iu iu 2433662 3月 30 17:41 LH_hgap.fa
-rwxrwxrwx 1 iu iu 159740 4月 1 2016 LH_hgap.fq
② iu@bielinux[result] rm -f LH_hgap.fq [12:13午後]
iu@bielinux[result] [12:13午後]
```

W11-2: Rは無理

①一番長い2,289,497 bpが原因で、ShortReadパッケージが提供するreadFastq関数では、polished_assembly.fastqファイルを読み込めない。それゆえ、R(正確にはここで示したやり方)ではFASTQファイルをコンティグごとに分割することができない

```
File Edit View Search Terminal Help
Loading required package: BiocParallel
Loading required package: Biostrings
Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for method 'S4Vectors'
Loading required package: IR
Loading required package: XML
Loading required package: RSQLite
Loading required package: Genomic
Loading required package: Genomic
Loading required package: Genomic
> fastq <- readFastq(in_f)
Error: Input/Output
file(s):
  polished_assembly.fastq
message: line too long pol
> fastq
Error: object 'fastq' not found
>
> q(save="no")
iu@bielinux[result] █
```

• W11-2:Rは無理
一番長い2,289,497 bpが原因で、ShortReadパッケージが提供するreadFastq関数では、polished_assembly.fastqファイルを読み込めない。

```
cd ~/Desktop/mac_share/result

pwd
ls -l
R -q

in_f <- "polished_assembly.fastq"
library(ShortRead)
fastq <- readFastq(in_f)
fastq

q(save="no")
```

#入力ファイル名を指定してin_fに格納
#パッケージの読み込み
#ファイルの読み込み
#fastqの中身を確認

#fastqの中身を確認

[12:14午後]



W11-3: FASTQ分割

①W10-3やW10-4のファイル分割のやり方と若干違うのは、このような記述の仕方でもOKであることを示すためです。FASTAとFASTQのサイズ比が1:2になっていることから妥当であると判断できます

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l polished_assembly.fast*      [12:21午後]
-rwxrwxrwx 1 iu iu 2474245  3月 29 19:12 polished_assembly.fasta
-rwxrwxrwx 1 iu iu 4867312  3月 29 19:12 polished_assembly.fastq
iu@bielinux[result] head -n 4 polished_assembly.fastq | tail -n 4 >
sequence1.fq
iu@bielinux[result] head -n 8 polished_assembly.fastq | tail -n 4 >
sequence2.fq
iu@bielinux[result] head -n 12 polished_assembly.fastq | tail -n 4
> sequence3.fq
iu@bielinux[result] head -n 16 polished_assembly.fastq | tail -n 4
> sequence4.fq
iu@bielinux[result] ls -l sequence*                    [12:21午後]
-rwxrwxrwx 1 iu iu 2289509  4月  1 11:27 sequence1.fa
-rwxrwxrwx 1 iu iu 4579015  4月  1  2016 sequence1.fq
-rwxrwxrwx 1 iu iu  86904   4月  1 11:56 sequence2.fa
-rwxrwxrwx 1 iu iu 173805   4月  1  2016 sequence2.fq
-rwxrwxrwx 1 iu iu  45865   4月  1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu  91727   4月  1  2016 sequence3.fq
-rwxrwxrwx 1 iu iu  11384   4月  1 11:57 sequence4.fa
-rwxrwxrwx 1 iu iu  22765   4月  1  2016 sequence4.fq
```



①

W11-3: FASTQ分割

FASTQファイルのdescription部分の行頭は①@および②+なので、念のため両方で調べている。③行数は全部4行

```
iu@bielinux[result] grep "^@" sequence*.fq [12:27午後]
sequence1.fq:@unitig_0|quiver
sequence2.fq:@unitig_2|quiver
sequence3.fq:@unitig_3|quiver
sequence4.fq:@unitig_1|quiver

iu@bielinux[result] grep "^+" sequence*.fq [12:27午後]
sequence1.fq:+
sequence2.fq:+
sequence3.fq:+
sequence4.fq:+

iu@bielinux[result] wc sequence*.fq [12:27午後]
  4      4 4579015 sequence1.fq
  4      4  173805 sequence2.fq
  4      4   91727 sequence3.fq
  4      4   22765 sequence4.fq
 16     16 4867312 total

iu@bielinux[result] █ [12:27午後]
```

W11-4: スコア分布

赤枠部分がFASTQファイル中の1文字表記のクオリティスコアを数値化(PHREDスコアに変換)して保存するコード

W11-4: スコア分布

「前処理 | クオリティチェック | [PHREDスコアに変換](#)」の例題3を参考にしています。par(mar=c(4, 4, 0, 0))で、余白の調整もしています。具体的には、図の下と左側を4行分、それ以外を0行分だけ開けるように指定しています。pngファイル作成(描画)時にいろいろオプション指定している。pch=20はプロット時のマーカーを「小さい黒丸」にせよ、cex=0.5は大きさを通常の0.5倍にせよ、type="p"は、「点プロット(デフォルト)」にせよ、という意味です。

```
R -q
in_f <- "sequence4.fq" #入力ファイル名を指定してin_fに格納
out_f1 <- "sequence4.png" #出力ファイル名を指定してout_f1に格納
out_f2 <- "sequence4.txt" #出力ファイル名を指定してout_f2に格納
param_fig <- c(700, 350) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
#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
fastq <- readFastq(in_f) #in_fで指定したファイルの読み込み
#本番(PHREDスコアに変換)
out <- as(quality(fastq), "matrix") #ASCIIコードのquality scoreをPHRED scoreに変換し
colnames(out) <- 1:ncol(out) #列名を付与
rownames(out) <- as.character(id(fastq)) #行名を付与
#ファイルに保存(pngファイル)
png(out_f1, pointsize=13, width=param_fig[1], height=param_fig[2]) #出力ファイルの各種パラメータを指定
par(mar=c(4, 4, 0, 0)) #下、左、上、右の順で余白(行)を指定
plot(x=1:ncol(out), y=out, pch=20, cex=0.5, #プロット
      type="p", xlab="position", ylab="PHRED score") #プロット
dev.off() #おまじない
#ファイルに保存(テキストファイル)
tmp <- cbind(colnames(out), as.vector(out)) #保存したい情報をtmpに格納
write.table(tmp, out_f2, sep="\t", append=F, quote=F, row.names=F, col.names=F) #tmpの
```

W11-5: 入出力の関係

これは、①sequence4.fq(一番短い11,372 bpのコンティグ)を入力ファイルとして、②2つのファイルを出力するコード。

W11-4: スコア分布

「前処理 | クオリティチェック | [PHREDスコアに変換](#)」の例題3を参考にしています。par(mar=c(4, 4, 0, 0))で、余白の調整もしています。具体的には、図の下と左側を4行分、それ以外を0行分だけ開けるように指定しています。pngファイル作成(描画)時にいろいろオプション指定している。pch=20はプロット時のマーカーを「小さい黒丸」にせよ、cex=0.5は大きさを通常の0.5倍にせよ、type="p"は、「点プロット(デフォルト)」にせよ、という意味です。

```
R -q
in_f <- "sequence4.fq" #入力ファイル名を指定してin_fに格納
out_f1 <- "sequence4.png" #出力ファイル名を指定してout_f1に格納
out_f2 <- "sequence4.txt" #出力ファイル名を指定してout_f2に格納
param_fig <- c(700, 350) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
#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
fastq <- readFastq(in_f) #in_fで指定したファイルの読み込み
#本番(PHREDスコアに変換)
out <- as(quality(fastq), "matrix") #ASCIIコードのquality scoreをPHRED scoreに変換し
colnames(out) <- 1:ncol(out) #列名を付与
rownames(out) <- as.character(id(fastq)) #行名を付与
#ファイルに保存(pngファイル)
png(out_f1, pointsize=13, width=param_fig[1], height=param_fig[2]) #出力ファイルの各種/
par(mar=c(4, 4, 0, 0)) #下、左、上、右の順で余白(行)を指定
plot(x=1:ncol(out), y=out, pch=20, cex=0.5, #プロット
      type="p", xlab="position", ylab="PHRED score") #プロット
dev.off() #おまじない
#ファイルに保存(テキストファイル)
tmp <- cbind(colnames(out), as.vector(out)) #保存したい情報をtmpに格納
write.table(tmp, out_f2, sep="\t", append=F, quote=F, row.names=F, col.names=F) #tmpの
```

コピー実行結果後に①lsで確認。②確かに指定した名前の2つのファイルが作成されていることがわかる

W11-6: 実行結果

W11-4: スコア分布

「前処理 | クオリティチェック | [PHREDスコアに変換](#)」の例題3を参考にしています。par(mar=c(4, 4, 0, 0))で、余白の調整もしています。具体的には、図の下と右側を、行分、それぞれを、行分が1増えるように指定しています。

pngファイル作成(描画)時に、cex=0.5は大きさを通常

```
R -q
in_f <- "sequence4.f
out_f1 <- "sequence4
out_f2 <- "sequence4
param_fig <- c(700,
#必要なパッケージをロー
library(ShortRead)
#入力ファイルの読み込み
fastq <- readFastq(i
#本番(PHREDスコアに変
out <- as(quality(fa
colnames(out) <- 1:n
rownames(out) <- as.
#ファイルに保存(pngフ
png(out_f1, pointsiz
par(mar=c(4, 4, 0, 0
plot(x=1:ncol(out),
type="p", xlab=
dev.off()
#ファイルに保存(テキス
tmp <- cbind(colname
write.table(tmp, out
```

```
File Edit View Search Terminal Help
) #出力ファイルの各種パラメータを指定
> par(mar=c(4, 4, 0, 0)) #下、左、上、右の順で余白
(行)を指定
> plot(x=1:ncol(out), y=out, pch=20, cex=0.5, #プロット
+ type="p", xlab="position", ylab="PHRED score") #プロット
> dev.off() #おまじない
null device
1
> #ファイルに保存(テキストファイル)
> tmp <- cbind(colnames(out), as.vector(out)) #保存したい情報をtmpに
格納
> write.table(tmp, out_f2, sep="\t", append=F, quote=F, row.names=F,
, col.names=F) #tmpの中身を指定したファイル名で保存
> q(save="no")
iu@bielinux[result] pwd [ 4:16午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence4* [ 4:16午後]
-rwxrwxrwx 1 iu iu 11384 4月 1 11:57 sequence4.fa
-rwxrwxrwx 1 iu iu 22765 4月 1 12:21 sequence4.fq
-rwxrwxrwx 1 iu iu 24763 4月 1 16:16 sequence4.png
-rwxrwxrwx 1 iu iu 86006 4月 1 2016 sequence4.txt
iu@bielinux[result]
```



W11-7: pngファイル

①pngファイルのほうは、②で横幅と縦幅のサイズを指定しているのが横長になっている。横軸はコンティグのposition、縦軸はPHREDスコア。数値が大きいほどクオリティが高い

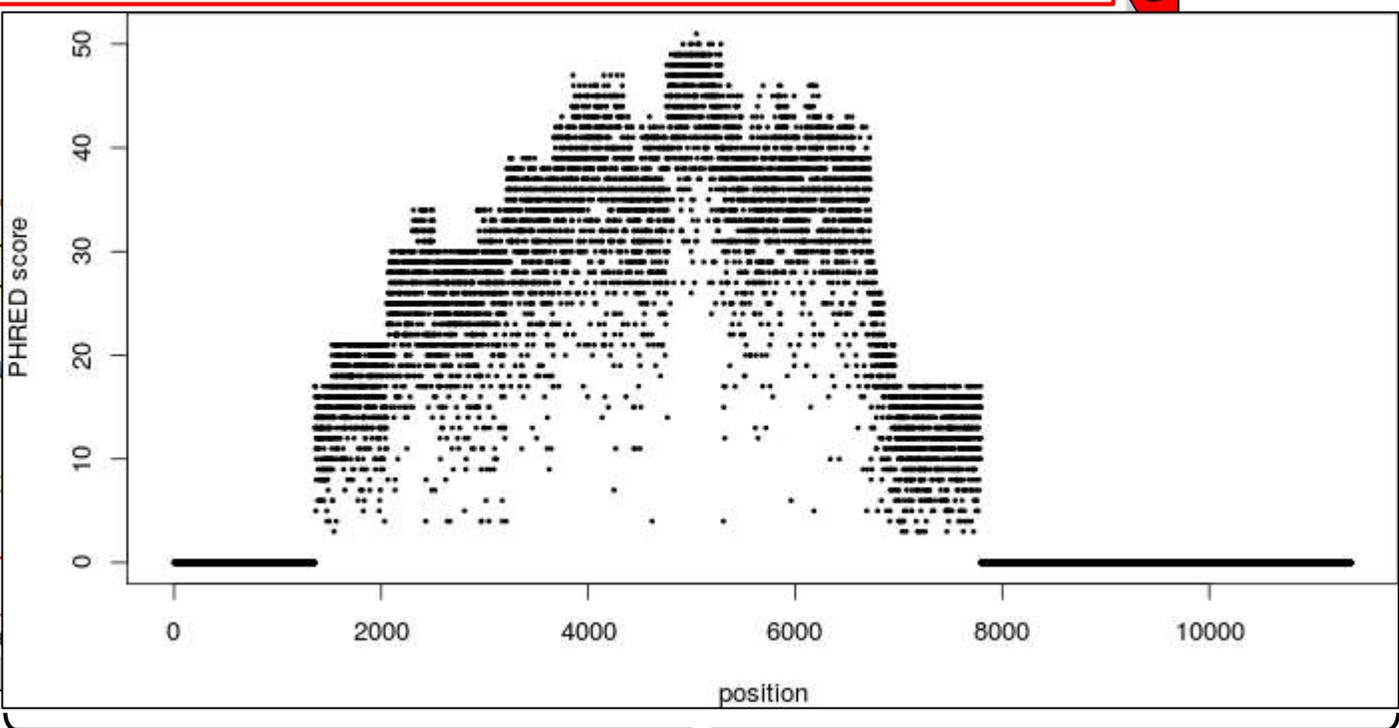
W11-4:スコア分布

「前処理 | クオリティチェック | PHREDスコアに変換」の例題3を参考にしています。par(mar=c(4, 4, 0, 0))で、余白の調整もしています。具体的には、図の下と左側を4行分、それ以外を0行分だけ開けるように指定しています。pngファイル作成(描画)時にいろいろオプション指定している。pch=20はプロット時のマーカーを「小さい黒丸」にせよ、cex=0.5は大きさを通常の0.5倍にせよ、type="p"は、「点プロット(デフォルト)」にせよ、という意味です。

```
R -q
in_f <- "sequence4.fq"
out_f1 <- "sequence4.png"
out_f2 <- "sequence4.txt"
param_fig <- c(700, 350)
```

```
#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
fastq <- readFastq(in_f)
#本番(PHREDスコアに変換)
out <- as(quality(fastq),
colnames(out) <- 1:ncol(out)
rownames(out) <- as.character(1:ncol(out))
#ファイルに保存(pngファイル)
png(out_f1, pointsize=13,
par(mar=c(4, 4, 0, 0))
plot(x=1:ncol(out), y=out,
type="p", xlab="position",
dev.off()
#ファイルに保存(テキストファイル)
tmp <- cbind(colnames(out), out)
write.table(tmp, out_f2, sep=",")
```

#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_f1に格納
#出力ファイル名を指定してout_f2に格納
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)



350 pixel

700 pixel

W11-8: テキストファイル

一番短い11,372 bpのコンティグなので、sequence4.txtは11,372行×2列のファイルになる。①最初の5行分と②最後の4行分を表示。③1列目はposition番号、④2列目がPHREDスコア

W11-4: スコア分布

「前処理 | クオリティチェック | PHREDスコアに変換」の例題3を参考にしています。調整もしています。具体的には、[例題3](#)の「行の長さ」を「列の長さ」に変更し、「行の長さ」を「列の長さ」に変更しています。

pngファイル作成(描画)時に、cex=0.5は大きさを通常

```
R -q
in_f <- "sequence4.f
out_f1 <- "sequence4
out_f2 <- "sequence4
param_fig <- c(700,
#必要なパッケージをロー
library(ShortRead)
#入力ファイルの読み込み
fastq <- readFastq(i
#本番(PHREDスコアに変
out <- as(quality(fa
colnames(out) <- 1:n
rownames(out) <- as.
#ファイルに保存(pngフ
png(out_f1, pointsiz
par(mar=c(4, 4, 0, 0
plot(x=1:ncol(out),
type="p", xlab=
dev.off()
#ファイルに保存(テキス
tmp <- cbind(colname
write.table(tmp, out
```

```
File Edit View Search Terminal Help
iu@bielinux[result] pwd [ 4:19午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence4* [ 4:19午後 ]
-rwxrwxrwx 1 iu iu 11384 4月 1 11:57 sequence4.fa
-rwxrwxrwx 1 iu iu 22765 4月 1 12:21 sequence4.fq
-rwxrwxrwx 1 iu iu 24763 4月 1 16:16 sequence4.png
-rwxrwxrwx 1 iu iu 86006 4月 1 16:16 sequence4.txt
iu@bielinux[result] head -n 5 sequence4.txt [ 4:19午後 ]
1 0
2 0
3 0
4 0
5 0
iu@bielinux[result] tail -n 4 sequence4.txt [ 4:19午後 ]
11369 0
11370 0
11371 0
11372 0
iu@bielinux[result] █ [ 4:19午後 ]
```



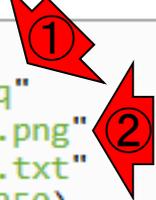
W11-9: sequence3.fq

①2番目に短い45,853 bpのファイル (sequence3.fq) に対しても同様な作業を実行。②pngファイルを眺めると、確かに③コンティグ両末端部分のクオリティが低いことがわかる

• W11-9: sequence3.fq

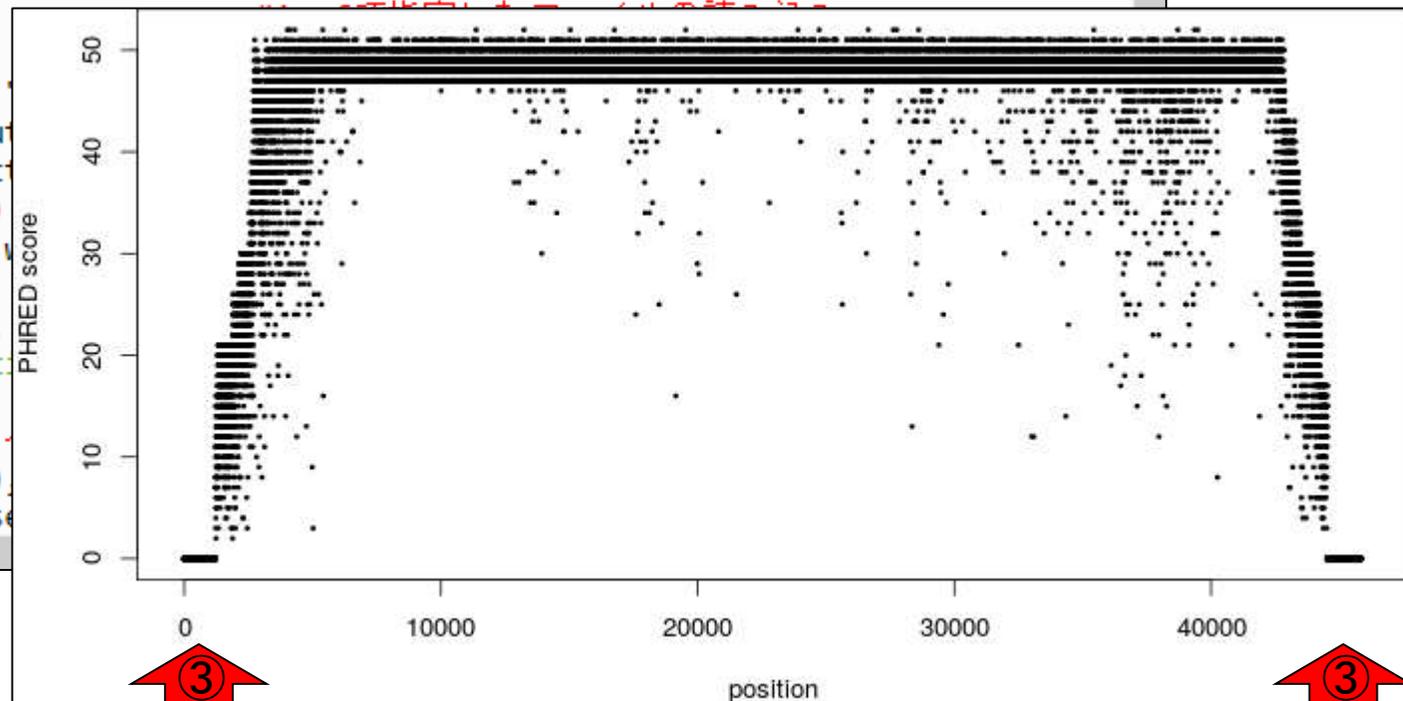
W11-4と基本的に同じで入出力のみ異なる。出力ファイルは、[sequence3.png](#)と[sequence3.txt](#)。

```
R -q
in_f <- "sequence3.fq"
out_f1 <- "sequence3.png"
out_f2 <- "sequence3.txt"
param_fig <- c(700, 350)
#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
fastq <- readFastq(in_f)
#本番(PHREDスコアに変換)
out <- as(quality(fastq),
colnames(out) <- 1:ncol(out)
rownames(out) <- as.character(1:nrow(out))
#ファイルに保存(pngファイル)
png(out_f1, pointsize=13,
par(mar=c(4, 4, 0, 0))
plot(x=1:ncol(out), y=out,
type="p", xlab="position",
dev.off()
#ファイルに保存(テキストファイル)
tmp <- cbind(colnames(out),
write.table(tmp, out_f2, se
```



#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_f1に格納
#出力ファイル名を指定してout_f2に格納
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#パッケージの読み込み



W11-10: sequence2.fq

①3番目に短い(2番目に長い)86,892 bpのファイル(sequence2.fq)に対して同様な作業を実行。②こちらも両末端部分のクオリティが低いことがわかる

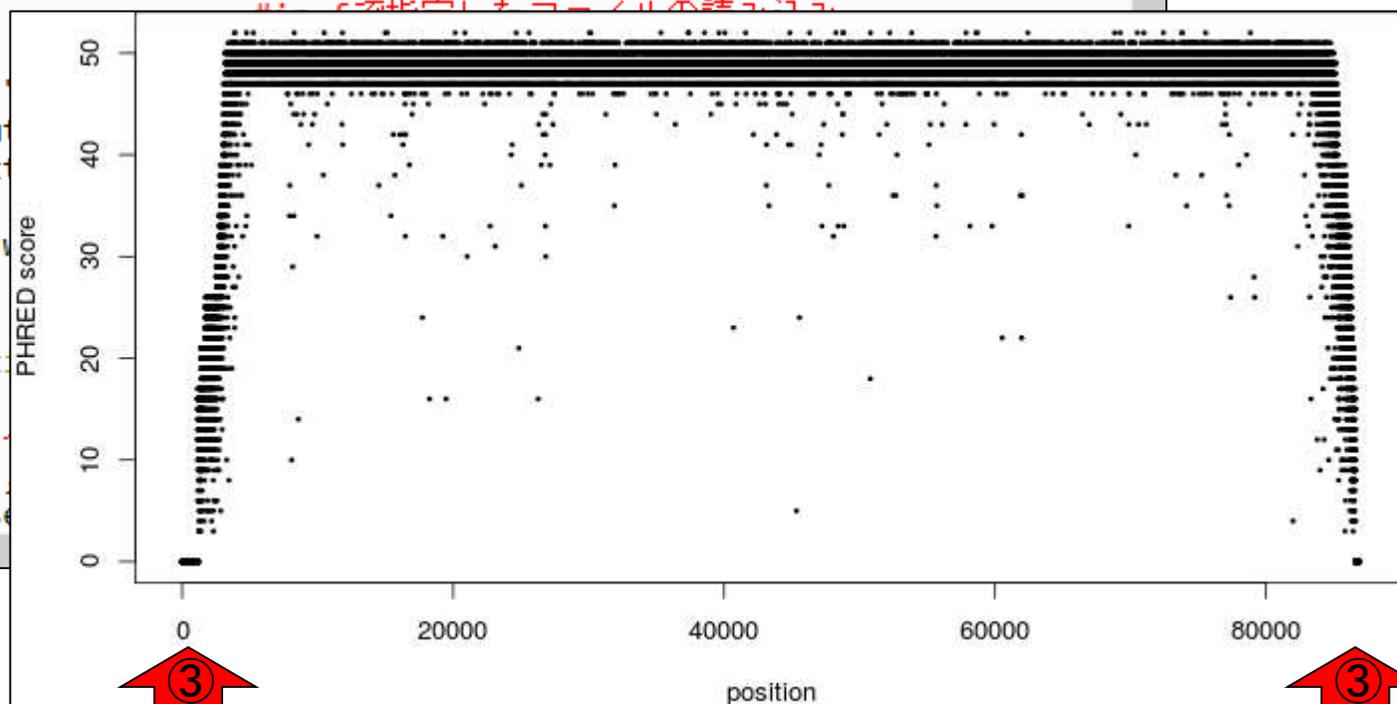
W11-10: sequence2.fq

W11-4と基本的に同じで入出力のみ異なる。出力ファイルは、[sequence2.png](#)と[sequence2.txt](#)。

```
R -q
in_f <- "sequence2.fq"
out_f1 <- "sequence2.png"
out_f2 <- "sequence2.txt"
param_fig <- c(700, 350)
#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
fastq <- readFastq(in_f)
#本番(PHREDスコアに変換)
out <- as(quality(fastq),
colnames(out) <- 1:ncol(out)
rownames(out) <- as.character(1:nrow(out))
#ファイルに保存(pngファイル)
png(out_f1, pointsize=13,
par(mar=c(4, 4, 0, 0))
plot(x=1:ncol(out), y=out,
type="p", xlab="position",
dev.off()
#ファイルに保存(テキストファイル)
tmp <- cbind(colnames(out),
write.table(tmp, out_f2, se
```



#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_f1に格納
#出力ファイル名を指定してout_f2に格納
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)
#パッケージの読み込み



W11-11: sequence1.fq

①最も長い2,289,497 bpのファイル (sequence1.fq)に対しても同様な作業を実行。②readFastq関数実行(FASTQファイル読み込み)部分でエラーが出る(爆)

W11-11: sequence1.fq

W11-4と基本的に同じで入出力のみ異なる。readFastq関数実行(FASTQファイル読み込み)部分でエラーが出る(爆)が原因とされます。このコンティグは2,289,497 bpの最も長いreadが読み込める1行本数の上限を超えていることが原因とされます。

```
R -q
in_f <- "sequence1.fq"
out_f1 <- "sequence1.fq.png"
out_f2 <- "sequence1.fq.txt"
param_fig <- c(700, 700, 700, 700)
#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
fastq <- readFastq(in_f)
#本番(PHREDスコアに変換)
out <- as(quality(fastq), "matrix")
colnames(out) <- 1:ncol(out)
rownames(out) <- as.character(id(fastq))
#ファイルに保存(pngファイル)
png(out_f1, pointsize=param_fig, par=mar=c(4, 4, 0, 0),
     type="p", xlab="Sequence", dev.off())
#ファイルに保存(テキストファイル)
tmp <- cbind(colnames(out), out)
write.table(tmp, out_f2, as.is=T)
```

```
> #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
Error: Input/Output
file(s):
  sequence1.fq
message: line too long sequence1.fq:1
> #本番(PHREDスコアに変換)
> out <- as(quality(fastq), "matrix") #ASCIIコードのquality scoreをPHRED scoreに変換し、データ構造をmatrixにした結果をoutに格納
Error in quality(fastq) :
  error in evaluating the argument 'x' in selecting a method for function 'quality': Error: object 'fastq' not found
> colnames(out) <- 1:ncol(out) #列名を付与
Error in ncol(out) :
  error in evaluating the argument 'x' in selecting a method for function 'ncol': Error: object 'out' not found
> rownames(out) <- as.character(id(fastq))#行名を付与
Error in id(fastq) :
  error in evaluating the argument 'object' in selecting a method for function 'id': Error: object 'fastq' not found
> #ファイルに保存(pngファイル)
```

W11-12: sequence3.txt

sequence3.pngでスコアが0よりも大きくなる境界部分を正確に把握すべくsequence3.txtを調査。①総塩基数は45,853 bp。②行頭と③行末はpngファイルの見た目通り、スコア0。

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3*
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 91727 4月 1 12:21 sequence3.fq
-rwxrwxrwx 1 iu iu 20878 4月 1 15:56 sequence3.png
-rwxrwxrwx 1 iu iu 398859 4月 1 15:26 sequence3.txt
iu@bielinux[result] wc sequence3.txt
45853 91706 398859 sequence3.txt
iu@bielinux[result] head -n 5 sequence3.txt
1      0
2      0
3      0
4      0
5      0
iu@bielinux[result] tail -n 5 sequence3.txt
45849  0
45850  0
45851  0
45852  0
45853  0
iu@bielinux[result] █
```

[5:55午後]

①

②

③

W11-12: sequence3.t

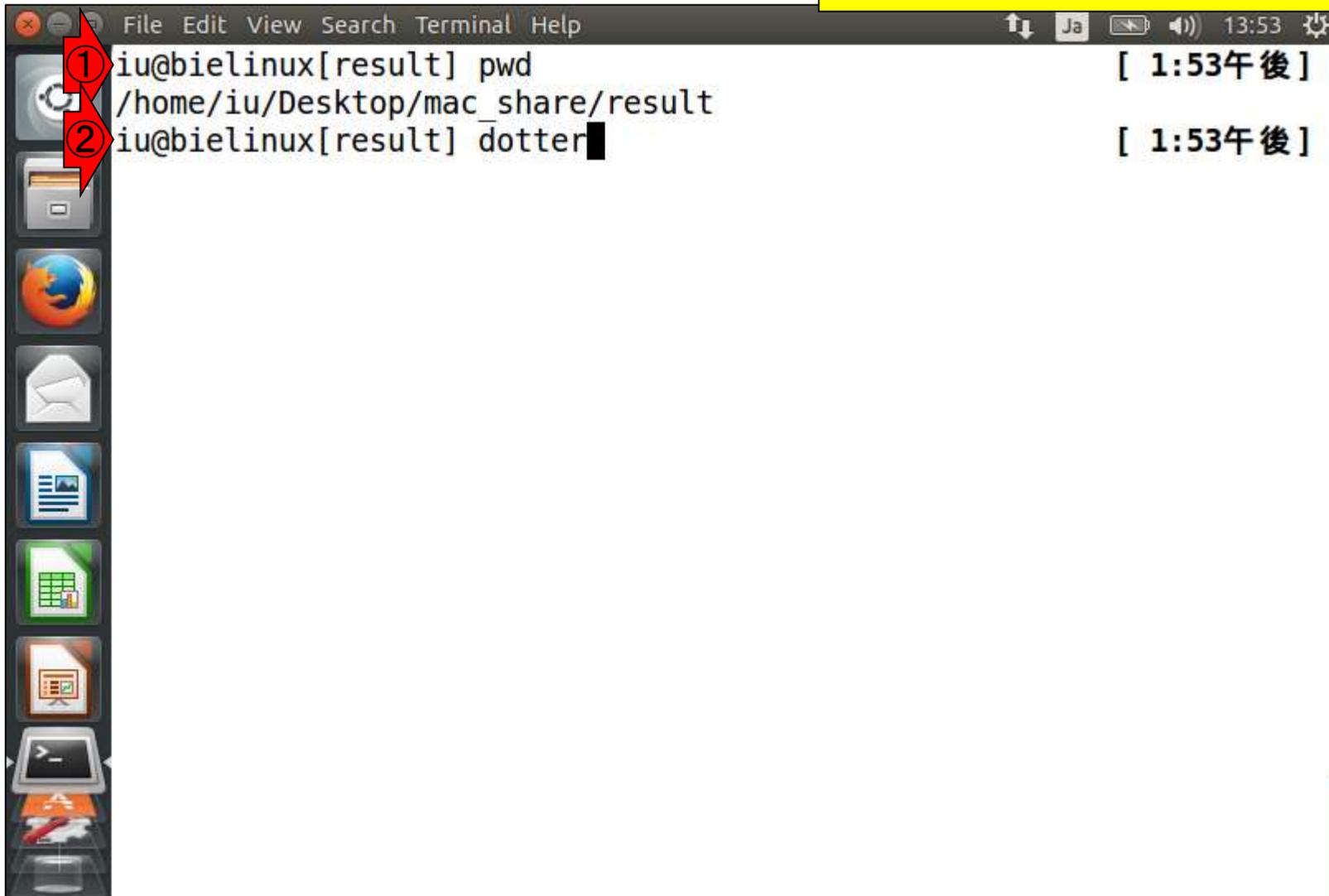
(もちろん裏でざっと眺めて境界領域がわかった上であるが...) ①最初の1,223 bpまでと②44,519 bp以降(最後の1,335 bp)がスコア0

```
iu@bielinux[result] head -n 1227 sequence3.txt | tail -n 8
1220      0
1221      0
1222      0
1223      0
1224      3
1225      7
1226     11
1227     11
iu@bielinux[result] head -n 44522 sequence3.txt | tail -n 8
44515     17
44516     16
44517     15
44518     13
44519      0
44520      0
44521      0
44522      0
iu@bielinux[result] █
```

[7:19午後]

W12-1 : dotter

①作業ディレクトリはどこでもよい。②比較したい2つの配列の類似度を視覚的に評価するために古くから用いられているドットプロット用プログラムdotter



A terminal window with a menu bar (File, Edit, View, Search, Terminal, Help) and a system tray (Ja, 13:53). The terminal shows the following commands and output:

```
iu@bielinux[result] pwd [ 1:53午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] dotter [ 1:53午後]
```

Red arrows labeled 1 and 2 point to the first and second lines of the terminal output, respectively.

W12-1 : dotter

①作業ディレクトリはどこでもよい。②比較したい2つの配列の類似度を視覚的に評価するために古くから用いられているドットプロット用プログラムdotter

```
File Edit View Search Terminal Help 13:59
iu@bielinux[result] pwd [ 1:53午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] dotter [ 1:53午後]

Dotter - Sequence dotplots with image enhancement tools.

Reference: Sonnhammer ELL & Durbin R (1995). A dot-matrix program
with dynamic threshold control suited for genomic DNA and protein
sequence analysis. Gene 167(2):GC1-10.

Usage: dotter [options] <horizontal_sequence> <vertical_sequence>
[X options]

Allowed types:
                Protein      -      Protein
                DNA          -      DNA
                DNA          -      Protein

Options:
-b <file>      Batch mode, write dotplot to <file>
-l <file>      Load dotplot from <file>
-m <float>     Memory usage limit in Mb (default 0.5)
```

W12-1 : dotter

```
File Edit View Search Terminal Help 14:00
-q <int>      Horizontal_sequence offset
-s <int>      Vertical_sequence offset

Some X options:
-acefont <font> Main font.
-font <font> Menu font.

See http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html for more info.

by Erik.Sonnhammer@cgb.ki.se
Version 3.1, compiled Mar 16 2010

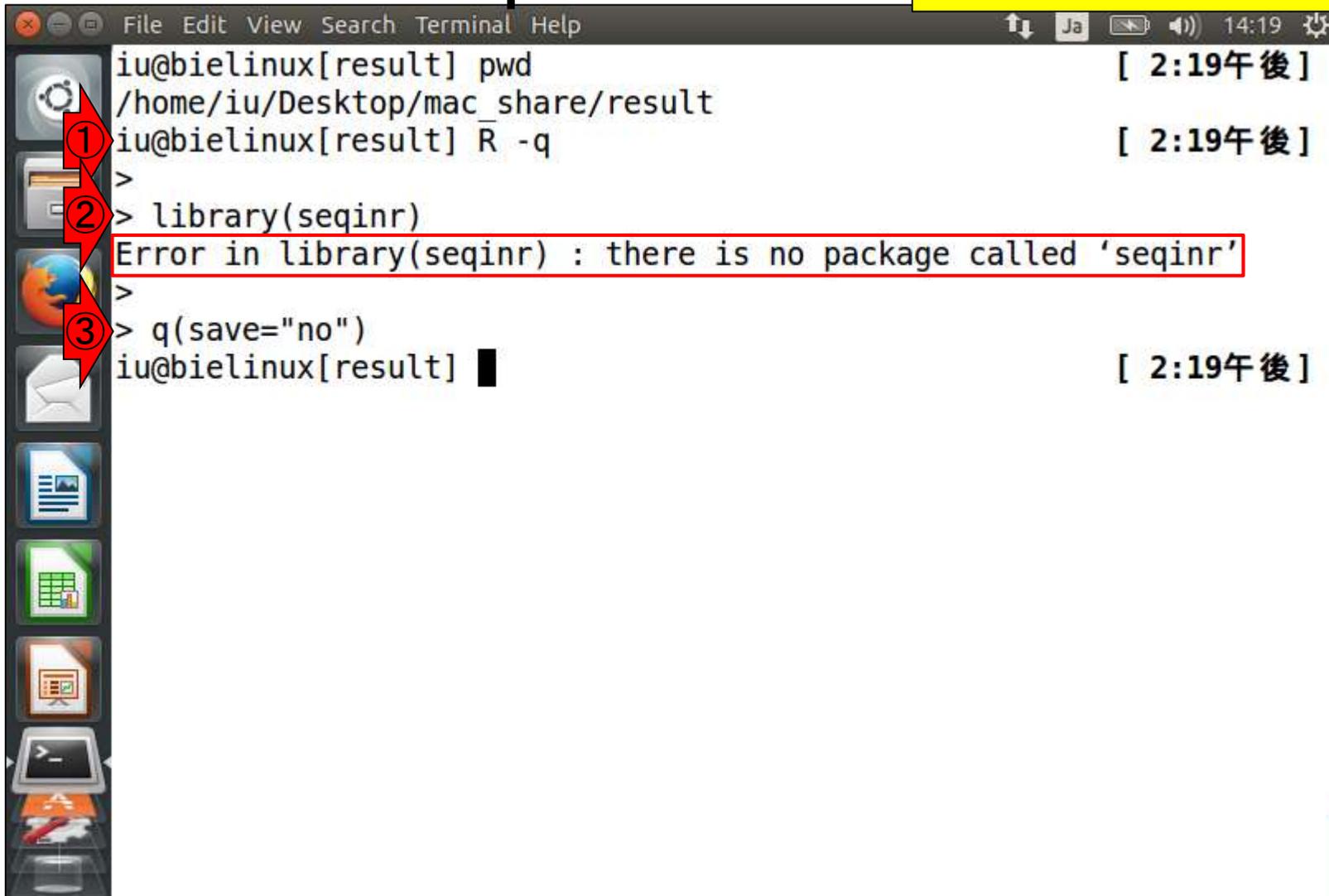
iu@bielinux[result] dotter -v [ 1:58午後 ]
dotter: invalid option -- 'v'

FATAL ERROR: Illegal option
iu@bielinux[result] dotter -h [ 2:00午後 ]
dotter: invalid option -- 'h'

FATAL ERROR: Illegal option
iu@bielinux[result] █ [ 2:00午後 ]
```

W12-2: seqinr

②seqinrというRパッケージは、デフォルトではBio-LinuxのR上にプレインストールされていないため、自分でインストールする必要がある



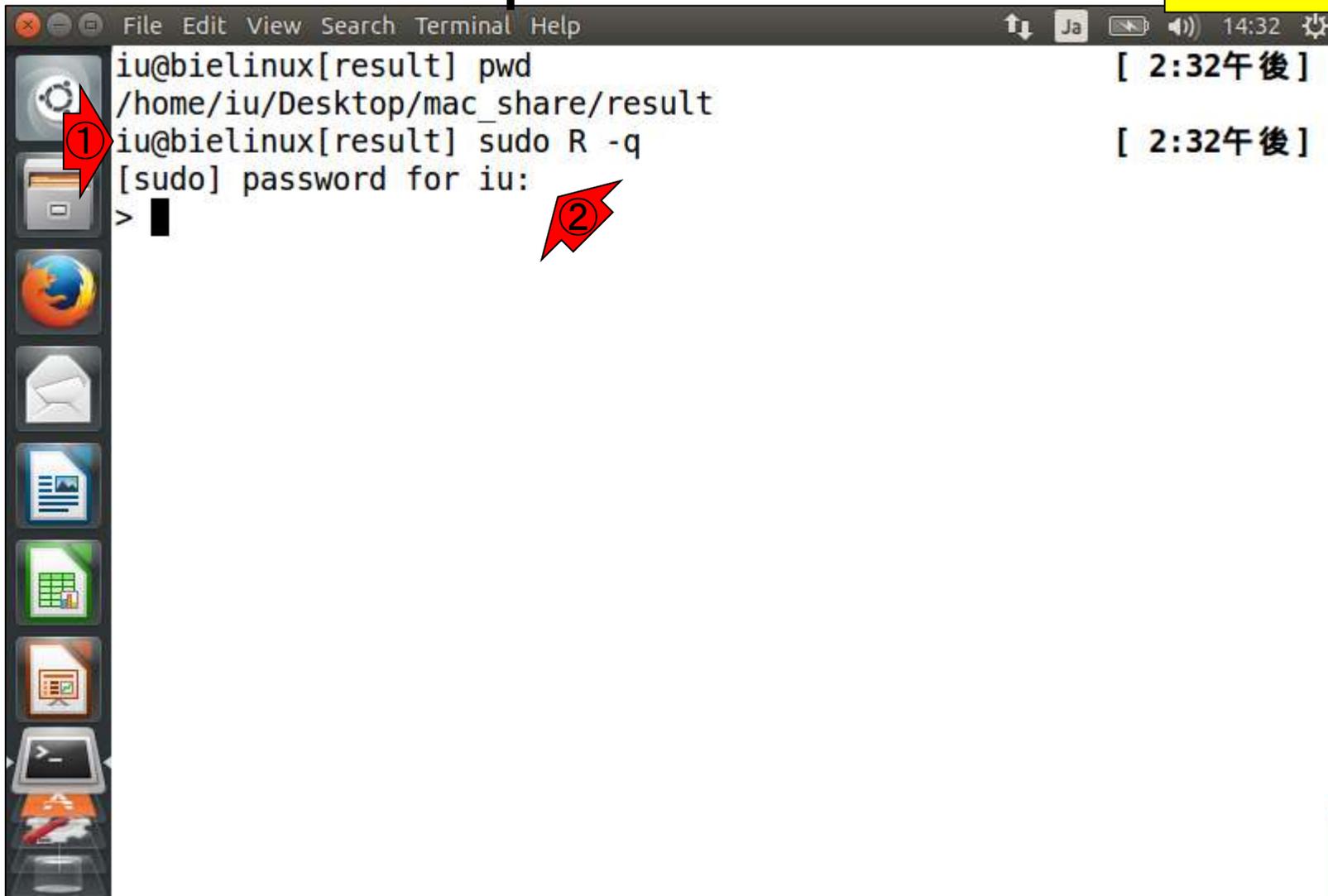
A terminal window showing the following commands and output:

```
iu@bielinux[result] pwd [ 2:19午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] R -q [ 2:19午後]
>
① > library(seqinr)
② Error in library(seqinr) : there is no package called 'seqinr'
③ > q(save="no")
iu@bielinux[result] [ 2:19午後]
```

The error message is highlighted with a red box. Red arrows on the left side of the terminal point to the first three lines of the command sequence.

W12-3: seqinrインストール

①root権限でRを起動。②パスワードを聞かれたら打ち込む(推奨手順通りだとpass1409)



```
iu@bielinux[result] pwd [ 2:32午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] sudo R -q [ 2:32午後]
[sudo] password for iu:
>
```

W12-3: seqinrインストール

①②の一連のコマンドを打ち込むとインストールが始まる。このあたりは第5回W7-4で示したQuasRパッケージのインストール手順と基本的に同じ

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] sudo R -q
[sudo] password for iu:
> source("http://bioconductor.org/biocLite.R")
Bioconductor version 3.1 (BiocInstaller 1.18.5), ?biocLite for help
A newer version of Bioconductor is available for this version of R,
?BiocUpgrade for help
> biocLite("seqinr")
BioC_mirror: http://bioconductor.org
Using Bioconductor version 3.1 (BiocInstaller 1.18.5), R version 3.
2.0.
Installing package(s) 'seqinr'
trying URL 'http://cran.rstudio.com/src/contrib/seqinr_3.1-3.tar.gz'
Content type 'application/x-gzip' length 2056857 bytes (2.0 MB)
=====
```

[2:32午後]

[2:32午後]

①

②

W12-3: seqinrインストール

①パッケージのアップデートをするかどうか聞かれている。(時間がかかるので)②nと打ち込んでリターン

```
File Edit View Search Terminal Help 14:43
installing to /usr/local/lib/R/site-library/seqinr/libs
** R
** data
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (seqinr)

The downloaded source packages are in
  '/tmp/RtmpKi6hmf/downloaded_packages'
Old packages: 'latticeExtra', 'ade4', 'digest', 'evaluate', 'ggplot2',
  'gplots', 'gtable', 'lme4', 'memoise', 'munsell', 'permute', 'Rcpp',
  'RcppEigen', 'RCurl', 'relimp', 'rgl', 'R.methodsS3', 'scales', 'sp',
  'vegan', 'XML', 'xtable', 'boot', 'Matrix', 'mgcv', 'nlme', 'nnet'

Update all/some/none? [a/s/n]: n
```



W12-3: seqinrインストール

①library関数を用いてseqinrパッケージのロードをリトライ。エラーメッセージが消えて、無事ロードできているようなので、②Rの終了。

```
File Edit View Search Terminal Help
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (seqinr)

The downloaded source packages are in
      '/tmp/RtmpKi6hmf/downloaded_packages'
Old packages: 'latticeExtra', 'ade4', 'digest', 'evaluate', 'ggplot2',
              'gplots', 'gtable', 'lme4', 'memoise', 'munsell', 'permute', 'Rcpp',
              'RcppEigen', 'RCurl', 'relimp', 'rgl', 'R.methodsS3', 'scales', 'sp',
              'vegan', 'XML', 'xtable', 'boot', 'Matrix', 'mgcv', 'nlme', 'nnet'

Update all/some/none? [a/s/n]: n
> library(seqinr)
Loading required package: ade4
> q(save="no")
iu@bielinux[result] █
```

[2:55午後]

W12-4: 入力ファイル

seqinrパッケージのdotPlot関数実行時に入力として用いるファイルhoge.faを作成する。①と③は作成したいhoge.faがないことを確認しているだけ

```
File Edit View Search Terminal Help
① iu@bielinux[result] pwd [ 3:51午後 ]
/home/iu/Desktop/mac_share/result
② iu@bielinux[result] ls -l hoge.fa [ 3:52午後 ]
ls: cannot access hoge.fa: No such file or directory
③ iu@bielinux[result] more hoge.fa [ 3:52午後 ]
hoge.fa: No such file or directory
iu@bielinux[result] █ [ 3:52午後 ]
```

W12-4: 入力ファイル

①echoコマンドは、第4回W9あたりでも利用している。ここでは任意の文字列を表示させているだけ。②リダイレクト(>)でhoge.faファイルを新規作成で書き込み。③ACTCGTCAGAという文字列をhoge.faに追加書き込み。④これでFASTA形式ファイルの完成

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l hoge.fa
ls: cannot access hoge.fa: No such file or directory
iu@bielinux[result] more hoge.fa
hoge.fa: No such file or directory
① iu@bielinux[result] echo ">test"
>test
② iu@bielinux[result] echo ">test" > hoge.fa
iu@bielinux[result] more hoge.fa
>test
③ iu@bielinux[result] echo "ACTCGTCAGA" >> hoge.fa
iu@bielinux[result] more hoge.fa
>test
ACTCGTCAGA
④ iu@bielinux[result] █
```

W12-5: dotPlot実行

W12-5: dotPlot実行

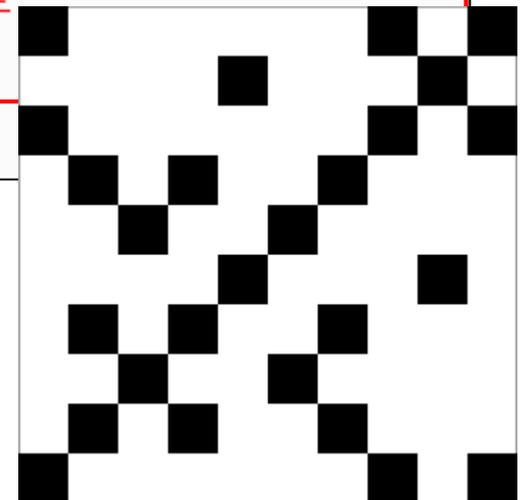
[seqinr](#)パッケージ中のdotPlot関数を用いてドットプロットを作成。

```
cd ~/Desktop/mac_share/result
```

```
R -q
```

```
in_f <- "hoge.fa"           #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.png"       #出力ファイル名を指定してout_fに格納
param_fig <- c(300, 300)   #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
#必要なパッケージをロード
library(seqinr)            #パッケージの読み込み
#入力ファイルの読み込み
hoge <- read.fasta(in_f, seqtype="DNA") #in_fで指定したファイルの読み込み
hoge                          #確認してるだけです
hoge[[1]]                   #確認してるだけです
#ファイルに保存(pngファイル)
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2]) #出力ファイルの各種パラメータを指定
par(mar=c(0, 0, 0, 0))     #下、左、上、右の順で余白(行)を指定
dotPlot(hoge[[1]], hoge[[1]], xlab="", ylab="") #プロット
dev.off()                  #おまじない
```

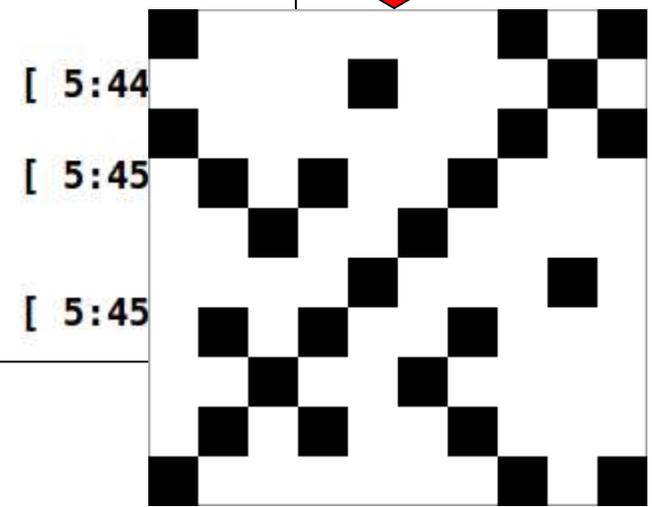
```
q(save="no")
```



コピー実行後に①lsで確認。②hoge1.png
が確かに作成されている。③中身はこれ

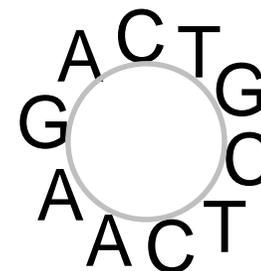
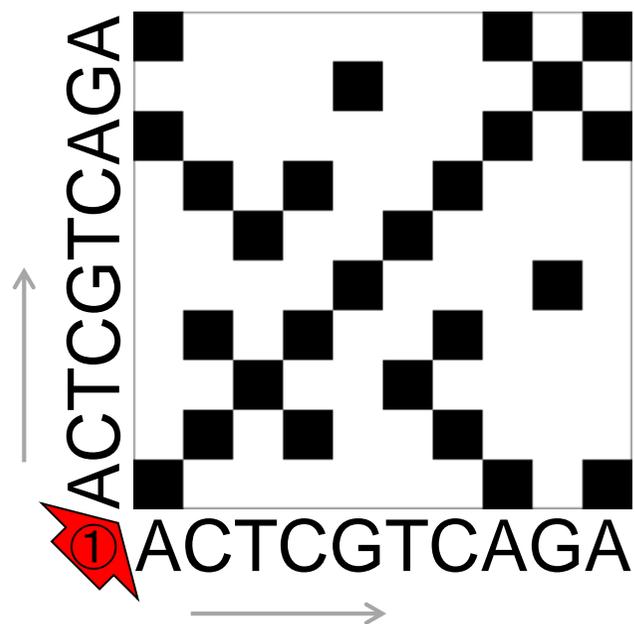
W12-5: dotPlot実行

```
File Edit View Search Terminal Help
[1] "test"
attr(,"Annot")
[1] ">test"
attr(,"class")
[1] "SeqFastadna"
> #ファイルに保存(pngファイル)
> png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])
#出力ファイルの各種パラメータを指定
> par(mar=c(0, 0, 0, 0)) #下、左、上、右の順で余白
(行)を指定
> dotPlot(hoge[[1]], hoge[[1]], xlab="", ylab="")#プロット
> dev.off() #おまじない
null device
      1
>
> q(save="no")
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
① iu@bielinux[result] ls -l hoge*
-rwxrwxrwx 1 iu iu 837  4月  6 17:44 hoge1.png ②
-rwxrwxrwx 1 iu iu  17  4月  6 16:40 hoge.fa
iu@bielinux[result] █
```



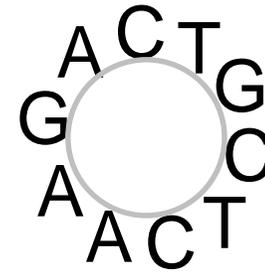
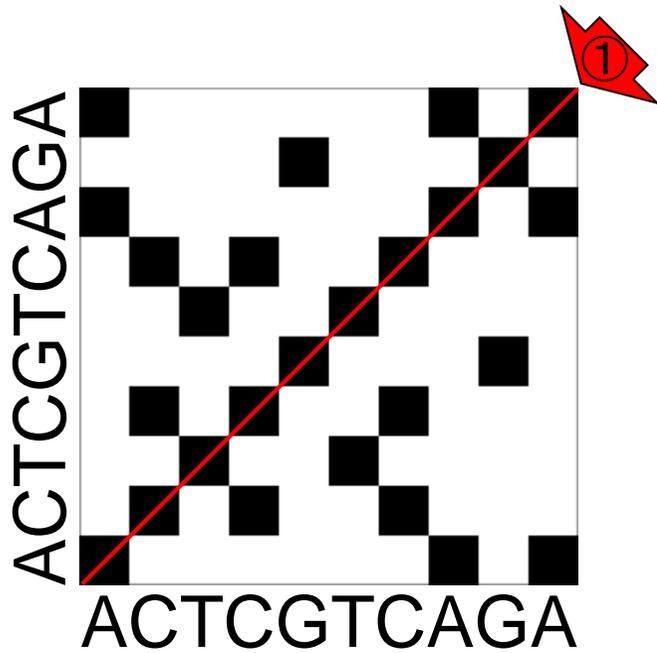
W12-6: 解説

ドットプロットの解説。seqinr中のdotPlot関数実行結果ファイルは、①左下を原点として比較する2つの配列を並べている。一致が黒、不一致が白。



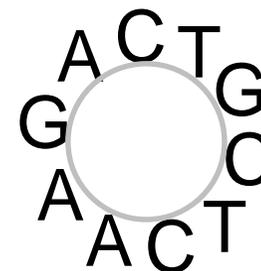
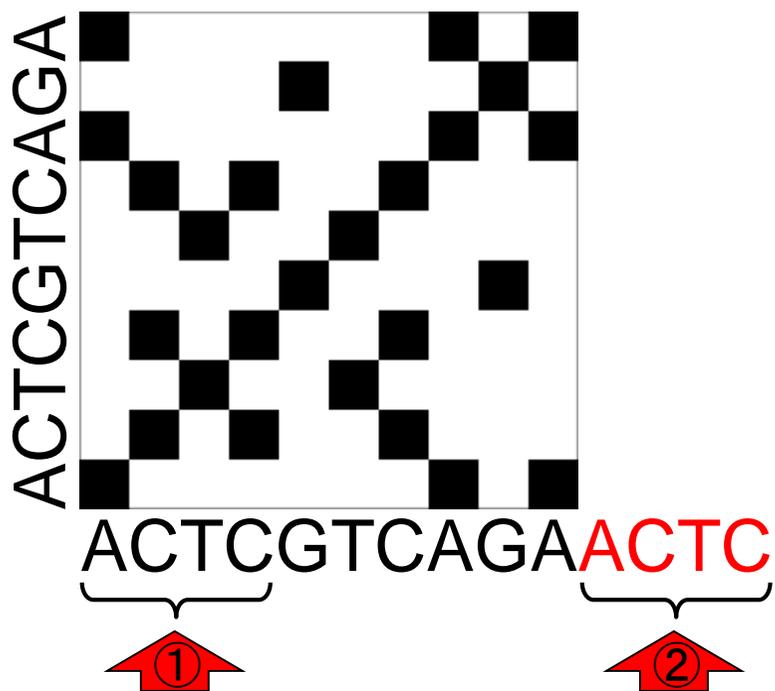
同一の配列を比較するときは、必ず対角線上の塩基が一致(つまり黒)する

W12-6: 解説



W12-7: 環状コンティグ例

アセンブリ結果として、①最初と②最後の末端部分が同じ配列の場合は、通常そのコンティグは環状と判断



W12-7: 環状コンティグ作成

両末端の4塩基がACTCで同じ、計14塩基からなるミニ環状コンティグファイル(hoge2.fa)を作成。最初の10塩基分は、W12-4で作成したhoge.faと同じ

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] echo ">test" > hoge2.fa
iu@bielinux[result] echo "ACTCGTCAGAACTC" >> hoge2.fa
iu@bielinux[result] more hoge2.fa
>test
ACTCGTCAGAACTC
iu@bielinux[result] █
```

[5:46午後]

[5:46午後]

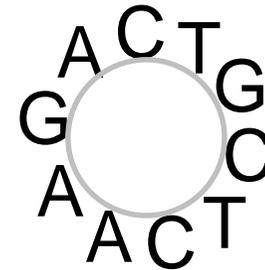
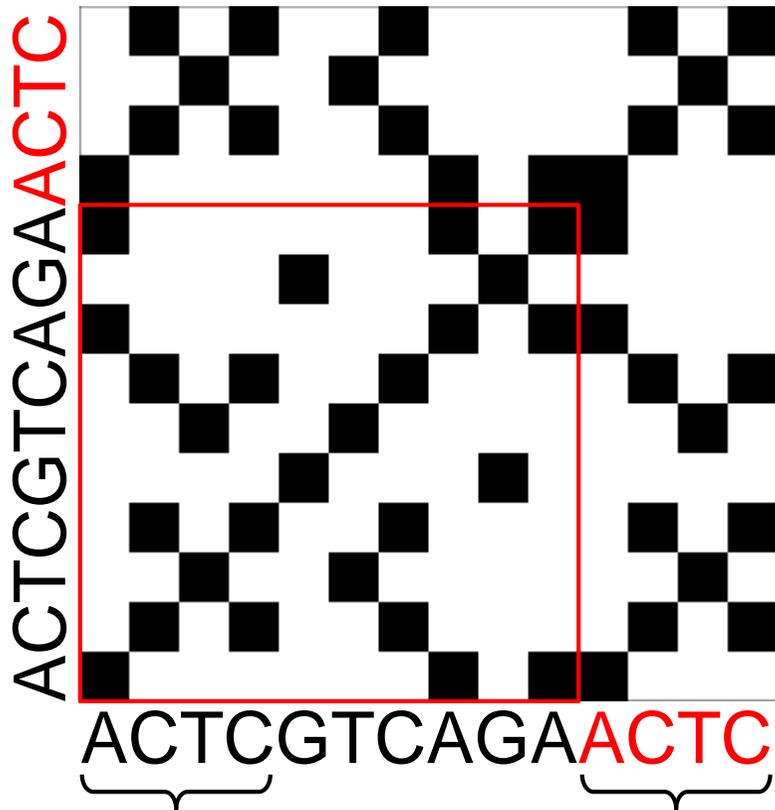
[5:46午後]

[5:46午後]

ACTCGTCAGAACTC

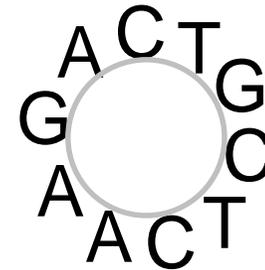
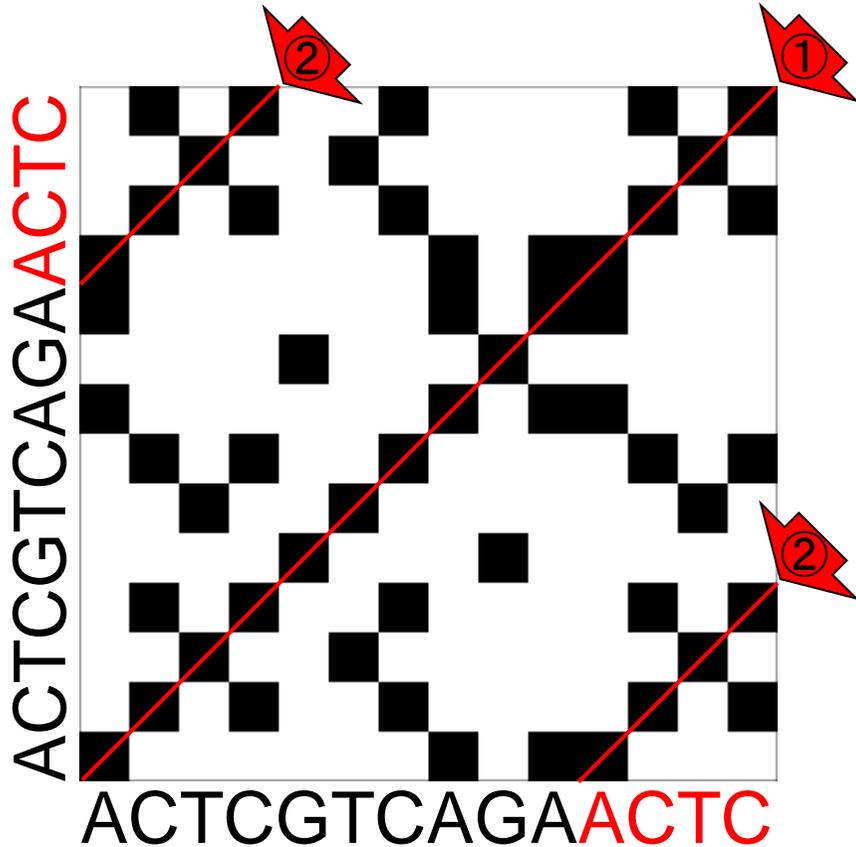
W12-8:ドットプロット

両末端の4塩基がACTCで同じ、計14塩基からなるミニ環状コンティグファイル(hoge2.fa)を入力として、再度ドットプロットを実行した結果。赤枠以外がACTC追加部分



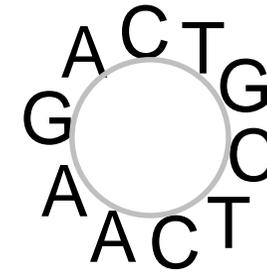
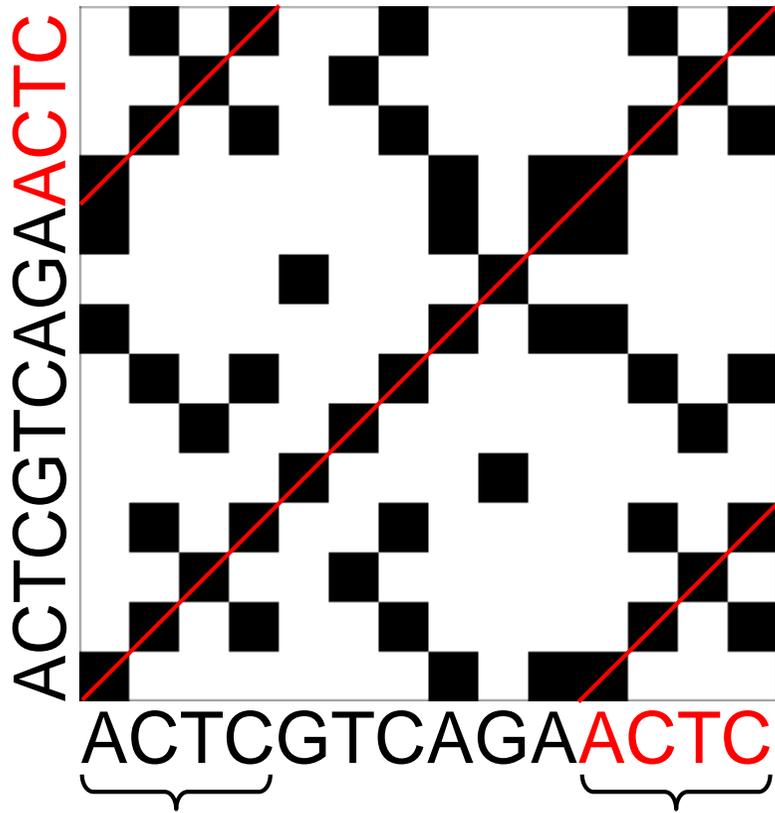
W12-8: 環状の場合

こんな感じに見えます。①対角線上にプロットされるのは同じですが、②対角線と平行に末端部分もプロットされるのが環状の特徴



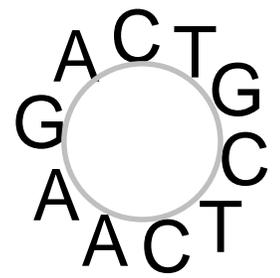
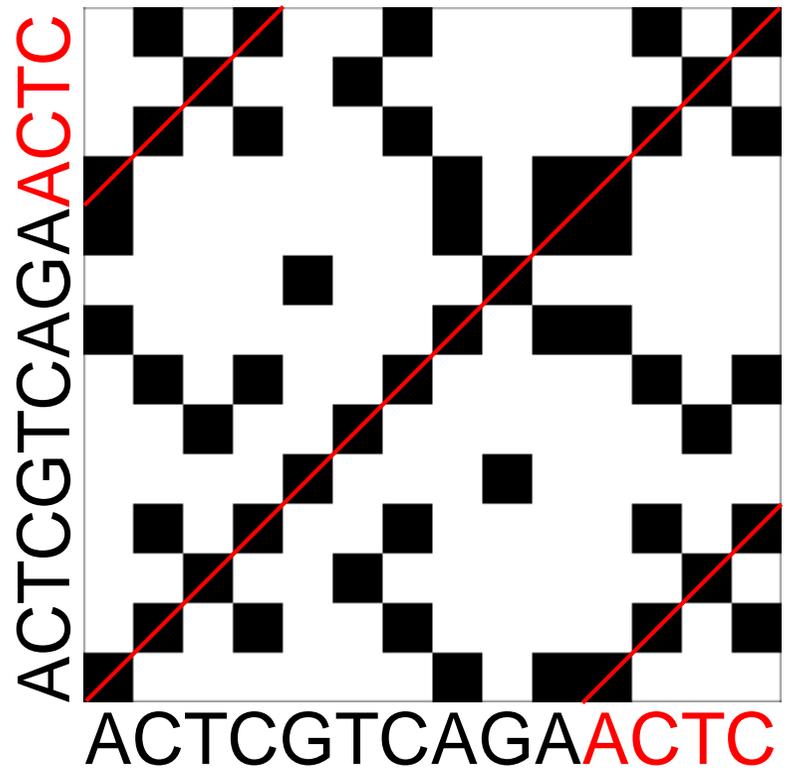
これは、コンティグの両末端が同じ配列であることを意味する。「重複する部分の除去」は、「complete genomeにする操作(finishing)」に相当する

W12-8: 環状の場合



重複除去(トリミング)の選択肢は、(この場合は結果的に同じになるが)①5通り存在する。通常(推奨)は、両末端はクオリティが低いので、アスタリスクのついた中央部分を残して両端をトリムする選択肢を採用する

W13-1: 重複除去



ACTCGTCAGAACTC
 ACTCGTCAGA
 CTCGTCAGAA
 TCGTCAGAAC*
 CGTCAGAACT
 GTCAGAACTC

} ①

W13-2: cutコマンド

特定の範囲の切り出しはcutコマンドを利用。① hoge2.faはsingle-FASTA形式。②「tail -n 1」で、最後の1行分のみ取り出している。③パイプで流してcutコマンドを実行し、3-12文字目を表示。これが④環状コンティグの重複除去後の塩基配列に相当する

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] more hoge2.fa
>test
ACTCGTCAGAACTC
iu@bielinux[result] tail -n 1 hoge2.fa
ACTCGTCAGAACTC
iu@bielinux[result] tail -n 1 hoge2.fa | cut -c 3-12
TCGTCAGAAC
iu@bielinux[result] █
```

```
ACTCGTCAGAACTC
ACTCGTCAGAA
CTCGTCAGAA
TCGTCAGAAC*
CGTCAGAACT
GTCAGAACTC
```



W13-3: 続cutコマンド

①最初の10塩基分のみ取り出しても、②最後の10塩基分のみ取り出しても、環状ゲノムの場合は③アスタリスクのついたやつと結果的に同じ

```
iu@bielinux[result] pwd [10:34午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] tail -n 1 hoge2.fa [10:34午後]
ACTCGTCAGAACTC
① iu@bielinux[result] tail -n 1 hoge2.fa | cut -c -10 [10:34午後]
ACTCGTCAGA
① iu@bielinux[result] tail -n 1 hoge2.fa | cut -c 1-10 [10:34午後]
ACTCGTCAGA
iu@bielinux[result] [10:34午後]
② iu@bielinux[result] tail -n 1 hoge2.fa | cut -c 5- [10:37午後]
GTCAGAACTC
② iu@bielinux[result] tail -n 1 hoge2.fa | cut -c 5-14 [10:37午後]
GTCAGAACTC
iu@bielinux[result] █
```



W13-3: 続cutコマンド

スタート地点をどこにするかという違いのみだから。本物の環状染色体の場合は、特定の遺伝子配列が先頭になるように回転させる慣例がある

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] tail -n 1 hoge2.fa
ACTCGTCAGAACTC
iu@bielinux[result] tail -n 1 hoge2.fa | cut -c -10
ACTCGTCAGA
iu@bielinux[result] tail -n 1 hoge2.fa | cut -c 1-10
ACTCGTCAGA
iu@bielinux[result]
iu@bielinux[result] tail -n 1 hoge2.fa | cut -c 5-
GTCAGAACTC
iu@bielinux[result] tail -n 1 hoge2.fa | cut -c 5-14
GTCAGAACTC
iu@bielinux[result] █
```



```
ACTCGTCAGAACTC
① ACTCGTCAGA
CTCGTCAGAA
TCGTCAGAAC* ③
CGTCAGAACT
GTCAGAACTC ②
```

W14-1 : sequence3.fa

W12-5と同様の手順で、①2番目に短い45,853 bpのファイル(sequence3.fa)を入力としてドットプロットを作成しようとする、②dotPlot関数実行部分でメモリ足りない系のエラーメッセージが出る。(最も短い11,372 bpのsequence4.faでやると同じ部分で20分以上変化がなかったので途中で止めた)

- W14-1: sequence3.faでdotPlot
最後のdotPlot関数実行部分でエラーが出ます。

```
cd ~/Desktop/mac_share/result
pwd
ls -l sequence3*
```

```
R -q
in_f <- "sequence3.fa"
out_f <- "sequence3.png"
param_fig <- c(350, 350)
#必要なパッケージをロード
library(seqinr)
#入力ファイルの読み込み
hoge <- read.fasta(in_f, seqtype="DNA")
hoge
hoge[[1]]
#ファイルに保存(pngファイル)
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])
par(mar=c(0, 0, 0, 0))
dotPlot(hoge[[1]], hoge[[1]], xlab="", ylab="")
dev.off()
q(save="no")
```

①

#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_fに格納
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)
#パッケージの読み込み
#in_fで指定したファイルの読み込み
#確認してるだけです
#確認してるだけです
#出力ファイルの各種パラメータを指定
#下、左、上、右の順で余白(行)を指定
#プロット
#おまじない

②

①この「cannot allocate...」がメモリ
足りない系のエラーメッセージです

W14-1 : sequence3.fa

```
File Edit View Search Terminal Help 00:34
"t" "t" "t"
[45829] "t" "t" "t" "a" "t" "c" "g" "c" "c" "a" "a" "c" "a" "t" "g"
"a" "t" "t"
[45847] "a" "a" "g" "c" "a" "c" "a"
attr(,"name")
[1] "sequence3"
attr(,"Annot")
[1] ">sequence3"
attr(,"class")
[1] "SeqFastadna"
> #ファイルに保存(pngファイル)
> png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])
#出力ファイルの各種パラメータを指定
> par(mar=c(0, 0, 0, 0)) #下、左、上、右の順で余白
(行)を指定
> dotPlot(hoge[[1]], hoge[[1]], xlab="", ylab="")#プロット
① Error: cannot allocate vector of size 15.7 Gb
> dev.off() #おまじない
null device
1
> q(save="no")
iu@bielinux[result] [12:29午前]
```

W14-2: dotter

①sequence3.fa同士のドットプロットをdotterで実行。画面はリターンキーを押して約10秒後の状態。

```
iu@bielinux[result] pwd [ 3:43午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3* [ 3:43午後 ]
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 91727 4月 1 12:21 sequence3.fq
-rwxrwxrwx 1 iu iu 358 4月 10 15:40 sequence3.png
-rwxrwxrwx 1 iu iu 398859 4月 1 15:26 sequence3.txt
iu@bielinux[result] dotter sequence3.fa sequence3.fa [ 3:43午後 ]

Detected sequence types: DNA vs. DNA
Karlin/Altschul statistics for these sequences and score matrix:
K = 0.162
Lambda = 0.177
=> Expected MSP score in a 100x100 matrix = 41.867
Expected residue score in MSP = 1.728
=> Expected MSP length = 24
45853 vs. 45853 residues => 2102.50 million dots. (Takes 2:02 minutes on an SGI MIPS R10000)
```

W14-2: dotter

計3つのウィンドウが立ち上がる。①Greyramp Toolはよくわかりませんが、ドットプロットのコントラスト調整用なのだろうと思います

The screenshot shows the 'Dotter - Alignment Tool' window. At the top, it displays the title 'Dotter sequence3 vs. sequence3'. Below this, there are four lines of sequence data:

```
sequence3: AACCGACCCCTTTTACAACACCTTCCC GGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTA
sequence3: AACCGACCCCTTTTACAACACCTTCCC GGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTA
RevComp: AAAGCTCTTGGTATAAATCCATT AACATGAGGTA AATTCCCATCAAATAGAAAGATCATACTAACTAAGGTATCAAAAT
sequence3: AACCGACCCCTTTTACAACACCTTCCC GGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTA
```

Below the sequences is a dot plot with a vertical axis ranging from 15000 to 25000. A red box highlights a 'Greyramp Tool' window that is open over the dot plot. The tool has a slider set to 40 and buttons for 'Close', 'Swap', and 'Undo'. A red arrow with the number '1' points to the tool's title bar.

W14-2: dotter

計3つのウィンドウが立ち上がる。①Alignment Toolは、比較している2つの配列のアラインメント結果を表示。②今比較しているのは同じ配列なので完全一致。③片方をReverse Complement(逆相補鎖)にしたものとの結果も表示されていることがわかる

The screenshot displays the 'Dotter - Alignment Tool' window. At the top, it shows 'Dotter sequence3 vs. sequence3'. The main area is divided into two sections. The top section shows two identical DNA sequences, 'sequence3:', with a red box highlighting the text. The bottom section shows 'RevComp:' followed by the reverse complement of the sequence, also with a red box highlighting the text. Below the text is a dot plot with a vertical axis ranging from 15000 to 25000. A 'Greyramp Tool' window is open in the bottom right corner, showing a grayscale gradient and a red square. Three red arrows with numbers 1, 2, and 3 point to the Alignment Tool title bar, the top sequence text, and the reverse complement text, respectively.

裏側に見えているのが主目的のドットプロット。①このあたりでクリックして、ドットプロットのウィンドウを手前に表示

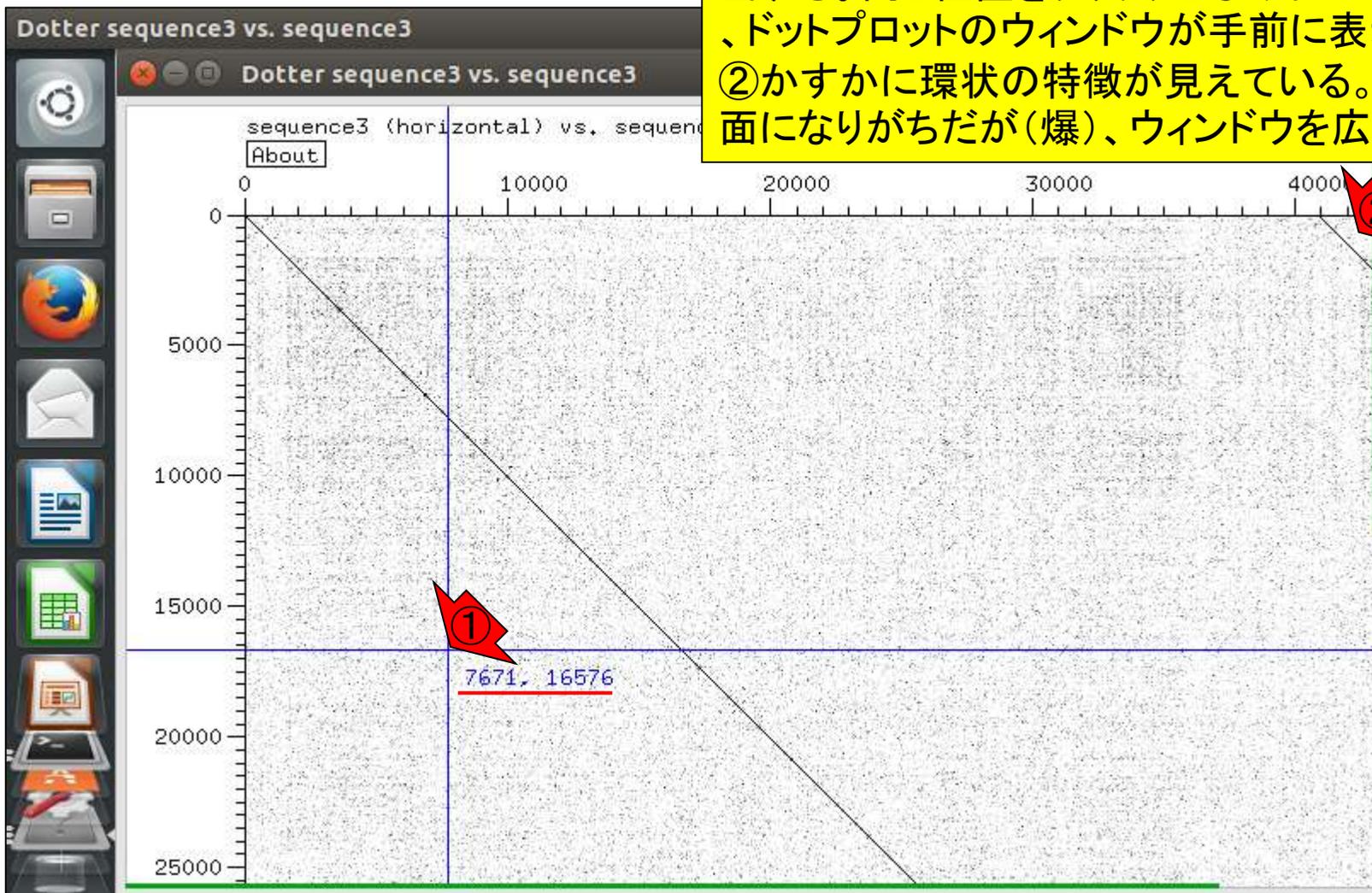
W14-2: dotter

The screenshot shows the 'Dotter - Alignment Tool' window. The top part displays sequence alignment for 'sequence3' and 'RevComp: sequence3'. The bottom part is a dot plot with a vertical axis from 15000 to 25000. A red arrow labeled '1' points to a specific location on the dot plot. A 'Greyramp Tool' window is overlaid on the bottom right, showing a gradient bar and a red square.

```
sequence3: AACCGACCCCTTTTACAACACCTTCCC GGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTA
sequence3: AACCGACCCCTTTTACAACACCTTCCC GGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTA
RevComp: AAAGCTCTTGGTATAAATCCATT AACATGAGGTA AATTCCCATCA AATAGAA GATCATACTAACTAAGGTATCAAAAAT
sequence3: AACCGACCCCTTTTACAACACCTTCCC GGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTA
```

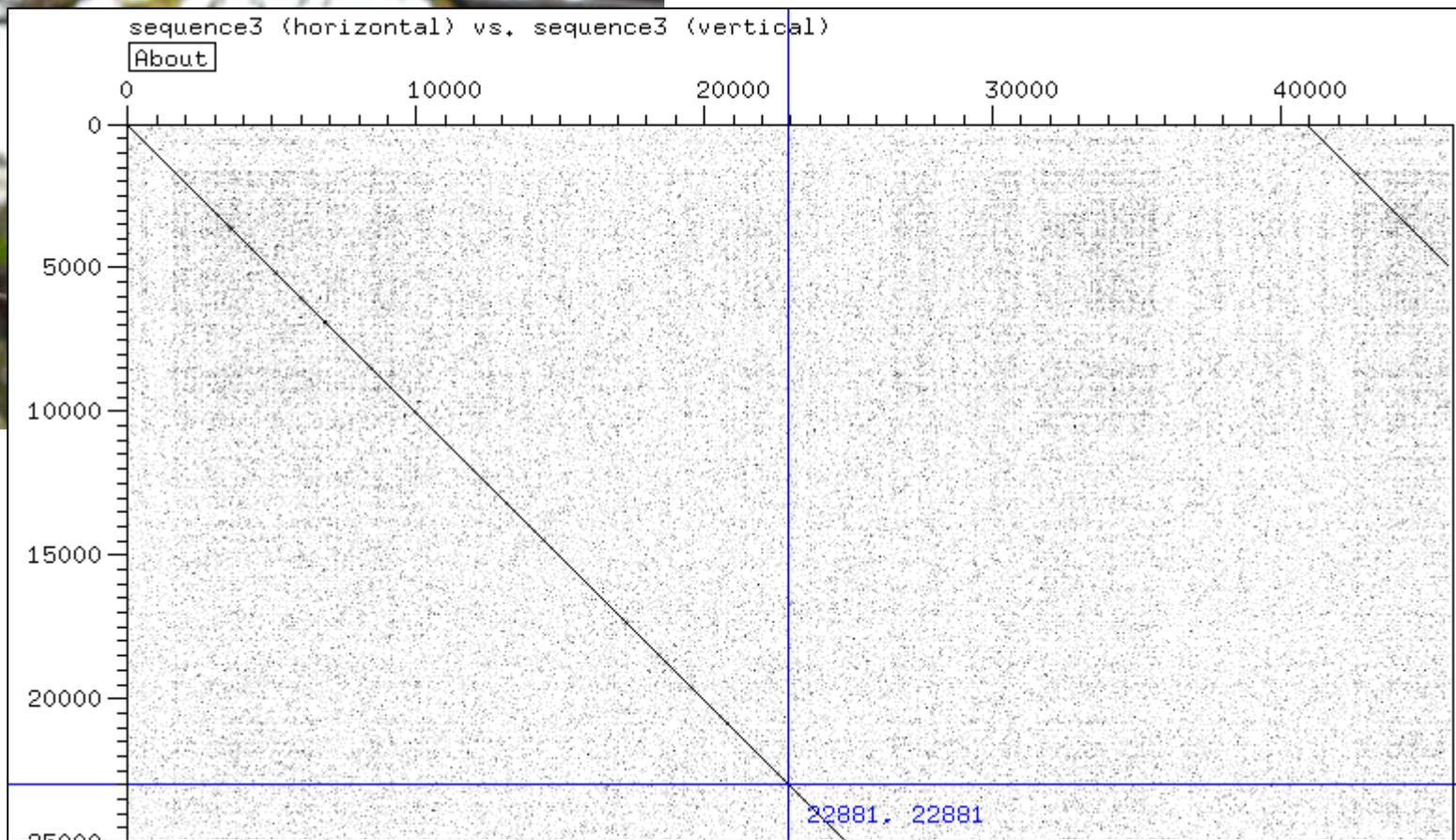
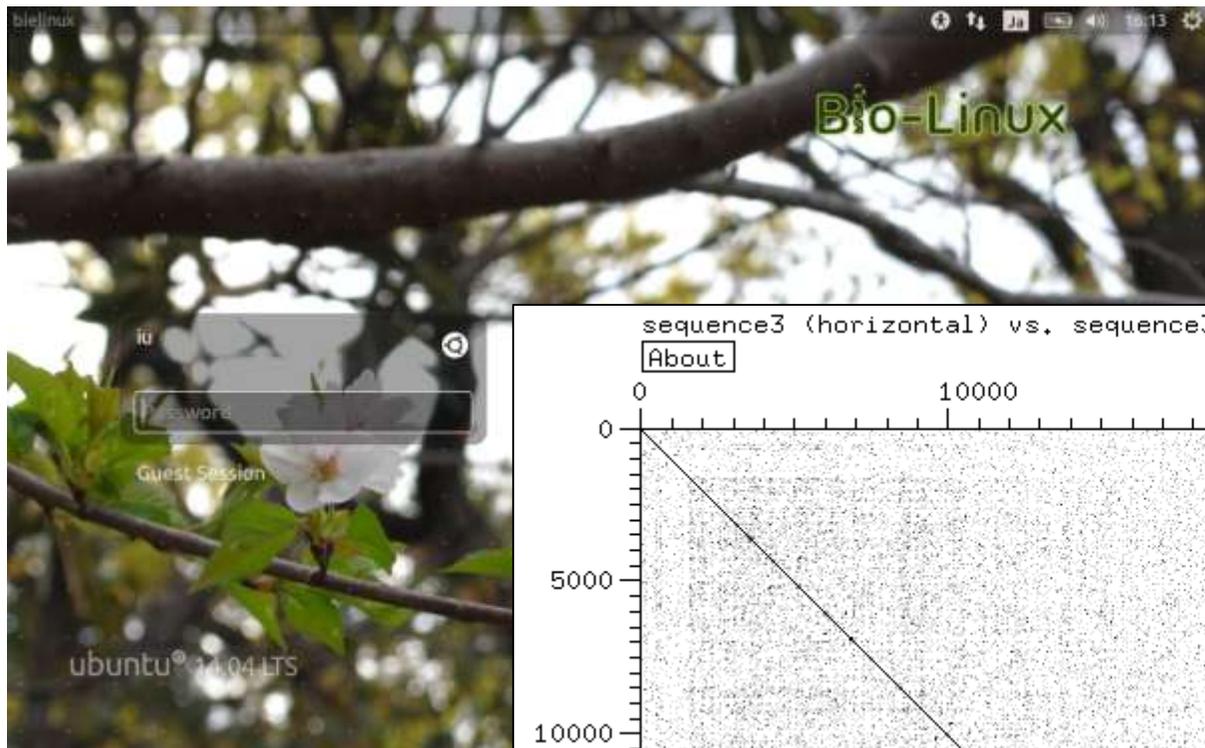
W14-2: dotter

赤下線部分に7671, 16576という数字が見えている。これは今比較している2つの配列の塩基番号(①の座標情報)に相当する。同じ位置をクリックしなければいけないわけではなく、ドットプロットのウィンドウが手前に表示されていればOK。②かすかに環状の特徴が見えている。バグってログイン画面になりがちだが(爆)、ウィンドウを広げて全体像を眺める

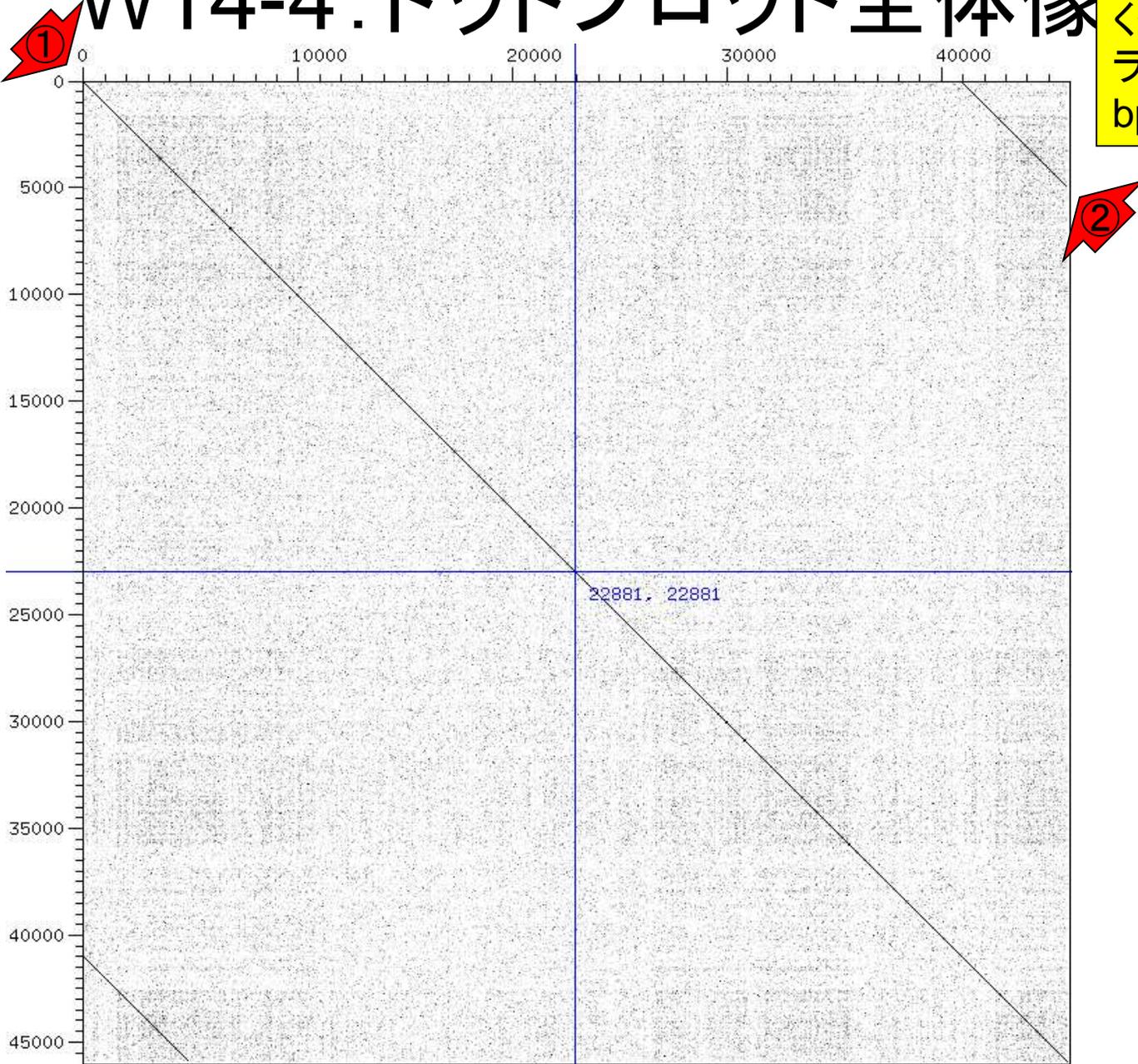


W14-3: バグった例

バグってログイン画面になっても、気を取り直して再挑戦しよう。右下のようなドットプロットが得られるはず



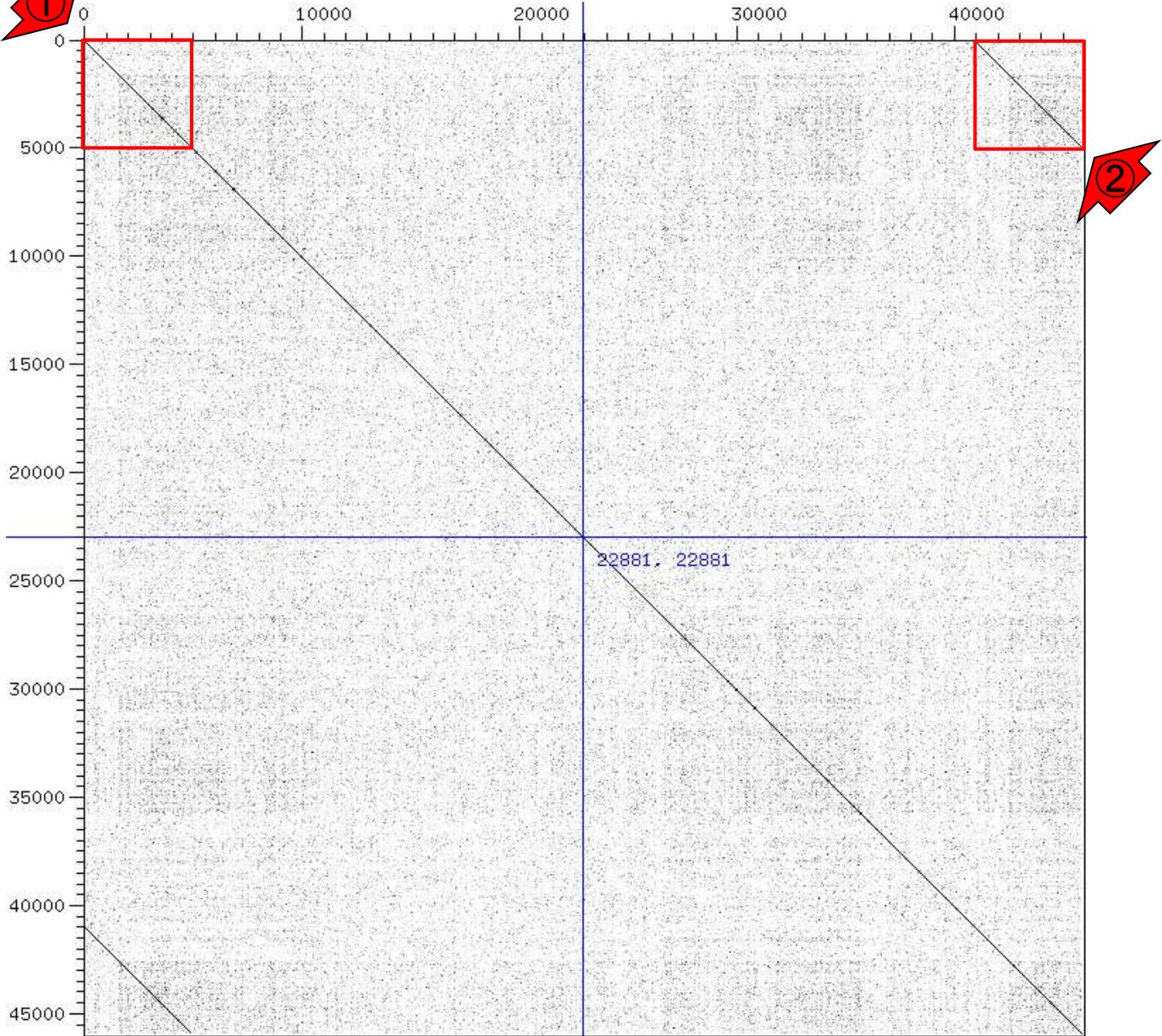
W14-4:ドットプロット全体像



dotterのドットプロットは①左上が原点のようだ。②右上(と左下)にもくっきりと「①の対角線」と平行したラインが見えているので、45,853 bpのsequence3は環状と判断

大まかには、全部で45,853 bpのうち、①最初と②最後の約5,000 bpが重複していると判断

W14-4:ドットプロット全体像



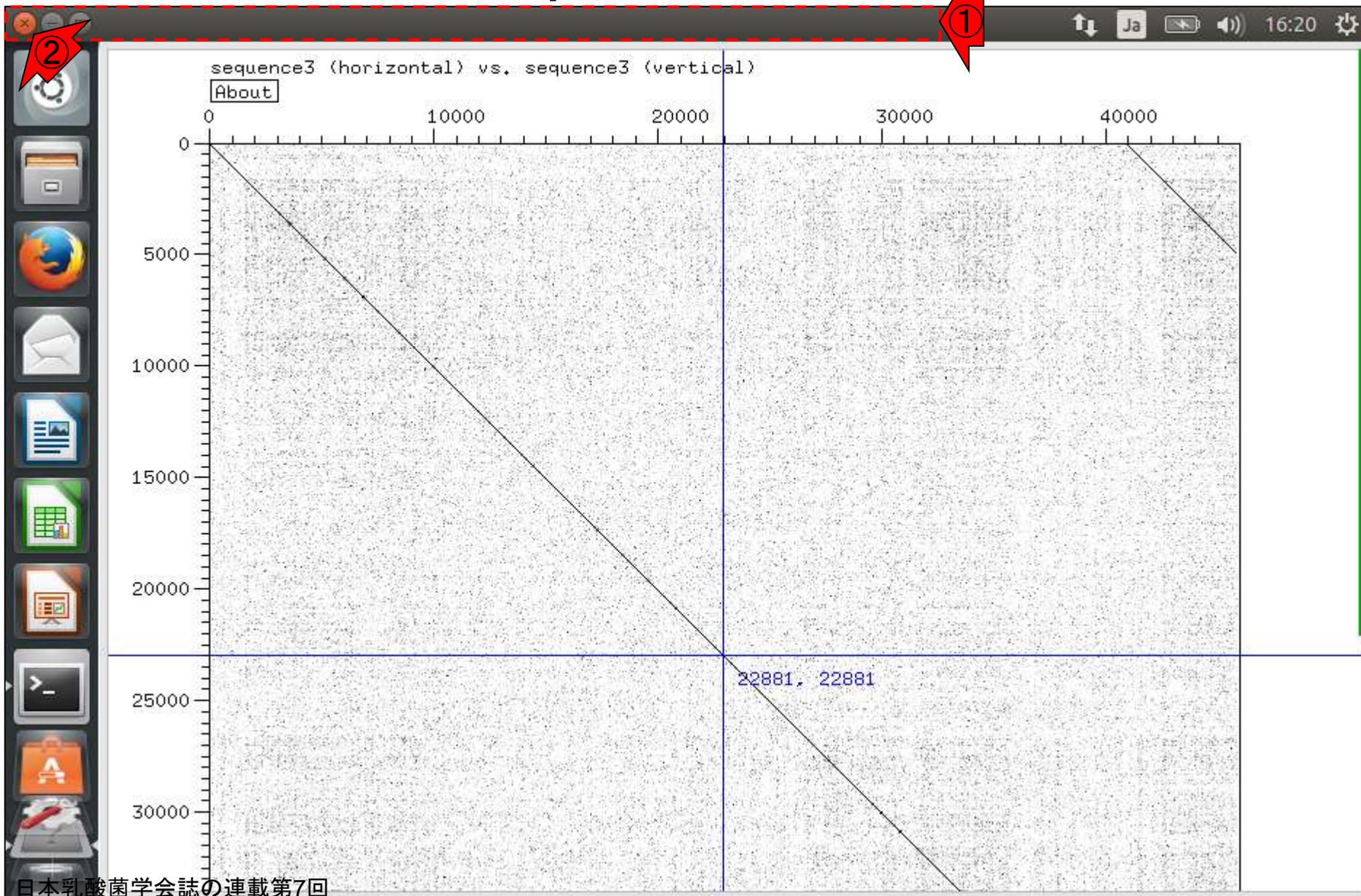
W14-5: dotter終了

Dotter実行結果を終了させるには、基本的に①該当するGUIの左上部分の×を押せばよい

The screenshot displays the 'Dotter - Alignment Tool' window. At the top, the title bar reads 'Dotter sequence3 vs. sequence3'. The main area shows two sequence alignments. The first alignment compares 'sequence3:' with 'sequence3:', showing a perfect match of the sequence: AACCGACCCCTTTTACAACACCTTCCCGGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTATTTGATGGGAATTTAC. The second alignment compares 'RevComp:' with 'sequence3:', showing a reverse complement match: AAAGCTCTTGGTATAAATCCATTACATGAGGTAATTTCCCATCAAATAGAAAGATCATACTAACTAAGGTATCAAATAGTAAGACAGCTTAGT AACCGACCCCTTTTACAACACCTTCCCGGCGGACTAAGCTGTCTTACTATTTTGATACCTTAGTTAGTATGATCTTCTATTTGATGGGAATTTAC. A red arrow with the number '1' points to the close button (X) in the top-left corner of the window. Below the alignment, a 'Greyramp Tool' dialog box is open, featuring a grayscale ramp, a red square, and buttons for 'Close', 'Swap', and 'Undo'. The dialog also has numerical input fields for '40' and '100'. The background shows a large dot plot with a vertical axis ranging from 15000 to 30000.

W14-5: dotter終了

①赤い点線の枠内にカーソルを移動させるとメニューバーが見られるようになるので、②×。第3回W6-3



①コマンド打ち込み可能状態
になっていることがわかります

W14-6: dotter終了後

```
iu@bielinux[result] pwd [ 4:16午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3* [ 4:16午後 ]
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 91727 4月 1 12:21 sequence3.fq
-rwxrwxrwx 1 iu iu 358 4月 10 15:40 sequence3.png
-rwxrwxrwx 1 iu iu 398859 4月 1 15:26 sequence3.txt
iu@bielinux[result] dotter sequence3.fa sequence3.fa [ 4:16午後 ]

Detected sequence types: DNA vs. DNA
Karlin/Altschul statistics for these sequences and score matrix:
K = 0.162
Lambda = 0.177
=> Expected MSP score in a 100x100 matrix = 41.867
Expected residue score in MSP = 1.728
=> Expected MSP length = 24
45853 vs. 45853 residues => 2102.50 million dots. (Takes 2:02 minutes on an SGI
MIPS R10000)

(dotter:24317): Gtk-WARNING **: GtkSpinButton: setting an adjustment with non-z
ero page size is deprecated

(dotter:24317): Gtk-WARNING **: GtkSpinButton: setting an adjustment with non-z
ero page size is deprecated
iu@bielinux[result] [ 4:24午後 ]
```



W15-1 : makeblastdb

①作業ディレクトリはどこでもよい。②
makeblastdbのバージョンは2.2.28+。③「
-h」で大まかな利用法(usage)を確認。④
より詳細な説明は「-help」で出るようだ

```
iu@bielinux[result] pwd [ 4:54午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] makeblastdb -version [ 4:54午後 ]
makeblastdb: 2.2.28+
Package: blast 2.2.28, build Jun  3 2013 11:17:14
iu@bielinux[result] makeblastdb -h [ 4:54午後 ]
USAGE
makeblastdb [-h] [-help] [-in input_file] [-input_type type]
             [-dbtype molecule_type] [-title database_title] [-parse_seqs]
             [-hash_index] [-mask_data mask_data_files] [-gi_mask]
             [-gi_mask_name gi_based_mask_names] [-out database_name]
             [-max_file_sz number_of_bytes] [-taxid TaxID] [-taxid_map TaxIDMa
pFile]
             [-logfile File_Name] [-version]
DESCRIPTION
Application to create BLAST databases, version 2.2.28+
Use '-help' to print detailed descriptions of command line arguments
iu@bielinux[result] [ 4:54午後 ]
```



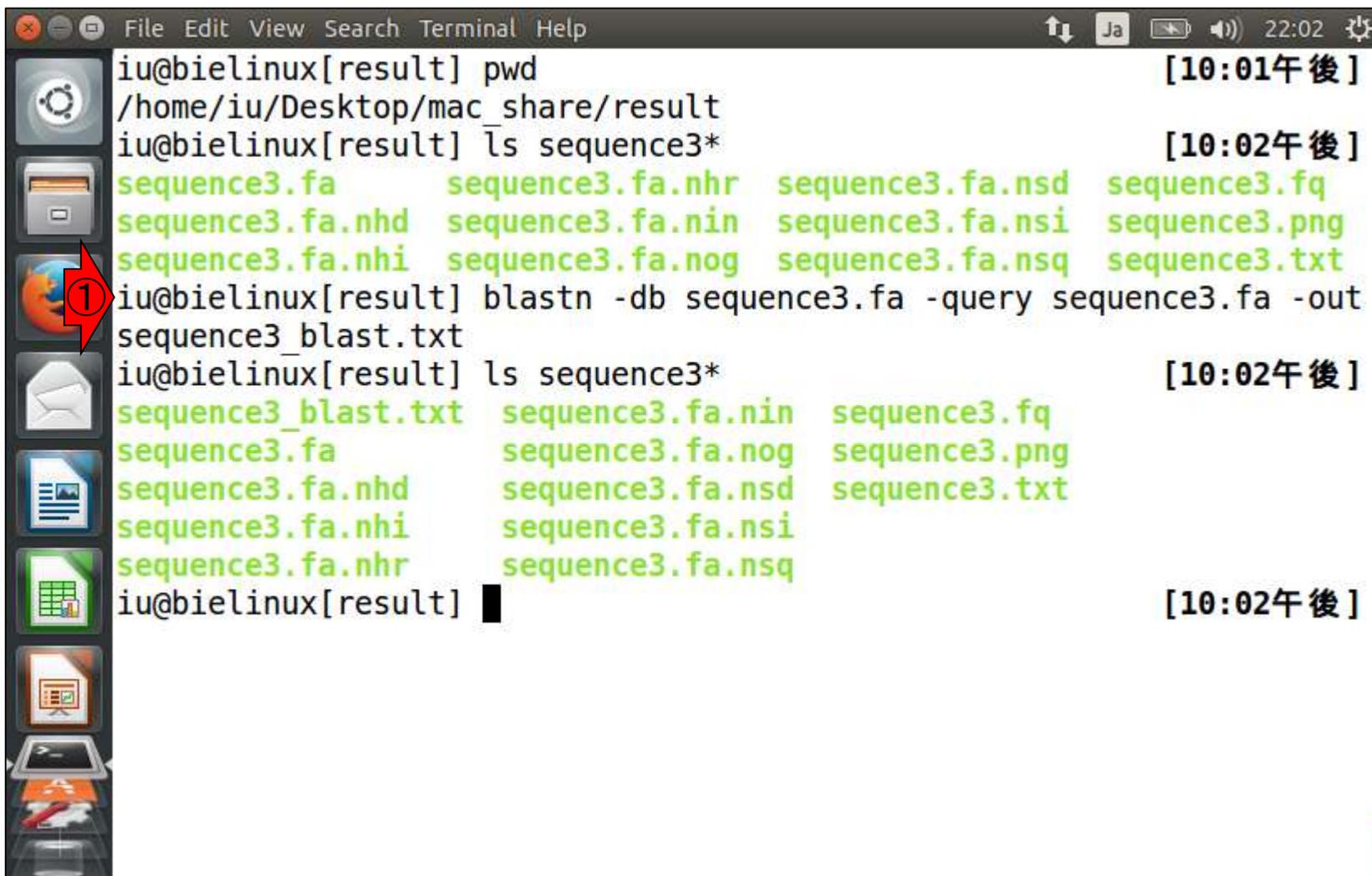
W15-1 : makeblastdb

①makeblastdb本番。入力はsequence3.fa。塩基配列であることを示すnuclを-dbtypeオプションで指定。②実行後は、sequence3.fa.n*というファイルが8個作成されている

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls sequence3.fa*
sequence3.fa
iu@bielinux[result] makeblastdb -in sequence3.fa -dbtype nucl -hash_index
Building a new DB, current time: 04/12/2016 21:27:26
New DB name: sequence3.fa
New DB title: sequence3.fa
Sequence type: Nucleotide
Keep Linkouts: T
Keep MBits: T
Maximum file size: 10000000000B
Adding sequences from FASTA; added 1 sequences in 0.001755 seconds.
iu@bielinux[result] ls sequence3.fa*
sequence3.fa sequence3.fa.nhr sequence3.fa.nsd
sequence3.fa.nhd sequence3.fa.nin sequence3.fa.nsi
sequence3.fa.nhi sequence3.fa.nog sequence3.fa.nsq
iu@bielinux[result]
```

①blastnを実行。DB側、query側はともにsequence3.fa。
。出力ファイル名はsequence3_blast.txt。計算は一瞬

W15-2: blastn



```
iu@bielinux[result] pwd [10:01午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls sequence3* [10:02午後]
sequence3.fa      sequence3.fa.nhr  sequence3.fa.nsd  sequence3.fq
sequence3.fa.nhd  sequence3.fa.nin  sequence3.fa.nsi  sequence3.png
sequence3.fa.nhi  sequence3.fa.nog  sequence3.fa.nsq  sequence3.txt
①iu@bielinux[result] blastn -db sequence3.fa -query sequence3.fa -out
sequence3_blast.txt
iu@bielinux[result] ls sequence3* [10:02午後]
sequence3_blast.txt  sequence3.fa.nin  sequence3.fq
sequence3.fa         sequence3.fa.nog  sequence3.png
sequence3.fa.nhd     sequence3.fa.nsd  sequence3.txt
sequence3.fa.nhi     sequence3.fa.nsi
sequence3.fa.nhr     sequence3.fa.nsq
iu@bielinux[result] [10:02午後]
```

W15-3: 結果を眺める

blastn実行結果ファイルを眺めるべく、① sequence3_blast.txtの最初の10行分(デフォルトが10行)を表示。②行数は3,852行

```
iu@bielinux[result] pwd [10:15午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3_blast.txt [10:15午後]
-rwxrwxrwx 1 iu iu 226098 4月 12 22:02 sequence3_blast.txt
iu@bielinux[result] head sequence3_blast.txt [10:15午後]
BLASTN 2.2.28+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb
Miller (2000), "A greedy algorithm for aligning DNA sequences", J
Comput Biol 2000; 7(1-2):203-14.

Database: sequence3.fa
iu@bielinux[result] wc sequence3_blast.txt [10:15午後]
3852 8793 226098 sequence3_blast.txt
iu@bielinux[result] [10:15午後]
```



W15-4: ヒット数

①BLAST結果ファイル中のヒット総数を把握したい場合は、"Score ="という文字列を含む行数をgrepで調べればよい

BLASTN 2.2.28+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

Database: sequence3.fa
1 sequences; 45,853 total letters

① **Score = 8.468e+04 bits (45853)**, Expect = 0.0
Identities = 45853/45853 (100%), Gaps = 0/45853 (0%)
Strand=Plus/Plus

Query= sequence3

Length=45853

Sequences producing significant alignments:

Score (Bits)	E Value
8.468e+04	0.0

sequence3

> sequence3
Length=45853

Score = 8.468e+04 bits (45853), Expect = 0.0
Identities = 45853/45853 (100%), Gaps = 0/45853 (0%)
Strand=Plus/Plus

```
Query 1 TTTTAGCGGCGGTGTTTGAAC TGCCGCAC TTCTCGAAACACAGTCAATCCTAATTGCCAA 60
      |||
Sbjct 1 TTTTAGCGGCGGTGTTTGAAC TGCCGCAC TTCTCGAAACACAGTCAATCCTAATTGCCAA 60
```

W15-4: ヒット数

- ① "Score =" という文字列を含む行を表示。
- ② その行数は10個。つまりヒット数は10

```
iu@bielinux[result] pwd [11:45午後]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3_blast.txt [11:45午後]
-rwxrwxrwx 1 iu iu 226098 4月 12 22:02 sequence3_blast.txt
iu@bielinux[result] grep "Score =" sequence3_blast.txt [11:45午後]
Score = 8.468e+04 bits (45853), Expect = 0.0
Score = 8844 bits (4789), Expect = 0.0
Score = 8844 bits (4789), Expect = 0.0
Score = 106 bits (57), Expect = 2e-23
Score = 93.5 bits (50), Expect = 2e-19
Score = 93.5 bits (50), Expect = 2e-19
Score = 82.4 bits (44), Expect = 3e-16
Score = 75.0 bits (40), Expect = 6e-14
Score = 75.0 bits (40), Expect = 6e-14
Score = 63.9 bits (34), Expect = 1e-10
iu@bielinux[result] grep -c "Score =" sequence3_blast.txt
10
iu@bielinux[result] [11:45午後]
```



W15-5: grep -n

①BLAST結果ファイル(sequence3_blast.txt)は3,852行だった。②grep実行時に-nをつけることで検索文字列(この場合"Score =")を含む行番号を表示。例えばセカンドヒットは3,092行目、サードヒットは3,425行目などというのがすぐにわかる

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3_blast.txt
-rwxrwxrwx 1 iu iu 226098  4月 12 22:02 sequence3_blast.txt
iu@bielinux[result] wc sequence3_blast.txt
 3852  8793 226098 sequence3_blast.txt
iu@bielinux[result] grep -n "Score =" sequence3_blast.txt
27: Score = 8.468e+04 bits (45853), Expect = 0.0
3092: Score = 8844 bits (4789), Expect = 0.0
3425: Score = 8844 bits (4789), Expect = 0.0
3758: Score = 106 bits (57), Expect = 2e-23
3771: Score = 93.5 bits (50), Expect = 2e-19
3784: Score = 93.5 bits (50), Expect = 2e-19
3797: Score = 82.4 bits (44), Expect = 3e-16
3806: Score = 75.0 bits (40), Expect = 6e-14
3815: Score = 75.0 bits (40), Expect = 6e-14
3824: Score = 63.9 bits (34), Expect = 1e-10
iu@bielinux[result]
```

[4:08午後]

W15-6: grep -A

①grep -Aオプションで一致した行を含め後ろの3行分を表示。
②1位は、一致領域が45,853 bpで100%一致、Gapsも0/45853。W14-4のドットプロットの③対角線の見栄えと一致

```
File Edit View Search Terminal Help
iu@bielinux[result] grep -A 3 "Score =" sequence3_blast.txt
Score = 8.468e+04 bits (45853), Expect = 0.0
Identities = 45853/45853 (100%), Gaps = 0/45853 (0%)
Strand=Plus/Plus

--

Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus

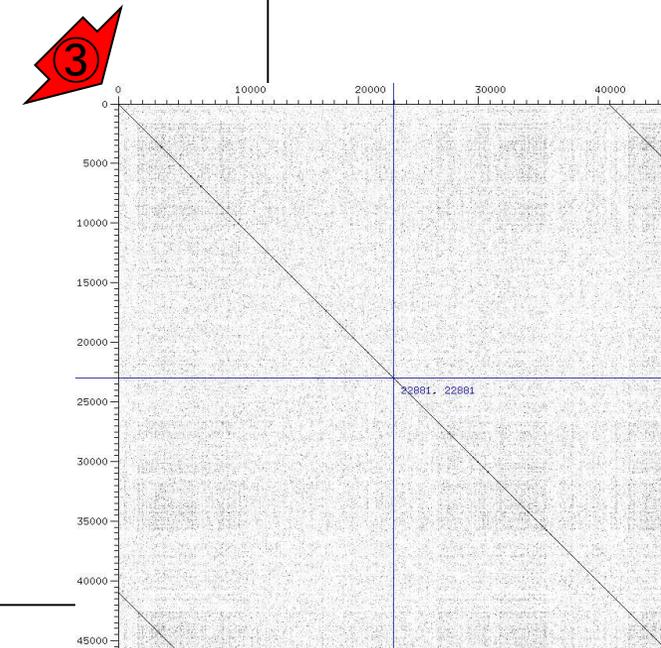
--

Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus

--

Score = 106 bits (57), Expect = 2e-23
Identities = 68/73 (93%), Gaps = 2/73 (3%)
Strand=Plus/Minus

--
```



W15-6: grep -A

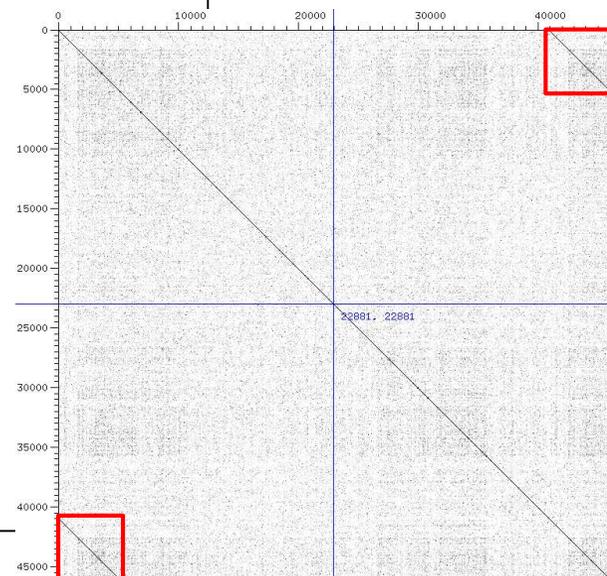
同率2位の①と②の2つのヒットは、一致領域が4,901 bpとなっており、W14-4のドットプロット上の2つの赤四角の見栄えと一致する。どちらがどちらに相当するかは、W15-5で得られた①セカンドヒットの3,092行目以降、②サードヒットの3,425行目以降のアラインメントを眺めればよい

```
iu@bielinux[result] grep -A 3 "Score =" seq
Score = 8.468e+04 bits (45853), Expect =
Identities = 45853/45853 (100%), Gaps = 0/
Strand=Plus/Plus
```

```
Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus
```

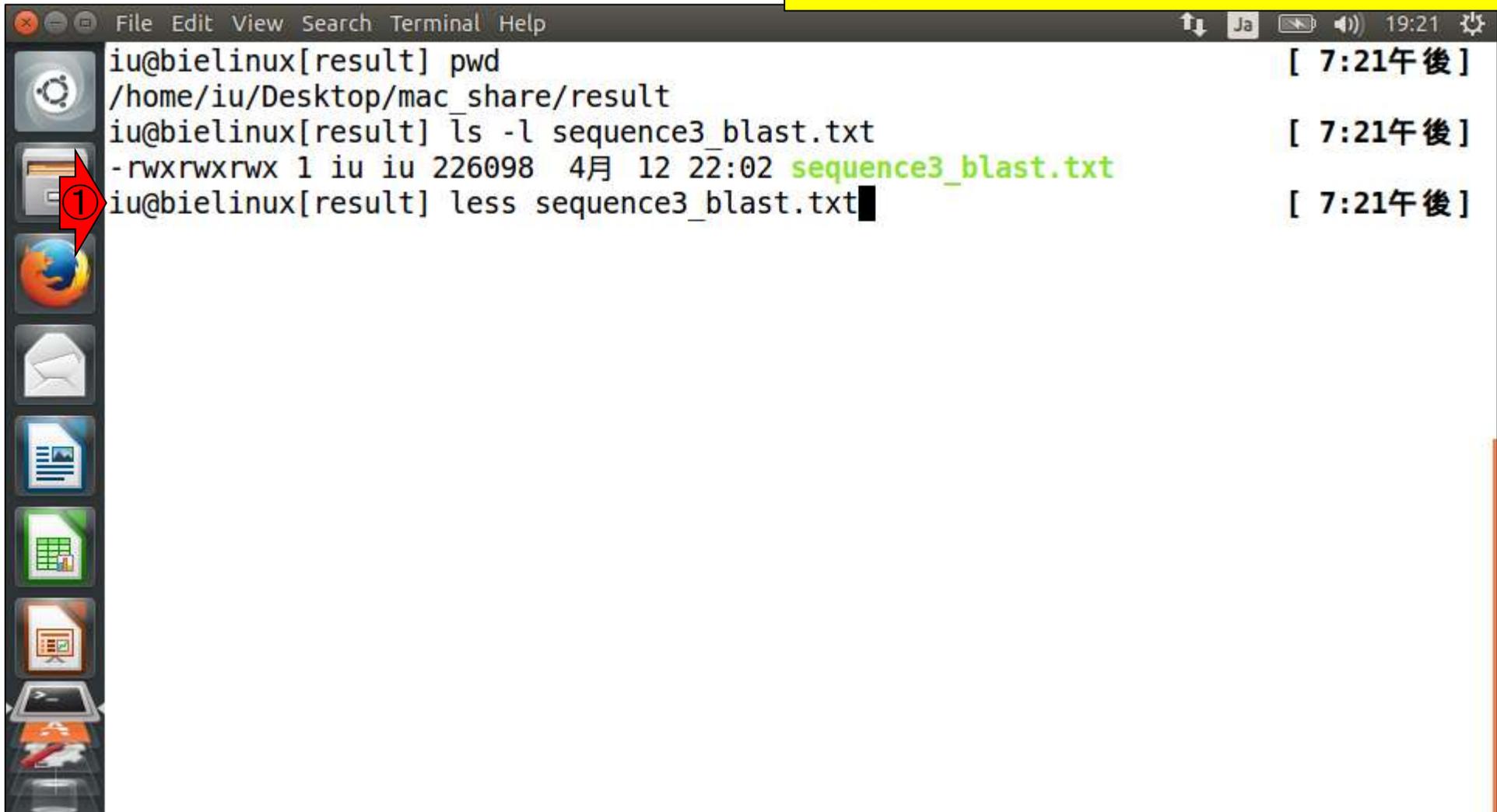
```
Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus
```

```
Score = 106 bits (57), Expect = 2e-23
Identities = 68/73 (93%), Gaps = 2/73 (3%)
Strand=Plus/Minus
```



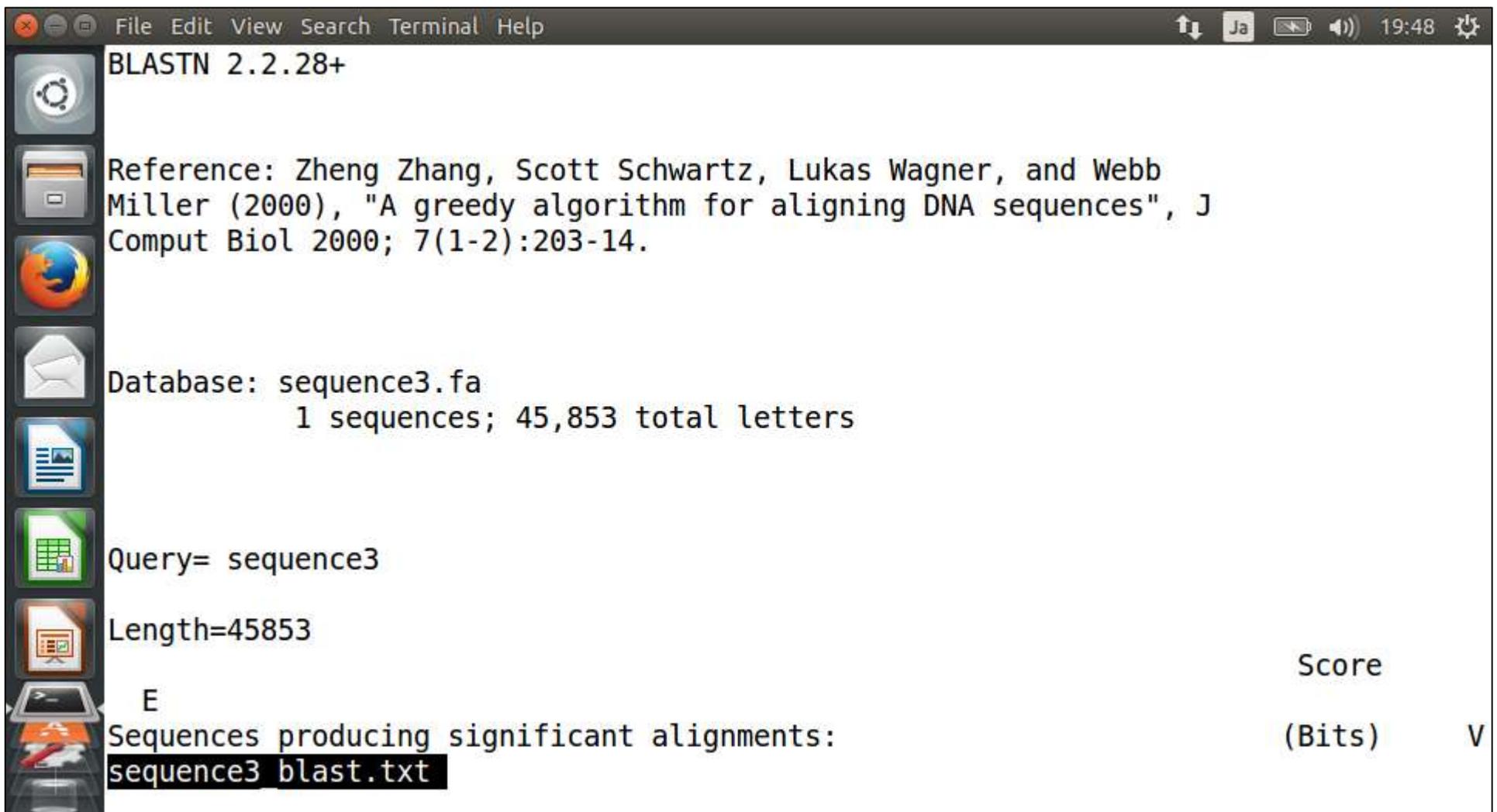
W15-7: less

lessコマンドでsequence3_blast.txtを開き、Score = で検索。画面の横幅を広めにとっておいたほうがよい。第3回のW14-6-2に文字列検索のやり方あり



```
iu@bielinux[result] pwd [ 7:21午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3_blast.txt [ 7:21午後 ]
-rwxrwxrwx 1 iu iu 226098 4月 12 22:02 sequence3_blast.txt
iu@bielinux[result] less sequence3_blast.txt [ 7:21午後 ]
```

W15-7: less



```
File Edit View Search Terminal Help
BLASTN 2.2.28+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb
Miller (2000), "A greedy algorithm for aligning DNA sequences", J
Comput Biol 2000; 7(1-2):203-14.

Database: sequence3.fa
          1 sequences; 45,853 total letters

Query= sequence3

Length=45853

                                     Score
E                                     (Bits)   V
Sequences producing significant alignments:
sequence3 blast.txt
```

W15-7: less

```

iu@bielinux[~/Desktop/mac_share/result]
BLASTN 2.2.28+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb
Miller (2000), "A greedy algorithm for aligning DNA sequences", J
Comput Biol 2000; 7(1-2):203-14.

Database: sequence3.fa
          1 sequences; 45,853 total letters

Query= sequence3
Length=45853

                                     Score
E                                     (Bits)      V
Sequences producing significant alignments:
/Score =

```



W15-7: less

①トップヒットのものが最初に見える。②全長の45,853 bp全てで完全一致なので、③queryの1-60番目の塩基とDB側(Sbjct; Subjectの意味)の1-60番目の塩基だけで眺めても完全一致となっていることがわかる

```
File Edit View Search Terminal Help
Score = 8.468e+04 bits (45853), Expect = 0.0
Identities = 45853/45853 (100%), Gaps = 0/45853 (0%)
Strand=Plus/Plus

Query 1 TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCAA 60
|||||
Sbjct 1 TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCAA 60

Query 61 TTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTAGCCA 120
|||||
Sbjct 61 TTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTAGCCA 120

Query 121 TTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTAT 180
|||||
Sbjct 121 TTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTAT 180

Query 181 GCTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCCGATCACAATAAGT 240
|||||
Sbjct 181 GCTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCCGATCACAATAAGT 240

:
```

W15-8: less

「n」と打って、2番目に一致するScore =が先頭行にくるページを表示した結果。①query配列の40,967番目の塩基がDB側配列の1番目の塩基と一致していることを意味する

```
iu@bielinux[~/Desktop/mac_share/result]
Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus

Query  40967  TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCCA  41026
      |||
Sbjct  1      TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTG-CCA  59

Query  41027  ATTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTAGCC  41086
      |||
Sbjct  60      ATTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTAGCC  119

Query  41087  ATTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTA  41146
      |||
Sbjct  120     ATTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTA  179

Query  41147  TGCCTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCCGATCACAATAA  41206
      ||
Sbjct  180     TG-CTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCCGATCACAATAA  238

:
```



W15-8: less

②や③のように、DB側(Sbjct)でところどころでGapが見られる。が、④全体で4,901 bpのアラインメントのうち31個だけGapがあった程度なので、実質的に無視でよい

```
iu@bielinux[~/Desktop/mac_share/result]
Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus

Query  40967  TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCCA  41026
      |||
Sbjct  1       TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTG-CCA  59

Query  41027  ATTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTAGCC  41086
      |||
Sbjct  60     ATTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTAGCC  119

Query  41087  ATTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTA  41146
      |||
Sbjct  120    ATTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTA  179

Query  41147  TGCCTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCGATCACAATAA  41206
      |||
Sbjct  180    TG-CTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCGATCACAATAA  238
```



W15-8: less

「n」と打って、3番目に一致するScore =が先頭行にくるページを表示した結果。①query配列の1番目の塩基がDB側配列の40,967番目の塩基と一致していることを意味する。2番目と3番目はQueryとSbjctが入れ替わっているだけ

```
iu@bielinux[~/Desktop/mac_share/result]
Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus
Query 1 TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTG-CCA 59
|||||
Sbjct 40967 TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCCA 41026
Query 60 ATTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTTAGCC 119
|||||
Sbjct 41027 ATTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTTAGCC 41086
Query 120 ATTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTA 179
|||||
Sbjct 41087 ATTACGGACACCTCCATCTTTTGATAGCGCTAACAAGTGCTACTTCAACAAATCCTTTTA 41146
Query 180 TG-CTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCGATCACAATAA 238
|||||
Sbjct 41147 TGCCTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCGATCACAATAA 41206
```



W15-9: less

上矢印キーを10回押し、10行分だけページ上部に移動した結果画面。セカンドヒットのラインメント結果の最後のほうを確認するのが目的

```
iu@bielinux[~/Desktop/mac_share/result]
Query 45753 AGAAAGAATTAACGAATTACGCAAAGAAGCCATTGATTACTCTACTAGAAAATCTTATGT 45812
|||||
Sbjct 4784 AGAAAGAATTAACGAATTACGCAAAGAAGCCATTGATTACTCTACTAGAAAATCTTATGT 4843

Query 45813 CACGACCAAATTATTTTTTATCGCCAACATGATTAAGCACA 45853
|||||
Sbjct 4844 CACGACCAAATTATTTTTTATCGCCAACATGATTAAGCACA 4884

Score = 8844 bits (4789), Expect = 0.0
Identities = 4869/4901 (99%), Gaps = 31/4901 (1%)
Strand=Plus/Plus

Query 1 TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTG-CCA 59
|||||
Sbjct 40967 TTTTAGCGGCGGTGTTTGAAGTCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCCA 41026
Query 60 ATTGCAATCAATAGTGACAATTTACCCCAAAAACCAGGGGTCTGTCGTTTAATTTTAGCC 119
|||||
```



W16-1:トリム候補領域

上矢印キーをさらに押し続け、(重複塩基数が4900 bp程度なのでその半分の)2400 - 2500番目付近を眺める。具体的には①の赤枠分くらいを眺め、どこにも mismatches や Gap がないことを確認

iu@bielinux[~/Desktop/mac_share/result]

```
Query 2390 GAATTATCAAGCTACGACTGGGGATTTCGATATAGTCCCTGGATTAGAACTGTTGATGAT 2419
|||||
Sbjct 43363 GAATTATCAAGCTACGACTGGGGATTTCGATATAGTCCCTGGATTAGAACTGTTGATGAT 43422

Query 2450 GAGCAGGGGTATTACTACTACATCATTCAAATGGGAACGGCACTTGGGAAAAAACAGAT 2509
|||||
Sbjct 43423 GAGCAGGGGTATTACTACTACATCATTCAAATGGGAACGGCACTTGGGAAAAAACAGAT 43482

Query 2510 CCGCGAATAGATCGTCAAATTTAACAGAATATCAAAAAGAAACCCCAATTGATCTAAGA 2569
|||||
Sbjct 43483 CCGCGAATAGATCGTCAAATTTAACAGAATATCAAAAAGAAACCCCAATTGATCTAAGA 43542

Query 2570 GAAGTAGTGCGCATTATCAAATATTGGAGAAAGGCTCATAACGCAGTATGTAAGCTTAAT 2629
|||||
Sbjct 43543 GAAGTAGTGCGCATTATCAAATATTGGAGAAAGGCTCATAACGCAGTATGTAAGCTTAAT 43602

Query 2630 TCTTATGCACTAGAGACCACGGTTCTTGACTTTATAGATACTAATCCAATATATTCCAAC 2689
|||||
Sbjct 43603 TCTTATGCACTAGAGACCACGGTTCTTGACTTTATAGATACTAATCCAATATATTCCAAC 43662
:
```



W16-1:トリム候補領域

①のところでトリムすることにする。左端にする理由は、上が2450番目、下が43423番目の塩基だとすぐにわかるから

```
iu@bielinux[~/Desktop/mac_share/result]
Query 2390 GAATTATCAAGCTACGACTGGGGATTTCGATATAGTCCCTGGATTTAGAACTGTTGATGAT 2449
      |||
Sbjct 43363 GAATTATCAAGCTACGACTGGGGATTTCGATATAGTCCCTGGATTTAGAACTGTTGATGAT 43422
      |||
Query 2450 GAGCAGGGGTATTACTACTACATCATTCAAATGGGAACGGCACTTGGGAAAAAACAGAT 2509
      |||
Sbjct 43423 GAGCAGGGGTATTACTACTACATCATTCAAATGGGAACGGCACTTGGGAAAAAACAGAT 43482
      |||
Query 2510 CCGCGAATAGATCGTCAAATTTAACAGAATATCAAAAAGAAACCCCAATTGATCTAAGA 2569
      |||
Sbjct 43483 CCGCGAATAGATCGTCAAATTTAACAGAATATCAAAAAGAAACCCCAATTGATCTAAGA 43542
      |||
Query 2570 GAAGTAGTGCGCATTATCAAATATTGGAGAAAGGCTCATAACGCAGTATGTAAGCTTAAT 2629
      |||
Sbjct 43543 GAAGTAGTGCGCATTATCAAATATTGGAGAAAGGCTCATAACGCAGTATGTAAGCTTAAT 43602
      |||
Query 2630 TCTTATGCACTAGAGACCACGGTTCTTGACTTTATAGATACTAATCCAATATATTCCAAC 2689
      |||
Sbjct 43603 TCTTATGCACTAGAGACCACGGTTCTTGACTTTATAGATACTAATCCAATATATTCCAAC 43662
      |||
:
```



W16-2:トリム後の配列

①2450番目の塩基をトリム後の1塩基目にする場合は、[2450, 43422 bp]を残せばよい。こうすることで、トリム後の塩基配列の最初のほうは①の赤枠のようになり、最後のほうは②のようになるはずである。③qで終了

```
iu@bielinux[~/Desktop/mac_share/result]
Query 2390 GAATTATCAAGCTACGACTGGGGATTTCGATATAGTCCCTGGATTAGAACTGTTGATGAT 2419
      |||
Sbjct 43363 GAATTATCAAGCTACGACTGGGGATTTCGATATAGTCCCTGGATTAGAACTGTTGATGAT 3422
      |||
Query 2450 ① GAGCAGGGGTATTACTACTACATCATTCAAATGGGAACGGCACTTGGGAAAAAACAGAT 2509
      |||
Sbjct 43423 GAGCAGGGGTATTACTACTACATCATTCAAATGGGAACGGCACTTGGGAAAAAACAGAT 43482
      |||
Query 2510 CCGCGAATAGATCGTCAAATTTAACAGAATATCAAAAAGAAACCCCAATTGATCTAAGA 2569
      |||
Sbjct 43483 CCGCGAATAGATCGTCAAATTTAACAGAATATCAAAAAGAAACCCCAATTGATCTAAGA 43542
      |||
Query 2570 GAAGTAGTGCGCATTATCAAATATTGGAGAAAGGCTCATAACGCAGTATGTAAGCTTAAT 2629
      |||
Sbjct 43543 GAAGTAGTGCGCATTATCAAATATTGGAGAAAGGCTCATAACGCAGTATGTAAGCTTAAT 43602
      |||
Query 2630 TCTTATGCACTAGAGACCACGGTTCTTGACTTTATAGATACTAATCCAATATATTCCAAC 2689
      |||
Sbjct 43603 TCTTATGCACTAGAGACCACGGTTCTTGACTTTATAGATACTAATCCAATATATTCCAAC 43662
      |||
:
```



W16-3:トリム実行

①まずはトリム後のFASTAファイル(ファイル名:sequence3_trimmed.fa)のdescription行を作成。W12-7とほぼ同じ

```
iu@bielinux[result] pwd [ 3:41午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3*.fa [ 3:41午後 ]
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
iu@bielinux[result] echo ">sequence3_trimmed" > sequence3_trimmed.fa [ 3:41午後 ]
iu@bielinux[result] more sequence3_trimmed.fa [ 3:41午後 ]
>sequence3_trimmed
iu@bielinux[result] █ [ 3:41午後 ]
```

W16-3:トリム実行

- ①トリム実行本番。W13-2とほぼ同じ。
- ② sequence3.faの最終行のみ取り出してパイプで流し、
- ③2450-43422文字目を抽出した結果を、
- ④sequence3_trimmed.faに追加書き込み

```
iu@bielinux[result] pwd [ 3:41午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3*.fa [ 3:41午後 ]
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
iu@bielinux[result] echo ">sequence3_trimmed" > sequence3_trimmed.fa [ 3:41午後 ]
iu@bielinux[result] more sequence3_trimmed.fa [ 3:41午後 ]
>sequence3_trimmed
iu@bielinux[result] tail -n 1 sequence3.fa | cut -c 2450-43422 >> sequence3_trimmed
ed.fa [ 3:41午後 ]
iu@bielinux[result] █
```



W16-4: moreで確認

①lsで確認。1 bp = 1 byte。ファイルサイズの的に妥当な印象を受ける。②moreでも確認

```
iu@bielinux[~/Desktop/mac_share/result]
iu@bielinux[result] pwd [ 3:41午後 ]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3*.fa [ 3:41午後 ]
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
iu@bielinux[result] echo ">sequence3_trimmed" > sequence3_trimmed.fa [ 3:41午後 ]
iu@bielinux[result] more sequence3_trimmed.fa [ 3:41午後 ]
>sequence3_trimmed
iu@bielinux[result] tail -n 1 sequence3.fa | cut -c 2450-43422 >> sequence3_trimm
ed.fa
① iu@bielinux[result] ls -l sequence3*.fa [ 3:41午後 ]
-rwxrwxrwx 1 iu iu 45865 4月 1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 40993 4月 16 15:41 sequence3_trimmed.fa
② iu@bielinux[result] more sequence3_trimmed.fa [ 3:42午後 ]
```


W16-4: moreで確認

スペースキーをガスガス押して最後まで表示し終わったところ。②最後の塩基配列の赤枠部分もW16-2と全く同じになっていることからうまくトリムできたと判断

```
File Edit View Search Terminal Help
GAGATGGCACACTGAACGTTGCAATGACAGCCCTAATCCAAGCTGAAGCACATATTCGTATCGGTGTTATTCCCATGGGAA
CGTTAATAATTTTGAACCCGCTACCAGTTACCAACTGACCCCAAGCGGCCATCGAACTAATTTGTACTIONAGCCGGCGA
CCCAAGCAGTCGGCATGCTAGTTTGAATCAACGCCGGGCAGTCGTGAGTTCACTGACATTCGGTAATTTGGCTGACATTT
CCAATGAAGTTCGACAATCTGAGAAGCAGCGATTTCGGTAAATTAAGCTATCTTTACCGTGCTATTCGCCATATTGGTCACA
ATAAGTCACTGCCAATTCGATATCAATCAAACACGAAGGAAGCCACACCTTAAAAACATGGTTCTGTTTGATTACAACGA
CCAAATCAGTCGGTGGGCACGTCTATAGCGGTCTGCTCCAGGAAAAATGCATATTAGTCTACTTAACAACATTGGCTGGC
GGCAAGTAATTCCATATATTTGGTTCGCACTAACGGGGAACCTTGCAAACTCCAAGGCCATTACACAGCTAACGGCAACCA
GTGCACGAATTACGAGTGCAACTGGTCAAGCCGTGACGACACGTATTGACGGCGACCCAGCGGTTAACTGCCAATTGAAT
TGACCTATTTGACAGACCGCTTCGAATTGATCGTACCAACAGTCATCGAATAACGACGTATTTTTTATTAATATATACATA
TATTAATTGCAAGAATTCATGTTAT
Query 2390 GAATTATCAAGCTACGACTGGGGATTCGATATAGTCCCTGGATTTAGAACTGTTGATGAT 2449
GACAAAAATCACAAATTAATTTTT
Sbjct 43363 GAATTATCAAGCTACGACTGGGGATTCGATATAGTCCCTGGATTTAGAACTGTTGATGAT 3422
GGATAATCATTGTACGCGGAAAGGC
Query 2450 GAGCAGGGGTATTACTACTACATCATTCCAATGGGAACGGCACTTGGGAAAAAACAGAT 2509
GAAAGAAATTCGATGTACAACGATG
Sbjct 43423 GAGCAGGGGTATTACTACTACATCATTCCAATGGGAACGGCACTTGGGAAAAAACAGAT 43482
AAAGCACGGGTTGAAAACGAGGATA
AAAACATAAATATAGATAGTAAAAA
GTCAAAAAACGAGGAAAACCTTGAATTATACTCGGACTCAGAATTTGCACTAAAAATGGGATCATTGCTCGAAAGACAC
AGATTAGACCTCTTGACGATGTTGACCAAATGATTATCTTTTCGGCAAAGGGGAGCACCGCTAATTTAGATACGTCTCAAT
GGAATCAGGTGTTTGTAAATGTTCCAGATAGCGCTCCAGAATTAAGAAAAATGGATGGAGAAAAATGGGCTTAGTTCTATAA
AAGTCTTGAATTATCTTAAACAGCTATTGAATGGAATATCGCAATATCAATCGGCAGATATTAAGGATTCAGCAAGCAC
TTAGACTGGAATTATCAAGCTACGACTGGGGATTCGATATAGTCCCTGGATTTAGAACTGTTGATGAT
iu@bielinux[result] 3:48午後]
```

W17-1: FASTQのトリム

①FASTQファイル(sequence3.fq)の場合は、2行目(塩基配列情報の行)と4行目(クオリティ情報の行) についてののみW16-3と同様な操作を行えばよい。②得られるファイルはsequence3_trimmed.fq

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3*.fq
-rwxrwxrwx 1 iu iu 91727  4月  1 12:21 sequence3.fq
iu@bielinux[result] wc sequence3.fq
  4      4 91727 sequence3.fq
iu@bielinux[result] head -n 1 sequence3.fq | tail -n 1 > sequence3_trimmed.fq
iu@bielinux[result] head -n 2 sequence3.fq | tail -n 1 | cut -c 2450-43422 >> sequence3_trimmed.fq
iu@bielinux[result] head -n 3 sequence3.fq | tail -n 1 >> sequence3_trimmed.fq
iu@bielinux[result] head -n 4 sequence3.fq | tail -n 1 | cut -c 2450-43422 >> sequence3_trimmed.fq
iu@bielinux[result] ls -l sequence3*.fq
-rwxrwxrwx 1 iu iu 91727  4月  1 12:21 sequence3.fq
-rwxrwxrwx 1 iu iu 81967  4月 16 2016 sequence3_trimmed.fq
iu@bielinux[result] █
```

[11:51午後]

[11:51午後]

[11:51午後]

[11:51午後]



①

W17-2:クオリティ分布

FASTQファイル(sequence3_trimmed.fq)を入力として、図1aおよびW11-9と同じようなクオリティスコア分布を作成。

W17-2:クオリティスコア分布

トリム後のFASTQ形式ファイル([sequence3_trimmed.fq](#))を入力として、図1aおよびW11-9と同じようなクオリティスコア分布を作成。出力ファイルは、[sequence3_trimmed.png](#)と[sequence3_trimmed.txt](#)。

```
cd ~/Desktop/mac_share/result

R -q
in_f <- "sequence3_trimmed.fq"           #入力ファイル名を指定してin_fに格納
out_f1 <- "sequence3_trimmed.png"       #出力ファイル名を指定してout_f1に格納
out_f2 <- "sequence3_trimmed.txt"       #出力ファイル名を指定してout_f2に格納
param_fig <- c(700, 350)                 #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
#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
fastq <- readFastq(in_f)                 #in_fで指定したファイルの読み込み
#本番(PHREDスコアに変換)
out <- as(quality(fastq), "matrix")       #ASCIIコードのquality scoreをPHRED scoreに変換し、データ相
colnames(out) <- 1:ncol(out)              #列名を付与
rownames(out) <- as.character(id(fastq)) #行名を付与
#ファイルに保存(pngファイル)
png(out_f1, pointsize=13, width=param_fig[1], height=param_fig[2]) #出力ファイルの各種パラメータを
par(mar=c(4, 4, 0, 0))                  #下、左、上、右の順で余白(行)を指定
plot(x=1:ncol(out), y=out, pch=20, cex=0.5, #プロット
      type="p", xlab="position", ylab="PHRED score") #プロット
dev.off()                                 #おまじない
#ファイルに保存(テキストファイル)
```

W17-2:クオリティ分布

```
File Edit View Search Terminal Help
> #ファイルに保存(pngファイル)
> png(out_f1, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイル
の各種パラメータを指定
> par(mar=c(4, 4, 0, 0)) #下、左、上、右の順で余白(行)を指定
> plot(x=1:ncol(out), y=out, pch=20, cex=0.5,#プロット
+ type="p", xlab="position", ylab="PHRED score")#プロット
> dev.off() #おまじない
null device
      1
> #ファイルに保存(テキストファイル)
> tmp <- cbind(colnames(out), as.vector(out))#保存したい情報をtmpに格納
> write.table(tmp, out_f2, sep="\t", append=F, quote=F, row.names=F, col.names=F)
#tmpの中身を指定したファイル名で保存
> q(save="no")
iu@bielinux[result] pwd [12:05午前]
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3_trimmed* [12:05午前]
-rwxrwxrwx 1 iu iu 40993 4月 16 15:41 sequence3_trimmed.fa
-rwxrwxrwx 1 iu iu 81967 4月 16 23:51 sequence3_trimmed.fq
-rwxrwxrwx 1 iu iu 19113 4月 17 2016 sequence3_trimmed.png
-rwxrwxrwx 1 iu iu 357641 4月 17 2016 sequence3_trimmed.txt
iu@bielinux[result] █ [12:05午前]
```



W17-2:クオリティ分布

①pngファイルを眺めているところ。W11-9で見られていた両側の低クオリティ領域がうまくトリムされていることがわかる。図3aと同じ

W17-2:クオリティスコア分布

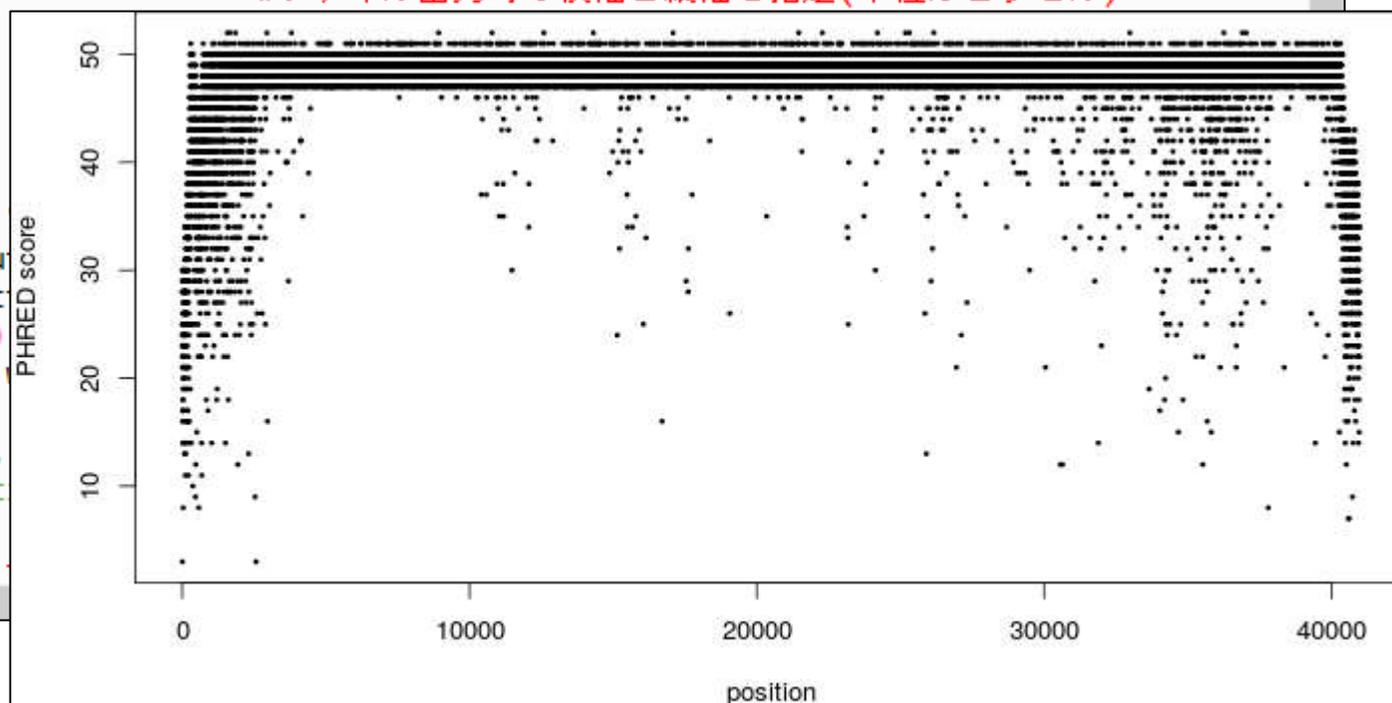
トリム後のFASTQ形式ファイル([sequence3_trimmed.fq](#))を入力として、図1aおよびW11-9と同じようなクオリティスコア分布を作成。出力ファイルは、[sequence3_trimmed.png](#)と[sequence3_trimmed.txt](#)。

```
cd ~/Desktop/mac_share/result
```

```
R -q
in_f <- "sequence3_trimmed.fq"
out_f1 <- "sequence3_trimmed.png"
out_f2 <- "sequence3_trimmed.txt"
param_fig <- c(700, 350)
#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
fastq <- readFastq(in_f)
#本番(PHREDスコアに変換)
out <- as(quality(fastq),
colnames(out) <- 1:ncol(out)
rownames(out) <- as.character(1:ncol(out))
#ファイルに保存(pngファイル)
png(out_f1, pointsize=13,
par(mar=c(4, 4, 0, 0))
plot(x=1:ncol(out), y=out,
type="p", xlab="position",
dev.off()
#ファイルに保存(テキストファイル)
```



#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_f1に格納
#出力ファイル名を指定してout_f2に格納
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)



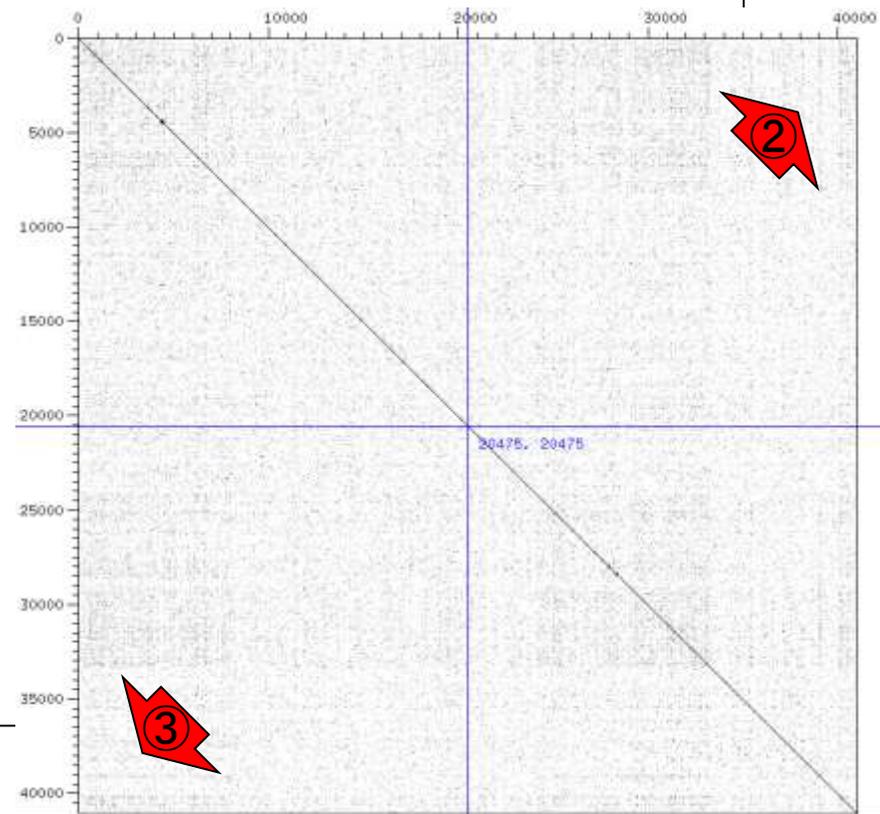
W17-3: ドットプロット

① sequence3.fa 同士のドットプロットを dotter で実行。トリム前 (W14-4) と違って、配列末端部分の一致領域 (② や ③ 付近の対角線のプロット) がなくなっていることがわかる。うまく重複領域をトリムできていることを意味する

```
iu@bielinux[result] pwd
/home/iu/Desktop/mac_share/result
iu@bielinux[result] ls -l sequence3*.fa
-rwxrwxrwx 1 iu iu 45865  4月  1 11:56 sequence3.fa
-rwxrwxrwx 1 iu iu 40993  4月 16 15:41 sequence3_trimmed.fa
iu@bielinux[result] dotter sequence3_trimmed.fa sequence3_trimmed.fa
```

[3:43午後]

[3:43午後]



W18-1: NCBI BLAST

W15で示したBLASTは、各種ウェブサービスでも実行可能。ここでは、[sequence3.fa](#)を入力としてNCBIで行う。①nucleotide blast

U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

BLAST finds regions of similarity between biological sequences. [more...](#)

New Try [SmartBLAST](#) for an improved protein-protein search

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested **GO**

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Cow](#)
- [Pig](#)
- [Dog](#)
- [Rabbit](#)
- [Chimp](#)
- [Guinea pig](#)
- [Fruit fly](#)
- [Honey bee](#)
- [Chicken](#)
- [Zebrafish](#)
- [Clawed frog](#)
- [Arabidopsis](#)
- [Rice](#)
- [Yeast](#)
- [Microbes](#)

Basic BLAST

Choose a BLAST program to run.

- ①** [nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast
- [protein blast](#) Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
- [blastx](#) Search **protein** database using a **translated nucleotide** query
- [tblastn](#) Search **translated nucleotide** database using a **protein** query
- [tblastx](#) Search **translated nucleotide** database using a **translated nucleotide** query

Your Recent Results [New!](#)

[All Recent results...](#)

News

[Searching Whole Genome Shotgun sequences](#)

It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer.

Wed, 20 Jan 2016 10:00:00 EST

[More BLAST news...](#)

Tip of the Day

[Use Genomic BLAST to see the genomic context](#)

If you are interested in the evolution of a particular gene or gene family it is often

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

W18-1: NCBI BLAST

同じ配列同士を比較したい場合は①
Align two or more sequencesにチェック

The screenshot shows the NCBI BLAST web interface. At the top, there is a navigation bar with the NIH logo, 'U.S. National Library of Medicine', 'NCBI', and 'Sign in to NCBI'. Below this is the 'BLAST' header with a 'blastn suite' link and navigation options: 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main section is titled 'Standard Nucleotide BLAST' and includes tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. A sub-header states 'BLASTN programs search nucleotide databases using a nucleotide query.' with links for 'Reset page' and 'Bookmark'. The 'Enter Query Sequence' section contains a text input field for 'Enter accession number(s), gi(s), or FASTA sequence(s)', a 'Clear' button, and a 'Query subrange' section with 'From' and 'To' input fields. Below this is an 'Or, upload file' section with a file upload button and a 'Job Title' input field. A red arrow with the number 1 points to the 'Align two or more sequences' checkbox. The 'Choose Search Set' section includes a 'Database' dropdown set to 'Nucleotide collection (nr/nt)', an 'Organism' input field, and several 'Exclude' and 'Limit to' checkboxes. The 'Program Selection' section at the bottom has 'Optimize for' radio buttons for 'Highly similar sequences (megablast)' and 'More dissimilar sequences (discontiguous megablast)'.

W18-1: NCBI BLAST

チェック後の状態。ここではホストOS (Windows)上のウェブブラウザで作業を行っており、入力ファイルはこの作業環境では「C:\Users\kadota\Desktop\share\result\sequence3.fa」にある。ウェブ資料通りだと、黒字部分はおそらくみんな同じで、灰色部分はヒトによって異なる

BLAST® >> blastn suite

Align Sequences Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [⌵](#)

From

To

Or, upload file [参照...](#) [⌵](#)

Job Title

Enter a descriptive title for your BLAST search [⌵](#)

Align two or more sequences [⌵](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Subject subrange [⌵](#)

From

To

Or, upload file [参照...](#) [⌵](#)

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm [⌵](#)

W18-2: ファイル指定

①query側、②DB側 (subject側) ともに「C:¥Users¥kadota¥Desktop¥share¥result¥sequence3.fa」を指定。灰色部分はヒトによって異なる

U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST® >> blastn suite Home Recent Results Saved Strategies Help

Align Sequences Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [From](#) [To](#)

Or, upload file [参照...](#)

Job Title [参照...](#)
Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Subject subrange [From](#) [To](#)

Or, upload file [参照...](#)

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

W18-2: ファイル指定

The screenshot shows the NCBI BLAST web interface. The browser address bar displays `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Meq...`. The page title is "Align Sequences Nucleotide BLAST". The "Enter Query Sequence" section includes a text input field for "Enter accession number(s), gi(s), or FASTA sequence(s)", a "Clear" button, and a "Query subrange" section with "From" and "To" input fields. Below this is the "Or, upload file" section with a text input field containing the file path `C:\Users\kadota\Desktop` and a "参照..." button. A "Job Title" input field is also present. The "Align two or more sequences" checkbox is checked. The "Enter Subject Sequence" section has a similar structure to the query section. The "Program Selection" section has radio buttons for "Highly similar sequences (megablast)", "More dissimilar sequences (discontiguous megablast)", and "Somewhat similar sequences (blastn)", with a "Choose a BLAST algorithm" link below. A red arrow with the number "1" points to the bottom right corner of the page.

W18-3: blastnを実行

①デフォルトはmegablastのようだが、ここでは当初やろうと思っていた②blastnを選択して、③BLASTを実行

The screenshot shows the NCBI Nucleotide BLAST web interface. The browser address bar displays "http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Meq...". The main form is divided into several sections:

- Query subrange:** Includes a text input for "Enter accession number(s), gi(s), or FASTA sequence(s)", a "Clear" link, and "From" and "To" range input fields.
- Upload and Job Title:** Features an "Or, upload file" button with a file path "C:\Users\kadota\Desktop" and a "参照..." button, and a "Job Title" input field with the prompt "Enter a descriptive title for your BLAST search".
- Align two or more sequences:** A checked checkbox.
- Enter Subject Sequence:** Mirrors the query section with its own "Enter accession number(s)..." input, "Clear" link, and "From" and "To" range fields.
- Program Selection:** Contains the "Optimize for" section with three radio button options:
 - Highly similar sequences (megablast)
 - More dissimilar sequences (discontiguous megablast)
 - Somewhat similar sequences (blastn)** (selected)A "Choose a BLAST algorithm" link is also present.
- BLAST Button:** A blue button labeled "BLAST" with the text "Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)". Below it is a checkbox for "Show results in a new window".
- Algorithm parameters:** A collapsed section indicated by a "+" icon.

Three red arrows with circled numbers point to specific elements: Arrow 1 points to the "Somewhat similar sequences (blastn)" radio button; Arrow 2 points to the "BLAST" button; Arrow 3 points to the "BLAST" button label.

At the bottom of the page, there is a footer with the text "BLAST is a registered trademark of the National Library of Medicine." and a row of links: "Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback" and "NCBI | NLM | NIH | DHHS".

W18-4: 実行結果

①query側と②DB側 (subject側) の配列情報。このプログラムを利用したときは③ Stephen et al. 1997の論文を引用すべし!

The screenshot shows the NCBI BLAST results page for a search of 'sequence3 (45853 letters)'. The page is titled 'BLAST Results' and includes navigation links like 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main content area displays 'Blast 2 sequences' with the following details:

Query ID	Subject ID
sequence3	sequence3
nucleic acid	nucleic acid
45853	45853

Below this, a 'Graphic Summary' section shows the 'Distribution of 49 Blast Hits on the Query Sequence'. A color key for alignment scores is provided:

Score Range	Color
<40	Black
40-50	Blue
50-80	Green
80-200	Magenta
>=200	Red

The query sequence is shown as a horizontal bar with a red segment at the end, indicating high alignment scores. The x-axis is labeled 'Query' and has markers at 1, 9000, 18000, 27000, 36000, and 45000.

③Stephen et al., *Nucleic Acids Res.*, **25**: 3389-3402, 1997

W18-4: 実行結果

様々な角度で結果を眺められる。①
Graphic Summaryは、デフォルトでは②
見えている。②を非表示にしたい場合は
①を押す。すると...

U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST® » blastn suite-2sequences » RID-HBH0YXTF114 Home Recent Results Saved Strategies Help

BLAST Results

Edit and Resubmit Save Search Strategies Formatting options Download YouTube How to read this page Blast re

Blast 2 sequences

sequence3 (45853 letters)

RID	HBH0YXTF114 (Expires on 04-20 13:37 pm)	Subject ID	Id Query_189369
Query ID	Id Query_189367	Description	sequence3
Description	sequence3	Molecule type	nucleic acid
Molecule type	nucleic acid	Subject Length	45853
Query Length	45853	Program	BLASTN 2.3.1+ Citation

Other reports: Search Summary

Graphic Summary ①

Distribution of 49 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Color key for alignment scores

<40	40-50	50-80	80-200	>=200
-----	-------	-------	--------	-------

Query

1 9000 18000 27000 36000 45000

②

+ Dot Matrix View

- Descriptions

W18-4: 実行結果

こんな感じになる。確かにデフォルトの画面では見えていたGraphic Summaryが消えた。同時に②のところは+に変わったこともわかる。①のところをクリックするたびに-と+が変わり、連動して表示と非表示も変わる。このノリで他の項目を眺める

BLAST® » blastn suite-2sequences » RID-HBH0YXTF114

BLAST Results

sequence3 (45853 letters)

RID [HBH0YXTF114](#) (Expires on 04-20 13:37 pm)

Query ID [Id|Query_189367](#)
Description [sequence3](#)
Molecule type [nucleic acid](#)
Query Length [45853](#)

Subject ID [Id|Query_189369](#)
Description [sequence3](#)
Molecule type [nucleic acid](#)
Subject Length [45853](#)
Program [BLASTN 2.3.1+](#) [Citation](#)

Reports: [Search Summary](#)

[Graphic Summary](#) **①**

[Dot Matrix View](#)

[Descriptions](#)

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [Graphics](#)

	Description	Max score	Total score	Query cover	E value	Ident	Access
<input type="checkbox"/>	sequence3	82691	1.015e+05	100%	0.0	100%	Query_189

[Alignments](#)

W18-5:ドットプロット

①Dot Matrix Viewはドットプロットのこと。デフォルトでは非表示になっているので、1回クリックして表示させる

The screenshot shows the NCBI BLAST results page for a BLASTN search. The query sequence is 'sequence3' (45853 letters) and the subject is 'RID-HBH0YXTF114'. The results section is currently collapsed, but the 'Dot Matrix View' link is highlighted with a red arrow and a circled '1', indicating it should be clicked to view the dot plot.

BLAST Results

sequence3 (45853 letters)

RID [HBH0YXTF114](#) (Expires on 04-20 13:37 pm)

Query ID Id|Query_189367
Description sequence3
Molecule type nucleic acid
Query Length 45853

Subject ID Id|Query_189369
Description sequence3
Molecule type nucleic acid
Subject Length 45853
Program BLASTN 2.3.1+ [Citation](#)

Other reports: [Search Summary](#)

[+ Graphic Summary](#)
[+ Dot Matrix View](#) ①
[- Descriptions](#)

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[↑↓ Alignments](#) [Download](#) [Graphics](#)

	Description	Max score	Total score	Query cover	E value	Ident	Access
<input type="checkbox"/>	sequence3	82691	1.015e+05	100%	0.0	100%	Query_189

[- Alignments](#)

W18-5:ドットプロット

表示させたところ。②半ページ分ほど下部に移動して全体を見られるようにする

U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST® » blastn suite-2sequences » RID-HBH0YXTF114 Home Recent Results Saved Strategies Help

BLAST Results

Edit and Resubmit Save Search Strategies Formatting options Download YouTube How to read this page Blast re

Blast 2 sequences

sequence3 (45853 letters)

RID [HBH0YXTF114](#) (Expires on 04-20 13:37 pm)

Query ID	Id Query_189367	Subject ID	Id Query_189369
Description	sequence3	Description	sequence3
Molecule type	nucleic acid	Molecule type	nucleic acid
Query Length	45853	Subject Length	45853
		Program	BLASTN 2.3.1+ Citation

Other reports: Search Summary

Graphic Summary

Dot Matrix View ①

Plot of Id|Query_189367 vs Id|Query_189369 ②

W18-5: ドットプロット

こんな感じ。②左上が原点のdotter (W14-4)と違って、③NCBI BLASTのドットプロットは左下が原点になっているが、細かいことは気にしない

Other reports: [Search Summary](#)

[Graphic Summary](#)

[Dot Matrix View](#) ①

Program BLASTN 2.3.1+ [Citation](#)

Plot of lc|Query_189367 vs lc|Query_189369 ③

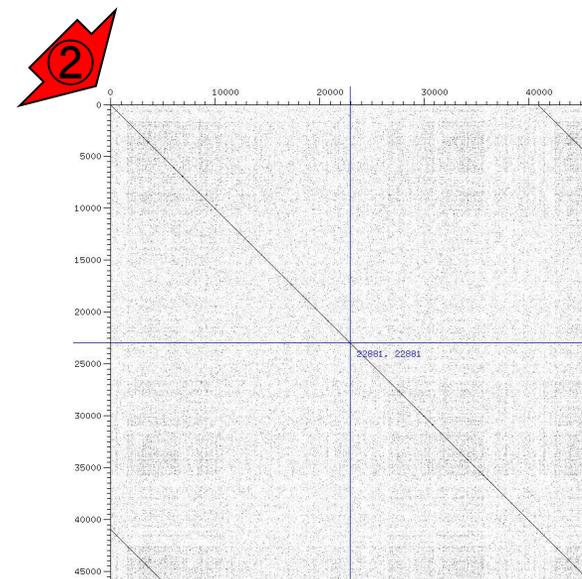
[Descriptions](#)

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

[Alignments](#) [Download](#) [Graphics](#)

Description	Max score	Total score	Query cover	E value	Ident	Access
-------------	-----------	-------------	-------------	---------	-------	--------



W18-6: アライメント

①Dot Matrix ViewとDescriptionsを非表示にし、②Alignmentsを眺める

Other reports: [Search Summary](#)

+ Graphic Summary

+ Dot Matrix View  ①

+ Descriptions

- Alignments ②

Download **Graphics** Sort by: E value ▼ Next Previous Descrip

sequence3
Sequence ID: lc|Query_189369 Length: 45853 Number of Matches: 49

Range 1: 1 to 45853 [Graphics](#) ▼ Next Match ▲ Previous Match **Related Information**

Score	Expect	Identities	Gaps	Strand
82691 bits(91706)	0.0	45853/45853(100%)	0/45853(0%)	Plus/Plus
Query 1	TTTTAGCGGCGGTGTTTGA	ACTGCCGCACTTCTCGAA	ACACAGTCAATCCTAAT	TGCCAA 60
Sbjct 1	TTTTAGCGGCGGTGTTTGA	ACTGCCGCACTTCTCGAA	ACACAGTCAATCCTAAT	TGCCAA 60
Query 61	TTGCAATCAATAGTGACA	ATTTACCCCAAAAACCAG	GGGCTGTGCGTTTAAT	TTTAGCCA 120
Sbjct 61	TTGCAATCAATAGTGACA	ATTTACCCCAAAAACCAG	GGGCTGTGCGTTTAAT	TTTAGCCA 120
Query 121	TTACGGACACCTCCATCT	TTTTGATAGCGCTAAC	AAGTGCTACTTCAACA	AATCCTTTTAT 180
Sbjct 121	TTACGGACACCTCCATCT	TTTTGATAGCGCTAAC	AAGTGCTACTTCAACA	AATCCTTTTAT 180
Query 181	GCTAATCACAATTACTGC	GGCTGAAGCGCCTGGG	CAGCAACGGTTCCGAT	CACAATAAGT 240
Sbjct 181	GCTAATCACAATTACTGC	GGCTGAAGCGCCTGGG	CAGCAACGGTTCCGAT	CACAATAAGT 240
Query 241	ATCATCGCCCCGCTCAC	GACTGGCAACAACGTC	ATTCCAGCAACTGTTG	GCCCGAACCAT 300
Sbjct 241	ATCATCGCCCCGCTCAC	GACTGGCAACAACGTC	ATTCCAGCAACTGTTG	GCCCGAACCAT 300
Query 301	CCAAAAGACAGCTGGCC	ACCAAGTACACCGAAT	AAGCCTGCGATTAAAT	TAATTGTCGTT 360

W18-6: アライメント

①ヒット数(ここではMatch数)は49。②トップヒット(ここではRange 1)はどう転んでもsequence3の全長配列間で100%一致のアライメントなので、③「Identities = 45853/45853 (100%)」。このあたりはBio-Linux上で実行したblastnの結果と同じ(W15-3)

Other reports: [Search Summary](#)

+ Graphic Summary

+ Dot Matrix View

+ Descriptions

- Alignments

Download Graphics Sort by: E value

sequence3
Sequence ID: lc|Query_189369 Length: 45853 Number of Matches: 49

Score	Expect	Identities	Gaps	Strand
82691 bits(91706)	0.0	45853/45853(100%)	0/45853(0%)	Plus/Plus

Range 1: 1 to 45853

Query	Score	Query	Score
1	60	61	120
121	180	181	240
241	300	301	360

W18-6: アライメント

①スコアはBio-Linux上で実行したblastnの結果と異なる(W15-3)が、計算方法はバージョンや提供サイトによって若干異なるのだろうと解釈して思考停止。重要なのは、セカンドヒットのアライメント結果がBio-Linuxのものと同じかどうかの確認

Other reports: [Search Summary](#)

[+ Graphic Summary](#)

[+ Dot Matrix View](#)

[+ Descriptions](#)

[- Alignments](#)

Download [Graphics](#) Sort by: [E value](#) [Next](#) [Previous](#) [Descrip](#)

sequence3
Sequence ID: lc|Query_189369 Length: 45853 Number of Matches: 49

Score	Expect	Identities	Gaps	Strand
82691 bits(91706)	0.0	45853/45853(100%)	0/45853(0%)	Plus/Plus

Range 1: 1 to 45853 [Graphics](#) [Next Match](#) [Previous Match](#)

Query	Subject	Score
TTTTAGCGGCGGTGTTTGAAGTCCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCAA	TTTTAGCGGCGGTGTTTGAAGTCCCGCACTTCTCGAAACACAGTCAATCCTAATTGCCAA	60
TTGCAATCAATAGTGAACAATTTACCCCAAAAACCAGGGGTCTGTGCGTTTAATTTAGCCA	TTGCAATCAATAGTGAACAATTTACCCCAAAAACCAGGGGTCTGTGCGTTTAATTTAGCCA	120
TTACGGACACCTCCATCTTTTGATAGCGCTAACAAAGTGCTACTTCAACAAATCCTTTTAT	TTACGGACACCTCCATCTTTTGATAGCGCTAACAAAGTGCTACTTCAACAAATCCTTTTAT	180
GCTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCCGATCACAATAAGT	GCTAATCACAATTACTGCGGCTGAAGCGCCTGGGCAGCAACGGTTCCGATCACAATAAGT	240
ATCATCGCCCCGCTCAGACTGGCAACAACGTCATTCCAGCAACTGTTGGCCCGAACCAT	ATCATCGCCCCGCTCAGACTGGCAACAACGTCATTCCAGCAACTGTTGGCCCGAACCAT	300
CCAAAAGACAGCTGGCCACCAAGTACACCGAATAAGCCTGCGATTAAATTAATTGTCGTT	CCAAAAGACAGCTGGCCACCAAGTACACCGAATAAGCCTGCGATTAAATTAATTGTCGTT	360

W18-7: セカンドヒット

① Next Matchを押すと次のヒットのアラインメント結果が見られる...はずだがそこに飛ばないので
② 「Range 1」の次の「Range 2」でページ内検索。おそらくInternet Explorerではうまく動かないが、firefoxなど他のブラウザではうまく動くのだろう

Other reports: [Search Summary](#)

+ Graphic Summary

+ Dot Matrix View

+ Descriptions

- Alignments

Download Graphics Sort by: E value Next Previous Descrip

sequence3
Sequence ID: lc|Query_189369 Length: 45853 Number of Matches: 49

Range 1: 1 to 45853 Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
82691 bits(91706)	0.0	45853/45853(100%)	0/45853(0%)	Plus/Plus

Query	Score	Subject	Score
1	60	1	60
61	120	61	120
121	180	121	180
181	240	181	240
241	300	241	300
301	360	301	360

W18-6: アラインメント

①「Range 2」でページ内検索。②「11件の一致」とあるが、これは総ヒット数(総Match数)が49あるため、Range 2以外にRange 20-29までの10個分があるため。③Score以外の結果は、Bio-Linux BLAST結果のセカンドヒットと全く同じ(W15-8)

Search results for Range 2 (1 to 4884):

Score	Expect	Identities	Gaps	Strand
8601 bits(9538)	0.0	4869/4901(99%)	31/4901(0%)	Plus/Plus

Alignment details for Range 2 (1 to 4884):

Query	Score	Expect	Identities	Gaps	Strand
40967	8601 bits(9538)	0.0	4869/4901(99%)	31/4901(0%)	Plus/Plus