

次世代シーケンサーデータの解析手法
第9回ゲノムアノテーションとその可視化、
DDBJへの登録

ウェブ資料

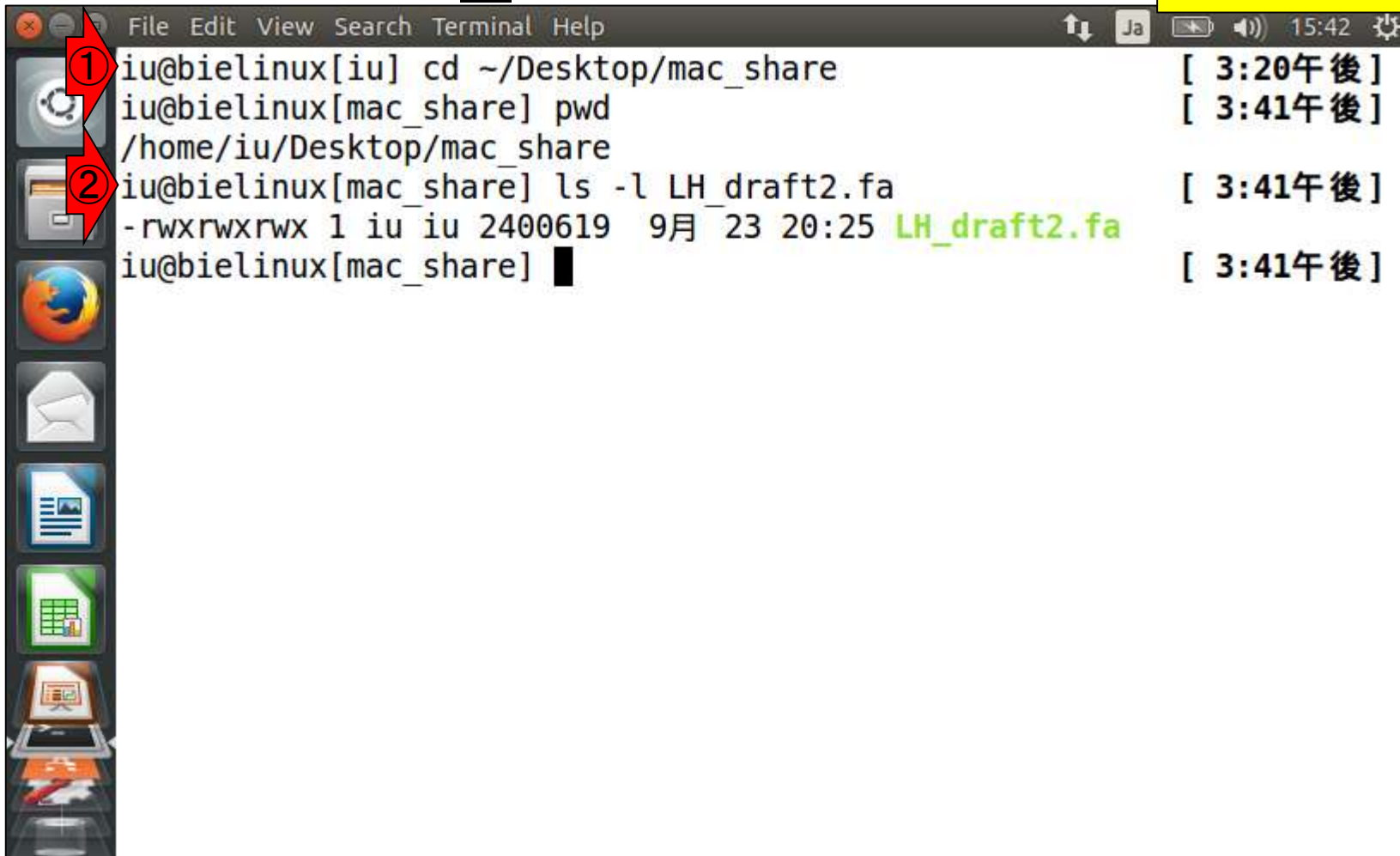
谷澤 靖洋、真島 淳、藤澤 貴智、李 慶範、
中村 保一*、清水 謙多郎、門田 幸二*

*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 : LH_draft2.fa

①共有フォルダ上に、②第9回で
利用するファイル(LH_draft2.fa)が
あるという前提で進めていきます



A terminal window with a dark background and a sidebar of application icons on the left. The terminal text is as follows:

```
iu@bielinux[iu] cd ~/Desktop/mac_share [ 3:20午後]
iu@bielinux[mac_share] pwd [ 3:41午後]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_draft2.fa [ 3:41午後]
-rwxrwxrwx 1 iu iu 2400619 9月 23 20:25 LH_draft2.fa
iu@bielinux[mac_share] █ [ 3:41午後]
```

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

W1-2:LH_draft2.fa

LH_draft2.faのACGTの存在比率を表示させた結果。GC含量が38.2%のLH_draft2.faの場合は、C or Gの存在比率の理論値が $38.2 / 2 = 19.1\%$ であり、実測値(Cが19.0%、Gが19.2%)であることから、Chargaff's second parity ruleが妥当であることがわかる

```
File Edit View Search Terminal Help
rownames, sapply, setdiff, sort, table, ta
unlist, unsplit

Loading required package: S4Vectors
Loading required package: stats4
Creating a generic function for 'nchar' from package 'base' in packa
ge 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
> #入力ファイルの読み込み
> fasta <- readDNASTringSet(in_f, format="fasta")#in_fで指定したファ
イルの読み込み
> #本番
> hoge <- alphabetFrequency(fasta)[,1:4] #A,C,G,Tの数を各配列ごと
にカウントした結果をhogeに格納
> colSums(hoge)/sum(as.numeric(hoge)) #A,C,G,Tの組成を表示
      A      C      G      T
0.3098775 0.1900621 0.1921449 0.3079155
>
> q(save="no")
iu@bielinux[mac_share] [ 1:51午後]
```

W2-1 : features.tsv

第8回の①W4-5の、②features.tsvをExcelで開いてdnaAの存在を確認。③annotation.gffを開いて確認でもよい

ゲノムアノテーション

- [MiGAP: Sugawara et al., 20th Int. Conf. Genome Informatics \(Kanagawa, Japan\), 2009](#)
- [RAST: Aziz et al., BMC Genomics, 2008, Overbee et al., Nucleic Acids Res., 2014, Bret](#)
- [DFAST: Tanizawa et al., Biosci Microbiota Food Health, 2016](#)
- [乳酸菌NGS連載第5回のPDF](#)
- [Prokka: Seemann T, Bioinformatics, 2014](#)

- W4-1: [DFAST](#)
- W4-2: [LH_hgap.fa](#)

ゲスト OS 上の LH_hgap.fa を、共有フォルダ 経由で ホスト OS にコピーしているだけです。

```
pwd
ls
cp LH_hgap.fa ../mac_share
ls -l ../mac_share
```



① W4-5: DFAST 実行結果 ([LH_hgap.fa](#))

- Genbank Flat File: [annotation.gbk](#)
- GFF3-formated File: [annotation.gff](#)
- Genome Fasta File: [genome.fna](#)
- Protein Fasta File: [protein.faa](#)
- CDS Fasta File: [cds.fna](#)
- RNA Fasta File: [rna.fna](#)
- Feature Table: [features.tsv](#)
- Genome Statistics: [statistics.txt](#)



BLASTの実行と可視化

- [dotter: Sonnhammer and Durbin, Gene, 1995](#)

W2-1 : features.tsv

①dnaAは、sequence1の[1441846, 1443162 bp]領域にあった(ことを一応確認)。ただこれは4配列からなるLH_hgap.faを入力として実行した結果

	A	B	C	D	E	F	G	H
1451	1450	LOCUS_01450	sequence1	complement(1432847..1434211)	CDS	hypothetical protein		ATGTTTA
1452	1451	LOCUS_01451	sequence1	complement(1434402..1436963)	CDS	DNA gyrase subunit A	gyrA	GTGGCTG
1453	1452	LOCUS_01452	sequence1	complement(1436991..1438934)	CDS	DNA gyrase subunit B	gyrB	GTGAGCC
1454	1453	LOCUS_01453	sequence1	complement(1438934..1440055)	CDS	DNA replication and repair protein RecF	recF	ATGATTT
1455	1454	LOCUS_01454	sequence1	complement(1440065..1440289)	CDS	S4-like RNA binding protein		GTGCAAG
1456	1455	LOCUS_01455	sequence1	complement(1440531..1441670)	CDS	DNA polymerase III subunit beta	dnaN	ATGAAAT
1457	1456	LOCUS_01456	sequence1	complement(1441846..1443162)	CDS	chromosomal replication initiator protein DnaA	dnaA	ACTG
1458	1457	LOCUS_01457	sequence1	1443654..1443788	CDS	50S ribosomal protein L34	rpmH	ATGAAGC
1459	1458	LOCUS_01458	sequence1	1443859..1444203	CDS	ribonuclease P	rnpA	ATGAGAA
1460	1459	LOCUS_01459	sequence1	1444219..1445052	CDS	preprotein translocase subunit YidC	yidC_2	GTGAAAA
1461	1460	LOCUS_01460	sequence1	1445219..1446019	CDS	hypothetical protein		ATGGCAA
1462	1461	LOCUS_01461	sequence1	1446195..1447583	CDS	tRNA modification GTPase	trmE	GTGGCAC
1463	1462	LOCUS_01462	sequence1	1447617..1449545	CDS	tRNA uridine 5-carboxymethylaminomethyl mod	gidA	ATGGAGC
1464	1463	LOCUS_01463	sequence1	complement(1449639..1450622)	CDS	aldo/keto reductase		ATGAATT

W2-2: LH_draft2.fa

LH_draft2.faを入力としてDFASTを実行した結果。
①features.tsvをダウンロードしてdnaAで文字列検索してもよい、②Featuresタブで調べてもよい

DFAST Analysis Archive Download Help

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : 3e0898bd-8b70-4b00-b4b3-7f16555a0090
Status : COMPLETE

```
[2016-12-15 15:13:49.566123] Job submitted.  
[2016-12-15 15:13:49.578419] Job started.  
[2016-12-15 15:17:03.495028] Job completed.
```

Result **Features** DDBJ Submission Log

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,330

Download Files

- Genbank Flat File : [annotation.gbk](#)
- GFF3-formated File : [annotation.gff](#)
- Genome Fasta File : [genome.fna](#)
- Protein Fasta File : [protein.faa](#)
- CDS Fasta File : [cds.fna](#)
- RNA Fasta File : [rna.fna](#)
- Feature Table : [features.tsv](#)**
- Genome Statistics : [statistics.txt](#)
- Zip Archive : [annotation.zip](#)

W2-2:LH_draft2.fa

①Featuresタブで、②dnaAで文字列検索した結果。③dnaAは、chromosome上の相補鎖側(complement)の[1436009, 1437325 bp]領域にあった。転写開始点oriCは、おそらくこのあたりだと予測されるであろう、と妄想

The screenshot shows the DFAST Job Result page. At the top, there is a navigation bar with 'DFAST', 'Analysis', 'Archive', 'Download', and 'Help'. Below this, a warning message states: 'Remember the current URL to access this page. The result will be deleted 30 days after your last visit.' A 'Delete this job now.' button is present, with a note 'Delete This procedure cannot be undone.' The job details are: Title: (blank), JobID: 3e0898bd-8b70-4b00-b4b3-7f16555a0090, Status: COMPLETE. A log shows three entries: '[2016-12-15 15:13:49.566123] Job submitted.', '[2016-12-15 15:13:49.578419] Job started.', and '[2016-12-15 15:17:03.495028] Job completed.' The 'Features' tab is selected, indicated by a red arrow with the number 1. Below the tabs, the 'Annotated Features' section shows a search for 'dnaA' in the search box, also indicated by a red arrow with the number 2. The search results table has one entry: No. 1496, LocusTag LOCUS_01443, Seq. ID chromosome, Location complement (1436009..1437325), Feature Type CDS, Product chromosomal replication initiator protein DnaA, Gene dnaA, Nucleotide View, Translation View, and Edit. A red arrow with the number 3 points to the 'Location' column of this entry. At the bottom, it says 'Showing 1 to 1 of 1 entries (filtered from 2,490 total entries)' and has 'Previous', '1', and 'Next' buttons.

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => [Delete](#) This procedure cannot be undone.

Title :
JobID : 3e0898bd-8b70-4b00-b4b3-7f16555a0090
Status : COMPLETE

[2016-12-15 15:13:49.566123] Job submitted.
[2016-12-15 15:13:49.578419] Job started.
[2016-12-15 15:17:03.495028] Job completed.

Result **Features** DDBJ Submission Log

Annotated Features

Show 25 entries Search: dnaA

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
1496	LOCUS_01443	chromosome	complement (1436009..1437325)	CDS	chromosomal replication initiator protein DnaA	dnaA	View	View	Edit

Showing 1 to 1 of 1 entries (filtered from 2,490 total entries)

[Previous](#) **1** [Next](#)

W3-1 : Ori-Finder

①Ori-Finderのトップページ。②入カファイルの形式はFASTA format

http://tubic.tju.edu.cn/Ori-Finder/ Ori-Finder

Welcome to Ori-Finder

TUBIC

- Home
- Introduction
- Examples
- Database
- Download
- Reference
- Links

Enter or Paste a Complete Genome Sequence in FASTA format. [example](#)

参照...

Running Options

参照... Upload a ptt file [example](#)

Escherichia coli Select species-specific DnaA boxes

Define your own DnaA boxes

1 Unmatch site(s)

Output Options

Output the distribution of DnaA boxes

Output AT GC RY MK disparity curves

Submit Clear

*If you want to predict the replication origins in archaeal genome, please visit the [Ori-Finder 2](#) web server [[Luo, Zhang and Gao, 2014](#)].

*For the convenience of users' query, the predicted *oriC* regions have been organized into a MySQL database, called Doric, which is freely available at <http://tubic.tju.edu.cn/doric/>

(c) 2007.TUBIC all rights reserved.

W3-2: seq1_draft.fa

①LH_draft2.faの行数は6。②grepで配列数が3であり、最初の配列がchromosomeであることを確認。③LH_draft2.faの最初の2行分(染色体配列部分)を抽出してseq1_draft.faに保存。④seq1_draft.faのファイルサイズは2,277,996 bytes

```
iu@bielinux[iu] cd ~/Desktop/mac_share
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_draft2.fa
-rwxrwxrwx 1 iu iu 2400619  9月 23 20:25 LH_draft2.fa
iu@bielinux[mac_share] wc LH_draft2.fa
  6          6 2400619 LH_draft2.fa
iu@bielinux[mac_share] grep ">" LH_draft2.fa
>chromosome
>plasmid1
>plasmid2
iu@bielinux[mac_share] head -n 2 LH_draft2.fa > seq1_draft.fa
iu@bielinux[mac_share] ls -l seq1_draft.fa
-rwxrwxrwx 1 iu iu 2277996 12月 16 13:56 seq1_draft.fa
iu@bielinux[mac_share] █
```

①

②

③

④

W3-3: Ori-Finder

Ori-Finderを実行すべく、①参照。②共有フォルダ中の、③seq1_draft.faを選択して、④開く

The image shows the Ori-Finder web interface and a Windows file explorer window. The web interface has a navigation menu on the left with links: Home, Introduction, Examples, Database, Download, Reference, and Links. The main area has a text input field for a genome sequence in FASTA format, a '参照...' button, and 'Running Options' and 'Output Options' sections. The file explorer window shows a list of files in a shared folder, with 'seq1_draft.fa' selected. Red arrows point to the '参照...' button, the file path, the selected file, and the '開く(O)' button.

名前	更新日時	サイズ	種類
seq1_draft.fa	2016/12/16 13:56	2,225 KB	FA ファイル
sequence1_blast.txt	2016/09/25 21:38	12,182 KB	テキストドキュ...
LH_draft2.fa	2016/09/23 20:25	2,345 KB	FA ファイル
out2.bam.bai.zip	2016/09/13 20:03	4 KB	ZIP ファイル
out2.bam.zip	2016/09/13 20:02	69,651 KB	ZIP ファイル
uniqout.sam.zip	2016/09/13 20:02	94,135 KB	ZIP ファイル
out.sam.zip	2016/09/13 19:42	113,662 KB	ZIP ファイル
out-unique.var.flt.vcf	2016/09/12 17:28	7 KB	VCF ファイル
out2.bam.bai	2016/09/12 17:26	8 KB	BAI ファイル

W3-3: Ori-Finder

①生物種ごとに微妙にDnaA boxの9-merの塩基配列が異なるようである。しかし、②Lactobacillusはなさそうなので、とりあえずデフォルトのままで③Submit

The screenshot shows the Ori-Finder web interface. The browser address bar displays <http://tubic.tju.edu.cn/Ori-Finder/>. The page title is "Welcome to Ori-Finder". On the left, there is a navigation menu for TUBIC with links to Home, Introduction, Examples, Database, Download, Reference, and Links. The main content area has a header "Enter or Paste a Complete Genome Sequence in FASTA format." with an "example" link. Below this is a text input field containing "C:\Users\kadota\Desktop" and a "参照..." button. A large empty text area follows. The "Running Options" section includes a "参照..." button, an "Upload a ptt file" button with an "example" link, a dropdown menu set to "Escherichia coli" with the option "Select species-specific DnaA boxes" checked, an unchecked checkbox for "Define your own DnaA boxes" with a text input field containing "ttatccaca", and a dropdown menu for "Unmatch site(s)" set to "1". The "Output Options" section has "Output the distribution of DnaA boxes" checked, and "Output AT", "GC", "RY", and "MK disparity curves" all checked. At the bottom of the form are "Submit" and "Clear" buttons. A note at the bottom states: "*If you want to predict the replication origins in archaeal genome, please visit the [Ori-Finder 2](#) web server [Luo, Zhang and Gao, 2014].". Another note says: "*For the convenience of users' query, the predicted *oriC* regions have been organized into a MySQL database, called Doric, which is freely available at <http://tubic.tju.edu.cn/doric/>". The footer reads "(c) 2007.TUBIC all rights reserved.".

②

①

③

W3-3: Ori-Finder

約10秒ほどで結果が得られるが、①Introductionのリンク先に入出力・オプション・パラメータに関する記載があるので、そこをみておくとよい。例えば②GC skewの結果のみを見たいのであれば、他のAT, RY, MKのところのチェックを外しておけばよい。尚、disparityはおそらくskewと同じ意味で使っているものと思われます

Enter or Paste a Complete Genome Sequence in FASTA format. [example](#)

[Home](#)
[Introduction](#)
[Examples](#)
[Database](#)
[Download](#)
[Reference](#)
[Links](#)

Running Options [参照...](#) Upload a ptt file [example](#)

Escherichia coli Select species-specific DnaA boxes
 Define your own DnaA boxes

1 Unmatch

Output Options
Output the distribution of DnaA boxes
Output AT GC RY MK disparity curves

*If you want to predict the replication origins in archaeal genome, please visit the [Ori-Finder 2](#) web server [[Luo, Zhang and Gao, 2014](#)].

*For the convenience of users' query, the predicted *oriC* regions have been organized into a MySQL database, called Doric, which is freely available at <http://tubic.tju.edu.cn/doric/>

(c) 2007.TUBIC all rights reserved.

W3-4: Ori-Finder

Ori-Finder実行結果。①新規タブ上に表示されている。②下のほうまで眺めることで…

The information of genome and oriC region	
Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	653 nt
OriC AT content	0.6769
The number of DnaA box	6
The location of oriC region	1396338..1396990 nt
The location of dnaA gene	-
The extremes of GC disparity	1435740 nt (minimum), 337722 nt (maximum)
The extremes of AT disparity	1334199 nt (minimum), 342895 nt (maximum)
The extremes of RY disparity	1437338 nt (minimum), 342895 nt (maximum)
The extremes of MK disparity	337276 nt (minimum), 1432968 nt (maximum)
Note	Note that the E. coli perfect DnaA box (ttatccaca) was searched for with no more than one mismatch. [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identified in the below sequence are capitalized and also marked in bold, if any
	<pre>aactaacgacctttcaacaagTTATCCACAgagTTTCCACAggttaagatatttttaat taaaaaataagtttcaaaactctattgacttcggttttttaaaaatattttTTATCCACAggt ttgcacagaaaacactcgttcgtgatagcagttacaacagttgatgtaccaacgattaag actactttcgtttggttggttagTTATCCACAagtggtgggcccgtcTGTGGATAAC ttttcaactgctataaacagttggttcggggggtttatactttcttaacaattatacca gaaagttccacagagTTTCCACAgattaattatttcaaaaaatacttgcaaaatgac tactgaactacttctctgtaacttatttggttcgttaaatttaaccatttttaggc tttttgccgattatactagataaaatataatttgcgttcgtcttctctaattatgcc ttttcagatacctatatgtgggtaatacaatataattccgaactttaacttctaaatag tattataaaacggagatccaatgcaacctgtggtttgccaatctccggtattttatatta tatatattgtagaacgttccccccagaaggtatcagtcctaagttttatgtgcc</pre>
The information of genome and oriC region	
Genome size	2277983 nt

W3-4: Ori-Finder

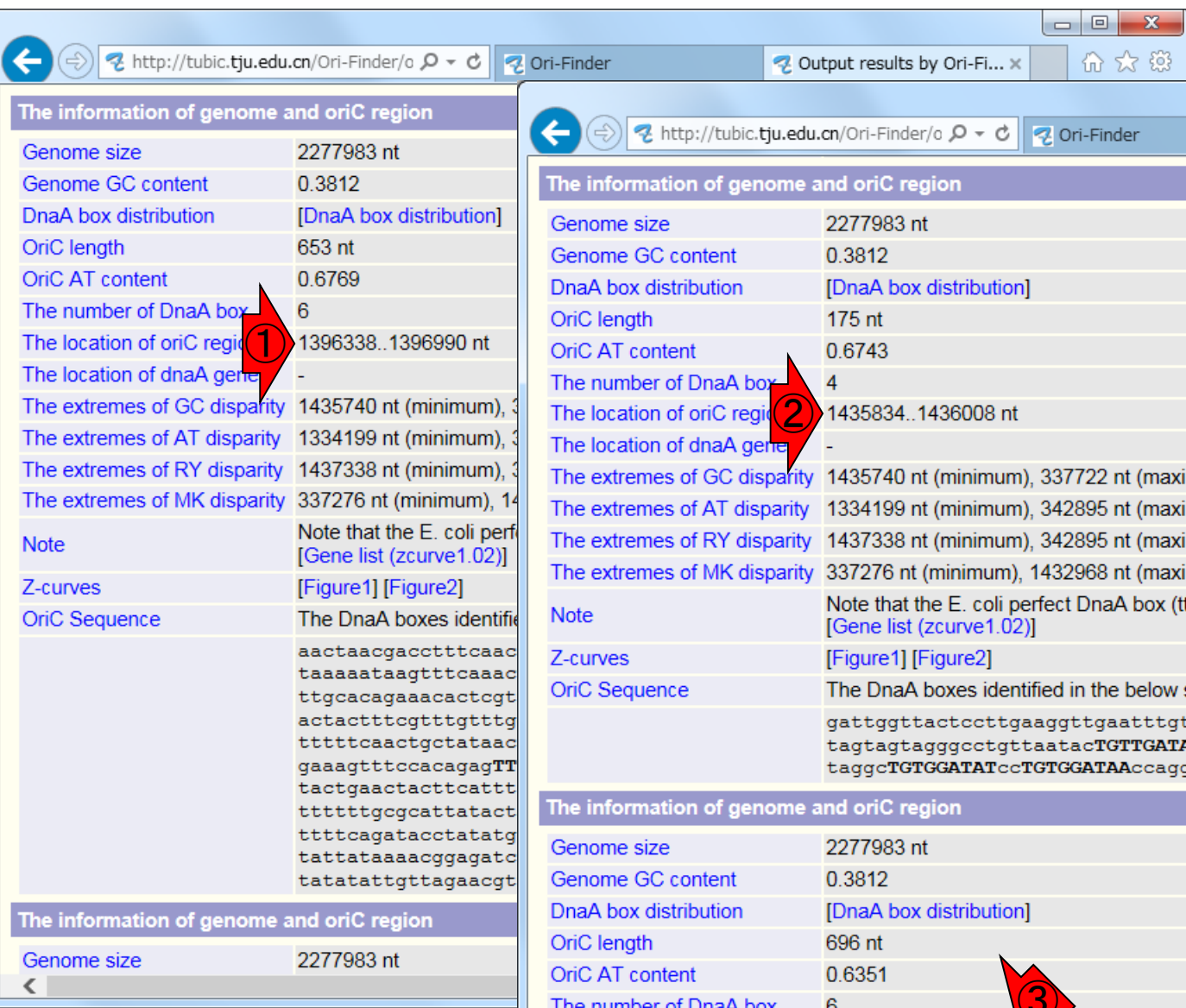
①一番下の状態。(同じような情報が全部で3つあったので)oriCの候補が、この場合は全部で3つあったことがわかります。例えば、ここで見えている3番目の候補のoriC領域は②[1437326, 1438021]

tagtagtagggcctgtaaacTGTGGATAAattaaagttcacaagaaggatacacggtttaggcTGTGGATATccTGTGGATAAccaggacaacttttaattagTTTTCCACAa

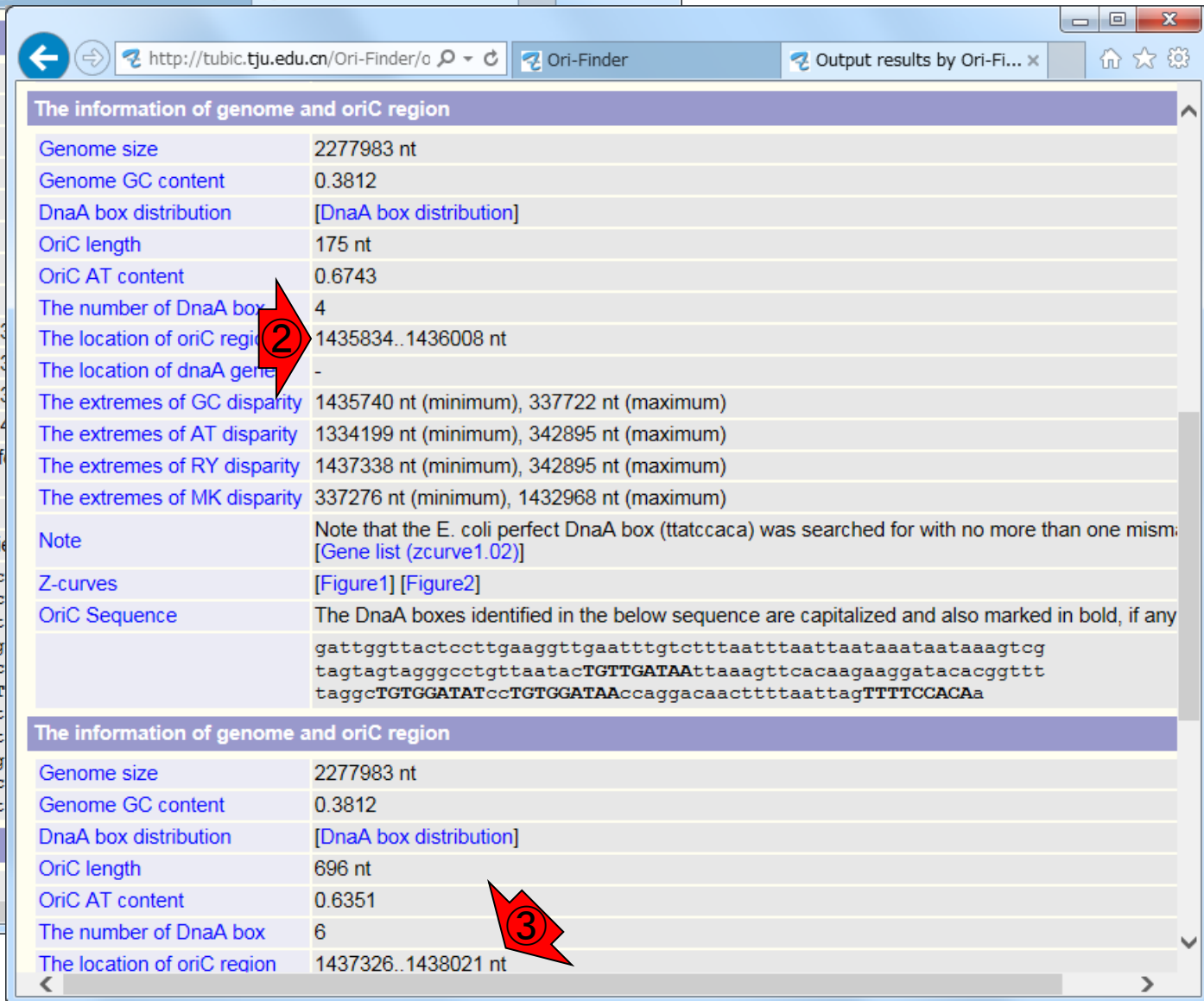
The information of genome and oriC region	
Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	696 nt
OriC AT content	0.6351
The number of DnaA box	6
The location of oriC region	1437326..1438021 nt
The location of dnaA gene	-
The extremes of GC disparity	1435740 nt (minimum), 337722 nt (maximum)
The extremes of AT disparity	1334199 nt (minimum), 342895 nt (maximum)
The extremes of RY disparity	1437338 nt (minimum), 342895 nt (maximum)
The extremes of MK disparity	337276 nt (minimum), 1432968 nt (maximum)
Note	Note that the E. coli perfect DnaA box (ttatccaca) was searched for with no more than one mism: [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identified in the below sequence are capitalized and also marked in bold, if any
	agtgactcctcctaataatgaaagtacaataaatagtttagcatgtcgagaaatagTTTTCC ACAagggttggtgaaatataatgtaataattgcacaactgtgtgtgtaaaTTATCCACAgc agtgtggaaagccaggtaaatgctaggctatacgaagTTATCCACAatTTTTTgtgcaaa gatacgtccgttgatcaaggcttggaggagTTATCCACTatTTTTTgcatagcTGTGGA TAAcctaatacgcagggtgttctTTTTgttcgtaaacggggaaaaaagcagttgaaaactt aaattgaatccacaggaaagaacaacgctgtggattgttttggtttaattTGTGGATAA tcaaaaatTTTTtattcTTTTtaggtgtttattatctTTTTtagaatgtgataagattgat gggtattgttgggaagataaaactTTTTcgactaataaaaatTTaaactaaattgttttagg aggTgcgaaatatagaagcgacatttcaacccaaaaaagcggcatcgttcacgtgttcacg gttccgtaaacgcatgagtactagtaaacggcctagtgatttagcacgggagacgtcaaa aaggcagaaaagtattatctgcatagggtcgtgacaaaactagctttgtcgcagccctt ttcatatctataggaatatttctggagattacgc

W3-4: Ori-Finder

3つのoriC候補領域の全貌。①[1396338, 1396990]、②[1435834, 1436008]、③[1437326, 1438021]



The information of genome and oriC region	
Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	653 nt
OriC AT content	0.6769
The number of DnaA boxes	6
The location of oriC region	1396338..1396990 nt
The location of dnaA gene	-
The extremes of GC disparity	1435740 nt (minimum), 337722 nt (maximum)
The extremes of AT disparity	1334199 nt (minimum), 342895 nt (maximum)
The extremes of RY disparity	1437338 nt (minimum), 342895 nt (maximum)
The extremes of MK disparity	337276 nt (minimum), 1432968 nt (maximum)
Note	Note that the E. coli perfect DnaA box (ttatccaca) was searched for with no more than one mismatch. [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identified in the below sequence are capitalized and also marked in bold, if any
<pre>aactaacgacctttcaac taaaaaataagtttcaaac ttgcacagaaacactcgt actactttcgtttgttg ttttcaactgctataaac gaaagtttccacagagTT tactgaactacttcattt tttttgcgccattatact ttttcagatacctatagc tattataaaaacggagatc tatatattgttagaacgt</pre>	



The information of genome and oriC region	
Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	175 nt
OriC AT content	0.6743
The number of DnaA boxes	4
The location of oriC region	1435834..1436008 nt
The location of dnaA gene	-
The extremes of GC disparity	1435740 nt (minimum), 337722 nt (maximum)
The extremes of AT disparity	1334199 nt (minimum), 342895 nt (maximum)
The extremes of RY disparity	1437338 nt (minimum), 342895 nt (maximum)
The extremes of MK disparity	337276 nt (minimum), 1432968 nt (maximum)
Note	Note that the E. coli perfect DnaA box (ttatccaca) was searched for with no more than one mismatch. [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identified in the below sequence are capitalized and also marked in bold, if any
<pre>gattggttactccttgaaggttgaatttgcctttaatttaattaataataaataaagtcg tagtagtagggcctgttaataacTGTTGATAAttaaagttcacaagaaggatacacggttt taggcTGTGGATATccTGTGGATAAaccaggacaacttttaattagTTTCCACAa</pre>	

The information of genome and oriC region	
Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	696 nt
OriC AT content	0.6351
The number of DnaA boxes	6
The location of oriC region	1437326..1438021 nt

W3-4: Ori-Finder

DFASTアノテーション結果(W2-2)より、dnaAは chromosome上の相補鎖側(complement)の[1436009, 1437325 bp]領域にあったことも踏まえ、③の3番目の候補領域[1437326, 1438021]が一番妥当と判断

The information of genome and oriC region

Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	653 nt
OriC AT content	0.6769
The number of DnaA box	6
The location of oriC region	1396338..1396990 nt
The location of dnaA gene	-
The extremes of GC disparity	1435740 nt (minimum), 337722 nt (maximum)
The extremes of AT disparity	1334199 nt (minimum), 342895 nt (maximum)
The extremes of RY disparity	1437338 nt (minimum), 342895 nt (maximum)
The extremes of MK disparity	337276 nt (minimum), 1432968 nt (maximum)
Note	Note that the E. coli perfect DnaA box (ttatccaca) was searched for with no more than one mismatch. [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identified in the below sequence are capitalized and also marked in bold, if any

The information of genome and oriC region

Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	175 nt
OriC AT content	0.6743
The number of DnaA box	4
The location of oriC region	1435834..1436008 nt
The location of dnaA gene	-
The extremes of GC disparity	1435740 nt (minimum), 337722 nt (maximum)
The extremes of AT disparity	1334199 nt (minimum), 342895 nt (maximum)
The extremes of RY disparity	1437338 nt (minimum), 342895 nt (maximum)
The extremes of MK disparity	337276 nt (minimum), 1432968 nt (maximum)
Note	Note that the E. coli perfect DnaA box (ttatccaca) was searched for with no more than one mismatch. [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identified in the below sequence are capitalized and also marked in bold, if any

The information of genome and oriC region

Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	696 nt
OriC AT content	0.6351
The number of DnaA box	6
The location of oriC region	1437326..1438021 nt



W3-5: Z-curve

http://tubic.tju.edu.cn/Ori-Finder/o Ori-Finder Output results by Ori-Fi...

```

tagtagtagggcctgttaataactTGTGGATAActaaagttcacaagaaggatacacggttt
taggcTGTGGATATccTGTGGATAAaccaggacaacttttaattagTTTCCACAa

```

The information of genome and oriC region

Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	696 nt
OriC AT content	0.6351
The number of DnaA box	6
The location of oriC region	1437326..1438021 nt
The location of dnaA gene	-
The extremes of GC disparity	1435740 nt (minimum), 337722 nt (maximum)
The extremes of AT disparity	1334199 nt (minimum), 342895 nt (maximum)
The extremes of RY disparity	1437338 nt (minimum), 342895 nt (maximum)
The extremes of MK disparity	337276 nt (minimum), 1432968 nt (maximum)
Note	Note that the E. coli perfect DnaA box (ttatccaca) was searched for with no more than one mism: [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identified in the below sequence are capitalized and also marked in bold, if any

```

agggactcctcctaataatgaaagtacaataaatagtttagcatgtcgcgagaaatagTTTTCC
ACAgggttggtgaaatataatgtaataattgcacaactgtgtgtgtaaaTTATCCACAgc
agtggtggaagccaggtaaatgctaggctatacgaagTTATCCACAtttttttgtgcaa
gatacgtccgttgatcaagcctggaggagTTATCCACTttttttgcatagcTGTGGA
TAActtaatcgacagggtgttcttttgcgtaaacaggggaaaaagcagttgaaaactt
aaattgaatccacaggaaagaacaacgctgtggattgttttggtttaatTGTGGATAAc
tcaaaaatttttattccttttaggtgtttattatcttttagaatgtgataagattgat
gggtattgtttggaagataaaacttttcgactaataaaaatataactaaattgttttagg
agtgcgaaatatgaagcgcacatttcaacccaaaaagcggcatcgttcacgtgttcacg
gttccgtaaacgcgatgagtactagtaacggcctagtgatttagcacgggagacgtcaaa
aaggcagaaaagtattatctgcatagggtcgtgacaaaactagctttgtcgcagccctt
ttcatatctataggaatatttctggagattacagc

```

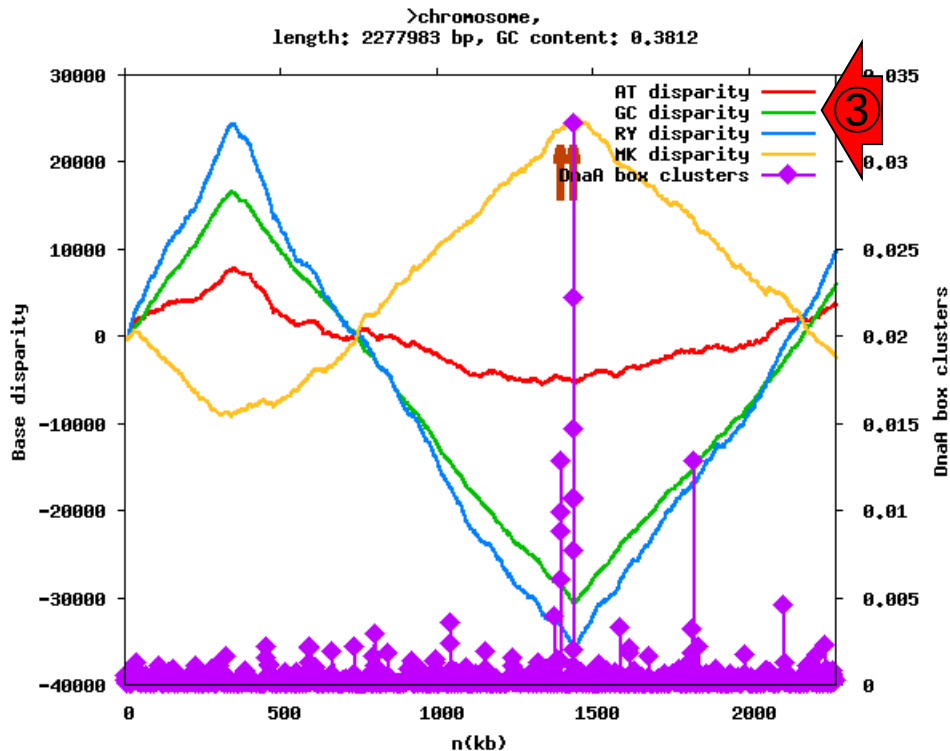
W3-5: Z-curve

①横軸は配列全体の長さに相当するようだ。②縦軸の Base disparityが、③緑の線で表されているGCの場合はGC skewに相当するものだろうと判断できる。disparityは格差とか不均衡という意味。Z-curve法の詳細がわからなくとも、候補領域[1437326, 1438021]周辺でGC disparityが極小値となっていることから、極小値を探しているのだろうと判断

the Z-curves for the inquired organism

Figure1 shows the Z-curves (AT, GC, RY and MK disparity curves) for the original sequence. Figure2 shows the Z-curves (AT, GC, RY and MK disparity curves) for the rotated sequence beginning and ending in the maximum of the GC disparity curve. Short vertical red line indicates the indicator gene (such as dnaA, dnaN, gidA, hemE etc) location, and short up vertical dark blue arrow indicates the identified oriC location. Purple peaks with the diamonds indicates the DnaA box clusters. [back to main]

Legend



W4-1: dnaAの開始コドン

① *dnaA* 遺伝子領域の最初の10塩基分を表示。相補鎖側なので② AATCAGTCACから、GTGACTGATTと読み解き、開始コドンが一般的なATGではなくGTGになっていることに気づく

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_draft2.fa
-rwxrwxrwx 1 iu iu 2400619  9月 23 20:25 LH_draft2.fa
iu@bielinux[mac_share] head -2 LH_draft2.fa | tail -1 | cut -c 14373
16-1437325
AATCAGTCAC
iu@bielinux[mac_share] █
```

[3:52午後]

[3:52午後]



W4-3: JQCH01000005.1

他の研究グループから出された *L. hokkaidonensis* の基準株のゲノム配列

The screenshot shows the NCBI GenBank entry for Lactobacillus hokkaidonensis strain DSM 26202 Scaffold5, whole genome shotgun sequence. The page includes a search bar, navigation links, and detailed sequence information.

GenBank: JQCH01000005.1

[FASTA](#) [Graphics](#)

Go to:

LOCUS JQCH01000005 96182 bp DNA linear BCT 06-NOV-2015

DEFINITION *Lactobacillus hokkaidonensis* strain DSM 26202 Scaffold5, whole genome shotgun sequence.

ACCESSION JQCH01000005 [JQCH01000000](#)

VERSION JQCH01000005.1

DBLINK BioProject: [PRJNA222257](#)
BioSample: [SAMN02797790](#)

KEYWORDS WGS.

SOURCE *Lactobacillus hokkaidonensis*

ORGANISM [Lactobacillus hokkaidonensis](#)
Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus.

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Related information

Assembly

BioProject

BioSample

W4-3: JQCH01000005.1

① chromosomal replication initiation protein で検索。② CDSをクリック

https://www.ncbi.nlm.nih.gov/nucore/JQCH01000005.1 Lactobacillus hokkaidon... x

検索: chromosomal replication initiation protein ① 前へ 次へ オプション 1件の一致

```
/transl_table=11
/product="stage III sporulation protein J"
/protein_id="KRO10455.1"
/translation="MKKHLKKYLTMASILSLALFLAACSNKPINSRSTGIWDKGIVYS
FSRAIIWLAHTFGGYGMGIIIFTIIVRIVILPLMIYQTKSMRKTQELQPKLKALQKKY
SSKDMETVRKQLQEEQKQLYSEAGVNPFFASILPMIVQLPVIYALYQAIWRTDVLGGSF
LWLELGHDPYFILPILAAVFTFFSSWLSMASQPEKNTMTSMTWVMPIMI FFMAMNI
SSAISLYWVVTNAFQVGQTLVIQNPFKINRERAEKEAAEKAHRKAVEKAKRNARKSRR
K"
gene complement (86295..86639)
/locus_tag="IV59_GL001855"
CDS complement (86295..86639)
/locus_tag="IV59_GL001855"
/codon_start=1
/transl_table=11
/product="ribonuclease P"
/protein_id="KRO10456.1"
/translation="MRKSYRVKKESEFQQVFETHNSVANRKFIIYTMEKPGQKHFRIG
ISVGKKIGNAVHRNWWKRRIRQSVLELKDQLKPCDFLVIARRSADQMSMAETKQHLT
HALKLAHLIDEH"
gene 87294..88652
/locus_tag="IV59_GL001856"
CDS 87294..88652
/locus_tag="IV59_GL001856"
/codon_start=1
/transl_table=11
/product="chromosomal replication initiation protein"
/protein_id="KRO10457.1"
/translation="MLNYLLYFHLGGVTVTDLETLDWTIKESFKNDVSSVTYENLIEP
```

①確かにATGから始まっていることがわかる

W4-3: JQCH01000005.1

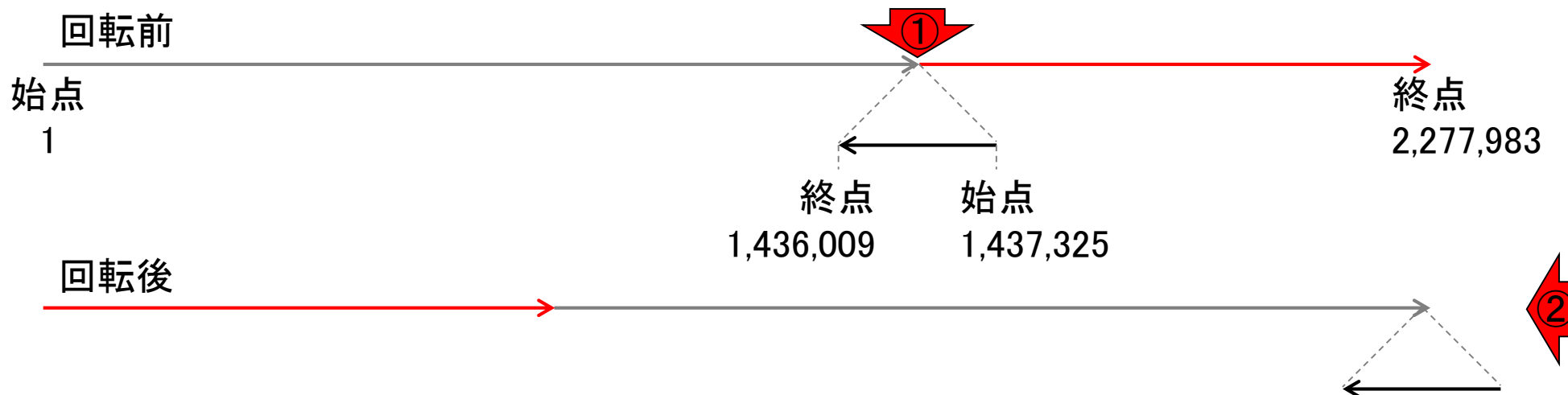
The screenshot shows a web browser window displaying the NCBI Nucleotide database entry for JQCH01000005.1. The search term is "chromosomal replication initiation protein". The sequence is shown in a table format with line numbers on the left. A red arrow with the number "1" points to the sequence "gacatgctaa" on line 87241, which is highlighted in brown. Below the sequence, a details box is open, showing the following information:

```
87294..88652  
/locus_tag="IV59_GL001856"  
/codon_start=1  
/transl_table= 11  
/product="chromosomal replication initiation protein"  
/protein_id=" KRO10457.1 "  
/translation="MLNYLLYFHLGGVTVDLETLWDTIKESFKNDVSSVTYENLIEP  
AKALSLANNQLTIELPTPFHKDYWRDSLTDKFRLYAFEATSEKIEPRFIIETEENTIT  
PEAPSKPSFKVNTALNPRYTFDTFVIGKGNQMAHAAALVVSEEPGKMYNPLFFYGGVG  
LGKTHLMAIIGNKMMASNAQTRVKYVTSEAFNTDFINAIQTGTQEQFRQYRKLDDLL  
VDDIQFADKEGTQEEFFHTFNALYEENKQIVLTSRDLPEI PKLQDRLVSRFKWGLS  
VDITPPDLETRIAILRSKAKADQIDIPNDTLTYIAGQIDSNVRELEGALSRVQAYSRL  
NKATINTSLASDALGKGLAGTSALT IPLIQEKVASYHYHSQTDLKGKRVKQIVVPR  
QIAMYLVRELTDNSLPKIGAEFGGKDHTTVLHSIDKIEEMLQSDHNLQEDVQNLKME  
LKP"
```

At the bottom of the browser window, there are navigation controls including "CDS", "Feature", "70 of 75", "JQCH01000005 : 1 segment", "Details", and "Display: FASTA GenBank Help".

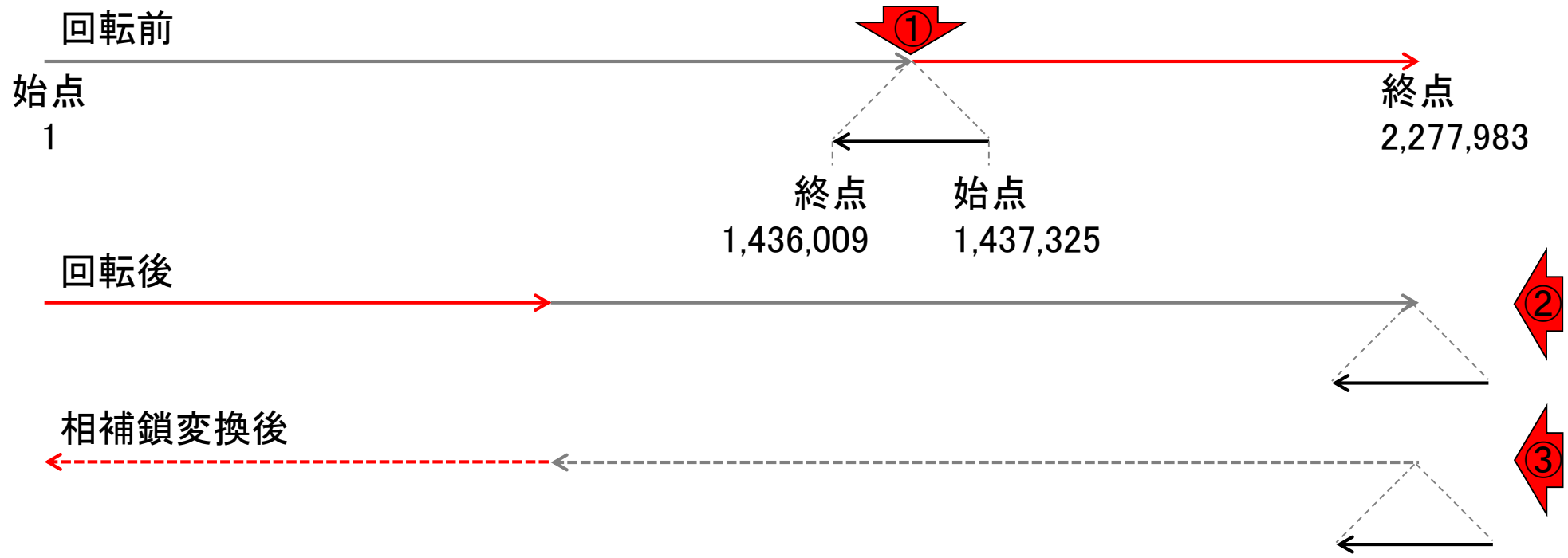
W4-4: 回転前後

染色体配列の回転前後の関係(イメージ図)。① DFASTアノテーション結果では *dnaA* 遺伝子が相補鎖側となっていたため(それゆえ矢印の向きが逆)、② 第1段階として、*dnaA* 転写開始点よりも100塩基上流まで含む灰色矢印領域[1, 1437425]と赤色矢印領域[1437426, 2277983]を入れ替える(回転させる)



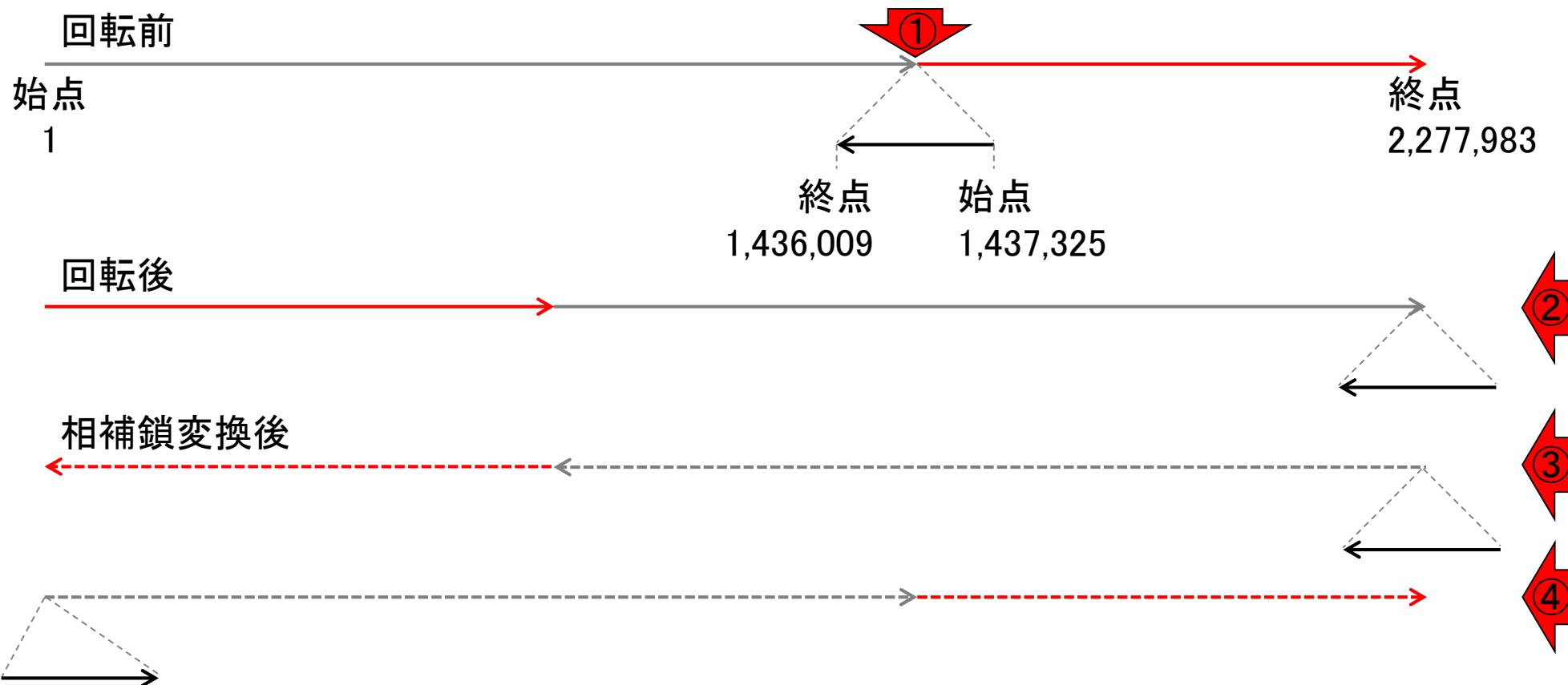
③第2段階として相補鎖変換を行う。点線は、実線で示した配列の相補鎖を表す

W4-4: 相補鎖変換後

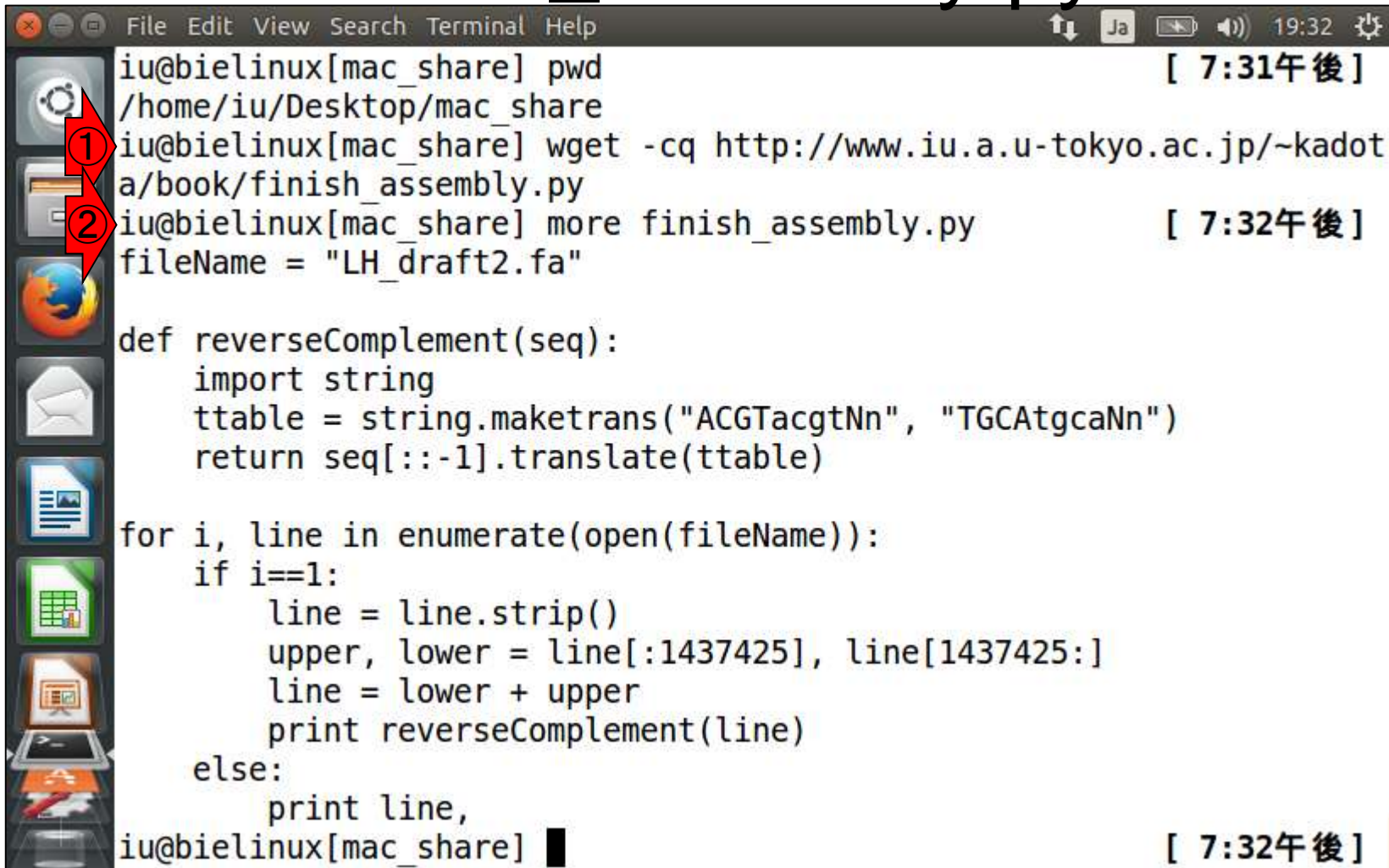


W4-4: 相補鎖変換後

④は③を逆向きにただけで、実質的に同じものです。それゆえ手順としてはreverse complementにしているが、逆相補鎖と強調する必要はない。ここまで行った配列を再度DFASTにかけたら、*dnaA*遺伝子が先頭の順鎖側にアノテーションされるはず(一種の答え合わせ)



W4-5: finish_assembly.py



```
iu@bielinux[mac_share] pwd [ 7:31午後 ]
/home/iu/Desktop/mac_share
① iu@bielinux[mac_share] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kadot
a/book/finish_assembly.py
② iu@bielinux[mac_share] more finish_assembly.py [ 7:32午後 ]
fileName = "LH_draft2.fa"

def reverseComplement(seq):
    import string
    ttable = string.maketrans("ACGTacgtNn", "TGCAtgcaNn")
    return seq[::-1].translate(ttable)

for i, line in enumerate(open(fileName)):
    if i==1:
        line = line.strip()
        upper, lower = line[:1437425], line[1437425:]
        line = lower + upper
        print reverseComplement(line)
    else:
        print line,
iu@bielinux[mac_share] [ 7:32午後 ]
```

W4-5: finish_assembly.py

①入力ファイルはLH_draft2.fa。②逆相補鎖(reverse complement)を得る関数を定義している

```
iu@bielinux[mac_share] pwd [ 7:31午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kadota/book/finish_assembly.py
iu@bielinux[mac_share] more finish_assembly.py [ 7:32午後 ]
fileName = "LH_draft2.fa" ①

def reverseComplement(seq):
    import string
    ttable = string.maketrans("ACGTacgtNn", "TGCAtgcaNn")
    return seq[::-1].translate(ttable) ②

for i, line in enumerate(open(fileName)):
    if i==1:
        line = line.strip()
        upper, lower = line[:1437425], line[1437425:]
        line = lower + upper
        print reverseComplement(line)
    else:
        print line,
iu@bielinux[mac_share] [ 7:32午後 ]
```

W4-5: finish_assemb

①LH_draft2.faは3配列からなるが、2行目の塩基配列のみ回転と相補鎖変換したいので、②ifと③elseで条件分岐させている。Pythonはi=0からスタートするので、i=1が入力ファイル中の2行目の情報のみ取り扱っていることに相当する

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kadota/book/finish_assembly.py
iu@bielinux[mac_share] more finish_assembly.py [ 7:32午後 ]
fileName = "LH_draft2.fa" ①

def reverseComplement(seq):
    import string
    ttable = string.maketrans("ACGTacgtNn", "TGCAtgcaNn")
    return seq[::-1].translate(ttable)

for i, line in enumerate(open(fileName)):
    ② if i==1:
        line = line.strip()
        upper, lower = line[:1437425], line[1437425:]
        line = lower + upper
        print reverseComplement(line)
    ③ else:
        print line,
iu@bielinux[mac_share] [ 7:32午後 ]
```

W4-5: finish_assembly

①ここに複製開始点にしたい塩基位置情報を与える。0番目からカウントするだとかそういったことは、結論としては考えなくてよい

```
iu@bielinux[mac_share] pwd [ 7:31午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wget -cq http://www.iu.a.u-tokyo.ac.jp/~kadota/book/finish_assembly.py
iu@bielinux[mac_share] more finish_assembly.py [ 7:32午後 ]
fileName = "LH_draft2.fa"

def reverseComplement(seq):
    import string
    ttable = string.maketrans("ACGTacgtNn", "TGCAtgcaNn")
    return seq[::-1].translate(ttable)

for i, line in enumerate(open(fileName)):
    if i==1:
        line = line.strip()
        upper, lower = line[:1437425], line[1437425:]
        line = lower + upper
        print reverseComplement(line)
    else:
        print line,
iu@bielinux[mac_share] [ 7:32午後 ]
```



W4-5: finish_asse

Pythonの範囲指定は、a:bで表現する。「a:b」は、a番目からb番目”の手前”までを抜き出す指定になる。「:b」は略記法であり、0番目からb番目の手前までの指定となる。また、「a:」は、a番目から最後まで指定となる。したがって①の「:1437425」は、0番目から1437425番目の手前までを指定していることになる。また、②の「1437425:」は、1437425番目から最後までを指定していることになる

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wget -cq http://www.biomedcentral.com/journal/14718/a/book/finish_assembly.py
iu@bielinux[mac_share] more finish_assembly.py [ 7:32午後 ]
fileName = "LH_draft2.fa"

def reverseComplement(seq):
    import string
    ttable = string.maketrans("ACGTacgtNn", "TGCAtgcaNn")
    return seq[::-1].translate(ttable)

for i, line in enumerate(open(fileName)):
    if i==1:
        line = line.strip()
        upper, lower = line[:1437425], line[1437425:]
        line = lower + upper
        print reverseComplement(line)
    else:
        print line,
iu@bielinux[mac_share] [ 7:32午後 ]
```



W4-6: 実行

①finish_assembly.pyを実行した結果をLH_complete.faというファイル名で保存。②ファイルサイズが変わっていないのが、うまくいっていることの傍証。理由は位置を入れ替えて相補鎖変換しているだけだから

```
iu@bielinux[mac_share] pwd [ 7:33午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l finish* LH_dra* [ 7:33午後 ]
-rwxrwxrwx 1 iu iu 418 12月 19 19:30 finish_assembly.py
-rwxrwxrwx 1 iu iu 2400619 9月 23 20:25 LH_draft2.fa
-rwxrwxrwx 1 iu iu 2400619 9月 10 16:52 LH_draft.fa
iu@bielinux[mac_share] python finish_assembly.py > LH_complete.fa
iu@bielinux[mac_share] ls -l finish* LH_* [ 7:33午後 ]
-rwxrwxrwx 1 iu iu 418 12月 19 19:30 finish_assembly.py
-rwxrwxrwx 1 iu iu 2400619 12月 19 2016 LH_complete.fa
-rwxrwxrwx 1 iu iu 2400619 9月 23 20:25 LH_draft2.fa
-rwxrwxrwx 1 iu iu 2400619 9月 10 16:52 LH_draft.fa
-rwxrwxrwx 1 iu iu 2433662 8月 30 21:22 LH_hgap.fa
iu@bielinux[mac_share] █ [ 7:33午後 ]
```



W4-7: 動作確認

動作確認。①回転および相補鎖変換前のLH_draft2.faファイルに対して、開始位置としたいdnaA遺伝子の転写開始点(1437325番目)の上流100塩基分の位置(1437425)から上流20塩基分の塩基配列を表示。②回転および相補鎖変換後のLH_complete.faファイルに対して、最初の20塩基分を表示。確かにうまく動作していることがわかる

```
File Edit View Search Terminal Help
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_dr
-rwxrwxrwx 1 iu iu 2400619 12月 19 19:33 LH_complete.fa
-rwxrwxrwx 1 iu iu 2400619  9月 23 20:25 LH_draft2.fa
① iu@bielinux[mac_share] head -2 LH_draft2.fa | tail -1 | cut -c 14374
06-1437425
ATGTAATATTGCACAACACTGT
② iu@bielinux[mac_share] head -2 LH_complete.fa | tail -1 | cut -c 1-2
0
ACAGTTGTGCAATATTACAT
iu@bielinux[mac_share] █ [ 7:33午後 ]
```

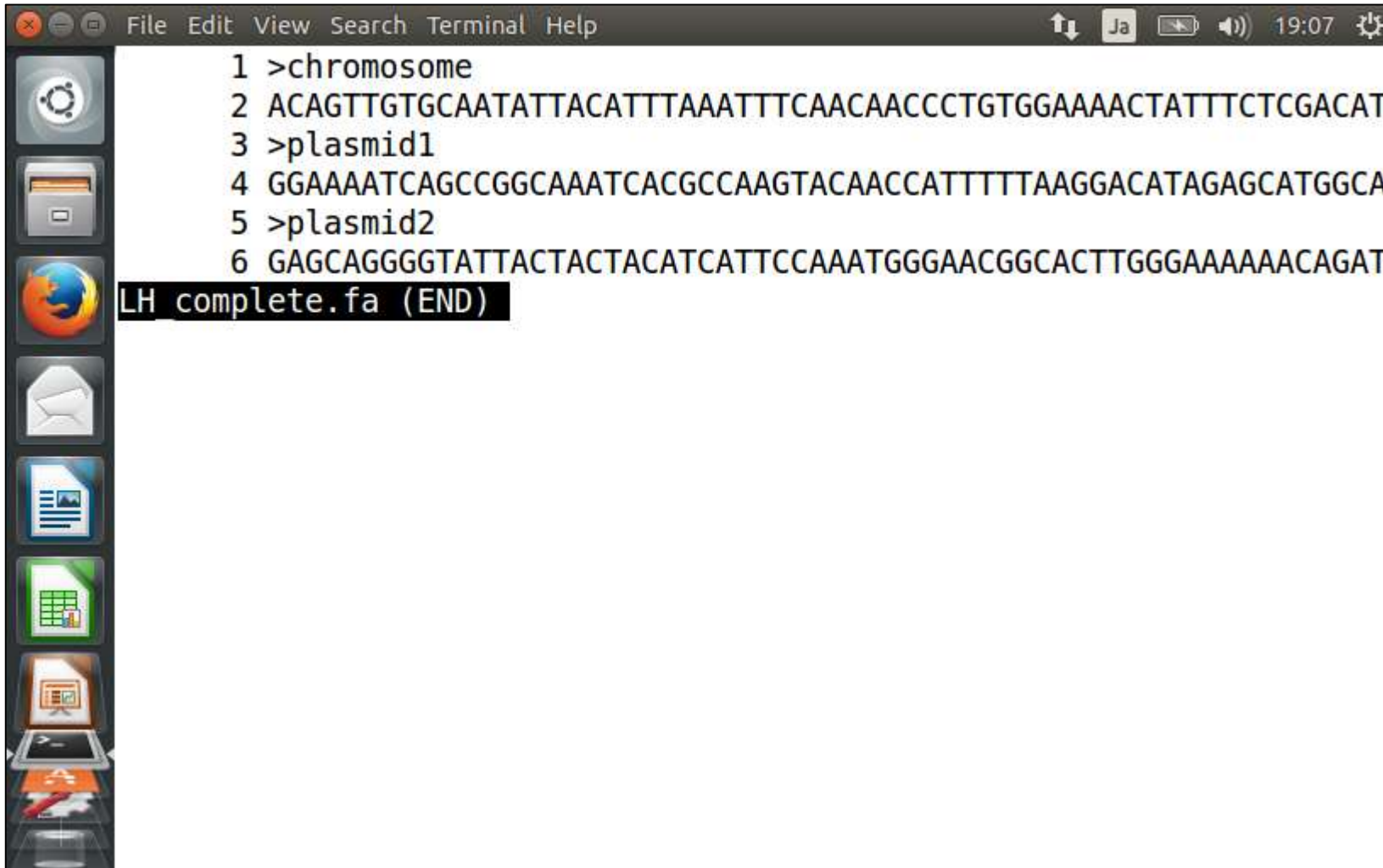
W4-7: 動作確認

動作確認。①less -N -Sで表示。一行が非常に長い場合には-Sが便利です。-Nは行番号の表示です。第8回W16のおさらいです

```
iu@bielinux[mac_share] pwd [ 7:33午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_draft2.fa LH_complete.fa
-rwxrwxrwx 1 iu iu 2400619 12月 19 19:33 LH_complete.fa
-rwxrwxrwx 1 iu iu 2400619  9月 23 20:25 LH_draft2.fa
iu@bielinux[mac_share] head -2 LH_draft2.fa | tail -1 | cut -c 14374
06-1437425
ATGTAATATTGCACAACACTGT
iu@bielinux[mac_share] head -2 LH_complete.fa | tail -1 | cut -c 1-2
0
ACAGTTGTGCAATATTACAT
iu@bielinux[mac_share] less -N -S LH_complete.fa [ 7:33午後 ]
```



W4-7: 動作確認



```
1 >chromosome
2 ACAGTTGTGCAATATTACATTTAAATTTCAACAACCCTGTGGAAACTATTTCTCGACAT
3 >plasmid1
4 GGAAAATCAGCCGGCAAATCACGCCAAGTACAACCATTTTAAAGGACATAGAGCATGGCA
5 >plasmid2
6 GAGCAGGGGTATTACTACTACATCATTCCAATGGGAACGGCACTTGGGAAAAAACAGAT
LH complete.fa (END)
```

W4-8: もう一つの手段

①まずLinuxコマンドを駆使して染色体配列のみを回転させた結果をhoge.faに格納

```
iu@bielinux[mac_share] pwd [ 7:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] echo ">chromosome" > hoge.fa [ 7:18午後 ]
iu@bielinux[mac_share] head -2 LH_draft2.fa | tail -1 | cut -c 14374
26-2277983 >> hoge.fa
iu@bielinux[mac_share] head -2 LH_draft2.fa | tail -1 | cut -c 1-143
7425 >> hoge.fa
iu@bielinux[mac_share] revseq -sequence hoge.fa -outseq LH_complete2
.fa
Reverse and complement a nucleotide sequence
iu@bielinux[mac_share] head -n 5 LH_complete2.fa [ 7:18午後 ]
>chromosome Reversed:
ACAGTTGTGCAATATTACATTTAAATTTCAACAACCCTGTGGAAAACATTTTCTCGACAT
GCTAAACTATTTATTGTACTTTTATTTAGGAGGAGTCACTGTGACTGATTTAGAAACACT
TTGGGACACAATTAAGAATCTTTTAAGAATGACGTTTCTAGCGTCACGTATGAAAATTT
GATTGAACCTGCTAAGGCACTGTCATTAGCGAACAACCAACTGACAATTGAATTACCAAC
iu@bielinux[mac_share] [ 7:18午後 ]
```



W4-8: もう一つの手

次に、Bio-Linuxにプレインストールされている、①EMBOSSのrevseqプログラムを用いてhoge.faのreverse complementを得た結果をLH_complete2.faに保存し、②最初の5行分を表示。revseq実行結果ファイルは、1行あたり60文字になっている点に注意。また、この段階ではまだLH_complete2.faには染色体配列しか格納されていない点にも注意

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] echo ">chromosome
iu@bielinux[mac_share] head -2 LH_draft2
26-2277983 >> hoge.fa
iu@bielinux[mac_share] head -2 LH_draft2.fa | tail -1 | cut -c 1-143
7425 >> hoge.fa
① iu@bielinux[mac_share] revseq -sequence hoge.fa -outseq LH_complete2
.fa
Reverse and complement a nucleotide sequence
② iu@bielinux[mac_share] head -n 5 LH_complete2.fa [ 7:18午後 ]
>chromosome Reversed:
ACAGTTGTGCAATATTACATTTAAATTTCAACAACCCTGTGGAAAACATTTTCTCGACAT
GCTAAACTATTTATTGTACTTTCATTTAGGAGGAGTCACTGTGACTGATTTAGAAACACT
TTGGGACACAATTAAGAATCTTTTAAGAATGACGTTTCTAGCGTCACGTATGAAAATTT
GATTGAACCTGCTAAGGCACTGTCATTAGCGAACAACCAACTGACAATTGAATTACCAAC
iu@bielinux[mac_share] [ 7:18午後 ]
```

W4-8: もう一つの手段

LH_draft2.fa中の最後の4行分がプラスミド配列に相当するので、それをLH_complete2.faに追加保存している。実行前後でファイルサイズが2,315,972 bytesから2,438,595 bytesになっていることから、確かに追加されたのだろうと判断

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_complete2.fa [ 7:31午後 ]
-rwxrwxrwx 1 iu iu 2315972 12月 20 19:18 LH_complete2.fa
iu@bielinux[mac_share] tail -n 4 LH_draft2.fa >> LH_complete2.fa
iu@bielinux[mac_share] ls -l LH_complete2.fa [ 7:31午後 ]
-rwxrwxrwx 1 iu iu 2438595 12月 20 19:31 LH_complete2.fa
iu@bielinux[mac_share] ls -l LH_complete.fa [ 7:31午後 ]
-rwxrwxrwx 1 iu iu 2400619 12月 19 19:33 LH_complete.fa
iu@bielinux[mac_share] █ [ 7:32午後 ]
```



W4-8: もう一つの手段

Pythonプログラム実行で得られた
LH_complete.faとファイルサイズよりも大
きくなるのは、revseq実行結果の段階で1
行60文字で改行が施されているためです

```
iu@bielinux[mac_share] pwd [ 7:31午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_complete2.fa [ 7:31午後 ]
-rwxrwxrwx 1 iu iu 2315972 12月 20 19:18 LH_complete2.fa
iu@bielinux[mac_share] tail -n 4 LH_draft2.fa >> LH_complete2.fa
iu@bielinux[mac_share] ls -l LH_complete2.fa [ 7:31午後 ]
-rwxrwxrwx 1 iu iu 2438595 12月 20 19:31 LH_complete2.fa
① iu@bielinux[mac_share] ls -l LH_complete.fa [ 7:31午後 ]
-rwxrwxrwx 1 iu iu 2400619 12月 19 19:33 LH_complete.fa
iu@bielinux[mac_share] [ 7:32午後 ]
```

W4-9:しつこく確認

本当にLH_complete.faとLH_complete2.faが同じ配列になっているかどうかを、ACGTの割合で確認。まずは①LH_complete.fa

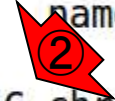
```
File Edit View Search Terminal Help
ge 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
> in_f <- "LH_complete.fa" #入力ファイル名を指定してin_fに格納
> fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
> fasta #確認してるだけです
A DNAStringSet instance of length 3
      width seq                      names
[1] 2277983 ACAGTTGTGCAATATTACATTT...CTGCTGTGGATAATTTACACAC chromosome
[2]   81630 GGAAAATCAGCCGGCAAATCAC...GTAGTAAACAGTGGACCACTTT plasmid1
[3]   40971 GAGCAGGGGTATTACTACTACA...TGGATTTAGAACTGTTGATGAT plasmid2
> hoge <- alphabetFrequency(fasta)[,1:4] #A,C,G,Tの数を各配列ごとにカウントした結果をhogeに格納
> colSums(hoge)/sum(as.numeric(hoge)) #A,C,G,Tの組成を表示
      A      C      G      T
0.3083008 0.1926315 0.1895755 0.3094922
>
```



W4-9:しつこく確認

本当にLH_complete.faとLH_complete2.faが同じ配列になっているかどうかを、ACGTの割合で確認。次は①LH_complete2.fa。完璧に同じっぽいことは②塩基配列の最初と最後のほうが同じことから推測可能。③Rの終了

```
File Edit View Search Terminal Help
> colSums(hoge)/sum(as.numeric(hoge)) #A,C,G,T
      A      C      G      T
0.3083008 0.1926315 0.1895755 0.3094922
> in_f <- "LH_complete2.fa" #入力ファイル名を指定してin_fに格納
> fasta <- readDNAStringSet(in_f, format="fasta") #in_fで指定したファイルの読み込み
> fasta #確認してるだけです
A DNAStringSet instance of length 3
      width seq
[1] 2277983 ACAGTTGTGCAATATTACATTT...CTGCTGTGGATAATTTACACAC chromosome Reversed:
[2] 81630 GGAAAATCAGCCGGCAAATCAC...GTAGTAAACAGTGGACCACTTT plasmid1
[3] 40971 GAGCAGGGGTATTACTACTACA...TGGATTTAGAACTGTTGATGAT plasmid2
> hoge <- alphabetFrequency(fasta)[,1:4] #A,C,G,Tの数を各配列ごとにカウントした結果をhogeに格納
> colSums(hoge)/sum(as.numeric(hoge)) #A,C,G,Tの組成を表示
      A      C      G      T
0.3083008 0.1926315 0.1895755 0.3094922
> q(save="no")
```

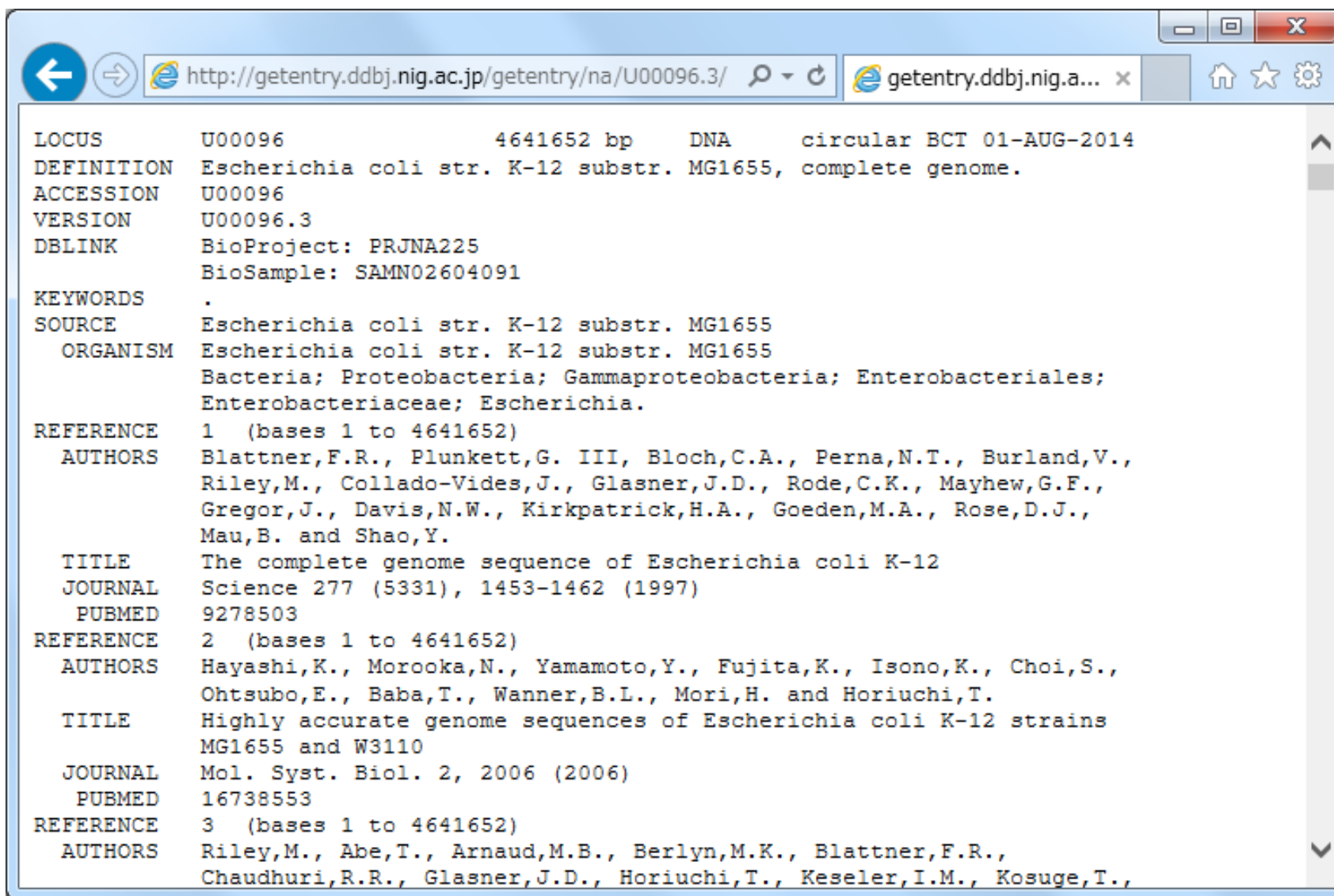


W4-10:ねちっこく確認

本当にLH_complete.faとLH_complete2.faが同じ配列になっているかどうかを、①Rのsetequal関数を用いて集合として等しいかどうかでも確認。TRUEが得られたので安心。②は簡単な文字列での動作確認

```
File Edit View Search Terminal Help
ge 'S4Vectors'
Loading required package: IRanges
Loading required package: XVector
> in_f <- "LH_complete.fa" #入力ファイル名を指定してin_fに格納
> fasta1 <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
>
> in_f <- "LH_complete2.fa" #入力ファイル名を指定してin_fに格納
> fasta2 <- readDNAStringSet(in_f, format="fasta")#in_fで指定したファイルの読み込み
>
① > setequal(fasta1, fasta2)
[1] TRUE
> setequal("AGGT", "ACGT")
[1] FALSE
② > setequal("AGGT", "AGGT")
[1] TRUE
> q(save="no")
iu@bielinux[mac_share] [ 8:15午後 ]
```

W5-1 : U00096



LOCUS U00096 4641652 bp DNA circular BCT 01-AUG-2014

DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.

ACCESSION U00096

VERSION U00096.3

DBLINK BioProject: PRJNA225
BioSample: SAMN02604091

KEYWORDS .

SOURCE Escherichia coli str. K-12 substr. MG1655

ORGANISM Escherichia coli str. K-12 substr. MG1655
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
Enterobacteriaceae; Escherichia.

REFERENCE 1 (bases 1 to 4641652)

AUTHORS Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
Mau,B. and Shao,Y.

TITLE The complete genome sequence of Escherichia coli K-12

JOURNAL Science 277 (5331), 1453-1462 (1997)

PUBMED 9278503

REFERENCE 2 (bases 1 to 4641652)

AUTHORS Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.

TITLE Highly accurate genome sequences of Escherichia coli K-12 strains
MG1655 and W3110

JOURNAL Mol. Syst. Biol. 2, 2006 (2006)

PUBMED 16738553

REFERENCE 3 (bases 1 to 4641652)

AUTHORS Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,

<http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/>

W5-1 : U00096

①U00096はこのentryの識別子(アクセッション番号)、②ピリオドの後の3はバージョン番号。これは最初の登録から2回更新されていることを意味する。登録直後はU00096.1、1回目の更新後にU00096.2、そして2回目の更新後にU00096.3となります

LOCUS U00096 4641652 bp DNA circular BCT 01-AUG-2014
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION U00096
VERSION U00096.3
DBLINK BioProject PRJNA225
BioSample SAMN02604091
KEYWORDS .
SOURCE Escherichia coli str. K-12 substr. MG1655
ORGANISM Escherichia coli str. K-12 substr. MG1655
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
Enterobacteriaceae; Escherichia.
REFERENCE 1 (bases 1 to 4641652)
AUTHORS Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
Mau,B. and Shao,Y.
TITLE The complete genome sequence of Escherichia coli K-12
JOURNAL Science 277 (5331), 1453-1462 (1997)
PUBMED 9278503
REFERENCE 2 (bases 1 to 4641652)
AUTHORS Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
TITLE Highly accurate genome sequences of Escherichia coli K-12 strains
MG1655 and W3110
JOURNAL Mol. Syst. Biol. 2, 2006 (2006)
PUBMED 16738553
REFERENCE 3 (bases 1 to 4641652)
AUTHORS Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,

W5-1 : U00096

①生物種名、②これが最初のリファレンス。
これは原著論文だが、そうでない場合もある

LOCUS U00096 4641652 bp DNA circular BCT 01-AUG-2014

DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.

ACCESSION U00096

VERSION U00096.3

DBLINK BioProject: PRJNA225
BioSample: SAMN02604091

KEYWORDS .

SOURCE Escherichia coli str. K-12 substr. MG1655

ORGANISM Escherichia coli str. K-12 substr. MG1655
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
Enterobacteriaceae; Escherichia.

REFERENCE 1 (bases 1 to 4641652)

AUTHORS Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
Mau,B. and Shao,Y.

TITLE The complete genome sequence of Escherichia coli K-12

JOURNAL Science 277 (5331), 1453-1462 (1997)

PUBMED 9278503

REFERENCE 2 (bases 1 to 4641652)

AUTHORS Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.

TITLE Highly accurate genome sequences of Escherichia coli K-12 strains
MG1655 and W3110

JOURNAL Mol. Syst. Biol. 2, 2006 (2006)

PUBMED 16738553

REFERENCE 3 (bases 1 to 4641652)

AUTHORS Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,

W5-1 : U00096

半ページほど下に移動。①2つめのリファレンス、②3つめ、③4つめ。こんな感じでこの配列に関する原著論文がどんどん追加されていることがわかる。かなり有名な大腸菌株なので、こういうことになる。また、③4つめのリファレンス番号が見えている時点で、U00096.3のバージョン番号と対応しているわけではないこともわかる

<http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/>

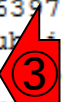
Enterobacteriaceae; Escherichia.

REFERENCE 1 (bases 1 to 4641652)
AUTHORS Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y.
TITLE The complete genome sequence of Escherichia coli K-12
JOURNAL Science 277 (5331), 1453-1462 (1997)
PUBMED 9278503

REFERENCE 2 (bases 1 to 4641652)
AUTHORS Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S., Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
TITLE Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110
JOURNAL Mol. Syst. Biol. 2, 2006 (2006)
PUBMED 1673553

REFERENCE 3 (bases 1 to 4641652)
AUTHORS Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T., Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H., Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
TITLE Escherichia coli K-12: a cooperatively developed annotation snapshot--2005
JOURNAL Nucleic Acids Res. 34 (1), 1-9 (2006)
PUBMED 16397293
REMARK Publication Status: Online-Only

REFERENCE 4 (bases 1 to 4641652)
AUTHORS Arnaud,M., Berlyn,M.K.B., Blattner,F.R., Galperin,M.Y., Glasner,J.D., Horiuchi,T., Kosuge,T., Mori,H., Perna,N.T., Plunkett,G. III, Riley,M., Rudd,K.E., Serres,M.H., Thomas,G.H. and



W5-1 : U00096

①REFERENCE 4の続き。②Unpublishedとなっているので、これが原著論文でないリファレンスの例。(publishされてから更新されていないだけという例もあるとは思いますが)

Snapshot 2003

JOURNAL Nucleic Acids Res. 34 (1), 1-9 (2006)
PUBMED 16397293
REMARK Publication Status: Online-Only

REFERENCE 4 (bases 1 to 4641652)
AUTHORS Arnaud,M., Berlyn,M.K.B., Blattner,F.R., Galperin,M.Y., Glasner,J.D., Horiuchi,T., Kosuge,T., Mori,H., Perna,N.T., Plunkett,G. III, Riley,M., Rudd,K.E., Serres,M.H., Thomas,G.H. and Wanner,B.L.
TITLE Workshop on Annotation of Escherichia coli K-12
JOURNAL Unpublished
REMARK Woods Hole, Mass., on 14-18 November 2003 (sequence corrections)

REFERENCE 5 (bases 1 to 4641652)
AUTHORS Glasner,J.D., Perna,N.T., Plunkett,G. III, Anderson,B.D., Bockhorst,J., Hu,J.C., Riley,M., Rudd,K.E. and Serres,M.H.
TITLE ASAP: Escherichia coli K-12 strain MG1655 version m56
JOURNAL Unpublished
REMARK ASAP download 10 June 2004 (annotation updates)

REFERENCE 6 (bases 1 to 4641652)
AUTHORS Hayashi,K., Morooka,N., Mori,H. and Horiuchi,T.
TITLE A more accurate sequence comparison between genomes of Escherichia coli K12 W3110 and MG1655 strains
JOURNAL Unpublished
REMARK GenBank accessions AG613214 to AG613378 (sequence corrections)

REFERENCE 7 (bases 1 to 4641652)
AUTHORS Perna,N.T.
TITLE Escherichia coli K-12 MG1655 yqiK-rfaE intergenic region, genomic sequence correction
JOURNAL Unpublished
REMARK GenBank accession AY605712 (sequence corrections)

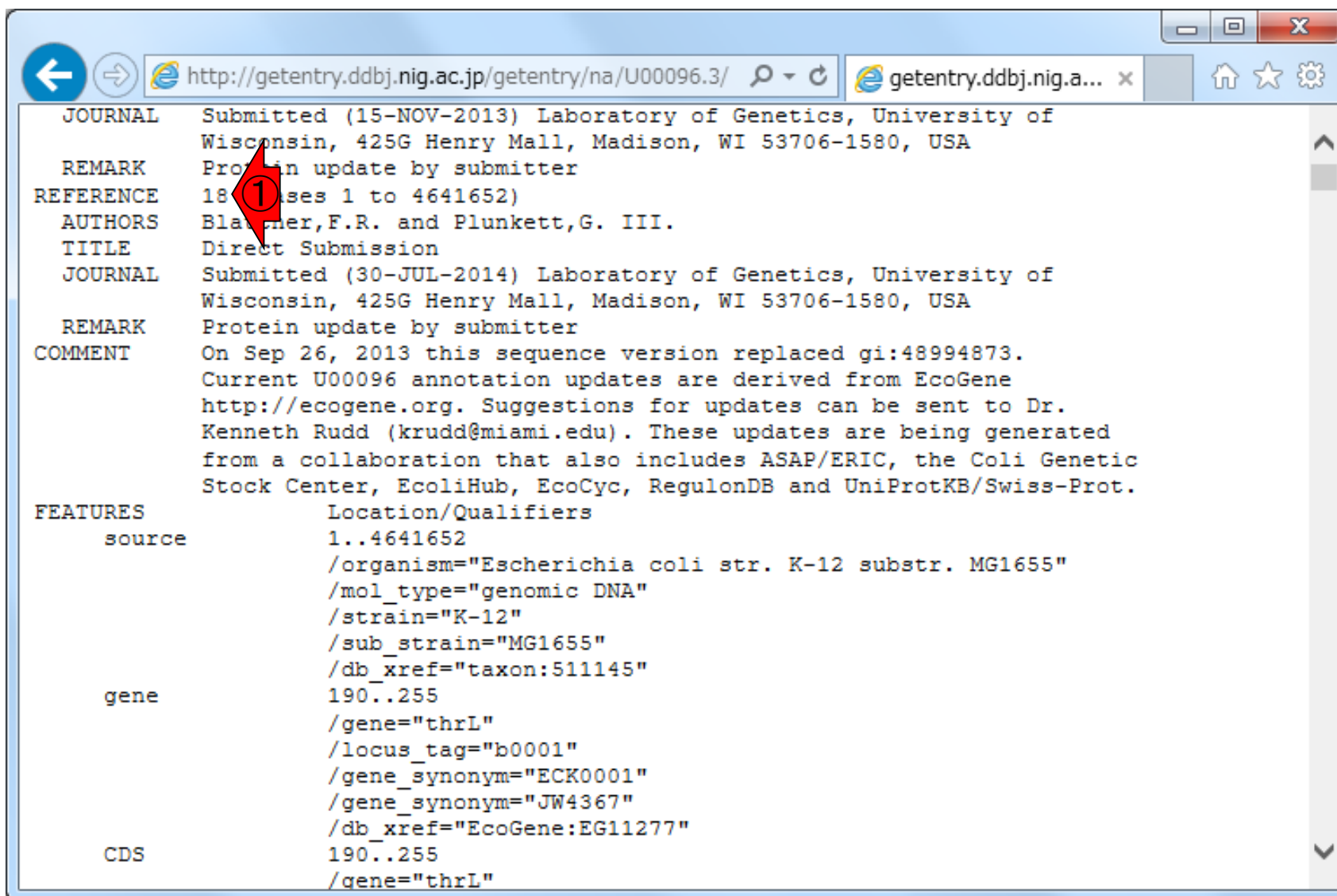
W5-2: U00096

①REFERENCE 9。②のAUTHORSの情報が塩基配列の登録者情報。基本的には、すぐ下のTITLEのところはDirect Submissionとなっており、JOURNALのところもジャーナル名やUnpublishedという記述でもないことを頼りに探せばよい

sequence correction
JOURNAL Unpublished
REMARK GenBank accession AY605712 (sequence corrections)
REFERENCE 8 (bases 1 to 4641652)
AUTHORS Rudd,K.E.
TITLE A manual approach to accurate translation start site annotation: an E. coli K-12 case study
JOURNAL Unpublished
REFERENCE 9 (bases 1 to 4641652)
AUTHORS Blattner,F.R. and Plunkett,G. III.
TITLE Direct Submission
JOURNAL Submitted (16-JAN-1997) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REFERENCE 10 (bases 1 to 4641652)
AUTHORS Blattner,F.R. and Plunkett,G. III.
TITLE Direct Submission
JOURNAL Submitted (02-SEP-1997) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REFERENCE 11 (bases 1 to 4641652)
AUTHORS Plunkett,G. III.
TITLE Direct Submission
JOURNAL Submitted (13-OCT-1998) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REFERENCE 12 (bases 1 to 4641652)
AUTHORS Plunkett,G. III.
TITLE Direct Submission
JOURNAL Submitted (10-JUN-2004) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REMARK Sequence update by submitter

W5-2:U00096

この場合は①REFERENCE 18まであるが、これは相当多いほうです



http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/

JOURNAL Submitted (15-NOV-2013) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA

REMARK Protein update by submitter

REFERENCE 18 (uses 1 to 4641652)

AUTHORS Blattner, F.R. and Plunkett, G. III.

TITLE Direct Submission

JOURNAL Submitted (30-JUL-2014) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA

REMARK Protein update by submitter

COMMENT On Sep 26, 2013 this sequence version replaced gi:48994873. Current U00096 annotation updates are derived from EcoGene <http://ecogene.org>. Suggestions for updates can be sent to Dr. Kenneth Rudd (krudd@miami.edu). These updates are being generated from a collaboration that also includes ASAP/ERIC, the Coli Genetic Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.

FEATURES

Location/Qualifiers

source 1..4641652
/organism="Escherichia coli str. K-12 substr. MG1655"
/mol_type="genomic DNA"
/strain="K-12"
/sub_strain="MG1655"
/db_xref="taxon:511145"

gene 190..255
/gene="thrL"
/locus_tag="b0001"
/gene_synonym="ECK0001"
/gene_synonym="JW4367"
/db_xref="EcoGene:EG11277"

CDS 190..255
/gene="thrL"

W5-2-1: AP014680

なお、比較的最近のエントリ(①AP014680)は、
②REFERENCE 1が登録者情報になっている

LOCUS AP014680 2277985 bp DNA circular BCT 03-APR-2015
DEFINITION Lactobacillus hokkaidonensis DNA, complete genome.
ACCESSION AP014680
VERSION AP014680.1
DBLINK BioProject:PRJDB1726
Sequence Read Archive:DRR024500, DRR024501
BioSample:SAMD00000344
KEYWORDS .
SOURCE Lactobacillus hokkaidonensis
ORGANISM Lactobacillus hokkaidonensis
Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae;
Lactobacillus.
REFERENCE 1 (bases 1 to 2277985)
AUTHORS Tohno,M. and Tanizawa,Y.
TITLE Direct Submission
JOURNAL Submitted (10-NOV-2014) to the DDBJ/EMBL/GenBank databases.
Contact:Masanori Tohno
National Agriculture and Food Research Organization, National
Institute of Livestock and Grassland Science, Animal Feeding and
Management Research Division; Senbonmatsu 768, Nasushiobara,
Tochigi 329-2793, Japan
REFERENCE 2
AUTHORS Tanizawa,Y., Tohno,M., Kaminuma,E., Nakamura,Y. and Arita,M.
TITLE Complete genome sequence and analysis of Lactobacillus
hokkaidonensis LOOC260T, a psychrotrophic lactic acid bacterium
isolated from silage
JOURNAL BMC Genomics 16, 240 (2015)
REMARK Publication Status: Online-Only
DOI:10.1186/s12864-015-1435-2

<http://getentry.ddbj.nig.ac.jp/getentry/na/AP014680>

W5-2-2: REFERENCE 1

The screenshot shows a web browser window displaying the DDBJ (DNA Data Bank of Japan) website. The address bar shows the URL <http://www.ddbj.nig.ac.jp/sub/reference1-j.html>. The page features the DDBJ logo and a search bar with the text "Google™ カスタム検索". A navigation menu includes links for "塩基配列の登録", "プロジェクトの登録", "塩基配列登録の前に", "Flat File の説明", and "お問い合わせ". The breadcrumb trail is "HOME > データ登録 > 塩基配列の登録 > DDBJ のデータ公開形式 (flat file) の説明 > REFERENCE 1". The page is updated as of 2015.11.9. A large black banner with the text "REFERENCE 1" is prominent. Below it, a red warning icon precedes the text "[重要] DDBJフラットファイルフォーマット改訂: E-mailアドレスと電話番号、FAX番号の非表示化(2007/8/1)". A section titled "REFERENCE 1" contains the following text: "データベースに登録した登録者の情報が記載されています。ただし古いデータに関してはこの限りではありません。またデータバンクが独自に構築する CON エントリの場合、登録者情報にあたる REFERENCE 1 は表示されない場合があります。"

<http://www.ddbj.nig.ac.jp/sub/reference1-j.html>

W5-3: FEATURES

http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/

```

JOURNAL Submitted (15-NOV-2013) Laboratory of Genetics, University of
Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REMARK Protein update by submitter
REFERENCE 18 (bases 1 to 4641652)
AUTHORS Blattner,F.R. and Plunkett,G. III.
TITLE Direct Submission
JOURNAL Submitted (30-JUL-2014) Laboratory of Genetics, University of
Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
REMARK Protein update by submitter
COMMENT On Sep 26, 2013 this sequence version replaced gi:48994873.
Current U00096 annotation updates are derived from EcoGene
http://ecogene.org. Suggestions for updates can be sent to Dr.
Kenneth Rudd (krudd@miami.edu). These updates are being generated
from a collaboration that also includes ASAP/ERIC, the Coli Genetic
Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.

FEATURES             Location/Qualifiers
   source             1..4641652
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
   gene               190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /gene_synonym="JW4367"
                     /db_xref="EcoGene:EG11277"
   CDS                190..255
                     /gene="thrL"
    
```

<http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/>

W5-4: source feature

① source featureは、配列全体に対する記述を行う特別なfeatureであり、各entryの最初に1つだけ存在

http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/

JOURNAL Submitted (15-NOV-2013) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA

REMARK Protein update by submitter

REFERENCE 18 (bases 1 to 4641652)

AUTHORS Blattner, F.R. and Plunkett, G. III.

TITLE Direct Submission

JOURNAL Submitted (30-JUL-2014) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA

REMARK Protein update by submitter

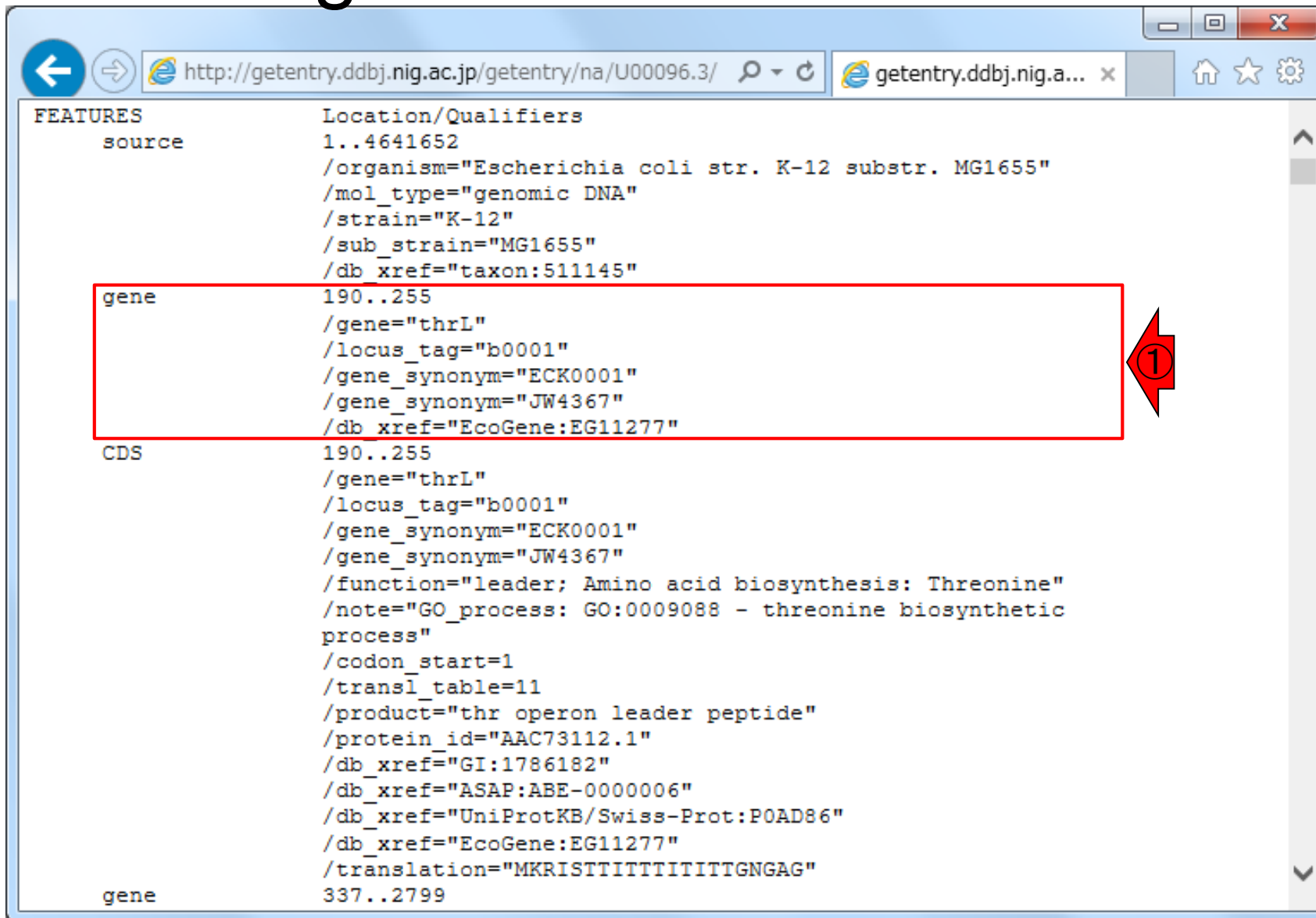
COMMENT On Sep 26, 2013 this sequence version replaced gi:48994873. Current U00096 annotation updates are derived from EcoGene <http://ecogene.org>. Suggestions for updates can be sent to Dr. Kenneth Rudd (krudd@miami.edu). These updates are being generated from a collaboration that also includes ASAP/ERIC, the Coli Genetic Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.

FEATURES Location/Qualifiers

source	1..4641652 /organism="Escherichia coli str. K-12 substr. MG1655" /mol_type="genomic DNA" /strain="K-12" /sub_strain="MG1655" /db_xref="taxon:511145"
gene	190..255 /gene="thrL" /locus_tag="b0001" /gene_synonym="ECK0001" /gene_synonym="JW4367" /db_xref="EcoGene:EG11277"
CDS	190..255 /gene="thrL"

①gene featureは、CDS、rRNA、tRNAなどの遺伝子領域を記載する際に用いられる

W5-5: gene feature



```
FEATURES             Location/Qualifiers
     source            1..4641652
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     gene              190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /gene_synonym="JW4367"
                     /db_xref="EcoGene:EG11277"
     CDS               190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /gene_synonym="JW4367"
                     /function="leader; Amino acid biosynthesis: Threonine"
                     /note="GO_process: GO:0009088 - threonine biosynthetic
                     process"
                     /codon_start=1
                     /transl_table=11
                     /product="thr operon leader peptide"
                     /protein_id="AAC73112.1"
                     /db_xref="GI:1786182"
                     /db_xref="ASAP:ABE-0000006"
                     /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                     /db_xref="EcoGene:EG11277"
                     /translation="MKRISTTITTTITITGNGAG"
     gene              337..2799
```

①CDS featureは、タンパクをコードする領域について記載している

W5-5: CDS feature

The screenshot shows a web browser window displaying the GenBank entry for Escherichia coli str. K-12 substr. MG1655. The URL is <http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/>. The page content is as follows:

```
FEATURES             Location/Qualifiers
     source            1..4641652
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
     gene              190..255
                     /db_xref="taxon:511145"
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /gene_synonym="JW4367"
                     /db_xref="EcoGene:EG11277"
     CDS                190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /gene_synonym="JW4367"
                     /function="leader; Amino acid biosynthesis: Threonine"
                     /note="GO_process: GO:0009088 - threonine biosynthetic
                     process"
                     /codon_start=1
                     /transl_table=11
                     /product="thr operon leader peptide"
                     /protein_id="AAC73112.1"
                     /db_xref="GI:1786182"
                     /db_xref="ASAP:ABE-0000006"
                     /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                     /db_xref="EcoGene:EG11277"
                     /translation="MKRISTTITTTITITGNGAG"
     gene              337..2799
```

A red box highlights the CDS feature, and a red arrow with the number 1 points to it.

W5-6: qualifier

featureの内容をさらに細かく記述したものがqualifier。①gene qualifier、②product qualifier、③translation qualifierなど

```
FEATURES             Location/Qualifiers
     source            1..4641652
                        /organism="Escherichia coli str. K-12 substr. MG1655"
                        /mol_type="genomic DNA"
                        /strain="K-12"
                        /sub_strain="MG1655"
     gene              190..255
                        /db_xref="taxon:511145"
                        /gene="thrL"
                        /locus_tag="b0001"
                        /gene_synonym="ECK0001"
                        /gene_synonym="JW4367"
                        /db_xref="EcoGene:EG11277"
     CDS               190..255
                        ① /gene="thrL"
                        /locus_tag="b0001"
                        /gene_synonym="ECK0001"
                        /gene_synonym="JW4367"
                        /function="leader; Amino acid biosynthesis: Threonine"
                        /note="GO_process: GO:0009088 - threonine biosynthetic
                        process"
                        /codon_start=1
                        /transl_table=11
                        ② /product="thr operon leader peptide"
                        /protein_id="AAC73112.1"
                        /db_xref="GI:1786182"
                        /db_xref="ASAP:ABE-0000006"
                        /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                        /db_xref="EcoGene:EG11277"
                        ③ /translation="MKRISTTITTTITITGNGAG"
     gene              337..2799
```


W5-7: qualifier

featureの内容をさらに細かく記述したものがqualifier。①experiment qualifierによって実験的に裏付けがなされていることが示されているものもある

検索: experiment 前へ 次へ オプション 100 より多い一致があります

```
/gene_synonym="JW0001"  
/gene_synonym="thrA1"  
/gene_synonym="thrA2"  
/gene_synonym="thrD"  
/db_xref="EcoGene:EG10998"  
CDS 337..2799  
/gene="thrA"  
/locus_tag="b0002"  
/gene_synonym="ECK0002"  
/gene_synonym="Hs"  
/gene_synonym="JW0001"  
/gene_synonym="thrA1"  
/gene_synonym="thrA2"  
/gene_synonym="thrD"  
/EC_number="1.1.1.3"  
/EC_number="2.7.2.4"  
/function="enzyme; Amino acid biosynthesis: Threonine"  
① /experiment="N-terminus verified by Edman degradation:  
PMID 354697,4562989"  
/note="bifunctional: aspartokinase I (N-terminal);  
homoserine dehydrogenase I (C-terminal);  
GO_component: GO:0005737 - cytoplasm;  
GO_process: GO:0009088 - threonine biosynthetic process;  
GO_process: GO:0009086 - methionine biosynthetic process;  
GO_process: GO:0009090 - homoserine biosynthetic process"  
/codon_start=1  
/transl_table=11  
/product="Bifunctional aspartokinase/homoserine  
dehydrogenase 1"  
/protein_id="AAC73113.1"
```

W5-8: featureの種類

検索: repeat_region 前へ 次へ オプション 100 より多い一致があります

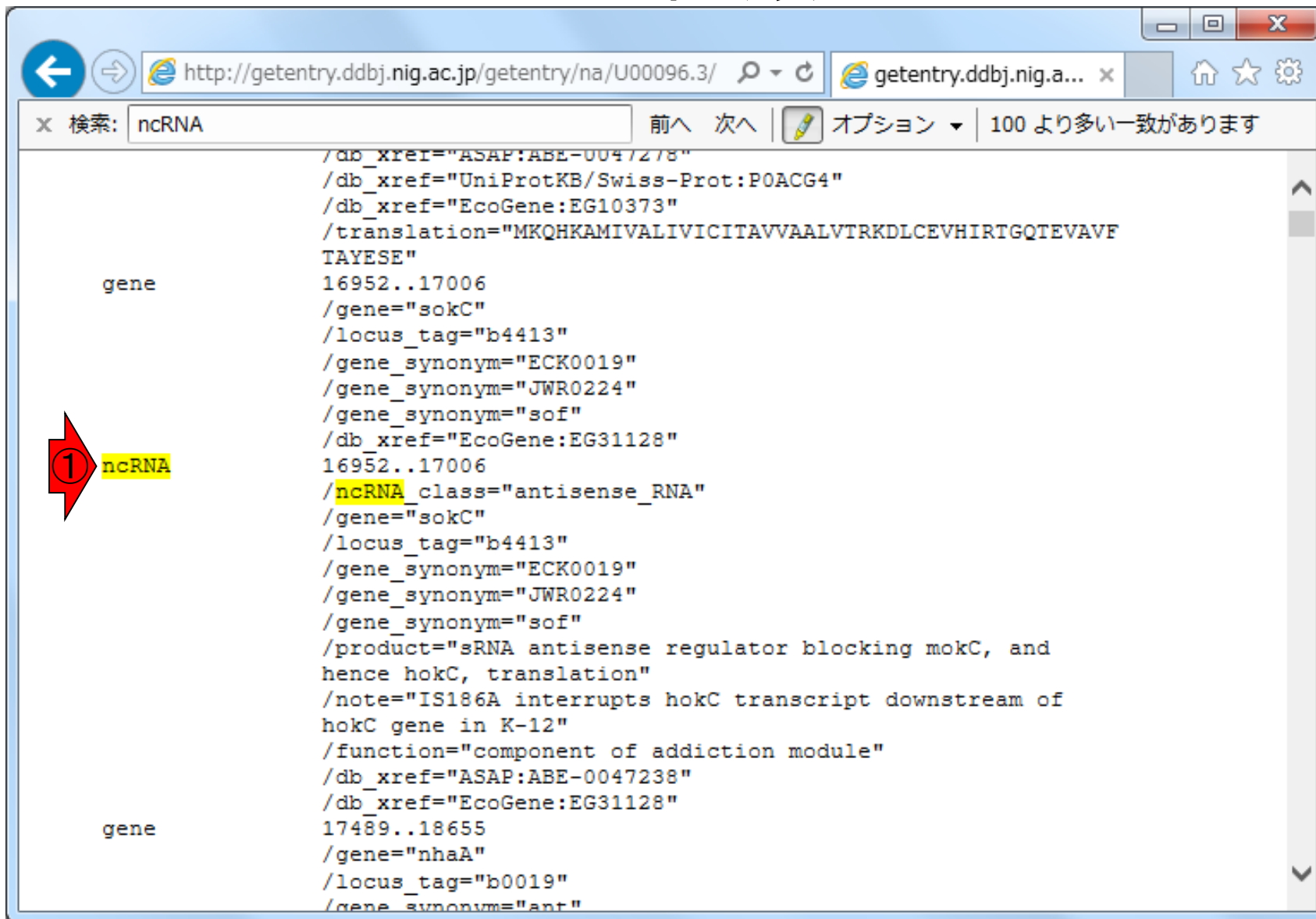
```

CDS
    /locus_tag="b0005"
    /gene_synonym="ECK0005"
    /gene_synonym="JW0004"
    /db_xref="EcoGene:EG14384"
    5234..5530
    /gene="yaaX"
    /locus_tag="b0005"
    /gene_synonym="ECK0005"
    /gene_synonym="JW0004"
    /codon_start=1
    /transl_table=11
    /product="DUF2502 family putative periplasmic protein"
    /protein_id="AAC73116.1"
    /db_xref="GI:1786186"
    /db_xref="ASAP:ABE-0000015"
    /db_xref="UniProtKB/Swiss-Prot:P75616"
    /db_xref="EcoGene:EG14384"
    /translation="MKKMQSIVLALSIVLVAPMAAQAAEITLVPSVKLQIGDRDRNGY
    YWDGGHWRDHGWWKQHYEWRGNRWHLHGPPPPPRHHKKAPHDHHGGHGPGKHHR"
    5565..5669
    /note="RIP1 (repetitive extragenic palindromic) element;
    contains 2 REP sequences and 1 IHF site"
gene
    complement(5683..6459)
    /gene="yaaA"
    /locus_tag="b0006"
    /gene_synonym="ECK0006"
    /gene_synonym="JW0005"
    /db_xref="EcoGene:EG10011"
CDS
    complement(5683..6459)

```

①ncRNA featureなどがあります。他には tRNA featureやrRNA featureが存在します

W5-9: featureの種類

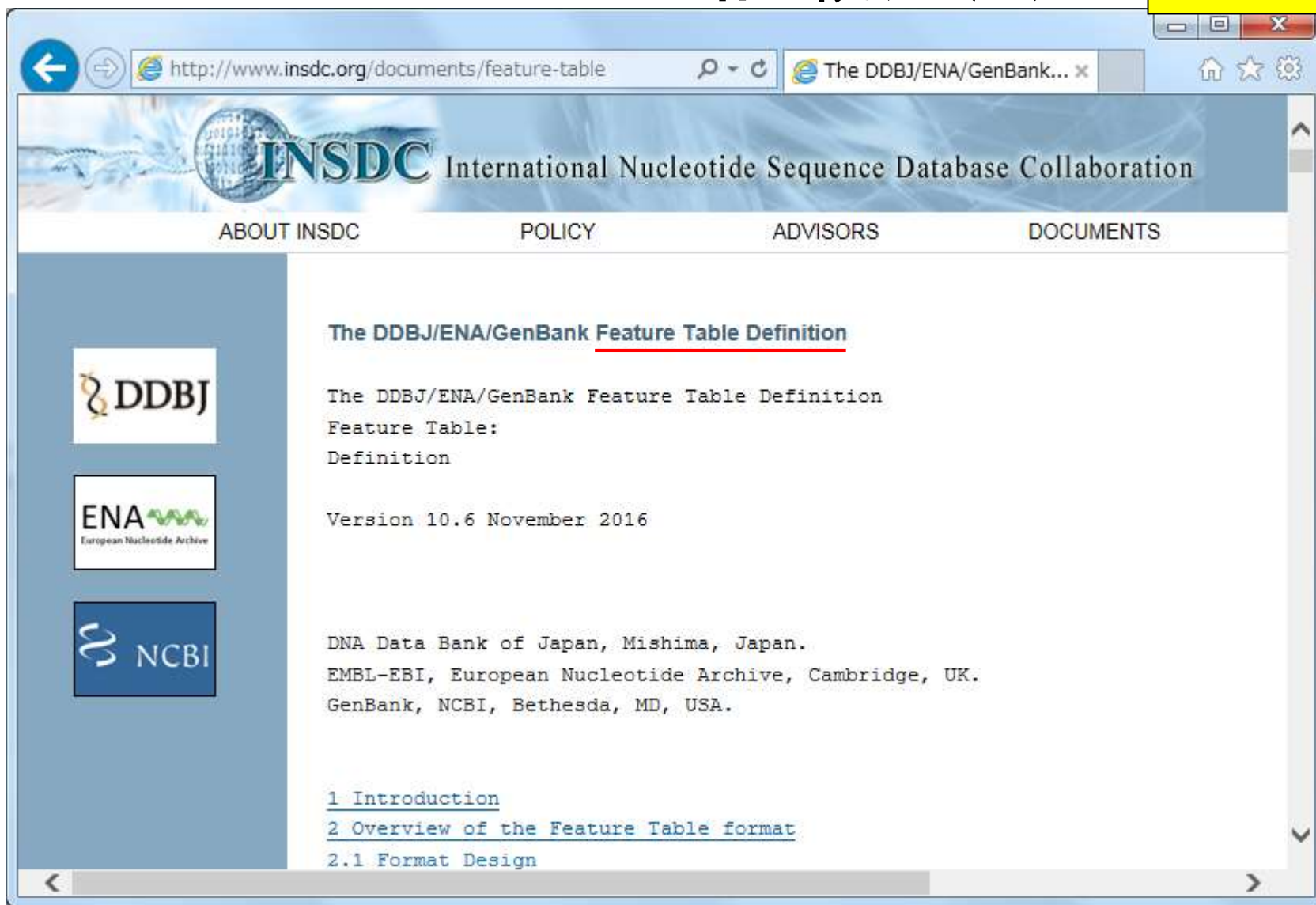


検索: ncRNA 前へ 次へ オプション 100 より多い一致があります

```
/db_xref="ASAP:ABE-0047278"  
/db_xref="UniProtKB/Swiss-Prot:P0ACG4"  
/db_xref="EcoGene:EG10373"  
/translation="MKQHKAMIVALIVICITAVVAALVTRKDLCEVHIRTGQTEVAVF  
TAYESE"  
gene 16952..17006  
/gene="sokC"  
/locus_tag="b4413"  
/gene_synonym="ECK0019"  
/gene_synonym="JWR0224"  
/gene_synonym="sof"  
/db_xref="EcoGene:EG31128"  
16952..17006  
/ncRNA_class="antisense_RNA"  
/gene="sokC"  
/locus_tag="b4413"  
/gene_synonym="ECK0019"  
/gene_synonym="JWR0224"  
/gene_synonym="sof"  
/product="sRNA antisense regulator blocking mokC, and  
hence hokC, translation"  
/note="IS186A interrupts hokC transcript downstream of  
hokC gene in K-12"  
/function="component of addiction module"  
/db_xref="ASAP:ABE-0047238"  
/db_xref="EcoGene:EG31128"  
gene 17489..18655  
/gene="nhaA"  
/locus_tag="b0019"  
/gene_synonym="ant"
```

W5-10: featureの記載方法

各featureの記載方法は、
INSDCのFeature Table
Definitionで定められている



<http://www.insdc.org/documents/feature-table>

W6-2: locus_tag

①U00096(W5-3と同じものです)上で見られるlocus_tag情報は昔の記述法であり、現在のものとは異なるので注意

検索: locus_tag 前へ 次へ オプション 100 より多い一致があります

JOURNAL Submitted (30-JUL-2014) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA

REMARK Protein update by submitter

COMMENT On Sep 26, 2013 this sequence version replaced gi:48994873. Current U00096 annotation updates are derived from EcoGene <http://ecogene.org>. Suggestions for updates can be sent to Dr. Kenneth Rudd (krudd@miami.edu). These updates are being generated from a collaboration that also includes ASAP/ERIC, the Coli Genetic Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.

FEATURES

	Location/Qualifiers
source	1..4641652 /organism="Escherichia coli str. K-12 substr. MG1655" /mol_type="genomic DNA" /strain="K-12" /sub_strain="MG1655" /db_xref="taxon:511145"
gene	190..255 /gene="thrL" ① /locus_tag="b0001" /gene_synonym="ECK0001" /gene_synonym="JW4367" /db_xref="EcoGene:EG11277"
CDS	190..255 ① /gene="thrL" /locus_tag="b0001" /gene_synonym="ECK0001" /gene_synonym="JW4367" /function="leader: Amino acid biosynthesis: Threonine"

<http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3/>

W6-3: AP014680

①accession番号AP014680を眺める。これはGenBankのエントリ

The screenshot shows the NCBI GenBank entry for *Lactobacillus hokkaidonensis* DNA, complete genome (AP014680.1). The browser address bar shows the URL <https://www.ncbi.nlm.nih.gov/nucleotide/AP014680.1>. The page title is "Lactobacillus hokkaidonensis DNA, complete genome". The accession number "AP014680.1" is highlighted with a red arrow and a circled "1". The page includes a search bar, navigation links, and a detailed description of the genome.

GenBank: AP014680.1

[FASTA](#) [Graphics](#)

Go to:

LOCUS AP014680 2277985 bp DNA circular BCT 03-APR-2015

DEFINITION *Lactobacillus hokkaidonensis* DNA, complete genome.

ACCESSION AP014680

VERSION AP014680.1

DBLINK BioProject: [PRJDB1726](#)
BioSample: [SAMD00000344](#)
Sequence Read Archive: [DRR024500](#), [DRR024501](#)

KEYWORDS .

SOURCE *Lactobacillus hokkaidonensis* JCM 18461
ORGANISM [Lactobacillus hokkaidonensis](#) JCM 18461
Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae;
Lactobacillus.

REFERENCE 1
AUTHORS Tanizawa, Y., Tohno, M., Kaminuma, E., Nakamura, Y. and Arita, M.

Change region shown

Customize view

Basic Features

Default features

Gene, RNA, and CDS features only

Display options

Show sequence

Show reverse complement

Update View

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

W6-3: AP014680

2014年11月10日に登録されたAP014680を眺めると、確かに現在のlocus_tag表記法(prefix_tag)どおり、① LOOC260_100010となっている。ここで、LOOC260はこの乳酸菌ゲノム配列上で統一して使われているlocus_tagのprefixに相当し、100010がtagに相当する

JOURNAL Submitted (10-NOV-2014) Contact: Masanori Tohno National Agriculture and Food Research Organization, National Institute of Livestock and Grassland Science, Animal Feeding and Management Research Division; Senbonmatsu 768, Nasushiobara, Tochigi 329-2793, Japan

COMMENT Annotated by GenomeRefine.
##Genome-Assembly-Data-START##
Assembly Method :: HGAP SMRTAnalysis v. 2.1
Genome Coverage :: 300X
Sequencing Technology :: PacBio RSII
##Genome-Assembly-Data-END##

FEATURES Location/Qualifiers
source 1..2277985
/organism="Lactobacillus hokkaidonensis JCM 18461"
/mol_type="genomic DNA"
/strain="LOOC260"
/db_xref="taxon:1291742"
/note="type strain of Lactobacillus hokkaidonensis"

gene 347..1676
/gene="dnaA"
/locus_tag="LOOC260_100010"

regulatory 347..352
/regulatory_class="ribosome_binding_site"
/gene="dnaA"
/locus_tag="LOOC260_100010"
/note="identified by MetaGeneAnnotator; putative"

CDS 360..1676
/gene="dnaA"
/locus_tag="LOOC260_100010"
/note="Bacterial dnaA protein helix-turn-helix; pfam08299; Bacterial dnaA protein; pfam00308; DnaA N-terminal domain; pfam11638;"

Genome
Identical RefSeq
Protein
PubMed
PubMed (Weighted)
Taxonomy

LinkOut to external resources

Ribosomal Database Project II [Ribosomal Database Project II]

SILVA LSU Database [SILVA]

SILVA SSU Database [SILVA]

Order PURB cDNA clone/Protein/Antibody/RNAi [OriGene]

Order MURC cDNA clone/Protein/Antibody/RNAi [OriGene]

Order MVK cDNA clone/Protein/Antibody/RNAi [OriGene]

W6-4: AP014680

赤枠が2つめのfeature。tagは、通常ゲノム中での出現順に通し番号を使用することが多いが、②ここでは後から追加・挿入されることを想定して10飛びの値を使っている(これもアリ)

https://www.ncbi.nlm.nih.gov/nucleotide/AP014680.1

```

/locus_tag="LOOC260_100010"
/note="identified by MetaGeneAnnotator; putative"
360..1676
/gene="dnaA"
/locus_tag="LOOC260_100010"
/note="Bacterial dnaA protein helix-turn-helix; pfam08299;
Bacterial dnaA protein; pfam00308;
DnaA N-terminal domain; pfam11638;
chromosomal replication initiation protein DnaA
[Lactobacillus plantarum WCFS1];
chromosomal replication initiator protein DnaA; TIGR00362;
gene_1;
identified by MetaGeneAnnotator; putative"
/codon_start=1
/transl_table=11
/product="chromosomal replication initiation protein DnaA"
/protein_id="BAP84581.1"

```

gene 1837..2991
/gene="dnaN"
/locus_tag="LOOC260_100020"
regulatory 1837..1842
/regulatory_class="ribosome_binding_site"
/gene="dnaN"
/locus_tag="LOOC260_100020"
CDS 1852..2991
/gene="dnaN"
/locus_tag="LOOC260_100020"
/note="Beta clamp domain. The beta subunit (processivity factor) of DNA polymerase III holoenzyme, referred to as the beta clamp, forms a ring shaped dimer that encircles

Order MURC cDNA clone/Protein/Antibody/RNAi [OriGene]
Order MVK cDNA clone/Protein/Antibody/RNAi [OriGene]

Recent activity
Turn Off Clear

- Lactobacillus hokkaidonensis DNA, complete genome Nucleotide
- 9278503[uid] (1) PubMed
- The complete genome sequence of Escherichia coli K-12. PubMed
- GenVisR: Genomic Visualizations in R. PubMed
- Cited In for PubMed (Select 22009675) (123) PubMed

See more...

W6-5: locus_tag公式情報

The screenshot shows a web browser window displaying the DDBJ (DNA Data Bank of Japan) website. The address bar shows the URL http://www.ddbj.nig.ac.jp/sub/locus_tag-j.html. The page title is "/locus_tag qualifier の記...". The DDBJ logo is visible in the top left, and there is a search bar with the text "Google™ カスタム検索" and a "Search" button. A navigation menu is present with links for "塩基配列の登録", "プロジェクトの登録", "塩基配列登録の前に", "Flat File の説明", and "お問い合わせ". The breadcrumb trail is "HOME > データ登録 > 塩基配列の登録 > Qualifier key の定義 > /locus_tag qualifier の記載法". The page content is titled "/locus_tag qualifier の記載法" and includes a section for "背景" (Background) with the following text:

/locus_tag qualifier は、2003年に導入されました。その導入当初は、ゲノムプロジェクトが その配列データを更新する際に feature 継承をするための追跡用 ID として自由度の高い記載を可能としていました。

しかし、The American Society for Microbiology からの要請を受けて、2005年の国際実務者会議において、/locus_tag qualifier の用法が再検討されました。その結果、/locus_tag qualifier を恒久的に一意な ID として維持していくことを目指して、配列データの登録時に当該ゲノム専用の prefix を割り当てることにより、/locus_tag を記載するように規則が変更されました。

DDBJ においては、ゲノム配列データ登録用に Mass Submission System (MSS) をご用意しております。MSS 申し込みフォームにおいて、適宜、指定していただければ、/locus_tag prefix 割り当てを検討いたします。

W6-6: gene feature

and Food Research Organization, National Institute of Livestock and Grassland Science, Animal Feeding and Management Research Division; Senbonmatsu 768, Nasushiobara, Tochigi 329-2793, Japan

COMMENT Annotated by GenomeRefine.

```
##Genome-Assembly-Data-START##
Assembly Method      :: HGAP SMRTAnalysis v. 2.1
Genome Coverage      :: 300X
Sequencing Technology :: PacBio RSII
##Genome-Assembly-Data-END##
```

FEATURES

Location/Qualifiers	Location/Qualifiers
source	1..2277985 /organism="Lactobacillus hokkaidonensis JCM 18461" /mol_type="genomic DNA" /strain="LOOC260" /db_xref="taxon:1291742" /note="type strain of Lactobacillus hokkaidonensis"
gene	347..1676 /gene="dnaA" /locus_tag="LOOC260_100010"
regulatory	347..352 /regulatory_class="ribosome_binding_site" /gene="dnaA" /locus_tag="LOOC260_100010" /note="identified by MetaGeneAnnotator; putative"
CDS	360..1676 /gene="dnaA" /locus_tag="LOOC260_100010" /note="Bacterial dnaA protein helix-turn-helix; pfam08299; Bacterial dnaA protein; pfam00308; DnaA N-terminal domain; pfam11638; chromosomal replication initiation protein DnaA"

Genome
Identical RefSeq
Protein
PubMed
PubMed (Weighted)
Taxonomy

LinkOut to external resources

- Ribosomal Database Project II [Ribosomal Database Project II]
- SILVA LSU Database [SILVA]
- SILVA SSU Database [SILVA]
- Order PURB cDNA clone/Protein/Antibody/RNAi [OriGene]
- Order MURC cDNA clone/Protein/Antibody/RNAi [OriGene]
- Order MVK cDNA clone/Protein/Antibody/RNAi [OriGene]

Recent activity

W6-6: gene feature

DDBJ形式では、①gene featureは存在しない。これは形式の違いの一例。DDBJで最初に登録された場合は、GenBankのほうでgene featureを自動的に付加する

Associated from [Linkage](#)

JOURNAL BMC Genomics 16, 240 (2015)
REMARK Publication Status: Online-Only
DOI:10.1186/s12864-015-1435-2
COMMENT Annotated by GenomeRefine

```
##Genome-Assembly-Data-START##  
Assembly Method      :: HGAP SMRTAnalysis v. 2.1  
Genome Coverage      :: 300X  
Sequencing Technology :: PacBio RSII  
##Genome-Assembly-Data-END##
```

FEATURES

Location/Qualifiers
source 1..2277985 /db_xref="taxon:1193095" /mol_type="genomic DNA" /note="type strain of Lactobacillus hokkaidonensis" /organism="Lactobacillus hokkaidonensis" /strain="LOOC260"
① regulatory 347..352 /gene="dnaA" /locus_tag="LOOC260_100010" /note="identified by MetaGeneAnnotator; putative" /regulatory_class="ribosome_binding_site"
CDS 360..1676 /codon_start=1 /gene="dnaA" /locus_tag="LOOC260_100010" /note="Bacterial dnaA protein helix-turn-helix; pfam08299" /note="Bacterial dnaA protein; pfam00308" /note="DnaA N-terminal domain; pfam11638" /note="chromosomal replication initiation protein DnaA [Lactobacillus plantarum WCFS1]" /note="chromosomal replication initiator protein DnaA; TIGR00362" /note="gene_1" /note="identified by MetaGeneAnnotator; putative"

W6-6: gene feature

```

DR      SILVA-LSU; AP014680.
DR      SILVA-SSU; AP014680.
DR      PubMed; 25879859.
XX
CC      Annotated by GenomeRefine
CC      ##Genome-Assembly-Data-START##
CC      Assembly Method      :: HGAP SMRTAnalysis v. 2.1
CC      Genome Coverage      :: 300X
CC      Sequencing Technology :: PacBio RSII
CC      ##Genome-Assembly-Data-END##
XX
FH      Key          Location/Qualifiers
FH
FT      source          1..2277985
FT                      /organism="Lactobacillus hokkaidonensis"
FT                      /strain="LOOC260"
FT                      /mol_type="genomic DNA"
FT                      /note="type strain of Lactobacillus hokkaidonensis"
FT                      /db_xref="taxon:1193095"
E ① FT      regulatory    347..352
FT                      /gene="dnaA"
FT                      /locus_tag="LOOC260_100010"
FT                      /note="identified by MetaGeneAnnotator; putative"
FT                      /regulatory_class="ribosome_binding_site"
FT      CDS             360..1676
FT                      /codon_start=1
FT                      /transl_table=11
FT                      /gene="dnaA"
FT                      /locus_tag="LOOC260_100010"
FT                      /product="chromosomal replication initiation protein DnaA"
FT                      /note="Bacterial dnaA protein helix-turn-helix; pfam08299"
FT                      /note="Bacterial dnaA protein; pfam00308"
FT                      /note="DnaA N-terminal domain; pfam11638"
FT                      /note="chromosomal replication initiation protein DnaA
FT                      [Lactobacillus plantarum WCFS1]"

```

W7-1: DFAST

DFASTトップ画面。2016年12月にバージョンアップ。
乳酸菌ファイル(LH_complete.fa)のアノテーションだと
わかっているので、①Lactic Acid Bacteriaをクリック

DFAST Analysis Archive Download Help

DFAST and DAGA are integrated genome annotation tools and resources. DFAST is an annotation platform for bacterial genomes. Its core annotation process is based on PROKKA and curated reference database tailored to specific organisms. It also generates DDBJ-compliant submission files for Mass Submission System (MSS) at DDBJ. DAGA is a genome archive that stores bacterial genomes obtained from DDBJ/ENA/GenBank and Sequence Read Archive (SRA). All the genomes deposited in DAGA are consistently annotated using DFAST. This website provides genome resources for *Lactic Acid Bacteria*.

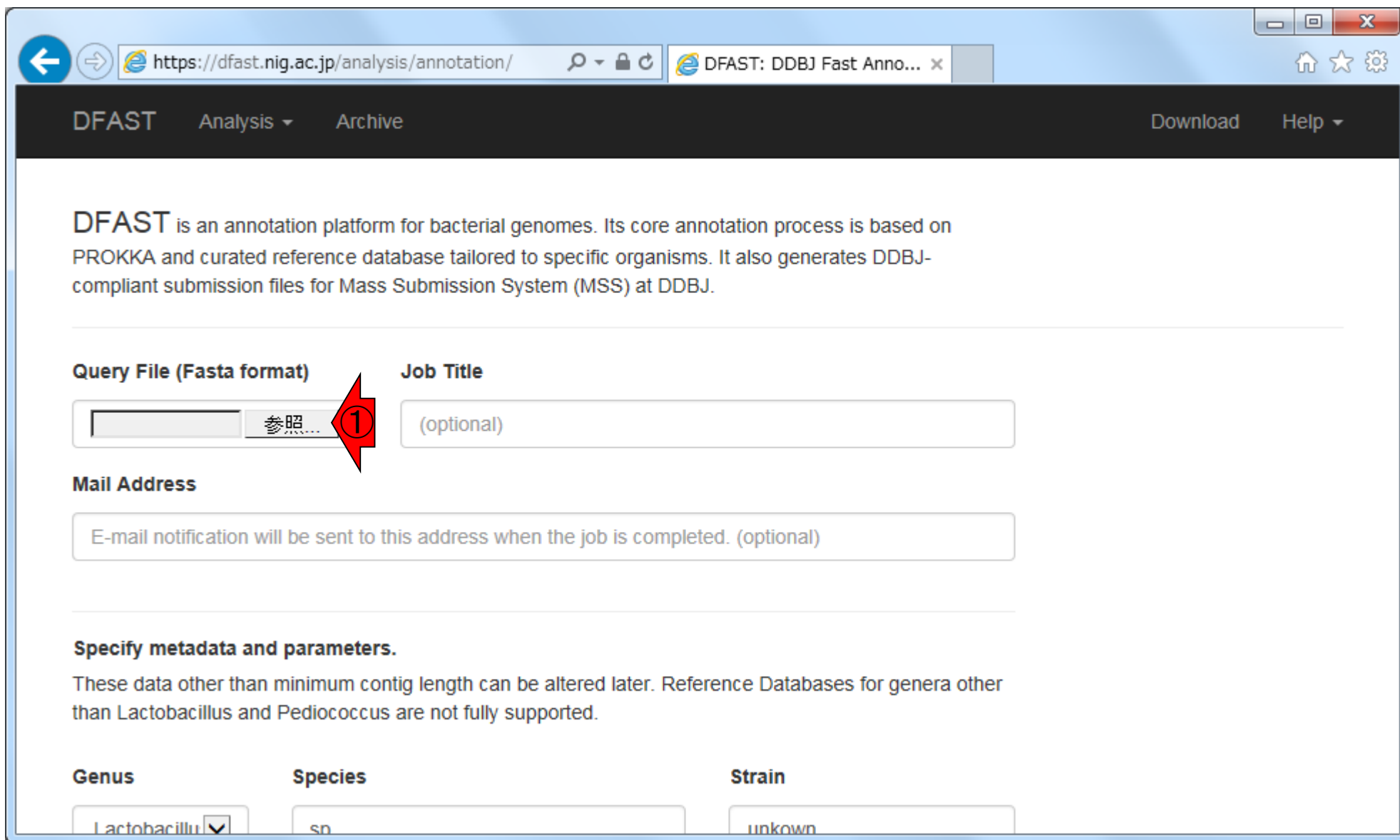
DFAST → DDBJ Fast Annotation and Submission Tool
Upload your Genome, Annotate, and Submit to DDBJ.

Select reference database to use.

Lactic Acid Bacteria ①

β-versions

W7-1: DFAST



DFAST is an annotation platform for bacterial genomes. Its core annotation process is based on PROKKA and curated reference database tailored to specific organisms. It also generates DDBJ-compliant submission files for Mass Submission System (MSS) at DDBJ.

Query File (Fasta format) **Job Title**

参照... ①

Mail Address

Specify metadata and parameters.

These data other than minimum contig length can be altered later. Reference Databases for genera other than Lactobacillus and Pediococcus are not fully supported.

Genus **Species** **Strain**

Lactobacillu...

W7-1: DFAST

①共有フォルダ上にある、②
LH_complete.faを選んで、③開く

The screenshot shows a web browser window with the URL `https://dfast.nig.ac.jp/analysis/annotation/`. The page title is "DFAST: DDBJ Fast Anno...". The browser window is overlaid with a file selection dialog box titled "アップロードするファイルの選択". The dialog box shows the path `C:\Users\kadota\Desktop\share` and a list of files. The file `LH_complete.fa` is selected. The "開く(O)" button is highlighted with a red arrow.

名前	更新日時	種類	サイズ
.RData	2016/12/20 19:55	R Workspace	686 KB
.Rhistory	2016/12/20 19:55	RHISTORY ファ...	2 KB
LH_complete2.fa	2016/12/20 19:31	FA ファイル	2,382 KB
hoge.fa	2016/12/20 19:19	FA ファイル	2,225 KB
LH_complete.fa	2016/12/19 19:33	FA ファイル	2,345 KB
finish_assembly.py	2016/12/19 19:30	PY ファイル	1 KB
seq1_draft.fa	2016/12/16 13:56	FA ファイル	2,225 KB
sequence1_blast.txt	2016/09/25 21:38	テキストドキュ...	12,182 KB
LH_draft2.fa	2016/09/23 20:25	FA ファイル	2,345 KB
out2.bam.bai.zip	2016/09/13 20:03	ZIP ファイル	4 KB
out2.bam.zip	2016/09/13 20:02	ZIP ファイル	69,651 KB

W7-1: DFAST

①Job Titleには任意の名称がつけられ、②Mail address欄にアドレスを入力しておけばジョブ完了時に通知メールを受け取ることができる

The screenshot shows the DFAST web interface. The browser address bar displays <https://dfast.nig.ac.jp/analysis/annotation/>. The page title is "DFAST: DDBJ Fast Anno...". The navigation menu includes "DFAST", "Analysis", "Archive", "Download", and "Help".

The main content area contains the following sections:

- DFAST is an annotation platform for bacterial genomes.** Its core annotation process is based on PROKKA and curated reference database tailored to specific organisms. It also generates DDBJ-compliant submission files for Mass Submission System (MSS) at DDBJ.
- Query File (Fasta format)**: A text input field containing "C:\Users\kadota" and a "参照..." button.
- Job Title**: A text input field with "(optional)" placeholder text. A red lightning bolt with the number "1" points to this field.
- Mail Address**: A text input field with "E-mail notification will be sent to this address when the job is completed. (optional)" placeholder text. A red lightning bolt with the number "2" points to this field.
- Specify metadata and parameters.** These data other than minimum contig length can be altered later. Reference Databases for genera other than Lactobacillus and Pediococcus are not fully supported.
- Genus**: A dropdown menu with "Lactobacillu" selected.
- Species**: A text input field with "sp" entered.
- Strain**: A text input field with "unkown" entered.

W7-1: DFAST

①ページ下部に移動。②Minimum Contig Lengthは設定値未満の短い断片配列を除くためのもの(デフォルトは200)。それ以外の設定項目は、アノテーション結果に影響を与えることはない。後から変更可能

E-mail notification will be sent to this address when the job is completed. (optional)

Specify metadata and parameters.
These data other than minimum contig length can be altered later. Reference Databases for genera other than Lactobacillus and Pediococcus are not fully supported.

Genus	Species	Strain
Lactobacillu <input type="button" value="v"/>	sp. ex) plantarum, delbrueckii subsp. bulgaricus	unkown

Locus Tag Prefix **Minimum Contig Length**

LOCUS 200

Perform Genome Assessment (optional)

W7-2: CheckMとANI

E-mail notification will be sent to this address when the job is completed. (optional)

Specify metadata and parameters.
These data other than minimum contig length can be altered later. Reference Databases for genera other than Lactobacillus and Pediococcus are not fully supported.

Genus Lactobacillus **Species** sp. **Strain** unkown
ex) plantarum, delbrueckii subsp. bulgaricus

Locus Tag Prefix LOCUS **Minimum Contig Length** 200

① Show Perform Genome Assessment (optional)

Run

W7-2: CheckMとANI

E-mail notification will be sent to this address when the job is completed. (optional)

Specify metadata and parameters.
These data other than minimum contig length can be altered later. Reference Databases for genera other than Lactobacillus and Pediococcus are not fully supported.

Genus Lactobacillu **Species** sp. **Strain** unkown
ex) plantarum, delbrueckii subsp. bulgaricus

Locus Tag Prefix LOCUS **Minimum Contig Length** 200

Perform Genome Assessment (optional)

Perform CheckM Calculation for Genome Quality Assessment

Select a Taxonomic Group for CheckM Calculation. (Currently, for Lactobacillus and Pediococcus)

Rank	Genus
------	-------

W7-2: CheckMとANI

①CheckMや②ANIの計算はオプションとなっているので、実行したい場合にはチェックをいれる必要がある

Locus Tag Prefix: LOCUS

Minimum Contig Length: 200

▲ Hide Perform Genome Assessment (optional)

Perform CheckM Calculation for Genome Quality Assessment

Select a Taxonomic Group for CheckM Calculation. (Currently, for Lactobacillus and Pediococcus)

Rank: Genus

Genus: Lactobacillus

Perform ANI Calculation for Taxonomic Assessment

Select Target Groups for ANI Calculation.

— not specified (calculate ANI against all representative genomes) —

Run

W7-3: CheckM

①CheckMを実行すべく、チェックを入れる。CheckM実行時にはどの系統群(taxonomic group)のマーカセットを使うか指定する必要があるが、通常は②Genus欄で選択した結果(この場合Lactobacillus)に基づいて自動で選択されるものを用いればよい

The screenshot shows the CheckM web interface. At the top, there are two input fields: "Locus Tag Prefix" with the value "LOCUS" and "Minimum Contig Length" with the value "200". Below these is a green button labeled "▲ Hide" and the text "Perform Genome Assessment (optional)". A red arrow with the number "1" points to a checked checkbox labeled "Perform CheckM Calculation for Genome Quality Assessment". Below this is the instruction "Select a Taxonomic Group for CheckM Calculation. (Currently, for Lactobacillus and Pediococcus)". There are two dropdown menus: "Rank" with "Genus" selected and "Genus" with "Lactobacillus" selected. A red arrow with the number "2" points to the "Lactobacillus" dropdown. Below these is an unchecked checkbox labeled "Perform ANI Calculation for Taxonomic Assessment" and the instruction "Select Target Groups for ANI Calculation.". A text input field contains the text "— not specified (calculate ANI against all representative genomes) —". At the bottom left is a "Run" button.

W7-4: ANI

①ANIの計算は、②何も指定しなければ全185菌種との間で計算を行う
(当然それだけの時間はかかる)

Locus Tag Prefix: LOCUS

Minimum Contig Length: 200

▲ Hide Perform Genome Assessment (optional)

Perform CheckM Calculation for Genome Quality Assessment

Select a Taxonomic Group for CheckM Calculation. (Currently, for Lactobacillus and Pediococcus)

Rank: Genus

Genus: Lactobacillus

Perform ANI Calculation for Taxonomic Assessment

Select Target Groups for ANI Calculation.

--- not specified (calculate ANI against all representative genomes) ---

Run

W7-4: ANI

(①チェックを入れると以下で系統群を指定可能になる)②計算時間を短縮したい場合には対象とする系統群を指定可能にしておくとい

Locus Tag Prefix: LOCUS

Minimum Contig Length: 200

▲ Hide Perform Genome Assessment (optional)

Perform CheckM Calculation for Genome Quality Assessment

Select a Taxonomic Group for CheckM Calculation. (Currently, for Lactobacillus and Pediococcus)

Rank: Genus

Genus: Lactobacillus

Perform ANI Calculation for Taxonomic Assessment

Select Target Groups for ANI Calculation.

Pediococcus

algidus

alimentarius

W7-5: 実行

https://dfast.nig.ac.jp/analysis/annotation/ DFAST: DDBJ Fast Anno... x

Locus Tag Prefix: LOCUS

Minimum Contig Length: 200

▲ Hide Perform Genome Assessment (optional)

Perform CheckM Calculation for Genome Quality Assessment

Select a Taxonomic Group for CheckM Calculation. (Currently, for Lactobacillus and Pediococcus)

Rank: Genus

Genus: Lactobacillus

Perform ANI Calculation for Taxonomic Assessment

Select Target Groups for ANI Calculation.

--- not specified (calculate ANI against all representative genomes) ---

Run

W7-5: 実行

The screenshot shows a web browser window with the following elements:

- Browser Address Bar:** <http://dfast.nig.ac.jp/analysis/annotation/cbbab016>
- Page Title:** DFAST - Job Result
- Navigation:** DFAST | Analysis | Archive | Download | Help
- Message:** Your annotation job has been submitted.
- Warning:** Remember the current URL to access this page. The result will be deleted 30 days after your last visit.
- Job Details:**
 - Title :
 - JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
 - Status : RUNNING
- Log:**
 - [2017-01-01 13:43:07.450512] Job submitted.
 - [2017-01-01 13:43:07.473965] Job started.
- Buttons:** Update Status, Delete (with warning: This procedure cannot be undone.)
- Navigation Tabs:** Result (selected), Features, DDBJ Submission, Log

W7-5: 実行中...

実行結果画面になるまでは①のような状態が続きます。PCを途中でオフすると実行結果画面に辿り着けなくなるので注意！実行終了をメールで知らせてもらうやり方がオススメだが、実行中でもURLをメモするなり、お気に入り追加しておけば実行中でもブラウザを閉じてOK

DFAST Analysis Archive Download Help

Your annotation job has been submitted.

Remember the current URL to access this page. The result will be deleted 30 days after your last visit. Delete this job now. => Delete This procedure cannot be undone.

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : RUNNING

Update Status Please wait patiently. Currently, 1 Jobs running / 0 Jobs waiting.

[2017-01-01 13:43:07.450512] Job submitted.
[2017-01-01 13:43:07.473965] Job started.

Result Features DDBJ Submission Log

CheckMとANIを計算したので若干時間がかかったが、それでも①約8分で計算終了

W7-6: 実行結果

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/cbbab016>. The page title is "DFAST - Job Result". The navigation bar includes "DFAST", "Analysis", "Archive", "Download", and "Help".

A warning message states: "Remember the current URL to access this page. The result will be deleted 30 days after your last visit." Below this is a "Delete this job now. => Delete" button with the note "This procedure cannot be undone."

Job details:

- Title :
- JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
- Status : COMPLETE

A log entry box is highlighted with a red border and a red arrow labeled "1":

```
[2017-01-01 13:43:07.450512] Job submitted.  
[2017-01-01 13:43:07.473965] Job started.  
[2017-01-01 13:50:54.541834] Job completed.
```

Navigation tabs include "Result" (selected), "Features", "DDBJ Submission", and "Log".

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%

Download Files

- Genbank Flat File : [annotation.gbk](#)
- GFF3-formated File : [annotation.gff](#)
- Genome Fasta File : [genome.fna](#)
- Protein Fasta File : [protein.faa](#)
- CDS Fasta File : [cds.fna](#)
- RNA Fasta File : [ma.fna](#)
- Feature Table : [features.tsv](#)
- Genome Statistics : [statistics.txt](#)

W7-6: 実行結果

①ジョブ完了の通知メール設定をしていない場合は、②このページのURLを覚えておかないと、結果画面へ再度アクセスすることができなくなる(ページを閉じるとアウト)。③通知メールの有無に関わらず、最後にアクセスした日から30日が経過すると削除されるので注意。④今すぐ消したければDelete

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/cbbab016>. The page title is "DFAST Analysis Archive". The main content area displays job details for JobID: cbbab016-21eb-4060-9cee-60b12d073d4a, which is in a "COMPLETE" status. A warning message states: "Remember the current URL to access this page. The result will be deleted 30 days after your last visit." Below this, there is a "Delete this job now. => Delete" button. A log of events is shown: [2017-01-01 13:43:07.450512] Job submitted, [2017-01-01 13:43:07.473965] Job started, and [2017-01-01 13:50:54.541834] Job completed. The page also features a "Genome Statistics" table and a "Download Files" section with various file formats for download.

① Remember the **current URL** to access this page. The result will be deleted 30 days after your last visit.

②

③ Delete this job now. => This procedure cannot be undone.

④

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : COMPLETE

[2017-01-01 13:43:07.450512] Job submitted.
[2017-01-01 13:43:07.473965] Job started.
[2017-01-01 13:50:54.541834] Job completed.

Result Features DDBJ Submission Log

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%

Download Files

- Genbank Flat File : [annotation.gbk](#)
- GFF3-formated File : [annotation.gff](#)
- Genome Fasta File : [genome.fna](#)
- Protein Fasta File : [protein.faa](#)
- CDS Fasta File : [cds.fna](#)
- RNA Fasta File : [ma.fna](#)
- Feature Table : [features.tsv](#)
- Genome Statistics : [statistics.txt](#)

①ページ下部に移動。②ゲノムの総塩基数や、③予測された遺伝子数などの基本的な統計値が示されている

W7-7: 基本統計量

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,331
No. of rRNA	12
No. of tRNA	56
No. of CRISPRS	1
Coding Ratio (%)	86.9%

You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

Download Files

- Genbank Flat File : [annotation.gbk](#)
- GFF3-formated File : [annotation.gff](#)
- Genome Fasta File : [genome.fna](#)
- Protein Fasta File : [protein.faa](#)
- CDS Fasta File : [cds.fna](#)
- RNA Fasta File : [rna.fna](#)
- Feature Table : [features.tsv](#)
- Genome Statistics : [statistics.txt](#)
- Zip Archive : [annotation.zip](#)

Genome Assessment

ANI Result : [Download](#) CheckM Result : [Download](#)

ANI TopHit	Lactobacillus hokkaidonensis LOOC260
ANI %	100.00%
Completeness %	99.30%
Contamination %	2.50%

W7-7: 基本統計量

経験的には、バクテリアゲノムでは①CDSの個数はゲノムサイズを1,000で割った値よりやや少ない程度(ゲノムサイズ2Mbpなら1,900程度)、②CDS領域の長さの和をゲノムサイズで割ったコーディング率(Coding ratio)は80%台後半の値を示すことが多い。NGS機器の種類によっては挿入・欠失のエラーが多くコドンの読み枠がずれてしまうためにCDSが極端に多くなったりコーディング率が下がったりすることもあるので注意

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,331
No. of rRNA	12
No. of tRNA	56
No. of CRISPRS	1
Coding Ratio (%)	86.9%



GFF3-formated File :	annotation.gff
Genome Fasta File :	genome.fna
Protein Fasta File :	protein.faa
CDS Fasta File :	cds.fna
RNA Fasta File :	rna.fna
Feature Table :	features.tsv
Genome Statistics :	statistics.txt
Zip Archive :	annotation.zip

Genome Assessment

ANI Result : [Download](#) CheckM Result : [Download](#)

ANI TopHit	Lactobacillus hokkaidonensis LOOC260
ANI %	100.00%
Completeness %	99.30%
Contamination %	2.50%

You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

W7-7: 基本統計量

①ドラフトゲノムではrRNAの予測ができないことがある。rRNAは通常はゲノム中に散在的に複数コピー存在する。このような散在反復領域はショートリードを使った *de novo* assembly で再現するのは難しく断片的な配列しか得られないことが多いため

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,331
No. of rRNA	12
No. of tRNA	56
No. of CRISPRS	1
Coding Ratio (%)	86.9%



Download Files

Genbank Flat File :	annotation.gbk
GFF3-formated File :	annotation.gff
Genome Fasta File :	genome.fna
Protein Fasta File :	protein.faa
CDS Fasta File :	cds.fna
RNA Fasta File :	rna.fna
Feature Table :	features.tsv
Genome Statistics :	statistics.txt
Zip Archive :	annotation.zip

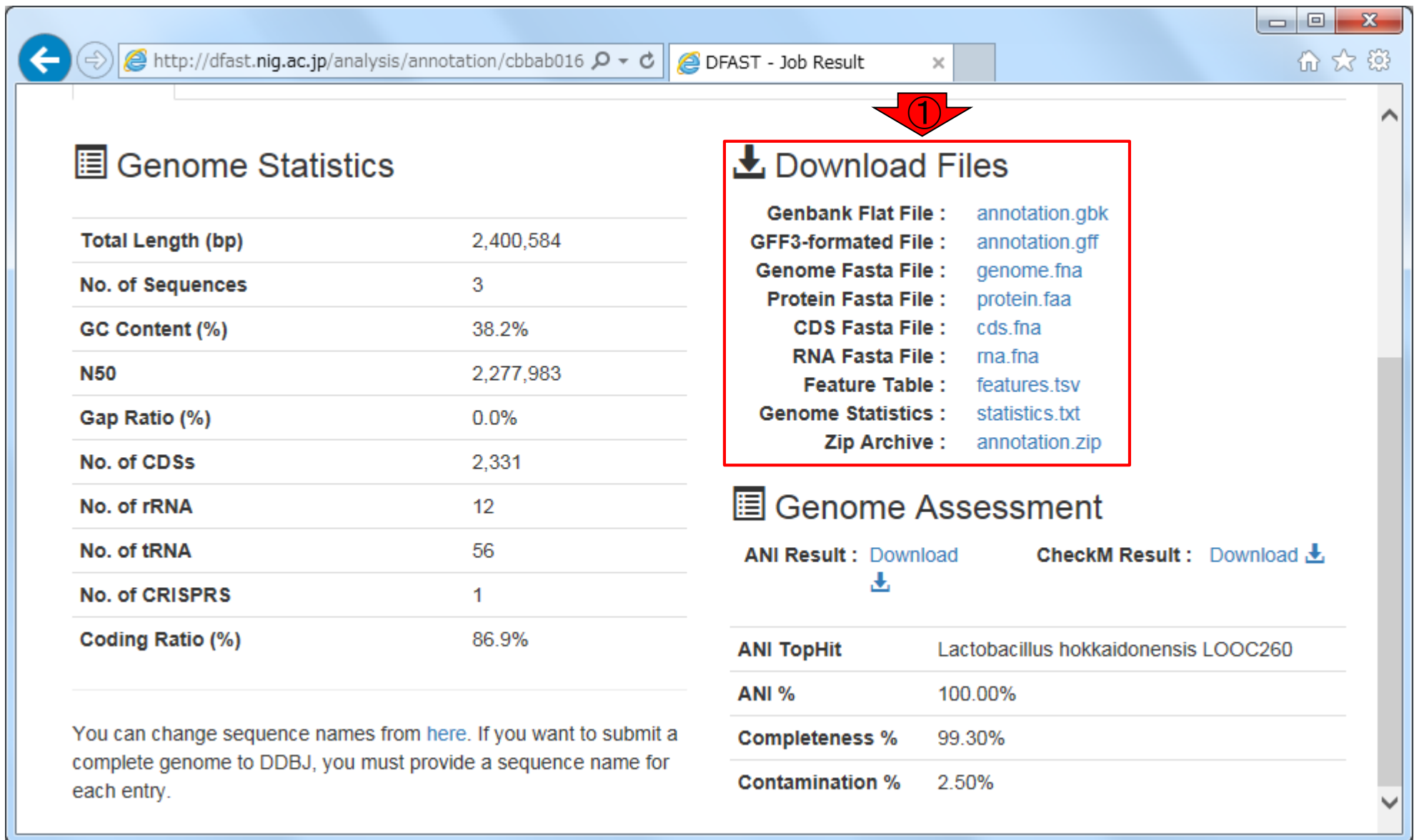
Genome Assessment

ANI Result : [Download](#) CheckM Result : [Download](#)

ANI TopHit	Lactobacillus hokkaidonensis LOOC260
ANI %	100.00%
Completeness %	99.30%
Contamination %	2.50%

You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

W7-8: ダウンロード



http://dfast.nig.ac.jp/analysis/annotation/cbbab016 DFAST - Job Result

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,331
No. of rRNA	12
No. of tRNA	56
No. of CRISPRS	1
Coding Ratio (%)	86.9%

You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

Download Files

- Genbank Flat File : [annotation.gbk](#)
- GFF3-formated File : [annotation.gff](#)
- Genome Fasta File : [genome.fna](#)
- Protein Fasta File : [protein.faa](#)
- CDS Fasta File : [cds.fna](#)
- RNA Fasta File : [rna.fna](#)
- Feature Table : [features.tsv](#)
- Genome Statistics : [statistics.txt](#)
- Zip Archive : [annotation.zip](#)

Genome Assessment

ANI Result : [Download](#) CheckM Result : [Download](#) ↓

ANI TopHit	Lactobacillus hokkaidonensis LOOC260
ANI %	100.00%
Completeness %	99.30%
Contamination %	2.50%

(オプションでCheckMとANIを実行した場合は)
①Genome Assessmentの結果が見られる

W7-9: CheckMとANI

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,331
No. of rRNA	12
No. of tRNA	56
No. of CRISPRS	1
Coding Ratio (%)	86.9%


You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

Download Files

Genbank Flat File :	annotation.gbk
GFF3-formated File :	annotation.gff
Genome Fasta File :	genome.fna
Protein Fasta File :	protein.faa
CDS Fasta File :	cds.fna
RNA Fasta File :	rna.fna
Feature Table :	features.tsv
Genome Statistics :	statistics.txt
Zip Archive :	annotation.zip

①

Genome Assessment

ANI Result : [Download](#) CheckM Result : [Download](#) 

ANI TopHit	Lactobacillus hokkaidonensis LOOC260
ANI %	100.00%
Completeness %	99.30%
Contamination %	2.50%

W7-10: ANI

①ANI計算結果は100.00%だった。基準株を中心に選定した185菌種(亜種も含む)のrepresentative genomesを対象として、菌種ごとに入力ファイル(LH_complete.fa)とのANIを計算し、最もANI値の高かったものを表示している。100.00%という完全一致の結果となった理由は、比較した185菌種の中に、入力と全く同じ*L. hokkaidonensis* LOOC260株が含まれていたため

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,331
No. of rRNA	12
No. of tRNA	56
No. of CRISPRS	1
Coding Ratio (%)	86.9%

You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

Download Files

Genbank Flat File :	annotation.gbk
GFF3-formated File :	annotation.gff
Genome Fasta File :	genome.fna
Protein Fasta File :	protein.faa
CDS Fasta File :	cds.fna
RNA Fasta File :	rna.fna
Feature Table :	features.tsv
Genome Statistics :	statistics.txt
Zip Archive :	annotation.zip

Genome Assessment

ANI Result : [Download](#)



CheckM Result : [Download](#)

ANI TopHit	Lactobacillus hokkaidonensis LOOC260
ANI %	100.00%
Completeness %	99.30%
Contamination %	2.50%

①

W7-11 : CheckM

①CheckMの結果。Completeness (99.30%)とContamination (2.50%)ともに、CheckMの基準(それぞれ>95%および<5%)を満たしている。我々の感覚としては、ほとんどのゲノムで1%未満のcontaminationである一方、比較的大きなプラスミドを持つ菌は高めのcontamination(2~3%程度)を示す。これはまさにその代表例だが、その感覚についての根拠となる原著論文を知らないので参考程度


Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%
No. of CDSs	2,331
No. of rRNA	12
No. of tRNA	56
No. of CRISPRS	1
Coding Ratio (%)	86.9%

You can change sequence names from [here](#). If you want to submit a complete genome to DDBJ, you must provide a sequence name for each entry.

Genbank Flat File :	annotation.gbk
GFF3-formated File :	annotation.gff
Genome Fasta File :	genome.fna
Protein Fasta File :	protein.faa
CDS Fasta File :	cds.fna
RNA Fasta File :	rna.fna
Feature Table :	features.tsv
Genome Statistics :	statistics.txt
Zip Archive :	annotation.zip

Genome Assessment

ANI Result : [Download](#) CheckM Result : [Download](#) 

ANI TopHit *Lactobacillus hokkaidonensis* LOOC260

ANI % 100.00%

Completeness % 99.30%

Contamination % 2.50%



W8-1 : Features

①ページ上部に再び移動し、②Featuresタブをクリック

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/cbbab016>. The page title is "DFAST - Job Result". The navigation bar includes "DFAST", "Analysis", "Archive", "Download", and "Help". A warning message states: "Remember the current URL to access this page. The result will be deleted 30 days after your last visit." Below this, a "Delete this job now. => Delete" button is present, with a note "This procedure cannot be undone." The job details are as follows:

Title :	
JobID :	cbbab016-21eb-4060-9cee-60b12d073d4a
Status :	COMPLETE

A log box contains the following entries:

```
[2017-01-01 13:43:07.450512] Job submitted.  
[2017-01-01 13:43:07.473965] Job started.  
[2017-01-01 13:50:54.541834] Job completed.
```

At the bottom, there are four tabs: "Result", "Features", "DDBJ Submission", and "Log". A red arrow labeled "2" points to the "Features" tab. On the right side, a red arrow labeled "1" points to the top of the page.

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%

Download Files

Genbank Flat File :	annotation.gbk
GFF3-formated File :	annotation.gff
Genome Fasta File :	genome.fna
Protein Fasta File :	protein.faa
CDS Fasta File :	cds.fna
RNA Fasta File :	ma.fna
Feature Table :	features.tsv
Genome Statistics :	statistics.txt

W8-2: dnaA

1番最初のfeatureは、①dnaAであり、②その開始点は101番目になっている。そうなるようなファイルを作成したのだから、そうなるあたり前(W4)

DFAST Analysis Archive Download Help

Remember the [current URL](http://dfast.nig.ac.jp/analysis/annotation/cbbab016) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => [Delete](#) This procedure cannot be undone.

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : COMPLETE

```
[2017-01-01 13:43:07.450512] Job submitted.  
[2017-01-01 13:43:07.473965] Job started.  
[2017-01-01 13:50:54.541834] Job completed.
```

Result Features DDBJ Submission Log

Annotated Features

Show 25 entries Search:

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
1	LOCUS_00001	chromosome	101..1417	CDS	chromosomal replication initiator protein DnaA	dnaA	View	View	Edit
2	LOCUS_00002	chromosome	1593..2732	CDS	DNA polymerase III	dnaN	View	View	Edit

W8-3: 配列情報取得

Viewボタンを押すと、その遺伝子の①塩基配列や②アミノ酸配列情報を得ることができる

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/cbbab016>. The page title is "DFAST - Job Result". The navigation bar includes "DFAST", "Analysis", "Archive", "Download", and "Help".

Remember the **current URL** to access this page. The result will be deleted 30 days after your last visit.
Delete this job now. => [Delete](#) This procedure cannot be undone.

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : COMPLETE

[2017-01-01 13:43:07.450512] Job submitted.
[2017-01-01 13:43:07.473965] Job started.
[2017-01-01 13:50:54.541834] Job completed.

Result Features DDBJ Submission Log

Annotated Features

Show entries Search:

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
1	LOCUS_00001	chromosome	101..1417	CDS	chromosomal replication initiator protein DnaA	dnaA	View ①	View ②	Edit
2	LOCUS_00002	chromosome	1593..2732	CDS	DNA polymerase III	dnaN	View	View	Edit

W8-4: Nucleotide

- ①塩基配列を見ているところ、
- ②ちょっとページ下部に移動

Remember the [current URL](http://dfast.nig.ac.jp/analysis/annotation/cbbab016) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : COMPLETE

[2017-01-01 13:43:07.450512] Job submitted.
[2017-01-01 13:43:07.473965] Job started.
[2017-01-01 13:50:54.541834] Job completed.

Result Features DDBJ Submission Log

Annotated Features

Show entries Search:

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
1	LOCUS_00001	chromosome	101..1417	CDS	chromosomal replication initiator	dnaA	<input type="button" value="View"/>	<input type="button" value="View"/>	<input type="button" value="Edit"/>
2	LOCUS_00002	chromosome	1593..2732	CDS	GTGACTGATTTAGAAACACTTTGGGACACAATTAAGAATCTTTAAGA				

W8-4: Nucleotide

①この塩基配列を入力として、NCBI BLASTを実行することができる

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/cbbab016>. The page displays a table of genomic loci and a text area containing a nucleotide sequence. A red arrow points to a 'BLAST' button in the text area.

Index	LOCUS	Chromosome	Coordinates	Type
5	LOCUS_00005	chromosome	4329..6272	CDS
6	LOCUS_00006	chromosome	6300..8861	CDS
7	LOCUS_00007	chromosome	9052..10416	CDS
8	LOCUS_00008	chromosome	10610..10906	CDS
9	LOCUS_00009	chromosome	10944..11474	CDS
10	LOCUS_00010	chromosome	11499..11735	CDS
11	LOCUS_00011	chromosome	complement (11971..12528)	CDS
12	LOCUS_00012	chromosome	12648..13400	CDS
13	LOCUS_00013	chromosome	13621..15642	CDS
14	LOCUS_00014	chromosome	15649..16101	CDS

```
AAGGTTAATACAGCGTTGAACCCACGTTACACATTTGACACATTTGTG
ATTGGCAAGGGCAACCAAATGGCCCATGCGGCTGCTTTAGTGGTTTC
TGAGGAACCCGGAAAAATGTATAATCCACTCTTCTTTACGGTGGTGT
TGGGTTAGGAAAGACCCATTTGATGCATGCAATTGGTAACAAAATGAT
GGCAAGCAATGCACAAACGCGCGTCAAGTACGTGACTAGTGAAGCTT
TCACCAACGACTTTATTAATGCCATTCAAACGGTACTCAAGAACAATT
TCGCCAAGAATACCGCAAACCTCGACTTATTATTAGTTGATGATATTCAA
TTTTTTGCCGATAAGGAAGGAACCCAAGAAGAATTCTTCCATACTTTCA
ACGCATTGTATGAAGAAAACAAGCAAATTGACTAACTTCCGATCGTTT
GCCAAATGAGATTCTAAGTTGCAGGATCGCCTTGTTTCTCGTTTTAAA
TGGGGATTGTCAGTTGATATTACACCACCCGATCTGGAACTAGAAATC
GCCATTTTACGCAGTAAAGCCAAAGCGGATCAAATCGACATTCCAAAC
GACACCTTAACTTATATTGCGGGTCAAATTGATTCCAATGTCCGTGAA
CTTGAGGGAGCATTATCACGAGTACAAGCGTACTCACGCCTTAATAAG
GCTACTATTAATACCAGTTTAGCGTCGGATGCCTTAAAAGGCTTAAAG
TTAGCTGGTACTAGTGCATTAACGATTCCGCTAATTCAAGAAAAAGTG
GCCAGCTATTATCATATTTCAAACTGATTTAAAAGGAAAAAACGGG
TTAAACAAATCGTCGTTCTAGACAAATTGCGATGTAAGTTAGTTCGTGA
ACTAACTGATAATTCATTACCTAAAATTGGAGCAGAGTTTGGTGGTAAG
GACCATACAACCTGTTTTACACTCGATTGATAAAATTGAAGAAATGCTCC
AATCCGATCATAATTTGCAAGAAGATGTCCAAACCTTAAAATGGAGCT
AAAGCCATAG
```

Blast this sequence at NCBI.

W8-5: Edit

① ページ上部に移動。② Editボタンを押すと、アノテーションされた遺伝子名、遺伝子シンボルを編集することができる

DFAST Analysis Archive Download Help

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.
Delete this job now. => This procedure cannot be undone.

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : COMPLETE

[2017-01-01 13:43:07.450512] Job submitted.
[2017-01-01 13:43:07.473965] Job started.
[2017-01-01 13:50:54.541834] Job completed.

Result Features DDBJ Submission Log

Annotated Features

Show entries Search:

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
1	LOCUS_00001	chromosome	101..1417	CDS	chromosomal replication initiator	dnaA	<input type="button" value="View"/>	<input type="button" value="View"/>	<input type="button" value="Edit"/>
2	LOCUS_00002	chromosome	1593..2732	CDS	GTGACTGATTTAGAAACACTTTGGGACACAATTAAGAATCTTTAAGA				

W8-5: Edit

①ページ上部に移動。②Editボタンを押すと、アノテーションされた遺伝子名、遺伝子シンボルを編集することができる。③Cancelしておく

The screenshot shows the DFAST web interface with the 'Edit a Feature' dialog box open. The dialog box contains the following information:

- Feature 1 : LOCUS_00001**
- Product:** chromosomal replication initiator protein DnaA
- Gene:** dnaA
- Note:** (empty text area)
- Highlight this Feature
- Search recommended names using [TogoAnnotator](#). Search
- Buttons: Save, Cancel

Background interface details:

- Browser: http://dfast.nig.ac.jp/analysis/annotation/cbbab016
- Page Title: DFAST - Job Result
- Navigation: DFAST, Analysis, Archive, Download, Help
- Table: Annotated Feature table with columns No., Locus Tag, and buttons for Translation and Edit.
- Context Menu: A menu is open over the table, showing 'Nucleotide' and a sequence: GTGACTGATTTAGAAACACTTTGGGACACAATTAAGAATCTTTAAGA.

W8-6: Tips

①Viewボタンを押して出現する配列のポップアップは、もう一度Viewボタンを押せば消えます

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => [Delete](#) This procedure cannot be undone.

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : COMPLETE

```
[2017-01-01 13:43:07.450512] Job submitted.  
[2017-01-01 13:43:07.473965] Job started.  
[2017-01-01 13:50:54.541834] Job completed.
```

Result Features DDBJ Submission Log

Annotated Features

Show entries Search:

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
1	LOCUS_00001	chromosome	101..1417	CDS	chromosomal replication initiator	dnaA	View	View	Edit
2	LOCUS_00002	chromosome	1593..2732	CDS	GTGACTGATTTAGAAACACTTTGGGACACAATTAAGAATCTTTAAGA				

W8-6: Tips

Remember the [current URL](http://dfast.nig.ac.jp/analysis/annotation/cbbab016) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
Status : COMPLETE

```
[2017-01-01 13:43:07.450512] Job submitted.
[2017-01-01 13:43:07.473965] Job started.
[2017-01-01 13:50:54.541834] Job completed.
```

Result **Features** DDBJ Submission Log

Annotated Features

Show entries Search:

No.	Locus Tag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
1	LOCUS_00001	chromosome	101..1417	CDS	chromosomal replication initiator protein DnaA	dnaA	<input type="button" value="View"/>	<input type="button" value="View"/>	<input type="button" value="Edit"/>
2	LOCUS_00002	chromosome	1593..2732	CDS	DNA polymerase III	dnaN	<input type="button" value="View"/>	<input type="button" value="View"/>	<input type="button" value="Edit"/>

W9-1 : DDBJ Submission

①DDBJ Submissionタブ、②こちらのページ。諸事情によりJobIDなどが変わっているが気にしない

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit. [Delete this job now.](#) => [Delete](#) This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.  
[2017-01-20 16:17:56.667496] Job started.  
[2017-01-20 16:25:41.084897] Job completed.
```

[Result](#) [Features](#) **[DDBJ Submission](#)** [Log](#)

1. Preparation for Submit.

You can create DDBJ Submission Files (sequence file and annotation file) required to submit the genome through DDBJ Mass Submission System (MSS). If you want to submit a complete genome, you must provide a sequence name for each entry at [this page](#).

Before submission, you need to register BioProject and BioSample. Please follow the instruction below. For detailed information, please refer to [DDBJ Handbooks](#). If necessary, raw sequence data should be deposited in SRA.

このページでは DDBJ Mass Submission System (MSS) を用いて塩基配列を登録するために必要な2種類のファイル (配列ファイルとアノテーションファイル) を作成できます。コンプリートゲノムを DDBJ に登録する場合には [こちらのページ](#) で配列名・配列種別 (染色体/プラスミド)・直鎖/環状の指定を行ってください。登録に先立ち、BioProject Database と BioSample Database への登録を次の手順に従って行います。詳細な手順は [DDBJ Handbooks](#) を参照してください。必要に応じてシーケンスデータの SRA への登録も行います。

W9-1 : DDBJ Submission

「こちらのページ」のリンク先。①ページ下部まで移動

The screenshot shows a web browser window with the URL `http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce`. The page title is "DFAST - Change Sequen...". The navigation bar includes "DFAST", "Analysis", "Archive", "Download", and "Help".

Remember the **current URL** to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.  
[2017-01-20 16:17:56.667496] Job started.  
[2017-01-20 16:25:41.084897] Job completed.
```

Result Features **DDBJ Submission** Log

Change Sequence Names

To submit a complete genome, you need to specify a sequence name, a type (chromosome or plasmid), and topology for each sequence.

Number of Sequences: 3

Sequence Status: draft complete

Sequence Prefix

W9-1 : DDBJ Submission

Browser address bar: <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce>

DFAST - Change Sequen... x

Delete this job now. => Delete This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.  
[2017-01-20 16:17:56.667496] Job started.  
[2017-01-20 16:25:41.084897] Job completed.
```

Result Features **DDBJ Submission** Log

Change Sequence Names

To submit a complete genome, you need to specify a sequence name, a type (chromosome or plasmid), and topology for each sequence.

Number of Sequences: 3

Sequence Status: draft complete

Sequence Prefix

Use only alphabetic characters.

Entry names will be **sequence1-sequence3**.

①

W9-2: 配列名の変更

①確かに入力は3配列だった。ドラフト配列のときは、(必須ではないが)②配列の接頭辞(Sequence Prefix)を③sequence##からcontig##のように変更することもできる。④我々の入力はcomplete配列

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8...>. The page title is "Change Sequence Names".

Job information:

- Title :
- JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
- Status : COMPLETE

Log messages:

```
[2017-01-20 16:17:56.643410] Job submitted.  
[2017-01-20 16:17:56.667496] Job started.  
[2017-01-20 16:25:41.084897] Job completed.
```

Navigation tabs: Result, Features, DDBJ Submission, Log

Change Sequence Names

To submit a complete genome, you need to specify a sequence name, a type (chromosome or plasmid), and topology for each sequence.

Number of Sequences: 3

Sequence Status: draft complete

③ Sequence Prefix ④

Use only alphabetic characters.

Entry names will be **sequence1-sequence3**.

Buttons: Save and Update, Reset

Red arrows and circled numbers 1-4 point to: 1. The instruction text; 2. The "complete" radio button; 3. The "Sequence Prefix" label and the text input field; 4. The "Save and Update" button.

W9-2: 配列名の変更

①completeにしたところ。②complete配列の登録時には、③配列の名称(Sequence Name; エントリー名)、④種別(Sequence Type; 染色体 or プラスミド)、⑤形状(Topology; 線状 or 環状)を指定する必要がある

http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8: ...

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.  
[2017-01-20 16:17:56.667496] Job started.  
[2017-01-20 16:25:41.084897] Job completed.
```

Result Features DDBJ Submission Log

Change Sequence Names

To submit a complete genome, you need to specify a sequence name, a type (chromosome or plasmid), and topology for each sequence.

Number of Sequences: 3

Sequence Status: draft complete

Sequence Name	Sequence Type	Topology
sequence1	chromosom	circular
sequence2	chromosom	circular
sequence3	chromosom	circular

W9-2: 配列名の変更

(入力配列の2, 3番目の配列がプラスミドだとわかっているので)赤枠部分をplasmidに変更。それ以外は特に変更する必要がないので、①Save and Update。せっかくコンプリートにしているのだから、②をchromosome、③をplasmid1、④をplasmid2などとしたほうがいいが、ここではやらない

← http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8: [2017- [2017-

status . COMPLETE

Result Features DDBJ Submission Log

Change Sequence Names

To submit a complete genome, you need to specify a sequence name, a type (chromosome or plasmid), and topology for each sequence.

Number of Sequences: 3

Sequence Status: draft complete

Sequence Name	Sequence Type	Topology
sequence1 ②	chromosome	circular
sequence2 ③	plasmid	circular
sequence3 ④	plasmid	circular

① Save and Update Reset

W9-2: 配列名の変更

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce>. The page title is "DFAST - Change Sequen...". The navigation bar includes "DFAST", "Analysis", "Archive", "Download", and "Help". A yellow notification box at the top states "Sequence Names have been successfully updated." with a close button (X). Two red arrows point to the notification: arrow ① points to the text, and arrow ② points to the close button. Below the notification, a warning message says "Remember the current URL to access this page. The result will be deleted 30 days after your last visit." and a "Delete this job now." button is present. The job details are: Title: (blank), JobID: 5359b058-11ef-40e7-8485-83fe7b7a1dce, Status: COMPLETE. A log box shows: [2017-01-20 16:17:56.643410] Job submitted., [2017-01-20 16:17:56.667496] Job started., [2017-01-20 16:25:41.084897] Job completed. The page has tabs for "Result", "Features", "DDBJ Submission", and "Log". The main heading is "Change Sequence Names" with a sub-heading "To submit a complete genome, you need to specify a sequence name, a type (chromosome or plasmid), and topology for each sequence." Below this, it shows "Number of Sequences: 3" and "Sequence Status: draft complete".

W9-2: 配列名の変更

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

[2017-01-20 16:17:56.643410] Job submitted.
[2017-01-20 16:17:56.667496] Job started.
[2017-01-20 16:25:41.084897] Job completed.

[Result](#) [Features](#) [DDBJ Submission](#) [Log](#)

Change Sequence Names

To submit a complete genome, you need to specify a sequence name, a type (chromosome or plasmid), and topology for each sequence.

Number of Sequences: 3

Sequence Status: draft complete

Sequence Name	Sequence Type	Topology
<input type="text" value="sequence1"/>	<input type="text" value="chromosom"/> ▼	<input type="text" value="circular"/> ▼

W9-3: アカウント取得など

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.
[2017-01-20 16:17:56.667496] Job started.
[2017-01-20 16:25:41.084897] Job completed.
```

Result Features **DDBJ Submission** Log

1. Preparation for Submit.

You can create DDBJ Submission Files (sequence file and annotation file) required to submit the genome through DDBJ Mass Submission System (MSS). If you want to submit a complete genome, you must provide a sequence name for each entry at [this page](#).

Before submission, you need to register BioProject and BioSample. Please follow the instruction below. For detailed information, please refer to [DDBJ Handbooks](#). If necessary, raw sequence data should be deposited in SRA.

このページでは DDBJ Mass Submission System (MSS) を用いて塩基配列を登録するために必要な2種類のファイル (配列ファイルとアノテーションファイル) を作成できます。コンプリートゲノムを DDBJ に登録する場合には [こちらのページ](#) で配列名・配列種別 (染色体/プラスミド)・直鎖/環状の指定を行ってください。

登録に先立ち、BioProject Database と BioSample Database への登録を次の手順に従って行います。詳細な手順は [DDBJ Handbooks](#) を参照してください。必要に応じてシーケンスデータの SRA への登録も行います。



W9-3: アカウ

赤枠あたりの話。実際の登録時には、①DDBJ登録アカウントの取得や②locus_tag_prefix情報(AP014680の場合は、LOOC260が予め取得したlocus_tag_prefixに相当)の取得などの作業があるが、本稿ではこのあたりの詳細には立ち入らない(第10回で詳述予定)。BioProjectやBioSample ID情報未取得のまま③の話に移行すると、W9-5で行き詰まる(空欄のままでも進めても差支えないが、話がややこしくなるのでやらない)

← http://dfast.nig.ac.jp/analysis/annotation/5359b

Mass Submission System (MSS). If you want to submit a genome, you must provide a sequence name for each page.

Before submission, you need to register BioProject and BioSample. Please follow the instruction below. For detailed information, please refer to [DDBJ Handbooks](#). If necessary, raw sequence data should be deposited in SRA.

1. Create a DDBJ submission account

Open the submission portal page [D-way](#), and create a new one.

2. Registration to the BioProject Database

Log-in at D-way, and create a new BioProject.

3. Registration to BioSample Database

Log-in at D-way, and create a new BioSample.

A unique [Locus Tag Prefix](#) is required for submitting an annotated genome, which can be registered during the submission procedure of BioProject or BioSample. To start using MSS, please apply from [MSS application form](#). Note that DFAST is not an official service of DDBJ. Please check the latest information at the [MSS guideline](#).

登録を次の手順に従って行います。詳細な手順は [DDBJ Handbooks](#) を参照してください。必要に応じてシーケンスデータの SRA への登録も行います。

1. DDBJ登録アカウントの取得

登録ポータル [D-way](#) でアカウント申請を行います。

2. BioProject Database への登録

D-way にログインし、新規 BioProject を登録します。

3. BioSample Database への登録

D-way にログインし、新規 BioSample を登録します。

ファイル作成時に必要な [Locus Tag Prefix](#) は BioProject もしくは BioSample 登録時に取得することができます。以上の準備ができたなら [申し込みフォーム](#) から MSS の利用申請を行います。MSSの詳細については DDBJ の [ガイドライン](#) を参照してください。(DFAST は DDBJ の公式サービスではありません。)

③ 2. Input Metadata

Input metadata by filling the form to create the submission file. Please refer to the [instruction](#) for more information for each item.

登録ファイル作成に必要なメタデータをフォームに入力します。入力項目の詳細な説明や作成例については [アノテーションファイル作成概説](#)

W9-4: Input Metadata

MSS application form. Note that DFAST is not an official service of DDBJ. Please check the latest information at the [MSS guideline](#).

① 2. Input Metadata

Input metadata by filling the form to create the submission file. Please refer to the [instruction](#) for more information for each item. If you want to add items not shown in the form, you can add them manually after downloading the annotation file.

登録ファイル作成に必要なメタデータをフォームに入力します。入力項目の詳細な説明や作成例については [アノテーションファイル作成概説](#) を参照してください。フォームにない項目はファイルダウンロード後に手動で修正を行ってください。メタデータを入力せずに空欄のままファイルを作成し、ダウンロード後に手動で入力することも可能です。

You can "Preview" the provided metadata. You can also import metadata from another DFAST job by providing the job ID in the text box below. By default, organism specific data, such as a species name or a strain name, are not imported. To enable this, check the "Override organism specific data".

メタデータ入力後 "Preview" ボタンで確認ができます。また、ジョブ ID を入力して他のジョブからメタデータをコピーすることができます。デフォルトでは生物種名や菌株名などのデータはコピーされませんが、"Override organism specific data" を有効にした場合、これらのデータも上書きされます。

▼ Show Form ② Preview

Enter Job ID to import another job's metadata.

: Override organism specific data.

3. Download Submission Files

W9-4: Input Metadata

MSS application form. Note that DFAST is not an official service of DDBJ. Please check the latest information at the [MSS guideline](#).

いては DDBJ の [ガイドライン](#) を参照してください。(DFAST は DDBJ の公式サービスではありません。)

2. Input Metadata

Input metadata by filling the form to create the submission file. Please refer to the [instruction](#) for more information for each item. If you want to add items not shown in the form, you can add them manually after downloading the annotation file.

登録ファイル作成に必要なメタデータをフォームに入力します。入力項目の詳細な説明や作成例については [アノテーションファイル作成概説](#) を参照してください。フォームにない項目はファイルダウンロード後に手動で修正を行ってください。メタデータを入力せずに空欄のままファイルを作成し、ダウンロード後に手動で入力することも可能です。

You can "Preview" the provided metadata. You can also import metadata from another DFAST job by providing the job ID in the text box below. By default, organism specific data, such as a species name or a strain name, are not imported. To enable this, check the "Override organism specific data".

メタデータ入力後 "Preview" ボタンで確認ができます。また、ジョブ ID を入力して他のジョブからメタデータをコピーすることができます。デフォルトでは生物種名や菌株名などのデータはコピーされませんが、"Override organism specific data" を有効にした場合、これらのデータも上書きされます。

: Override organism specific data.


Genus* Species*

W9-4: Input Metadata

You can "Preview" the provided metadata. You can also import metadata from another DFAST job by providing the job ID in the text box below. By default, organism specific data, such as a species name or a strain name, are not imported. To enable this, check the "Override organism specific data".

ださい。メタデータを入力せずに空欄のままファイルを作成し、ダウンロード後に手動で入力することも可能です。

メタデータ入力後 "Preview" ボタンで確認ができます。また、ジョブ ID を入力して他のジョブからメタデータをコピーすることができます。デフォルトでは生物種名や菌株名などのデータはコピーされませんが、"Override organism specific data" を有効にした場合、これらのデータも上書きされます。


▲ Hide Form Preview Enter Job ID to import another job's metadata. 

: Override organism specific data.

Genus* **Species***


Lactobacillus sp.

Strain* **Type Strain** **Culture Collection**

unkown NO  ex) JCM:00001

Locus Tag Prefix **BioProject*** **BioSample*** **Sequence Read Archive**

LOCUS ex) PRJDB90001 ex) SAMD900000000 ex) DRR999900



W9-5: 情報未取得だと...

入力欄の記入例に従って入力するわけだが、W9-3で述べたように①の情報を未取得のままだと、ここを埋められない。このデータは実質的にAP014680なので、対応する情報を埋めるとこんな感じ。とりあえずこれで話を進める

http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8: DF

You can "Preview" the provided metadata. You can also import metadata from another DFAST job by providing the job ID in the text box below. By default, organism specific data, such as a species name or a strain name, are not imported. To enable this, check the "Override organism specific data".

Hide Form Preview Enter Job ID to import another job's metadata.

: Override organism specific data.

Genus* Lactobacillus Species* hokkaidonensis

Strain* LOOC260 Type Strain YES Culture Collection JCM:18461

Locus Tag Prefix LOOC260 BioProject* PRJDB1726 BioSample* SAMD00000344 Sequence Read Archive DRR024500, DRR024501

ださい。メタデータロード後に手動で

メタデータ入力後 "Preview" ボタンで確認ができます。また、ジョブ ID を入力して他のジョブからメタデータをコピーすることができます。デフォルトでは生物種名や菌株名などのデータはコピーされませんが、"Override organism specific data" を有効にした場合、これらのデータも上書きされます。



W10-1: 適宜Save

①ページ下部に移動して、
適宜②Saveしておきましょう

http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8: DFAS: DDBJ Fast Anno... x

ex) Draft genome sequence of *Lactobacillus plantarum* strain 7006x. 2017

Authors* Status
ex) Mishima,H.; Yamada,T.; Robertson,G.R.; Kim,K.-H.; Park,C.S.; Shizuoka,T. Unpublishe ▾

Assembly

Assembly Method* **Sequencing Technology*** **Genome Coverage***
ex) Velevt v. 2.0, SPAdes v. 3.6.1, Plat ex) Illumina MiSeq, Ion PGM, PacBio S ex) 120x

Hold Date **Comment**
ex) 20160801

Save ②

3. Download Submission Files

You can download the sequence file and the annotation file. You 作成したファイルをダウンロードできます。Parser および

W10-2: Error

入力項目中にエラーとなった箇所があれば、適宜指摘してくれます。これまで入力した中で、1か所間違いがあることがわかります。① Sequence Read Archiveの箇所ですね。②ページ下部に移動し、該当箇所を確認

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8>. The page title is "DFAST Analysis Archive". A red error message box at the top states: "Error in the Sequence Read Archive field - Invalid Format. ex)DRR999999". A red arrow labeled "1" points to the error message, and another red arrow labeled "2" points to the bottom of the page. Below the error message, there is a warning: "Remember the current URL to access this page. The result will be deleted 30 days after your last visit." and a "Delete" button. The page content includes job details: Title, JobID (5359b058-11ef-40e7-8485-83fe7b7a1dce), and Status (COMPLETE). A log shows job submission, start, and completion times. The "DDBJ Submission" tab is active, showing instructions for creating DDBJ Submission Files and a note about the DDBJ Mass Submission System (MSS).

Error in the Sequence Read Archive field - Invalid Format. ex)DRR999999

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :

JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce

Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.
[2017-01-20 16:17:56.667496] Job started.
[2017-01-20 16:25:41.084897] Job completed.
```

Result Features **DDBJ Submission** Log

1. Preparation for Submit.

You can create DDBJ Submission Files (sequence file and annotation file) required to submit the genome through DDBJ Mass Submission System (MSS). If you want to submit a complete genome, you must provide a sequence name for each entry at [this page](#).

このページでは DDBJ Mass Submission System (MSS) を用いて塩基配列を登録するために必要な2種類のファイル (配列ファイルとアンテーションファイル) を作成できます。コンプリートゲノムを DDBJ に登録する場合には [こちらのページ](#) で配列名・配列種別 (染色体/プラスミド)・直鎖/環状の指定を行ってください。

W10-2: Error

①エラーは、Sequence Read Archive (SRA) field。ここは、複数個のSRA IDの入力を想定していないことに由来するエラー。今この画面上では、単純にテキスト形式の「DDBJ登録用ファイル」を作成しているだけなので、とりあえず1つ分だけのIDを入力しておけばよい。後で保存したファイルを、直接テキストエディタでいじって、複数IDを入力したい場合は直接ファイルに書き込む、という作業を行うことで対応することができる

The screenshot shows a web form for DDBJ registration. The URL is <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7>. The form includes several fields:

- Genus***: Lactobacillus
- Species***: hokkaidonensis
- Strain***: LOOC260
- Type Strain**: YES (dropdown menu)
- Culture Collection**: JCM:18461
- Locus Tag Prefix**: LOOC260
- BioProject***: PRJDB1726
- BioSample***: SAMD00000344
- Sequence Read Archive**: DRR24500, DRR024501 (indicated by a red arrow and the number 1)
- Submitter**: (empty)
- Submitters***: ex) Robertson,G.R.; Mishima,H.; Kim,K.-H.

Buttons: Hide Form, Preview, Enter Job ID to, : Override organism specific data.

Must be separated with semicolons.

W10-3: Error対処

①よく考えたら、このアセンブリ結果は4セル分のPacBioデータのうちの1つ分(DRR054113)だけで得られたものだったので、DRR054113を入力し、Save

が、"Override organism specific data"を有効にした場合、これらのデータも上書きされます。

▲ Hide Form

Preview

Enter Job ID to import another job's metadata.

: Override organism specific data.

Genus*

Lactobacillus

Species*

hokkaidonensis

Strain*

LOOC260

Type Strain

YES



Culture Collection

JCM:18461

Locus Tag Prefix

LOOC260

BioProject*

PRJDB1726

BioSample*

SAMD00000344

Sequence Read Archive

DRR054113



Submitter

Submitters*

ex) Robertson,G.R.; Mishima,H.; Kim,K.-H.

Must be separated with semicolons

W10-3: Error対処

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.
[2017-01-20 16:17:56.667496] Job started.
[2017-01-20 16:25:41.084897] Job completed.
```

Result Features **DDBJ Submission** Log

1. Preparation for Submit.

You can create DDBJ Submission Files (sequence file and annotation file) required to submit the genome through DDBJ Mass Submission System (MSS). If you want to submit a complete genome, you must provide a sequence name for each entry at [this page](#).

Before submission, you need to register BioProject and BioSample. Please follow the instruction below. For detailed information, please refer to [DDBJ Handbooks](#). If necessary, raw sequence data should be deposited in SRA.


このページでは DDBJ Mass Submission System (MSS) を用いて塩基配列を登録するために必要な2種類のファイル (配列ファイルとアノテーションファイル) を作成できます。コンプリートゲノムを DDBJ に登録する場合には [こちらのページ](#) で配列名・配列種別 (染色体/プラスミド)・直鎖/環状の指定を行ってください。

登録に先立ち、BioProject Database と BioSample Database への登録を次の手順に従って行います。詳細な手順は [DDBJ Handbooks](#) を参照してください。必要に応じてシーケンスデータの SRA への登録も行います。

W10-4: Preview

①Previewボタンを押して、
これまでの入力内容を確認

が、"Override organism specific data" を有効にした場合、これらのデータも上書きされます。

▲ Hide Form Preview Enter Job ID to import another job's metadata. 

: Override organism specific data.

Genus* **Species***

Lactobacillus hokkaidonensis

Strain* **Type Strain** **Culture Collection**

LOOC260 YES JCM:18461

Locus Tag Prefix **BioProject*** **BioSample*** **Sequence Read Archive**

LOOC260 PRJDB1726 SAMD00000344 DRR054113

Submitter

Submitters*

ex) Robertson,G.R.; Mishima,H.; Kim,K.-H.

Must be separated with semicolons.

W10-4: Preview

①入力内容が反映されているようだ。②
ページ下部に移動して他の箇所も確認

が、"Override organism specific data" を有効にした場合、これらのデータも上書きされます

Preview for Common Entry

Entry	Feature	Location	Qualifier	Value
COMMON	DBLINK		project	PRJDB1726
			biosample	SAMD00000344
			sequence read archive	DRR054113
	SUBMITTER		contact	
			ab_name	
			email	
			phone	
			fax	
			institute	
			department	
			country	
			state	

W10-4: Preview

①一番下まで移動したところ。②のあたりも入力内容が反映されているようだ。③Close

		line	Annotated at D-FAST https://dfast.nig.ac.jp	
	ST_COMMENT	tagset_id	Genome-Assembly-Data	
		Assembly Method		
		Genome Coverage		
		Sequencing Technology		
	source	1..E	organism	Lactobacillus hokkaidonensis
			strain	LOOC260
			note	type strain of Lactobacillus hokkaidonensis
			culture_collection	JCM:18461
			mol_type	genomic DNA
			ff_definition	@[organism]@@ DNA, complete genome, strain: @[strain]@@

W10-5: Submitter

①次はSubmitter情報。②
ちょっとページ下部に移動

が、"Override organism specific data" を有効にした場合、これらのデータも上書きされます。

▲ Hide Form Preview Enter Job ID to import another job's metadata.

: Override organism specific data.

Genus* Lactobacillus **Species*** hokkaidonensis

Strain* LOOC260 **Type Strain** YES **Culture Collection** JCM:18461


Locus Tag Prefix LOOC260 **BioProject*** PRJDB1726 **BioSample*** SAMD00000344 **Sequence Read Archive** DRR054113

Submitter

Submitters*

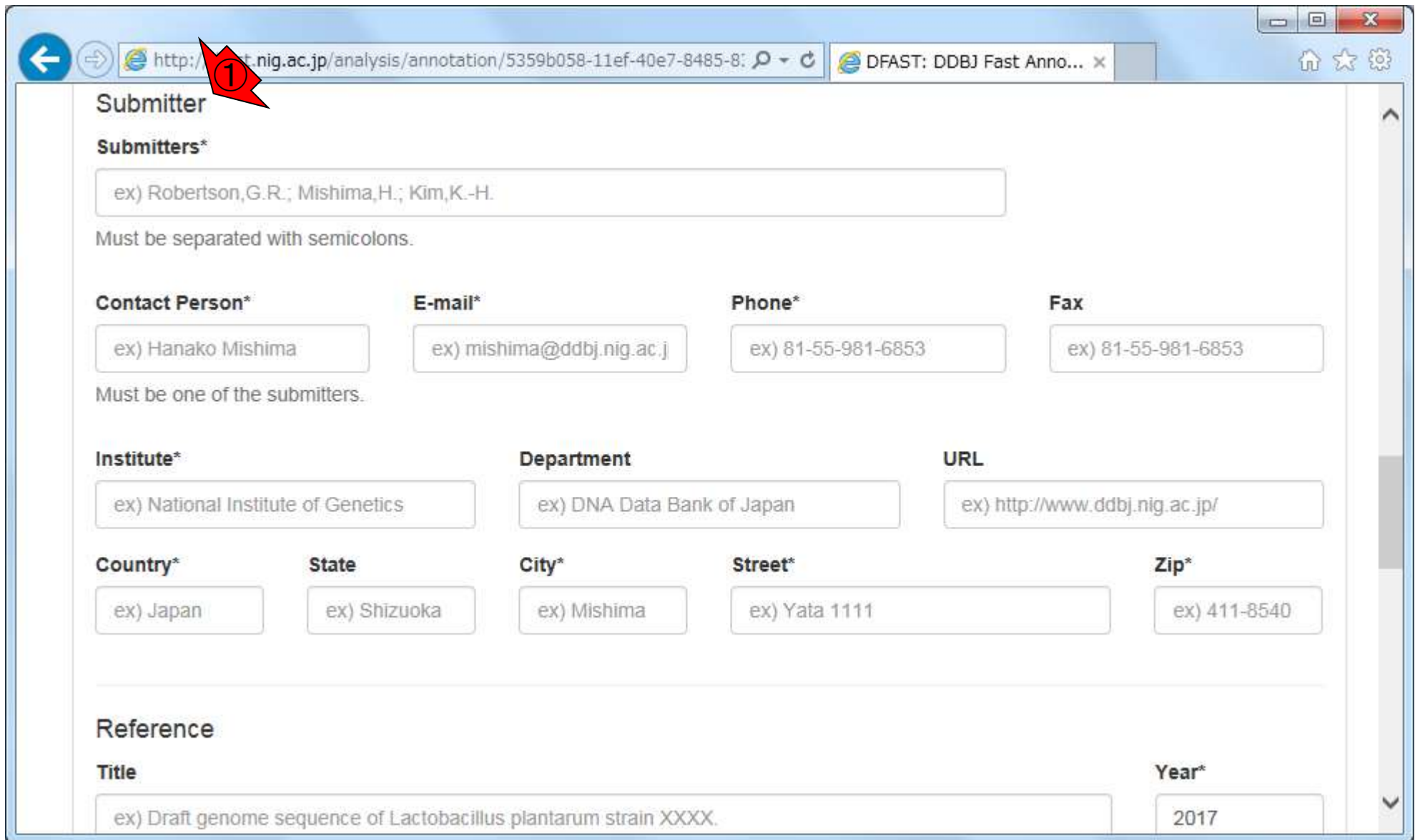
ex) Robertson,G.R.; Mishima,H.; Kim,K.-H.

Must be separated with semicolons.



①Submitter部分。実際のAP014680の入力内容と一部対応させると…

W10-5: Submitter



Submitter

Submitters*

ex) Robertson,G.R.; Mishima,H.; Kim,K.-H.

Must be separated with semicolons.

Contact Person* **E-mail*** **Phone*** **Fax**

ex) Hanako Mishima ex) mishima@ddbj.nig.ac.j ex) 81-55-981-6853 ex) 81-55-981-6853

Must be one of the submitters.

Institute* **Department** **URL**

ex) National Institute of Genetics ex) DNA Data Bank of Japan ex) http://www.ddbj.nig.ac.jp/

Country* **State** **City*** **Street*** **Zip***

ex) Japan ex) Shizuoka ex) Mishima ex) Yata 1111 ex) 411-8540

Reference

Title **Year***

ex) Draft genome sequence of Lactobacillus plantarum strain XXXX. 2017

W10-5: Submitter

こんな感じ。①の部分のみAP014680と同じにして、②それ以外は例示されているもので埋めている

Submitter

Submitters*

Tohno, M.; Tanizawa, Y.

Must be separated with semicolons.

Contact Person* **E-mail*** **Phone*** **Fax**

Hanako Mishima mishima@ddbj.nig.go.jp 81-55-981-6853 81-55-981-6853

Must be one of the submitters.

Institute* **Department** **URL**

National Institute of Genetics DNA Data Bank of Japan http://www.ddbj.nig.ac.jp

Country* **State** **City*** **Street*** **Zip***

Japan Shizuoka Mishima Yata 111 411-8540

Reference

Title **Year***

ex) Draft genome sequence of Lactobacillus plantarum strain XXXX. 2017

W10-5: Submitter

- ①このあたりまで移動。
- ②適宜Saveしましょう

Hold Date: ex) 20160801

Comment: [Empty text box]

Save

3. Download Submission Files

You can download the sequence file and the annotation file. You can also check the file format and translated amino acid sequence using [Parser](#) and [transChecker](#) (© DDBJ). If you have edited the annotation and metadata, files are automatically updated.

作成したファイルをダウンロードできます。[Parser](#) および [transChecker](#) (© DDBJ) を使用してファイル書式およびアミノ酸翻訳配列のチェックを行うことができます。アノテーションやメタデータを編集した場合、変更内容は自動的に反映されます。

Format Check

[Sequence \(Genomic Fasta File\)](#) [Annotation \(TAB-separated Table\)](#) [TAB-separated Table without annotation](#) [# Use this to submit a genome without annotation.](#)

4. Submit Your Genome.

Send submission files to DDBJ MSS in email attachments ([mass at ddbi.nig.ac.jp](mailto:mass@ddbi.nig.ac.jp)). Alternatively, you can send the URL of this page

作成したファイルに問題がなければ、ファイルをメールに添付し DDBJ MSS ([mass at ddbi.nig.ac.jp](mailto:mass@ddbi.nig.ac.jp)) 宛てに送付します。また、ファイルを添

W10-6: Format Check

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8>. The page has a form at the top with a 'Hold Date' field (containing 'ex) 20160801') and a 'Comment' field. Below the form is a 'Save' button. The main content area is titled '3. Download Submission Files' and contains instructions in both English and Japanese. The English text states: 'You can download the sequence file and the annotation file. You can also check the file format and translated amino acid sequence using Parser and transChecker (© DDBJ). If you have edited the annotation and metadata, files are automatically updated.' The Japanese text states: '作成したファイルをダウンロードできます。Parser および transChecker (© DDBJ) を使用してファイル書式およびアミノ酸翻訳配列のチェックを行うことができます。アノテーションやメタデータを編集した場合、変更内容は自動的に反映されます。' Below the text are four buttons: 'Format Check', 'Sequence (Genomic Fasta File)', 'Annotation (TAB-separated Table)', and 'TAB-separated Table without annotation # Use this to submit a genome without annotation.' A red arrow labeled '1' points to the section title, and another red arrow labeled '2' points to the 'Format Check' button. Below this section is a '4. Submit Your Genome.' section with instructions in both English and Japanese.

W10-6: Format Check

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8>. The page has a form with a 'Hold Date' field containing 'ex) 20160801' and an empty 'Comment' field. Below the form is a 'Save' button. The main content area is titled '3. Download Submission Files' and contains instructions in English and Japanese. At the bottom of this section, there are four buttons: 'Format Check', 'Sequence (Genomic Fasta File)', 'Annotation (TAB-separated Table)', and 'TAB-separated Table without annotation'. A red arrow with the number '1' points to the 'Format Check' button. Below this section is another section titled '4. Submit Your Genome.' with instructions in English and Japanese.

Hold Date
ex) 20160801

Comment

Save

3. Download Submission Files

You can download the sequence file and the annotation file. You can also check the file format and translated amino acid sequence using [Parser](#) and [transChecker](#) (© DDBJ).
If you have edited the annotation and metadata, files are automatically updated.

作成したファイルをダウンロードできます。[Parser](#) および [transChecker](#) (© DDBJ) を使用してファイル書式およびアミノ酸翻訳配列のチェックを行うことができます。
アノテーションやメタデータを編集した場合、変更内容は自動的に反映されます。

Format Check Sequence (Genomic Fasta File) Annotation (TAB-separated Table) TAB-separated Table without annotation
Use this to submit a genome without annotation.

4. Submit Your Genome.

Send submission files to DDBJ MSS in email attachments (mass at ddbi.nig.ac.jp). Alternatively, you can send the URL of this page

作成したファイルに問題がなければ、ファイルをメールに添付し DDBJ MSS (mass at ddbi.nig.ac.jp) 宛てに送付します。また、ファイルを添

W10-6: Format Check

約1分ほどでこのような画面になる。①Submitters情報のところで間違っているようだ。②×を押す

Format Check for DDBJ Submission Files

Check For Annotation File

jParser (Ver. 6.54) started.

reading sequence.....

reading sequence.....

reading sequence.....

reading sequence.....

reading sequence.....

reading sequence.....

reading annotation.....

reading annotation.....

JP0114:WAR:STX:ANN:Line [5]: The value format of [ab_name] [Tohno, M.] is possibly wrong.

JP0114:WAR:STX:ANN:Line [6]: The value format of [ab_name] [Tanizawa, Y.] is possibly wrong.

JP0078:ER1:STX:ANN:Line [17]: Invalid length [] for [ab_name] qualifier, it must be modified following [1-64].

JP0078:ER1:STX:ANN:Line [18]: Invalid length [] for [title] qualifier, it must be modified following [1-255].

JP0157:WAR:STX:ANN:Line [24]: There is no value for [Assembly Method] qualifier in [ST_COMMENT] feature.

JP0157:WAR:STX:ANN:Line [25]: There is no value for [Genome Coverage] qualifier in [ST_COMMENT] feature.

JP0157:WAR:STX:ANN:Line [26]: There is no value for [Sequencing Technology] qualifier in [ST_COMMENT] feature.

W10-7: 修正

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce/subm>. The page contains a form with a "Hold Date" field (example: 20160801) and a "Comment" field. A "Save" button is located below the form. Below the form is a section titled "3. Download Submission Files" which provides instructions on downloading sequence and annotation files. It includes a "Format Check" button and links for "Sequence (Genomic Fasta File)", "Annotation (TAB-separated Table)", and "TAB-separated Table without annotation". A red arrow with the number 1 points to the top of the page, indicating the location of the Submitters information mentioned in the header.

W10-7: 修正

①Submittersのところ。②の部分にスペースがあるとダメです。また、③Contact Personのところの④を見ると、one of the submittersを指定しなければいけないと書かれている

The screenshot shows a web browser window with the URL <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce/subm>. The form contains the following fields and annotations:

- Submitter**: Input field containing "Tohno, M.; Tanizawa, Y.". Two red arrows labeled "②" point to the spaces after "M." and "Y.". A red arrow labeled "①" points to the semicolon separator.
- Must be separated with semicolons.**: Text instruction below the Submitter field.
- Contact Person***: Input field containing "Hanako Mishima". A red arrow labeled "④" points to this field.
- E-mail***: Input field containing "mishima@ddbj.nig.go.jp". A red arrow labeled "③" points to this field.
- Phone***: Input field containing "81-55-981-6853".
- Fax**: Input field containing "81-55-981-6853".
- Must be one of the submitters.**: Text instruction below the Contact Person field.
- Institute***: Input field containing "National Institute of Genetics".
- Department**: Input field containing "DNA Data Bank of Japan".
- URL**: Input field containing "http://www.ddbj.nig.ac.jp".
- Country***: Input field containing "Japan".
- State**: Input field containing "Shizuoka".
- City***: Input field containing "Mishima".
- Street***: Input field containing "Yata 111".
- Zip***: Input field containing "411-8540".
- Reference**: Section header.
- Title**: Input field containing "ex) Draft genome sequence of Lactobacillus plantarum strain XXXX.". A red arrow labeled "④" points to this field.
- Year***: Input field containing "2017".

W10-7: 修正

①の部分のスペースを消し、②AP014680の実際のContact PersonでもあるMasanori Tohnoに変更して、Saveしたのち、再度Format Checkを実行…

The screenshot shows a web browser window displaying a form for DFAST (DDBJ Fast Annotation System). The form is titled "Submitter" and contains several fields. Two red arrows labeled "①" point to the "Submitter" field, which contains the text "Tohno,M.; Tanizawa,Y.". A red arrow labeled "②" points to the "Contact Person*" field, which contains the text "Masanori Tohno". Below the "Submitter" field, there is a note: "Must be separated with semicolons." Below the "Contact Person*" field, there is a note: "Must be one of the submitters." The form also includes fields for "E-mail*", "Phone*", "Fax", "Institute*", "Department", "URL", "Country*", "State", "City*", "Street*", "Zip*", and "Reference". The "Reference" section has a "Title" field containing "ex) Draft genome sequence of Lactobacillus plantarum strain XXXX." and a "Year*" field containing "2017".

Submitter

Submitter
① ①
Tohno,M.; Tanizawa,Y.
Must be separated with semicolons.

Contact Person* E-mail* Phone* Fax
Masanori Tohno ② mishima@ddbj.nig.go.jp 81-55-981-6853 81-55-981-6853
Must be one of the submitters.

Institute* Department URL
National Institute of Genetics DNA Data Bank of Japan http://www.ddbj.nig.ac.jp

Country* State City* Street* Zip*
Japan Shizuoka Mishima Yata 111 411-8540

Reference

Title Year*
ex) Draft genome sequence of Lactobacillus plantarum strain XXXX. 2017

W10-8: Format Check

Submittersのところのエラーが
消えたことがわかる。こんな感じ
で他の箇所も適宜修正していく

Format Check for DDBJ Submission Files

Check For Annotation File

jParser (Ver. 6.54) started.

reading sequence.....

reading sequence.....

reading sequence.....

reading sequence.....

reading sequence.....

reading sequence.....

reading annotation.....

reading annotation.....

JP0078:ER1:STX:ANN:Line [17]: Invalid length [] for [ab_name] qualifier, it must be modified following [1-64].

JP0078:ER1:STX:ANN:Line [18]: Invalid length [] for [title] qualifier, it must be modified following [1-255].

JP0157:WAR:STX:ANN:Line [24]: There is no value for [Assembly Method] qualifier in [ST_COMMENT] feature.

JP0157:WAR:STX:ANN:Line [25]: There is no value for [Genome Coverage] qualifier in [ST_COMMENT] feature.

JP0157:WAR:STX:ANN:Line [26]: There is no value for [Sequencing Technology] qualifier in [ST_COMMENT] feature.

jParser (Ver. 6.54) finished.

W10-9: 修正

このあたりは空欄のままなので、*のついた必須事項部分を中心に、AP014680の情報を参考にして埋める

Reference

Title
ex) Draft genome sequence of Lactobacillus plantarum strain XXXX.

Year*
2017

Authors*
ex) Mishima,H.; Yamada,T.; Robertson,G.R.; Kim,K.-H.; Park,C.S.; Shizuoka,T.

Status
Unpublishe ▾

Assembly

Assembly Method*
ex) Velevt v. 2.0, SPAdes v. 3.6.1, Plat

Sequencing Technology*
ex) Illumina MiSeq, Ion PGM, PacBio S

Genome Coverage*
ex) 120x

Hold Date
ex) 20160801

Comment

Save

W10-9: 修正

①は、第7回W8-1やW9-2からわかる。②自分がどんな機器でデータを取得したかくらいは知っておこう。③は第7回原稿PDFのp104右上あたりで算出している。④は論文のタイトルを書くところ。とりあえずComplete genome...としている

The screenshot shows a web browser window with a URL: <http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce/subn>. The browser tabs include "DFAST: DDBJ Fast An...".

The form is titled "Reference" and contains the following fields:

- Title:** Complete genome sequence of Lactobacillus hokkaidonensis LOOC260T (marked with ④)
- Year*:** 2017
- Authors*:** Tohno,M.; Tanizawa,Y.
- Status:** Unpublishe (dropdown menu)

The form is titled "Assembly" and contains the following fields:

- Assembly Method*:** HGAP v. 2.2.0 (marked with ①)
- Sequencing Technology*:** PacBio RSII (marked with ②)
- Genome Coverage*:** 65x (marked with ③)

The form also includes:

- Hold Date:** ex) 20160801
- Comment:** (empty text area)
- Save:** (button)

W10-10: Format Check

Assembly

Assembly Method* **Sequencing Technology*** **Genome Coverage***

Hold Date **Comment**

Save ①

3. Download Submission Files

You can download the sequence file and the annotation file. You can also check the file format and translated amino acid sequence using [Parser](#) and [transChecker](#) (© DDBJ).
If you have edited the annotation and metadata, files are automatically updated.

作成したファイルをダウンロードできます。[Parser](#) および [transChecker](#) (© DDBJ) を使用してファイル書式およびアミノ酸翻訳配列のチェックを行うことができます。
アノテーションやメタデータを編集した場合、変更内容は自動的に反映されます。

Format Check ② **Sequence (Genomic Fasta File)** **Annotation (TAB-separated Table)** **TAB-separated Table without annotation # Use this to submit a genome without**

W10-10: Format Check

①特にエラーはでてないっぽい。②ページ下部までチェック

Format Check for DDBJ Submission Files

Check For Annotation File

jParser (Ver. 6.54) started.

- reading sequence.....
- reading sequence.....
- reading sequence.....
- reading sequence.....
- reading sequence.....
- reading annotation.....
- reading annotation.....

jParser (Ver. 6.54) finished.

Check For Translation

transChecker (Ver. 2.19) started at Mon Jan 23 14:19:04 JST 2017

Reading Sequence File(s).....finished.

Reading Annotation File(s) and Checking translation error....

..finished.

TransChecker (Ver. 2.19) finished at Mon Jan 23 14:19:06 JST 2017

W10-10: Format Check

① ページ下部まで移動して、② 全体を概観。大丈夫そうなので③ Close

Assembly
Assembly M
HGAP v. 2

Hold Date
ex) 201608

Save

3. Download

You can download the assembly file. You can also check the assembly file using Parser. If you have error, it will be automatically corrected.

Format Check

File) Table) # Use this to submit a genome without

Check For Annotation File

jParser (Ver. 6.54) started.
reading sequence.....
reading sequence.....
reading sequence.....
reading sequence.....
reading sequence.....
reading annotation.....
reading annotation.....
jParser (Ver. 6.54) finished.

Check For Translation

transChecker (Ver. 2.19) started at Mon Jan 23 14:19:04 JST 2017
Reading Sequence File(s).....finished.
Reading Annotation File(s) and Checking translation error....
..finished.
TransChecker (Ver. 2.19) finished at Mon Jan 23 14:19:06 JST 2017

Close

① ② ③

W10-11: とりあえず完成

①と②がDDBJ登録用ファイル。W9-3およびW9-5でも述べたように、まだ仮想の値でデータを一通り埋めたただけなのでダウンロードしてもしょうがない。ここまであえてやったのは、②をクリックして得られる情報を見せたかったから

← http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce/subn

Assembly

Assembly Method*	Sequencing Technology*	Genome Coverage*
HGAP v. 2.2.0	PacBio RSII	65x

Hold Date	Comment
ex) 20160801	

[Save](#)

3. Download Submission Files

You can download the sequence file and the annotation file. You can also check the file format and translated amino acid sequence using [Parser](#) and [transChecker](#) (© DDBJ).
If you have edited the annotation and metadata, files are automatically updated.

作成したファイルをダウンロードできます。[Parser](#) および [transChecker](#) (© DDBJ) を使用してファイル書式およびアミノ酸翻訳配列のチェックを行うことができます。
アノテーションやメタデータを編集した場合、変更内容は自動的に反映

Format Check	Sequence (Genomic Fasta File)	Annotation (TAB-separated Table)	TAB-separated Table without annotation # Use this to submit a genome without
------------------------------	-----------------------------------------------	--------------------------------------------------	----------------------------------------------------------------------------------------------



W10-12: Annotation

DDBJ形式ファイルが作成できていることがわかります。AP014680の情報とかなり似ていることがわかります

```
COMMON DBLINK      project PRJDB1726
                    biosample      SAMD00000344
                    sequence read archive DRR054113
SUBMITTER           contact Masanori Tohno
                    ab_name Tohno,M.
                    ab_name Tanizawa,Y.
                    email  mishima@dccb.jnig.go.jp
                    phone  81-55-981-6853
                    fax    81-55-981-6853
                    institute National Institute of Genetics
                    department DNA Data Bank of Japan
                    country Japan
                    state   Shizuoka
                    city    Mishima
                    street  Yata 111
                    zip     411-8540
REFERENCE           ab_name Tohno,M.
                    ab_name Tanizawa,Y.
                    title   Direct Submission
                    status  Unpublished
                    year    2017
COMMENT            line   Annotated using prokka 1.11 from http://www.vicbioinformatics.com.
                    line   Annotated at D-FAST https://dfast.nig.ac.jp
ST_COMMENT         tagset_id      Genome-Assembly-Data
                    Assembly Method HGAP v. 2.2.0
                    Genome Coverage 65x
                    Sequencing Technology PacBio RSII
sequencel         source 1..2277983      organism      Lactobacillus hokkaidonensis
                    strain LOOC260
                    mol_type      genomic DNA
                    culture_collection JCM:18461
                    note    type strain of Lactobacillus hokkaidonensis
                    ff definition @@[organism]@@ DNA, complete genome, strain: @@[strain]@@
```

W10-12: Annotation

①のあたりなどは自分が入力した情報となっております

```
COMMON  DBLINK      project PRJDB1726
        biosample     SAMD00000344
        sequence read archive  DRR054113
SUBMITTER contact Masanori Tohno
        ab_name Tohno,M.
        ab_name Tanizawa,Y.
        email  mishima@dodbj.nig.go.jp
        phone  81-55-981-6853
        fax    81-55-981-6853
        institute      National Institute of Genetics
        department     DNA Data Bank of Japan
        country Japan
        state  Shizuoka
        city   Mishima
        street Yata 111
        zip    411-8540
REFERENCE ab_name Tohno,M.
        ab_name Tanizawa,Y.
        title  Direct Submission
        status Unpublished
        year   2017
COMMENT   line  Annotated using prokka 1.11 from http://www.vicbiinformatics.com.
        line  Annotated at D-FAST https://dfast.nig.ac.jp
ST_COMMENT tagset id      Genome-Assembly-Data
        Assembly Method HGAP v. 2.2.0
        Genome Coverage 65x
        Sequencing Technology PacBio RSII
sequencel source 1..2277983 organism Lactobacillus hokkaidonensis
        strain LOOC260
        mol_type genomic DNA
        culture_collection JCM:18461
        note type strain of Lactobacillus hokkaidonensis
        ff definition @@[organism]@@ DNA, complete genome, strain: @@[strain]@@
```

W10-12: Annotation

①などは自分で入力した覚えはないもの。D-FAST経由でDDBJ登録用ファイルを作成したからこそ自動で付加してくれるものの例です

COMMON DBLINK project PRJDB1726
biosample SAMD00000344
sequence read archive DRR054113

SUBMITTER contact Masanori Tohno
ab_name Tohno,M.
ab_name Tanizawa,Y.
email mishima@dccb.j.nig.go.jp
phone 81-55-981-6853
fax 81-55-981-6853
institute National Institute of Genetics
department DNA Data Bank of Japan
country Japan
state Shizuoka
city Mishima
street Yata 111
zip 411-8540

REFERENCE ab_name Tohno,M.
ab_name Tanizawa,Y.
title Direct Submission
status Unpublished
year 2017

COMMENT line Annotated using prokka 1.11 from <http://www.vicbioinformatics.com>.
line Annotated at D-FAST <https://dfast.nig.ac.jp>

ST_COMMENT tagset_id Genome-Assembly-Data
Assembly Method HGAP v. 2.2.0
Genome Coverage 65x
Sequencing Technology PacBio RSII

sequencel source 1..2277983 organism Lactobacillus hokkaidonensis
strain LOOC260
mol_type genomic DNA
culture_collection JCM:18461
note type strain of Lactobacillus hokkaidonensis
ff definition @@[organism]@@ DNA, complete genome, strain: @@[strain]@@

W11-1: DNAPlotter

wellcome trust
sanger
institute

SCIENCE

DNAPlotter

- Overview
- Download
- Learn
- License

Overview

DNAPlotter can be used to generate images of circular and linear DNA maps to display regions and features of interest. The images can be inserted into a document or printed out directly. As this uses [Artemis](#) it can read in the common file formats EMBL, GenBank and GFF3.

Download and Installation

Tool type

- [Annotation](#)

Screenshots

W11-2: 入力ファイル

必要な入力ファイルは、①DFASTアノテーションのResultタブから得られる、②GenBank形式ファイル(annotation.gbk)のみ。③ゲノム配列のFASTAファイル(genome.fna)とアノテーション情報ファイル(annotation.gff)の組合せでも原理的には読み込めるはず。現実的にはgffファイルの読込時にエラーが出てだめだったが、まずはそれを示す

Remember the [current URL](http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-8) to access this page. The result will be deleted 30 days after your last visit.

Delete this job now. => This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.  
[2017-01-20 16:17:56.667496] Job started.  
[2017-01-20 16:25:41.084897] Job completed.
```

①

Result Features DDBJ Submission Log

Genome Statistics

Total Length (bp)	2,400,584
No. of Sequences	3
GC Content (%)	38.2%
N50	2,277,983
Gap Ratio (%)	0.0%

Download Files

- Genbank Flat File : [annotation.gbk](#)
- GFF3-formatted File : [annotation.gff](#)**
- Genome Fasta File : [genome.fna](#)**
- Protein Fasta File : [protein.faa](#)
- CDS Fasta File : [cds.fna](#)
- RNA Fasta File : [ma.fna](#)
- Feature Table : [features.tsv](#)
- Genome Statistics : [statistics.txt](#)

②

③

W11-3: インストール

①ページ下部に移動し、②Download and Installationが見られるようにする

The screenshot shows a web browser window with the URL <http://www.sanger.ac.uk/science/tools/dnaplotter>. The page header includes the Wellcome Trust Sanger Institute logo and a search bar. The main content area features a navigation menu on the left with links for Overview, Download, Learn, and License. The main text area contains an 'Overview' section and a 'Download and Installation' section. A 'Tool type' section lists 'Annotation'. A 'Screenshots' section displays a circular DNA map. Two red arrows with numbers 1 and 2 indicate the steps: arrow 1 points to the hamburger menu icon, and arrow 2 points to the 'Download and Installation' link.

wellcome trust
sanger
institute

SCIENCE

DNAPlotter

- Overview
- Download
- Learn
- License

Overview

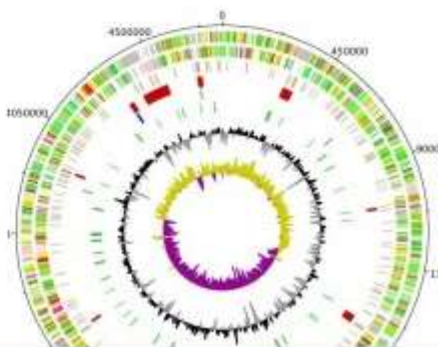
DNAPlotter can be used to generate images of circular and linear DNA maps to display regions and features of interest. The images can be inserted into a document or printed out directly. As this uses [Artemis](#) it can read in the common file formats EMBL, GenBank and GFF3.

Download and Installation

Tool type

- [Annotation](#)

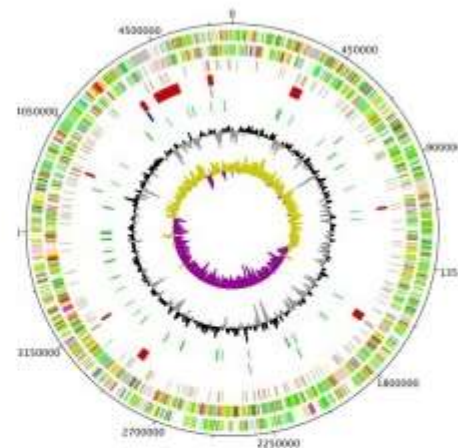
Screenshots



W11-3: インストール

①好きなものをダウンロードしてください。以降は②Windowsの例で説明します。尚、私(門田)のWindows環境では、セキュリティに関する設定上の問題かはわかりませんが、③のやり方ではうまく起動させることができませんでした

The screenshot shows the DNAPlotter website interface. On the left is a navigation menu with links for Overview, Download, Learn, License, Contact, Related, and Publications. The main content area is titled "Download and Installation" and contains the following text: "The latest release of DNAPlotter can be downloaded from our ftp site:" followed by a list of operating systems: UNIX, MacOS, and Windows. A red arrow labeled "2" points to the "Windows" link. To the right of the list is a bracket and a red arrow labeled "1" pointing to the list. Below this, it says "It can also be launched with Java Web Start:" followed by a blue "Launch" button with a right-pointing arrow. A red arrow labeled "3" points to the "Launch" button. At the bottom, it says "For further download and installation instructions,". On the right side of the page, there is a circular genome plot labeled "Example showing S. typhi genome" and a section titled "Related Tools" with a sub-item "Artemis" described as "A free genome browser and annotation tool that allows visualisation of sequence features,".



Example showing *S. typhi* genome

Related Tools

- **Artemis**
A free genome browser and annotation tool that allows visualisation of sequence features,

W11-4: ダウンロードと実行

- ①Windows版のjarファイルを
- ②デスクトップなどに保存して、
- ③ダブルクリックで実行

The screenshot shows a web browser window at <http://www.sanger.ac.uk/science/tools/dnaplotter>. The page includes a navigation menu on the left with links for Overview, Download, Learn, License, Contact, and Related. The main content area features a "Download and Installation" section with the following text:

DNAPlotter can be used to generate images of circular and linear DNA maps to display regions and features of interest. The images can be inserted into a document or printed out directly. As this uses [Artemis](#) it can read in the common file formats EMBL, GenBank and GFF3.

Download and Installation

The latest release of DNAPlotter can be downloaded from our ftp site:

- [UNIX](#)
- [MacOS](#)
- [Windows](#)

It can also be launched with Java Web Start:

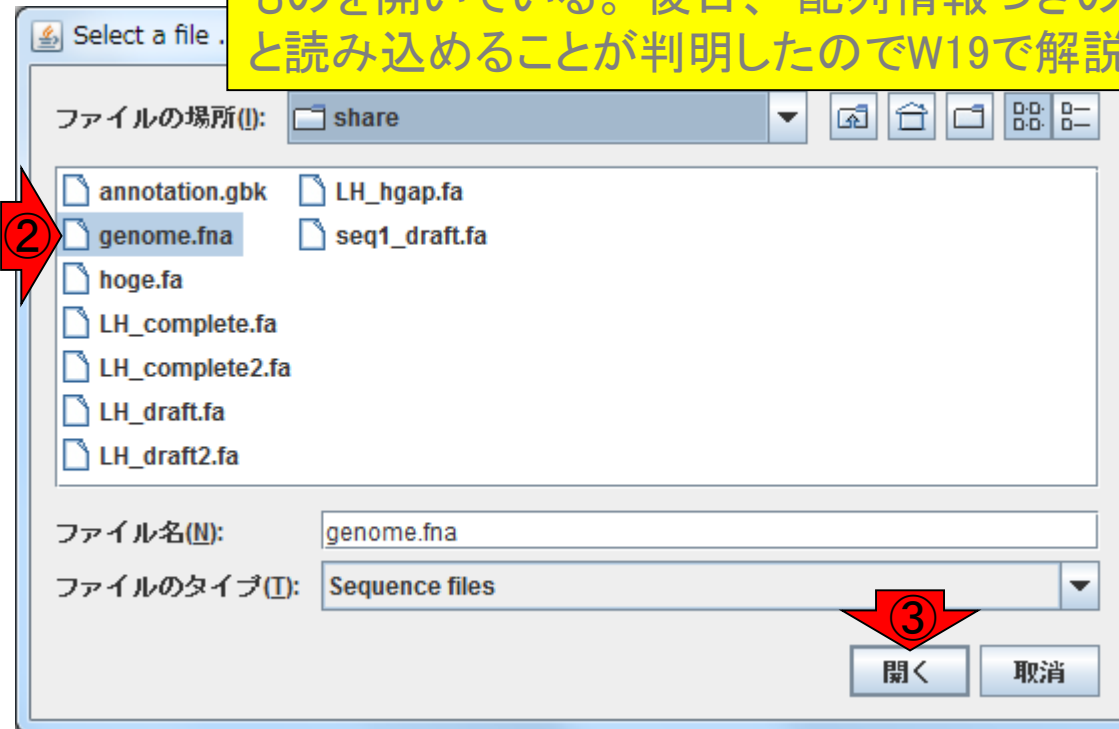
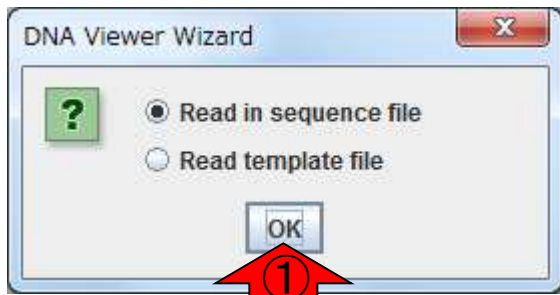
On the right side, there is a "Screenshots" section with an image of a circular DNA map and the caption "Example showing *S. typhi* genome". Below this is a "Related Tools" section with a link to "Artemis".

Annotations on the screenshot indicate the following steps:

- A red arrow with the number "1" points to the "Windows" link in the list.
- A red arrow with the number "2" points to the "保存(S)" (Save) button in a file dialog box that appears at the bottom of the browser window. The dialog box text reads: "ftp.sanger.ac.uk から dnaplotter.jar (7.08 MB) を開くか、または保存しますか?" (Do you want to open dnaplotter.jar (7.08 MB) from ftp.sanger.ac.uk, or save it?).
- A red arrow with the number "3" points to a file icon labeled "dnaplotter.jar" on the desktop.

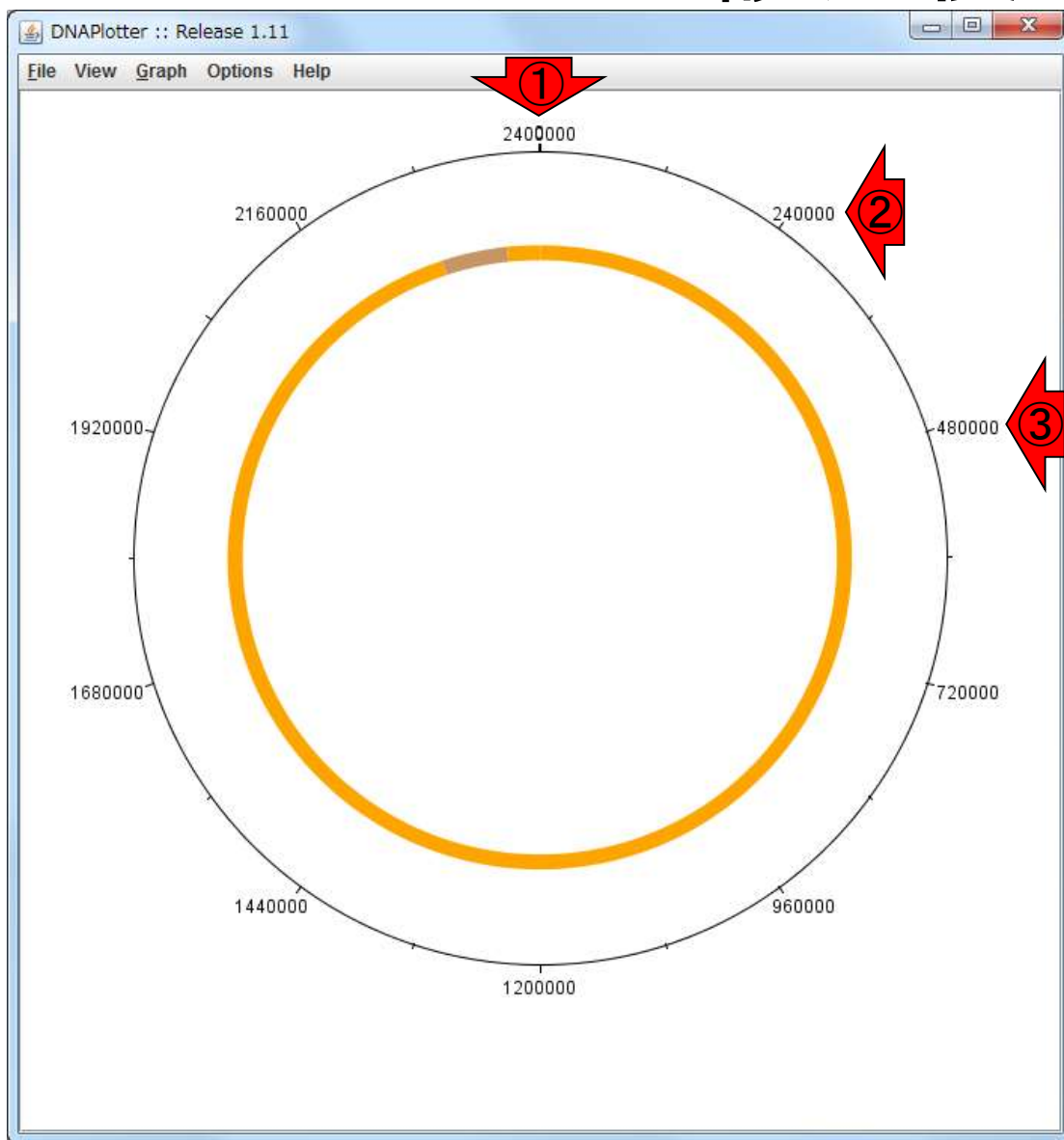
W11-4: 実行

まずはゲノム配列のFASTAファイル(genome.fna)とアノテーション情報ファイル(annotation.gff)の組合せで開く試みを行う。結論としてはFASTAファイルは読み込めるがGFFファイル読み込み時にエラーが出ました。①OK、②genome.fnaを③開く。ここではDesktop - shareフォルダ上に保存していたものを開いている。後日、“配列情報付きのgffファイル”だと読み込めることが判明したのでW19で解説しています



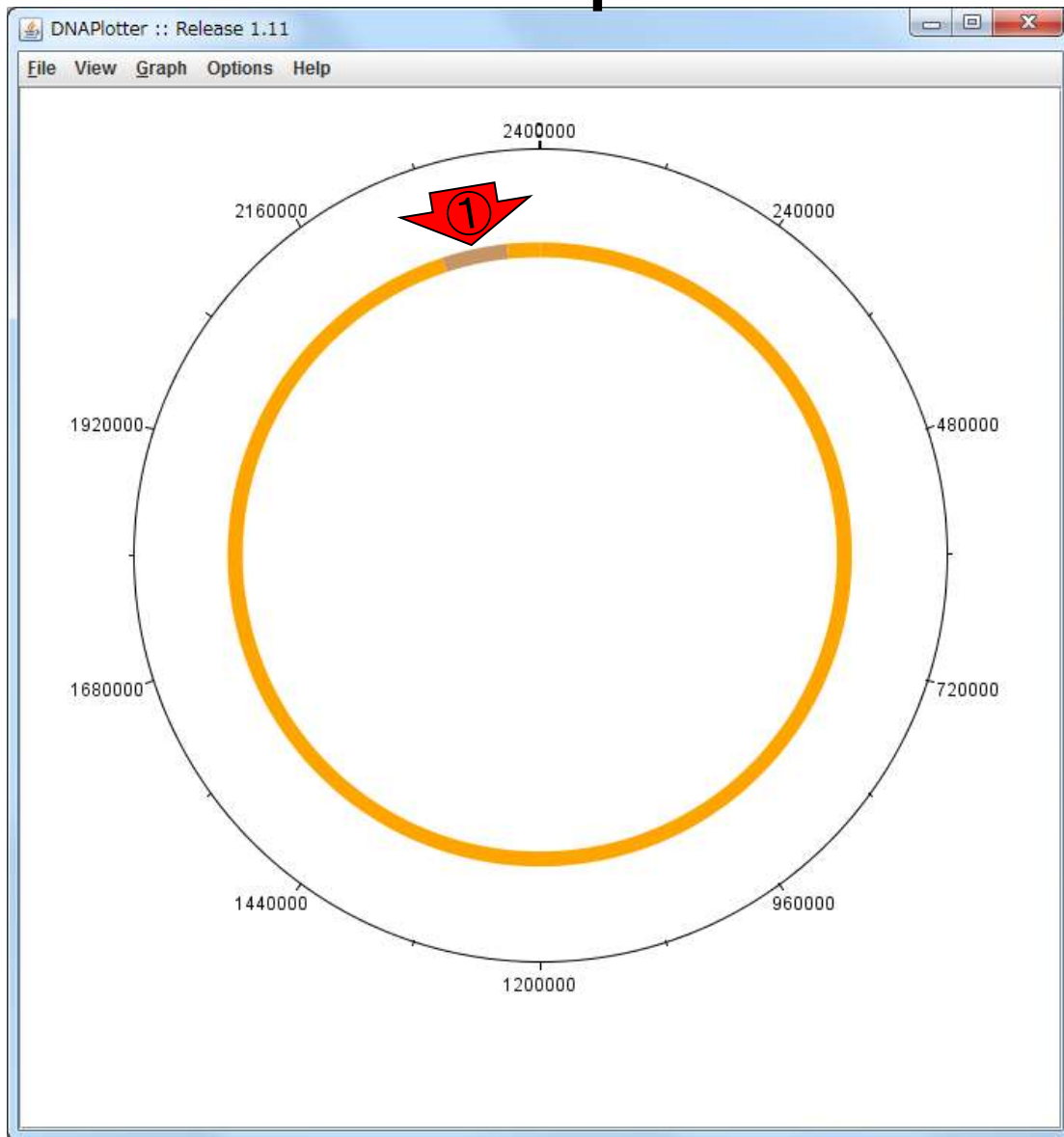
W11-5: ゲノム読込後

すぐにこんな感じのものが描画されます。このファイルは、計3配列からなり、ゲノムサイズのトータルは2,400,584 bpでした(W7-7)。それゆえ①の場所が2,400,584に近い2400000という値になっているのだろうと判断。10等分した位置を示しているようなので、②が240000、③が480000のような感じになっているのだろう



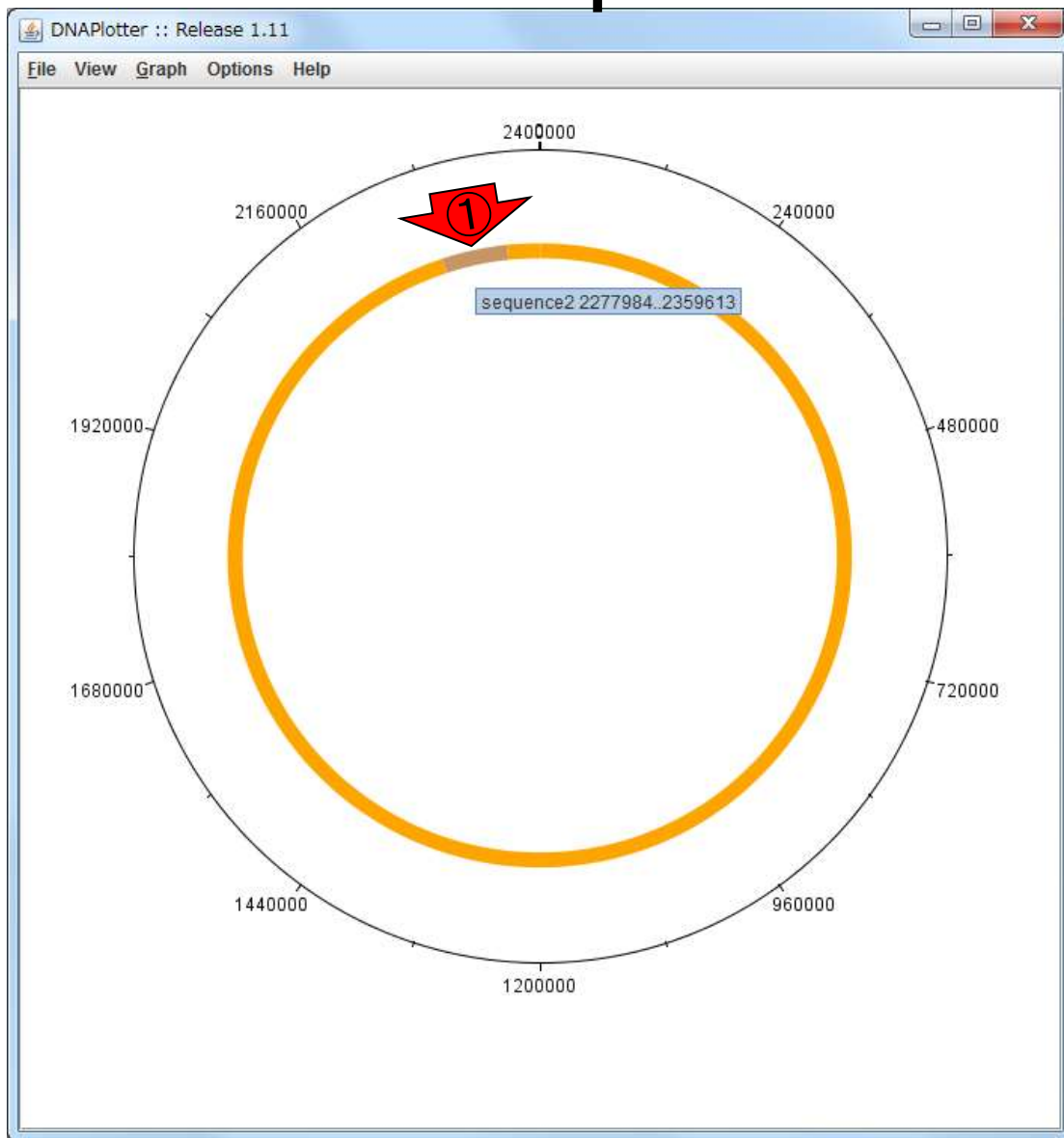
W11-6: sequence2の領域

①の部分の色の違いはなんだろうと思いつつ、このあたりにマウスオーバーすると…



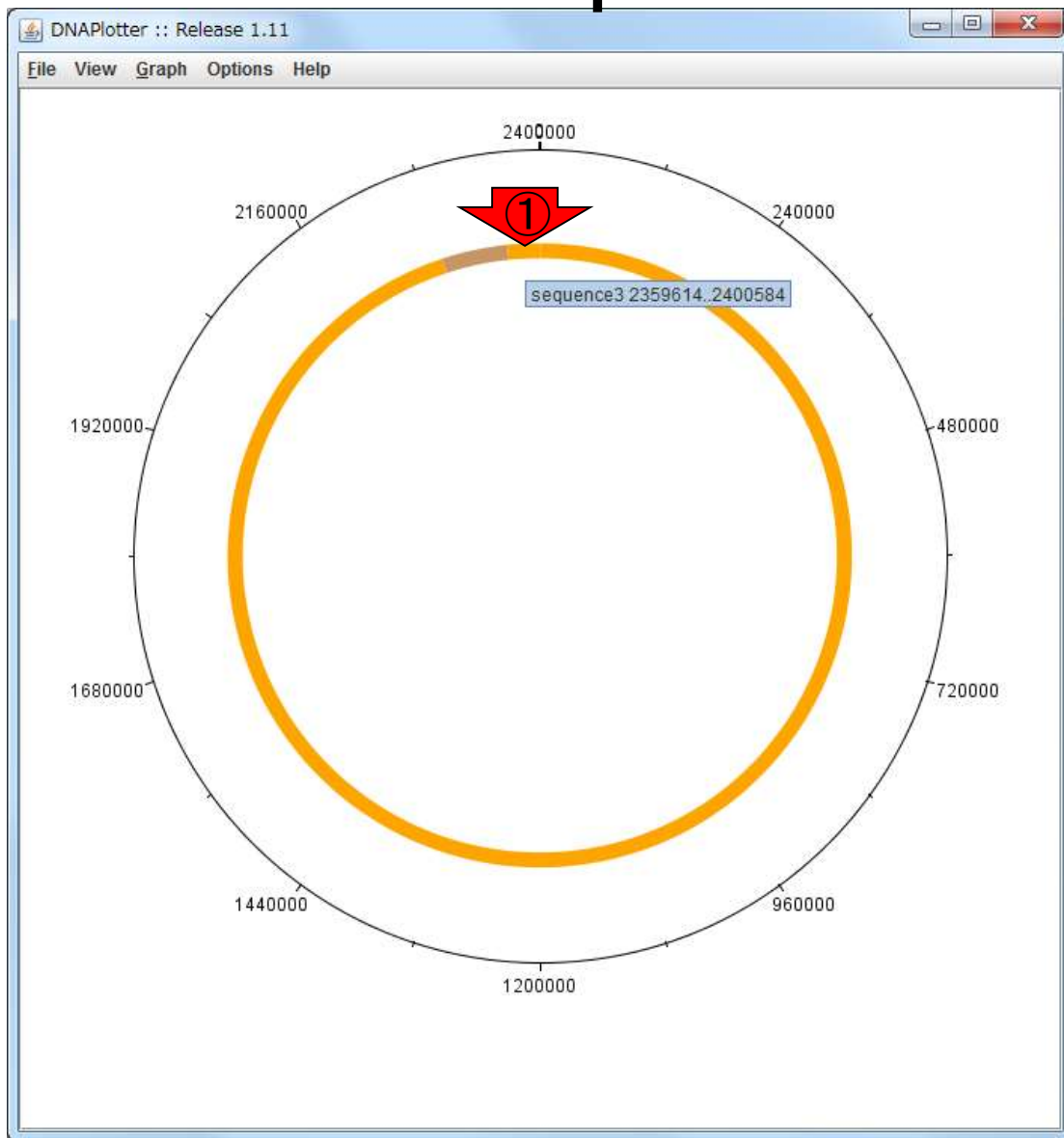
W11-6: sequence2の領域

「Sequence2 2277984..2359613」が見えます。確かに全部で3配列からなるので、色分けしているのだろうと判断。①の部分がsequence2の領域[2277984, 2359613]なのでしょう



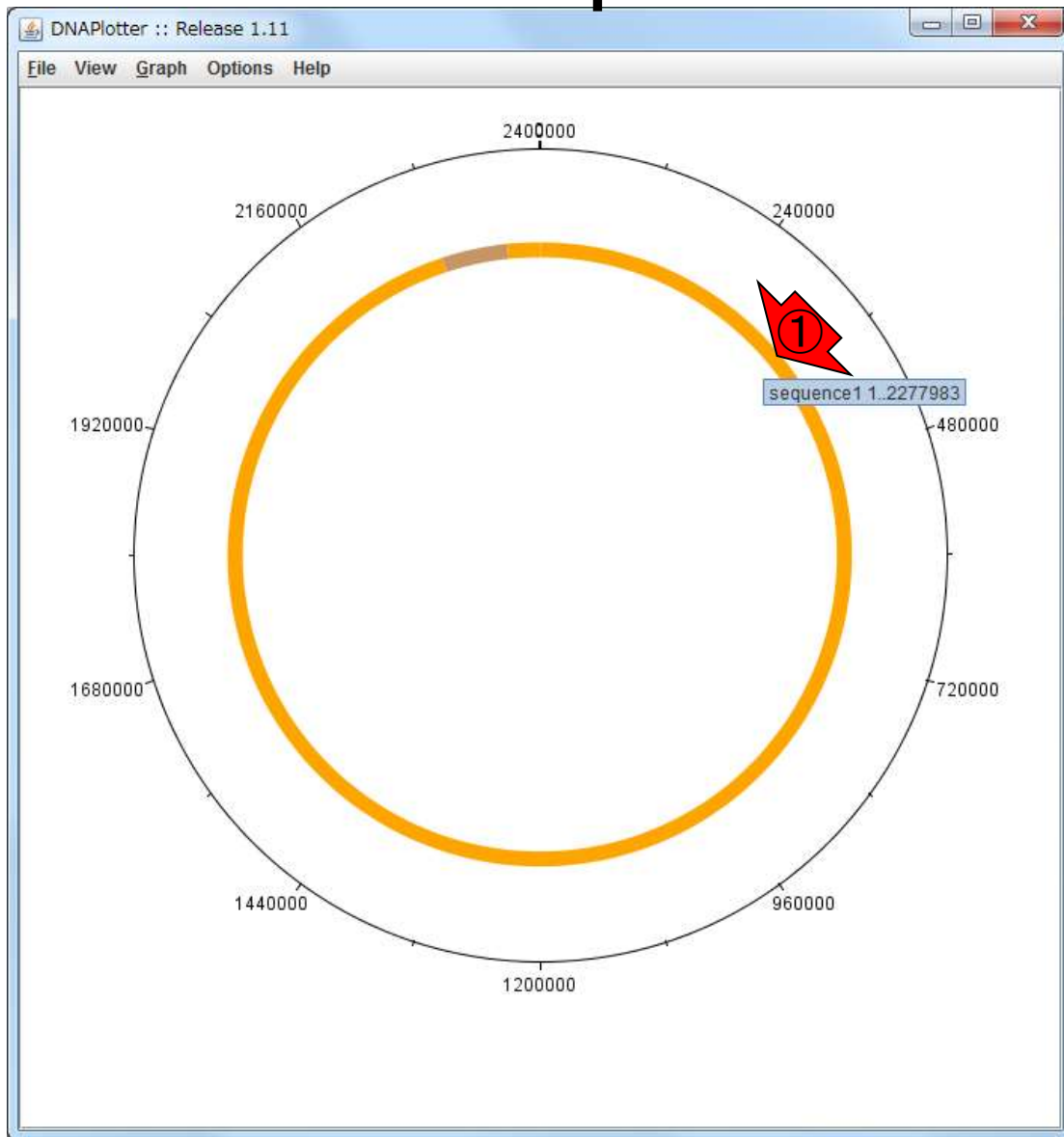
W11-6: sequence3の領域

予想通り、①のあたりをマウスオーバーすると「Sequence3 2359614..2400584」が見えます



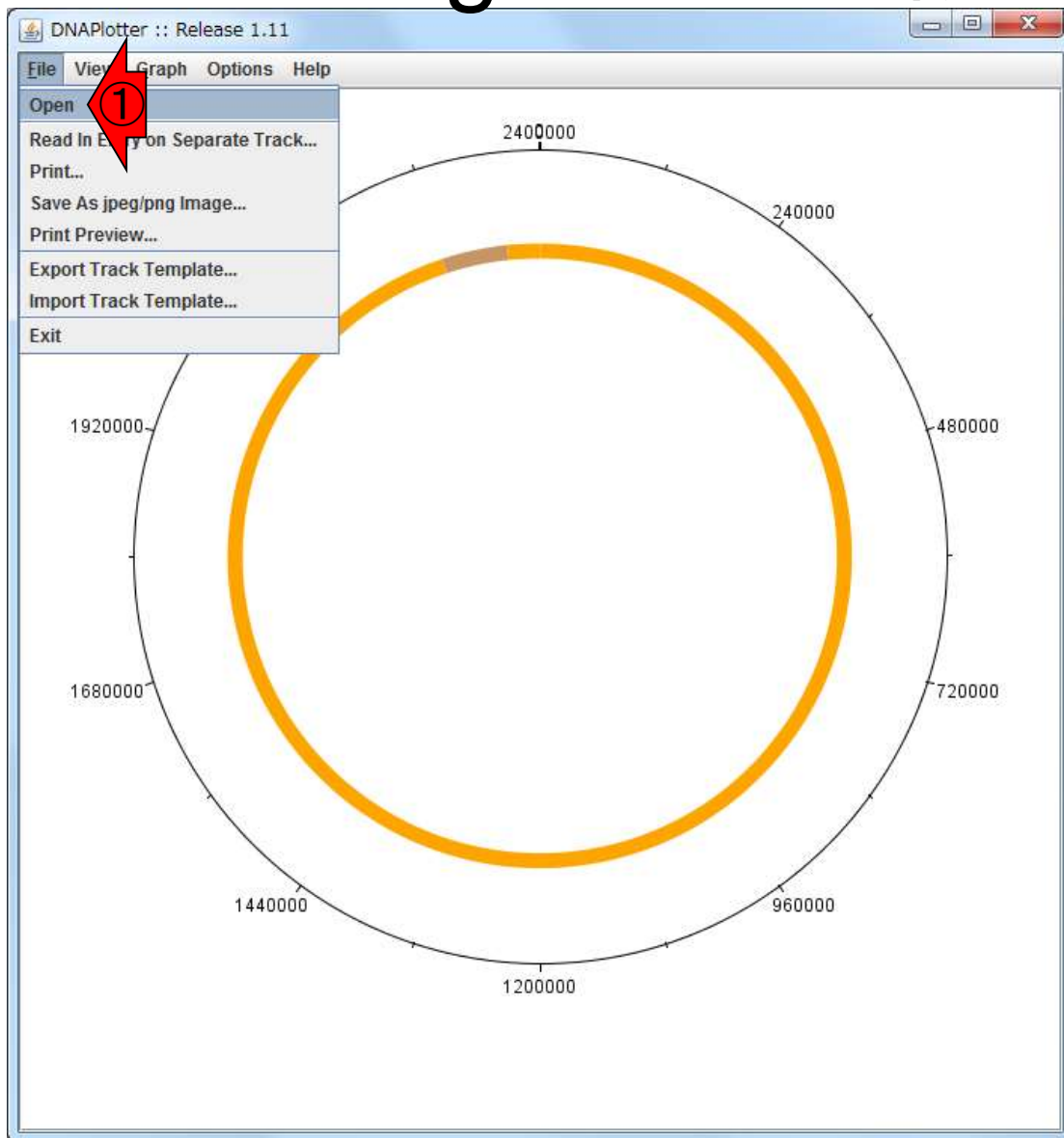
W11-6: sequence1の領域

予想通り、①のあたりをマウスオーバーすると「Sequence1 1..2277983」が見えます



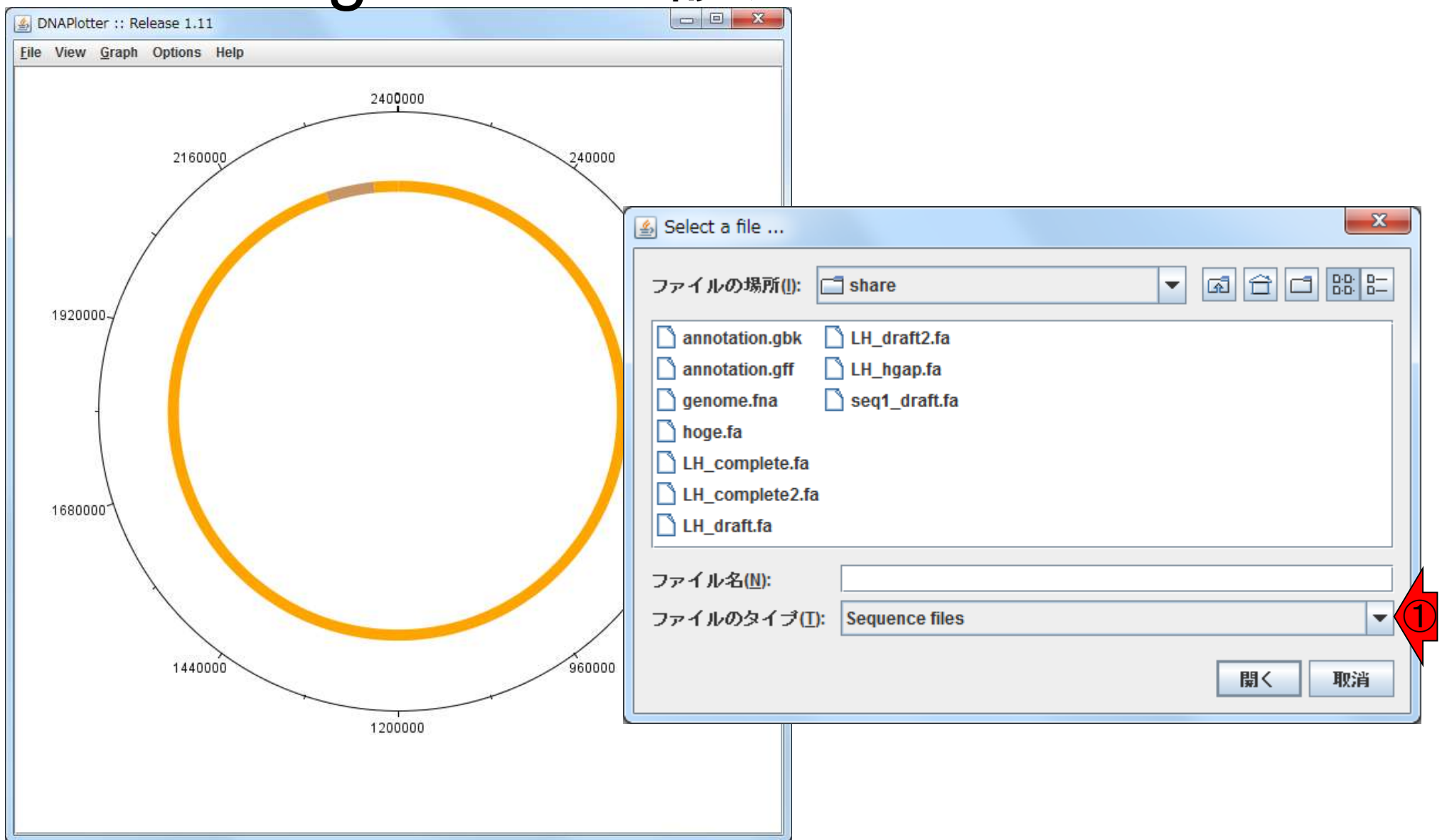
W12-1: gffファイル読込

次は、アノテーションファイル(annotation.gff)の読込。Fileメニューの①Open



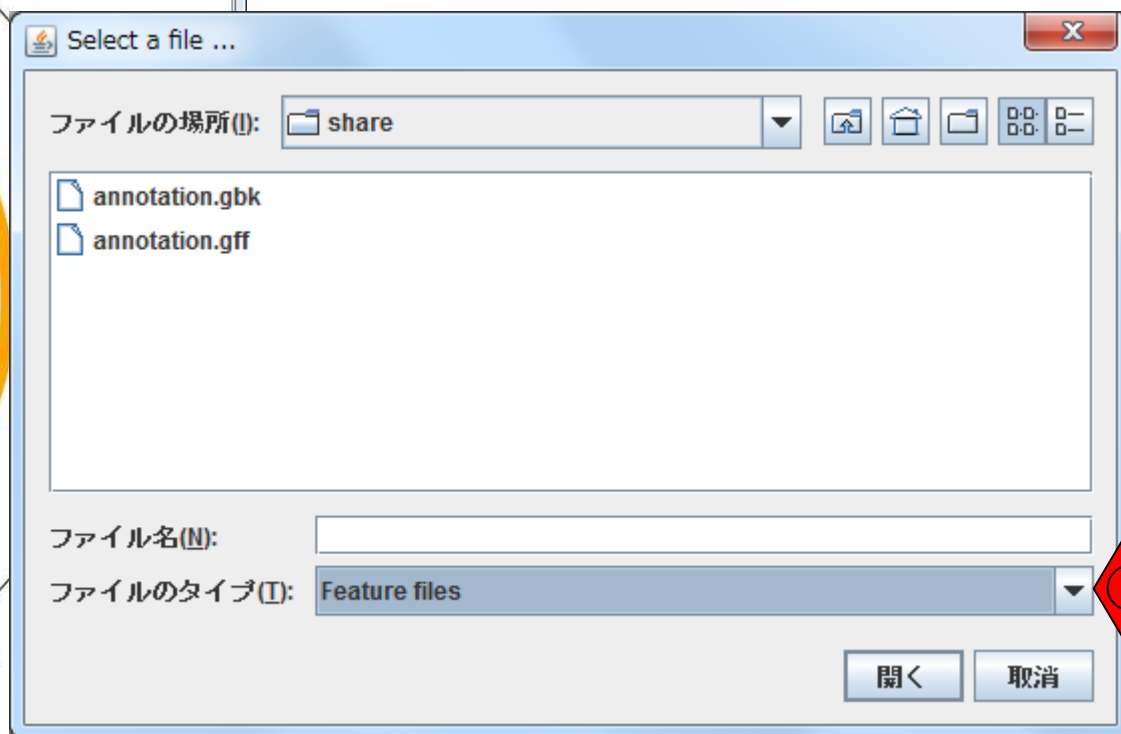
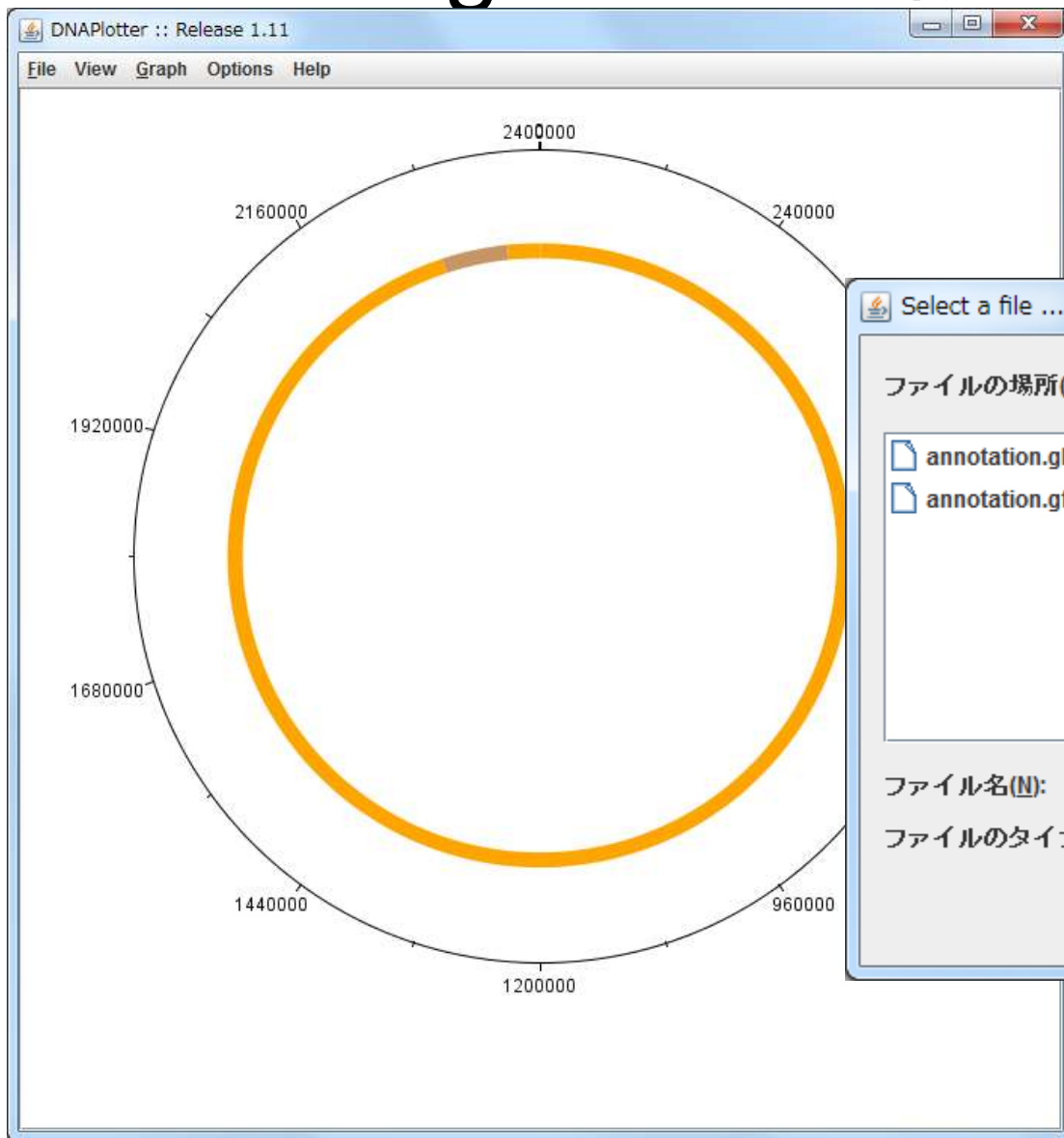
ファイルのタイプを①
Sequence filesから...

W12-1 : gffファイル読込



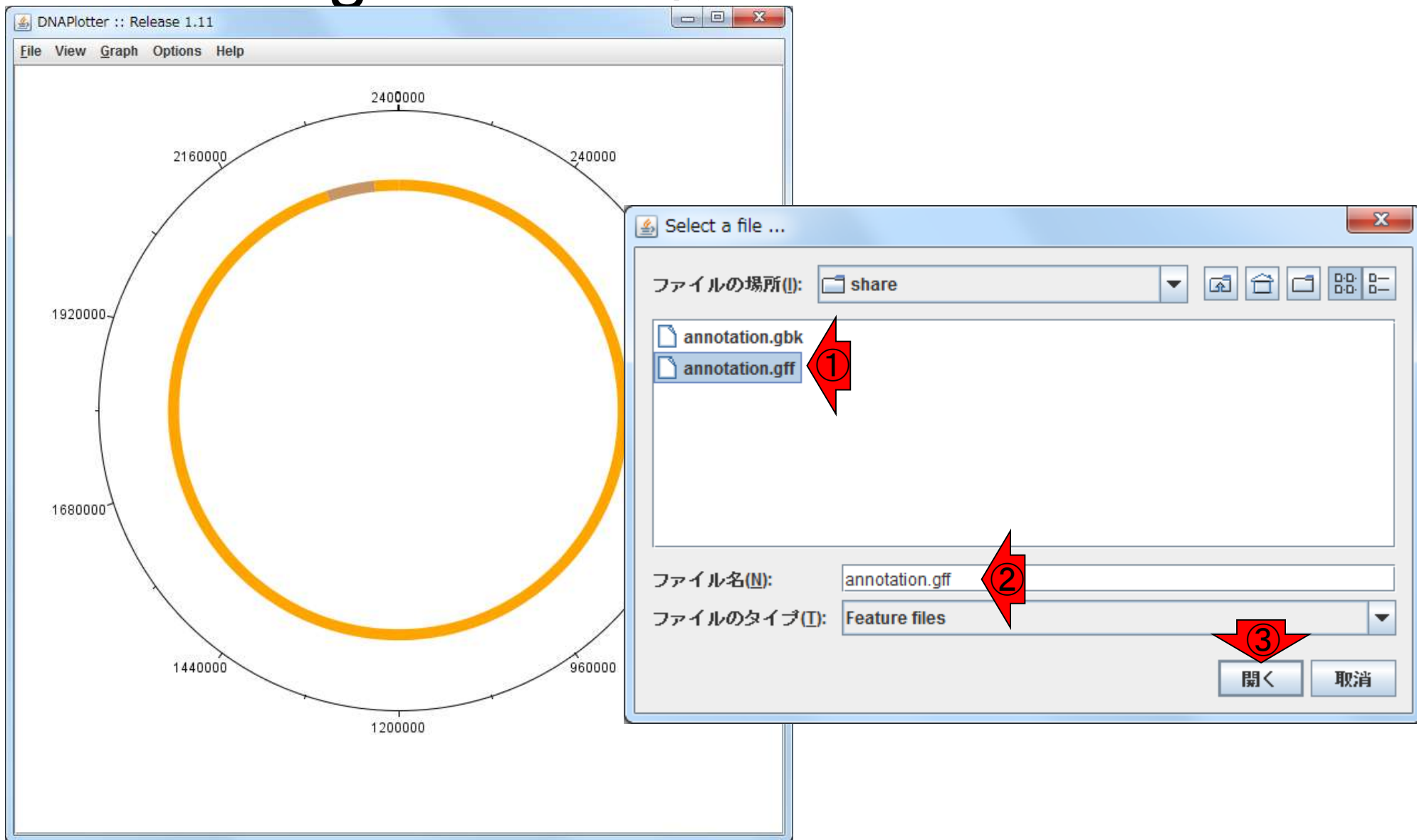
W12-1: gffファイル読込

ファイルのタイプを①Feature filesに変更すると、ターゲットファイルを絞り込める

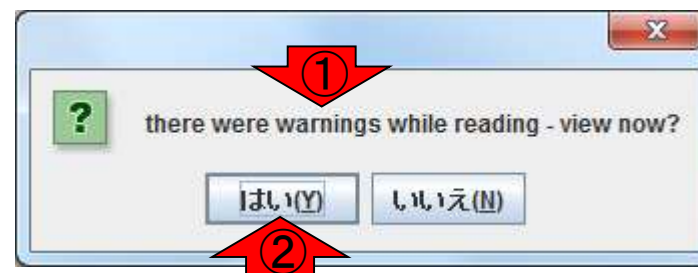
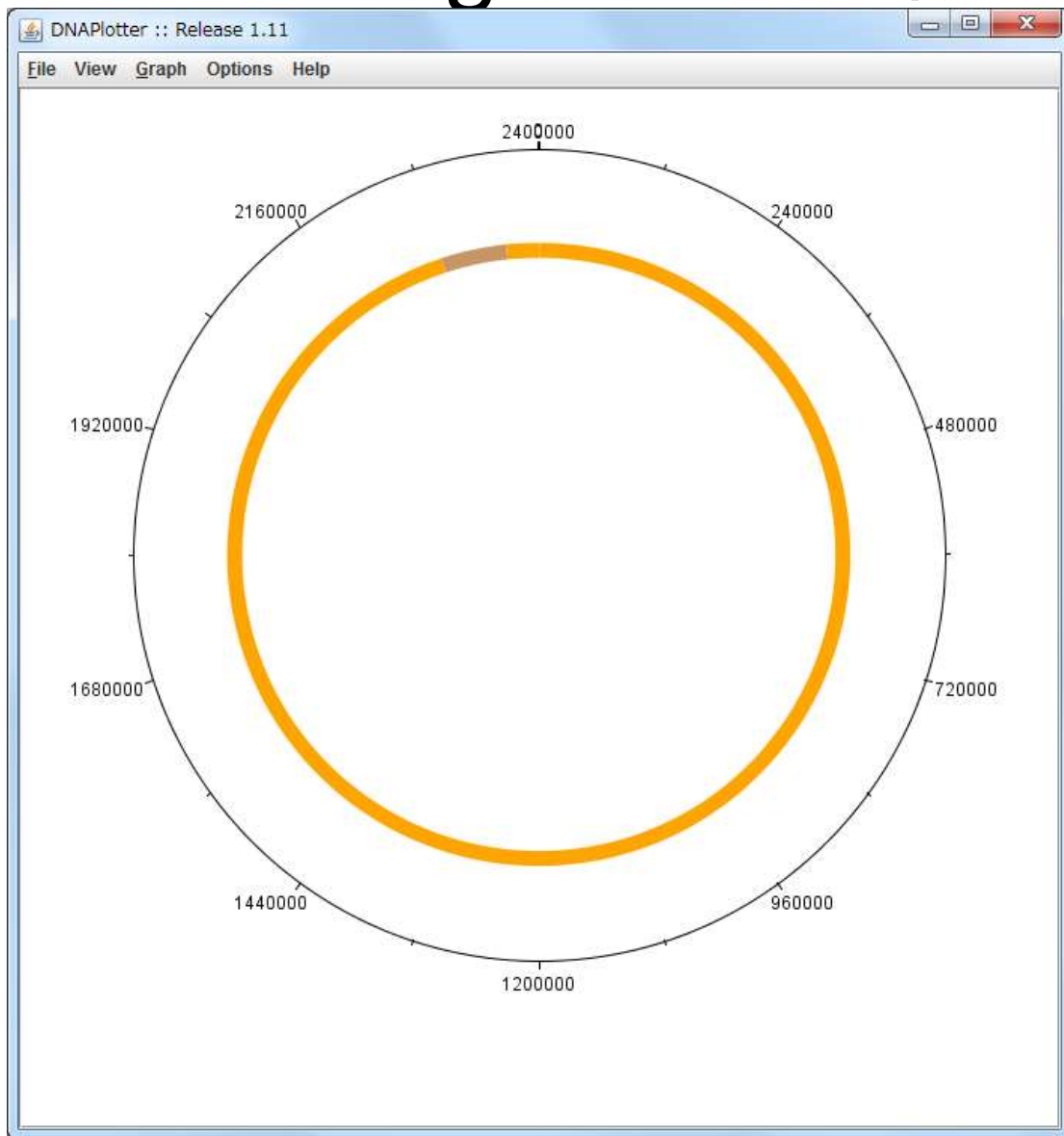


W12-1: gffファイル読込

①annotation.gffを選択して、②のところに表示させてから、③開く



W12-1: gffファイル読込



W12-1: gffファイル読

①なんだか、can'tなどのネガティブなメッセージが
②Log Viewerにて、③No sequence found!となつて
てしまいます。①に原因が書かれてはいますが、
何をどうすればいいのかさっぱりわからないので、
速やかに次の手段に移行します。④OK、⑤Close

The screenshot shows the DNAPlotter interface. On the left is a circular plot with a yellow arc and genomic coordinates (2160000, 1920000, 1680000, 1440000). In the center is a 'Log Viewer' window with the following text:

```
File  
while reading from annotation.gff: tRNA can't have inference as a qualifier  
while reading from annotation.gff: sig_peptide is not a valid key  
while reading from annotation.gff: tRNA can't have gene as a qualifier
```

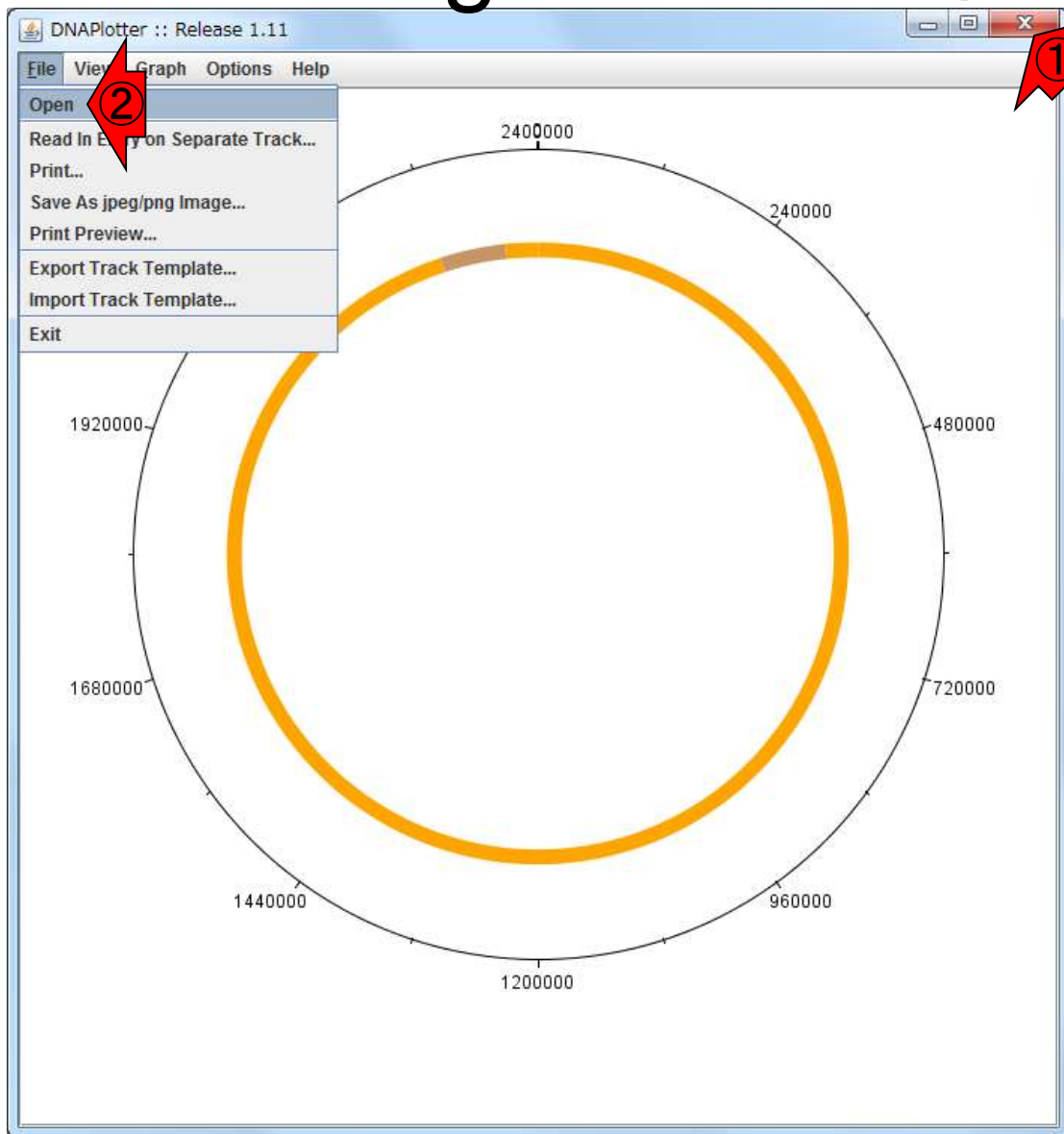
On the right is a 'Sequence Missing' dialog box with a warning icon and the text 'No sequence found!'. Below the dialog is an 'OK' button. At the bottom of the main window are 'Close' and 'Clear' buttons.

Red arrows with circled numbers point to the following elements:

- ①: A bracket grouping the error messages in the Log Viewer.
- ②: The Log Viewer window title bar.
- ③: The 'No sequence found!' text in the dialog box.
- ④: The 'OK' button in the dialog box.
- ⑤: The 'Close' button in the main window.

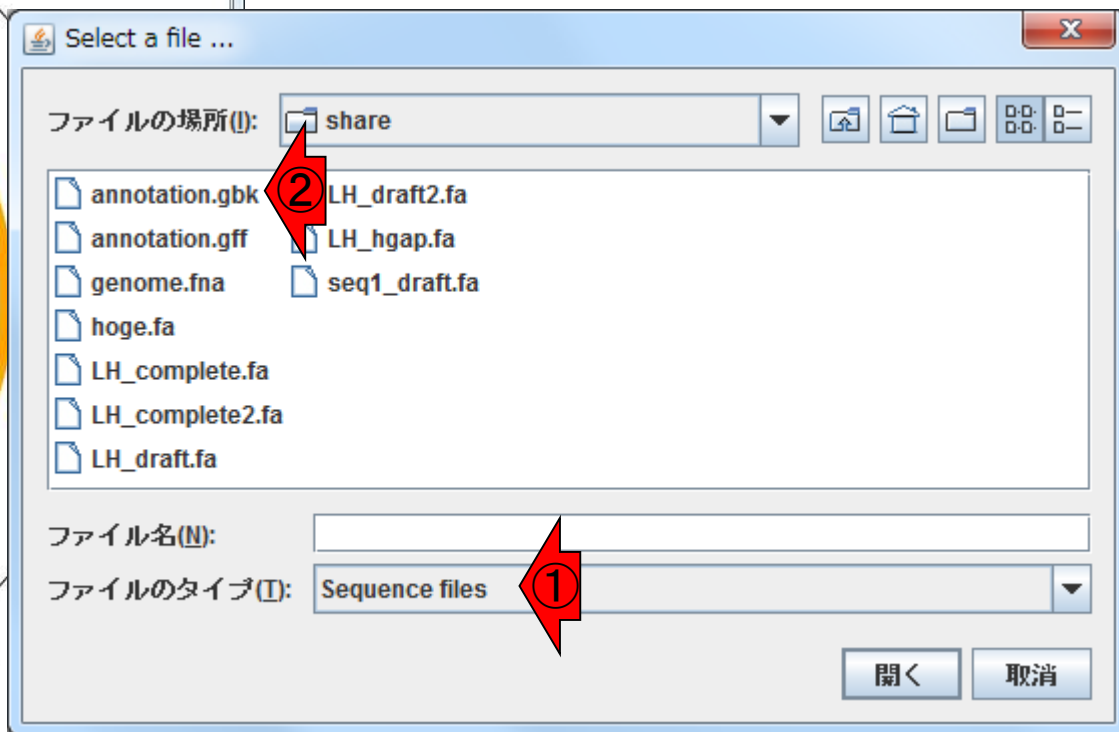
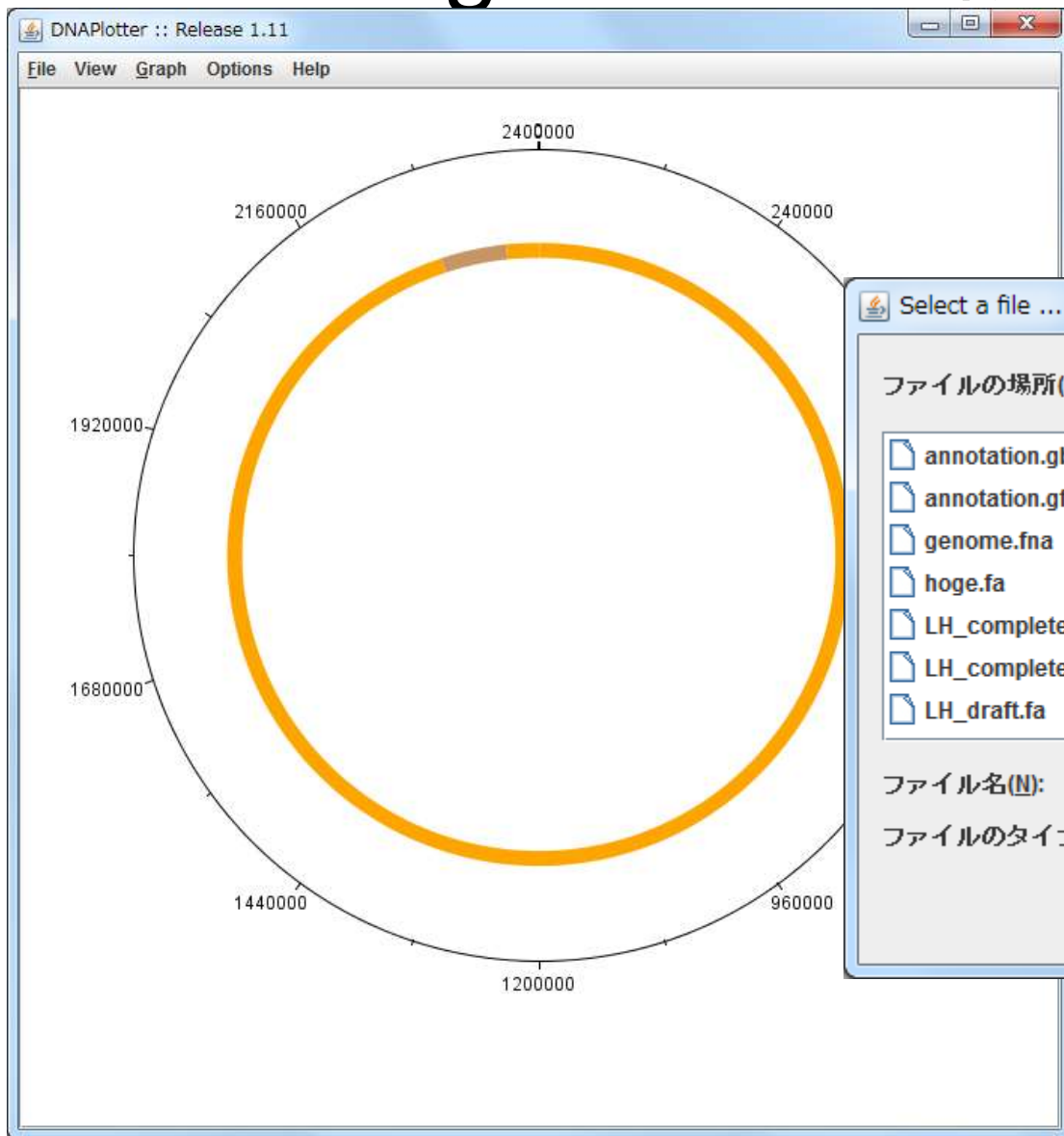
W12-2: gbkファイル読込

①一旦終了させて、再度DNAPlotterを起動しなおしてからでもよいが、FASTAファイル(*genome.fna*)を読み込んだ状態のままでもよいし、Fileメニューの② Openでもよい。ここでは後者で説明

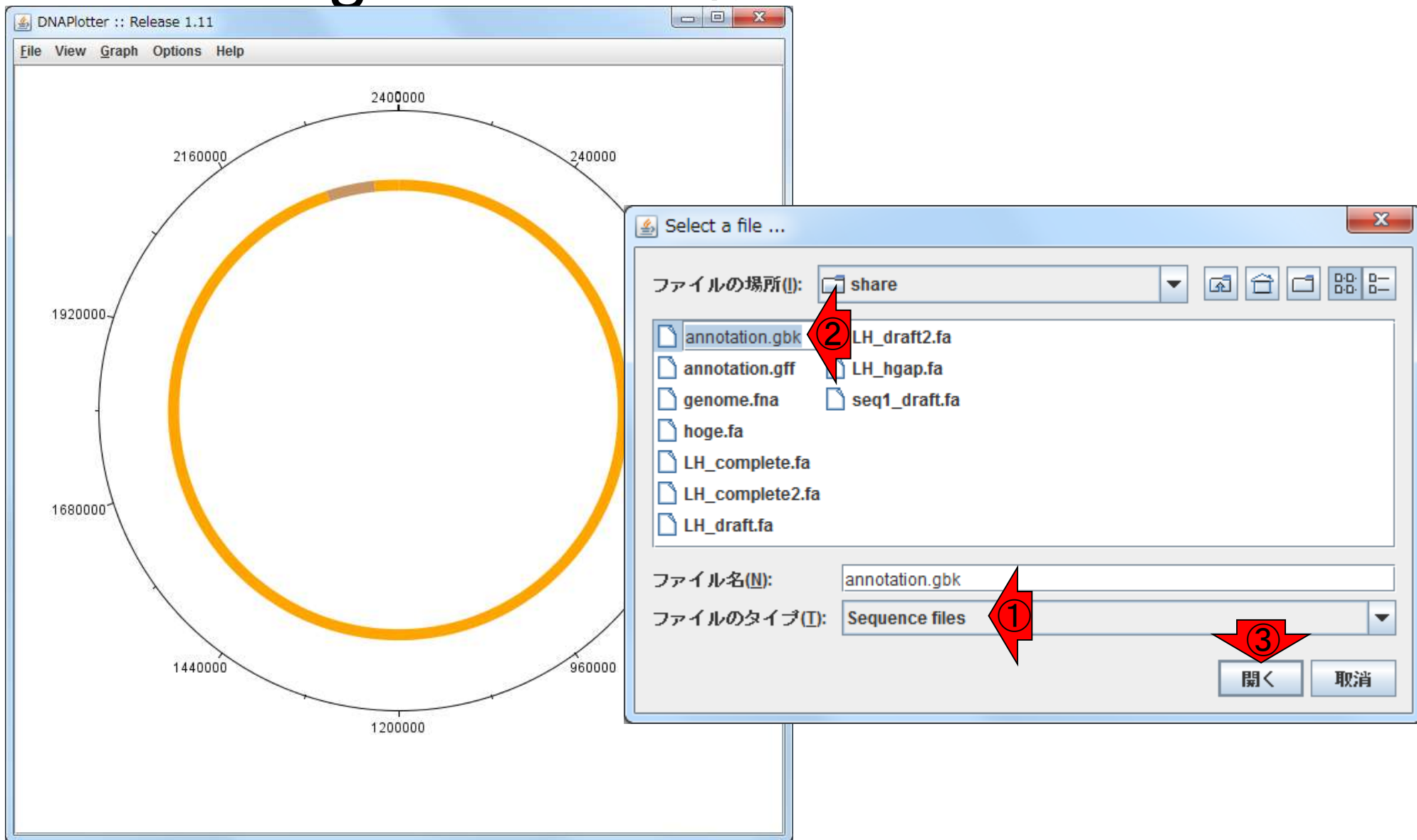


W12-2: gbkファイル読込

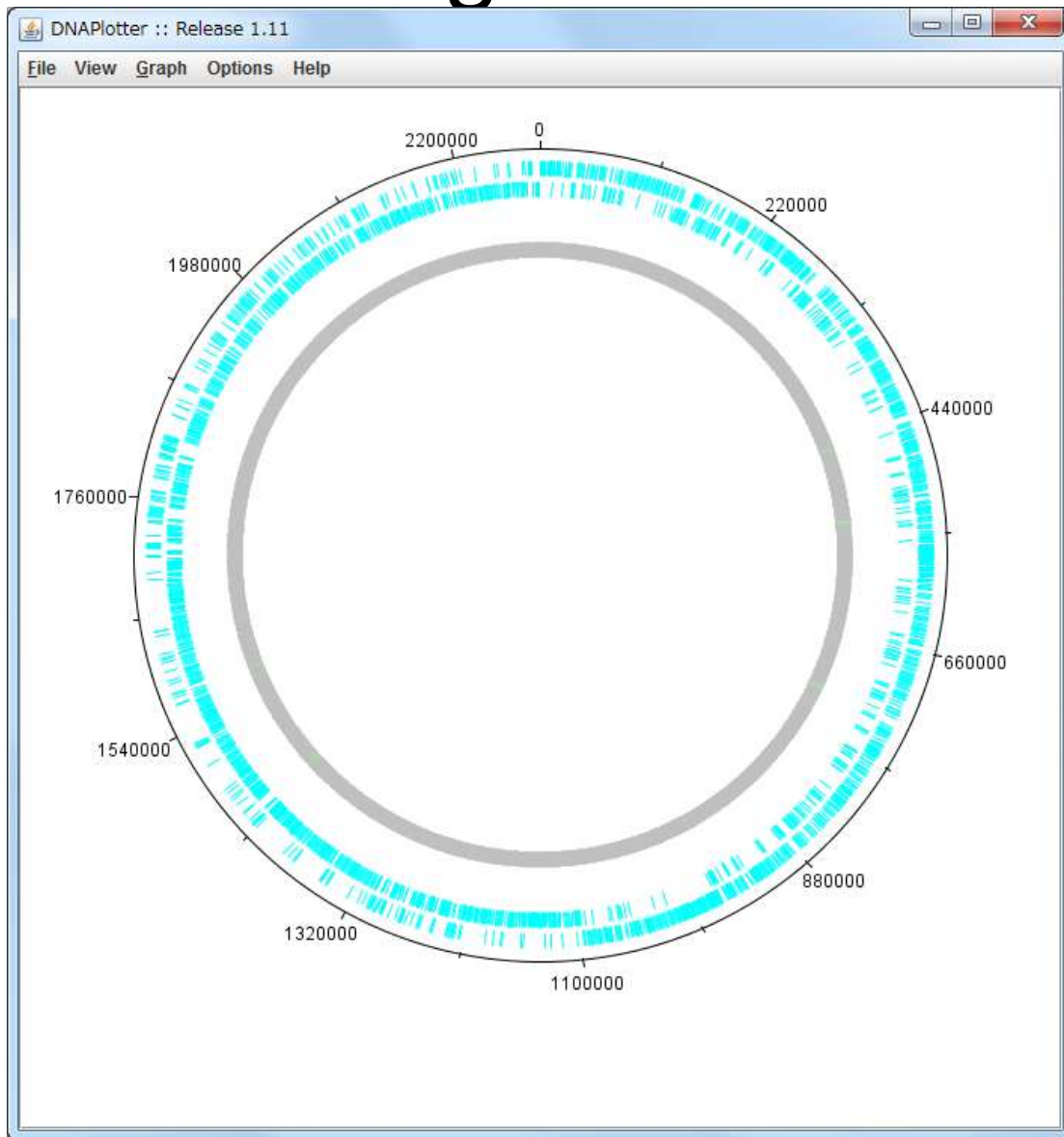
①ファイルのタイプがSequence filesのままでよいので、



W12-2: gbkファイル読込

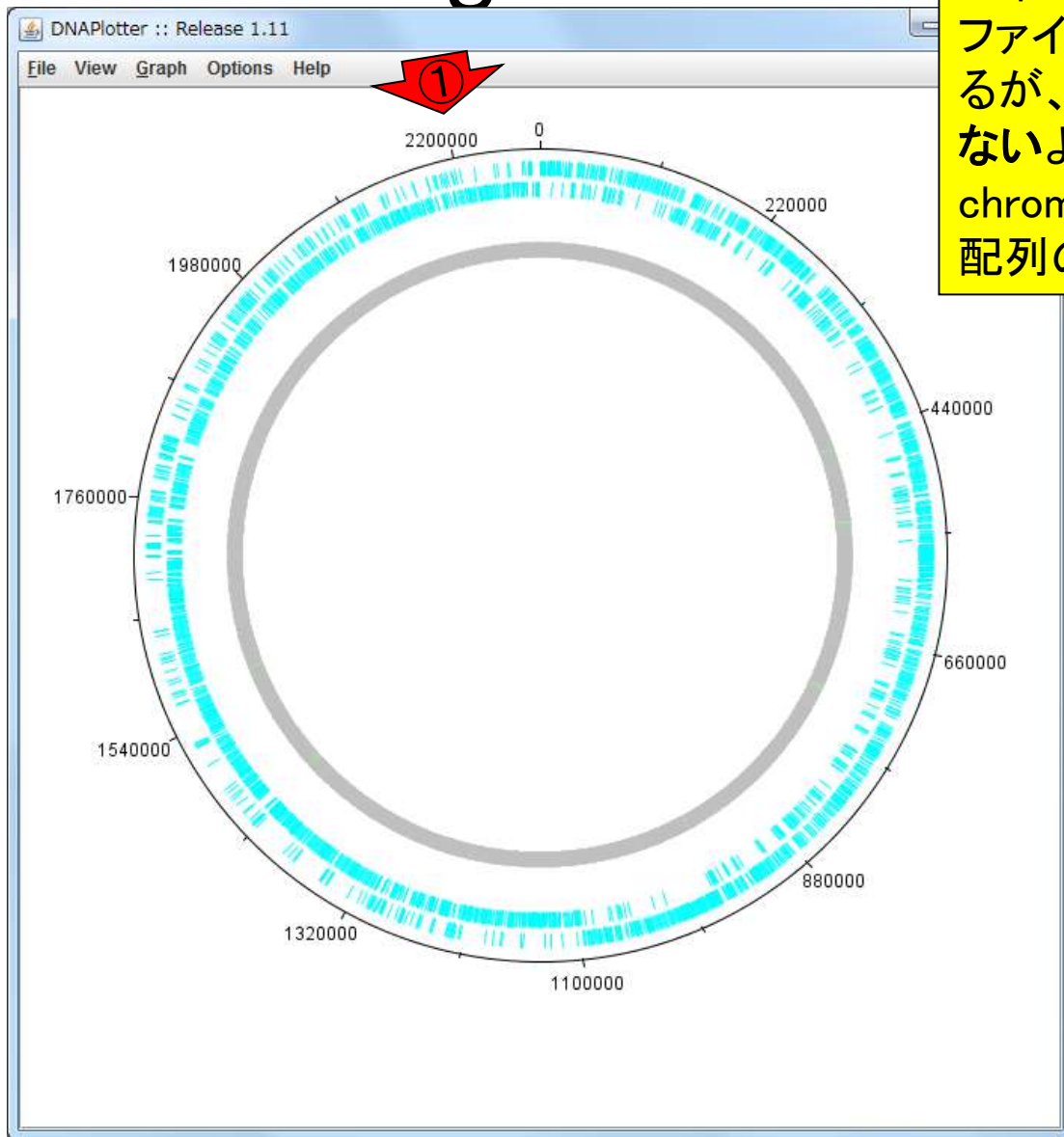


W12-3: gbk読込後



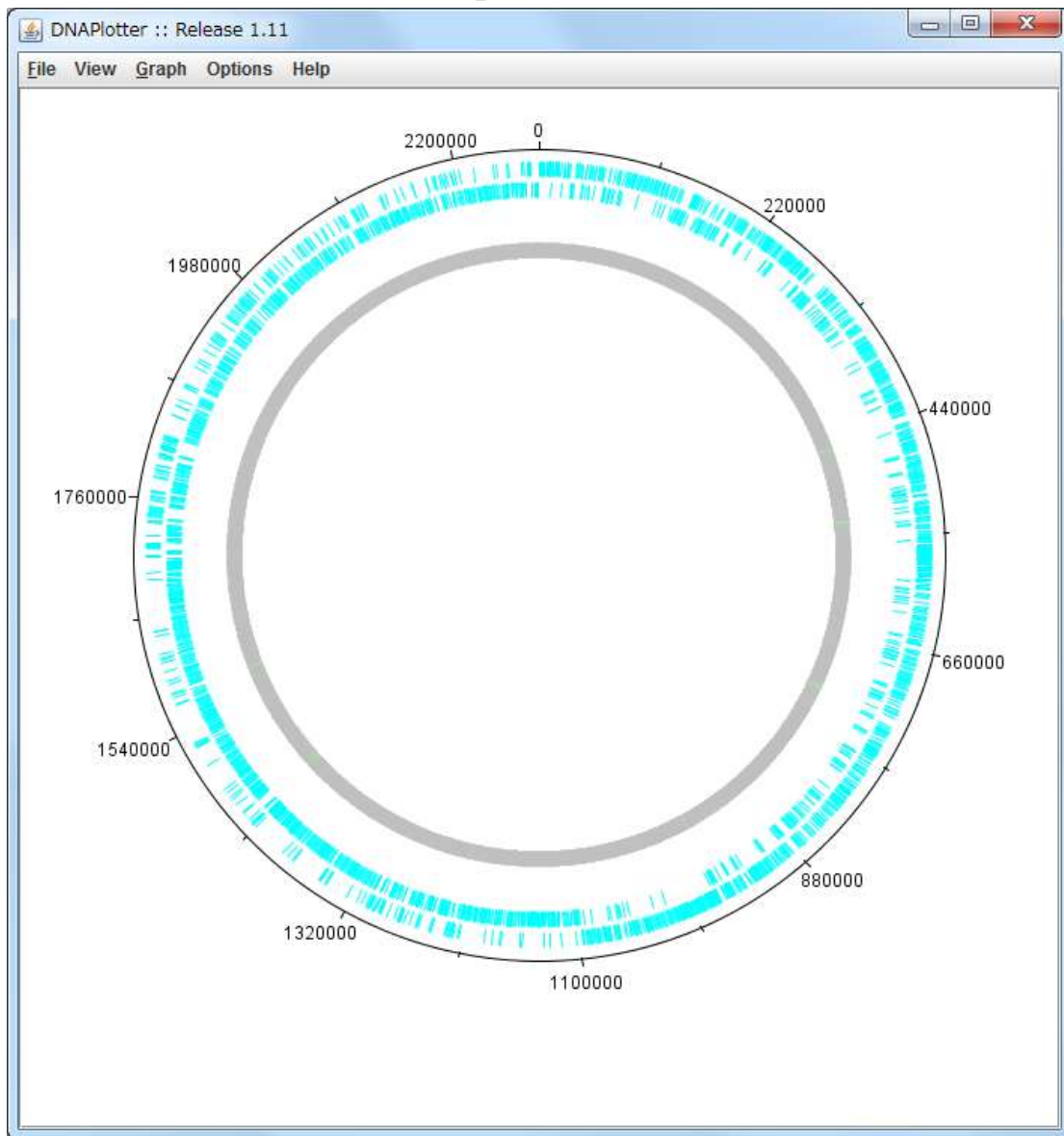
W12-3: gbk読込後

①このあたりを見ればわかりますが、sequence1の情報のみしか読み込んでいないことがわかります。sequence1の配列長は2,277,983 bpだからです。入力ファイル(annotation.gbk)自体は3配列分の情報があるが、DNAPlotterは最初の配列のみしか読み込まないようです。この場合は、sequence1という名のchromosome配列のみです。とりあえずchromosome配列のみ眺めたいのでこのまま話を進める

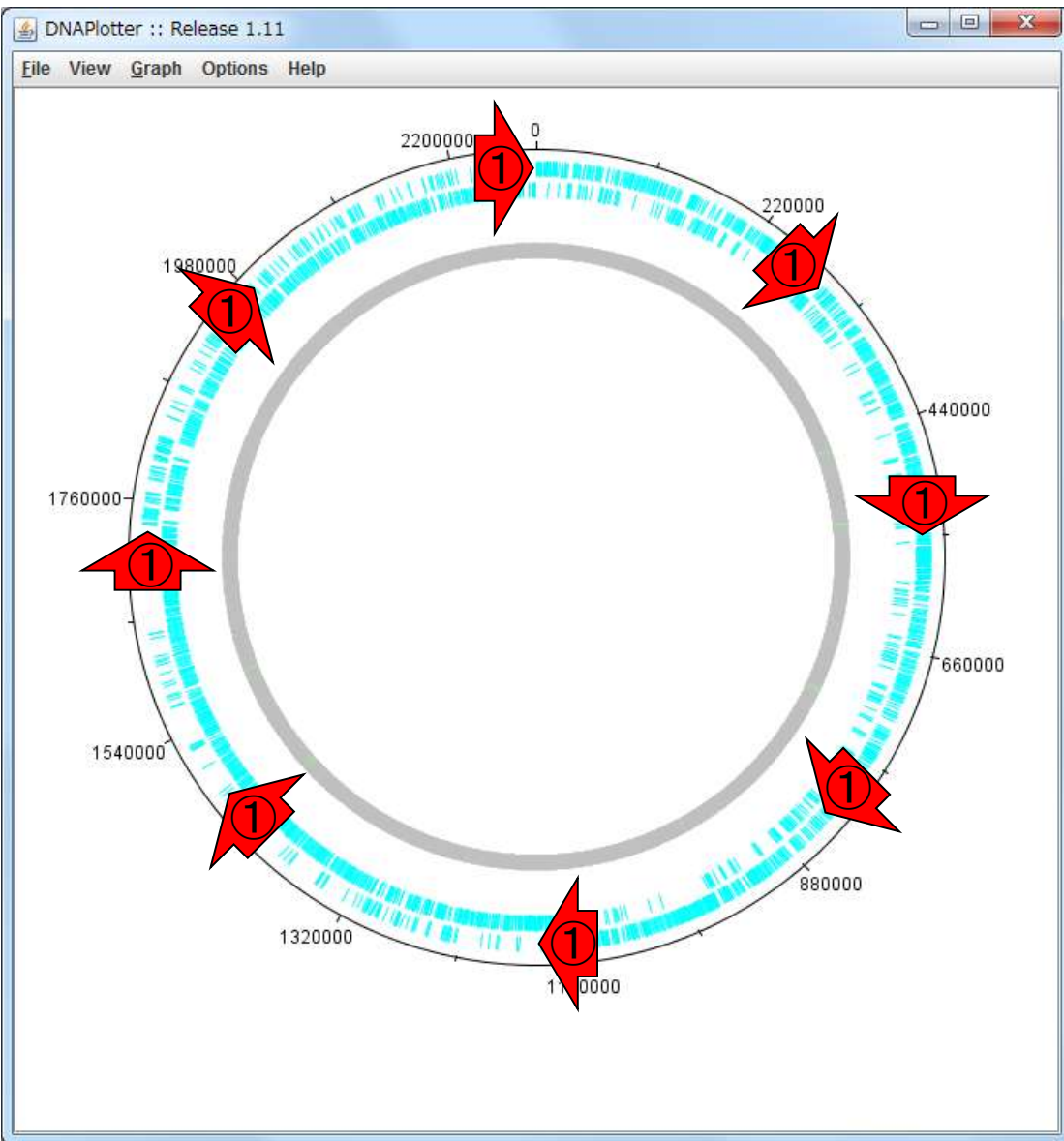


W12-4: 解説

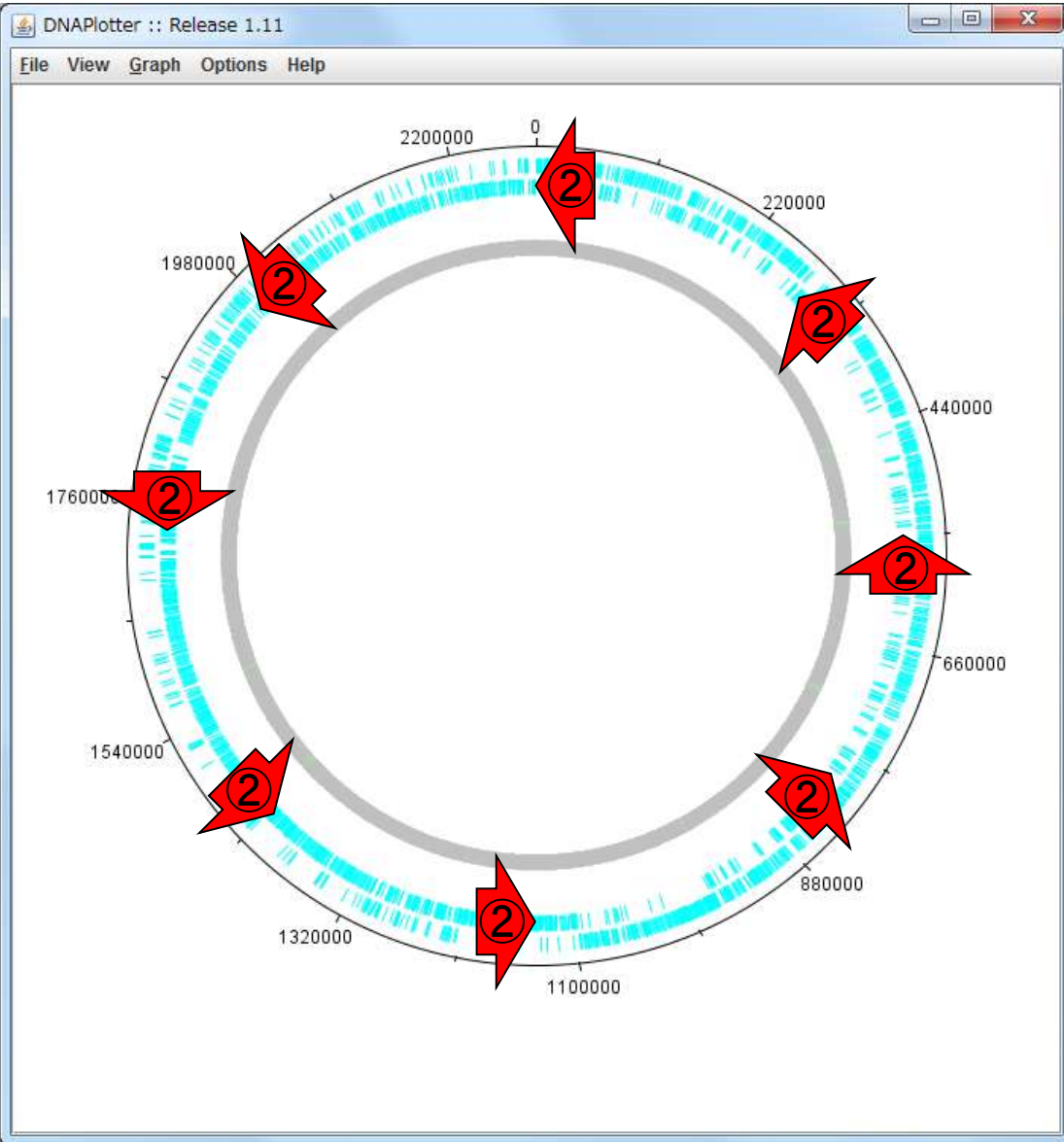
一つ一つ解説していきます。まず、水色の2つの輪っかが見られます



W12-4: 解説

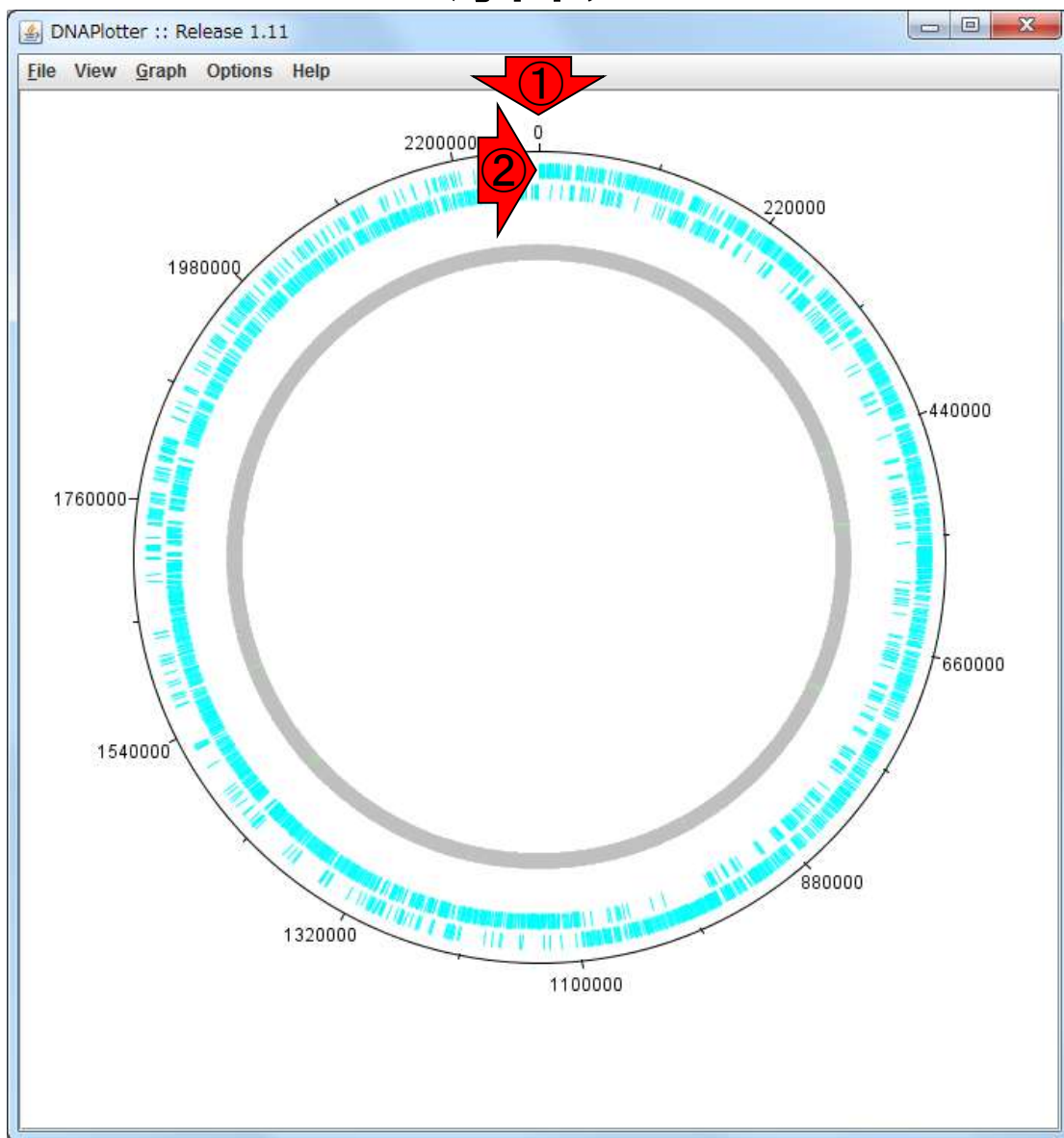


W12-4: 解説



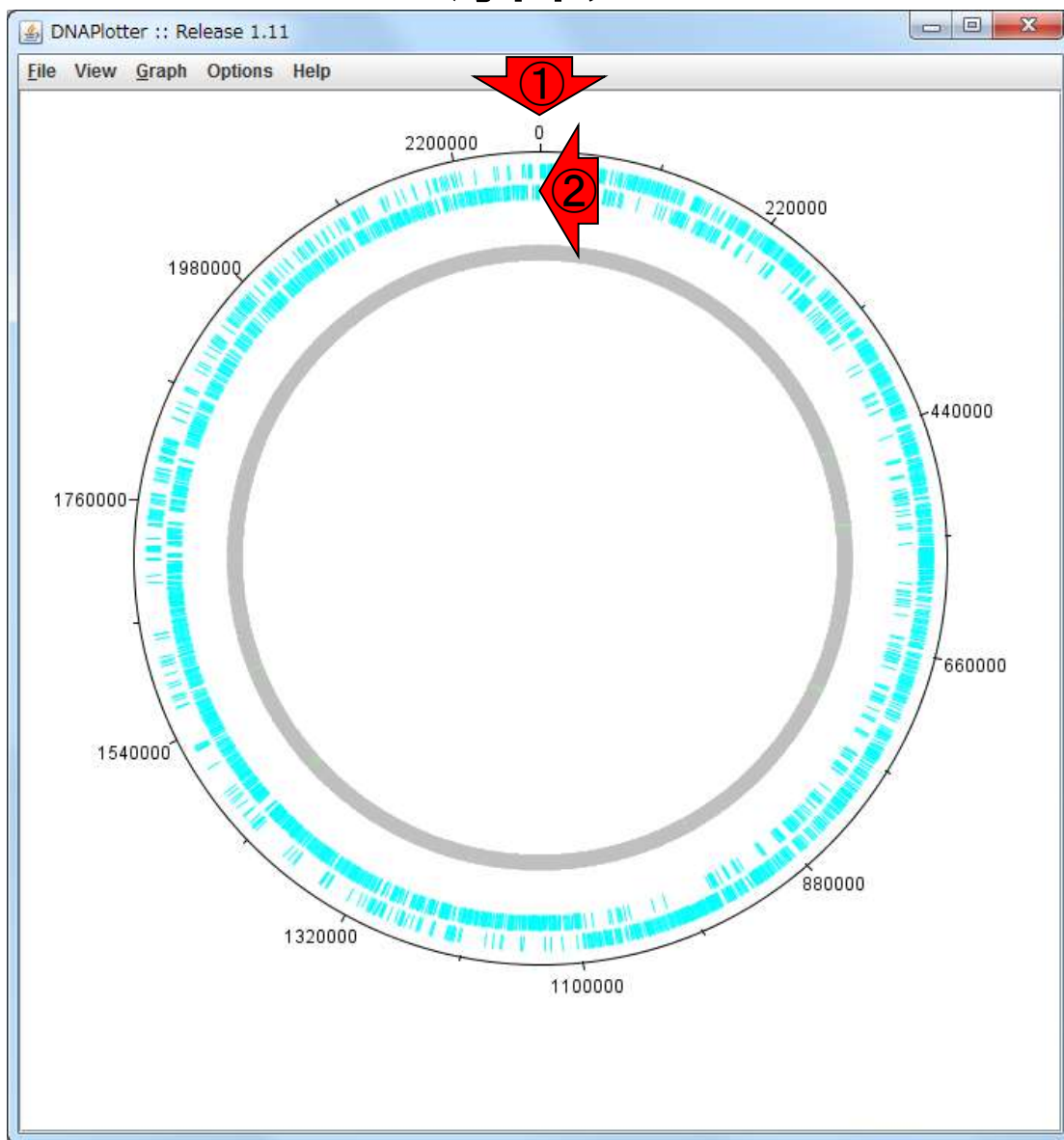
W12-4: 解説

この配列は複製開始点が①の場所になるように回転させたものでした(W4)。したがって、複製開始点が開いて、②時計回りに複製が進む場合は順鎖がリーディング鎖



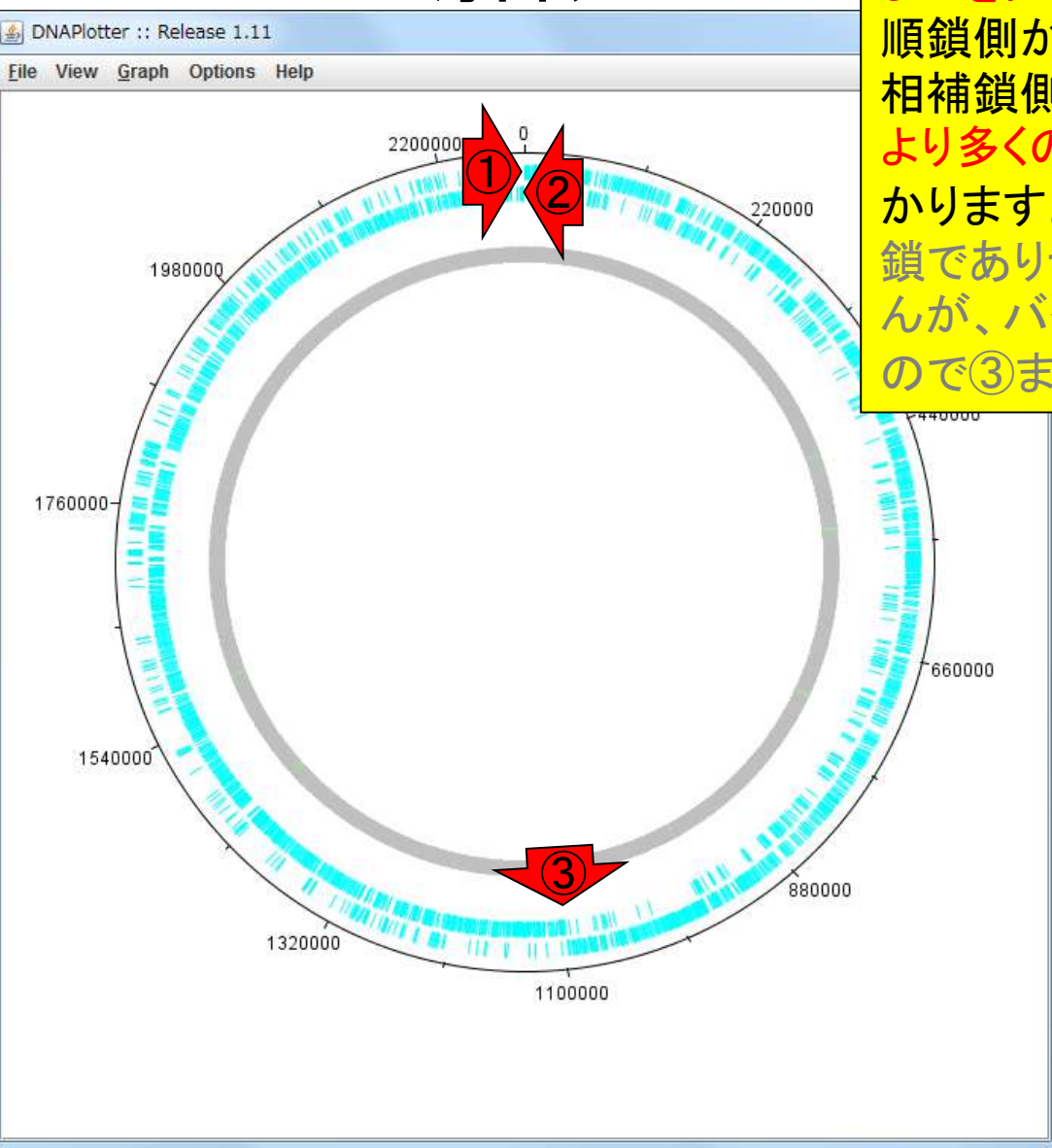
W12-4: 解説

②反時計回りに複製が進む場合は
相補鎖側がリーディング鎖になります



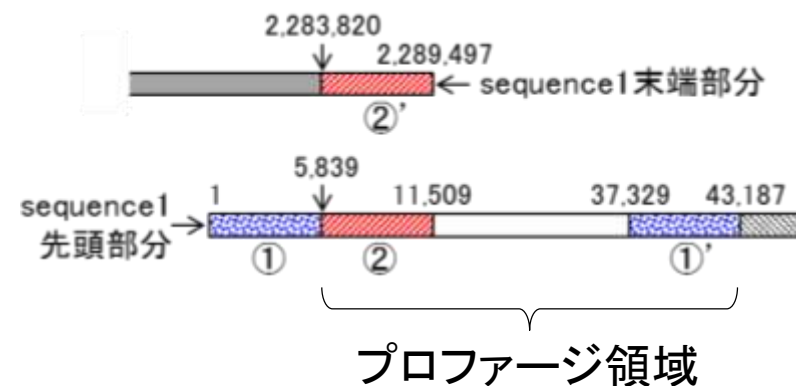
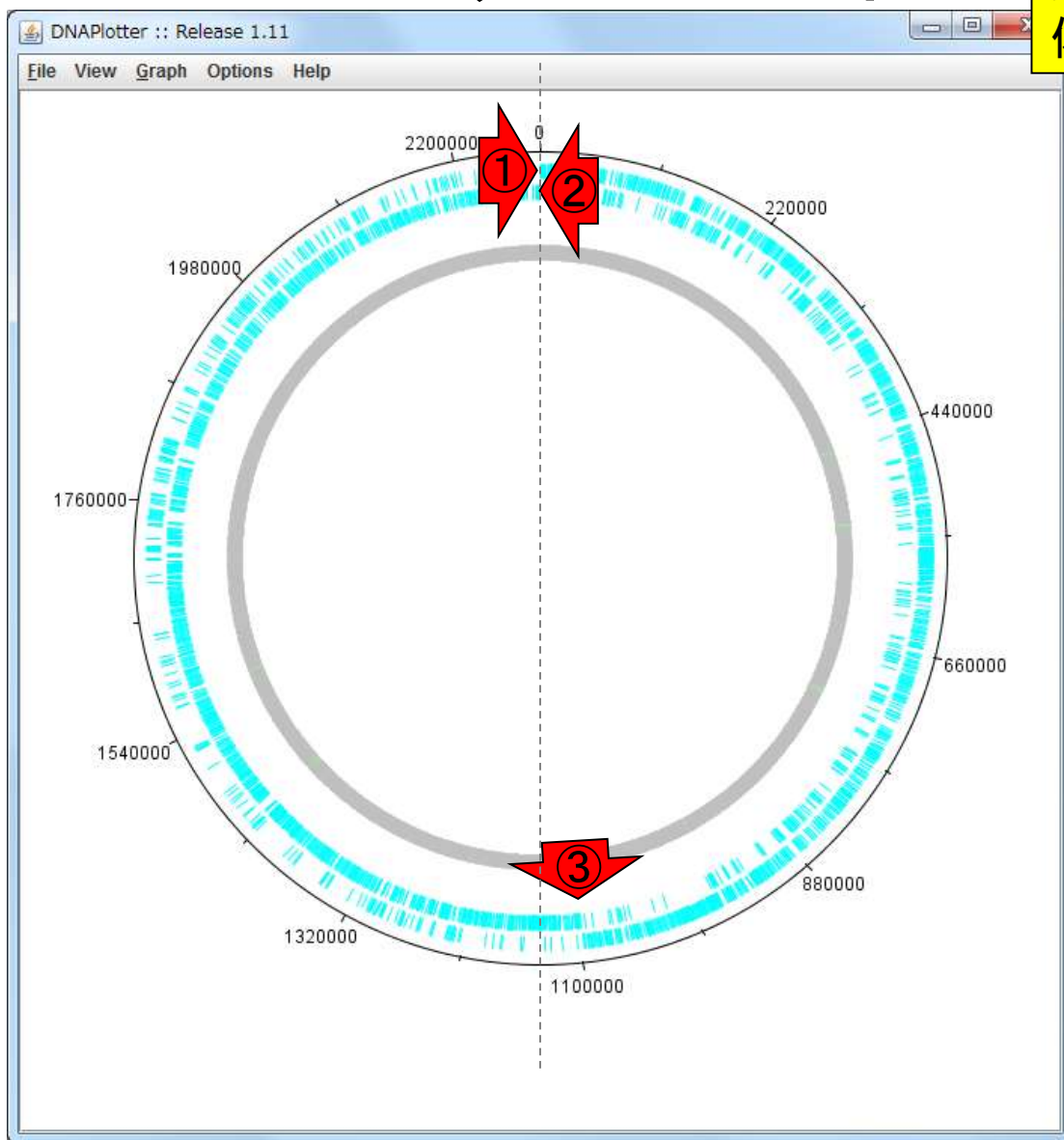
W12-4: 解説

実際の複製では、①時計回りと②反時計回りの複製が同時に進むので、③のあたりが分かれ目になります。③までを区切りとして考えると、①時計回りに複製が進む順鎖側がリーディング鎖、②反時計回りに複製が進む相補鎖側がリーディング鎖となります。リーディング鎖により多くの遺伝子(CDS)が存在する傾向がはっきりとわかります。最後までぐるっと一周すれば全部リーディング鎖でありラギング鎖じゃないかと思われるかもしれませんが、バクテリアの場合は複製開始点が1か所のみなので③までを区切りとして考えるのでよいと思います



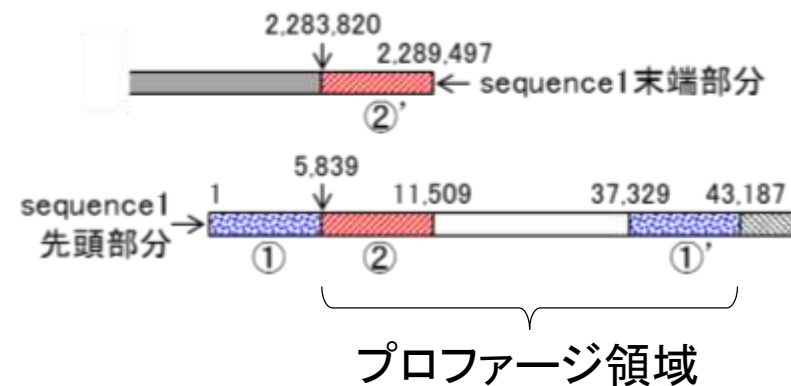
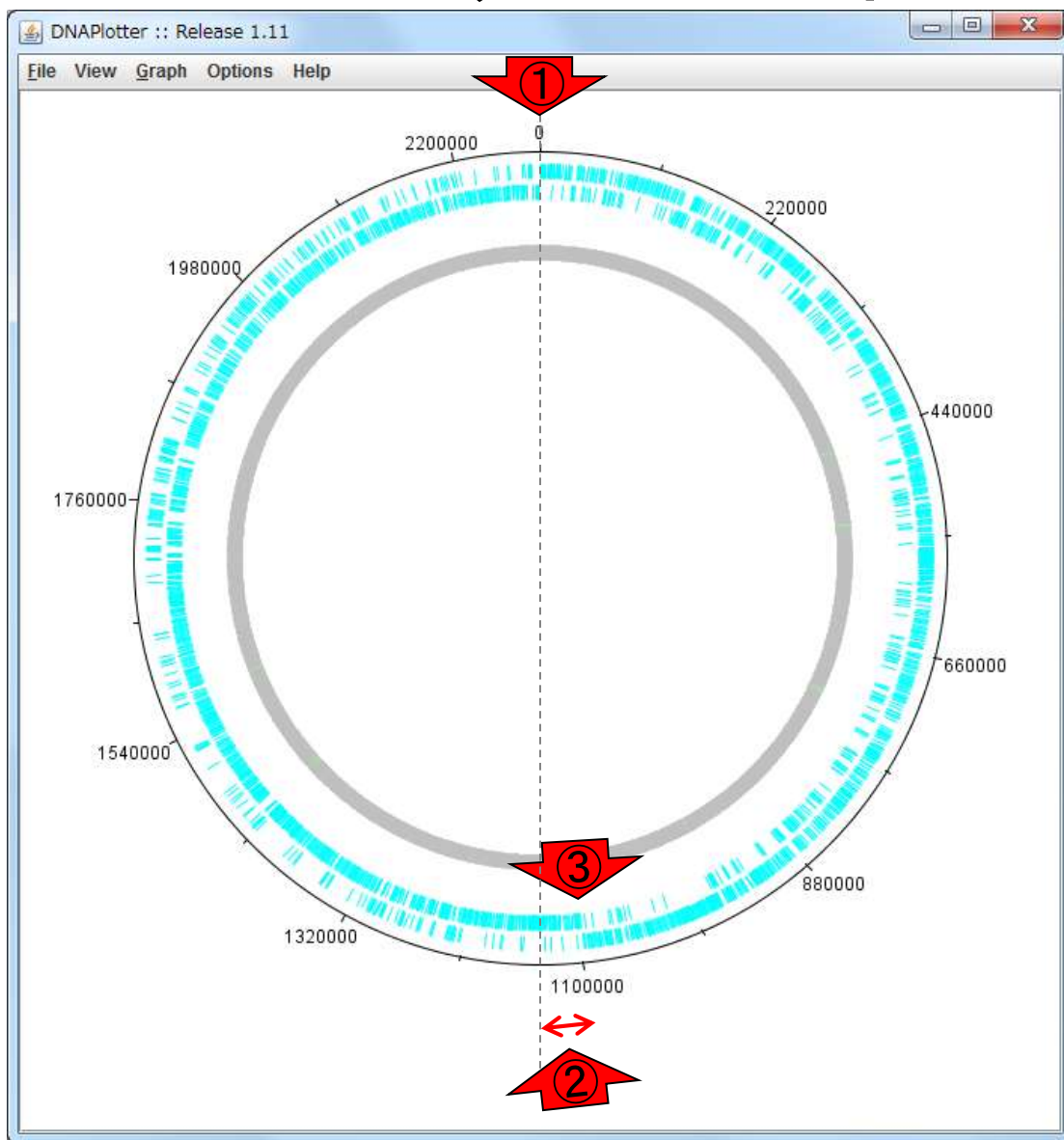
W12-5: ずれの理由

③の位置が若干右にずれているのは、第8回W10あたりでも議論したプロファージのようなごく最近染色体に組み込まれた領域や、逆に染色体から失われた領域の影響によると思われます



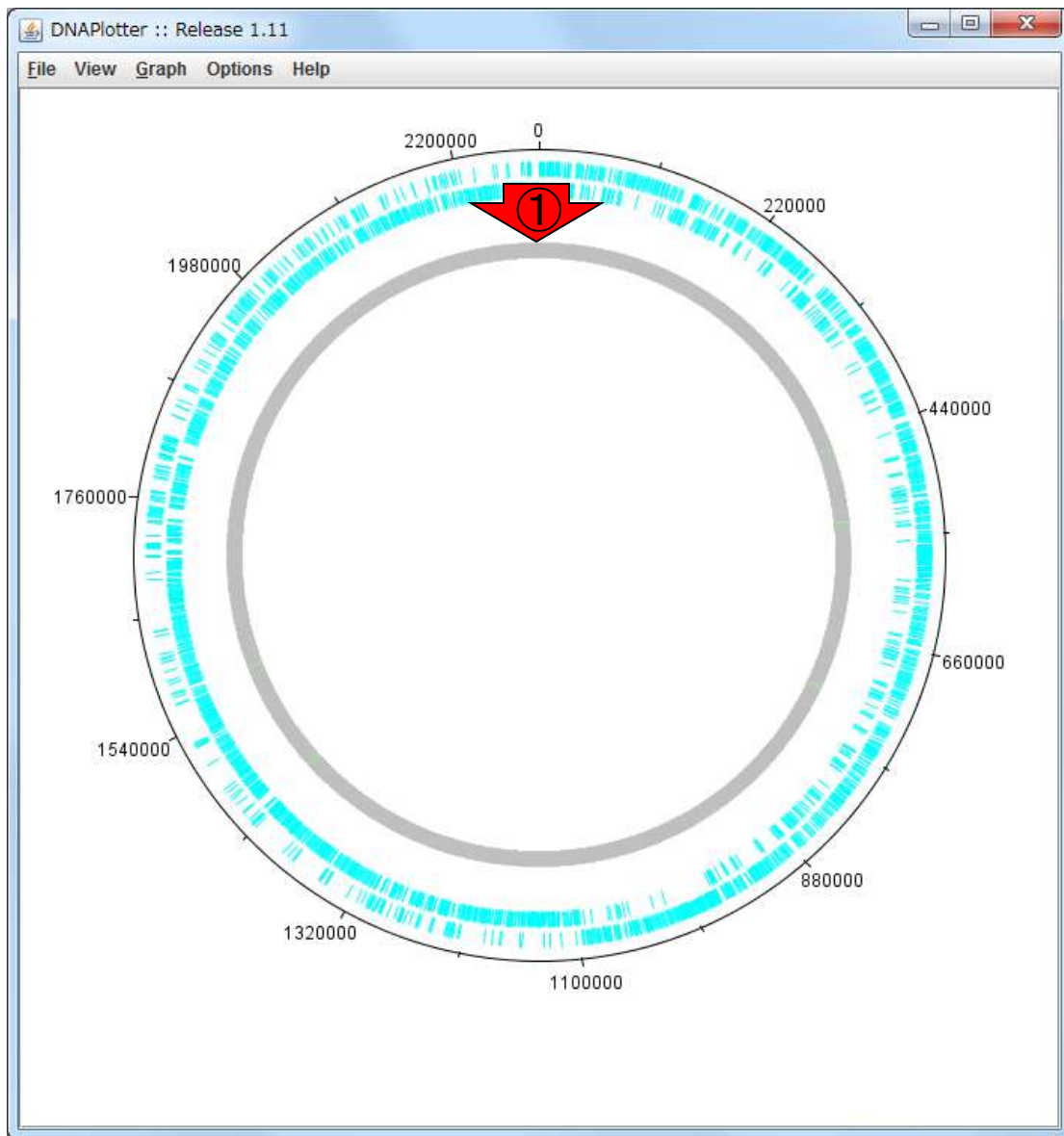
W12-5: ずれの理由

①sequence1の配列長は2,277,983 bpなので、
 ②の長さが40,000 bp弱。プロファージ領域が
 $(43,187 - 5,839 + 1) = 37,349$ bp程度なので、
 ③程度の若干のずれは妥当と当初判断しま
 したが、それだと③が現在の時計の長針で28
 分ごろの位置ではなく32分ごろの位置になら
 ないとオカシイのではないかと感じておりま
 す。ですから、プロファージ領域のみが原因
 ではないと思われます



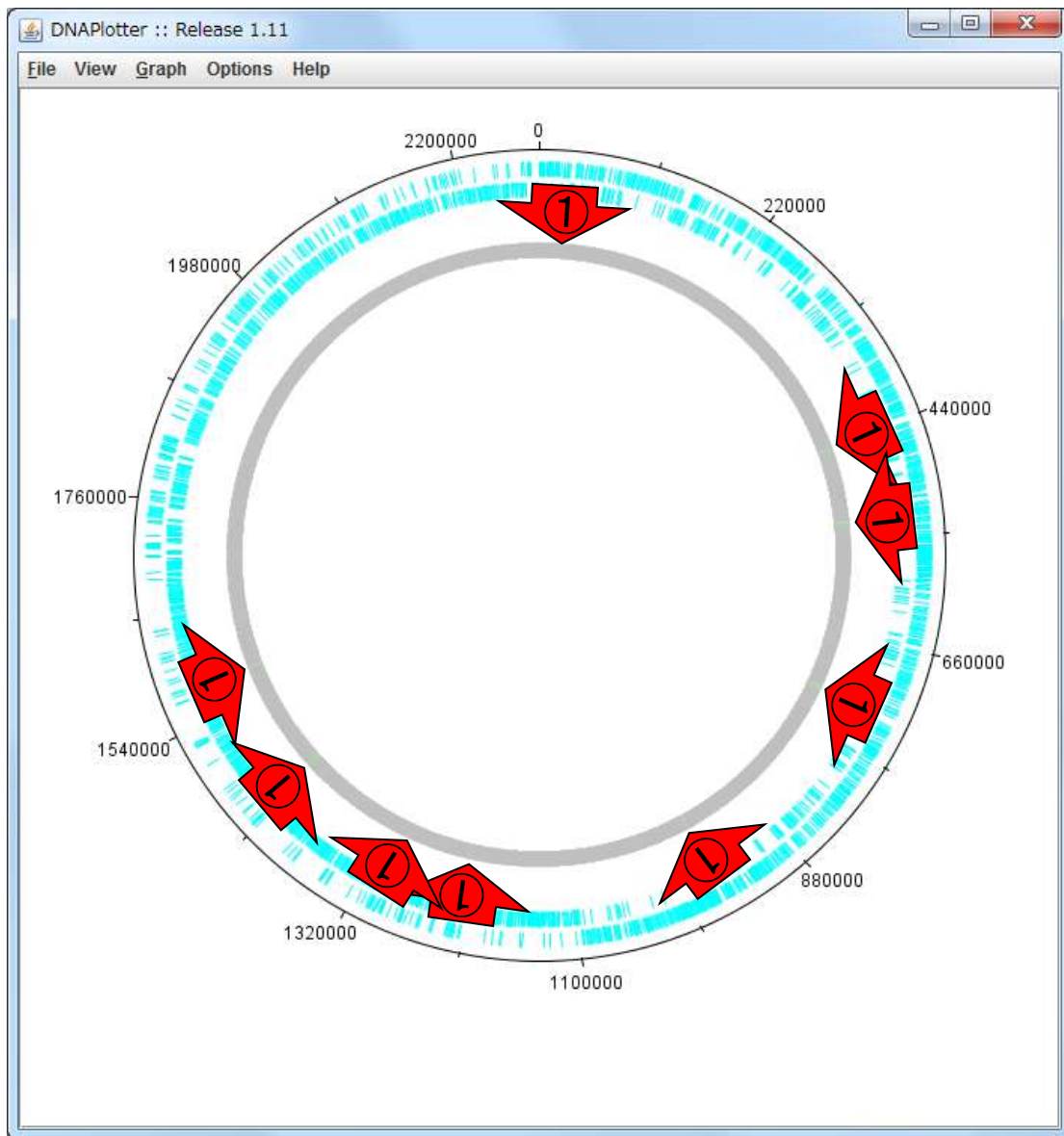
W12-6:tRNA

①灰色の輪っかは、tRNA
の位置情報を表します



W12-6:tRNA

よく見ると、①のあたりに緑色の線があります。これがtRNAの位置情報です。解像度の関係で見えていないものも多数あると思われます



W12-7: DFAST

CDSとtRNAが別々に表示できるのは、例えば DFASTの①Featureタブの、②Feature Typeのところでは種類別に分けられているからです。この Feature Typeの分類分けはGFFファイルや GenBank形式ファイル中にも書き込まれているので、場合分けできるのでしよう

Remember the [current URL](#) to access this page. The result will be deleted 30 days after your last visit. [Delete this job now.](#) => [Delete](#) This procedure cannot be undone.

Title :
JobID : 5359b058-11ef-40e7-8485-83fe7b7a1dce
Status : COMPLETE

```
[2017-01-20 16:17:56.643410] Job submitted.  
[2017-01-20 16:17:56.667496] Job started.  
[2017-01-20 16:25:41.084897] Job completed.
```

[Result](#) **Features** [DDBJ Submission](#) [Log](#)

Annotated Features

Show entries Search:

No.	Locus Tag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation
1	LOOC260_00001	sequence1	101..1417	CDS	chromosomal replication initiator protein DnaA	dnaA	View	View
2	LOOC260_00002	sequence1	1593..2732	CDS	DNA polymerase III subunit beta	dnaN	View	View

W12-7: DFAST

DFASTの①Featureタブ内で、②tRNAなどで検索すると、③Feature TypeがtRNAとなっているものがみられます。④のあたりとかで種類別にソートするほうがいいかも…

Result Features **Features** DDBJ Submission Log

Annotated Features

Show 100 entries

Search: tRNA

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation
21	LOOC260_00021	sequence1	complement (23521..23595)	tRNA	tRNA-Lys		View	
129	LOOC260_00127	sequence1	139985..140056	tRNA	tRNA-Arg		View	
203	LOOC260_00199	sequence1	214943..216745	CDS	threonyl-tRNA synthetase	thrS_1	View	View
342	LOOC260_00330	sequence1	348523..349779	CDS	tyrosyl-tRNA synthetase	tyrS	View	View
368	LOOC260_00354	sequence1	380925..381941	CDS	tryptophanyl-tRNA synthetase	trpS	View	View
375	LOOC260_00361	sequence1	387557..389578	CDS	methionyl-tRNA synthetase	metG	View	View
414	LOOC260_00397	sequence1	423345..423902	CDS	peptidyl-tRNA hydrolase	pth	View	View
420	LOOC260_00403	sequence1	430546..431943	CDS	tRNA(Ile)-lysidine synthetase	mesJ	View	View
424	LOOC260_00407	sequence1	435737..436747	CDS	tRNA-dihydrouridine	dusA	View	View

例えば①のtRNAは23500番目の塩基あたりに相当するので…

W12-8: 23500番目あたり

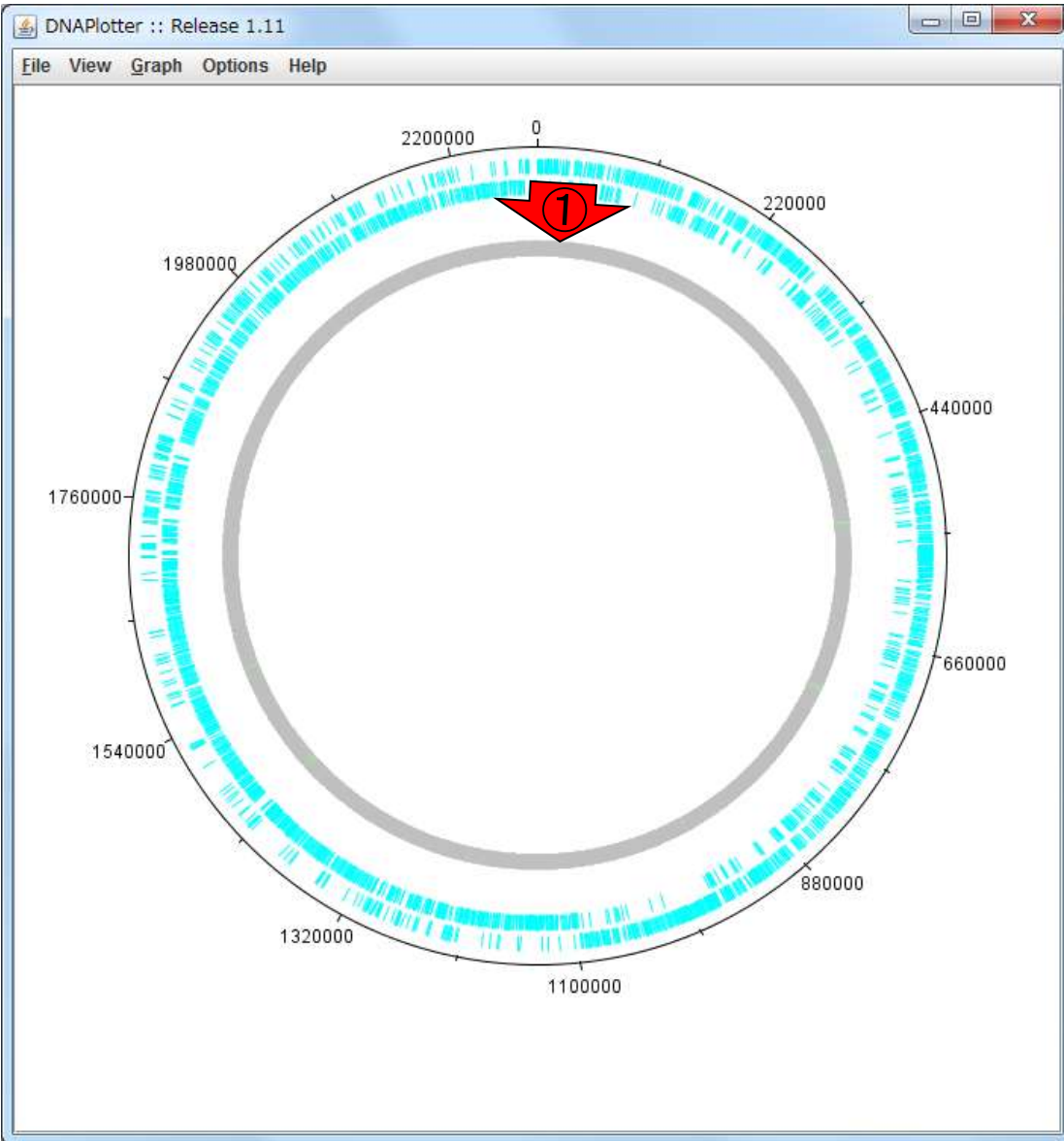
Result Features DDBJ Submission Log

Annotated Features

Show 100 entries Search: tRNA

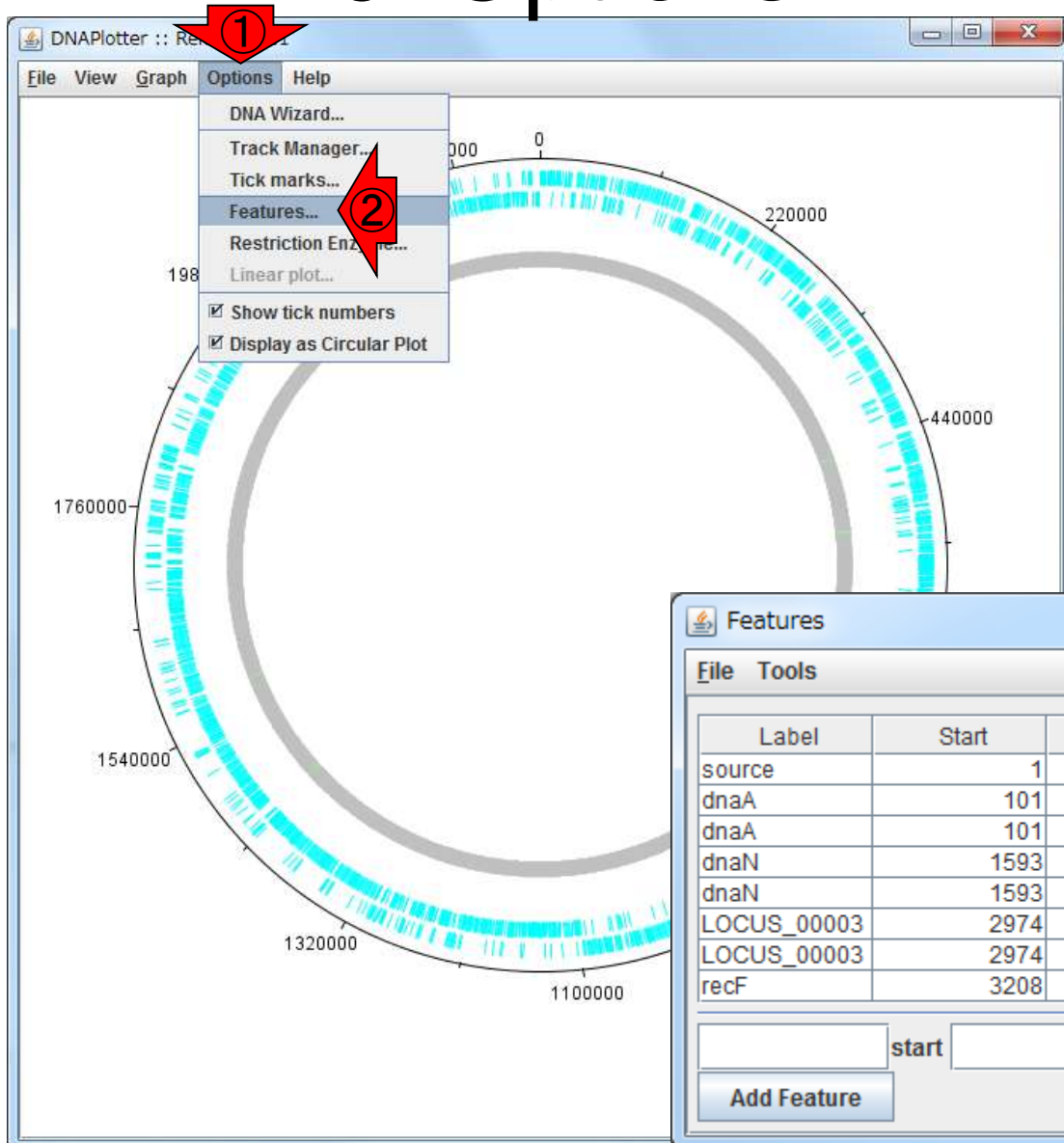
No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation
21	LOOC260_00021	sequence1	complement (23521..23595)	tRNA	tRNA-Lys		View	
129	LOOC260_00127	sequence1	139985..140056	tRNA	tRNA-Arg		View	
203	LOOC260_00199	sequence1	214943..216745	CDS	threonyl-tRNA synthetase	thrS_1	View	View
342	LOOC260_00330	sequence1	348523..349779	CDS	tyrosyl-tRNA synthetase	tyrS	View	View
368	LOOC260_00354	sequence1	380925..381941	CDS	tryptophanyl-tRNA synthetase	trpS	View	View
375	LOOC260_00361	sequence1	387557..389578	CDS	methionyl-tRNA synthetase	metG	View	View
414	LOOC260_00397	sequence1	423345..423902	CDS	peptidyl-tRNA hydrolase	pth	View	View
420	LOOC260_00403	sequence1	430546..431943	CDS	tRNA(Ile)-lysidine synthetase	mesJ	View	View
424	LOOC260_00407	sequence1	435737..436747	CDS	tRNA-dihydrouridine	dusA	View	View

W12-8:23500番目あたり



W12-9: Options

①Optionsの②Featuresで、より詳細なFeatures (CDS or tRNA)の情報をみることができます。どれが水色で表示されていて、どれが緑色で表示されているかなどです



The 'Features' dialog box is shown, displaying a table of genomic features. The table has columns for Label, Start, End, Colour, Line width, Arrow head, and Arrow tail. Below the table, there are input fields for start, stop, and line width, along with an 'Add Feature' button.

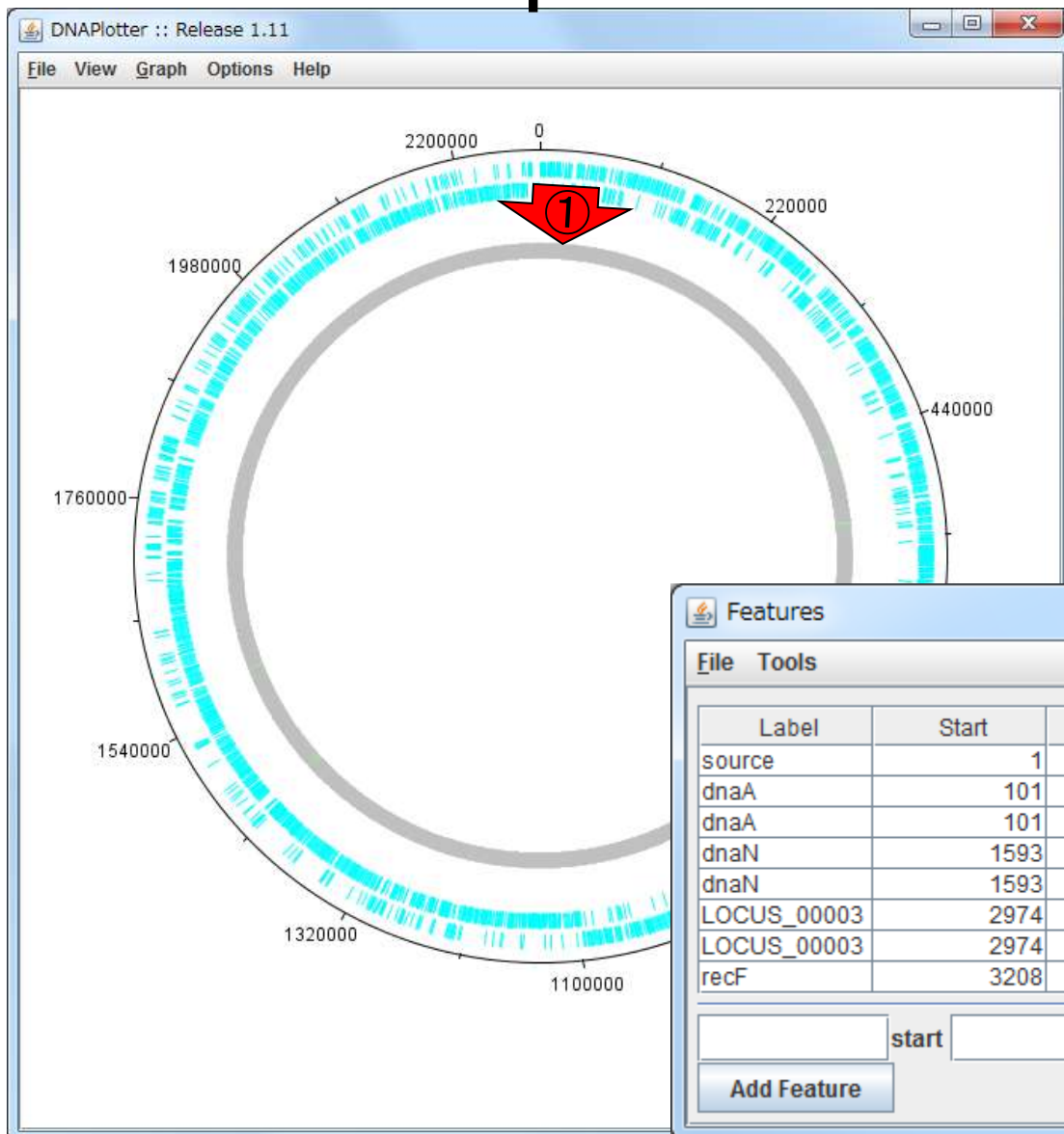
Label	Start	End	Colour	Line width	Arrow head	Arrow tail
source	1	2277983		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaA	101	1417		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaA	101	1417		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaN	1593	2732		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaN	1593	2732		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00003	2974	3198		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00003	2974	3198		10	<input type="checkbox"/>	<input type="checkbox"/>
recF	3208	4329		10	<input type="checkbox"/>	<input type="checkbox"/>

start: [] stop: [] line width: []

Add Feature

W12-9: Options

①緑色に見えている23500番目あたりを実際に調べてみる。②のカーソルを下に移動



Features

File Tools

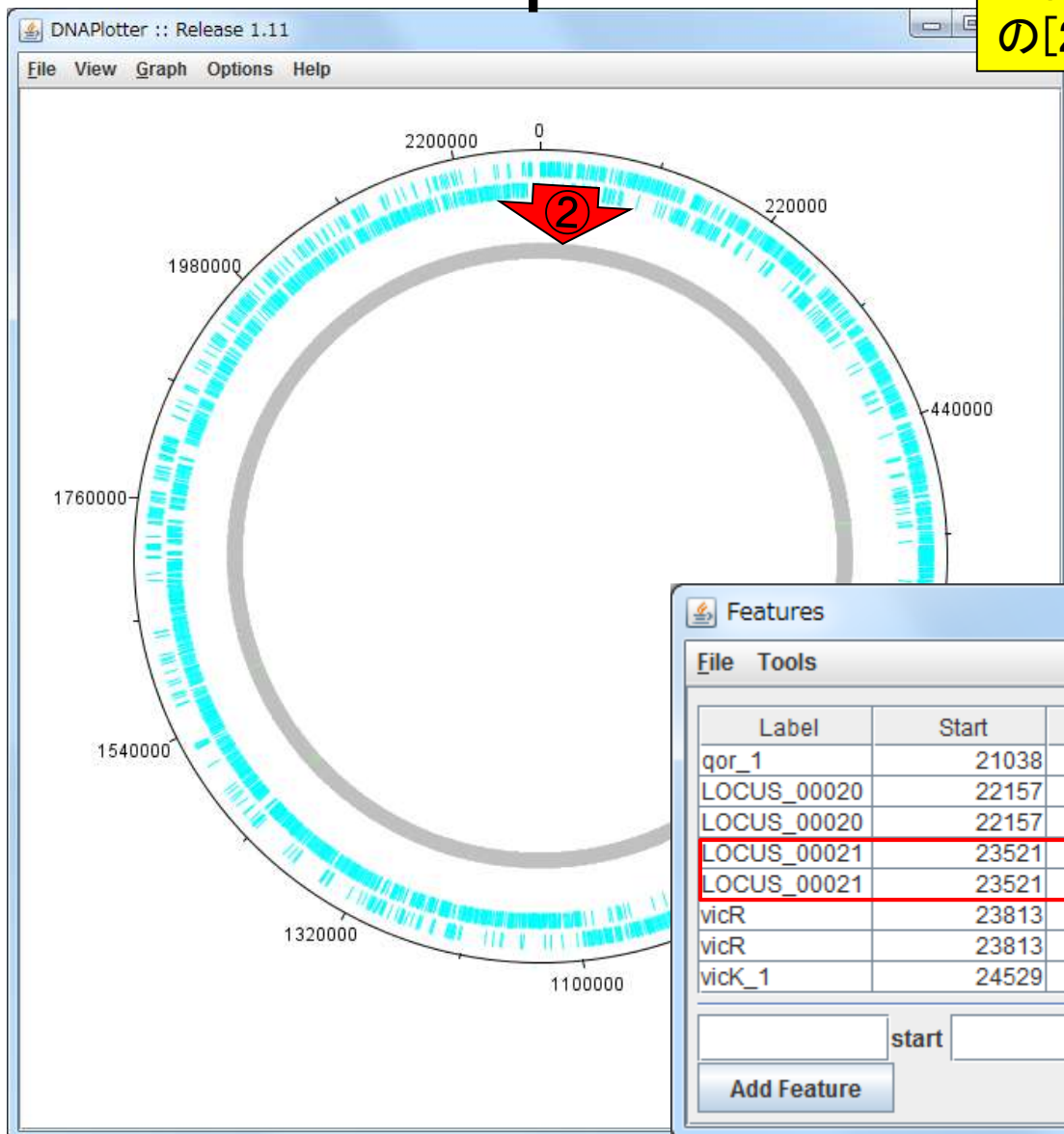
Label	Start	End	Colour	Line width	Arrow head	Arrow tail
source	1	2277983		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaA	101	1417		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaA	101	1417		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaN	1593	2732		10	<input type="checkbox"/>	<input type="checkbox"/>
dnaN	1593	2732		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00003	2974	3198		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00003	2974	3198		10	<input type="checkbox"/>	<input type="checkbox"/>
recF	3208	4329		10	<input type="checkbox"/>	<input type="checkbox"/>

start stop 1 line width

Add Feature

W12-9: Options

①カーソルを少し下に移動させ、だいたい23500番目あたりでColourが緑色になっているところを探した結果。②の緑色は、③赤枠で示すLOCUS_00021の[23521, 23595]が見えているのだろう



Features

File Tools

Label	Start	End	Colour	Line width	Arrow head	Arrow tail
qor_1	21038	21961		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00020	22157	22915		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00020	22157	22915		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00021	23521	23595		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00021	23521	23595		10	<input type="checkbox"/>	<input type="checkbox"/>
vicR	23813	24514		10	<input type="checkbox"/>	<input type="checkbox"/>
vicR	23813	24514		10	<input type="checkbox"/>	<input type="checkbox"/>
vicK_1	24529	26391		10	<input type="checkbox"/>	<input type="checkbox"/>

start stop 1 line width

Add Feature

③

①

W12-9: DFASTで確認

The screenshot shows the DFAST web interface. The main table lists annotated features. A red box highlights the entry for LocusTag LOOC260_00021, which is a tRNA-Lys gene. A red arrow labeled '1' points to this entry. An inset window titled 'Features' shows a detailed view of the feature. A red box highlights the entry for LOCUS_00021, which is a tRNA-Lys gene. A red arrow labeled '2' points to this entry. The 'Features' window also shows a table with columns for Label, Start, End, Colour, Line width, Arrow head, and Arrow tail. Below the table, there are input fields for start, stop, and line width, and an 'Add Feature' button.

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation
21	LOOC260_00021	sequence1	complement (23521..23595)	tRNA	tRNA-Lys		View	
129	LOOC260_00127	sequence1	139985..140050	tRNA	tRNA-Phe		View	
203	LOOC260_00199	sequence1	214944..215009	tRNA	tRNA-Phe		View	
342	LOOC260_00330	sequence1	348523..348588	tRNA	tRNA-Phe		View	
368	LOOC260_00354	sequence1	380923..380988	tRNA	tRNA-Phe		View	
375	LOOC260_00361	sequence1	387553..387618	tRNA	tRNA-Phe		View	
414	LOOC260_00397	sequence1	423343..423408	tRNA	tRNA-Phe		View	
420	LOOC260_00403	sequence1	430543..430608	tRNA	tRNA-Phe		View	
424	LOOC260_00407	sequence1	435733..435798	tRNA	tRNA-Phe		View	

Label	Start	End	Colour	Line width	Arrow head	Arrow tail
qor_1	21038	21961	Cyan	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00020	22157	22915	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00020	22157	22915	Cyan	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00021	23521	23595	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00021	23521	23595	Green	10	<input type="checkbox"/>	<input type="checkbox"/>
vicR	23813	24514	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>
vicR	23813	24514	Cyan	10	<input type="checkbox"/>	<input type="checkbox"/>
vicK_1	24529	26391	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>

W12-9: DFASTで確認

The screenshot displays the DFAST web interface for feature annotation. The main table lists features with columns for No., LocusTag, Seq. ID, Location, Feature Type, Product, Gene, Nucleotide, and Translation. A search filter for 'tRNA' is applied. The entry for LOOC260_00127 (tRNA-Arg) is highlighted with a red box and a red arrow labeled '1'. An inset window titled 'Features' provides a detailed view of the features, including a table with columns for Label, Start, End, Colour, Line width, Arrow head, and Arrow tail. The entry for LOCUS_00127 (green bar) is highlighted with a red box and a red arrow labeled '2'. Below the table in the inset window, there are input fields for 'start', 'stop', 'line width', and an 'Add Feature' button.

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation
21	LOOC260_00021	sequence1	complement (23521..23595)	tRNA	tRNA-Lys		View	
129	LOOC260_00127	sequence1	139985..140056	tRNA	tRNA-Arg		View	
203	LOOC260_00199	sequence1	214943..216745	CDS	threonyl-tRNA synthetase	thrS_1	View	View
342	LOOC260_00330	sequence1	348522..349770	CDS	tRNA synthetase	thrS_1	View	View
368	LOOC260_00354	sequence1	380924..381672	CDS	tRNA synthetase	thrS_1	View	View
375	LOOC260_00361	sequence1	387552..388300	CDS	tRNA synthetase	thrS_1	View	View
414	LOOC260_00397	sequence1	423344..424092	CDS	tRNA synthetase	thrS_1	View	View
420	LOOC260_00403	sequence1	430544..431292	CDS	tRNA synthetase	thrS_1	View	View
424	LOOC260_00407	sequence1	435732..436480	CDS	tRNA synthetase	thrS_1	View	View
425	LOOC260_00408	sequence1	436784..437532	CDS	tRNA synthetase	thrS_1	View	View

Label	Start	End	Colour	Line width	Arrow head	Arrow tail
ddpX	138548	139105	Cyan	10	<input type="checkbox"/>	<input type="checkbox"/>
ada	139117	139623	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>
ada	139117	139623	Cyan	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00127	139985	140056	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00127	139985	140056	Green	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00128	140164	140301	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00128	140164	140301	Cyan	10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00129	140837	141076	Grey	10	<input type="checkbox"/>	<input type="checkbox"/>

W12-9: DFASTで確認

http://dfast.nig.ac.jp/analysis/annotation/5359b058-11ef-40e7-8485-83fe7b7a1dce/feat

Line	LOCUS	Sequence	Start	End	Type	Name	Gene	View	View
425	LOOC260_00408	sequence1	436782..438275		CDS	lysyl-tRNA synthetase	lysU	View	View
429	LOOC260_00412	sequence1	443594..443668		tRNA	tRNA-Asn		View	
430	LOOC260_00413	sequence1	443713..443787		tRNA	tRNA-Thr		View	
433	LOOC260_00416	sequence1	446553..446635		tRNA	tRNA-Tyr		View	
434	LOOC260_00417	sequence1	446640..446713		tRNA	tRNA-Gln		View	
446	LOOC260_00429	sequence1	461973..462060		tRNA	tRNA-Leu		View	
466	LOOC260_00447	sequence1	479060..480550		CDS	glutamyl-tRNA synthetase	glnS	View	View
467	LOOC260_00448	sequence1	480810..480810						
497	LOOC260_00476	sequence1	505450..505450						
530	LOOC260_00508	sequence1	532540..532540						
531	LOOC260_00509	sequence1	532650..532650						
532	LOOC260_00510	sequence1	532740..532740						
536	LOOC260_00513	sequence1	535080..535080						
537	LOOC260_00514	sequence1	535170..535170						
538	LOOC260_00515	sequence1	535250..535250						
586	LOOC260_00563	sequence1	577230..577230						

Features

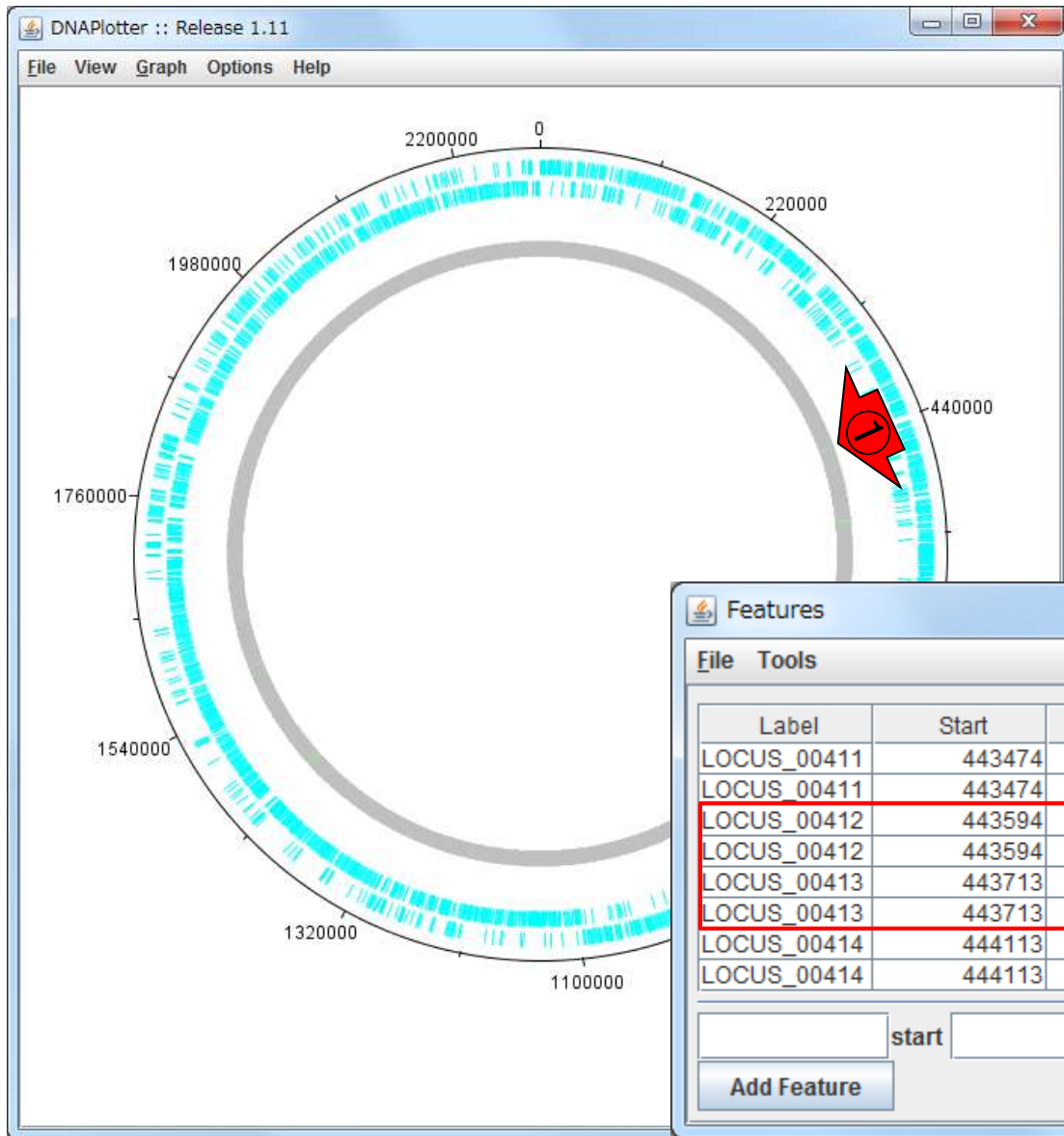
File Tools

Label	Start	End	Colour	Line width	Arrow head	Arrow tail
LOCUS_00411	443474	443586		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00411	443474	443586		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00412	443594	443668		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00412	443594	443668		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00413	443713	443787		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00413	443713	443787		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00414	444113	445609		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00414	444113	445609		10	<input type="checkbox"/>	<input type="checkbox"/>

start stop line width

Add Feature

W12-9: DFASTで確認



Features

File Tools

Label	Start	End	Colour	Line width	Arrow head	Arrow tail
LOCUS_00411	443474	443586		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00411	443474	443586		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00412	443594	443668		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00412	443594	443668		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00413	443713	443787		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00413	443713	443787		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00414	444113	445609		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00414	444113	445609		10	<input type="checkbox"/>	<input type="checkbox"/>

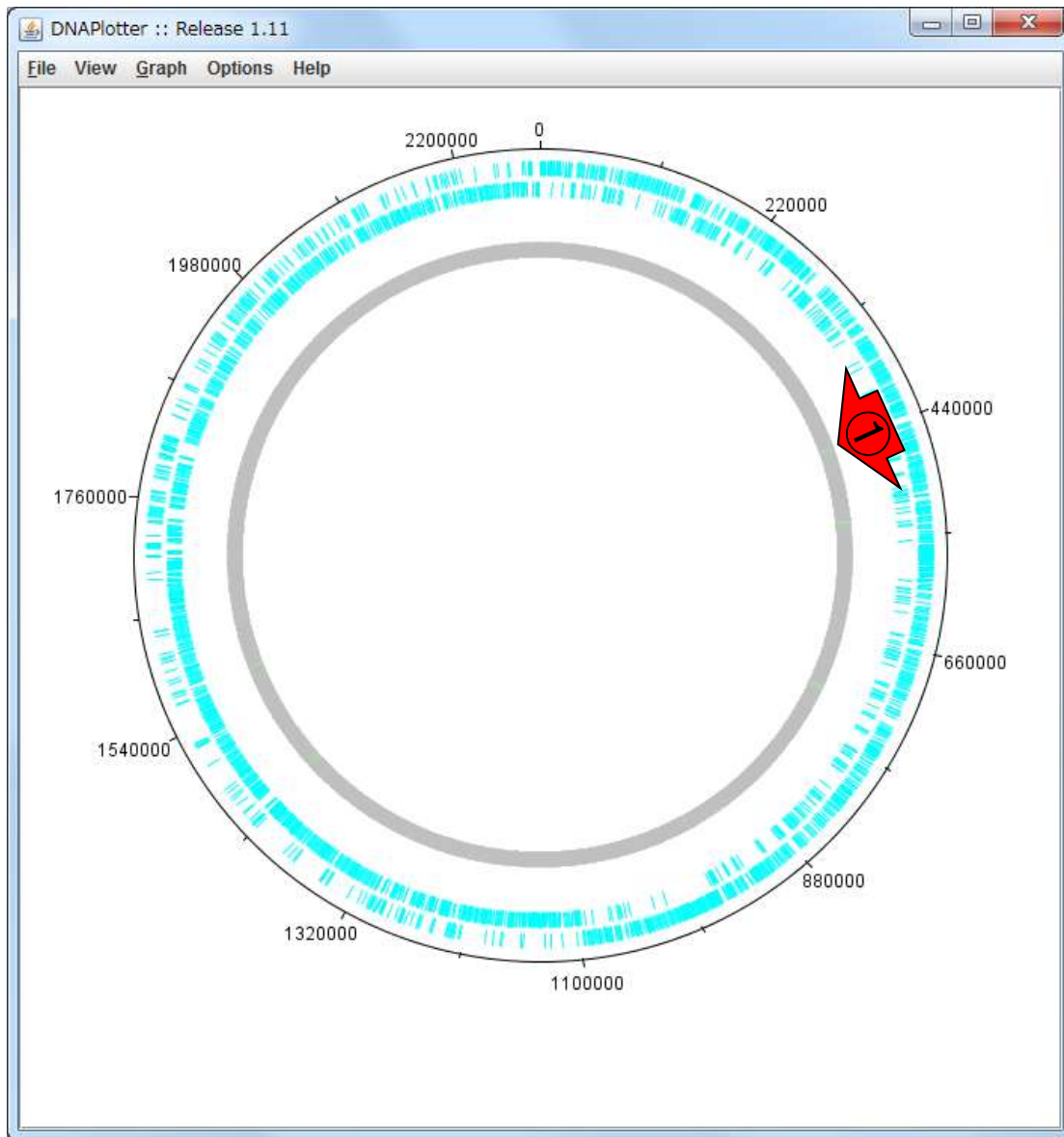
2

start stop 1 line width

Add Feature

W12-10: ちなみに...

マニュアルには①の部分で緑色で示されるものはrRNAとtRNAだと書かれていますが、実際にはtRNAだけなのです



W12-10: ちなみに...

例えば、DFASTの①Featureタブ上で、②rRNAで検索して得られる、③は...

Result Features **DFAST - Job Result** DDBJ Submission Log

Annotated Features

Show 100 entries Search: rRNA

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
35	LOOC260_00035	sequence1	41089..41568	CDS	rRNA methyltransferase		View	View	Edit
210	LOOC260_00206	sequence1	226598..227323	CDS	16S rRNA methyltransferase GidB	gidB	View	View	Edit
343	LOOC260_00331	sequence1	350155..351730	rRNA	16S ribosomal RNA		View		Edit
344	LOOC260_00332	sequence1	351956..354877	rRNA	23S ribosomal RNA		View		Edit
345	LOOC260_00333	sequence1	354977..355089	rRNA	5S ribosomal RNA		View		Edit
426	LOOC260_00409	sequence1	438651..440226	rRNA	16S ribosomal RNA		View		Edit
427	LOOC260_00410	sequence1	440453..443374	rRNA	23S ribosomal RNA		View		Edit

W12-10: ちなみに...

Result Features DDBJ Submission Log

Annotated Features

Show 100 entries Search: rRNA

No.	LocusTag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	Edit
35	LOOC260_00035	sequence1	41089..41568	CDS	rRNA methyltransferase		View	View	Edit
210	LOOC260_00206	sequence1	226598..227000	CDS	16S rRNA	16S	View	View	Edit
343	LOOC260_00331	sequence1	350155..351730	CDS	tyrS	tyrS	View	View	Edit
344	LOOC260_00332	sequence1	351956..354877	CDS	tyrS	tyrS	View	View	Edit
345	LOOC260_00333	sequence1	354977..355089	CDS	tyrS	tyrS	View	View	Edit
426	LOOC260_00409	sequence1	438651..439100	CDS	tyrS	tyrS	View	View	Edit
427	LOOC260_00410	sequence1	440451..440900	CDS	tyrS	tyrS	View	View	Edit

Features

File Tools

Label	Start	End	Colour	Line width	Arrow head	Arrow tail
tyrS	348523	349779		10	<input type="checkbox"/>	<input type="checkbox"/>
tyrS	348523	349779		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00331	350155	351730		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00331	350155	351730		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00332	351956	354877		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00332	351956	354877		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00333	354977	355089		10	<input type="checkbox"/>	<input type="checkbox"/>
LOCUS_00333	354977	355089		10	<input type="checkbox"/>	<input type="checkbox"/>

start stop 1 line width

Add Feature

W12-10: ちなみに...

wellcome trust
sanger
institute

SCIENCE

DNAPlotter

- Overview
- Download
- Learn
- License

Overview

DNAPlotter can be used to generate images of circular and linear DNA maps to display regions and features of interest. The images can be inserted into a document or printed out directly. As this uses [Artemis](#) it can read in the common file formats EMBL, GenBank and GFF3.

Download and Installation

Tool type

- [Annotation](#)

Screenshots

W12-10: ちなみに...

①のように書いてあったからなのですが、②のところでも書いてあるように、色を変えることができます。なので、①で書いてあるのはデフォルトではなく、変更後の色使いなのではと思います。このようにいろいろいじってみないと分からないことも多々あります

http://www.sanger.ac.uk/science/tools/dnaplotter

Overview

Download

Learn

License

Contact

Related

Publications

Learn and Support

The screenshot on the right shows the *S. typhi* genome and the tracks from the outside represent:

- Forward CDS
- Reverse CDS
- pseudogenes
- Salmonella Pathogenicity Island (red)
- repeat regions (blue)
- rRNA and tRNA (green) ①
- %GC plot
- GC skew ([GC]/[G+C])

Click [here](#) to open this in DNAPlotter.

Editing Features ②

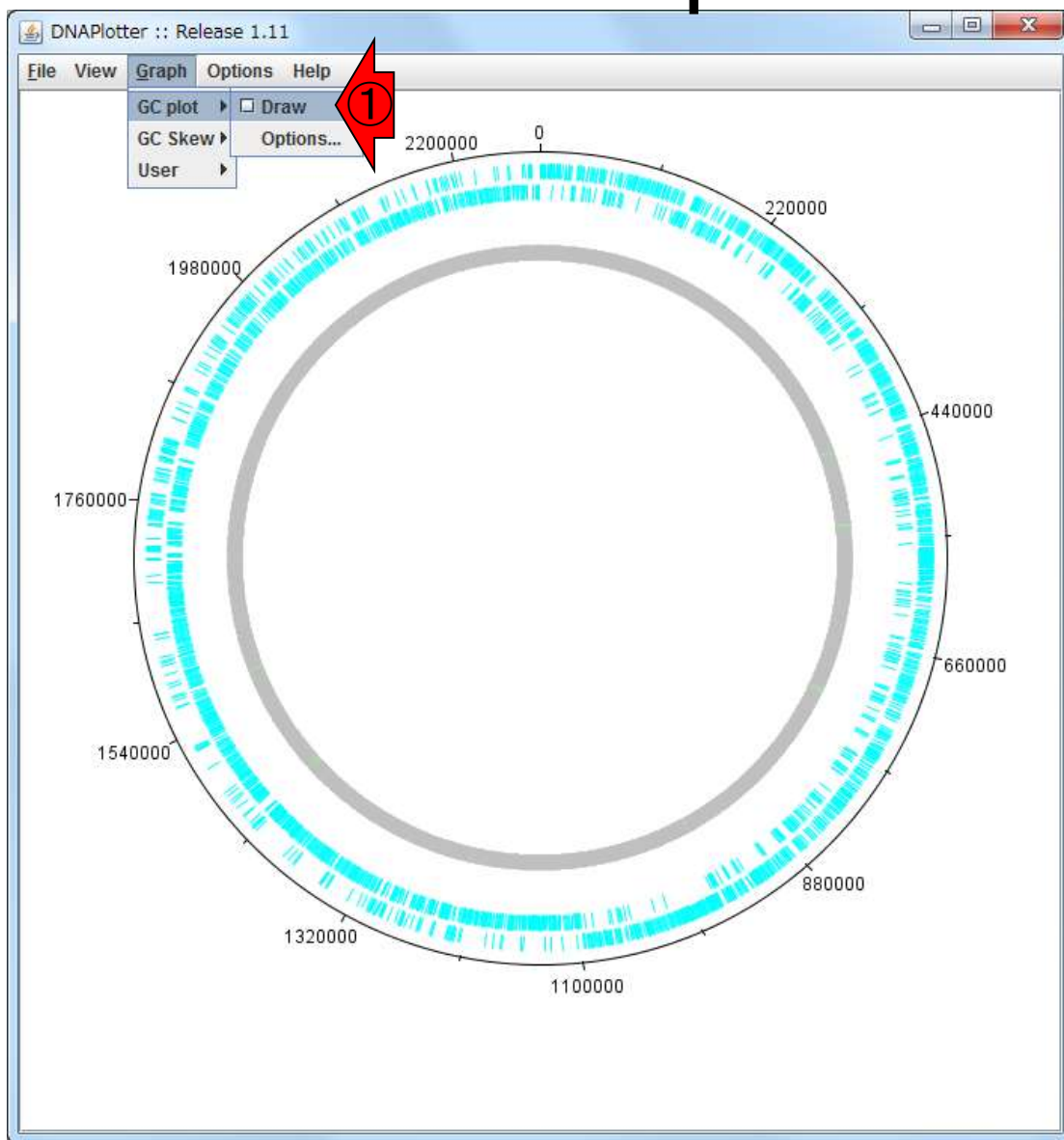
Each feature drawn can be double clicked on to open up a properties window. This allows the label, start and end coordinates, colour and line width to be changed. There is also an option to show or hide the label.

translation.

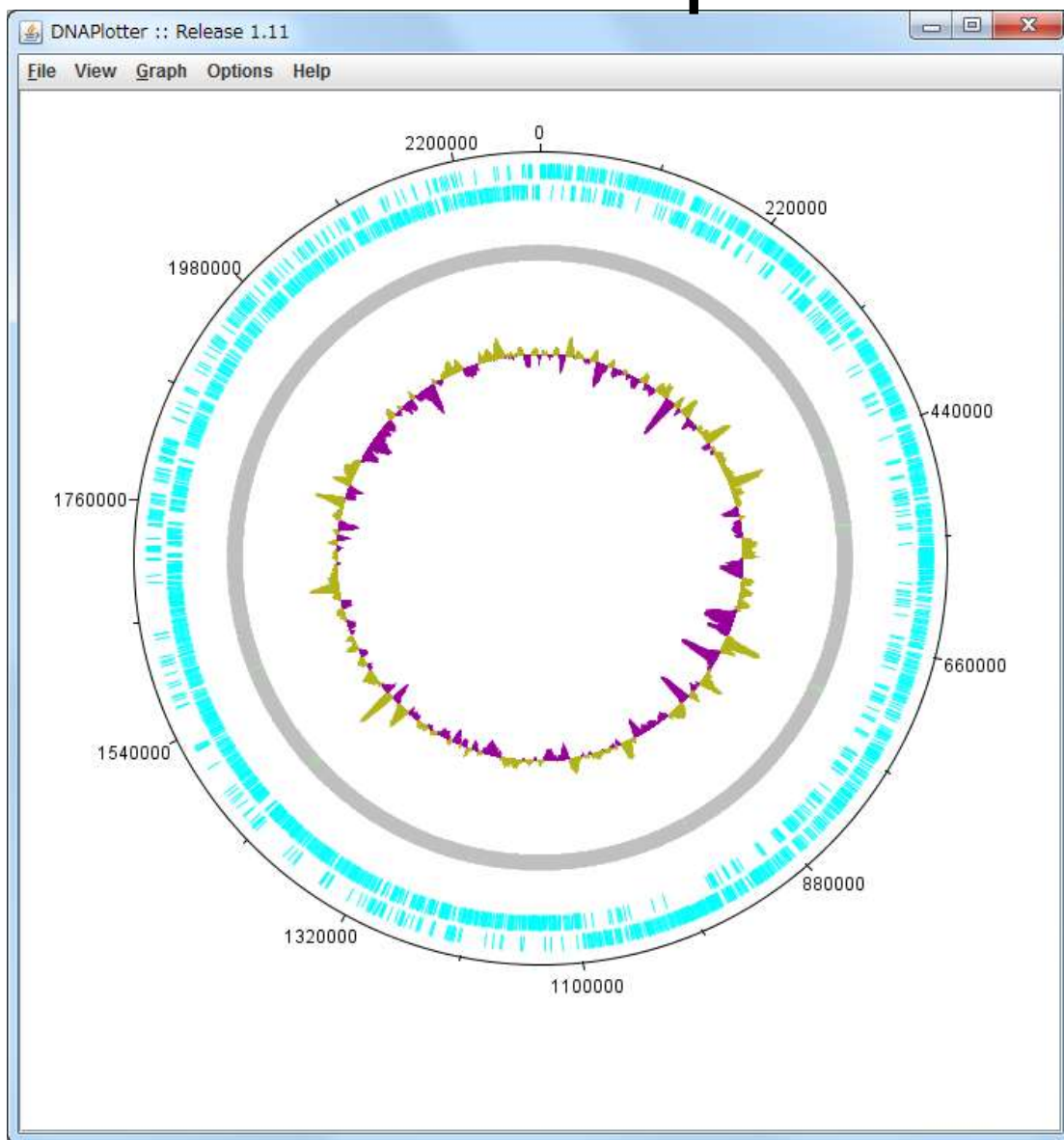
- **Artemis Comparison Tool (ACT)**
A Java application for displaying pairwise comparisons between two or more DNA sequences. It can be used to identify and analyse regions of similarity and difference between genomes and to explore conservation of synteny, in the context of the entire sequences and their annotation.
- **BamView**
An interactive Java application for visualising read-alignment data stored in BAM files.
- **Web-Artemis**
A lightweight web-based version of the popular genome visualisation and annotation tool.

W12-11: GC plot

①Graph - GC plot - Draw。これでGC plotを追加で描画することができます

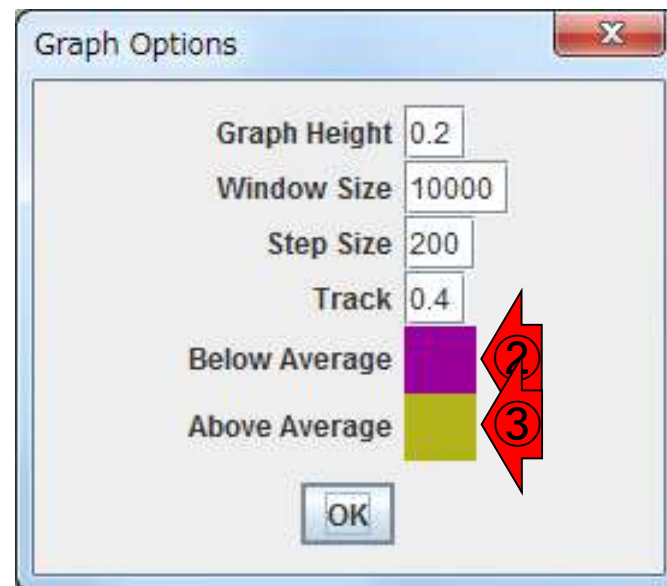
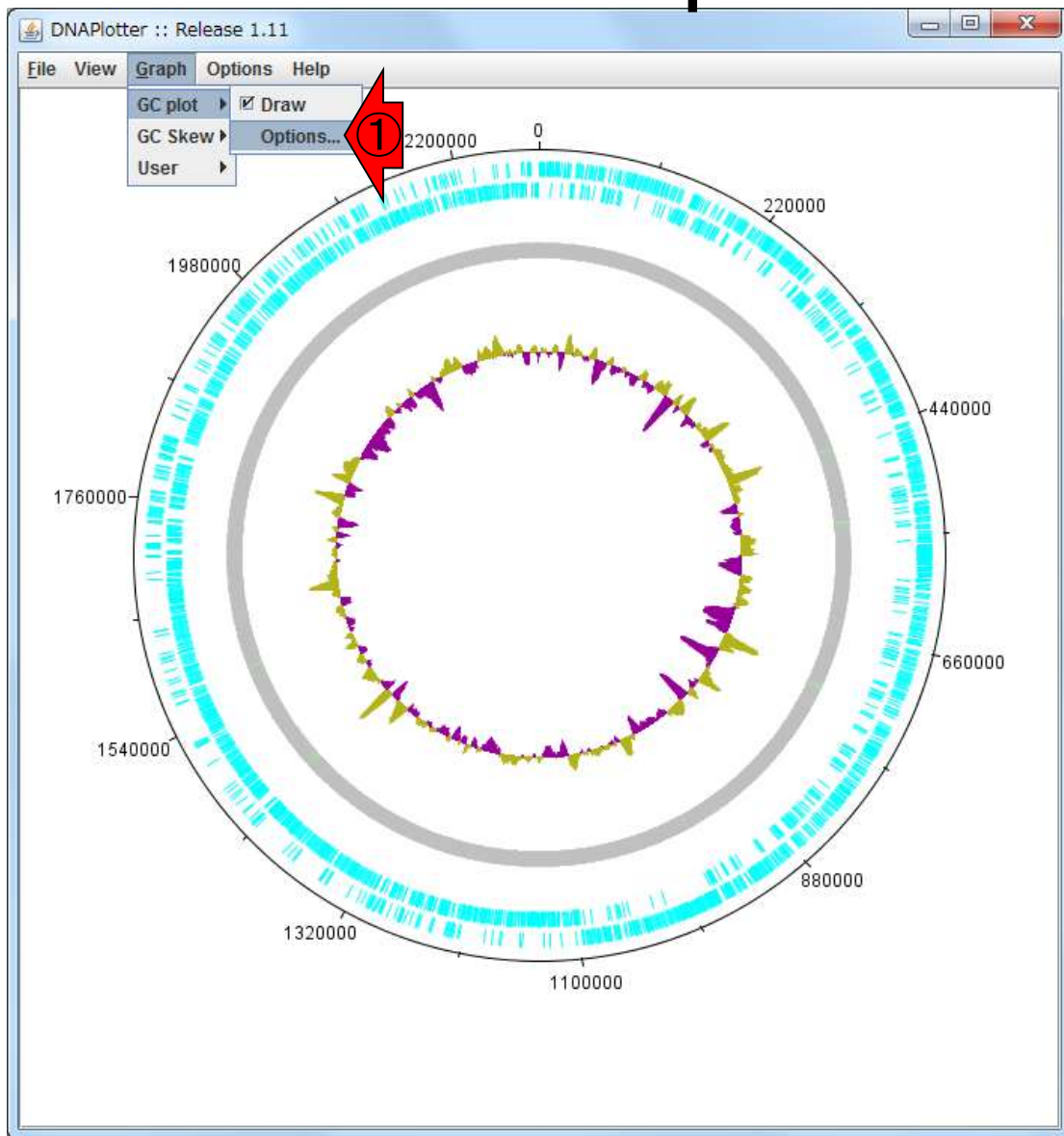


W12-11: GC plot



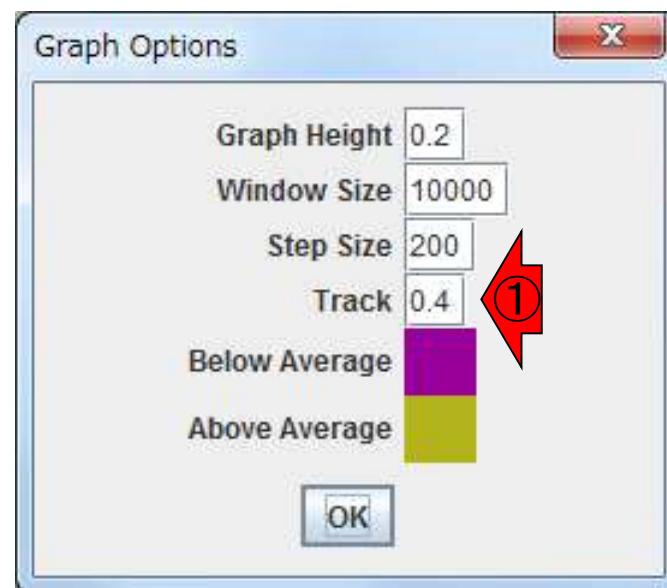
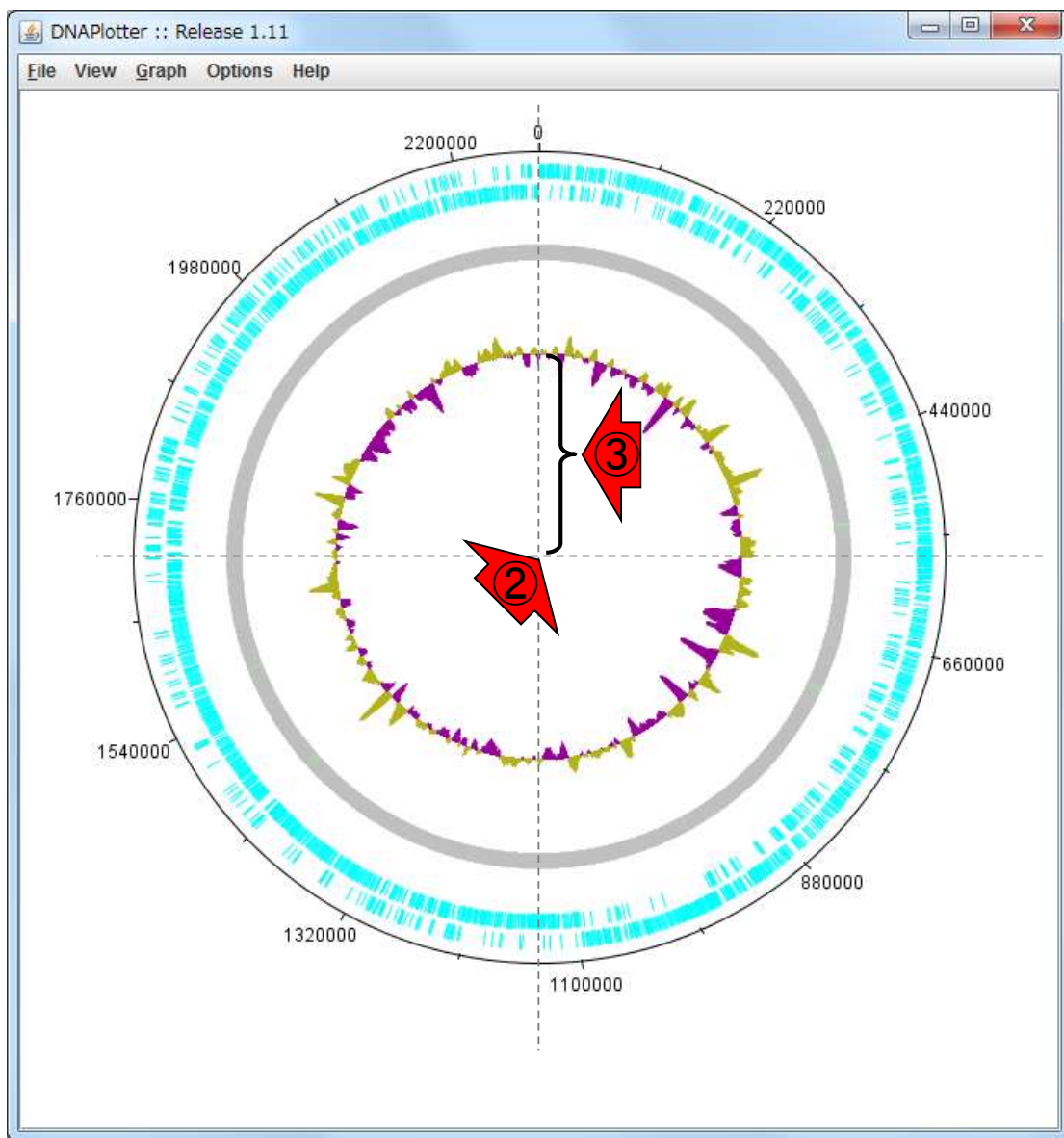
W12-11: GC plot

①Graph - GC plot - Options...で、色使いの意味を知ることができます。②GC含量が平均よりも低い領域は赤紫色?!、③平均よりも高い領域は黄土色?!で表されていることがわかります

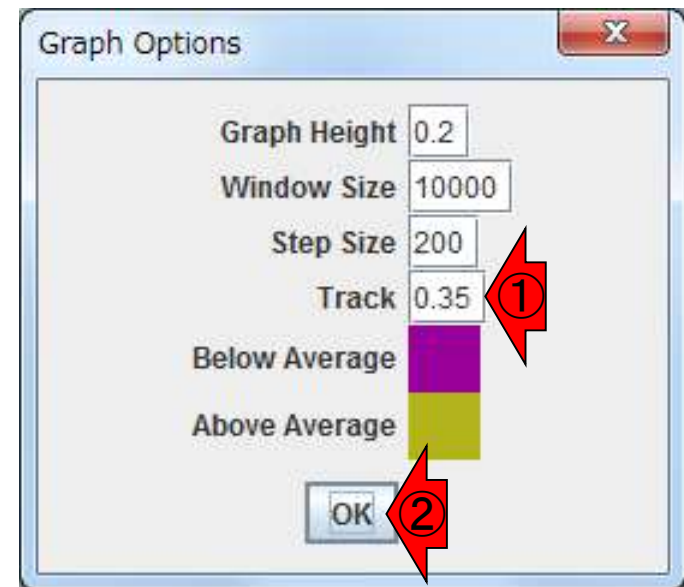
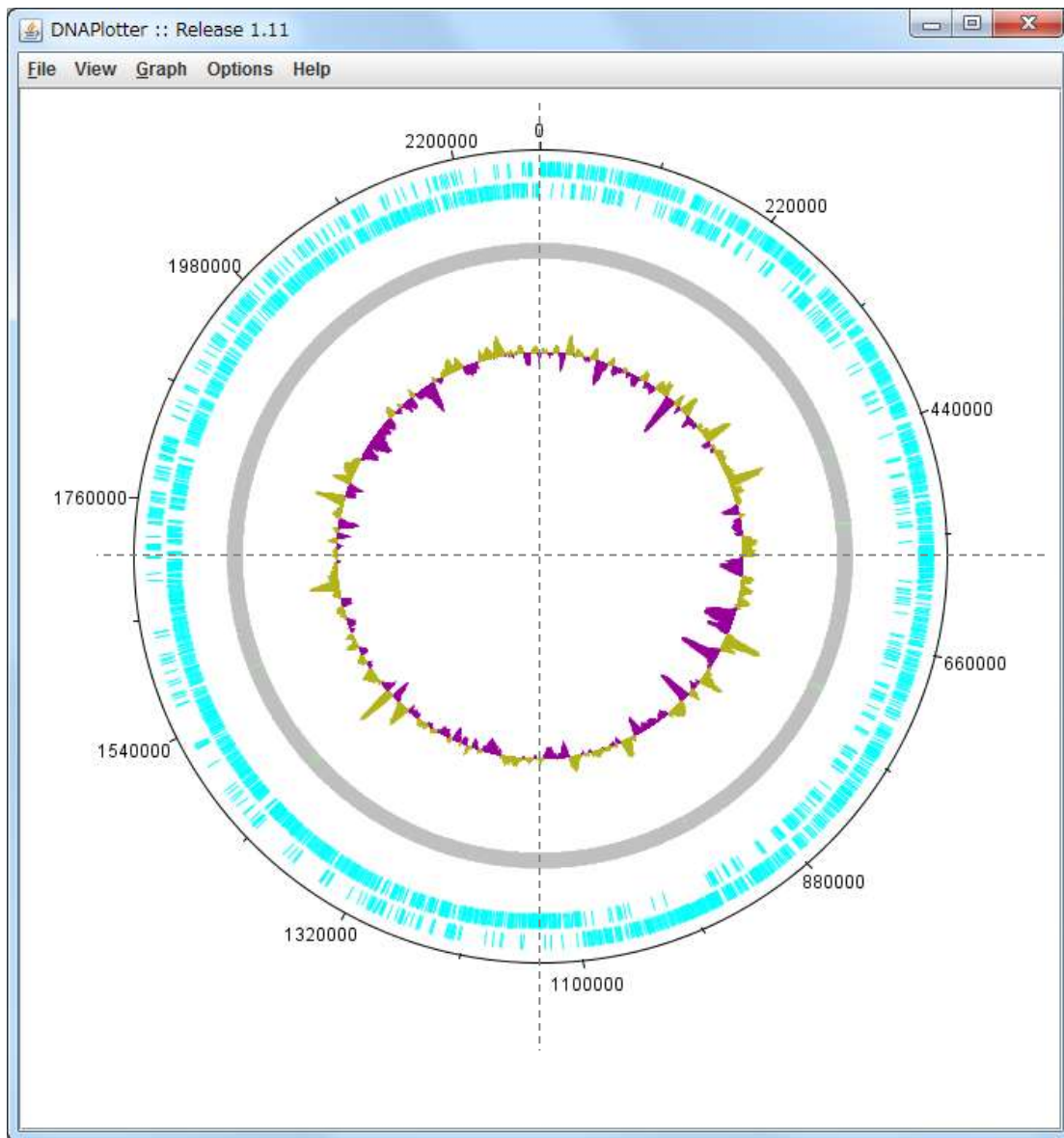


W12-12: Track

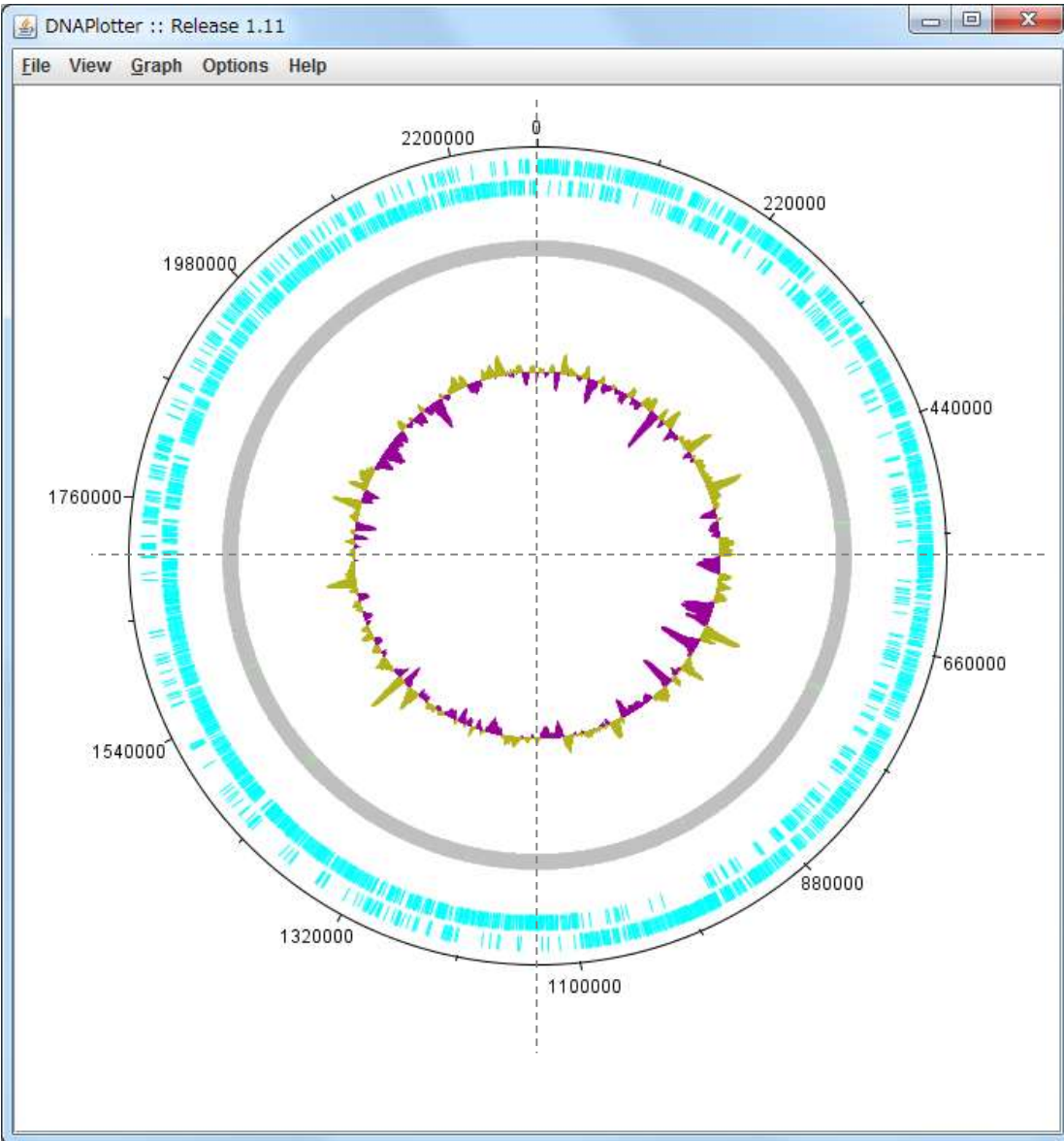
①Track = 0.4というのは、②中心からの③相対的な距離と解釈すればよい



W12-12: Track

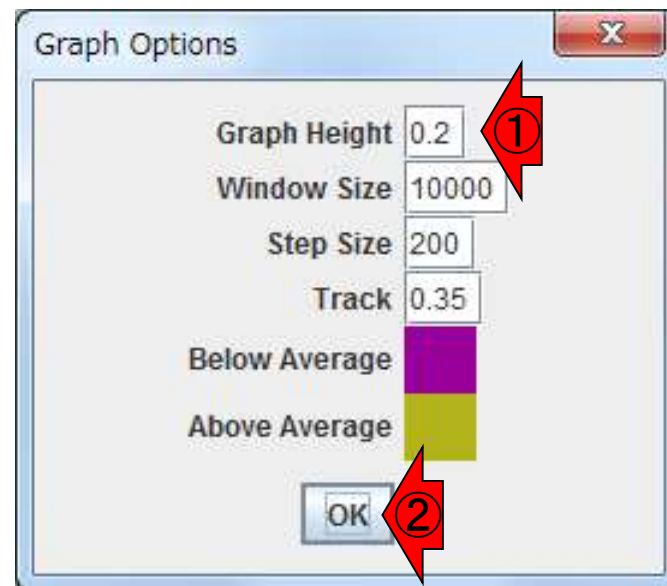
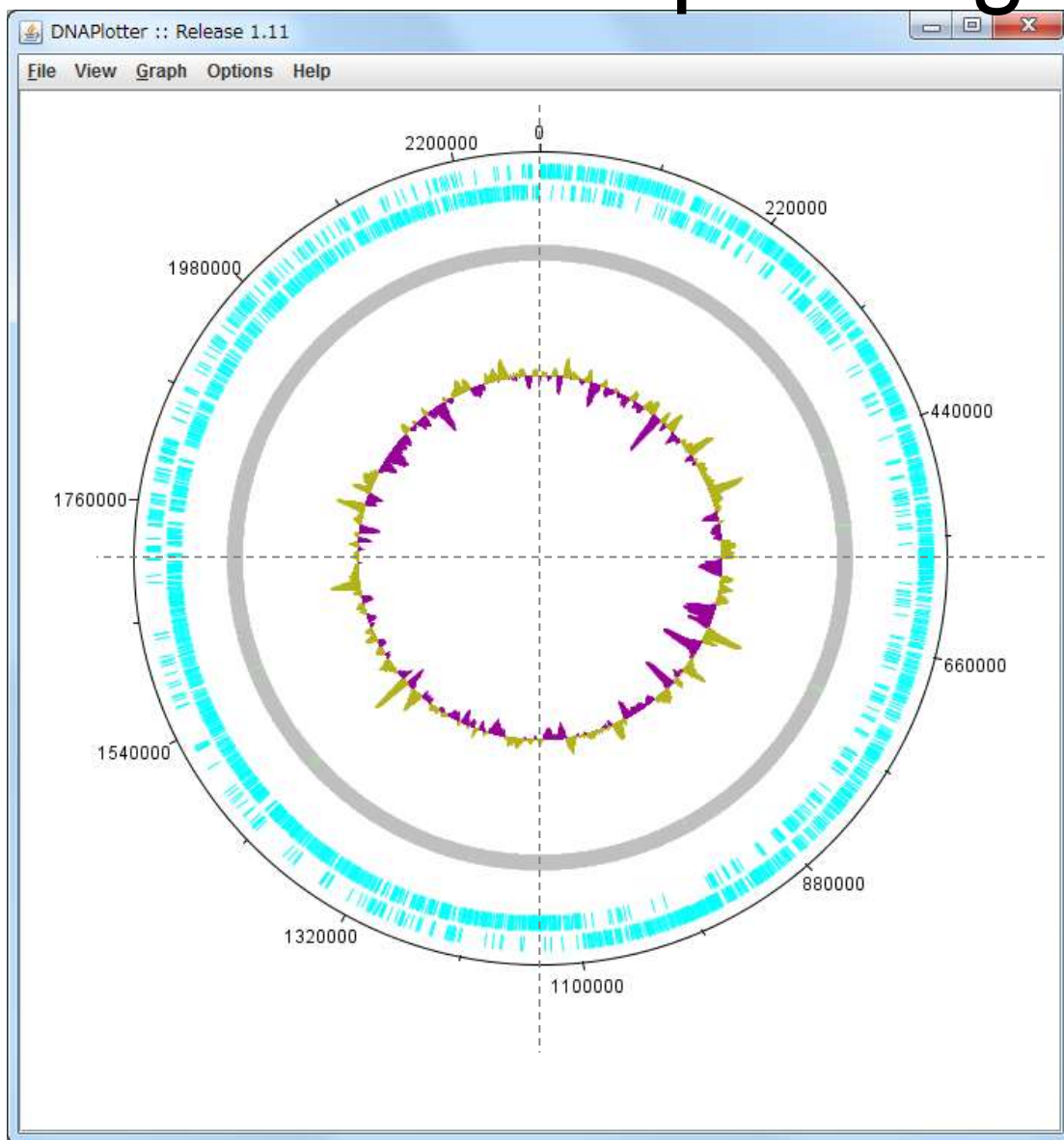


W12-12: Track

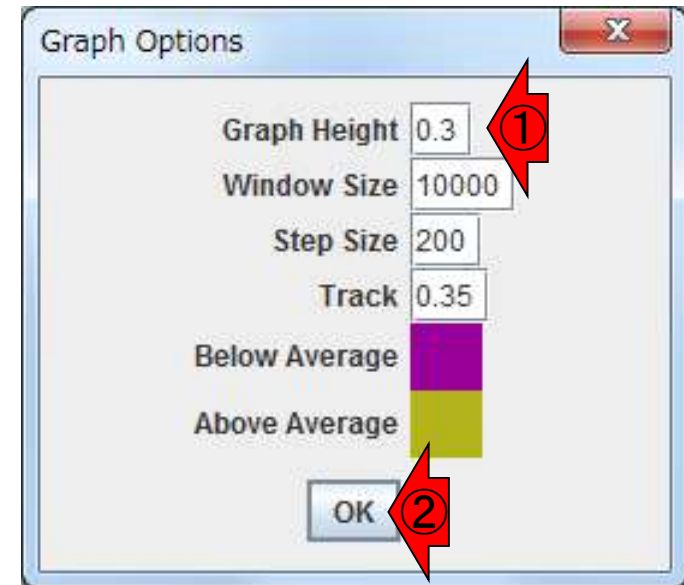
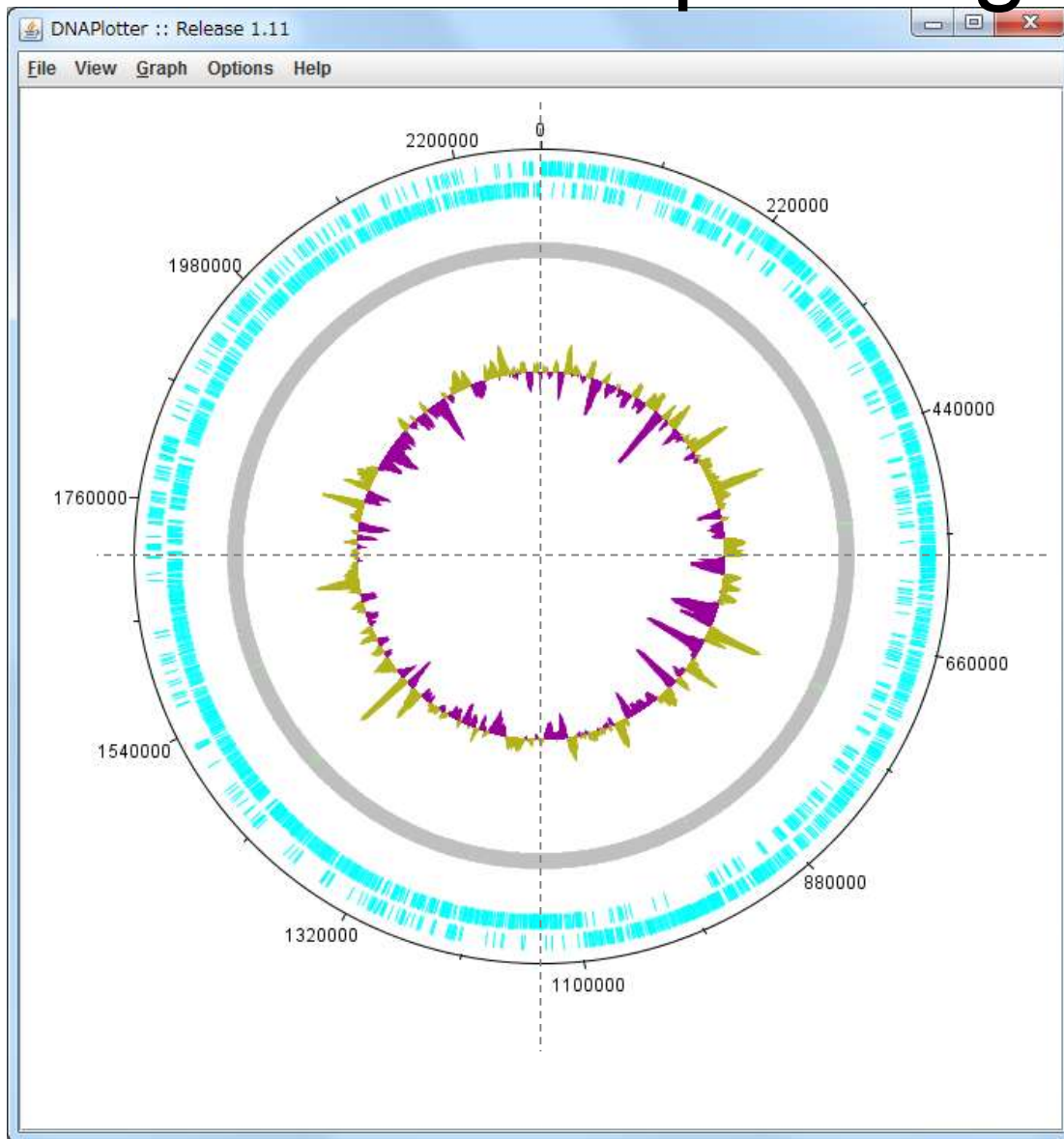


W12-13: Graph Height

①Graph Heightは高さの調整するところなのでしょ。例えば①Graph Height = 0.3にすると、0.2のときに比べて上下に伸びる感じに見えるのだからと妄想しながらいじるのです

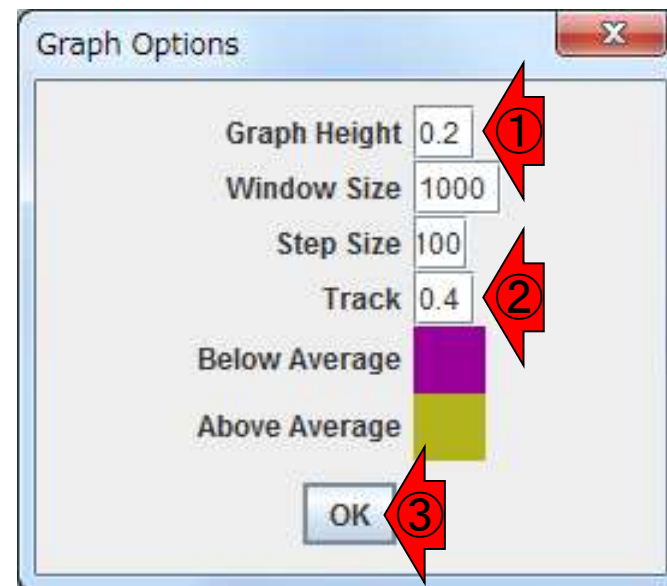
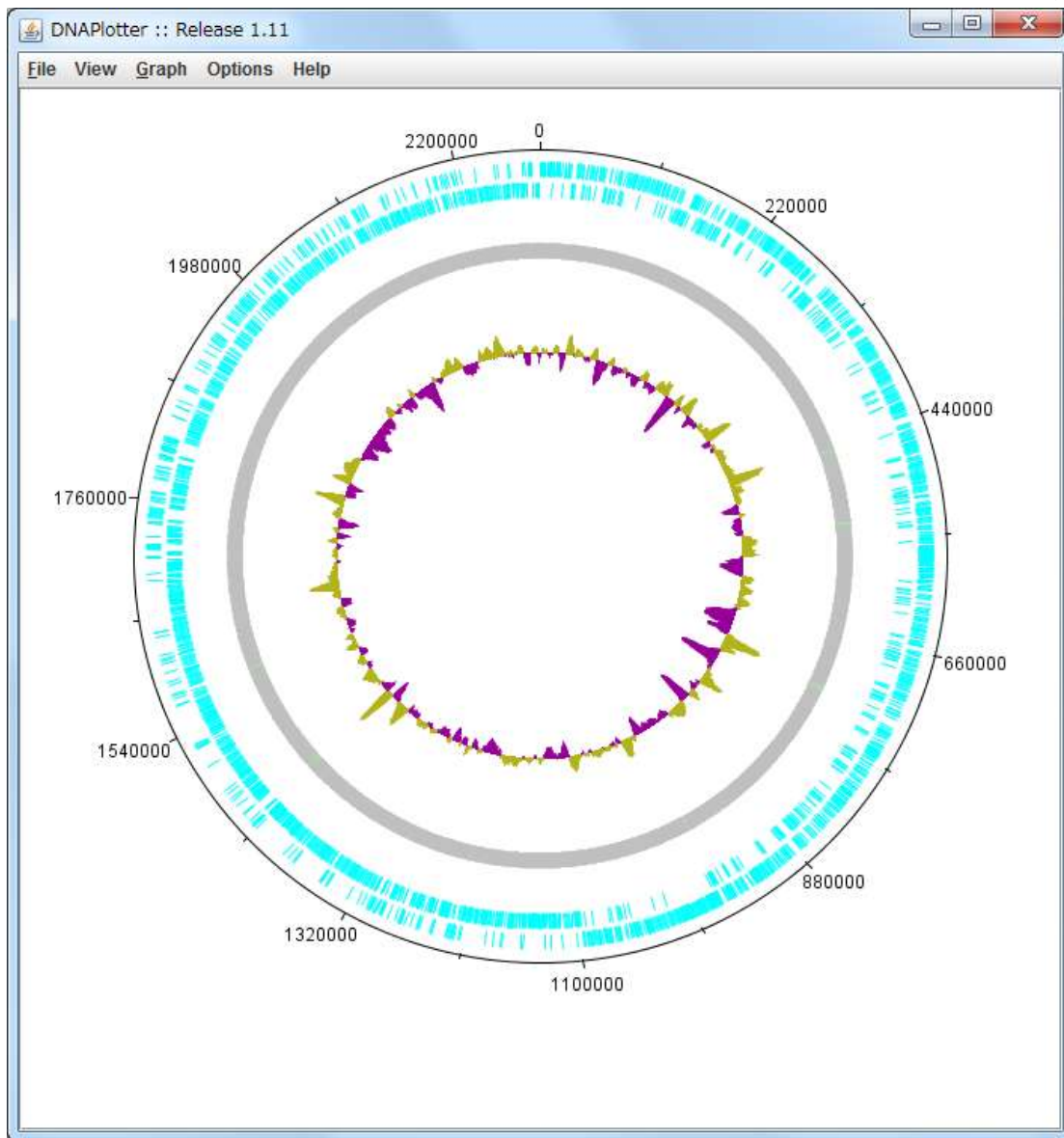


W12-13: Graph Height



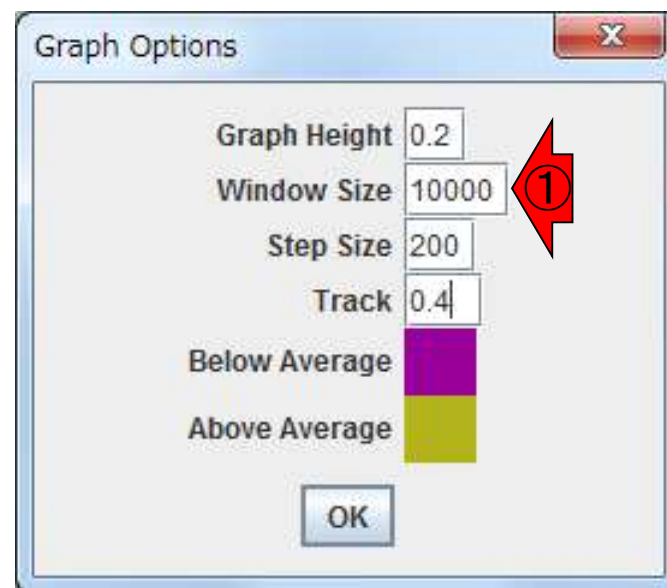
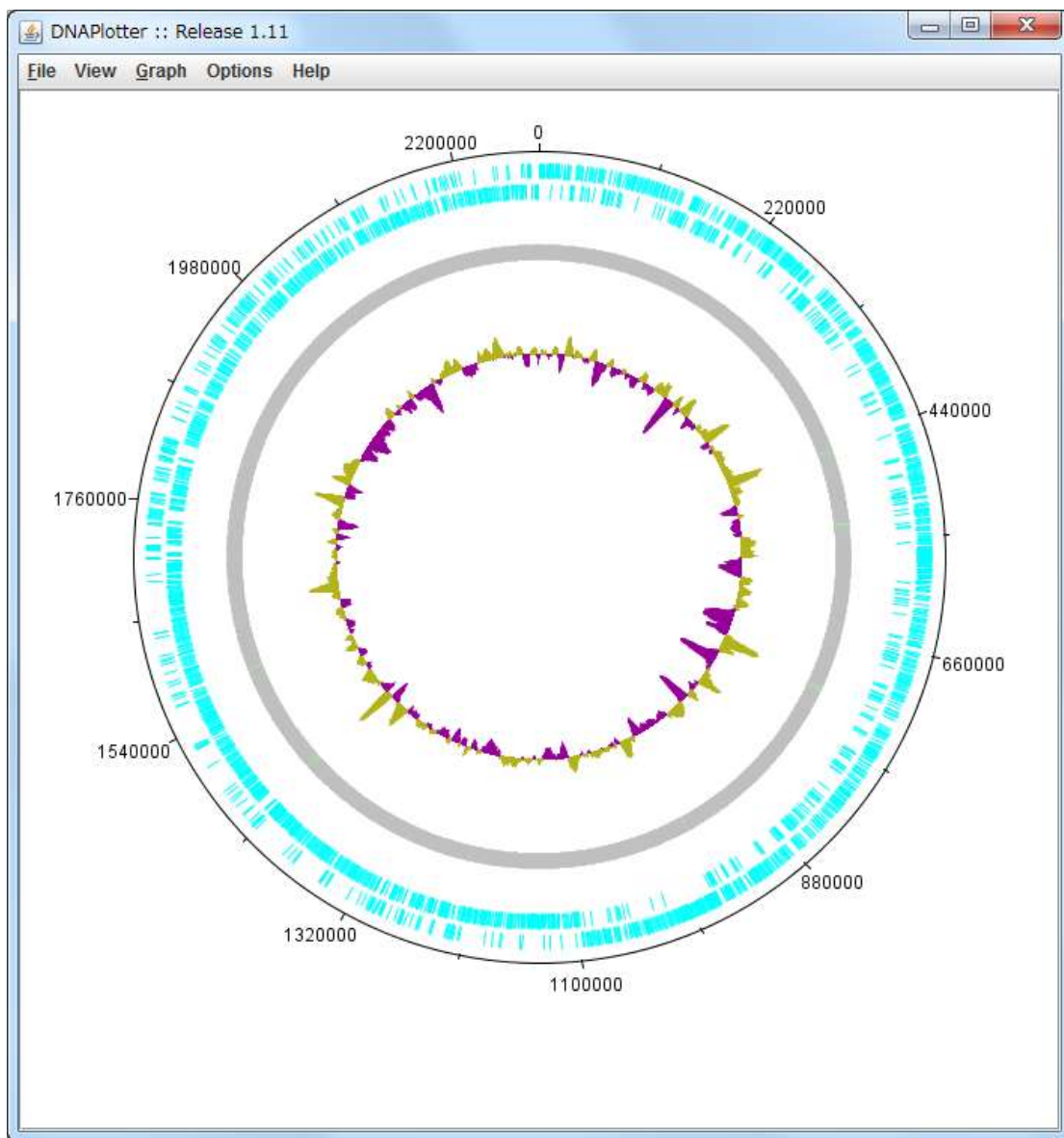
W12-14: デフォルトに戻す

一旦デフォルトの①0.2
と②0.4に戻して、③OK



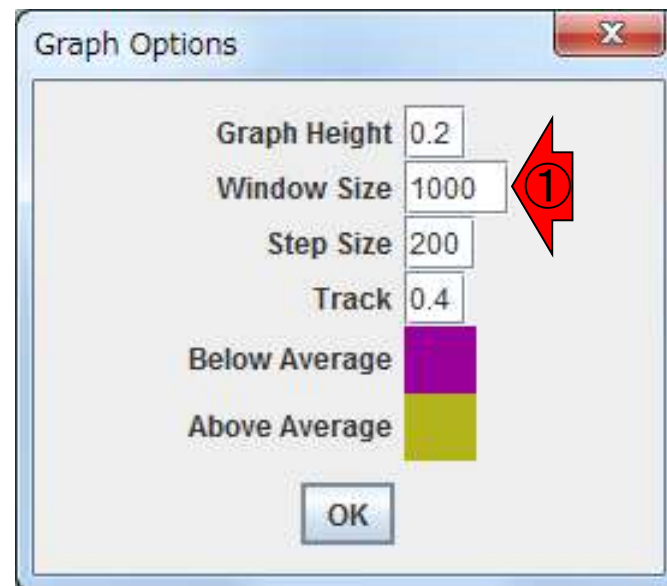
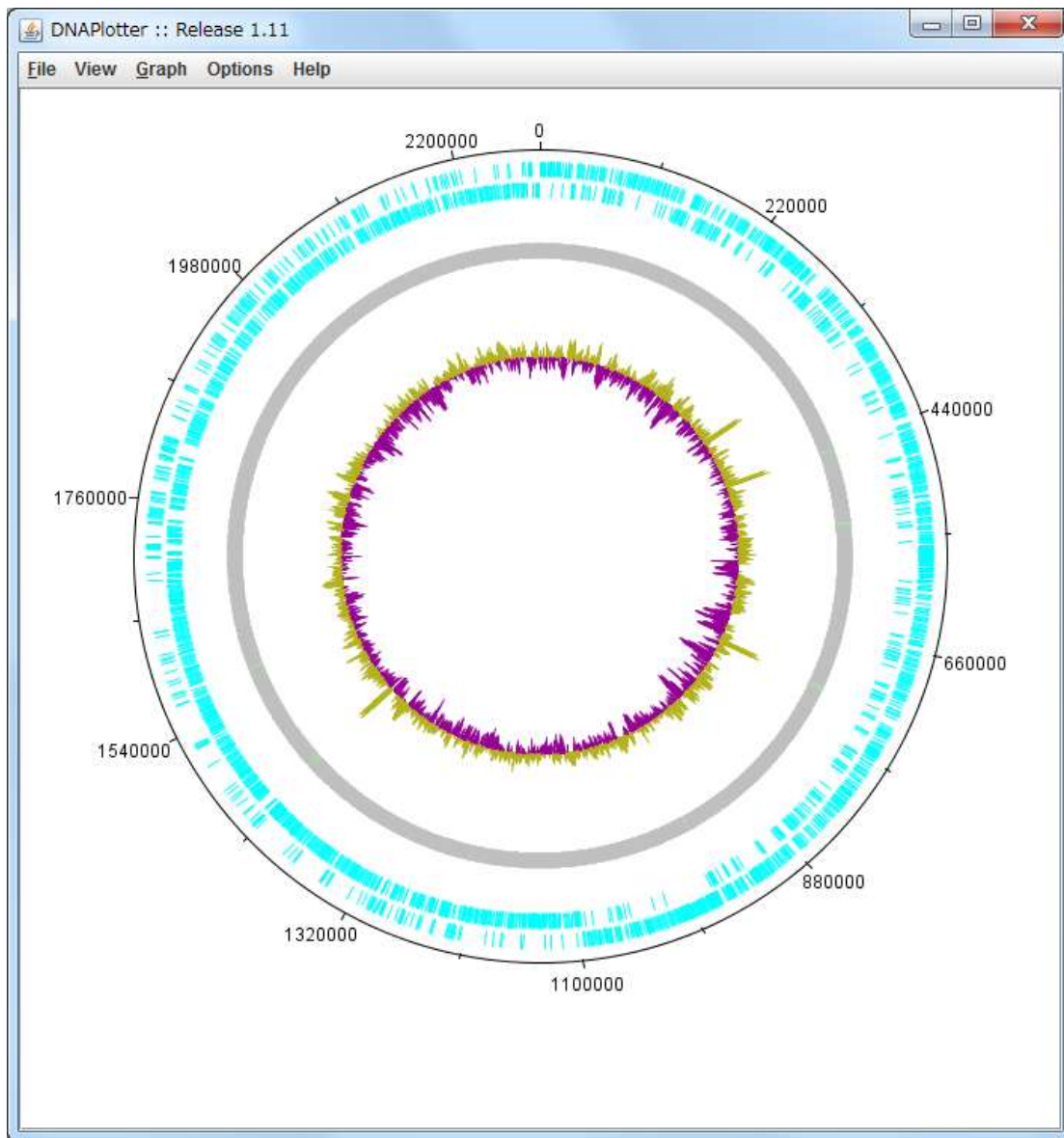
W12-15: Window Size

GC plotの描画目的は、どの領域がGC含量が高いかなど、GC含量分布の全体像を把握すること。①Window Sizeは、Window (つまり幅)を10000塩基として、その範囲の平均のGC含量を計算せよという指定



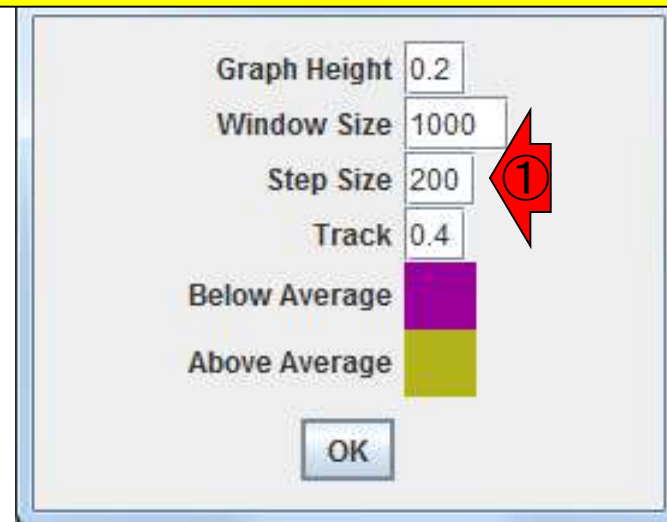
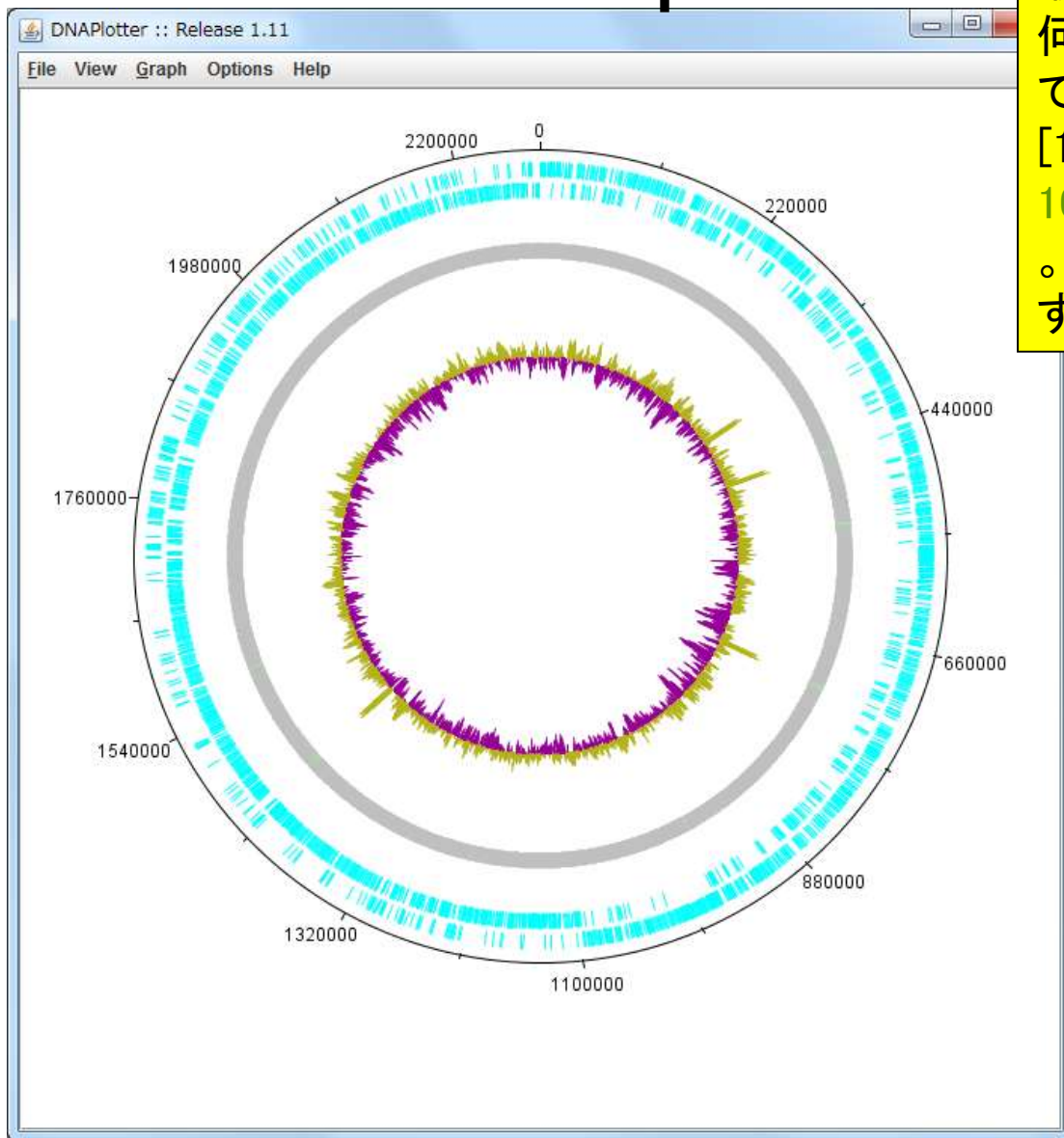
W12-15: Window Size

なので、例えば①Window Size = 1000にすると、その範囲が1/10になるので、得られるプロットの解像度が上がります。つまりなだらかさの度合いが減少する、という意味です



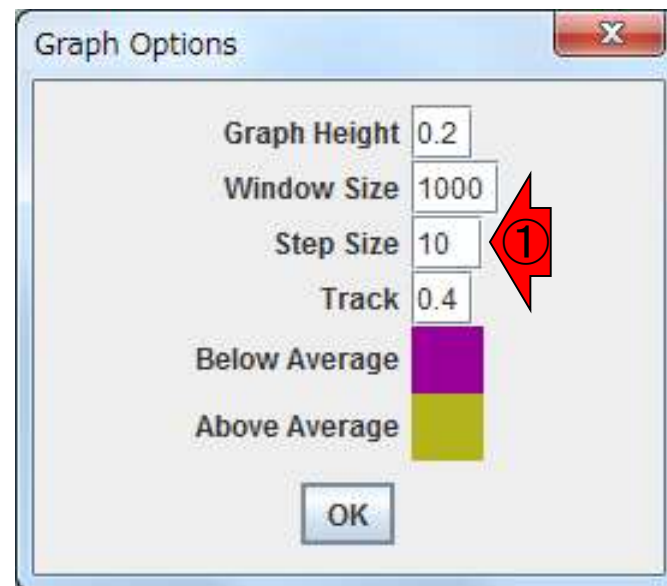
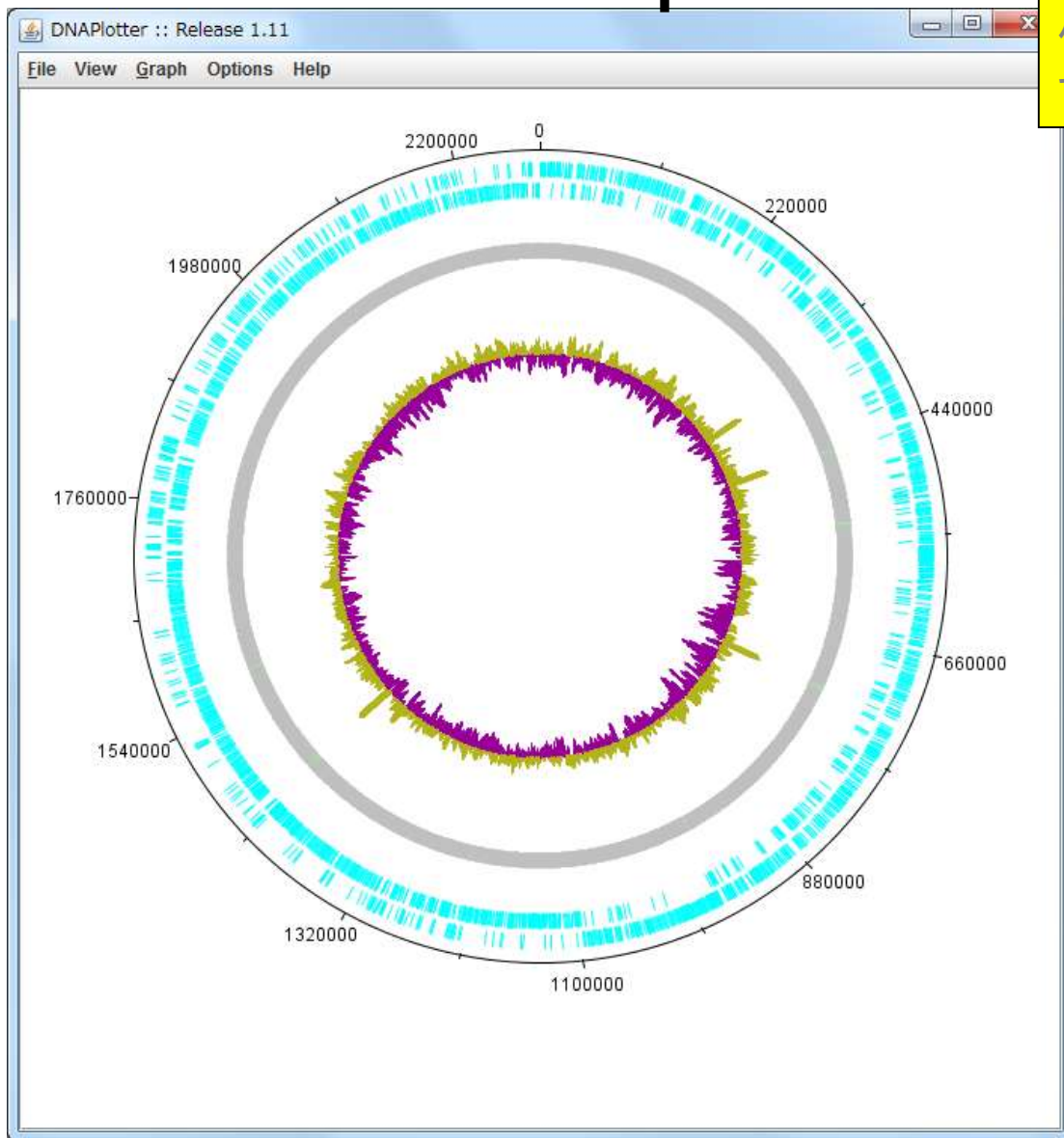
W12-16: Step Size

Step Sizeは、どれだけ計算をさぼるか、といった意味合いのパラメータです。具体的には、例えば最初に領域[1, 1000]でGC含量を計算し、次に何塩基分とばした位置から計算するかということです。Step Size = 200だと、次に計算する領域が[1+200, 1000+200]、そしてその次が[1+200+200, 1000+200+200]の領域を計算するという意味です。Step Sizeを小さくするほど正確な計算になりますが、その分だけ時間がかかります



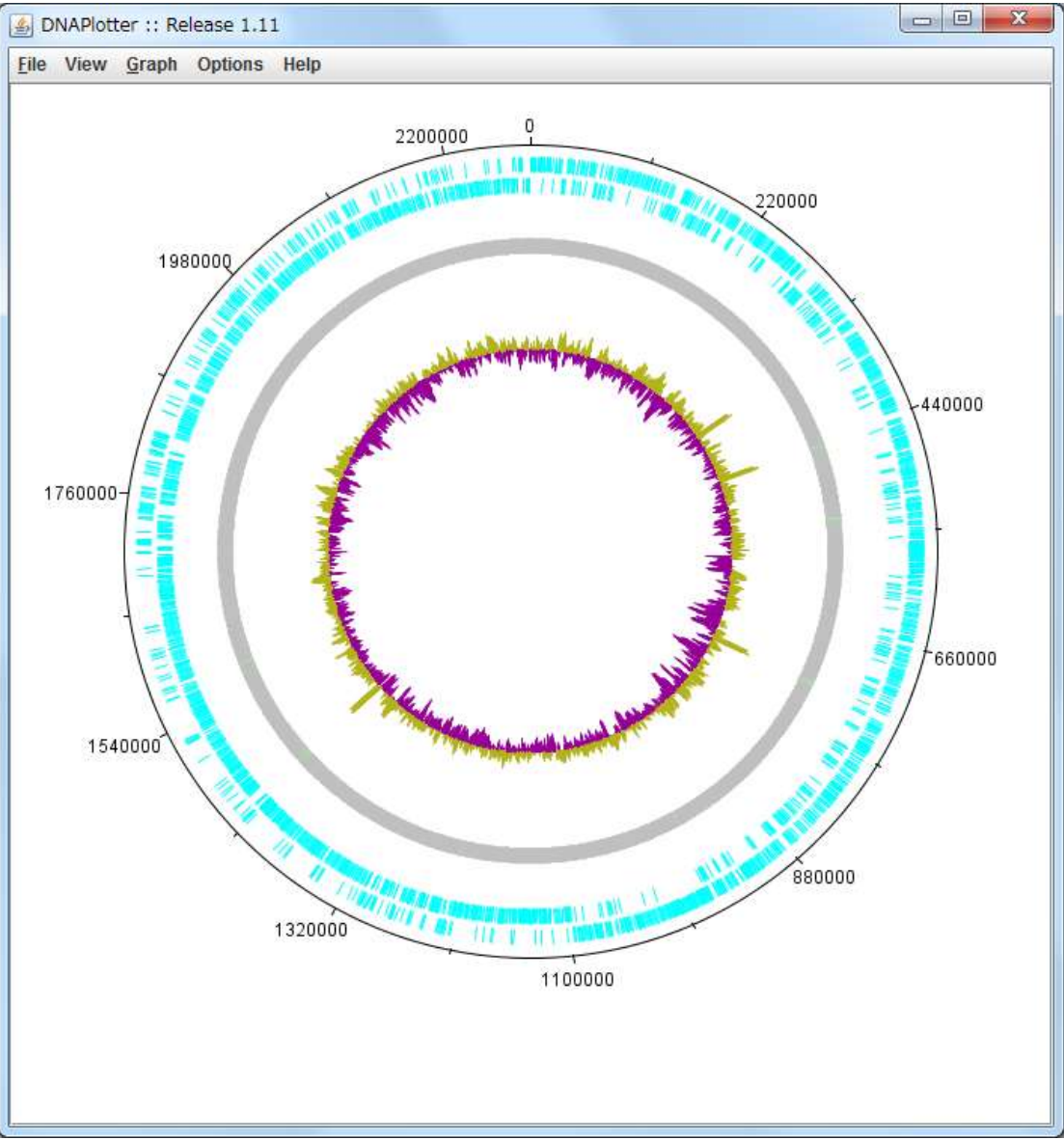
W12-16: Step Size

例えばStep Sizeを200から①10に変更することは、単純に計算時間が $200/10 = 20$ 倍かかることを意味します。このあたりはプログラム側のデフォルトに任せるのが一般的だと思いますが、気になるヒト用に解説したまでです



一旦デフォルトの①100
に戻して、②OK

W12-17: デフォルトに戻す



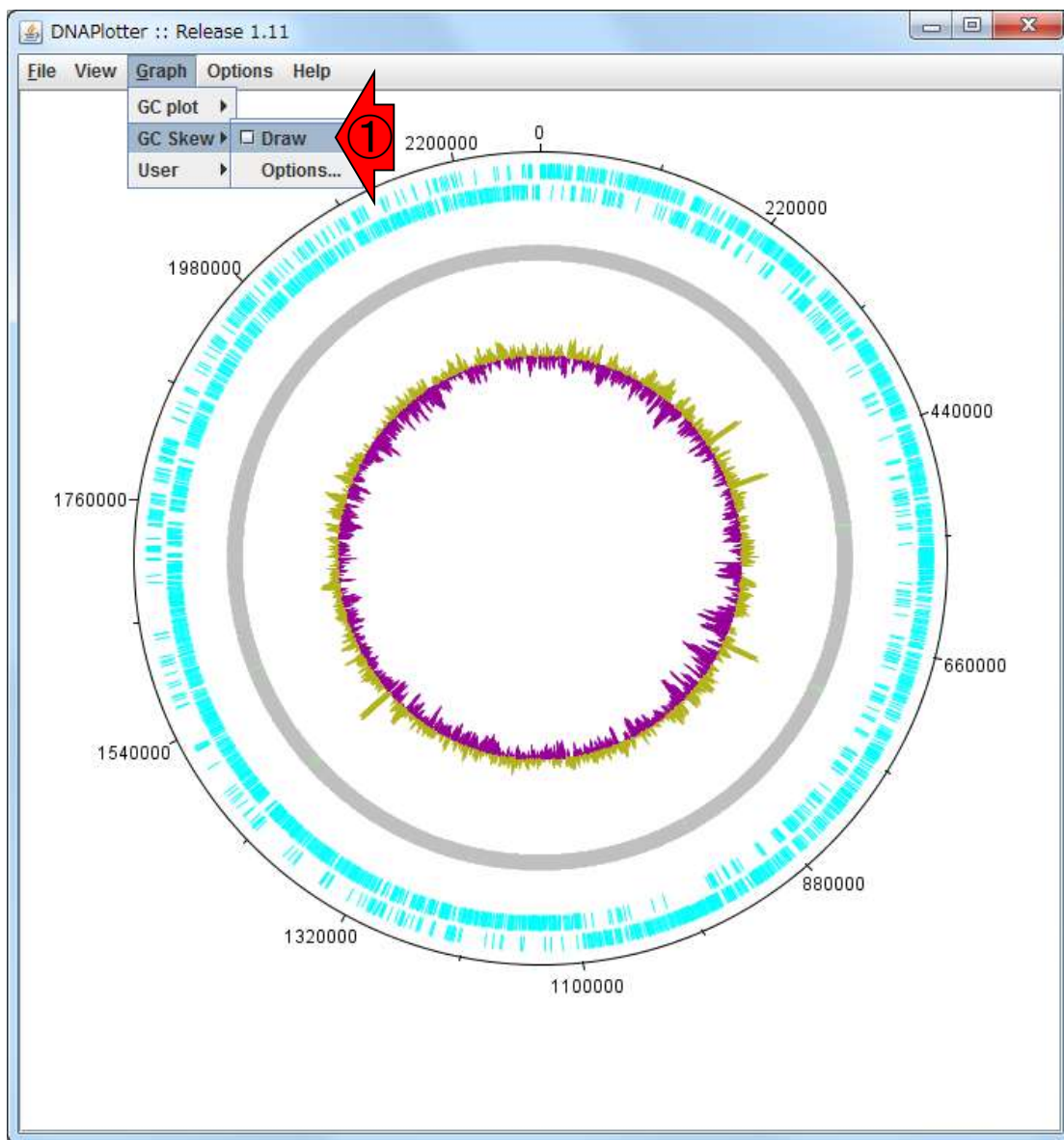
The "Graph Options" dialog box is shown with the following settings:

Graph Height	0.2
Window Size	1000
Step Size	100
Track	0.4
Below Average	
Above Average	

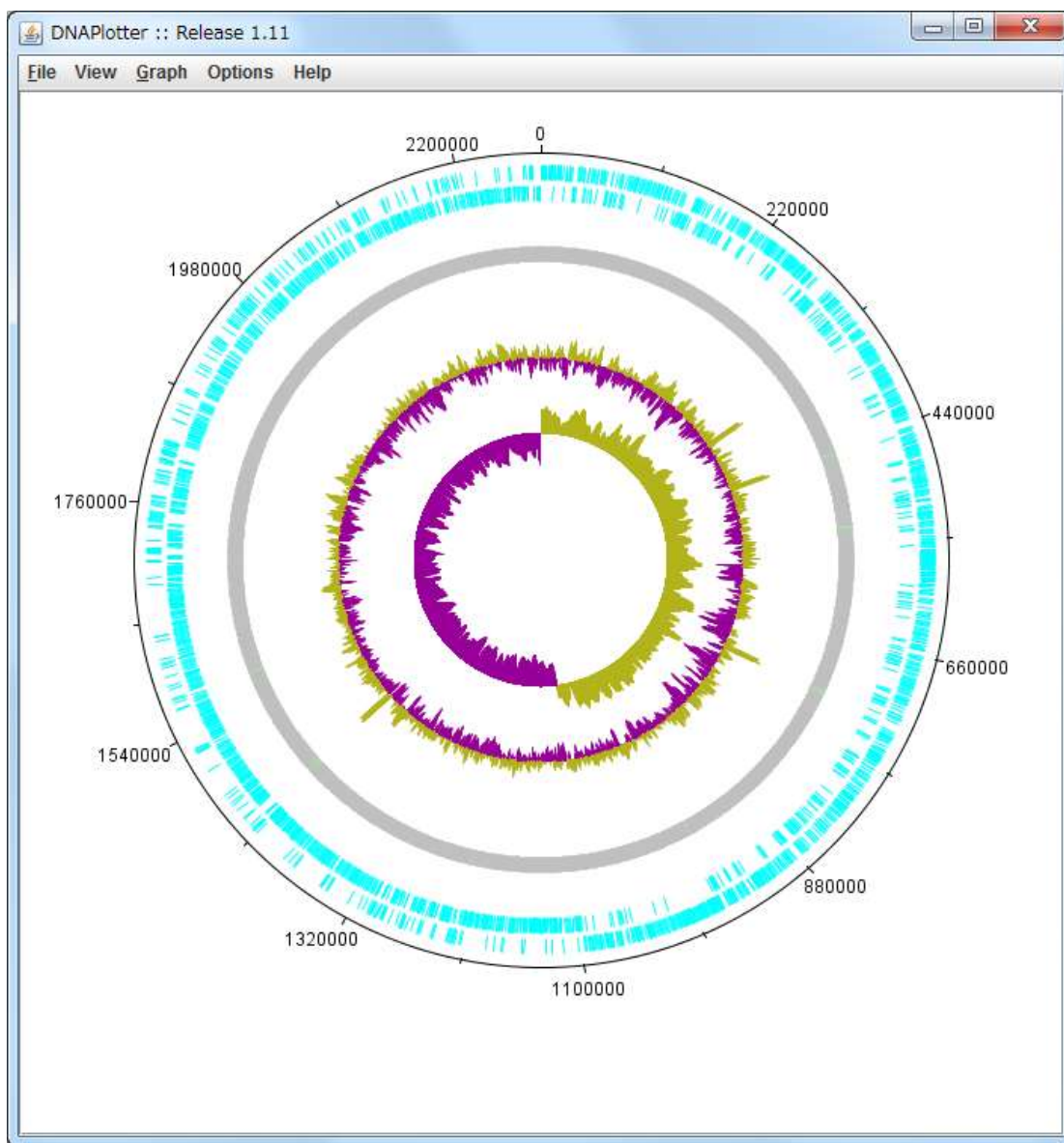
Red arrows with circled numbers 1 and 2 point to the "Step Size" field and the "OK" button, respectively.

W12-18: GC skew

①Graph - GC skew - Draw。これでGC skewを追加で描画することができます

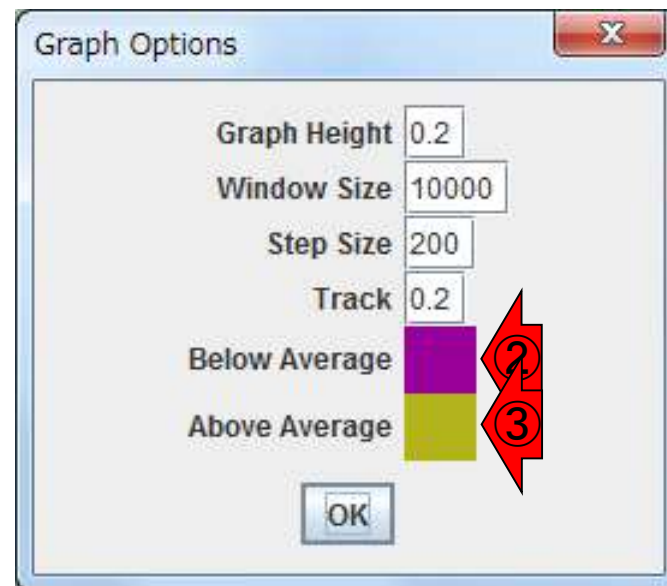
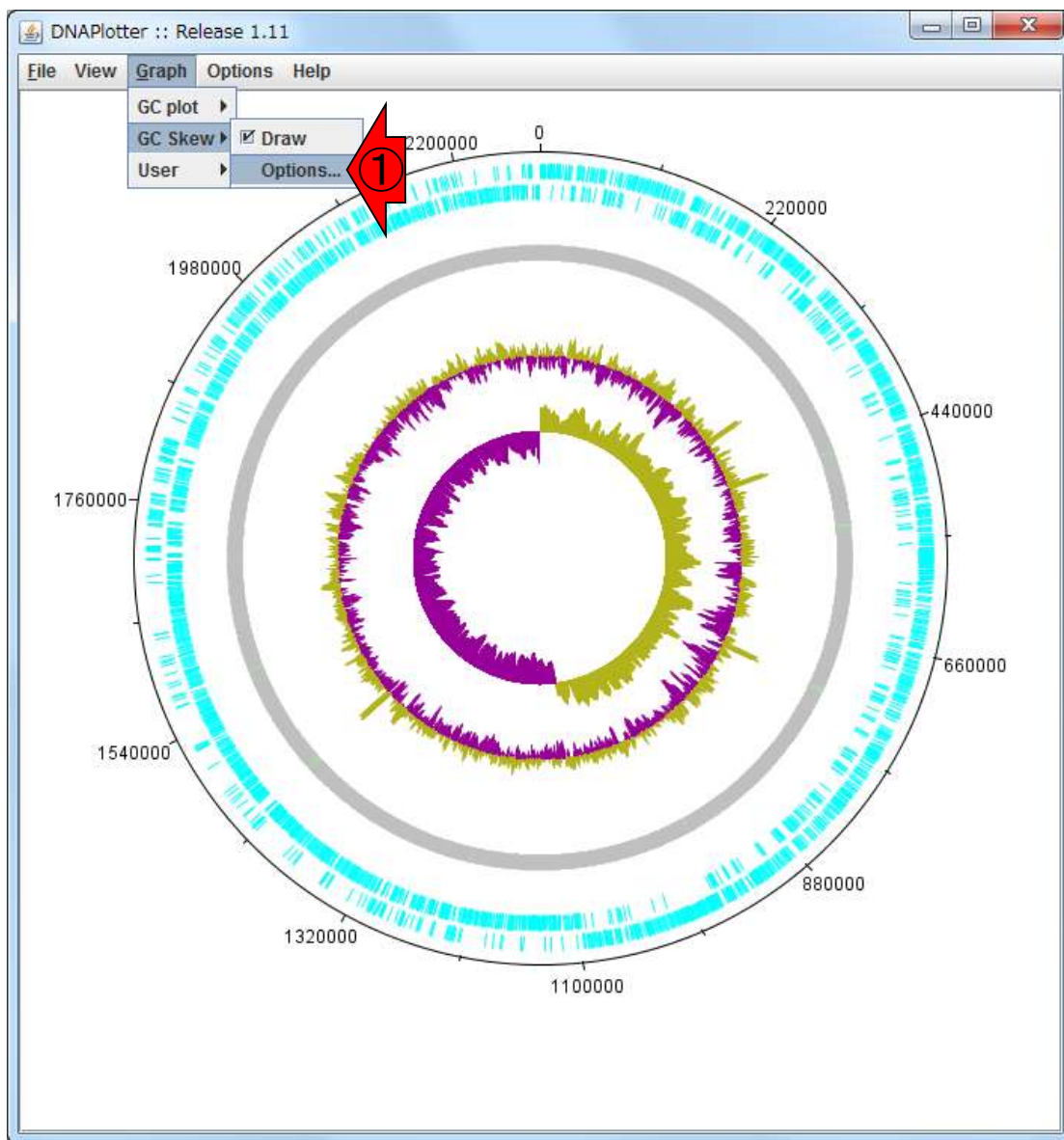


W12-18: GC skew



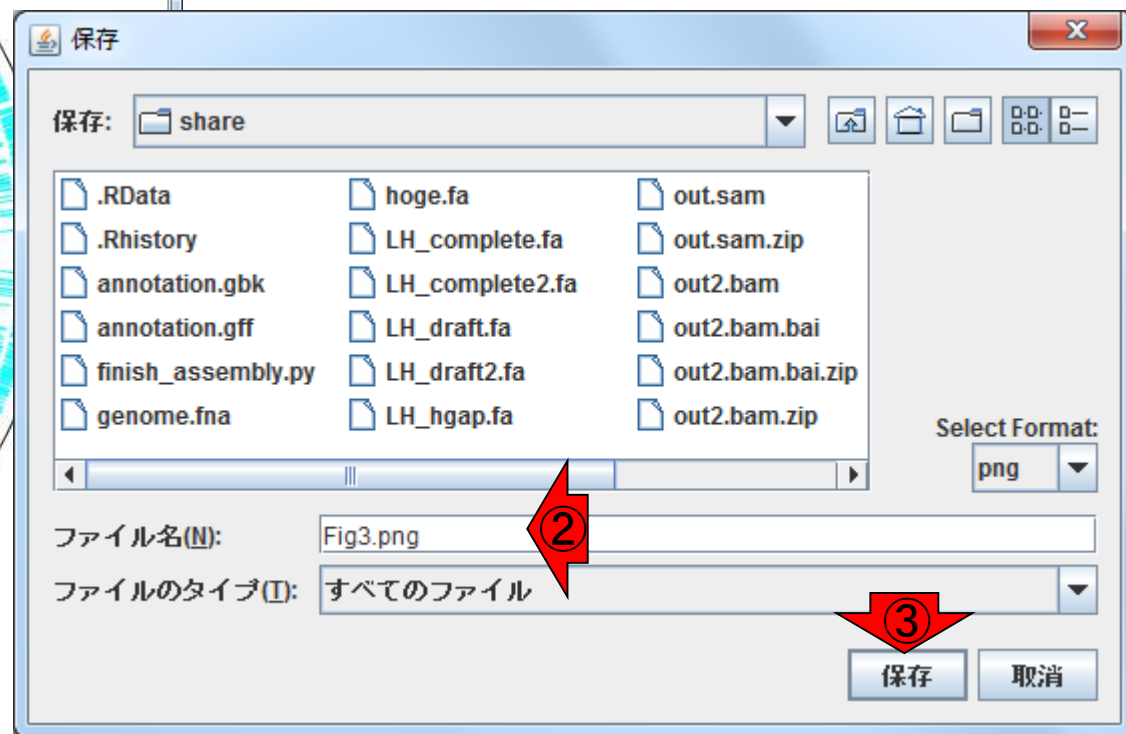
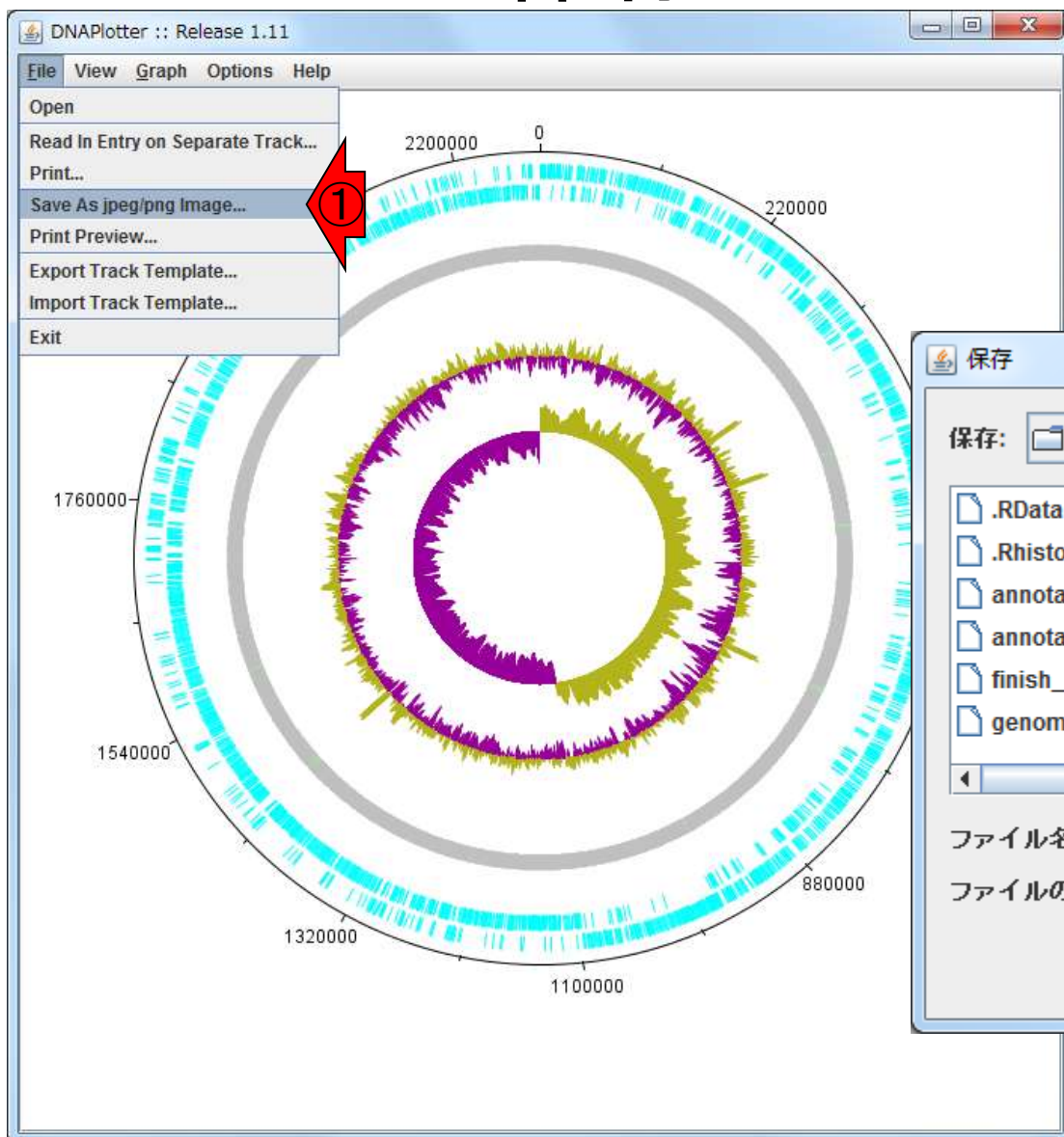
W12-18: GC skew

① Graph - GC skew - Options...で、色使いの意味を知ることができます。
② GC skewが平均よりも低い領域は赤紫色?!、③平均よりも高い領域は黄土色?!で表されていることがわかります



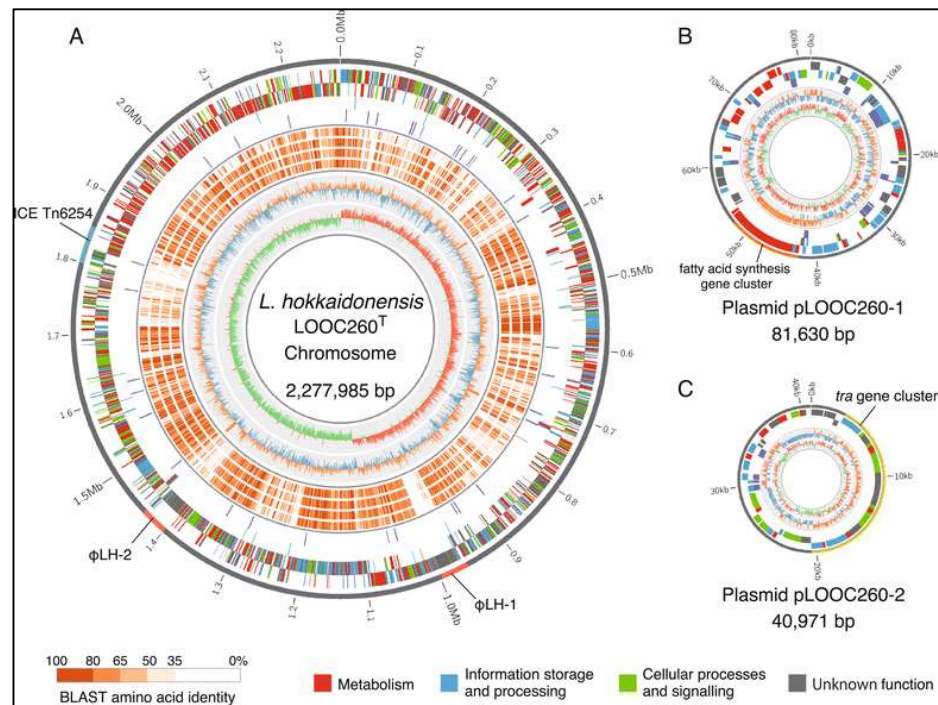
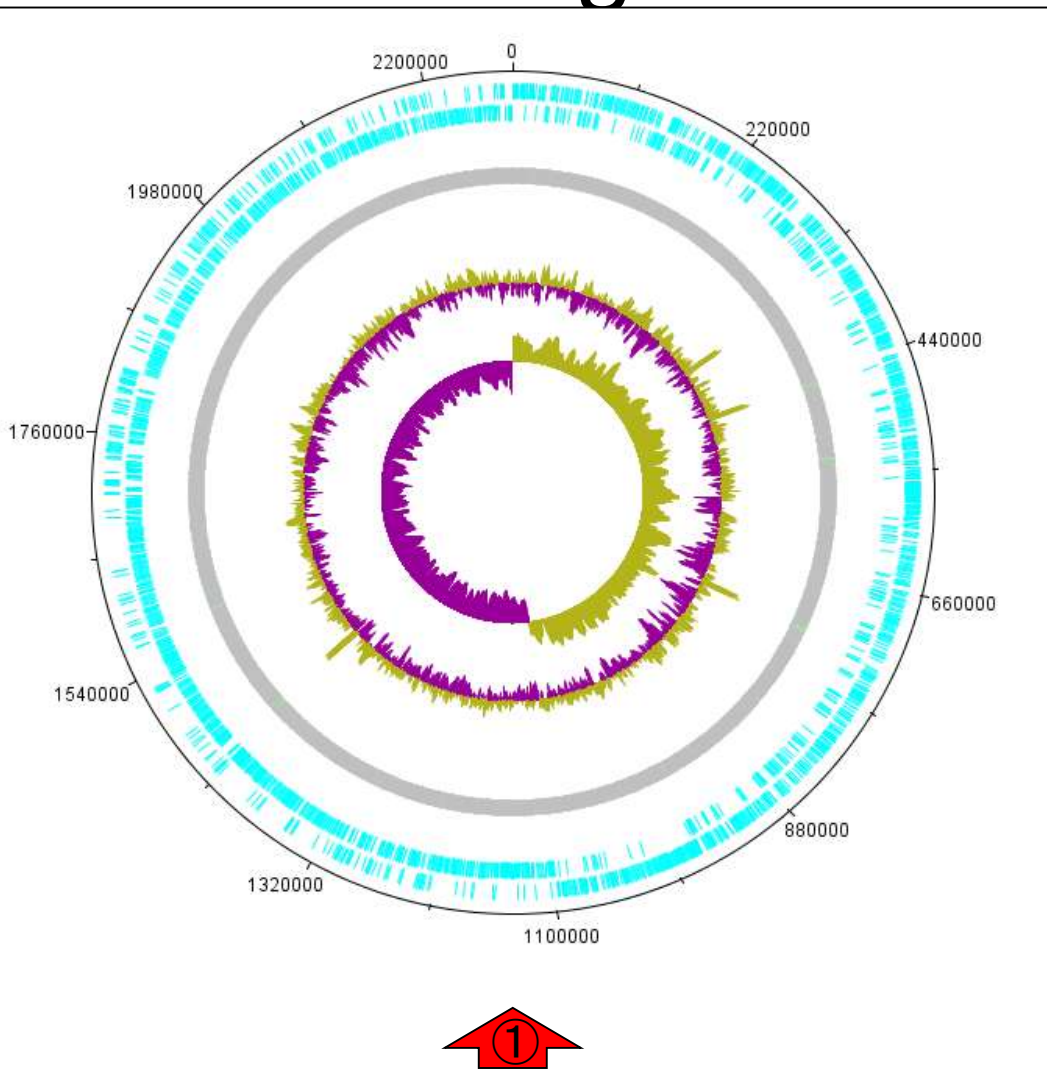
W12-19: 保存

①File – Save As jpeg/png Image...で保存
できます。ここでは②Fig3.pngとして保存
しました。原稿中のFig. 3と同じものです



W12-20: Fig. 3

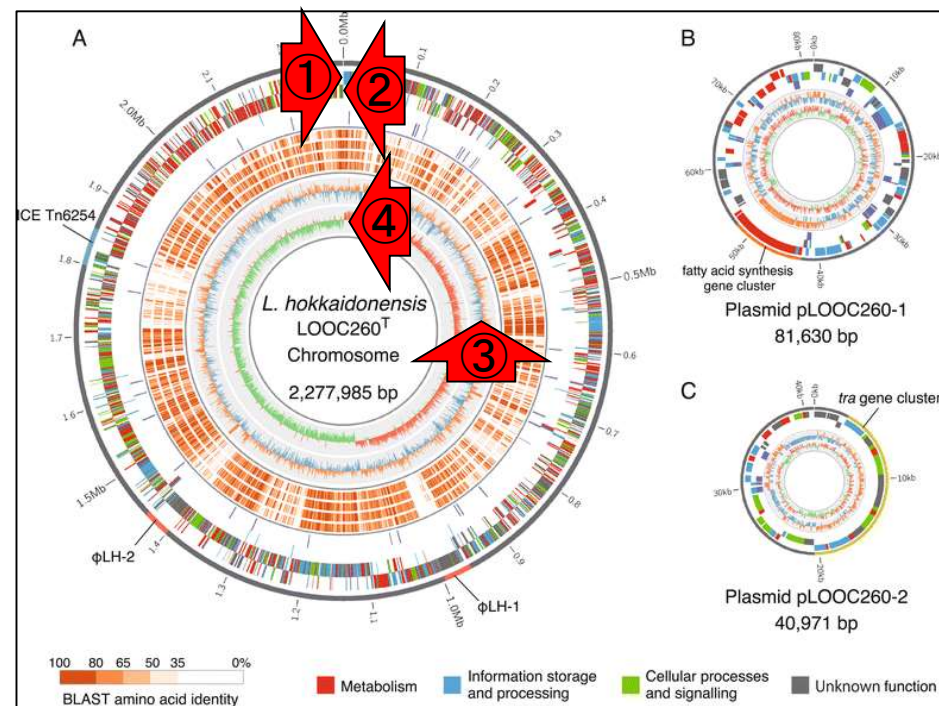
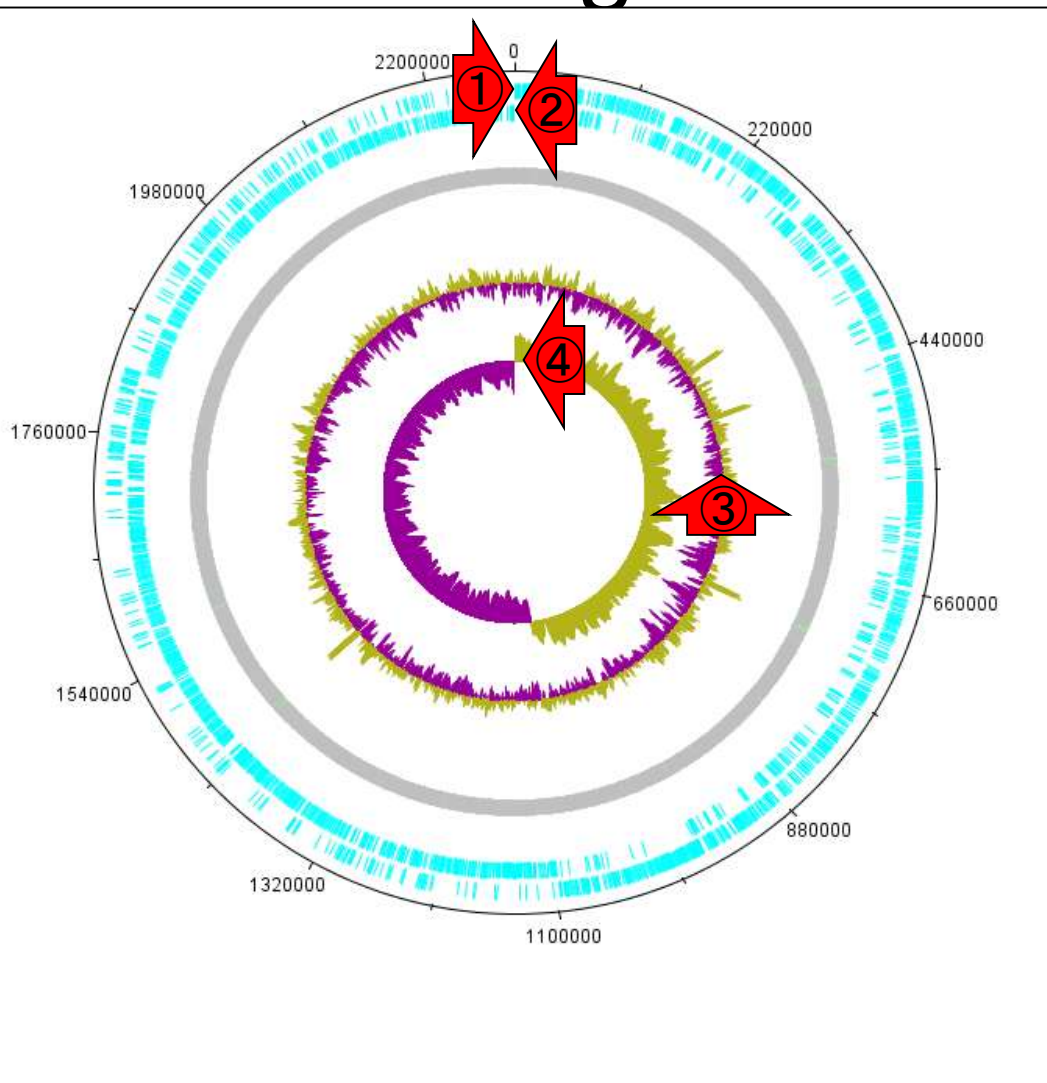
①第9回原稿中のFig. 3です。②原著論文のFig. 1と似た傾向になっている図が得られていることがわかります



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

W12-20: Fig. 3

具体的には、①②リーディング鎖上に多くのCDSが存在する傾向、③GC plot、④GC skewの分布などです



W13-1 : annotation.gbk

目的: 3配列分が含まれるannotation.gbkファイルのGenBank形式を把握し、個別の配列のみ抽出すること。まずはGenBank形式を把握するため、①headで概観。赤枠が実行結果部分

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l annotation.gbk
-rwxrwxrwx 1 iu iu 5241683 1月 21 12:39 annotation.gbk
iu@bielinux[mac_share] head annotation.gbk
LOCUS          sequence1                2277983 bp      DNA              BCT
20-JAN-2017
DEFINITION     Lactobacillus sp. strain unkown.
ACCESSION     sequence1
VERSION       sequence1
KEYWORDS       .
SOURCE        Lactobacillus sp
ORGANISM      Lactobacillus sp. Unclassified. .
COMMENT       Annotated using prokka 1.12-beta from
              https://github.com/tseemann/prokka.
iu@bielinux[mac_share]
```

[4:34午後]

[4:34午後]

[4:36午後]

①

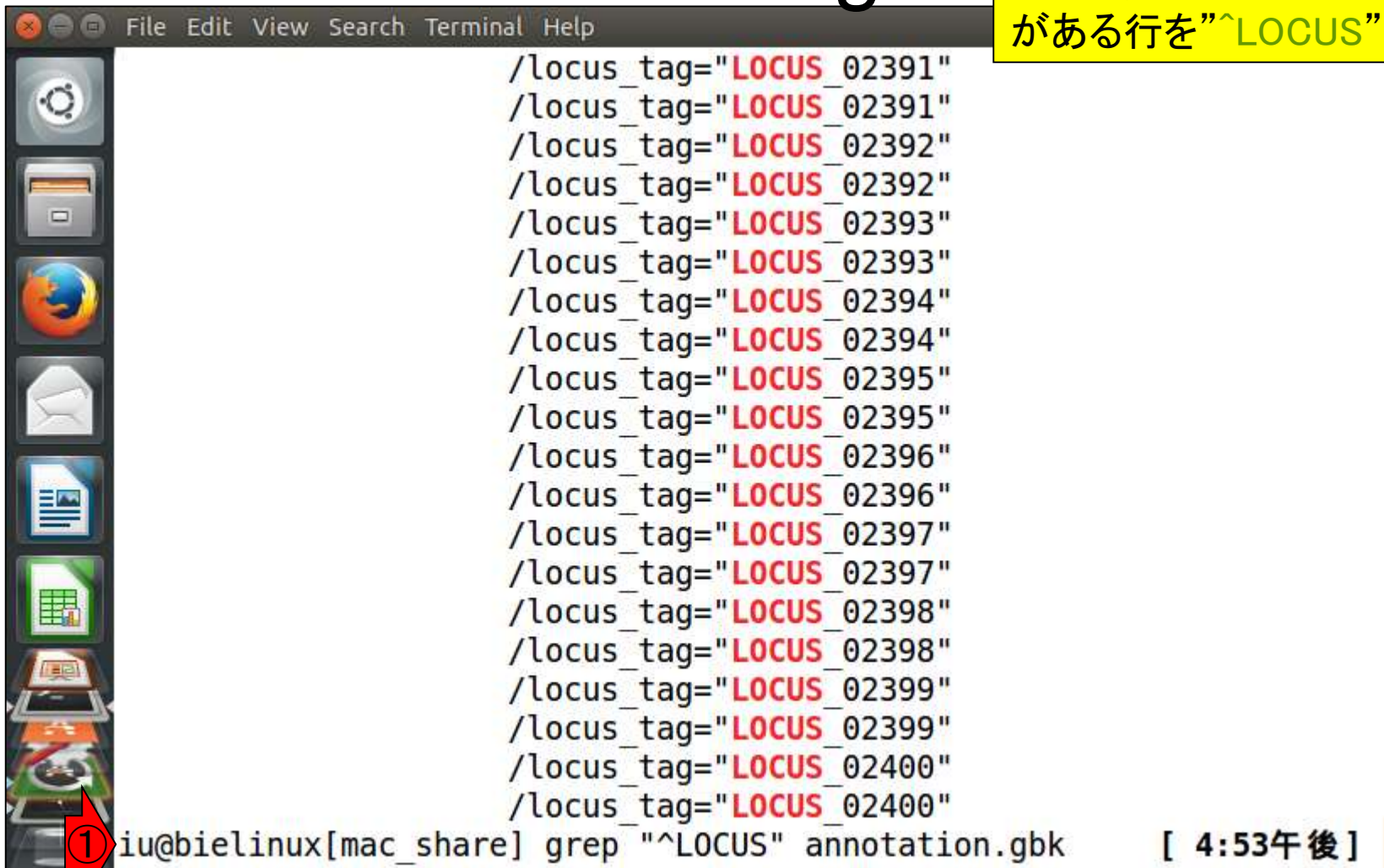
W13-1 : annotation.gbk

①1番最初の行がLOCUSという文字列を含んでいるので、②annotation.gbkに対して"LOCUS"という文字を含む行をgrepで表示

```
File Edit View Search Terminal Help
iu@bielinux[mac_share] pwd [ 4:34午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l annotation.gbk [ 4:34午後 ]
-rwxrwxrwx 1 iu iu 5241683 1月 21 12:39 annotation.gbk
iu@bielinux[mac_share] head annotation.gbk [ 4:34午後 ]
LOCUS          sequence1          2277983 bp      DNA          BCT
20-JAN-2017
DEFINITION    Lactobacillus sp. strain unkown.
ACCESSION     sequence1
VERSION       sequence1
KEYWORDS      .
SOURCE        Lactobacillus sp
ORGANISM      Lactobacillus sp. Unclassified. .
COMMENT       Annotated using prokka 1.12-beta from
              https://github.com/tseemann/prokka.
iu@bielinux[mac_share] grep "LOCUS" annotation.gbk [ 4:36午後 ]
```

W13-1 : annotation.gbk

LOCUSを含む行は、annotation.gbk中に大量に存在するようだ。これだと配列間の区切りが分からないので、①行頭にLOCUSがある行を"^LOCUS"としてgrepする



```
File Edit View Search Terminal Help
/locus_tag="LOCUS_02391"
/locus_tag="LOCUS_02391"
/locus_tag="LOCUS_02392"
/locus_tag="LOCUS_02392"
/locus_tag="LOCUS_02393"
/locus_tag="LOCUS_02393"
/locus_tag="LOCUS_02394"
/locus_tag="LOCUS_02394"
/locus_tag="LOCUS_02395"
/locus_tag="LOCUS_02395"
/locus_tag="LOCUS_02396"
/locus_tag="LOCUS_02396"
/locus_tag="LOCUS_02397"
/locus_tag="LOCUS_02397"
/locus_tag="LOCUS_02398"
/locus_tag="LOCUS_02398"
/locus_tag="LOCUS_02399"
/locus_tag="LOCUS_02399"
/locus_tag="LOCUS_02400"
/locus_tag="LOCUS_02400"
① iu@bielinux[mac_share] grep "^LOCUS" annotation.gbk [ 4:53午後 ]
```

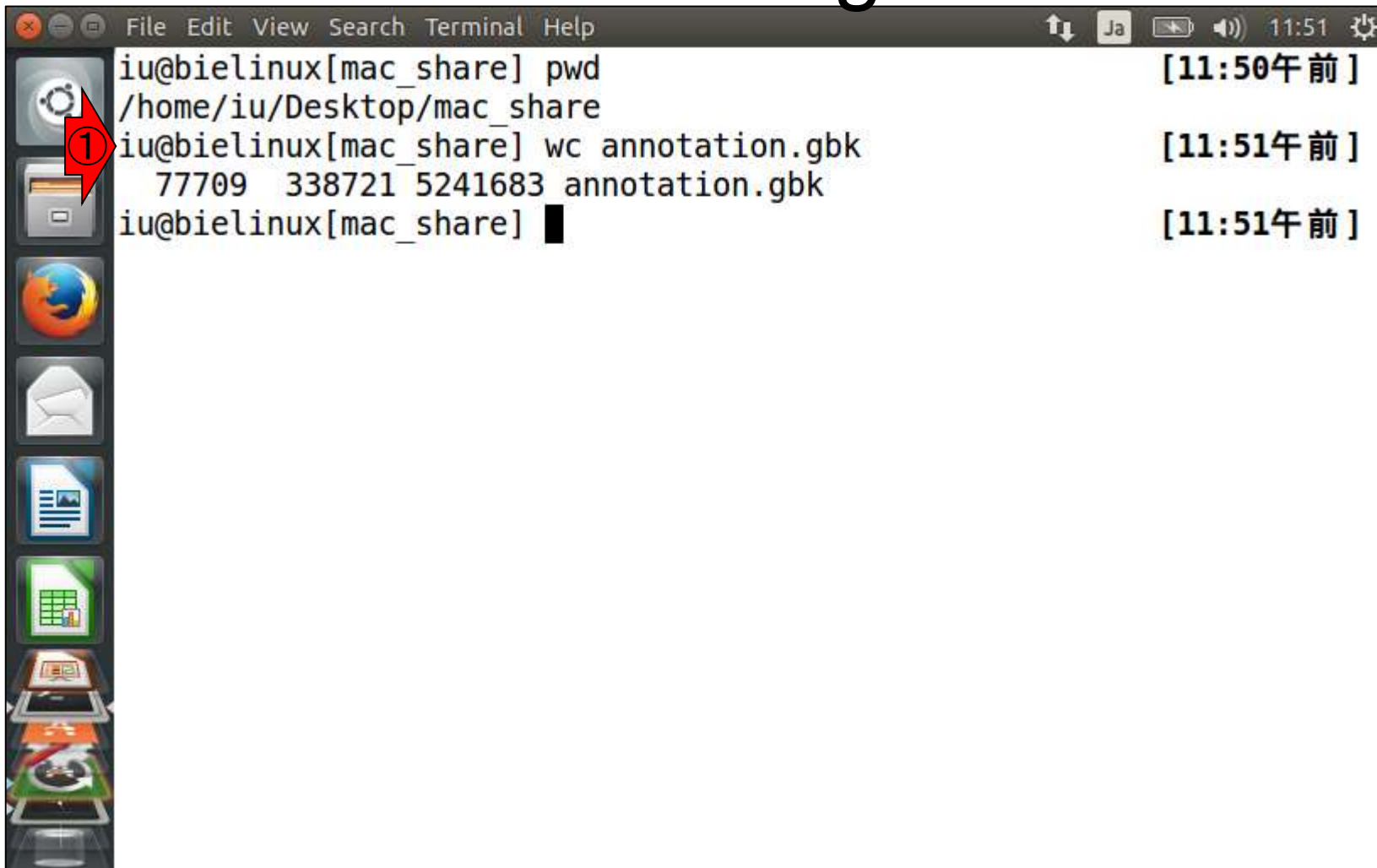
W13-1 : annotation.gb

①行頭にLOCUSがある行を”^LOCUS”としてgrepした結果。配列ごとの最初の行を得るのは、”^LOCUS”で十分だということが分かった

```
File Edit View Search Terminal Help 11:42
/locus_tag="LOCUS_02394"
/locus_tag="LOCUS_02395"
/locus_tag="LOCUS_02395"
/locus_tag="LOCUS_02396"
/locus_tag="LOCUS_02396"
/locus_tag="LOCUS_02397"
/locus_tag="LOCUS_02397"
/locus_tag="LOCUS_02398"
/locus_tag="LOCUS_02398"
/locus_tag="LOCUS_02399"
/locus_tag="LOCUS_02399"
/locus_tag="LOCUS_02400"
/locus_tag="LOCUS_02400"
① iu@bielinux[mac_share] grep "^LOCUS" annotation.gbk [ 4:53午後 ]
LOCUS      sequence1      2277983 bp      DNA      BCT
20-JAN-2017
LOCUS      sequence2      81630 bp      DNA      BCT
20-JAN-2017
LOCUS      sequence3      40971 bp      DNA      BCT
20-JAN-2017
iu@bielinux[mac_share] [11:23午前]
```


W13-2: annotation.gbk

①wcコマンドで、annotation.gbkの
行数が77709行であることを把握



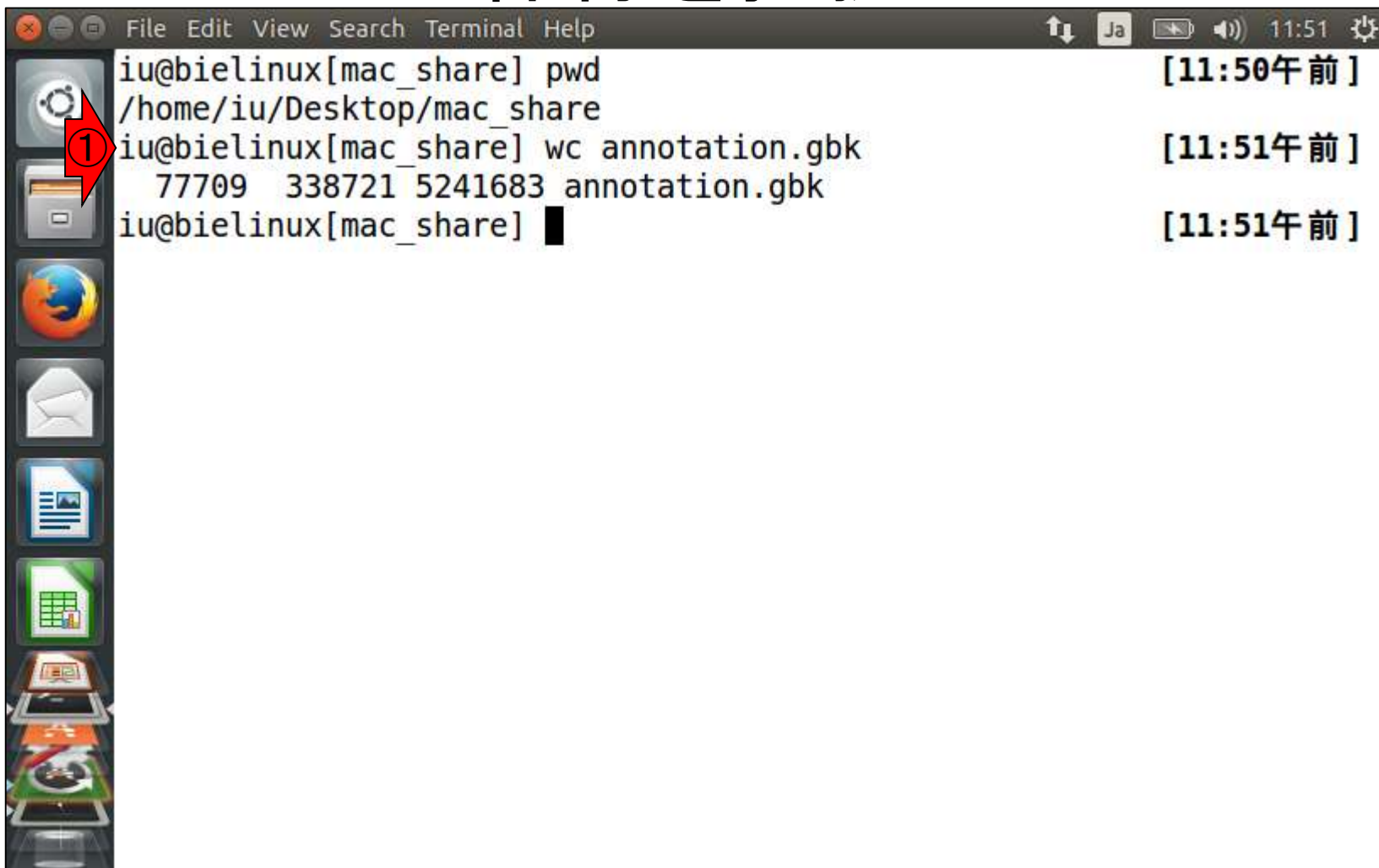
The image shows a terminal window with a dark background and a light-colored text area. The window title bar includes 'File Edit View Search Terminal Help' and system icons for volume, network, and battery. The terminal content shows the following sequence of commands and outputs:

```
iu@bielinux[mac_share] pwd [11:50午前]
/home/iu/Desktop/mac_share
① iu@bielinux[mac_share] wc annotation.gbk [11:51午前]
77709 338721 5241683 annotation.gbk
iu@bielinux[mac_share] [11:51午前]
```

A red arrow with the number '1' points to the 'wc' command line. The left sidebar of the terminal window contains various application icons, including a gear, a folder, a globe, an envelope, a document, a spreadsheet, and a presentation.

W13-2: 全体像を把握

①wcコマンドで、annotation.gbkの
行数が77709行であることを把握



The image shows a terminal window with a dark background and a light-colored text area. The window title bar includes 'File Edit View Search Terminal Help' and system icons for volume, network, and battery. The terminal content is as follows:

```
iu@bielinux[mac_share] pwd [11:50午前]
/home/iu/Desktop/mac_share
① iu@bielinux[mac_share] wc annotation.gbk [11:51午前]
77709 338721 5241683 annotation.gbk
iu@bielinux[mac_share] [11:51午前]
```

A red arrow with the number '1' points to the second line of the terminal output, which is the command 'wc annotation.gbk'.

W13-2: 全体像を把握

①grep -nで^LOCUSを満たす行番号を把握(第6回W15-5)。sequence1が1行目から、sequence2が73530行目から、そしてsequence3が76315行目からスタートしていることがわかる

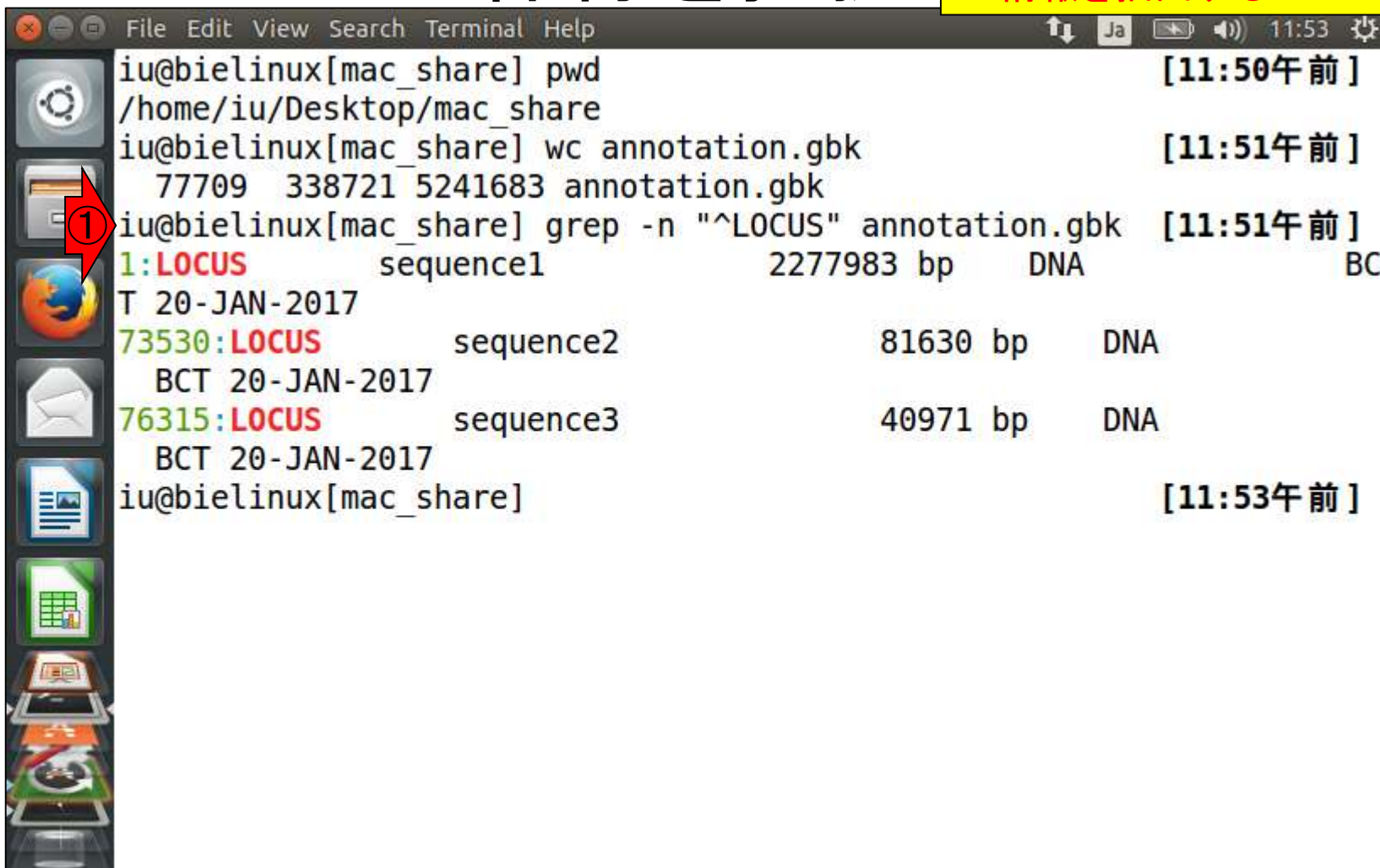
```
iu@bielinux[mac_share] pwd [11:50午前]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation.gbk [11:51午前]
 77709  338721 5241683 annotation.gbk
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [11:51午前]
1:LOCUS          sequence1          2277983 bp    DNA          BC
T 20-JAN-2017
73530:LOCUS      sequence2          81630 bp    DNA
BCT 20-JAN-2017
76315:LOCUS      sequence3          40971 bp    DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] [11:53午前]
```



①

W13-2: 全体像を把握

実はここまでの情報を用いれば、**head**と**tail**コマンドを組み合わせて目的の配列のみの情報を抽出することが可能(第3回W19-3)



```
iu@bielinux[mac_share] pwd [11:50午前]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation.gbk [11:51午前]
 77709  338721 5241683 annotation.gbk
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [11:51午前]
1:LOCUS          sequence1          2277983 bp      DNA          BC
T 20-JAN-2017
73530:LOCUS      sequence2          81630 bp      DNA
BCT 20-JAN-2017
76315:LOCUS      sequence3          40971 bp      DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] [11:53午前]
```

W13-3: さらに把握

①grep -A 2で、2行分だけ下側を余分に表示させている(第6回W15-6)。

```
iu@bielinux[mac_share] pwd [12:31午後]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -A 2 "^LOCUS" annotation.gbk
LOCUS          sequence1          2277983 bp      DNA              BCT
20-JAN-2017
DEFINITION      Lactobacillus sp. strain unkown.
ACCESSION       sequence1
--
LOCUS          sequence2          81630 bp       DNA              BCT
20-JAN-2017
DEFINITION      Lactobacillus sp. strain unkown.
ACCESSION       sequence2
--
LOCUS          sequence3          40971 bp       DNA              BCT
20-JAN-2017
DEFINITION      Lactobacillus sp. strain unkown.
ACCESSION       sequence3
iu@bielinux[mac_share] [12:31午後]
```



①

W13-3: さらに把握

```
iu@bielinux[mac_share] pwd [12:33午後]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n -A 2 "^LOCUS" annotation.gbk
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
2-DEFINITION    Lactobacillus sp. strain unkown.
3-ACCESSION     sequence1
--
73530:LOCUS     sequence2          81630 bp      DNA
BCT 20-JAN-2017
73531-DEFINITION Lactobacillus sp. strain unkown.
73532-ACCESSION sequence2
--
76315:LOCUS     sequence3          40971 bp      DNA
BCT 20-JAN-2017
76316-DEFINITION Lactobacillus sp. strain unkown.
76317-ACCESSION sequence3
iu@bielinux[mac_share] [12:33午後]
```

W13-3: さらに把握

よく見ると、①”^LOCUS”との一致行は
:となっていて、②-A 2に相当する下側
2行分は-となっている。芸が細かいw

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n -A 2 "^LOCUS" annotation.gbk
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
2-DEFINITION    Lactobacillus sp. strain unkown.
3-ACCESSION     sequence1
--
73530:LOCUS     sequence2          81630 bp      DNA
BCT 20-JAN-2017
73531-DEFINITION Lactobacillus sp. strain unkown.
73532-ACCESSION sequence2
--
76315:LOCUS     sequence3          40971 bp      DNA
BCT 20-JAN-2017
76316-DEFINITION Lactobacillus sp. strain unkown.
76317-ACCESSION sequence3
iu@bielinux[mac_share]
```

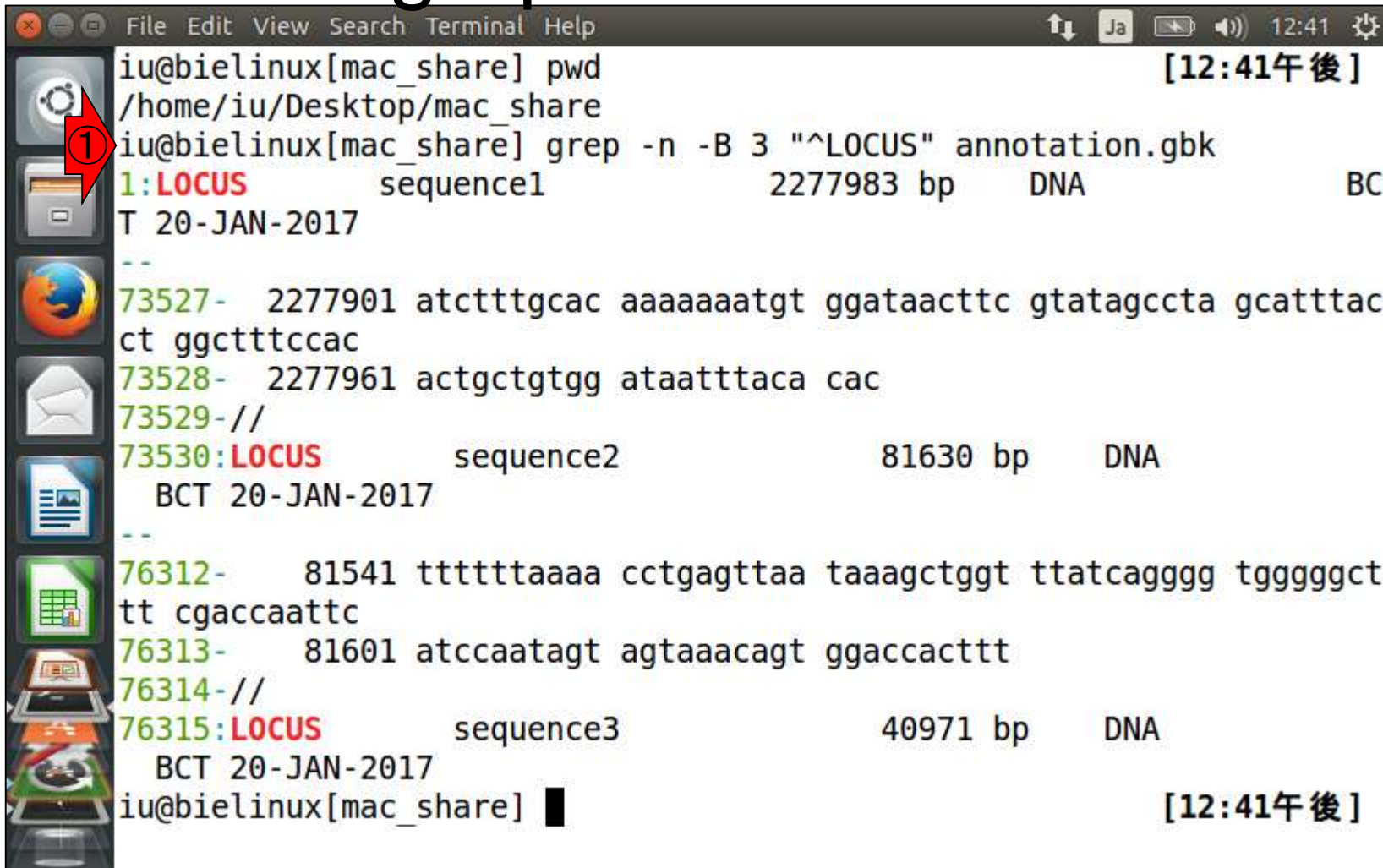


[12:33午後]

[12:33午後]

①grep -B 3で、3行分だけ一致行の上側を余分に表示させている

W13-4: grep -B



```
iu@bielinux[mac_share] pwd [12:41午後]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n -B 3 "^LOCUS" annotation.gbk
1:LOCUS          sequence1          2277983 bp      DNA          BC
T 20-JAN-2017
--
73527-  2277901 atctttgcac aaaaaaatgt ggataacttc gtatagccta gcatttac
ct ggctttccac
73528-  2277961 actgctgtgg ataattaca cac
73529-//
73530:LOCUS          sequence2          81630 bp      DNA
BCT 20-JAN-2017
--
76312-  81541 ttttttaaaa cctgagtaa taaagctggt ttatcagggg tgggggct
tt cgaccaattc
76313-  81601 atccaatagt agtaaacagt ggaccacttt
76314-//
76315:LOCUS          sequence3          40971 bp      DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] █ [12:41午後]
```


W13-4: grep -B

①sequence1に相当する1行目の上側はないので妥当。②sequence2と③sequence3の上側3行分が表示されていることがわかる

```
iu@bielinux[mac_share] pwd [12:41午後]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n -B 3 "^LOCUS" annotation.gbk
1:LOCUS          sequence1          2277983 bp      DNA          BC
T 20-JAN-2017
--
73527-  2277901 atctttgcac aaaaaaatgt ggataacttc gtatagccta gcatttac
ct ggctttccac
73528-  2277961 actgctgtgg ataattaca cac
73529-//
73530:LOCUS          sequence2          81630 bp      DNA
BCT 20-JAN-2017
--
76312-  81541 ttttttaaaa cctgagttaa taaagctggt ttatcagggg tgggggct
tt cgaccaattc
76313-  81601 atccaatagt agtaaacagt ggaccacttt
76314-//
76315:LOCUS          sequence3          40971 bp      DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] [12:41午後]
```



W13-4: grep -B

①//が各配列の最後を表すのでしょ。本当はGenBank形式がこうなっていることを実務担当者は知っているので、厳密にはそうだったことを確認しているだけ

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n -B 3 "^LOCUS" annotation.gbk
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
--
73527-  2277901 atctttgcac aaaaaaatgt ggataacttc gtatagccta gcatttac
ct ggctttccac
73528-  2277961 actgctgtgg ataattaca cac
73529-//
73530:LOCUS          sequence2          81630 bp      DNA
BCT 20-JAN-2017
--
76312-  81541 ttttttaaaa cctgagtaa taaagctggt ttatcagggg tgggggct
tt cgaccaattc
76313-  81601 atccaatagt agtaaacagt ggaccacttt
76314-//
76315:LOCUS          sequence3          40971 bp      DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] █
```

[12:41午後]

[12:41午後]



W13-4: tailで行末を表示

①tail -n 5でsequence3の最後の5行分を念のため表示。確かに//が区切りとなっている

```
File Edit View Search Terminal Help 13:58
73528- 2277961 actgctgtgg ataattaca cac
73529-//
73530:LOCUS          sequence2          81630 bp      DNA
      BCT 20-JAN-2017
      --
76312- 81541 ttttttaaaa cctgagtaa taaagctggt ttatcagggg tgggggct
      tt cgaccaattc
76313- 81601 atccaatagt agtaaacagt ggaccacttt
76314-//
76315:LOCUS          sequence3          40971 bp      DNA
      BCT 20-JAN-2017
iu@bielinux[mac_share] tail -n 5 annotation.gbk          [ 1:58午後 ]
      40741 tggaatcagg tgtttgtaa tgttccagat agcgctccag aattaaagaa aat
      ggatgga
      40801 gaaatgggc ttagttctat aaaagtcttg aattatctta aacagctatt gaa
      tggaata
      40861 tcgcaatatc aatcggcaga tattaaaagg agtcagcaag cacttagact gga
      attatca
      40921 agctacgact ggggattcga tatagtccct ggatttagaa ctgttgatga t
      //
iu@bielinux[mac_share] █          [ 1:58午後 ]
```



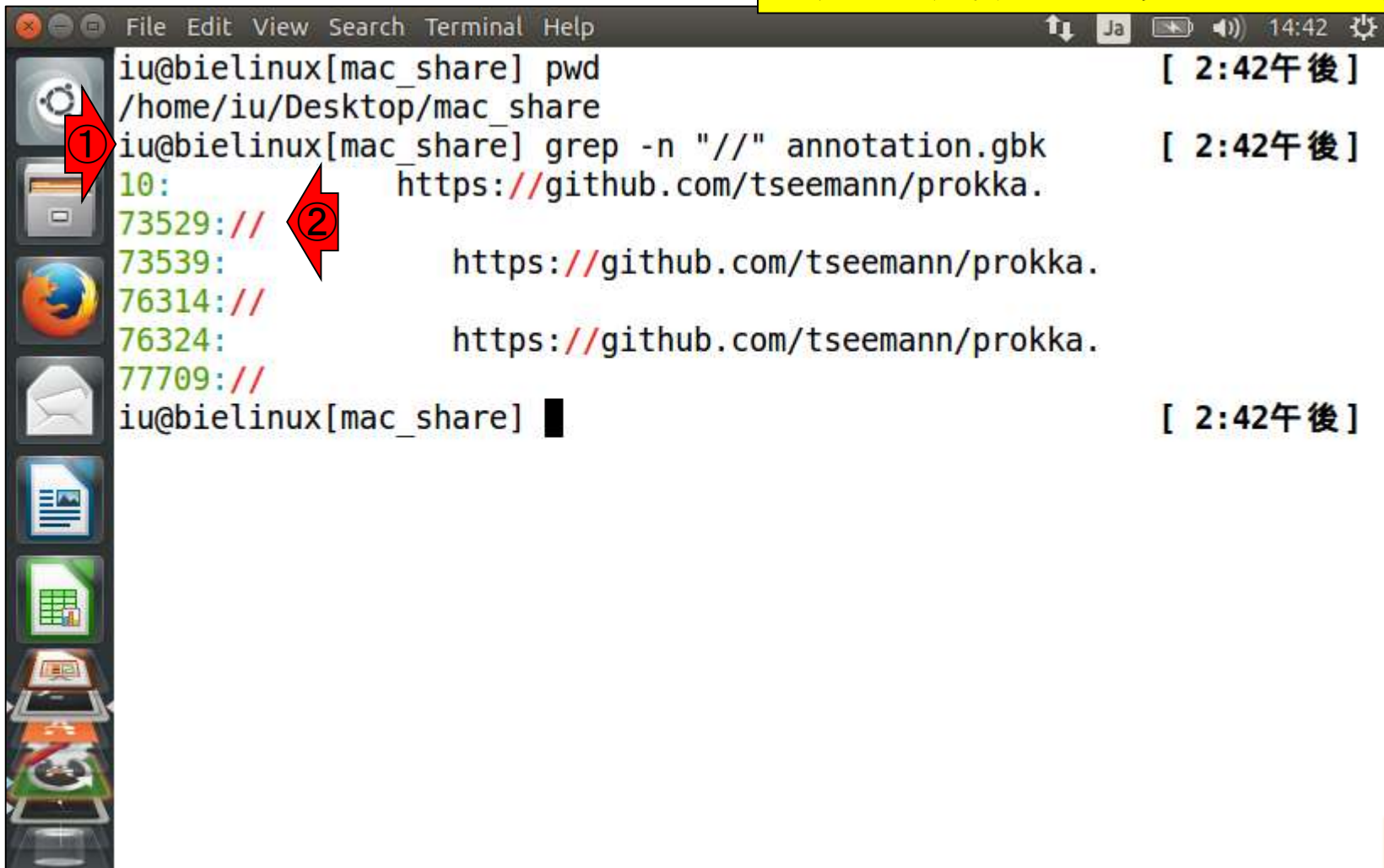
W14-1:awk

annotation.gbkファイル中の最初の配列であるsequence1のみをawkコマンドを用いて抽出するやり方を伝授。主目的はawkの存在を知ってもらうこと。ここでは、各配列の最終行が//であることを利用して//行になる(手前)までの部分を別ファイルに保存するやり方を示す

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "//" annotation.gbk [ 2:42午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] [ 2:42午後 ]
```

W14-1:awk

①grepでannotation.gbk中の//を含む行を、行番号つきで表示。ここでは、行全体が//のみからなるかどうかを判定して、そうでない場合のみ出力するというを行う



```
iu@bielinux[mac_share] pwd [ 2:42午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "//" annotation.gbk [ 2:42午後 ]
10: https://github.com/tseemann/prokka.
73529:// https://github.com/tseemann/prokka.
73539: https://github.com/tseemann/prokka.
76314:// https://github.com/tseemann/prokka.
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] [ 2:42午後 ]
```

W14-1:awk

もう少し具体的にいうと、①が最初に条件を満たす行。awkを利用してやろうとしていることは、条件を満たさない行は、ず〜っと画面上に表示していき、条件を満たしたら画面表示をやめて抜ける。そのため、②や③自体は条件を満たすが、①になった段階で作業を終了しているので無関係。そういう簡単なawkコマンドを伝授しようとしているのです

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "/
10: https://github.com/tseemann/prokka.
73529:// ①
73539: https://github.com/tseemann/prokka.
76314:// ②
76324: https://github.com/tseemann/prokka.
77709:// ③
iu@bielinux[mac_share] █
```

[2:42午後]

W14-1 : awk

①awkコマンドの実行部分。赤下線部分のannotation.gbkが入カファイル

```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "/" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
① iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' annotation.gbk > annotation_seq1.gbk [ 3:18午後 ]
iu@bielinux[mac_share] █
```

W14-1 : awk

②がリダイレクト(>)で実行結果をannotation_seq1.gbkというファイルで保存するという指令部分

```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "/" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
① iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' annotation.gbk > annotation_seq1.gbk [ 3:18午後 ]
iu@bielinux[mac_share] █ ②
```


W14-1 : awk

```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "/" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
1 iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' annotation.gbk > annotation_seq1.gbk [ 3:18午後 ]
iu@bielinux[mac_share] █
```

W14-2: 解説

awkで入力ファイルを1行1行読み込んで①\$0に格納して行ごとに処理を行うのが基本。まず最初の1行分のみを\$0に格納し、\$0が②//と同じかどうかを判定しているのが、赤下線部分

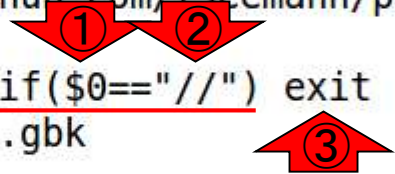
```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "//" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' ann
otation.gbk > annotation_seq1.gbk [ 3:18午後 ]
iu@bielinux[mac_share] █
```



もし①\$0が②//と同じであれば(つまり条件を満たせば)、③作業終了

W14-2: 解説

```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "//" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' annotation.gbk > annotation_seq1.gbk [ 3:18午後 ]
iu@bielinux[mac_share] █
```



W14-2: 解説

④条件を満たさなければ(それがif elseのelseに相当する部分)、⑤\$0の中身を⑥printせよ(つまり出力せよ)、です

```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "/" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' annotation.gbk > annotation_seq1.gbk
iu@bielinux[mac_share] [ 3:18午後 ]
```

W14-2: 解説

awkコマンドで①~⑥の一連の条件判定や操作を行ごとに行っています。だから、awkの処理対象である⑦annotation.gbk中の、⑧73529行目で「\$0=="//"」という条件を満たすまでは…

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "//" annotation.gbk
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' annotation.gbk > annotation_seq1.gbk
iu@bielinux[mac_share]
```

④が成立するので、延々と⑤\$0の中身(つまり行の中身)が⑥printされるのです

W14-2: 解説

```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "/" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529:// ← ⑧
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' ann
otation.gbk > annotation_seq1.gbk
iu@bielinux[mac_share] [ 3:18午後 ]
```

④ ↓
⑤ ↓
⑥ ↑

W14-3: 行数確認

①wcで行数を確認。⑨annotation_seq1.gbkの行数は73528行。⑧//からなる73529行目の1つ手前の行までからなるのだろう、つまりうまくいっているのだろうということが⑨73528という数値からも予想できる

```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "//" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529:// ←⑧ https://github.com/tseemann/prokka.
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' annotation.gbk > annotation_seq1.gbk
iu@bielinux[mac_share] wc annotation* [ 3:18午後 ]
 77709  338721  5241683 annotation.gbk
   2492   38582   511967 annotation.gff
 73528  320936  4966072 annotation_seq1.gbk ←⑨
153729  698239 10719722 total
iu@bielinux[mac_share] █ [ 3:53午後 ]
```


W14-4:awk

一見すごくややこしいが、①で指定した文字列と合致する行(の手前)までを出力したい場合に、②全体のスクリプトをテンプレートとして利用できる。コピーして必要な部分のみを変更して利用すればよい、ということ

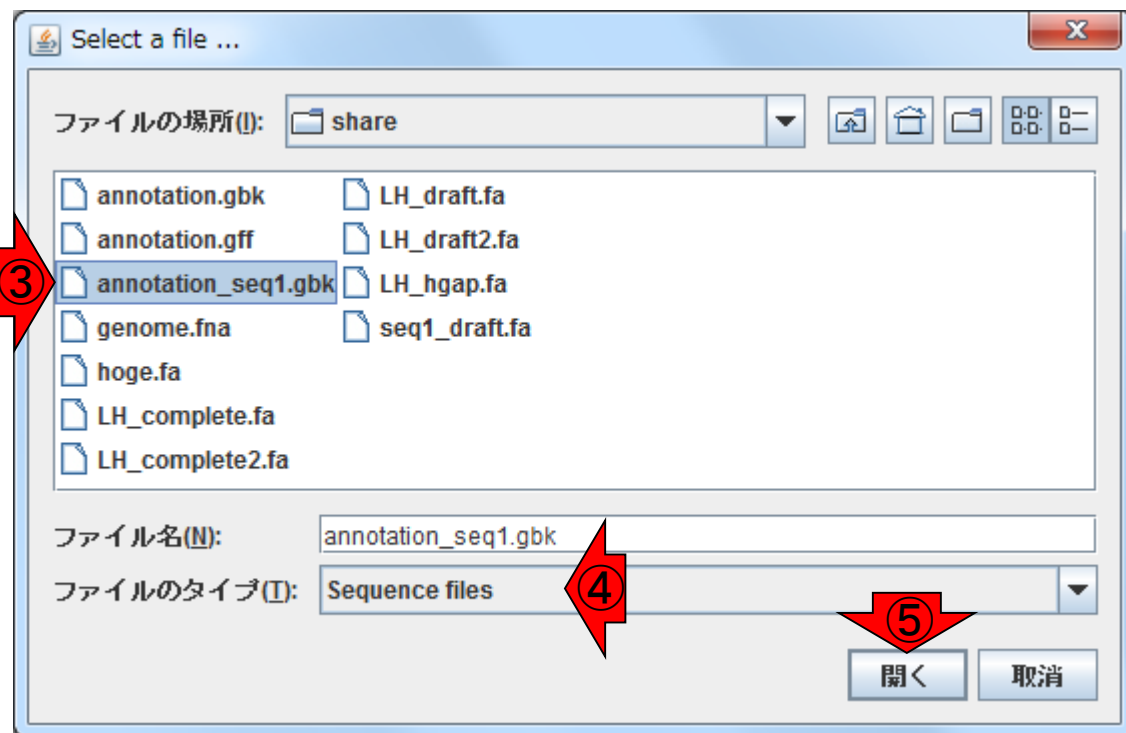
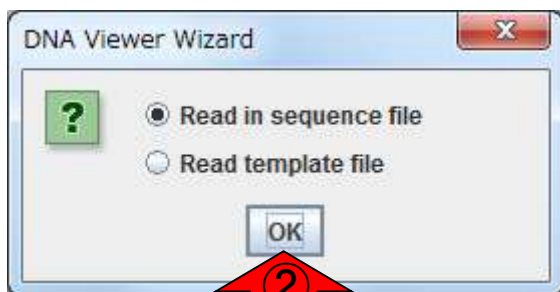
```
iu@bielinux[mac_share] pwd [ 3:18午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "/" annotation.gbk [ 3:18午後 ]
10: https://github.com/tseemann/prokka.
73529://
73539: https://github.com/tseemann/prokka.
76314://
76324: https://github.com/tseemann/prokka.
77709://
② iu@bielinux[mac_share] awk '{if($0=="//") exit ; else print $0}' ann [ 3:18午後 ]
otation.gbk > annotation_seq1.gbk
iu@bielinux[mac_share] wc annotation* [ 3:18午後 ]
 77709  338721  5241683 annotation.gbk
  2492   38582   511967 annotation.gff
 73528  320936  4966072 annotation_seq1.gbk
153729  698239 10719722 total
iu@bielinux[mac_share] [ 3:53午後 ]
```

W15-1 : DNAPlotter

①②DNAPlotterを起動して、作成した③ annotation_seq1.gbkを、④ファイルのタイプはSequence filesのままよいので、⑤開く

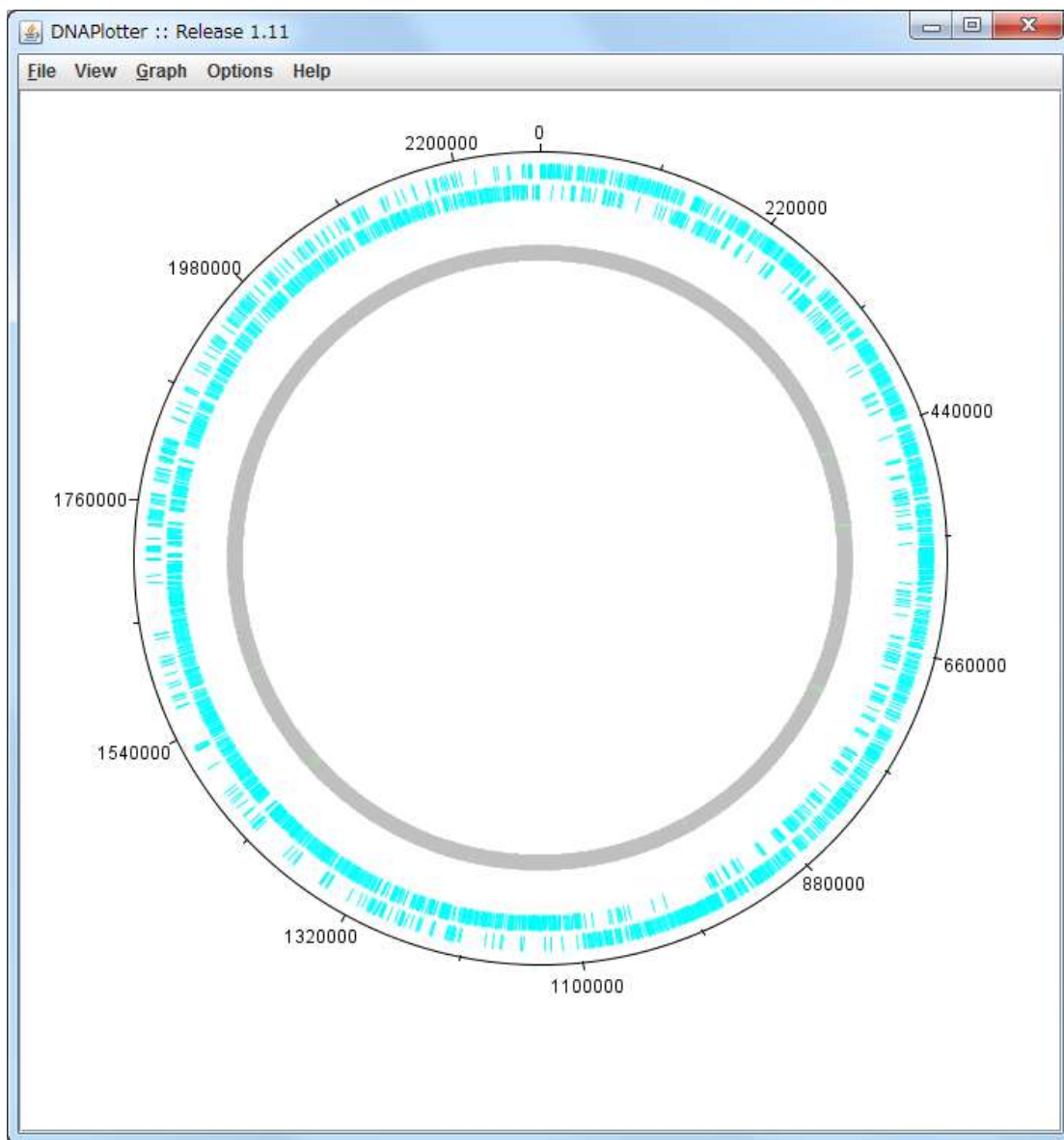


dnaplotter.jar



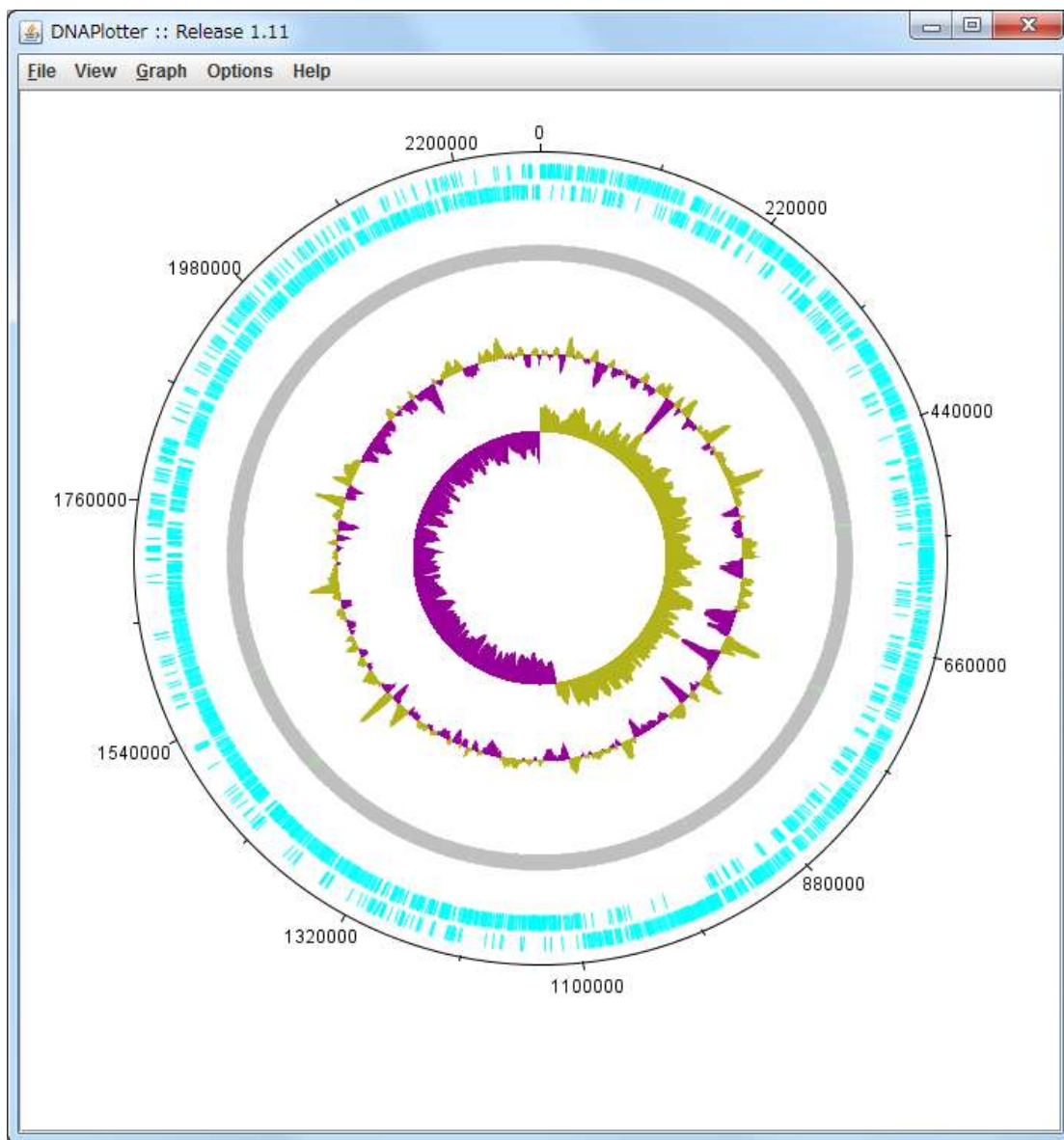
W15-1 : DNAPlotter

問題なく開きます。//という行がなくても問題ないという証拠を示したまでです



W15-1 : DNAPlotter

GC plotやGC skewを表示させることができ、その結果はannotation.gbkを読み込ませて作成したもの(W12-20)と同じ



W15-2:tail

①annotation_seq1.gbkの最終行3行分、および
②“^LOCUS sequence2”という条件を満たす行の上側4行分を表示。確かにawkで意図通りの部分まで抽出できていることがわかる

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] tail -n 3 annotation_seq1.gbk
2277841 ttatccacag ctatgcaaaa aatagtggat aactcctcca agccttgata caa
cggacgt
2277901 atctttgcac aaaaaaatgt ggataacttc gtatagccta gcatttacct ggc
tttccac
2277961 actgctgtgg ataattaca cac
iu@bielinux[mac_share] grep -B 4 "^LOCUS           sequence2" annotation
.gbk
2277841 ttatccacag ctatgcaaaa aatagtggat aactcctcca agccttgata caa
cggacgt
2277901 atctttgcac aaaaaaatgt ggataacttc gtatagccta gcatttacct ggc
tttccac
2277961 actgctgtgg ataattaca cac
//
LOCUS           sequence2                           81630 bp       DNA                           BCT
20-JAN-2017
iu@bielinux[mac_share]
```

W16-1 : sequence1

W13-2でheadとtailコマンドを組み合わせることで目的の配列のみの情報を抽出することが可能(第3回W19-3)と書いたが、そのやり方を示す。まずはW13-2と同じく、annotation.gbkに対して①wc、および②grep -n “^LOCUS”を実行

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
① iu@bielinux[mac_share] wc annotation.gbk [ 5:10午後 ]
77709 338721 5241683 annotation.gbk
② iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 5:10午後 ]
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
73530:LOCUS      sequence2          81630 bp      DNA
BCT 20-JAN-2017
76315:LOCUS      sequence3          40971 bp      DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] █ [ 5:10午後 ]
```

W16-1 : sequence1

sequence1についてはheadコマンドのみで抽出可能。W14で作成したannotation_seq1.gbkとのファイルの同一性を後で確認すべく、①最初の73528行分のみをannotation.gbkから抽出した結果をhoge_seq1.gbkとして保存。②diffで同一性を確認。何も表示されないなので同一である

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation.gbk
 77709  338721 5241683 annotation.gbk
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 5:10午後 ]
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
73530:LOCUS          sequence2          81630 bp      DNA
BCT 20-JAN-2017
76315:LOCUS          sequence3          40971 bp      DNA
BCT 20-JAN-2017
① iu@bielinux[mac_share] head -n 73528 annotation.gbk > hoge_seq1.gbk
iu@bielinux[mac_share] wc hoge_seq1.gbk [ 5:22午後 ]
 73528  320936 4966072 hoge_seq1.gbk
② iu@bielinux[mac_share] diff annotation_seq1.gbk hoge_seq1.gbk
iu@bielinux[mac_share] █ [ 5:22午後 ]
```

W16-2: sequence2

①の結果から、annotation.gbk中で、sequence2は②73530から、③(76315 - 1)=76314行目に相当することがわかる。このあたりの情報を利用

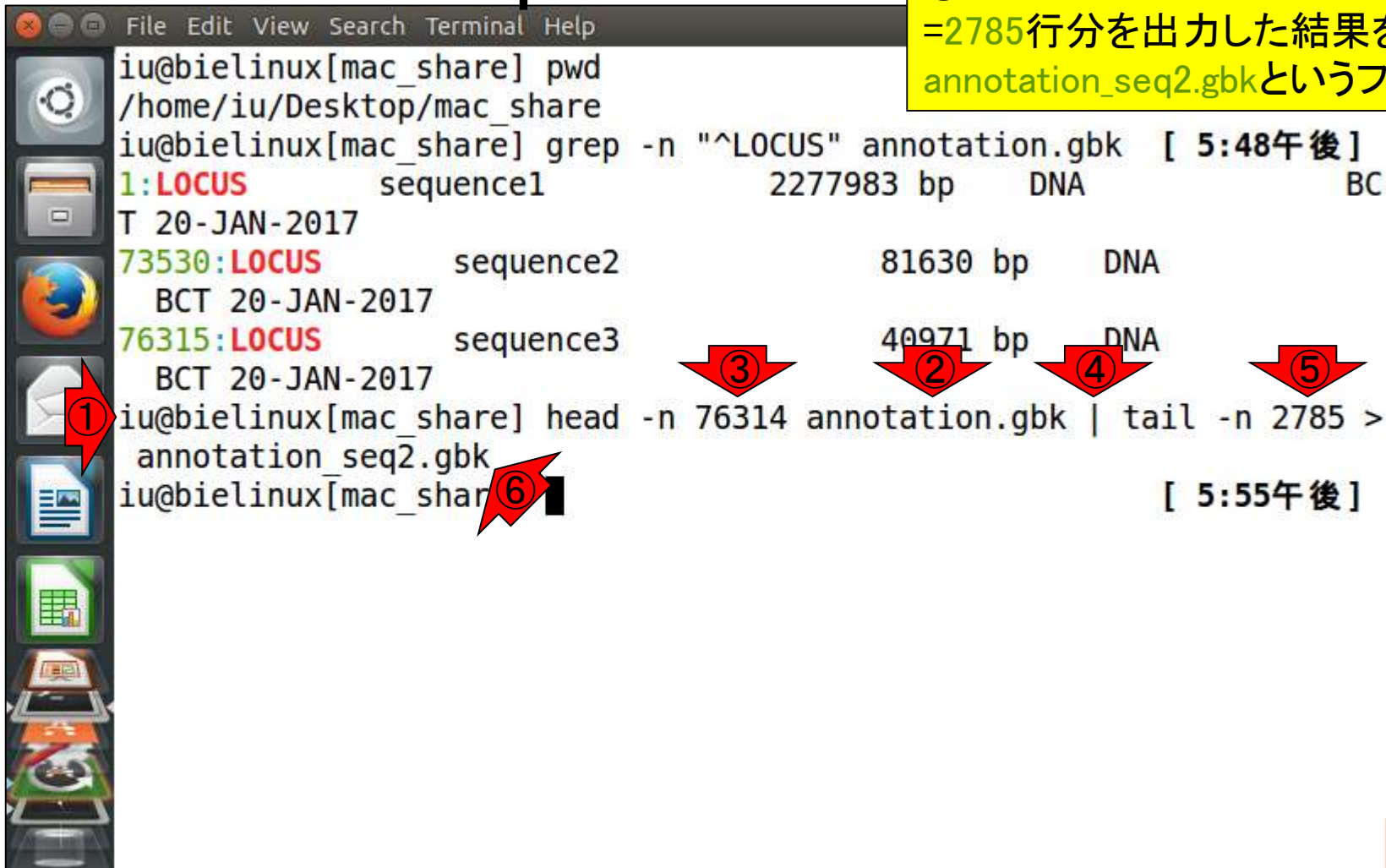
```
iu@bielinux[mac_share] pwd [ 5:48午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 5:48午後 ]
1:LOCUS          sequence1          2277983 bp      DNA            BC
T 20-JAN-2017
2:73530:LOCUS    sequence2          81630 bp      DNA
BCT 20-JAN-2017
3:76315:LOCUS    sequence3          40971 bp      DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] [ 5:48午後 ]
```



W16-2: sequence2

①は、②annotation.gbkからheadコマンドで最初の③76314行を抽出した結果を、④パイプして、⑤tailコマンドで最後の(76314 - 73530 + 1) = 2785行分を出力した結果を、リダイレクトで⑥annotation_seq2.gbkというファイル名で保存

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 5:48午後 ]
1:LOCUS      sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
73530:LOCUS  sequence2          81630 bp      DNA
BCT 20-JAN-2017
76315:LOCUS  sequence3          40971 bp      DNA
BCT 20-JAN-2017
① iu@bielinux[mac_share] head -n 76314 annotation.gbk | tail -n 2785 >
annotation_seq2.gbk [ 5:55午後 ]
⑥
```



W16-2: sequence2

このあたりはただの確認です。得られた `annotation_seq2.gbk` の②最初の2行と③最後の2行を表示。赤枠部分を眺めてうまく抽出できていることを確認しています

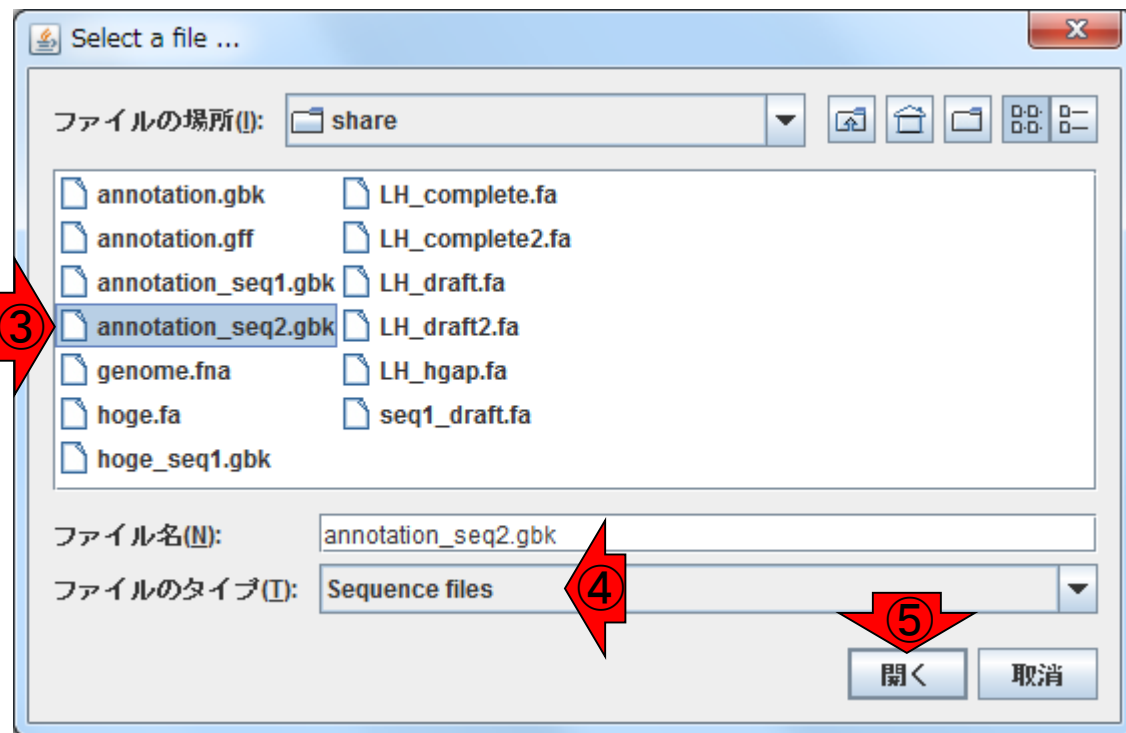
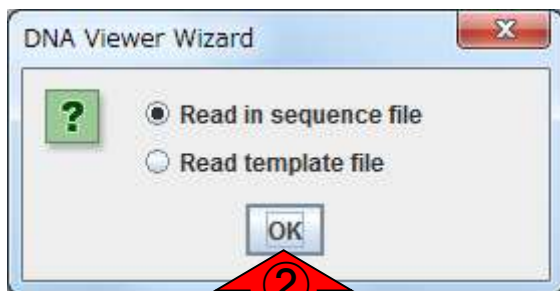
```
iu@bielinux[mac_share] pwd [ 5:48午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 5:48午後 ]
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
73530:LOCUS      sequence2          81630 bp       DNA
BCT 20-JAN-2017
76315:LOCUS      sequence3          40971 bp       DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] head -n 76314 annotation.gbk | tail -n 2785 >
annotation_seq2.gbk
① iu@bielinux[mac_share] wc annotation_seq2.gbk [ 5:55午後 ]
2785 11824 183552 annotation_seq2.gbk
② iu@bielinux[mac_share] head -n 2 annotation_seq2.gbk [ 6:02午後 ]
LOCUS          sequence2          81630 bp       DNA              BCT
20-JAN-2017
DEFINITION    Lactobacillus sp. strain unkown.
③ iu@bielinux[mac_share] tail -n 2 annotation_seq2.gbk [ 6:02午後 ]
81601 atccaatagt agtaaacagt ggaccacttt
//
iu@bielinux[mac_share] [ 6:02午後 ]
```

W16-3: DNAPlotter

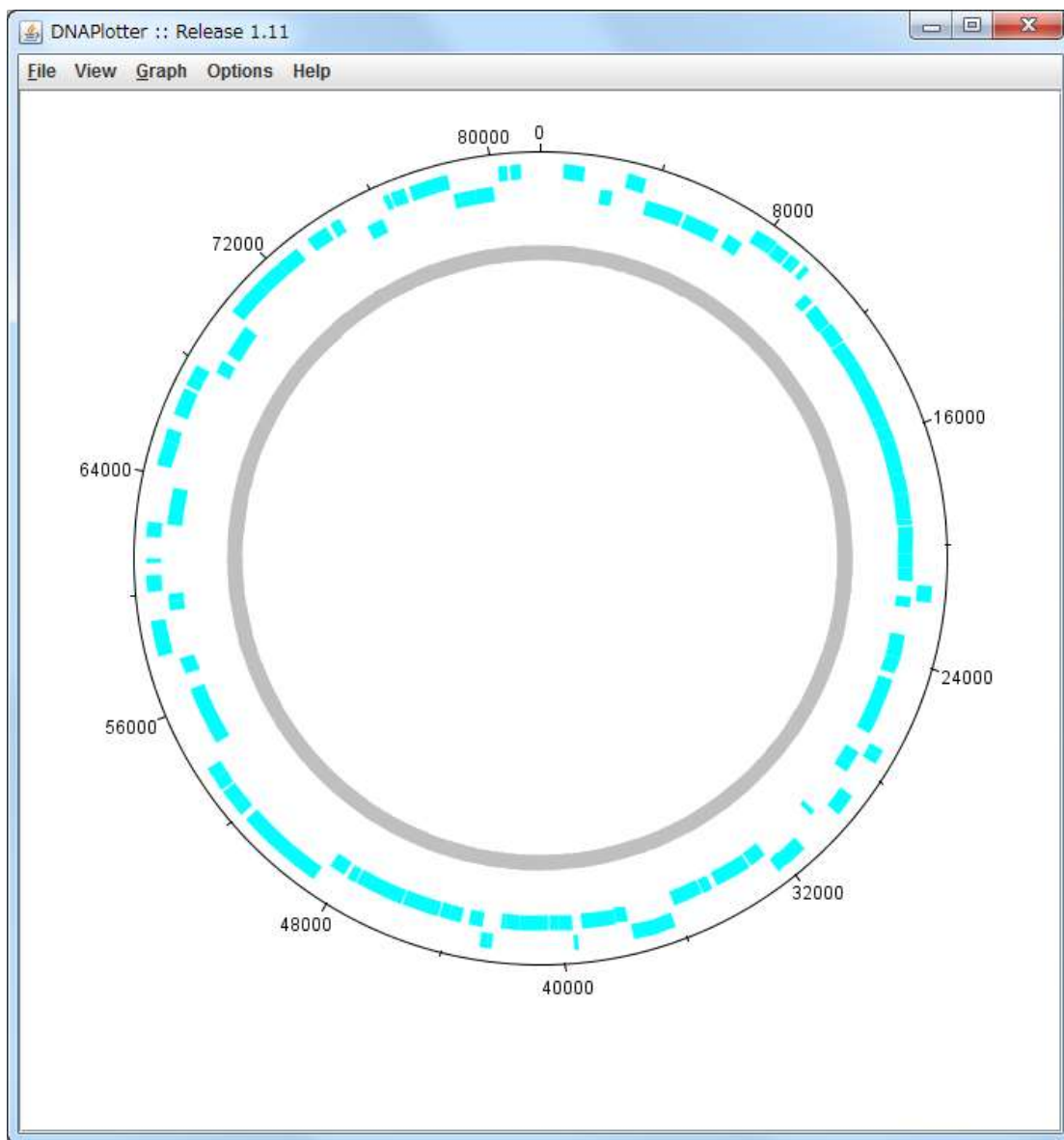
①②DNAPlotterを起動して、作成した③
annotation_seq2.gbkを、④ファイルのタイプ
はSequence filesのままでよいので、⑤開く



dnaplotter.jar

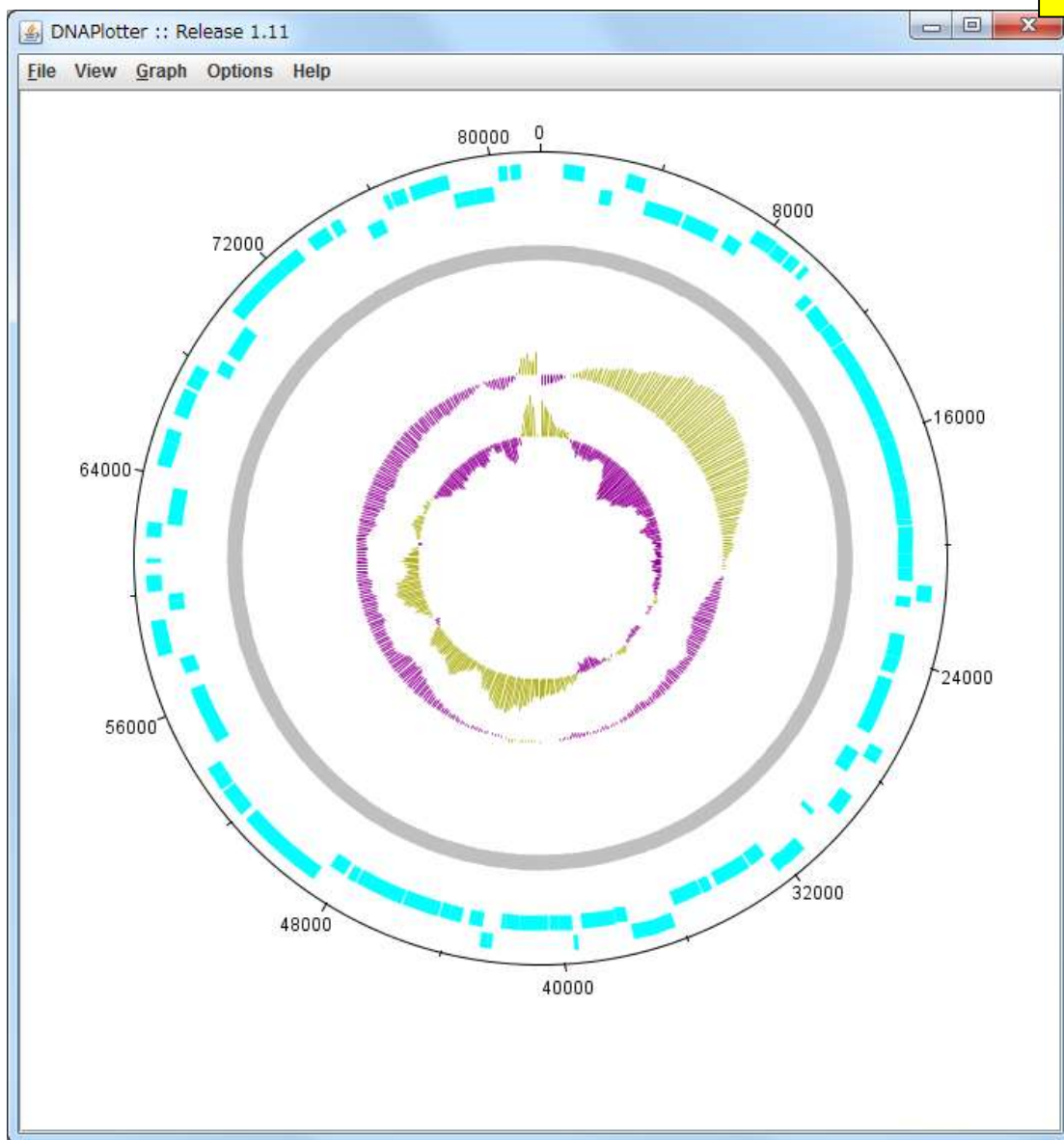


W16-3: DNAPlotter



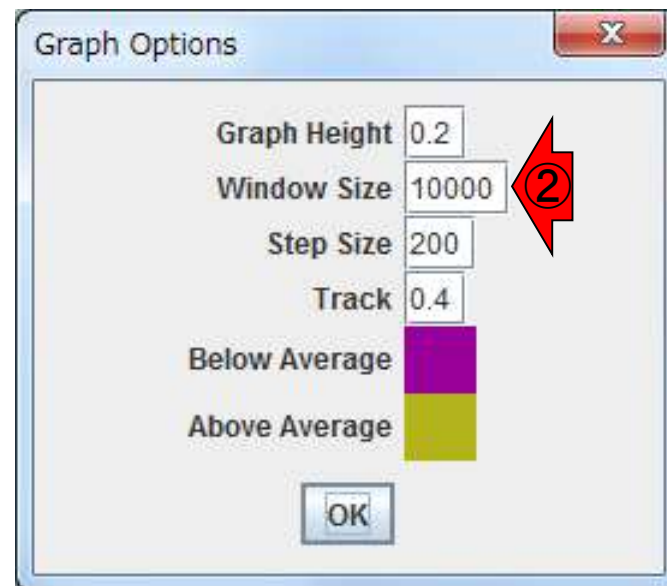
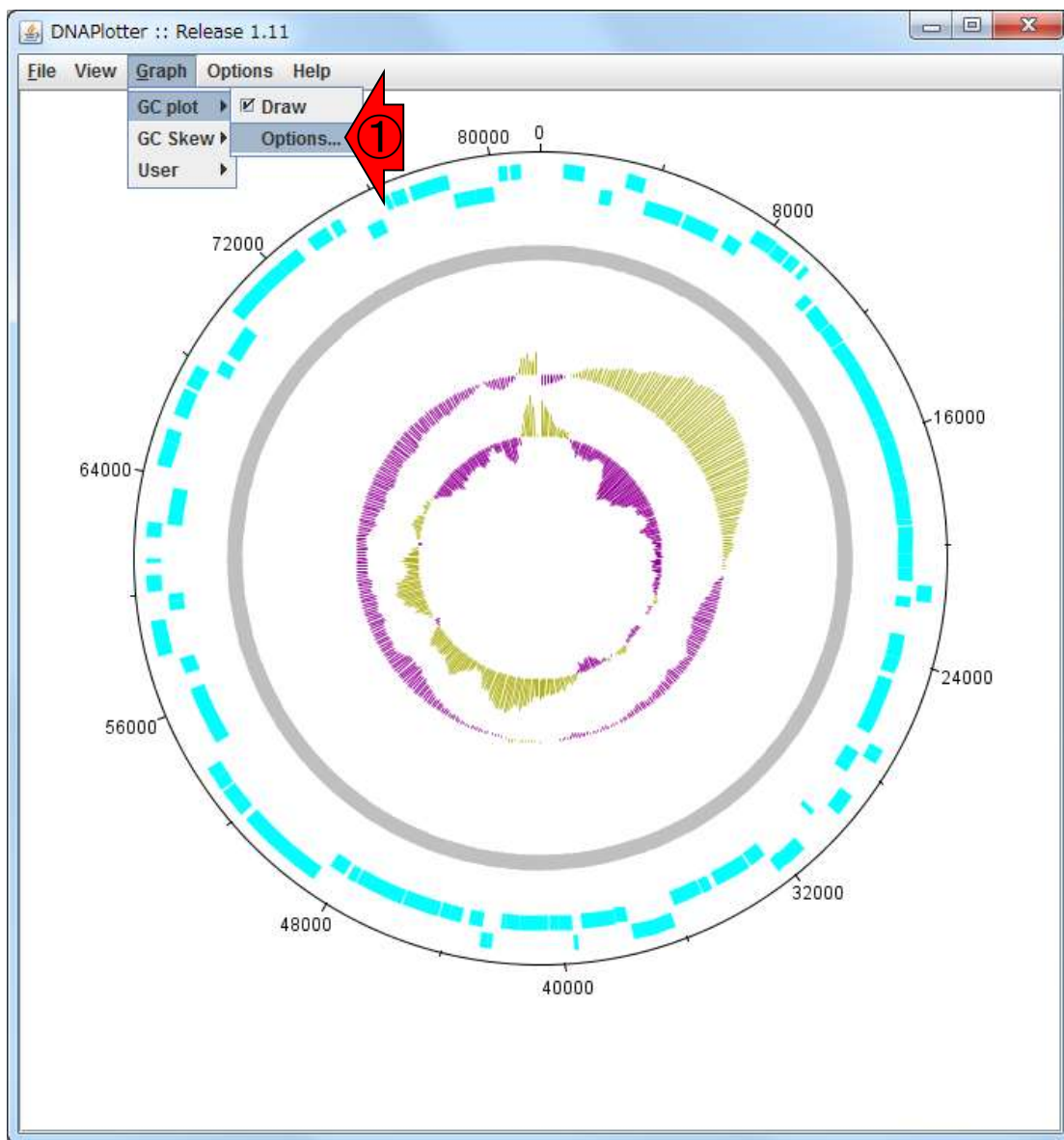
W16-3: DNAPlotter

GC plotとGC skewを表示させた結果。ちゃんと表示できたことに最初は感動したが...、あまりに平滑化しすぎてるんじゃないかという疑念を抱く



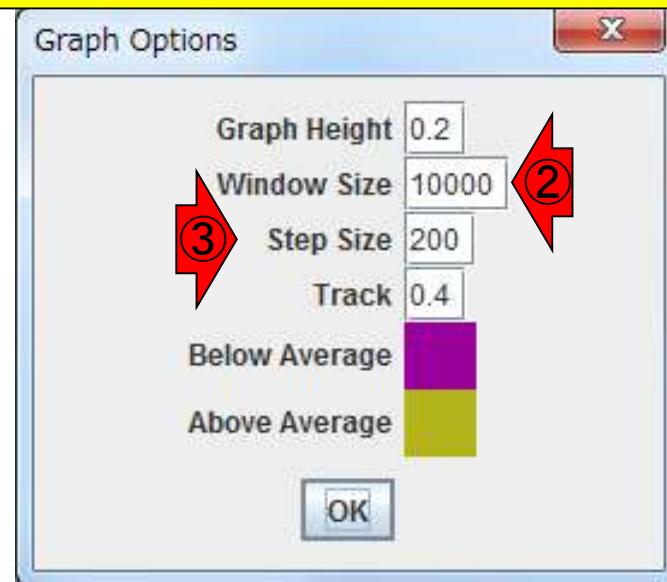
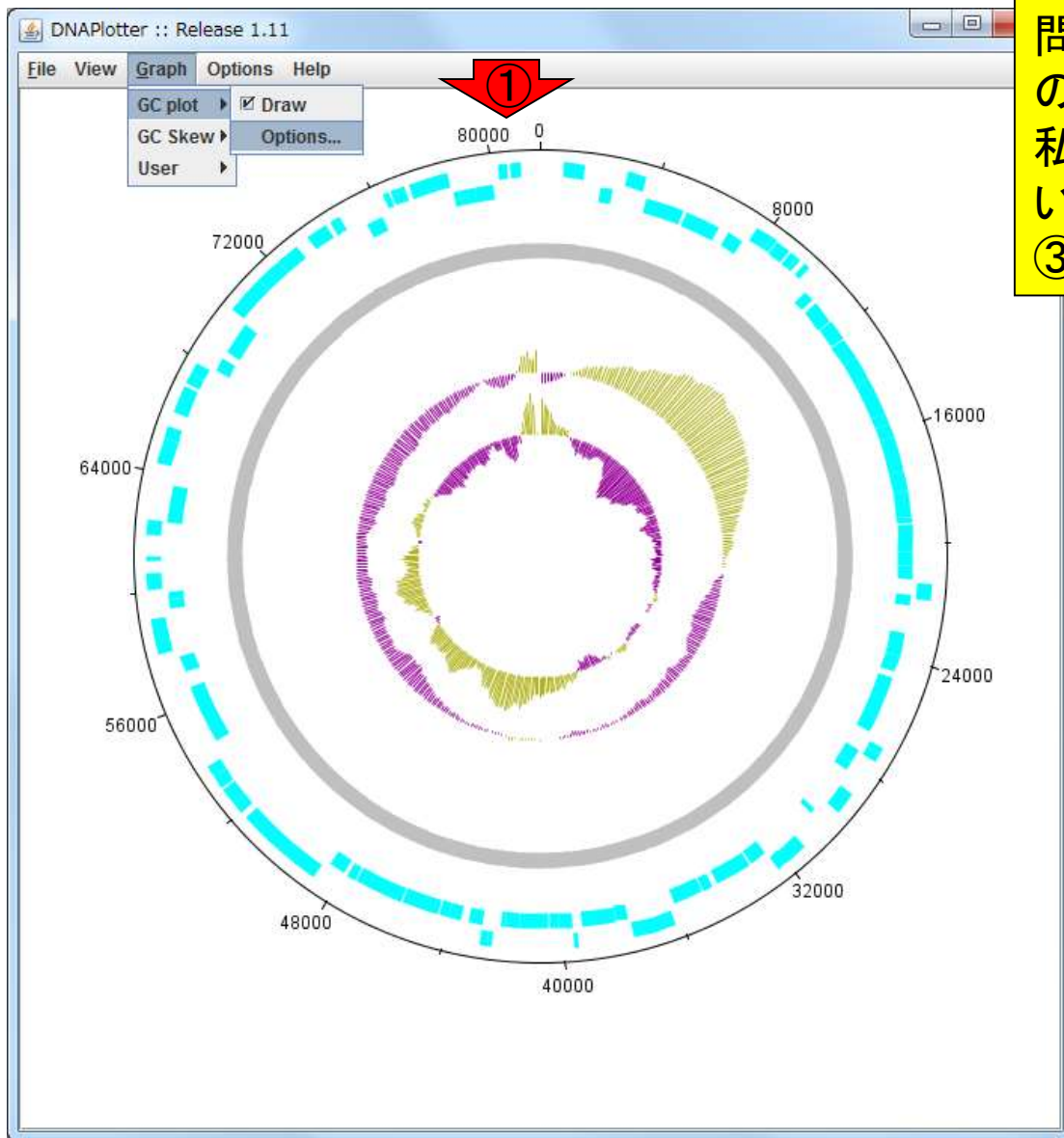
W16-3: DNAPlotter

① Graph - GC plot - Options...。② Window Sizeは、Window(つまり幅)を10000塩基とって、その範囲の平均のGC含量を計算せよという指定だった(W12-15)



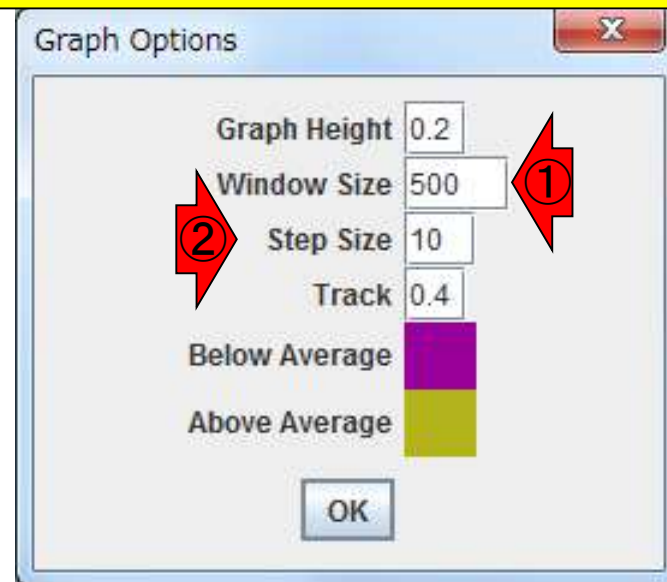
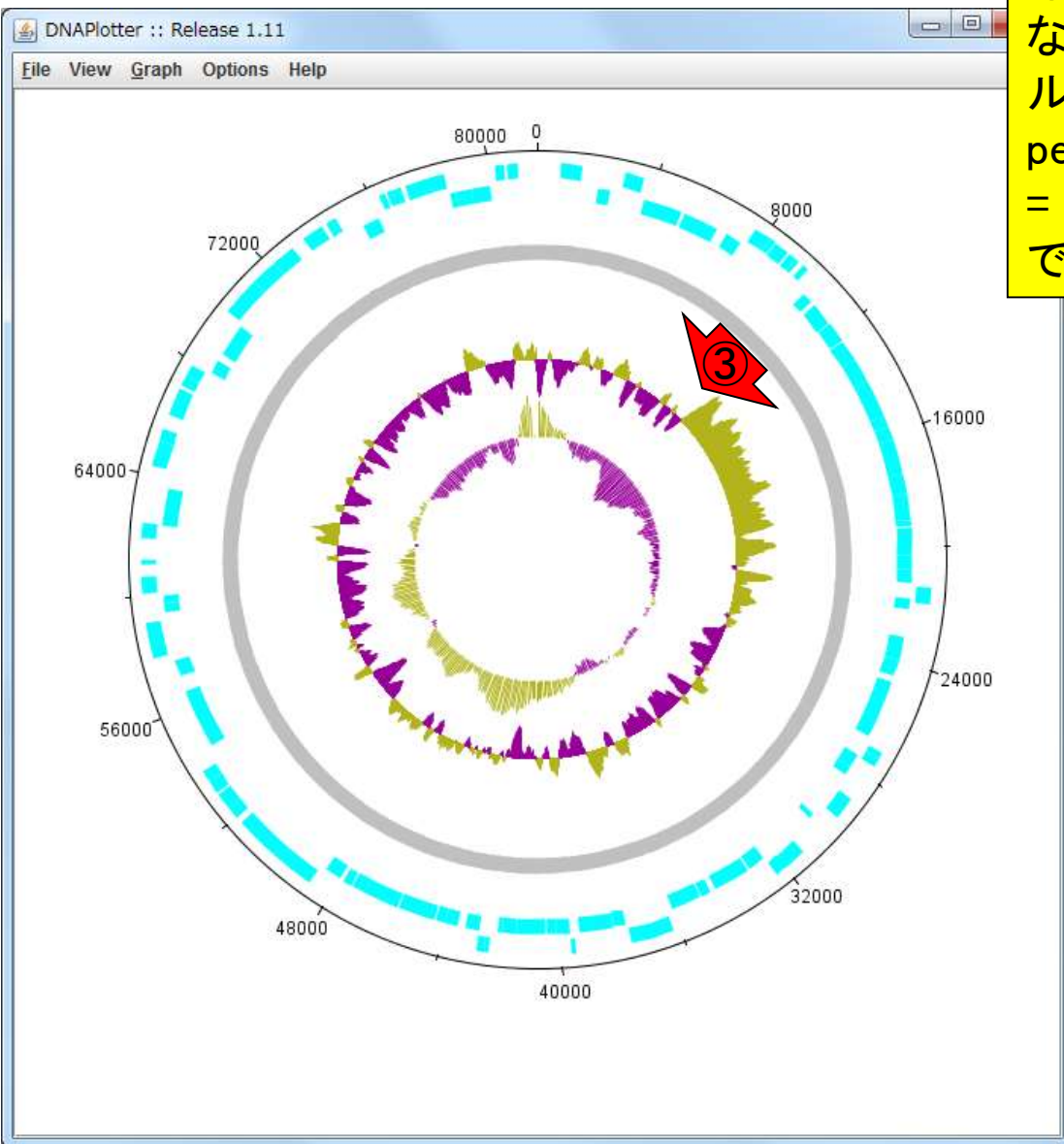
W16-3: DNAPlotter

①しかしsequence2は配列長が81,630 bpの小さい環状プラスミド。2,277,983 bpの染色体ゲノム
のときであれば②Window Size = 10000でも特に問題なかったであろうが、配列長8万ちょっとのものに対しては明らかに大きすぎる値(というのが私の常識的な感覚)。1000とか500とかのほうがいいんじゃないだろうか…。またそれに伴って、
③Step Sizeも20とか10くらいですかねえ…



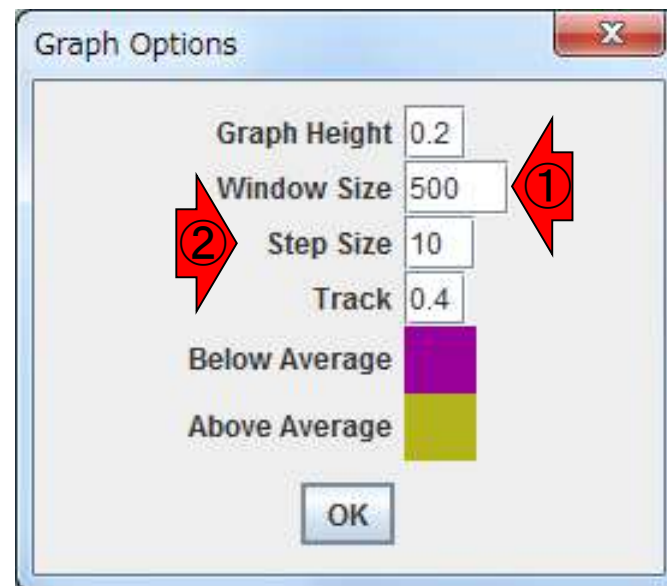
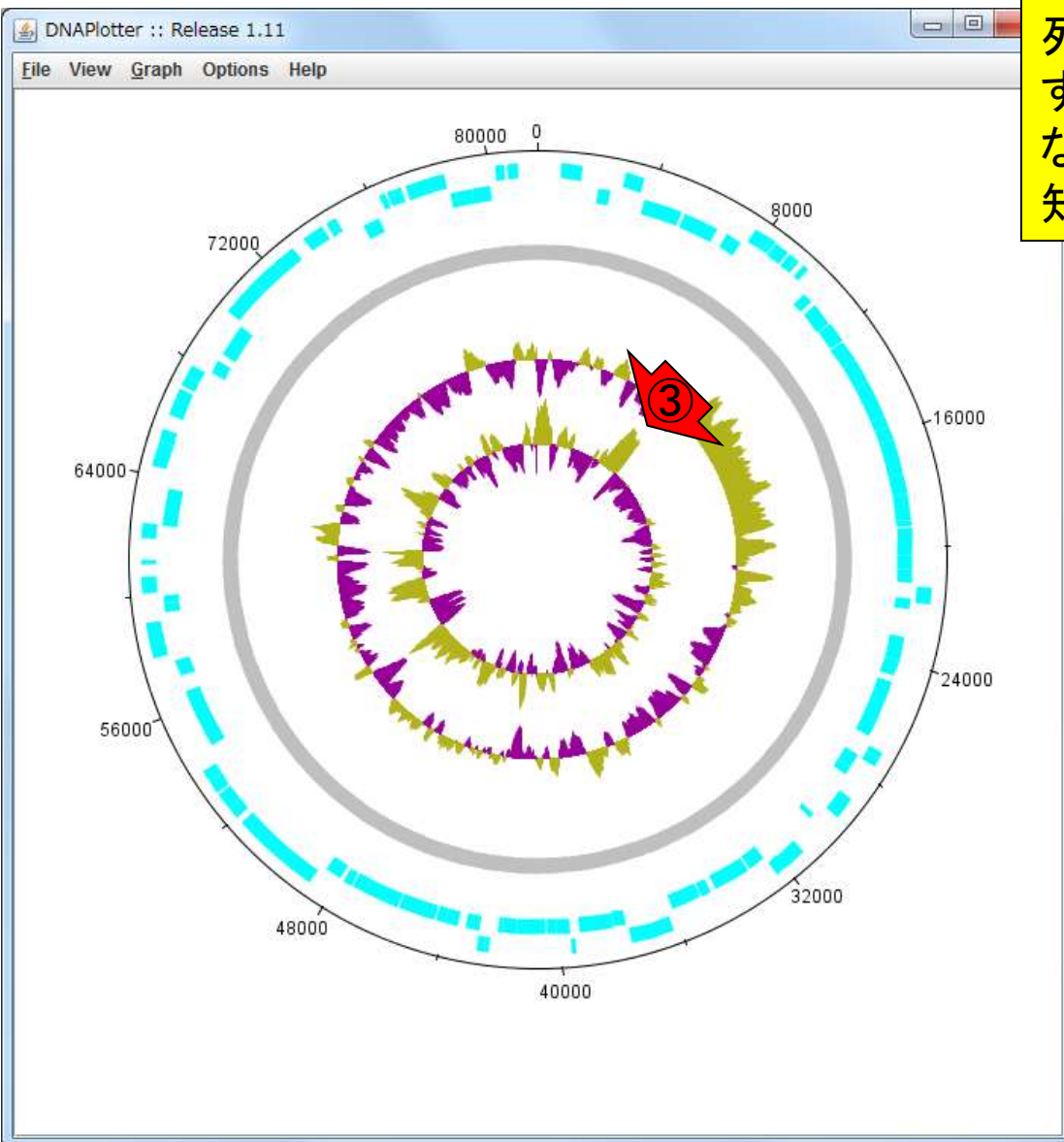
W16-3: DNAPlotter

①Window Size = 500、②Step Size = 10で再描画した結果。特に③のあたりがはっきりとGC含量分布の見栄えとして異なっているのがわかるのではないのでしょうか。投稿論文で示すときは、「DNAPlotter was performed with Window Size = 500 and Step Size = 10 for GC plot」みたいな感じで書いておくべきでしょう(英語はテキストです)。



W16-3: DNAPlotter

GC skewについても、①Window Size = 500、② Step Size = 10で再描画。例えば③のあたりの見え方が変わっているのがわかります。全体の配列長とWindow Size(およびそれに関連して変更するStep Size)の関係についてのガイドライン的なものはどこかにあるのかもしれませんが、私は知りません



W16-4: sequence3

sequence3については、①で得られる②77709という情報と、③で得られる④76315という情報を利用する。基本的にはtailコマンドのみで、 $(77709 - 76315 + 1) = 1395$ 行分を抽出すればよい

```
iu@bielinux[mac_share] pwd [ 8:09午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation.gbk [ 8:09午後 ]
 77709 38721 5241683 annotation.gbk
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 8:09午後 ]
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
73530:LOCUS      sequence2          81630 bp      DNA
BCT 20-JAN-2017
76315:LOCUS      sequence3          40971 bp      DNA
BCT 20-JAN-2017
iu@bielinux[mac_share] █ [ 8:09午後 ]
```



W16-4: sequence3

```
iu@bielinux[mac_share] pwd [ 8:17午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation.gbk [ 8:17午後 ]
 77709  338721 5241683 annotation.gbk
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 8:17午後 ]
1:LOCUS          sequence1          2277983 bp      DNA          BC
T 20-JAN-2017
73530:LOCUS          sequence2          81630 bp      DNA
BCT 20-JAN-2017
76315:LOCUS          sequence3          40971 bp      DNA
BCT 20-JAN-2017
① iu@bielinux[mac_share] tail -n 1395 annotation.gbk > annotation_seq3
.gbk
iu@bielinux[mac_share] █ [ 8:17午後 ]
```

W16-4: sequence3

一応得られたannotation_seq3.gbkに対して、②headと③tailで最初と最後の2行分を実行して確認。LOCUSから始まって、//で終わっているのが大丈夫そうだと判断

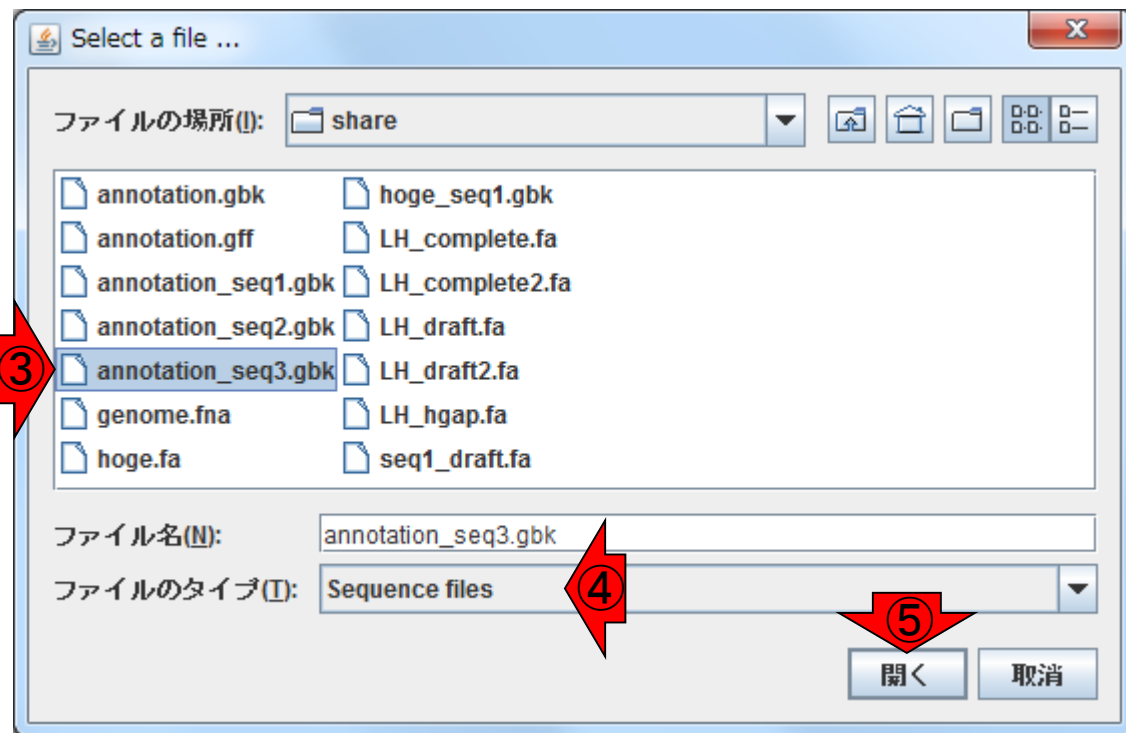
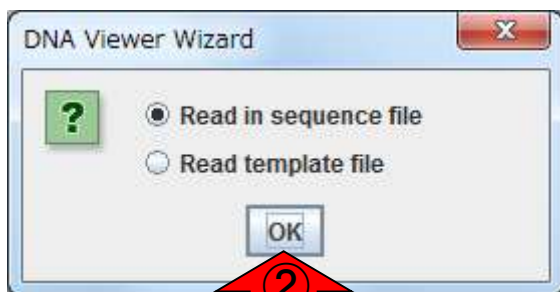
```
iu@bielinux[mac_share] pwd [ 8:17午後 ]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation.gbk [ 8:17午後 ]
 77709  338721 5241683 annotation.gbk
iu@bielinux[mac_share] grep -n "^LOCUS" annotation.gbk [ 8:17午後 ]
1:LOCUS          sequence1          2277983 bp      DNA              BC
T 20-JAN-2017
73530:LOCUS      sequence2          81630 bp      DNA
BCT 20-JAN-2017
76315:LOCUS      sequence3          40971 bp      DNA
BCT 20-JAN-2017
① iu@bielinux[mac_share] tail -n 1395 annotation.gbk > annotation_seq3
.gbk
② iu@bielinux[mac_share] head -n 2 annotation_seq3.gbk [ 8:17午後 ]
LOCUS          sequence3          40971 bp      DNA              BCT
20-JAN-2017
DEFINITION  Lactobacillus sp. strain unkown.
③ iu@bielinux[mac_share] tail -n 2 annotation_seq3.gbk [ 8:18午後 ]
 40921 agctacgact ggggattcga tatagtcct ggatttagaa ctgttgatga t
//
iu@bielinux[mac_share] [ 8:18午後 ]
```

W16-5: DNAPlotter

①②DNAPlotterを起動して、作成した③
annotation_seq3.gbkを、④ファイルのタイプ
はSequence filesのままでよいので、⑤開く

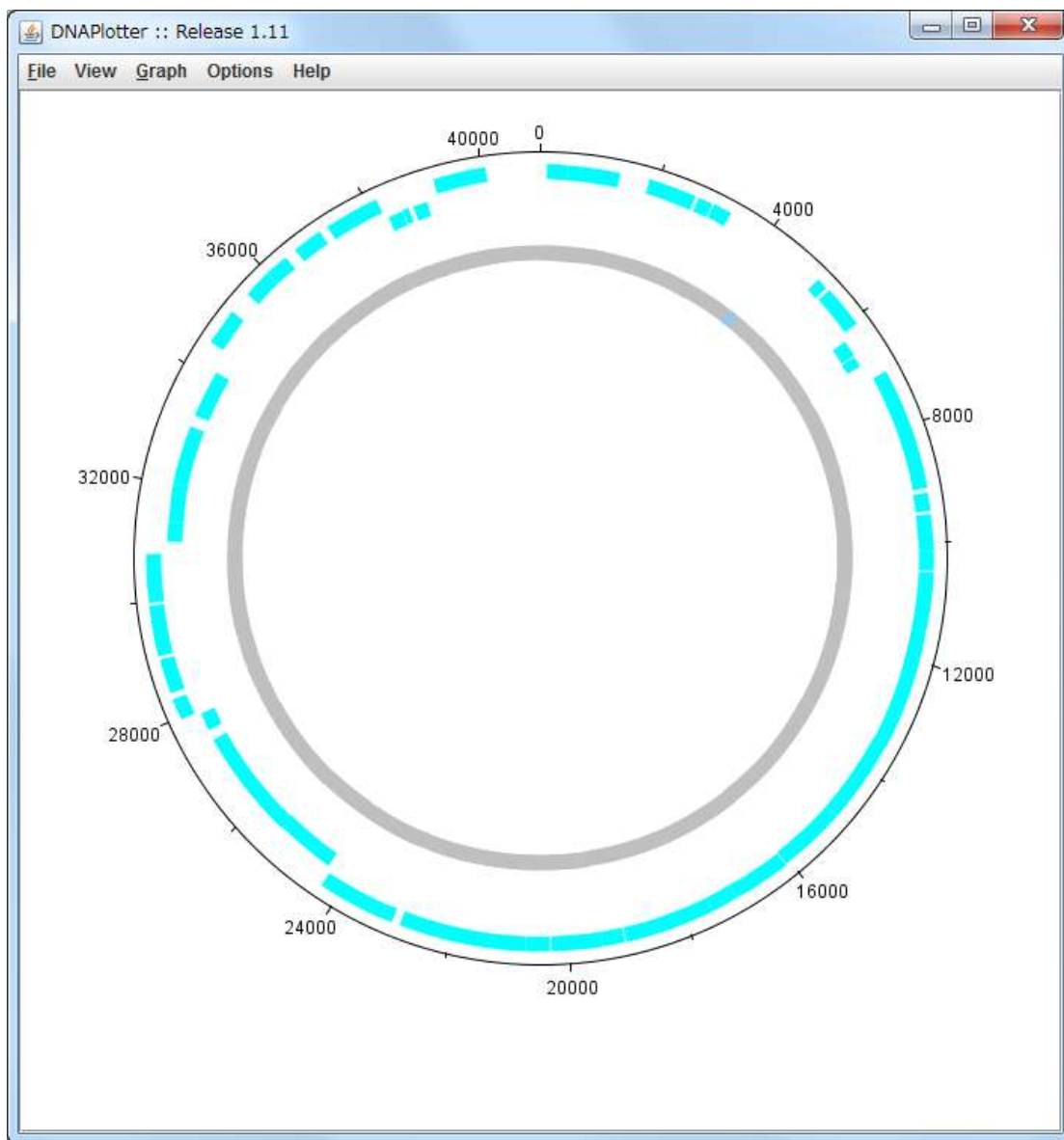


dnaplotter.jar



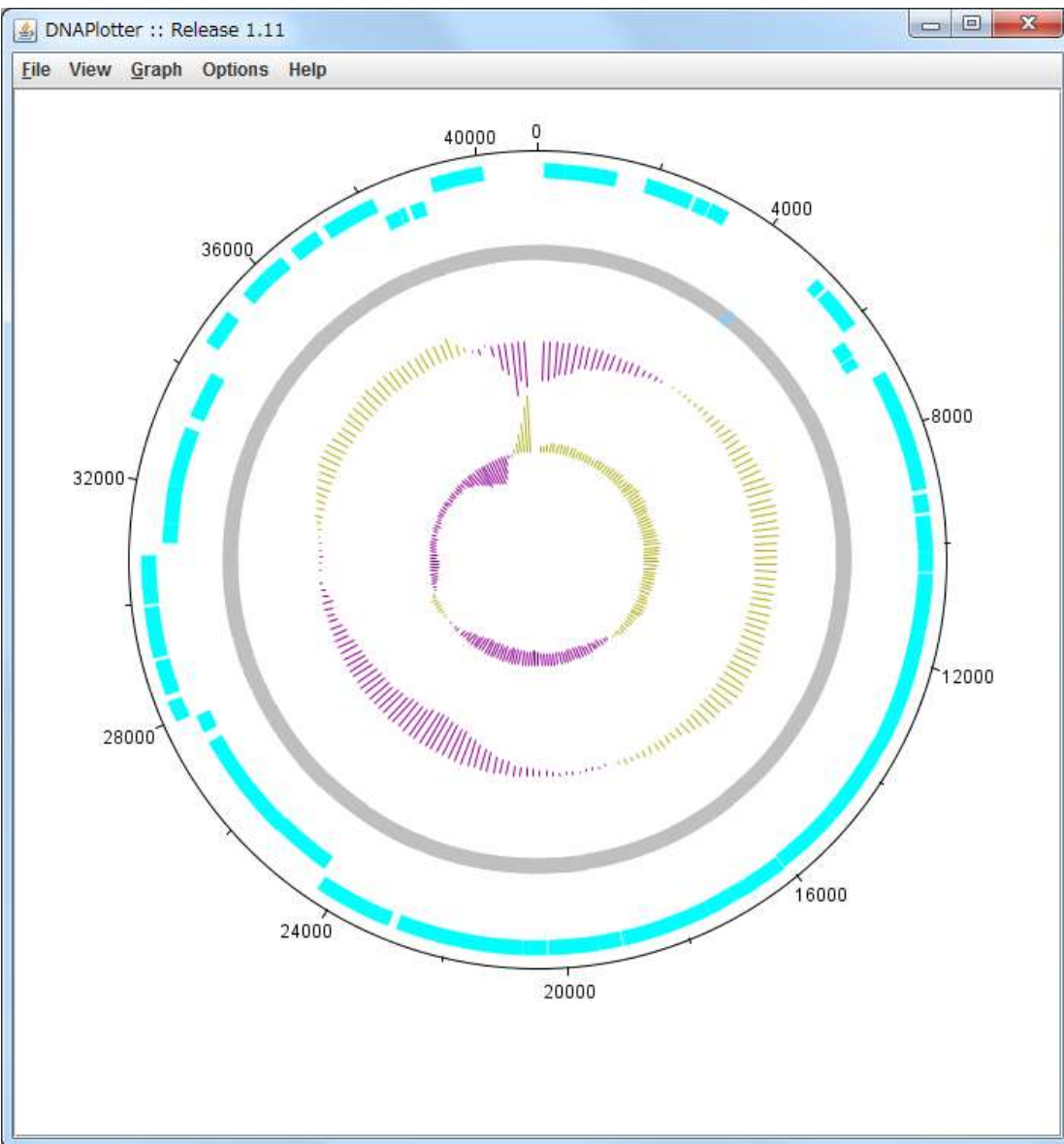
W16-5: DNAPlotter

sequence2のときほどの感動はないw。
①sequence3の配列長(40,971 bp)は、
sequence2 (81,630 bp)の半分程度

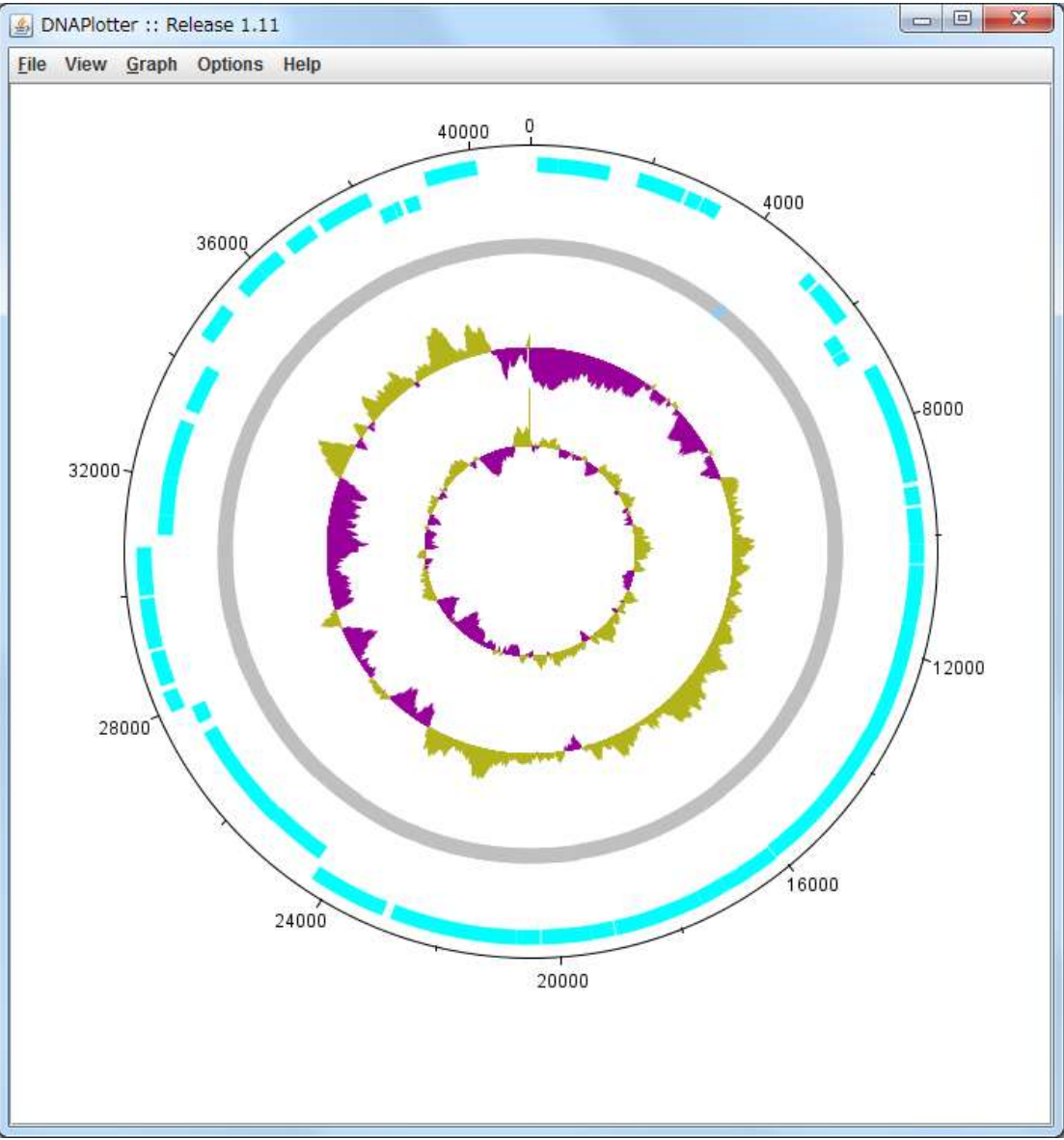


W16-5: DNAPlotter

GC plotとGC skewを表示させた結果。デフォルトのWindow Size = 10000およびStep Size = 200のままなので、sequence2のときと同じくWindow Size = 500、②Step Size = 10にする

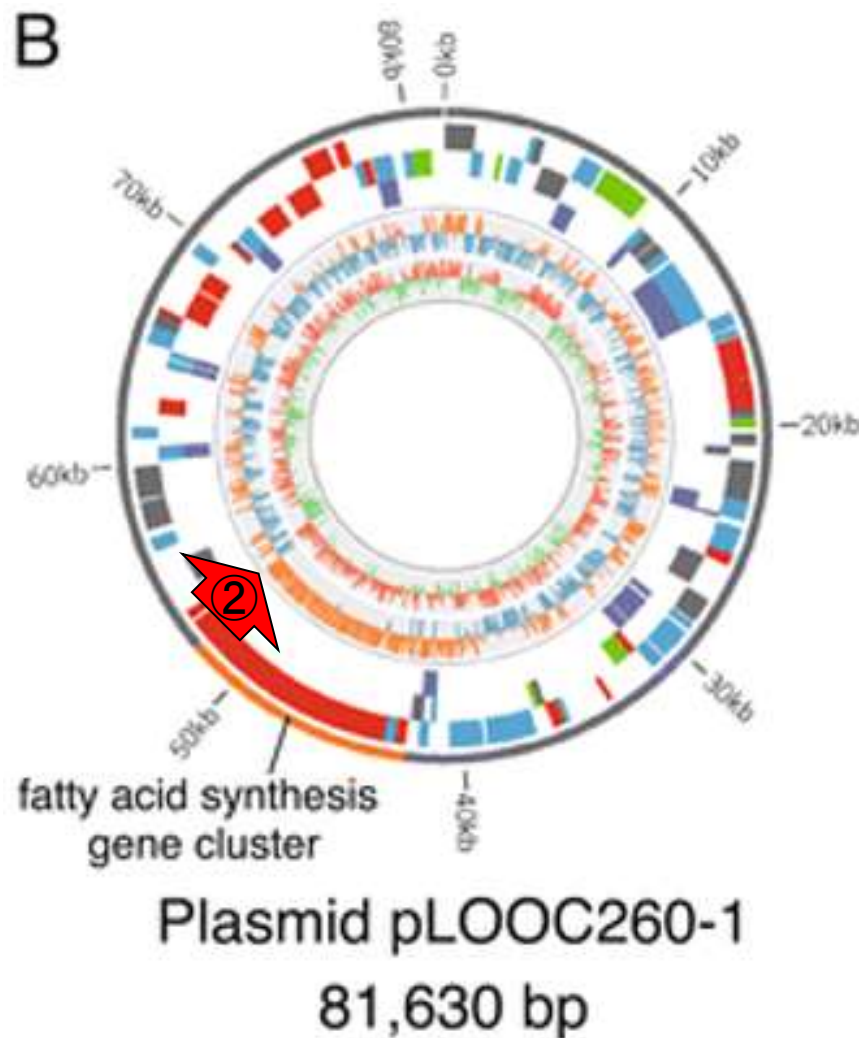
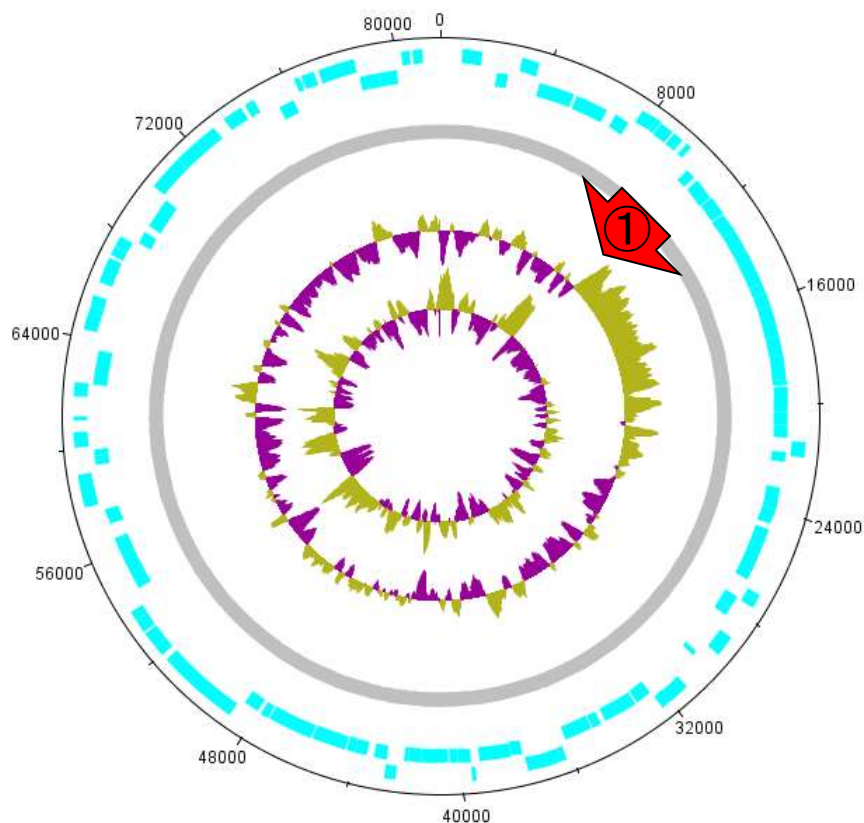


W16-5: DNAPlotter



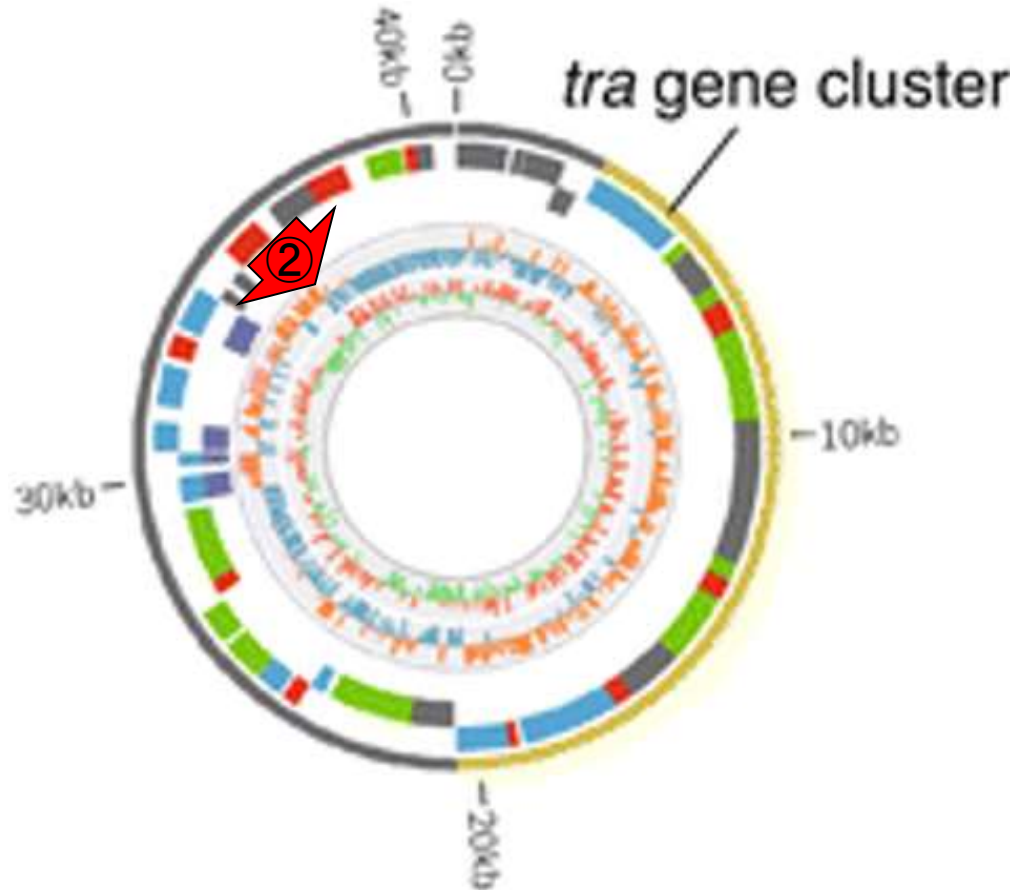
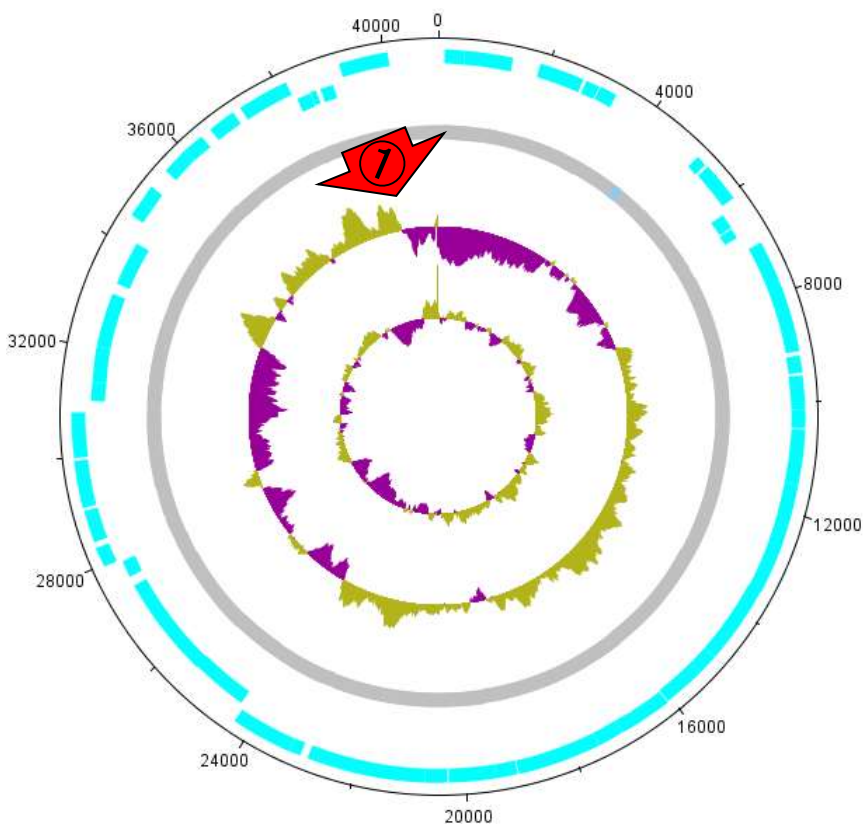
W17-1: 原著論文と比較

①sequence2の結果を、②原著論文
中のFig. 1Bと比較。矢印の位置同士
が対応していると思われる。鎖の向き
が異なるようです。配列長は同じ



W17-2: 原著論文と比較

①sequence3の結果を、②原著論文
中のFig. 1Cと比較。矢印の位置
同士が対応していると思われる。
これは向きは同じ。配列長も同じ



Plasmid pLOOC260-2
40,971 bp

W18-1: 調子が...

ときどきBio-Linuxの調子が悪くなるときがあります。経験上、このような状況になったときはovaファイルを入れ替えるのが一番手っ取り早いです。とりあえず状況をもう少し詳細にみていきます

The image shows the Oracle VM VirtualBox Manager interface on the left and a terminal window for a BioLinux8 virtual machine on the right.

Oracle VM VirtualBox マネージャー

メニュー: ファイル(E) 仮想マシン(M) ヘルプ(H)

ボタン: 新規(N) 設定(S) 破棄 起動(T)

仮想マシン: BioLinux8 (電源オフ)

一般

名前: BioLinux8
オペレーティングシステム: Ubuntu (64)

システム

メインメモリー: 2048 MB
プロセッサ: 2
起動順序: フロッピー, 光学, ハードディスク
アクセラレーション: VT-x/AMD-V, ネイティブ仮想化

ディスプレイ

ビデオメモリー: 12 MB
リモートデスクトップサーバー: 無効
ビデオキャプチャーしたファイル: C:\Users\...
ビデオキャプチャーの属性: フレーム

ストレージ

コントローラー: IDE
IDE セカンダリマスター: [光学ドライブ]
コントローラー: SATA
SATA ポート 0: BioLinux8-

オーディオ

ホストドライバー: Windows DirectSound
コントローラー: ICH AC97

BioLinux8 [実行中] - Oracle VM VirtualBox

メニュー: ファイル 仮想マシン 表示 入力 デバイス ヘルプ

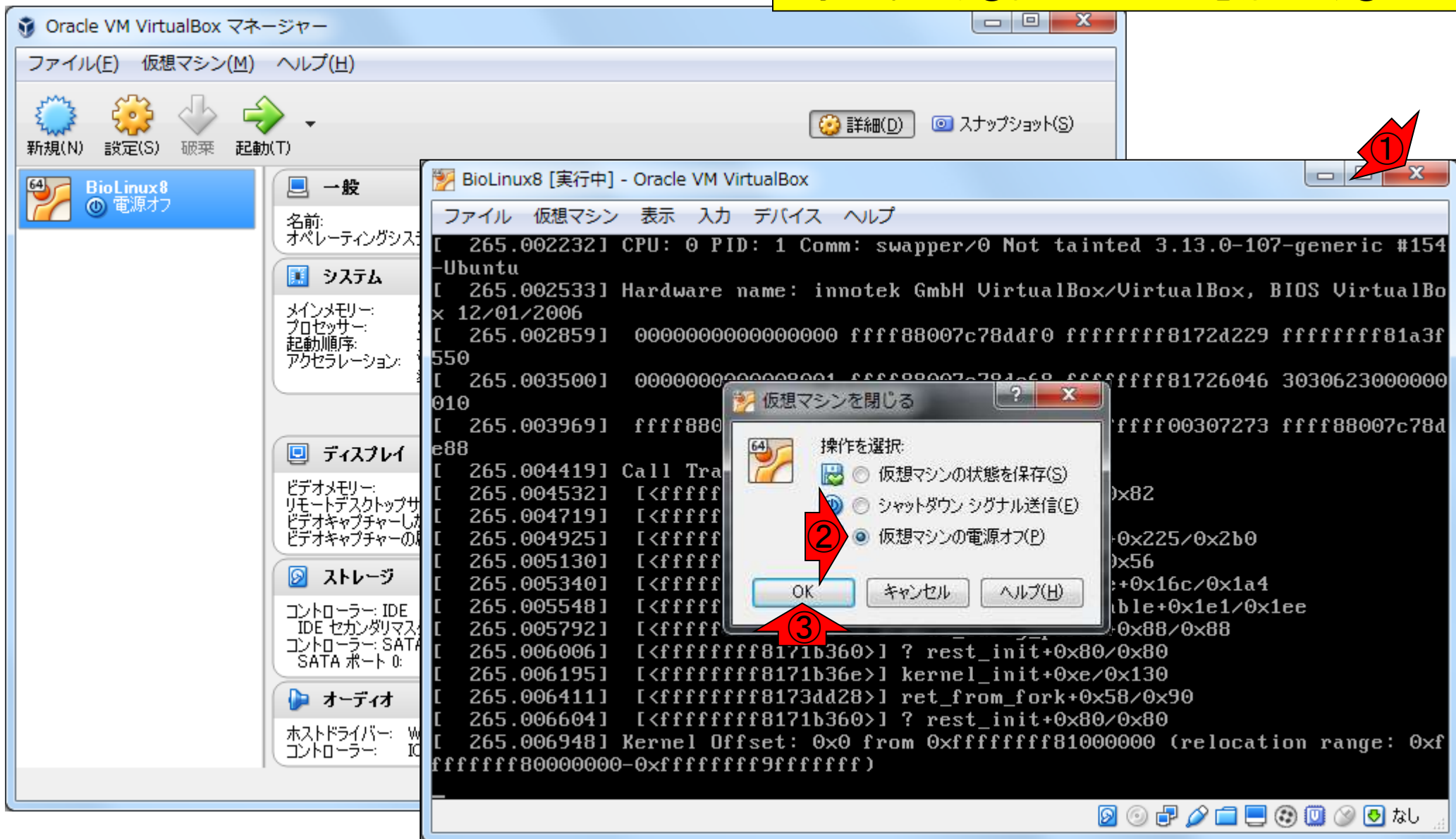
GNU GRUB version 2.02~beta2-9ubuntu1.6

```
*Bio-Linux 8
Advanced options for Bio-Linux 8
Memory test (memtest86+)
Memory test (memtest86+, serial console 115200)
```

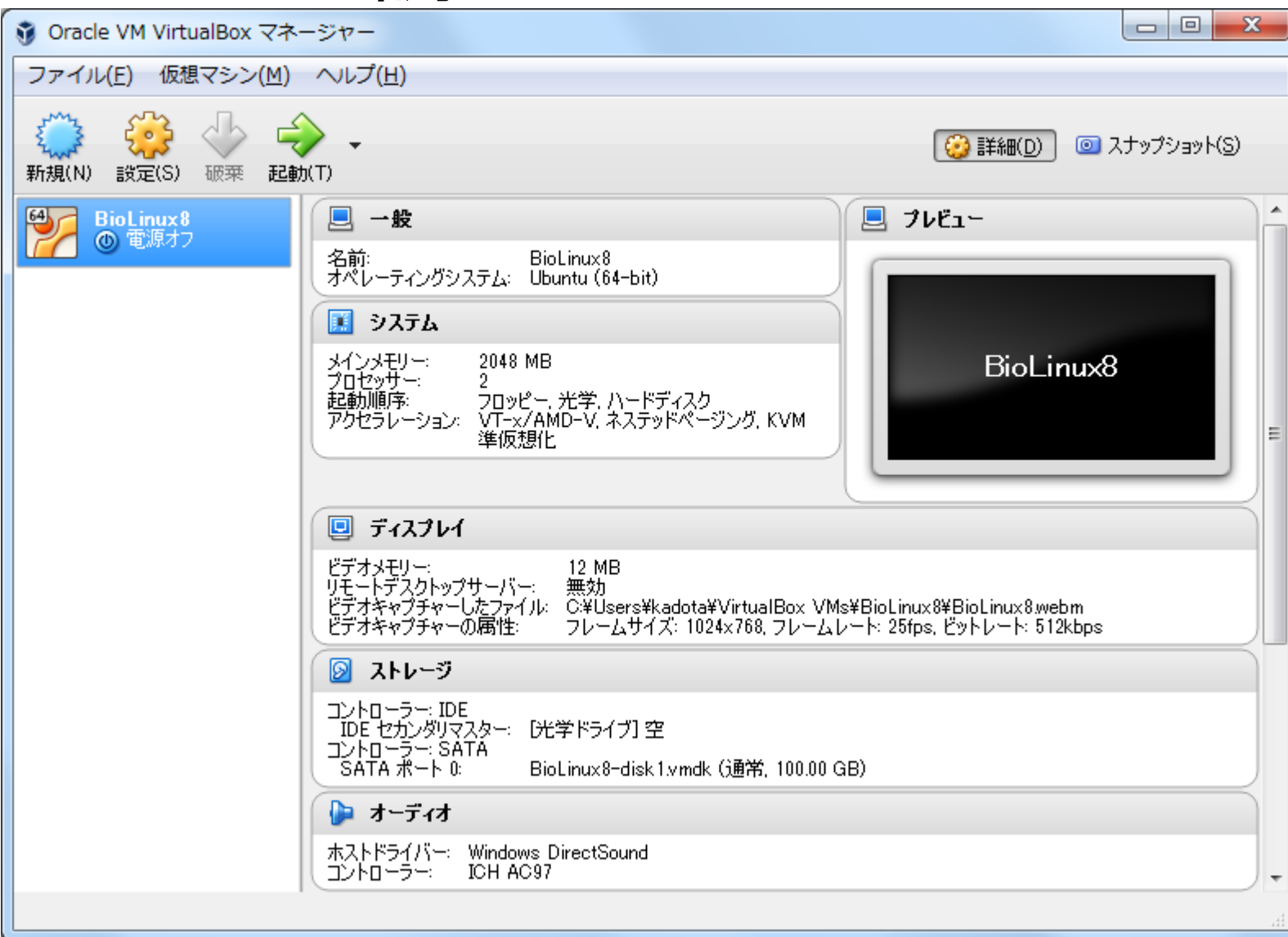
Use the ↑ and ↓ keys to select which entry is highlighted.
Press enter to boot the selected OS, `e` to edit the commands
before booting or `c` for a command-line.

W18-1: 調子が...

こんな感じになって、ログイン画面がいつまでたってもでてきません(爆)。こういうときは①×で強制終了。②仮想マシンの電源オフ、③OK



W18-1: 調子が...



W18-2: ovaの再install

①第6回のページ、②ovaファイルからのインストール手順。この中で例として「乳酸菌連載第5回終了時点のovaファイル希望」と書いているからだと思いますが、第5回のリクエストしか来ませんが、もちろん第6回でも第7回でも大丈夫です

(Rで)塩基配列解析

(last modified 2017/01/27, since 2010)

このウェブページのR関連部分は、[インストール](#)についての推奨手順 ([Windows2015.04.04版](#)と

[Macintosh2015.04.03版](#))に従ってインストールしてください。インストール済みの環境を再インストールする場合は、[再インストール](#)をご覧ください。

しています。初心者の方は、本ウェブページを

- 書籍 | [トランスクリプトーム解析 | 4.3.4 他の実験デザイン\(3群間\)](#) (last modified 2014/04/28)
- 書籍 | [日本乳酸菌学会誌 | 第1回イントロダクション](#) (last modified 2016/12/13)
- 書籍 | [日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ](#) (last modified 2016/12/22)
- 書籍 | [日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで](#) (last modified 2015/11/26)
- 書籍 | [日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール](#) (last modified 2015/12/07)
- 書籍 | [日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC](#) (last modified 2016/06/03)
- 書籍 | [日本乳酸菌学会誌 | 第6回ゲノムアセンブリ](#) (last modified 2016/07/05)
- 書籍 | [日本乳酸菌学会誌 | 第7回ロングリードアセンブリ](#) (last modified 2016/06/17)
- 書籍 | [日本乳酸菌学会誌 | 第8回アセンブリ後の解析](#) (last modified 2017/01/26) **NEW**
- 書籍 | [日本乳酸菌学会誌 | 第9回アセンブリ後の解析](#) (last modified 2016/11/10)
- 書籍 | [日本乳酸菌学会誌 | 第9回ゲノムアノテーションとその可視化](#) [DBIへの登録](#) (last modified 2017/01/31)
- イントロ | 一般 | [ランダム](#)

What's new?

- 「書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ」の検索文字列が「m(_)_m(2017/01/27)」から「m(_)_m(2017/01/27)」に変更されています。

書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ

日本乳酸菌学会誌の第6回分です。Linuxコマンドのリンク先は主に日経BP社様です。原稿PDFの「はじめに」項目のところにtypoがあります。誤:するであれば、正:する"の"であれば、ですねm(_)_m。また、「配列長によるフィルタリング」項目の最後の文章「これらについては、第7回で詳述する予定である。」についてですが、これは「これらについては、"第8回以降"で詳述する予定である。」と読み替えてください。第7回ドラフト原稿作成時点で、これらの内容は含まれていないためです(2016年4月23日追加)。

- [原稿PDF](#)
- ウェブ資料PDF
 - [Windows用](#)(2016.06.17版; 約20MB)
 - [Macintosh用](#)

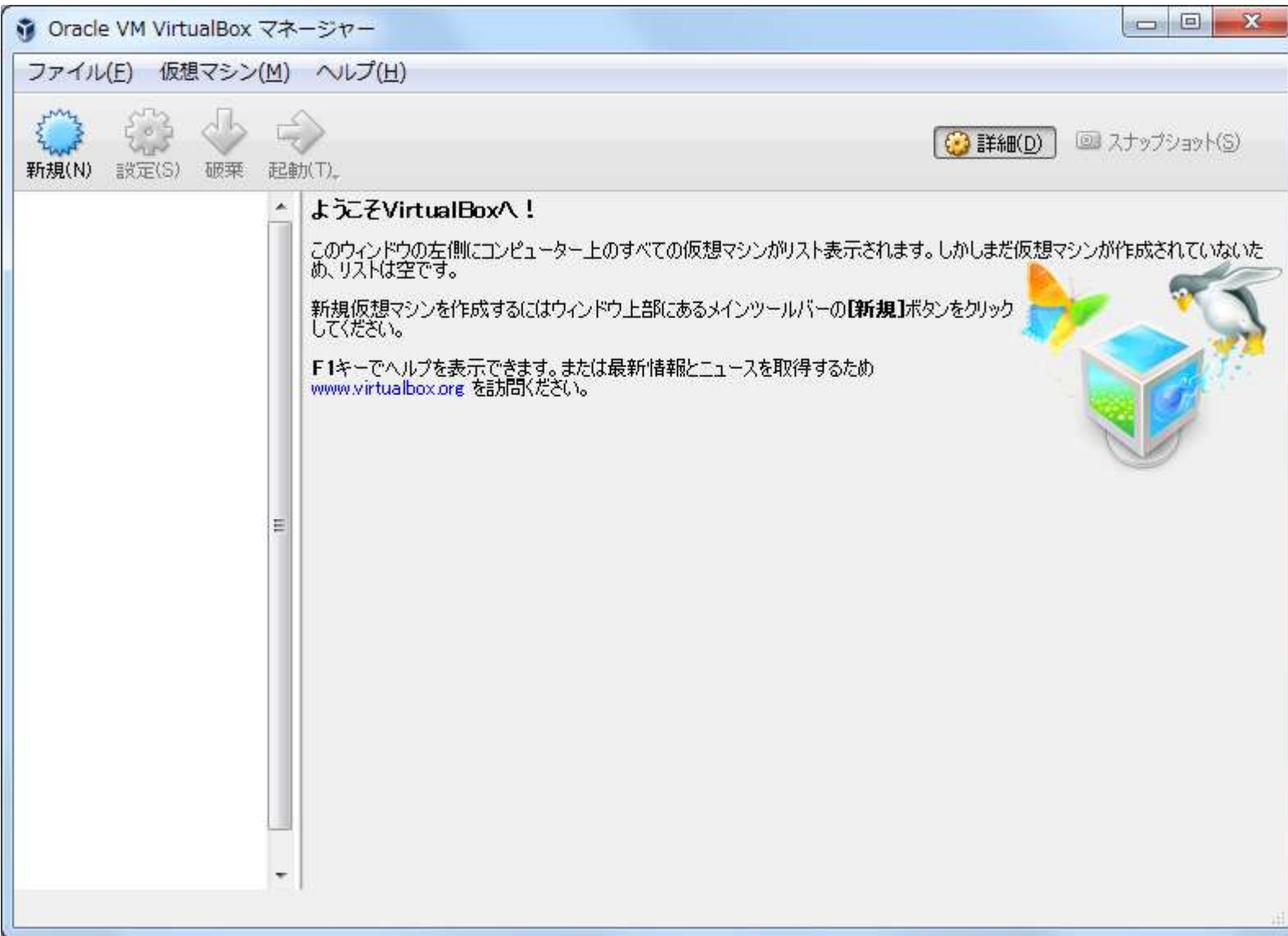
- (共有フォルダ設定情報を含む)連載第3回終了時点以降のovaファイルからのインストール手順:
 - [Windows用](#)(2015.12.28版; 約2MB)
 - [Macintosh用](#)(2015.12.28版; 約2MB)

W18-3: ova除去

「ovaファイルからのインストール手順」PDFの内容に準拠して、ここでは連載第4回終了時点のovaファイルが手元にあるという前提でざっくりとovaの再インストールを行う。①のあたりで右クリックして、②除去、③すべてのファイルを除去

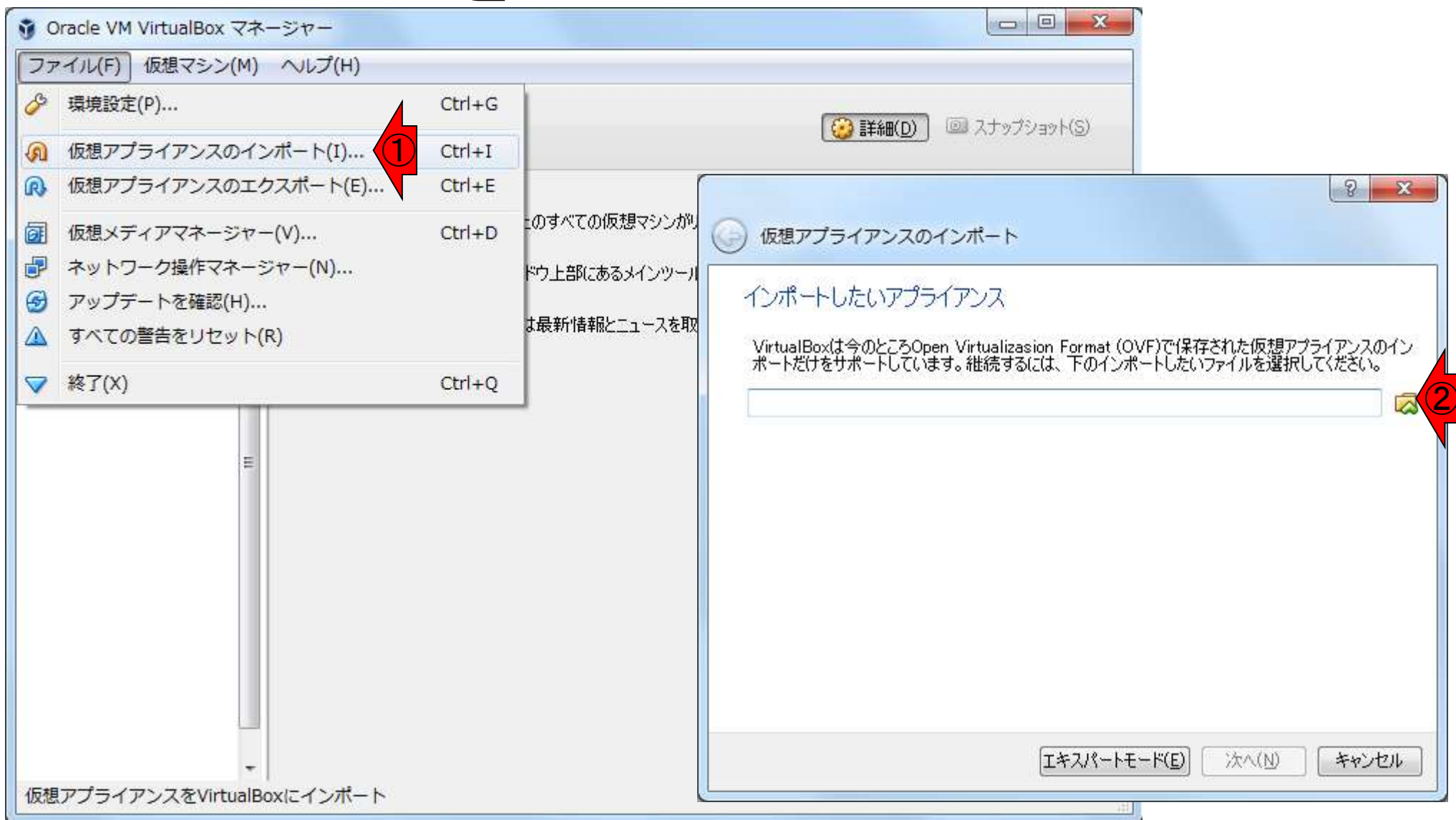
The screenshot shows the Oracle VM VirtualBox Manager interface. The main window displays a list of virtual machines, with 'BioLinux8' selected. A context menu is open over the 'BioLinux8' VM, showing options such as '設定(S)...', 'クローン(O)...', '除去(R)...', 'グループ(U)', '起動(T)', '一時停止(P)', 'リセット(R)', '閉じる(C)', '保存状態を破棄(I)...', 'ログを参照(L)...', '最新の情報に更新(F)', 'エクスプローラーに表示(H)', 'デスクトップにショートカットを作成(E)', and 'ソート(S)'. The '除去(R)...' option is highlighted, and a red arrow points to it. A dialog box titled 'VirtualBox - 質問' (VirtualBox - Question) is open, asking for confirmation to delete the VM. The dialog text reads: '以下の仮想マシンを除去しようとしています: BioLinux8. 仮想マシンを構成するファイルをハードディスクから削除しますか? 他の仮想マシンで使用されていない仮想ハードディスクも削除します。' (The following virtual machine is to be removed: BioLinux8. Do you want to delete the files that make up the virtual machine from the hard disk? Other virtual hard disks not used by other virtual machines will also be deleted.) The 'すべてのファイルを削除' (Delete all files) button is highlighted with a red arrow. The status bar at the bottom of the main window reads '選択した仮想マシンを除去' (Remove selected virtual machine).

W18-3: ova除去



W18-4: ovaをインポート

①仮想アプライアンスのインポート、②のところを押す



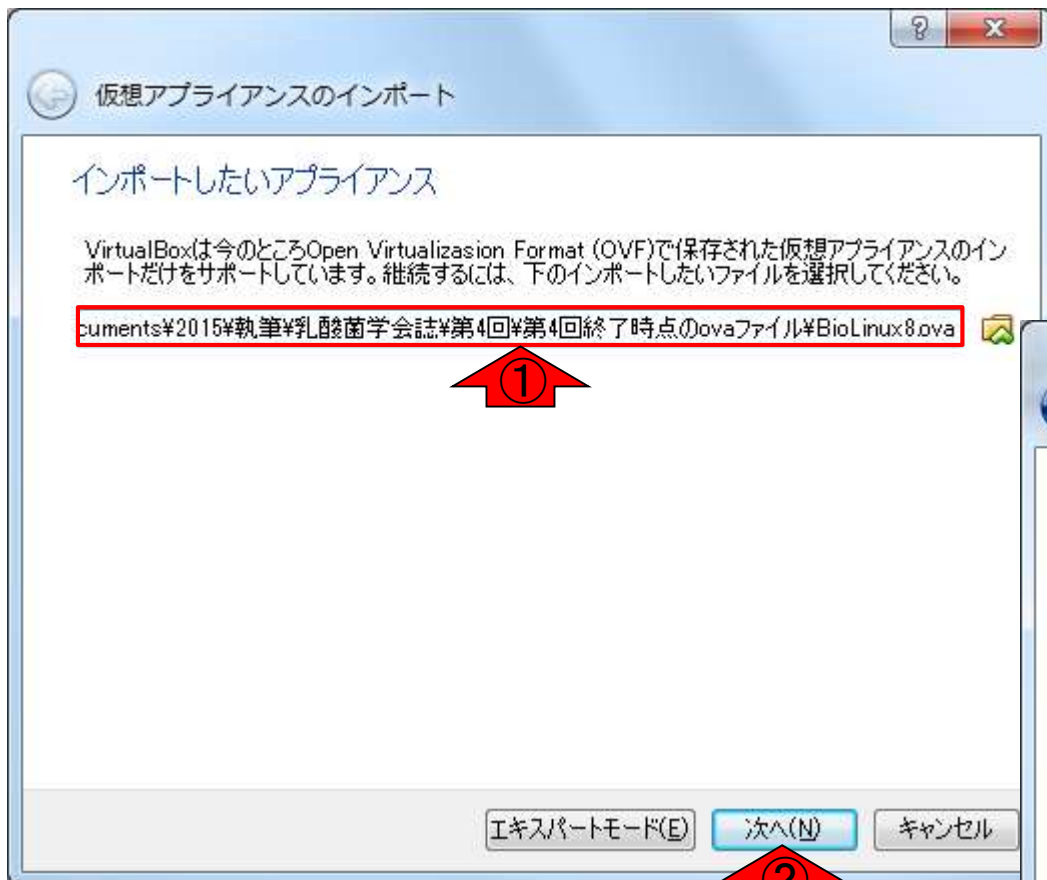
W18-5: BioLinux8.ovaを選択

インポートしたいovaファイル(ここでは第4回終了時点のovaファイル)を選択して、
②開く。このovaファイルは、消すと動作しなくなります。消さないよう注意しましょう

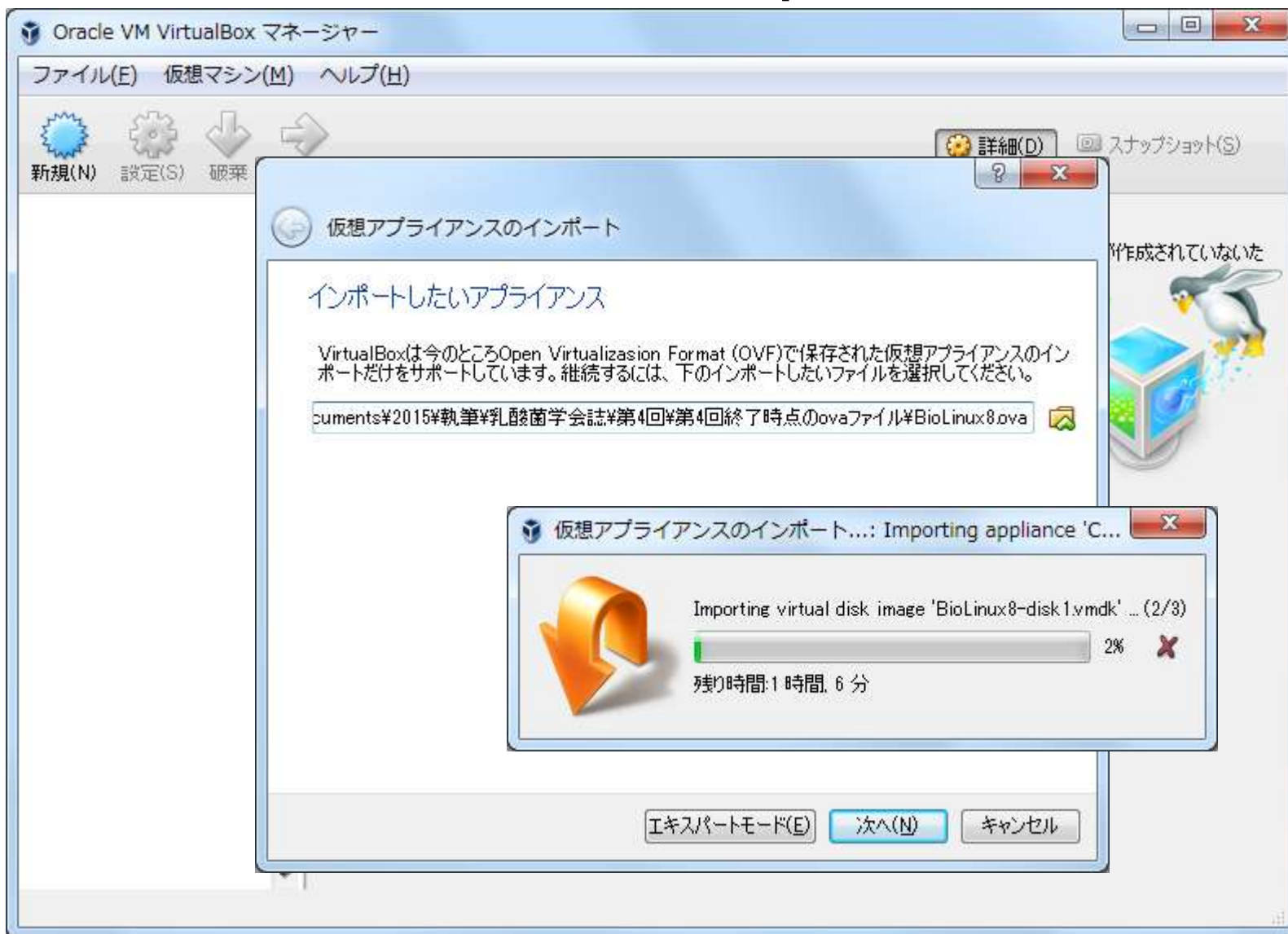


W18-6: ovaをインポート

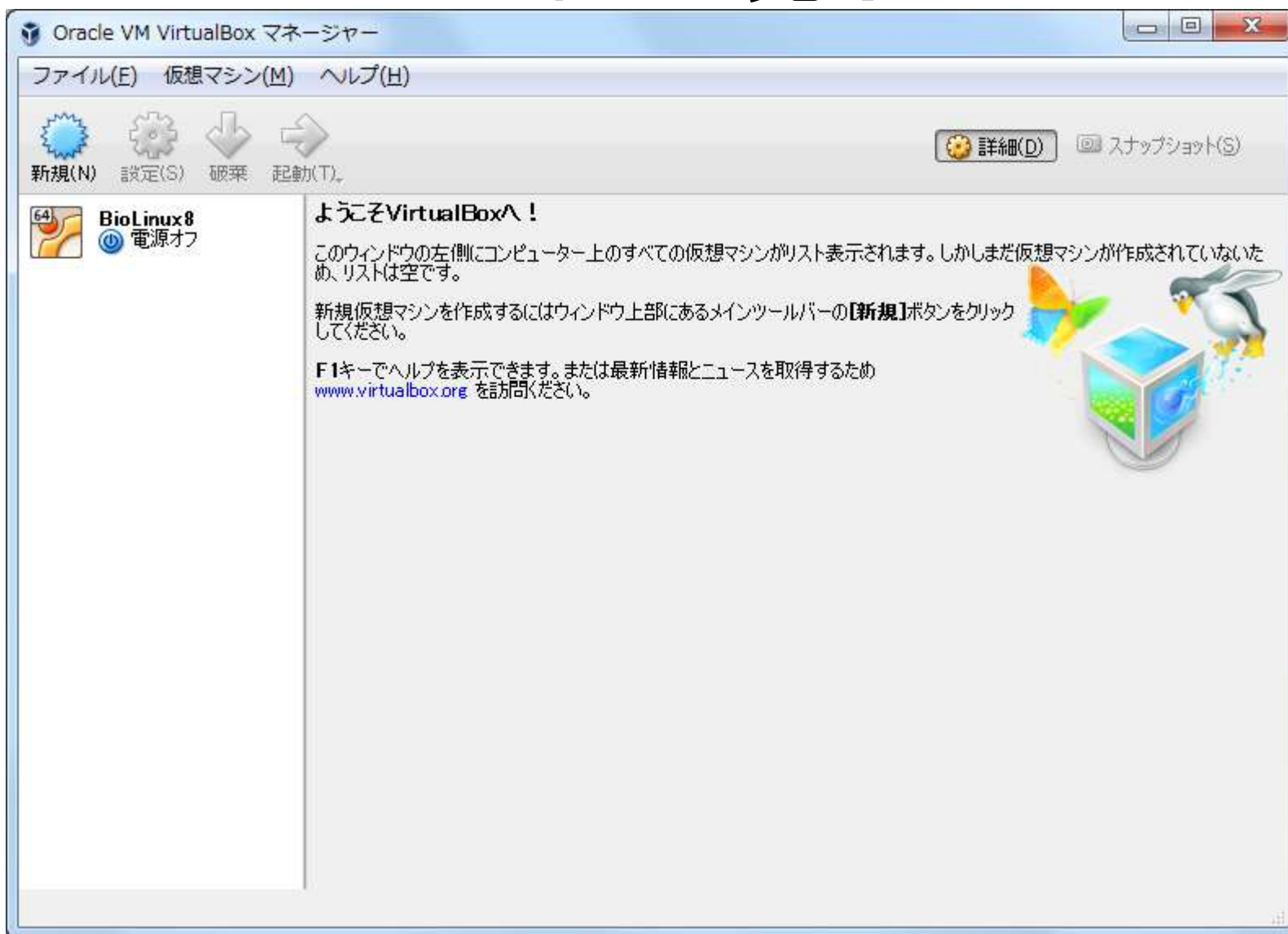
インポートしたいovaファイル(ここでは第4回終了時点のovaファイル)を選択して、②開く。このovaファイルは、消すと動作しなくなります。消さないよう注意しましょう



W18-7: インポート中

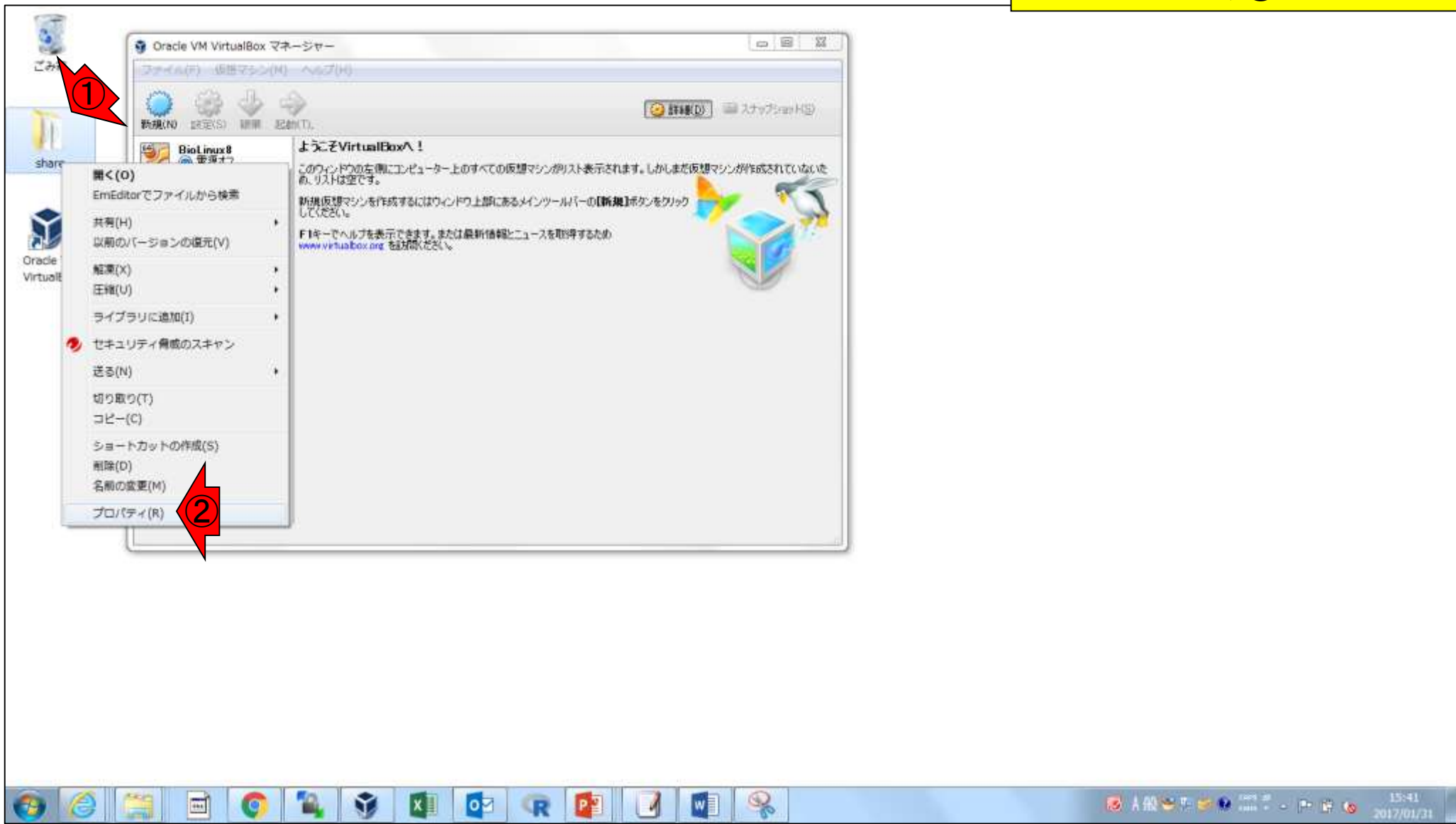


W18-8: インポート完了



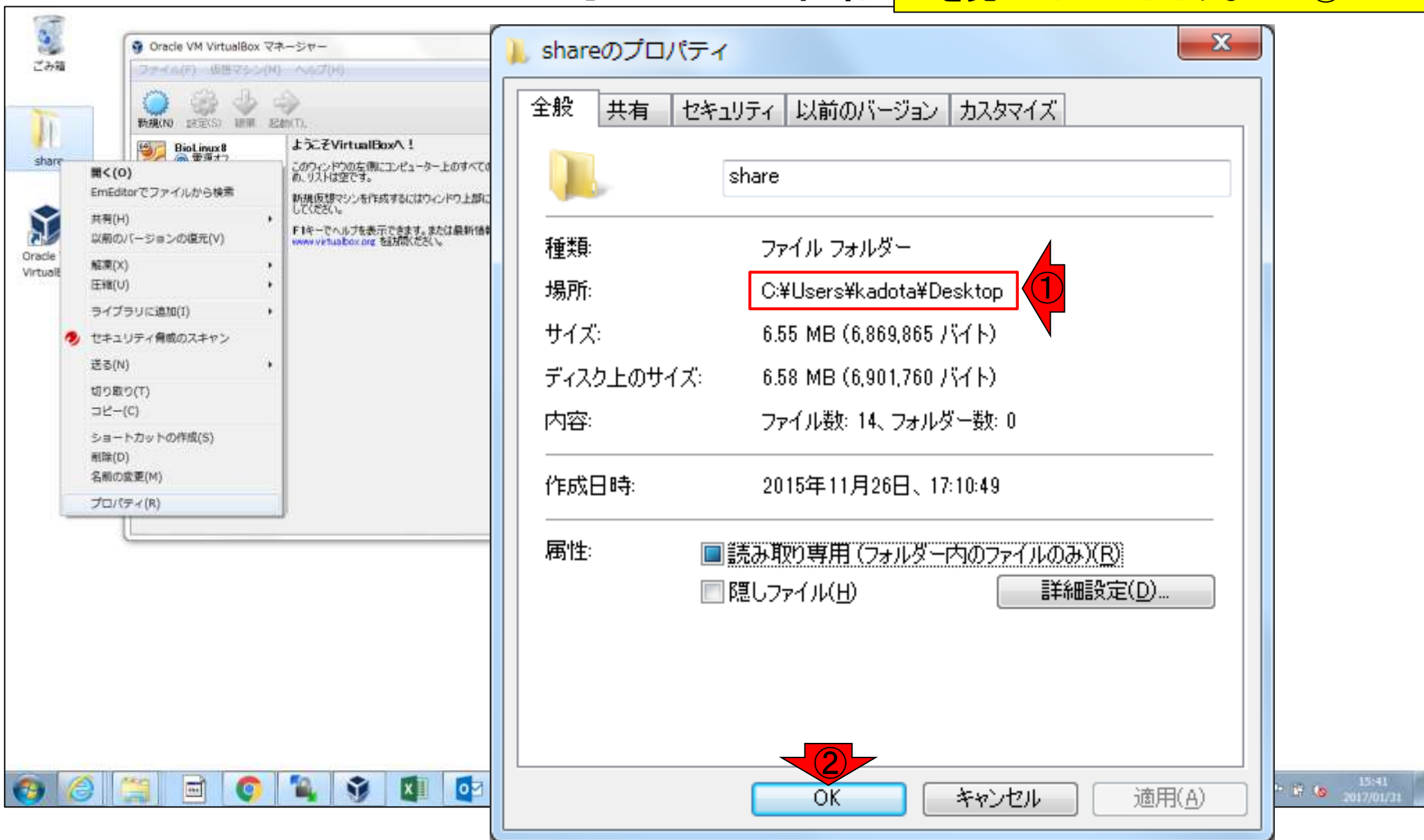
W18-9: shareフォルダ確認

①ホストOSのデスクトップ上にあるはずのshareフォルダ上で右クリックし、②プロパティ



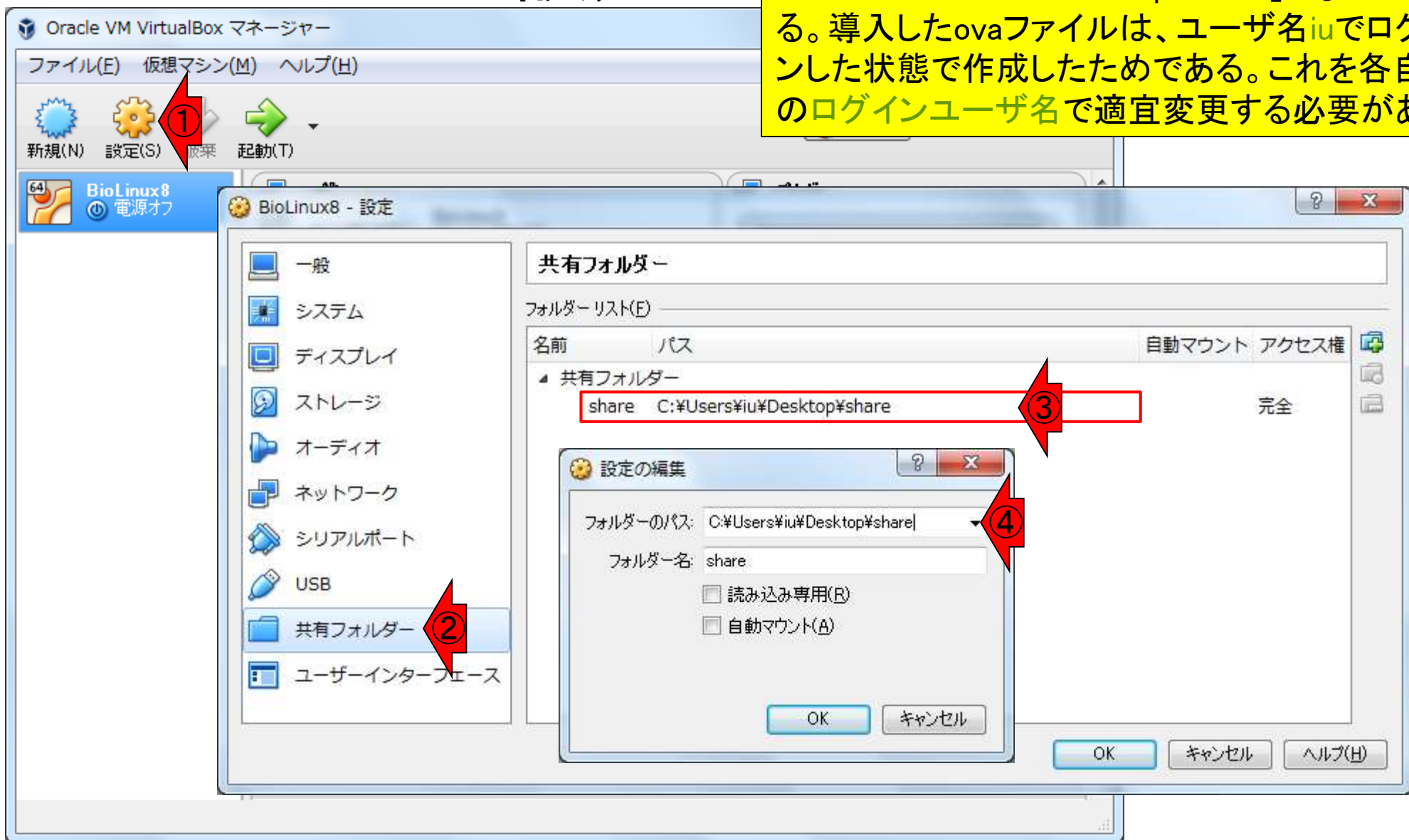
W18-9: shareフォルダ確認

ユーザ名kadotaでログインしている私の環境では①のように見える。パスを見せたかったただけなので②OK



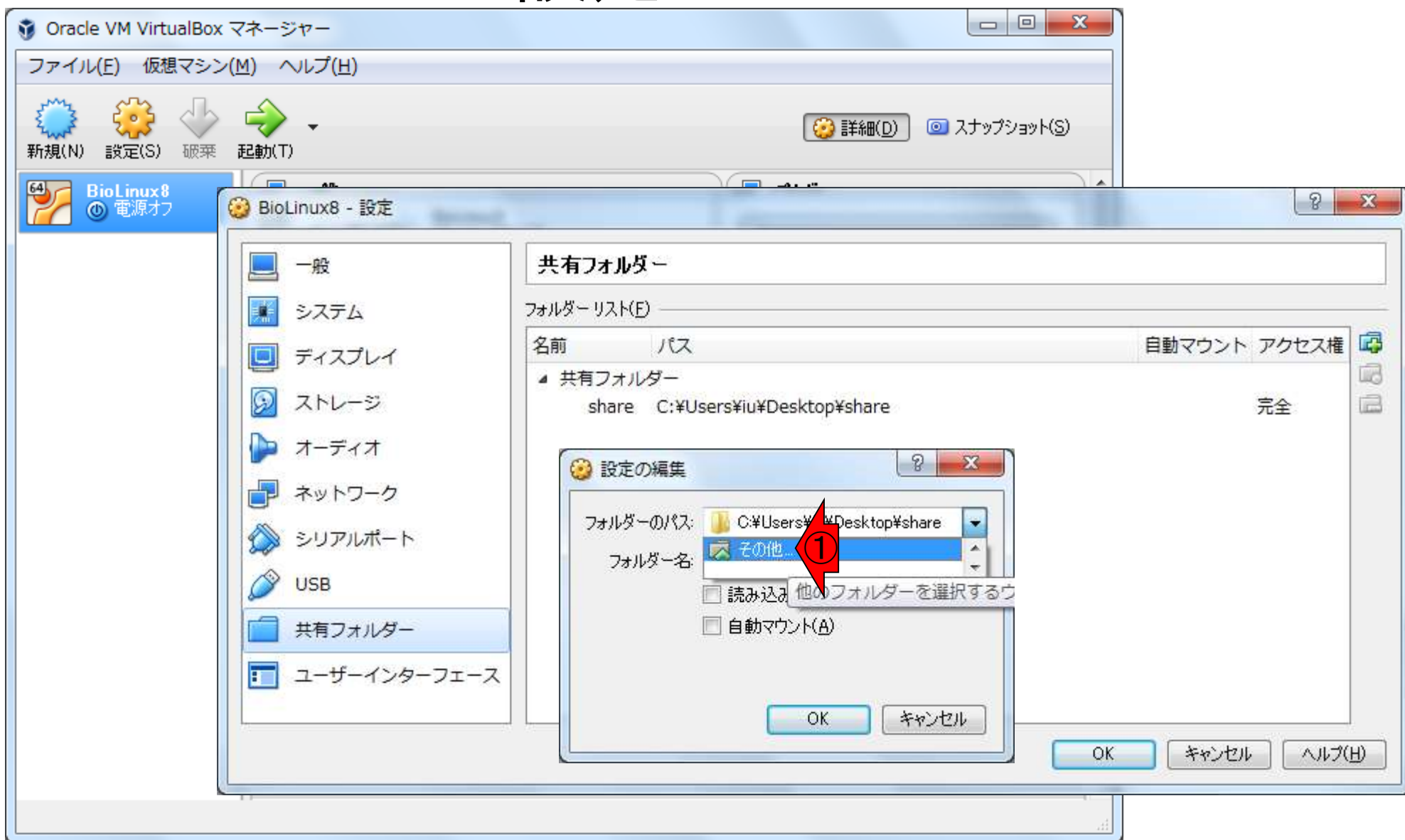
W18-10: パス設定

①設定、②共有フォルダー、③赤枠でダブルクリック。④フォルダーのパスのところがデフォルトでは「C:¥Users¥iu¥Desktop¥share」となっている。導入したovaファイルは、ユーザ名iuでログインした状態で作成したためである。これを各自のログインユーザ名で適宜変更する必要がある



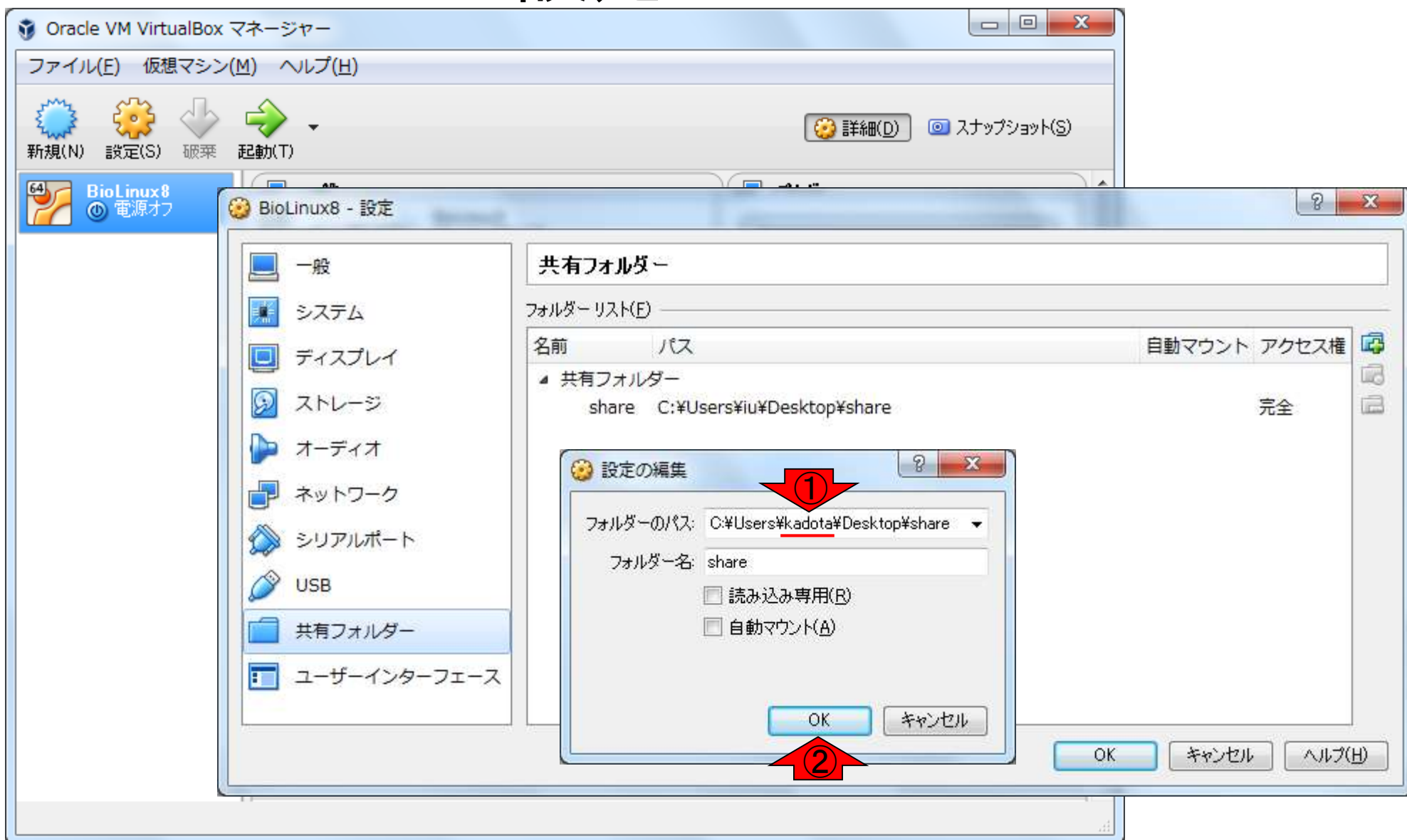
W18-10: パス設定

①「その他」を選んで、本来の「デスクトップ - share」のパスを指定



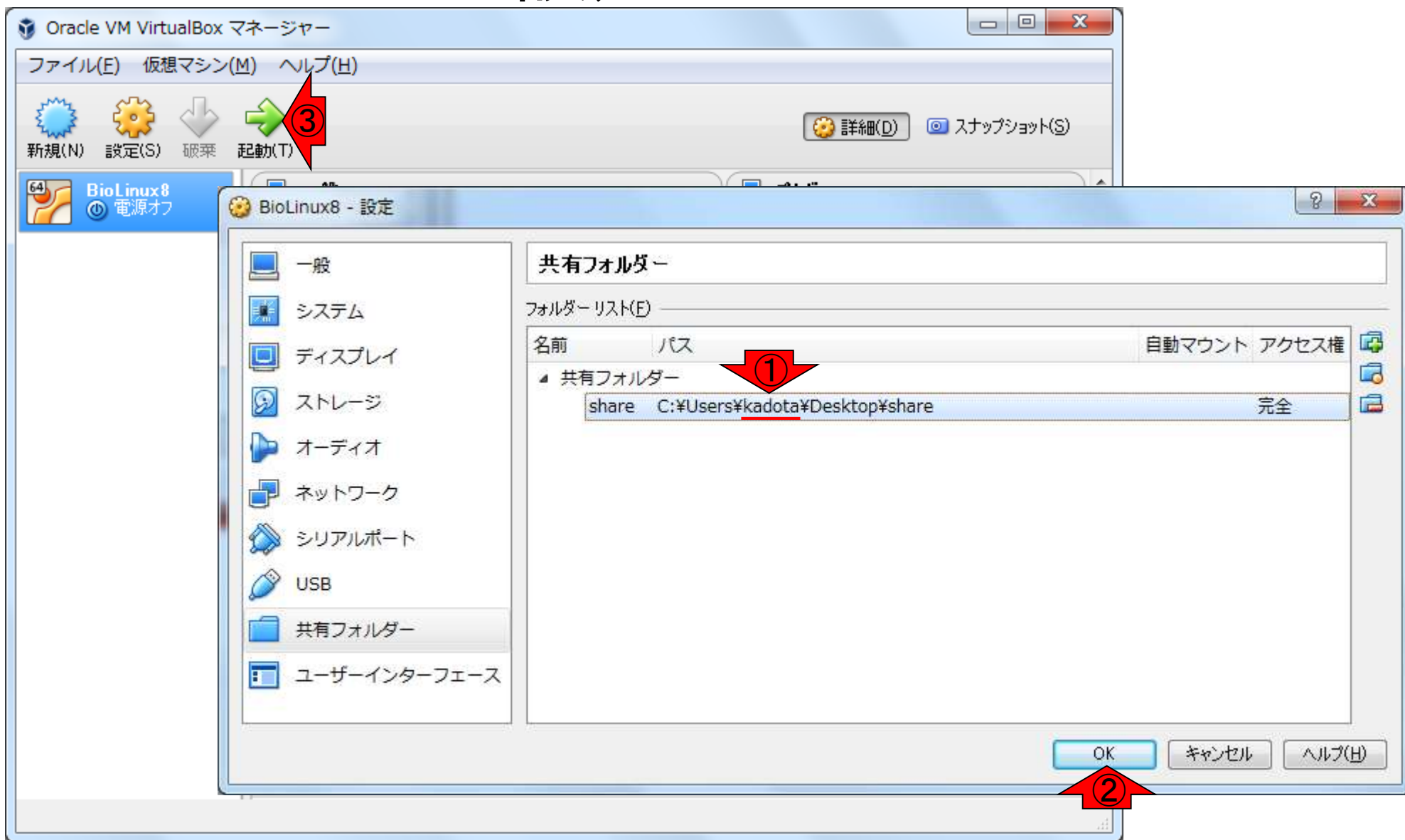
W18-10: パス設定

ユーザ名kadotaのPC環境
では①のようになる。②OK



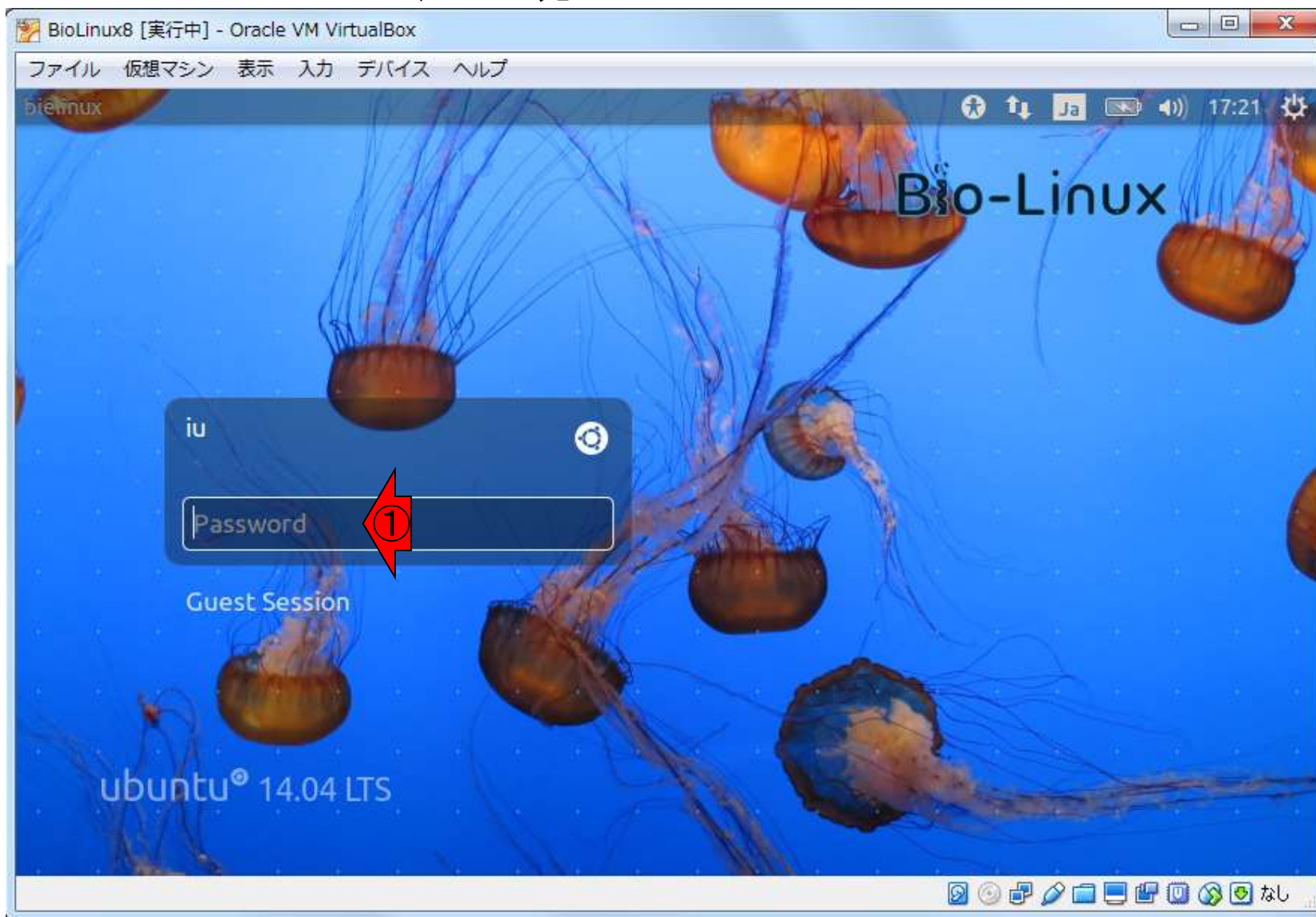
W18-10: パス設定

①の部分が変更されていることがわかります。②OK、③起動



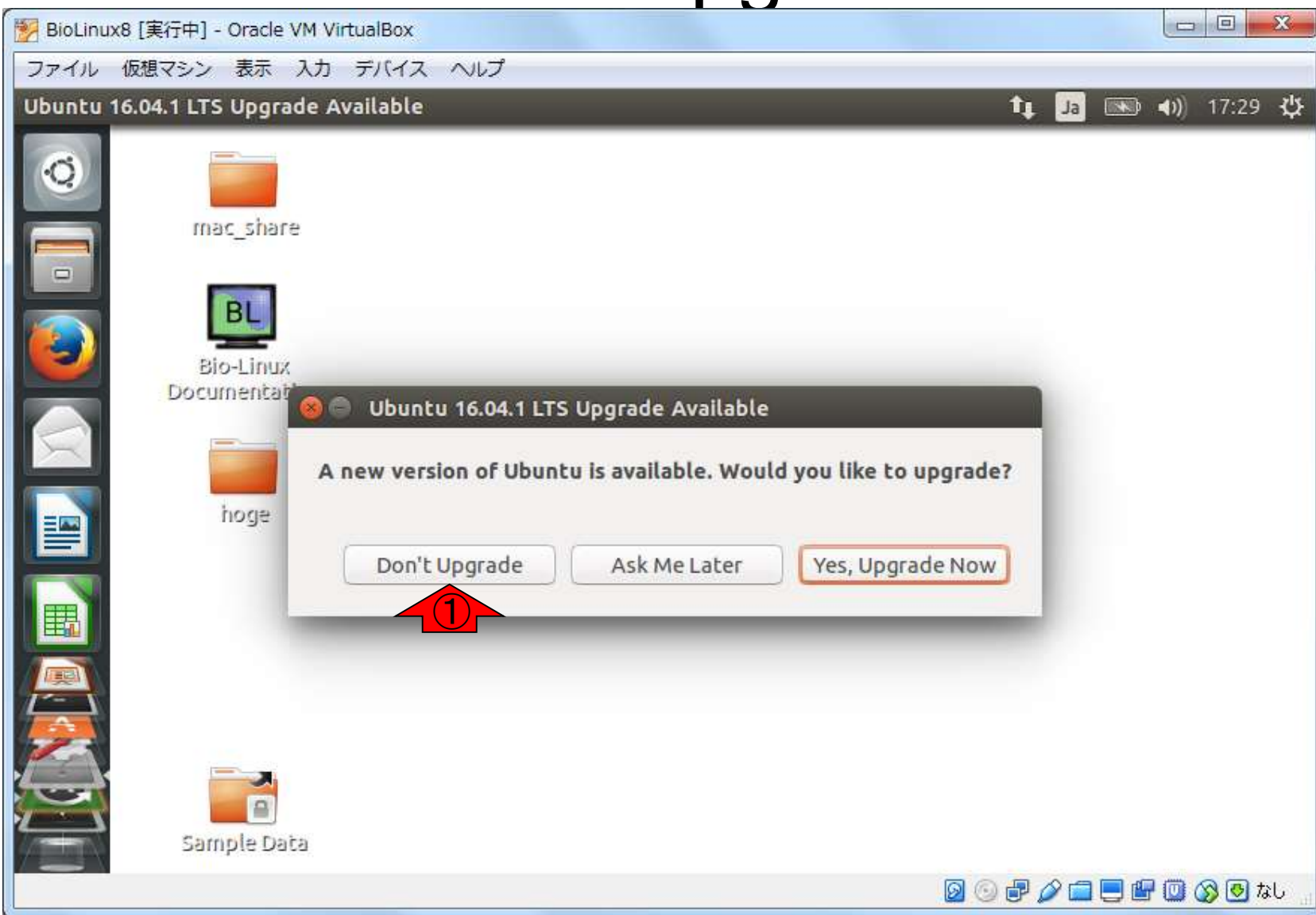
W18-11: 起動

無事復活。①パスワード
(pass1409)を入力して起動

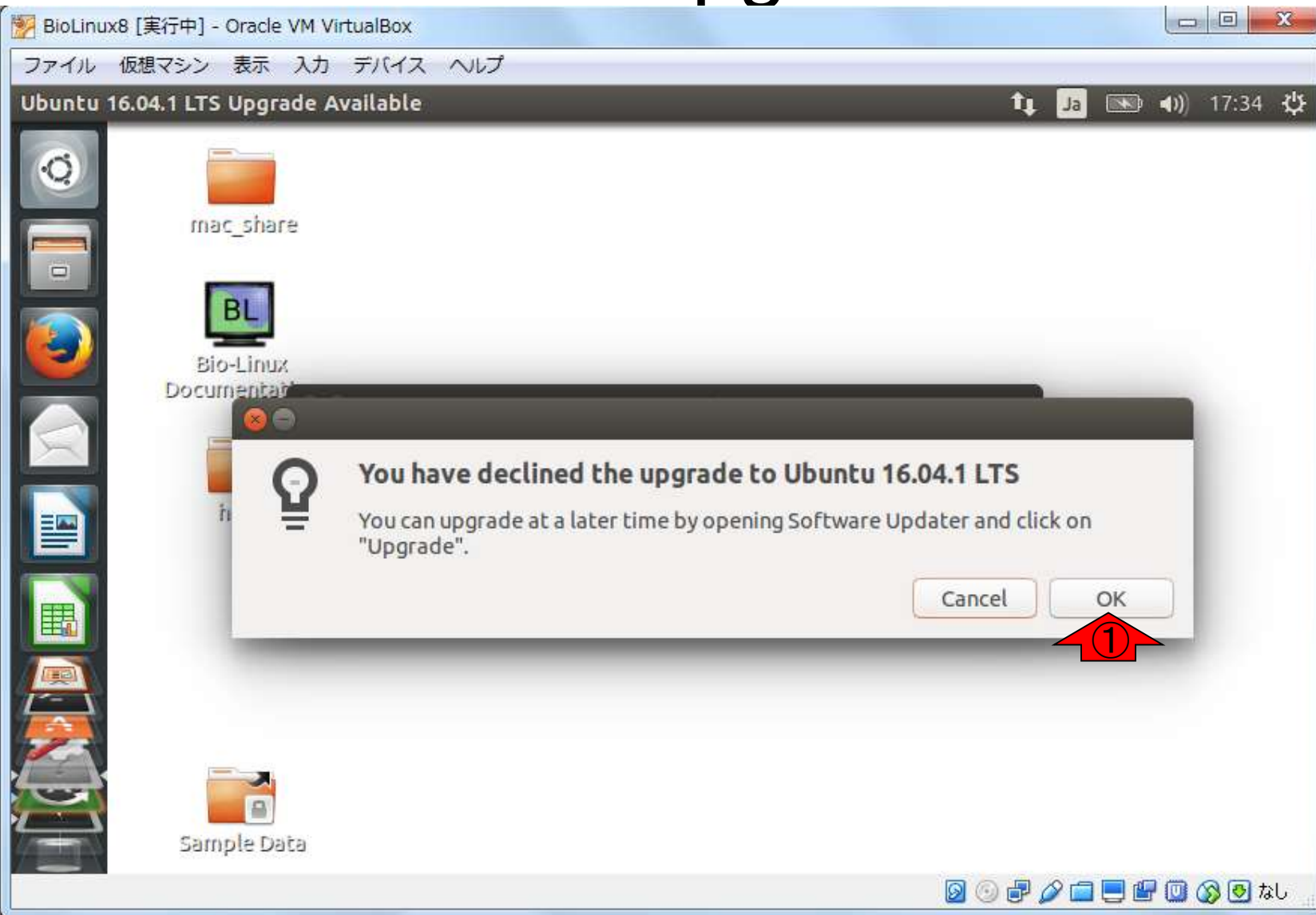


W18-11: Don't Upgrade

アップグレードしてもいいのかもしれませんが、私はいつも①Don't Upgrade

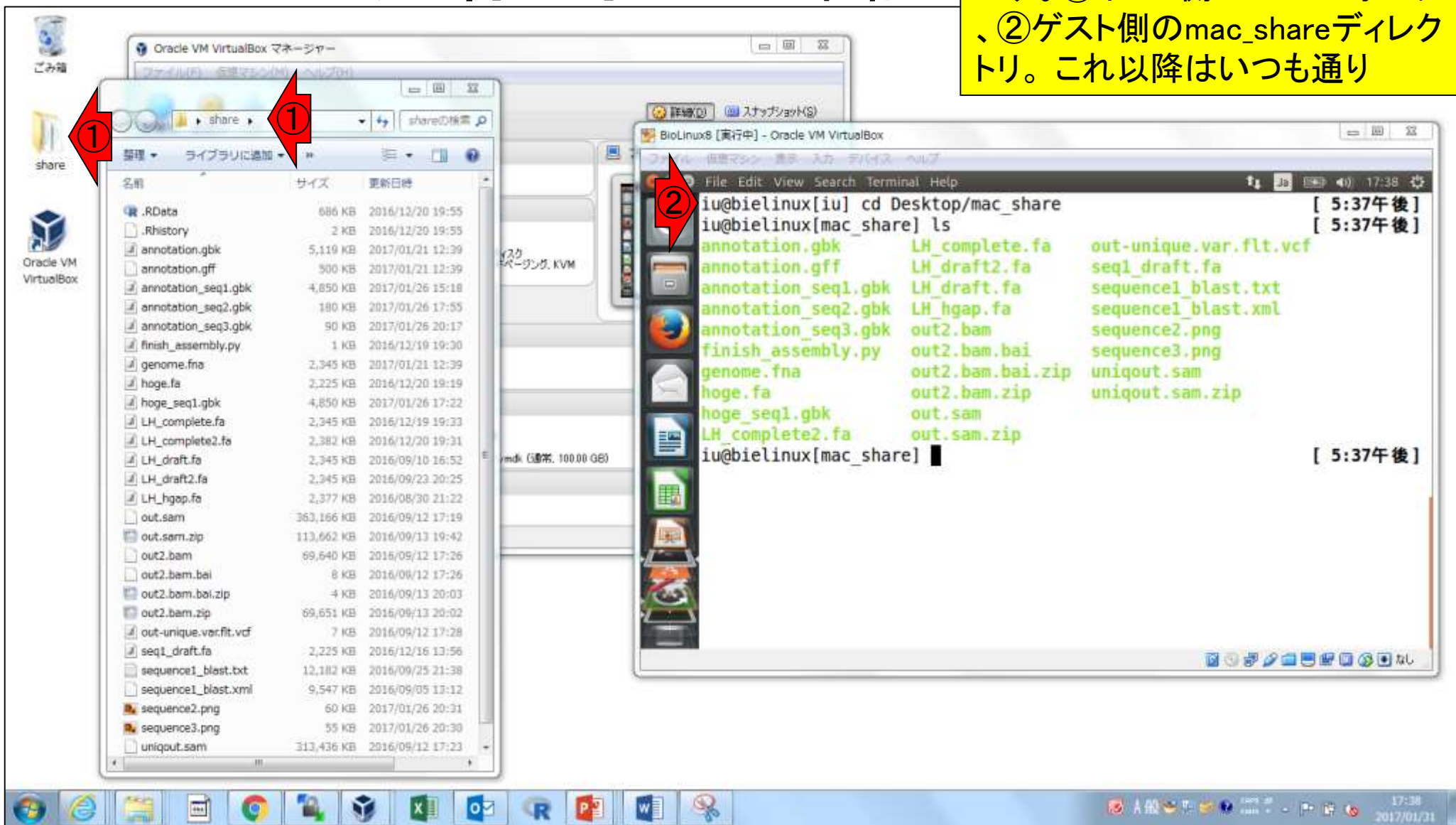


W18-11: Don't Upgrade



W18-12: 共有フォルダ確認

共有フォルダがちゃんと機能しているかどうかを確認しているだけです。①ホスト側のshareフォルダ、②ゲスト側のmac_shareディレクトリ。これ以降はいつも通り



W19-1: 配列情報つきgff

W11-4で「”配列情報付きのgffファイル”だと読み込めることが判明した」について解説。まず、annotation.gffの全体像を把握。①行数は2,493行、②1行目がヘッダ一行で、2行目以降からsequence1の情報がスタートし、`^sequence1`でgrepすればよさそうと知る。③念のためlessで確認。-Nは行番号表示(第8回W16-4)。-Sは1行分を折り返さずに表示(第8回W16-6)

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls *.gff
annotation.gff
① iu@bielinux[mac_share] wc annotation.gff
  2492  38582 511967 annotation.gff
② iu@bielinux[mac_share] head -n 3 annotation.gff
##gff-version 3
sequence1      Prodigal:2.6    CDS      101    1417    .      +      0
ID=LOCUS_00001;product=chromosomal replication initiator protein DnaA;inference=ab initio prediction:Prodigal:2.6,similar to AA sequence:INSD:BAP84581.1;locus_tag=LOCUS_00001;gene=dnaA
sequence1      Prodigal:2.6    CDS      1593   2732    .      +      0
ID=LOCUS_00002;product=DNA polymerase III subunit beta;inference=ab initio prediction:Prodigal:2.6,similar to AA sequence:INSD:BAP84582.1;locus_tag=LOCUS_00002;gene=dnaN
③ iu@bielinux[mac_share] less -N -S annotation.gff [ 6:37午後]
```


W19-2: less復習

```

iu@bielinux[~/Desktop/mac_share]
1 ##gff-version 3
2 sequence1      Prodigal:2.6    CDS      101      1417      .      +
3 sequence1      Prodigal:2.6    CDS     1593     2732      .      +
4 sequence1      Prodigal:2.6    CDS     2974     3198      .      +
5 sequence1      Prodigal:2.6    CDS     3208     4329      .      +
6 sequence1      Prodigal:2.6    CDS     4329     6272      .      +
7 sequence1      Prodigal:2.6    CDS     6300     8861      .      +
8 sequence1      Prodigal:2.6    CDS     9052    10416     .      +
9 sequence1      Prodigal:2.6    CDS    10610    10906     .      +
10 sequence1     Prodigal:2.6    CDS    10944    11474     .      +
11 sequence1     Prodigal:2.6    CDS    11499    11735     .      +
12 sequence1     Prodigal:2.6    CDS    11971    12528     .      -
13 sequence1     Prodigal:2.6    CDS    12648    13400     .      +
14 sequence1     Prodigal:2.6    CDS    13621    15642     .      +
15 sequence1     Prodigal:2.6    CDS    15649    16101     .      +
16 sequence1     Prodigal:2.6    CDS    16123    17508     .      +
17 sequence1     Prodigal:2.6    CDS    17584    18762     .      +
18 sequence1     Prodigal:2.6    CDS    18769    19089     .      +
19 sequence1     Prodigal:2.6    CDS    19279    21024     .      +
20 sequence1     Prodigal:2.6    CDS    21038    21961     .      +
annotation.gff

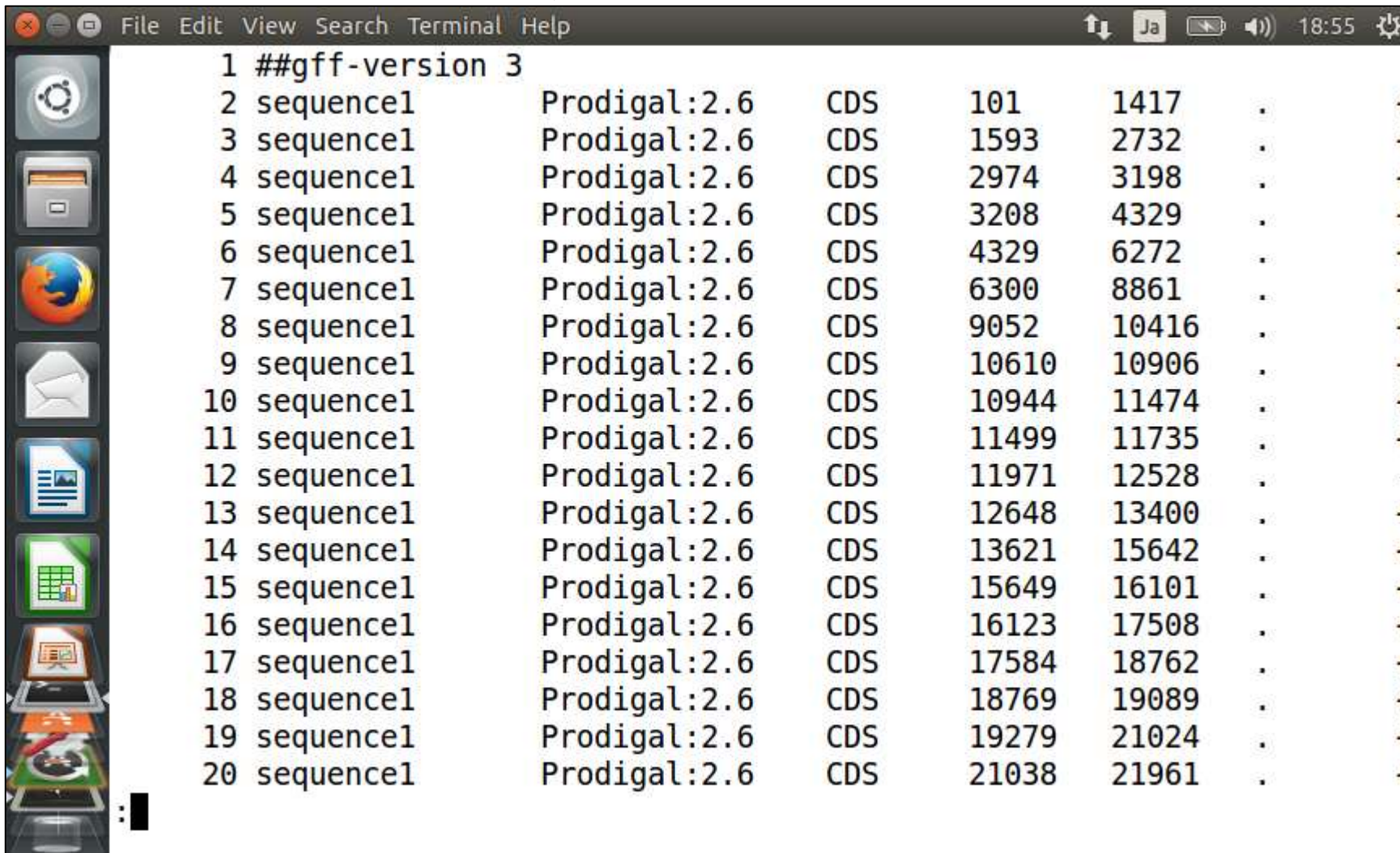
```

W19-2: less 復習

大文字のGを打ち込んで最終行を表示させたところ。sequence3については`^sequence3`でgrepすればよいと確信を得る

```
iu@bielinux[~/Desktop/mac_share]
2473 sequence3      Prodigal:2.6      CDS      25476      27407      .      -
2474 sequence3      Prodigal:2.6      CDS      27593      27904      .      -
2475 sequence3      Prodigal:2.6      CDS      27973      28335      .      +
2476 sequence3      Prodigal:2.6      CDS      28440      29024      .      +
2477 sequence3      Prodigal:2.6      CDS      29073      29936      .      +
2478 sequence3      Prodigal:2.6      CDS      29968      30798      .      +
2479 sequence3      Prodigal:2.6      CDS      31000      31353      .      -
2480 sequence3      Prodigal:2.6      CDS      31362      32036      .      -
2481 sequence3      Prodigal:2.6      CDS      32041      33087      .      -
2482 sequence3      Prodigal:2.6      CDS      33272      34114      .      -
2483 sequence3      Prodigal:2.6      CDS      34513      34827      .      +
2484 sequence3      Prodigal:2.6      CDS      34808      35131      .      +
2485 sequence3      Prodigal:2.6      CDS      35502      35987      .      +
2486 sequence3      Prodigal:2.6      CDS      36000      36389      .      +
2487 sequence3      Prodigal:2.6      CDS      36596      37096      .      +
2488 sequence3      Prodigal:2.6      CDS      37239      38168      .      +
2489 sequence3      Prodigal:2.6      CDS      38236      38502      .      -
2490 sequence3      Prodigal:2.6      CDS      38518      38637      .      -
2491 sequence3      Prodigal:2.6      CDS      38709      38957      .      -
2492 sequence3      Prodigal:2.6      CDS      39169      40065      .      +
(END)
```

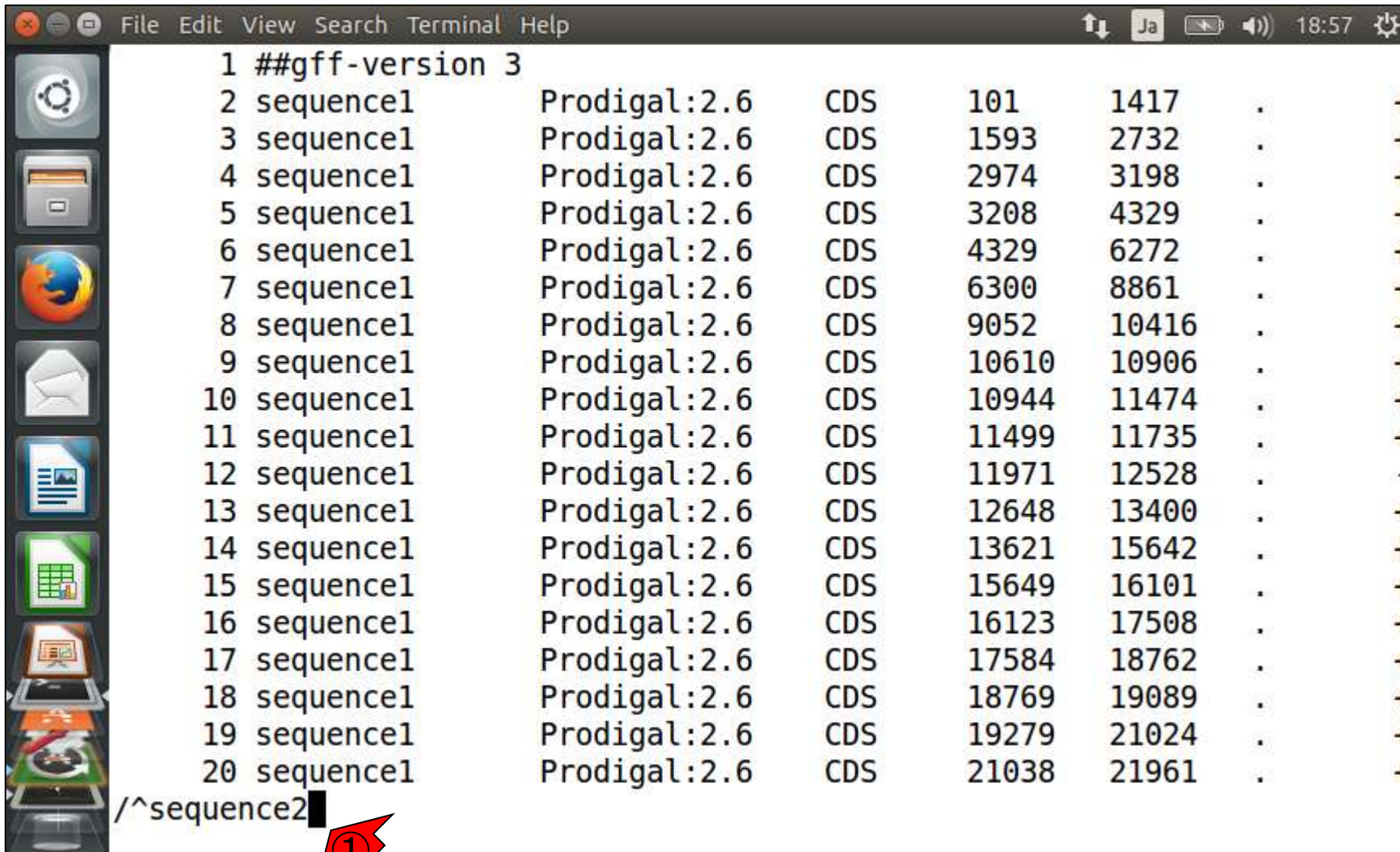
W19-2: less 復習



```
File Edit View Search Terminal Help 1 ##gff-version 3 2 sequence1 Prodigal:2.6 CDS 101 1417 . + 3 sequence1 Prodigal:2.6 CDS 1593 2732 . + 4 sequence1 Prodigal:2.6 CDS 2974 3198 . + 5 sequence1 Prodigal:2.6 CDS 3208 4329 . + 6 sequence1 Prodigal:2.6 CDS 4329 6272 . + 7 sequence1 Prodigal:2.6 CDS 6300 8861 . + 8 sequence1 Prodigal:2.6 CDS 9052 10416 . + 9 sequence1 Prodigal:2.6 CDS 10610 10906 . + 10 sequence1 Prodigal:2.6 CDS 10944 11474 . + 11 sequence1 Prodigal:2.6 CDS 11499 11735 . + 12 sequence1 Prodigal:2.6 CDS 11971 12528 . - 13 sequence1 Prodigal:2.6 CDS 12648 13400 . + 14 sequence1 Prodigal:2.6 CDS 13621 15642 . + 15 sequence1 Prodigal:2.6 CDS 15649 16101 . + 16 sequence1 Prodigal:2.6 CDS 16123 17508 . + 17 sequence1 Prodigal:2.6 CDS 17584 18762 . + 18 sequence1 Prodigal:2.6 CDS 18769 19089 . + 19 sequence1 Prodigal:2.6 CDS 19279 21024 . + 20 sequence1 Prodigal:2.6 CDS 21038 21961 . +
```

W19-2: less 復習

①「/^sequence2」と打ってリターン。grepと同じで
行頭がsequence2となっているものをハイライト



```
File Edit View Search Terminal Help 18:57
1 ##gff-version 3
2 sequence1      Prodigal:2.6    CDS      101      1417      .      +
3 sequence1      Prodigal:2.6    CDS     1593     2732      .      +
4 sequence1      Prodigal:2.6    CDS     2974     3198      .      +
5 sequence1      Prodigal:2.6    CDS     3208     4329      .      +
6 sequence1      Prodigal:2.6    CDS     4329     6272      .      +
7 sequence1      Prodigal:2.6    CDS     6300     8861      .      +
8 sequence1      Prodigal:2.6    CDS     9052    10416     .      +
9 sequence1      Prodigal:2.6    CDS    10610    10906     .      +
10 sequence1     Prodigal:2.6    CDS    10944    11474     .      +
11 sequence1     Prodigal:2.6    CDS    11499    11735     .      +
12 sequence1     Prodigal:2.6    CDS    11971    12528     .      -
13 sequence1     Prodigal:2.6    CDS    12648    13400     .      +
14 sequence1     Prodigal:2.6    CDS    13621    15642     .      +
15 sequence1     Prodigal:2.6    CDS    15649    16101     .      +
16 sequence1     Prodigal:2.6    CDS    16123    17508     .      +
17 sequence1     Prodigal:2.6    CDS    17584    18762     .      +
18 sequence1     Prodigal:2.6    CDS    18769    19089     .      +
19 sequence1     Prodigal:2.6    CDS    19279    21024     .      +
20 sequence1     Prodigal:2.6    CDS    21038    21961     .      +
/^sequence2
```



W19-2: less 復習

させたところ。①sequence2のアノテーション情報は、2,340行目からスタートしているのだろう

```
iu@bielinux[~/Desktop/mac_share] 18:59
① 2340 sequence2      Prodigal:2.6    CDS      764      1486      .      +
    2341 sequence2      Prodigal:2.6    CDS     2095     2574      .      -
    2342 sequence2      Prodigal:2.6    CDS     2931     3560      .      +
    2343 sequence2      Prodigal:2.6    CDS     3780     5156      .      -
    2344 sequence2      Prodigal:2.6    CDS     5272     6477      .      -
    2345 sequence2      Prodigal:2.6    CDS     6837     7421      .      -
    2346 sequence2      Prodigal:2.6    CDS     7619     8179      .      +
    2347 sequence2      Prodigal:2.6    CDS     8191     8376      .      +
    2348 sequence2      Prodigal:2.6    CDS     8404     8946      .      +
    2349 sequence2      Prodigal:2.6    CDS     8961     9212      .      +
    2350 sequence2      Prodigal:2.6    CDS     9224     9364      .      +
    2351 sequence2      Prodigal:2.6    CDS     9562     9777      .      +
    2352 sequence2      Prodigal:2.6    CDS    10240    10608      .      -
    2353 sequence2      Prodigal:2.6    CDS    10767    11576      .      -
    2354 sequence2      Prodigal:2.6    CDS    11586    12341      .      -
    2355 sequence2      Prodigal:2.6    CDS    12366    13136      .      -
    2356 sequence2      Prodigal:2.6    CDS    13137    13967      .      -
    2357 sequence2      Prodigal:2.6    CDS    13951    15324      .      -
    2358 sequence2      Prodigal:2.6    CDS    15330    15746      .      -
    2359 sequence2      Prodigal:2.6    CDS    15748    16173      .      -
```

W19-2: less 復習

キーボードの上矢印キーを1回押したところ。
①「sequence2のアノテーション情報は2,340行目からスタートしている」ことを確信

```
iu@bielinux[~/Desktop/mac_share] 19:02
2339 sequence1      SignalP:4.1      sig_peptide      2277497 2277592 .
2340 sequence2      Prodigal:2.6     CDS      764      1486      .      +
2341 sequence2      Prodigal:2.6     CDS      2095     2574      .      -
2342 sequence2      Prodigal:2.6     CDS      2931     3560      .      +
2343 sequence2      Prodigal:2.6     CDS      3780     5156      .      -
2344 sequence2      Prodigal:2.6     CDS      5272     6477      .      -
2345 sequence2      Prodigal:2.6     CDS      6837     7421      .      -
2346 sequence2      Prodigal:2.6     CDS      7619     8179      .      +
2347 sequence2      Prodigal:2.6     CDS      8191     8376      .      +
2348 sequence2      Prodigal:2.6     CDS      8404     8946      .      +
2349 sequence2      Prodigal:2.6     CDS      8961     9212      .      +
2350 sequence2      Prodigal:2.6     CDS      9224     9364      .      +
2351 sequence2      Prodigal:2.6     CDS      9562     9777      .      +
2352 sequence2      Prodigal:2.6     CDS      10240    10608     .      -
2353 sequence2      Prodigal:2.6     CDS      10767    11576     .      -
2354 sequence2      Prodigal:2.6     CDS      11586    12341     .      -
2355 sequence2      Prodigal:2.6     CDS      12366    13136     .      -
2356 sequence2      Prodigal:2.6     CDS      13137    13967     .      -
2357 sequence2      Prodigal:2.6     CDS      13951    15324     .      -
2358 sequence2      Prodigal:2.6     CDS      15330    15746     .      -
```

W19-2: less 復習

```

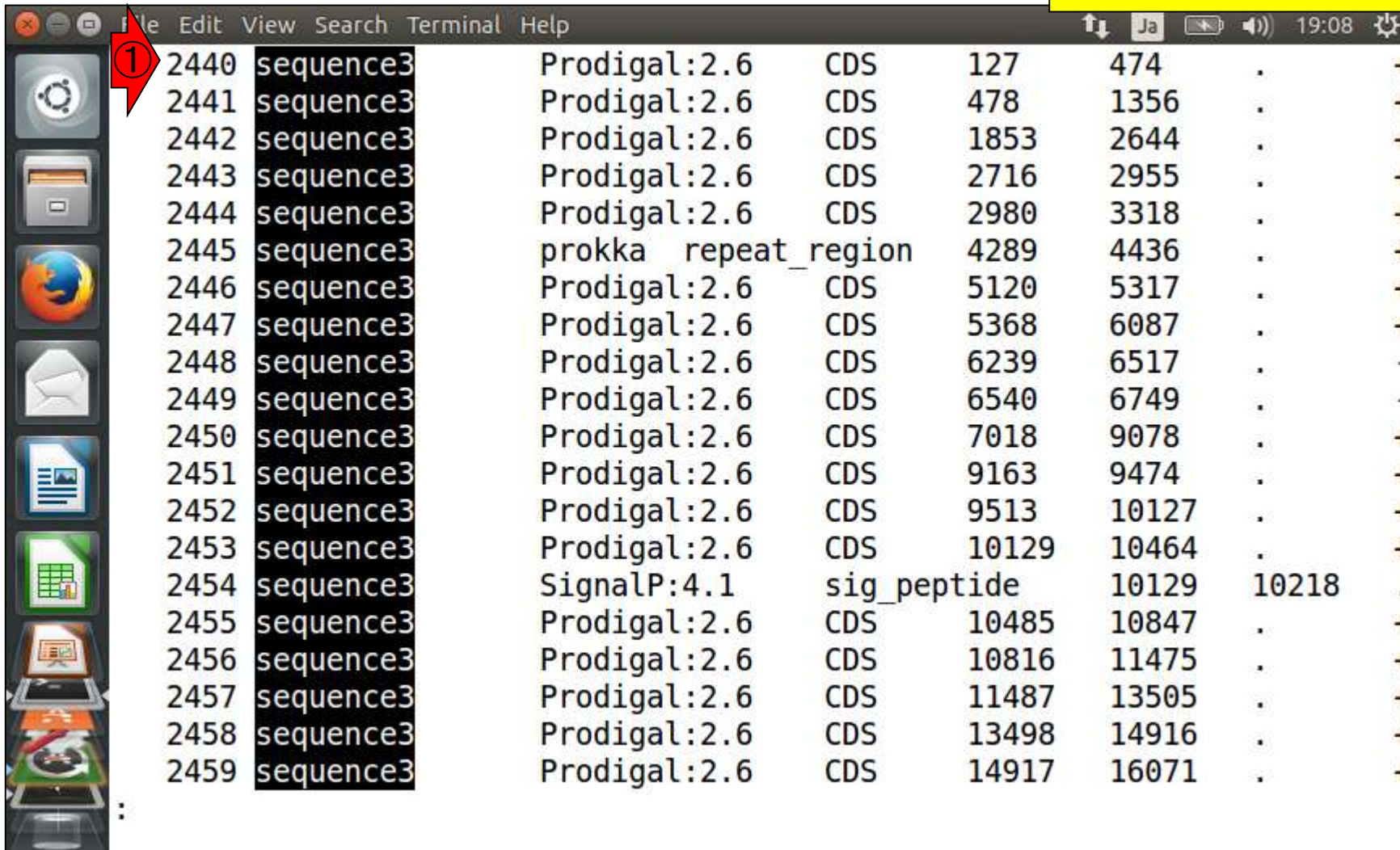
File Edit View Search Terminal Help
2339 sequence1      SignalP:4.1      sig_peptide      2277497 2277592 .
2340 sequence2      Prodigal:2.6     CDS       764    1486    .    +
2341 sequence2      Prodigal:2.6     CDS      2095    2574    .    -
2342 sequence2      Prodigal:2.6     CDS      2931    3560    .    +
2343 sequence2      Prodigal:2.6     CDS      3780    5156    .    -
2344 sequence2      Prodigal:2.6     CDS      5272    6477    .    -
2345 sequence2      Prodigal:2.6     CDS      6837    7421    .    -
2346 sequence2      Prodigal:2.6     CDS      7619    8179    .    +
2347 sequence2      Prodigal:2.6     CDS      8191    8376    .    +
2348 sequence2      Prodigal:2.6     CDS      8404    8946    .    +
2349 sequence2      Prodigal:2.6     CDS      8961    9212    .    +
2350 sequence2      Prodigal:2.6     CDS      9224    9364    .    +
2351 sequence2      Prodigal:2.6     CDS      9562    9777    .    +
2352 sequence2      Prodigal:2.6     CDS     10240   10608    .    -
2353 sequence2      Prodigal:2.6     CDS     10767   11576    .    -
2354 sequence2      Prodigal:2.6     CDS     11586   12341    .    -
2355 sequence2      Prodigal:2.6     CDS     12366   13136    .    -
2356 sequence2      Prodigal:2.6     CDS     13137   13967    .    -
2357 sequence2      Prodigal:2.6     CDS     13951   15324    .    -
2358 sequence2      Prodigal:2.6     CDS     15330   15746    .    -
/sequence3

```



W19-2: less 復習

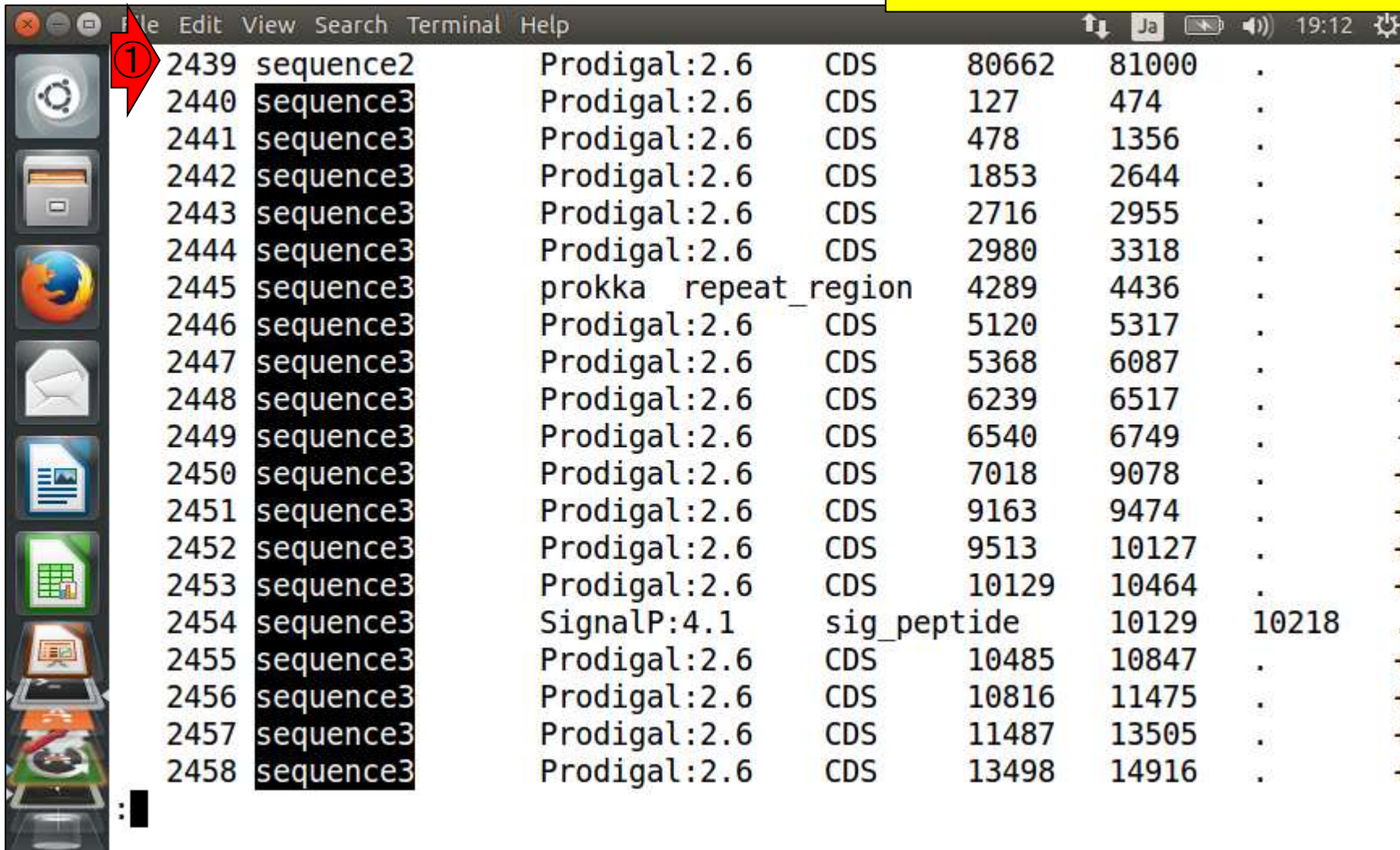
検索時に行頭を表す^をつけなかったが、この場合はおそらくsequence3からはじまるものばかりなのだろう



```
File Edit View Search Terminal Help
① 2440 sequence3 Prodigal:2.6 CDS 127 474 . +
2441 sequence3 Prodigal:2.6 CDS 478 1356 . +
2442 sequence3 Prodigal:2.6 CDS 1853 2644 . +
2443 sequence3 Prodigal:2.6 CDS 2716 2955 . +
2444 sequence3 Prodigal:2.6 CDS 2980 3318 . +
2445 sequence3 prokka repeat_region 4289 4436 . +
2446 sequence3 Prodigal:2.6 CDS 5120 5317 . +
2447 sequence3 Prodigal:2.6 CDS 5368 6087 . +
2448 sequence3 Prodigal:2.6 CDS 6239 6517 . -
2449 sequence3 Prodigal:2.6 CDS 6540 6749 . -
2450 sequence3 Prodigal:2.6 CDS 7018 9078 . +
2451 sequence3 Prodigal:2.6 CDS 9163 9474 . +
2452 sequence3 Prodigal:2.6 CDS 9513 10127 . +
2453 sequence3 Prodigal:2.6 CDS 10129 10464 . +
2454 sequence3 SignalP:4.1 sig_peptide 10129 10218 .
2455 sequence3 Prodigal:2.6 CDS 10485 10847 . +
2456 sequence3 Prodigal:2.6 CDS 10816 11475 . +
2457 sequence3 Prodigal:2.6 CDS 11487 13505 . +
2458 sequence3 Prodigal:2.6 CDS 13498 14916 . +
2459 sequence3 Prodigal:2.6 CDS 14917 16071 . +
```


W19-2: less 復習

キーボードの上矢印キーを1回押したところ。
①1つ上の2,439行目がsequence2の情報であることを念のため確認しているだけ。qで抜ける



```
File Edit View Search Terminal Help
① 2439 sequence2      Prodigal:2.6   CDS      80662    81000    .        +
    2440 sequence3      Prodigal:2.6   CDS       127      474     .        +
    2441 sequence3      Prodigal:2.6   CDS       478     1356    .        +
    2442 sequence3      Prodigal:2.6   CDS      1853     2644    .        +
    2443 sequence3      Prodigal:2.6   CDS      2716     2955    .        +
    2444 sequence3      Prodigal:2.6   CDS      2980     3318    .        +
    2445 sequence3      prokka repeat_region 4289     4436    .        +
    2446 sequence3      Prodigal:2.6   CDS      5120     5317    .        +
    2447 sequence3      Prodigal:2.6   CDS      5368     6087    .        +
    2448 sequence3      Prodigal:2.6   CDS      6239     6517    .        -
    2449 sequence3      Prodigal:2.6   CDS      6540     6749    .        -
    2450 sequence3      Prodigal:2.6   CDS      7018     9078    .        +
    2451 sequence3      Prodigal:2.6   CDS      9163     9474    .        +
    2452 sequence3      Prodigal:2.6   CDS      9513    10127    .        +
    2453 sequence3      Prodigal:2.6   CDS     10129    10464    .        +
    2454 sequence3      SignalP:4.1    sig_peptide 10129    10218    .
    2455 sequence3      Prodigal:2.6   CDS     10485    10847    .        +
    2456 sequence3      Prodigal:2.6   CDS     10816    11475    .        +
    2457 sequence3      Prodigal:2.6   CDS     11487    13505    .        +
    2458 sequence3      Prodigal:2.6   CDS     13498    14916    .        +
:
```

W19-3: ヘッダ一部分

①grepの結果からもわかるように、“^##”で annotation.gff中のヘッダ行情報を取り出せる。従って、このヘッダ行とgrep “^sequence1”の行を合わせることで、sequence1のみの情報からなるgffファイルを得ることができる。つまり…

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls *.gff
annotation.gff
iu@bielinux[mac_share] wc annotation.gff
2492 38582 511967 annotation.gff
iu@bielinux[mac_share] head -n 3 annotation.gff
##gff-version 3
sequence1      Prodigal:2.6   CDS    101    1417    .      +      0
ID=LOCUS_00001;product=chromosomal replication initiator protein DnaA;inf
erence=ab initio prediction:Prodigal:2.6,similar to AA sequence:INSD:BAP8
4581.1;locus_tag=LOCUS_00001;gene=dnaA
sequence1      Prodigal:2.6   CDS    1593   2732    .      +      0
ID=LOCUS_00002;product=DNA polymerase III subunit beta;inference=ab initi
o prediction:Prodigal:2.6,similar to AA sequence:INSD:BAP84582.1;locus_ta
g=LOCUS_00002;gene=dnaN
iu@bielinux[mac_share] less -N -S annotation.gff
iu@bielinux[mac_share] grep -n "^##" annotation.gff
1:##gff-version 3
iu@bielinux[mac_share]
```



W19-4: sequence1のgff

①最初のgrep “^##”で、gffのヘッダ一行を取り出してannotation_seq1.gffに保存。次にgrep “^sequence1”を満たす行をannotation_seq1.gffに追加保存

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep "^##" annotation.gff > annotation_seq1.gff
iu@bielinux[mac_share] grep "^sequence1" annotation.gff >> annotation_seq1.gff
iu@bielinux[mac_share]
```

[11:13午前]

[11:13午前]



①

W19-4: sequence1のgff

①lsで確認。annotation_seq1.gffの中身はsequence1の情報がほとんどを占めるので、ファイルサイズの減少度合い(511,967 → 481,561 bytes)的にも妥当

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] grep "^##" annotation.gff > annotation_seq1.gff
iu@bielinux[mac_share] grep "^sequence1" annotation.gff >> annotation_seq1.gff
iu@bielinux[mac_share] ls -l annotation*.gff
-rwxrwxrwx 1 iu iu 511967 1月 21 12:39 annotation.gff
-rwxrwxrwx 1 iu iu 481561 2月 1 11:13 annotation_seq1.gff
iu@bielinux[mac_share]
```

[11:13午前]

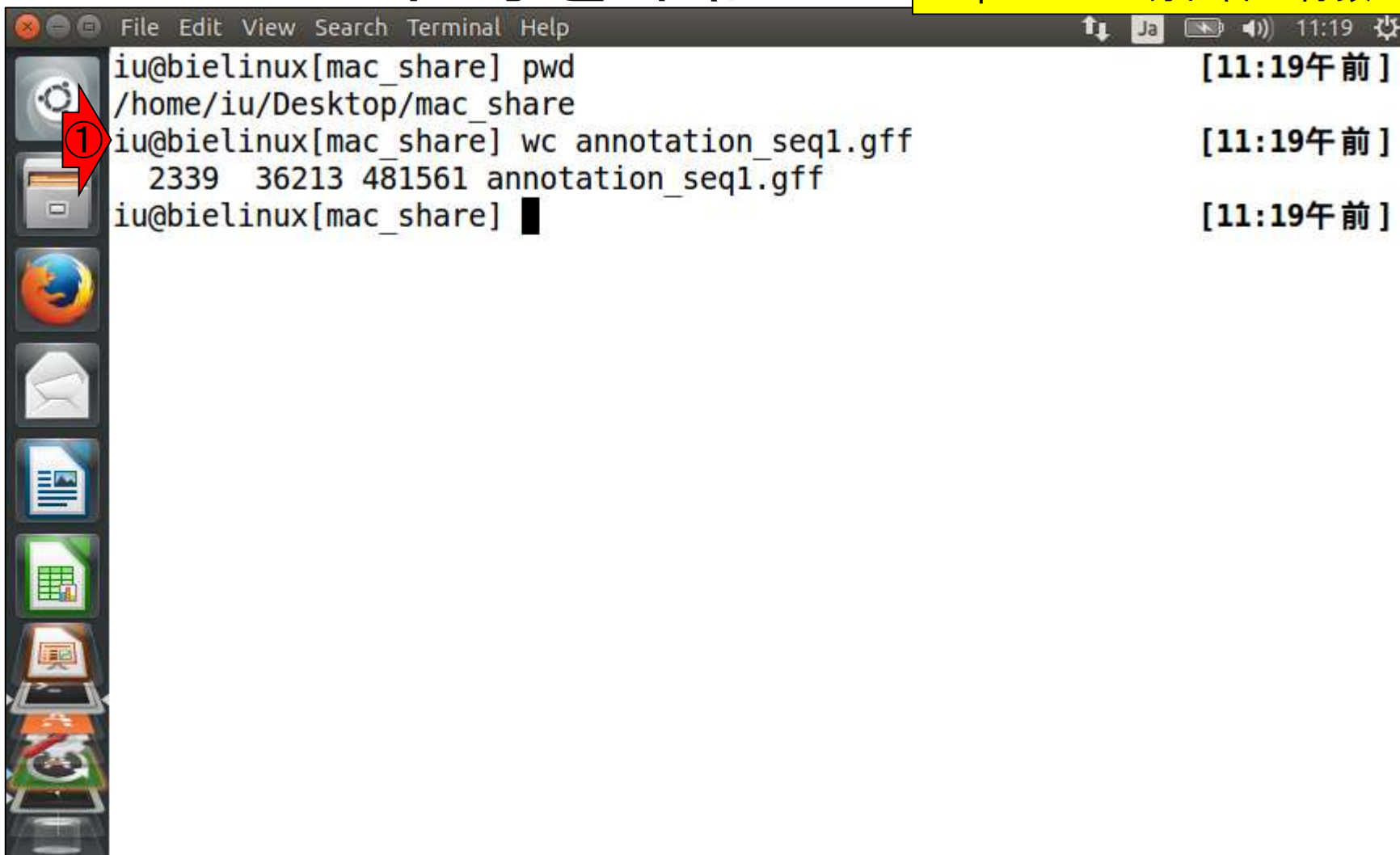
[11:13午前]

[11:14午前]



W19-5: 中身を確認

① annotation_seq1.gffは2,339行。lessで annotation.gffを眺めた結果のsequence1と sequence2の切れ目の行数と同じであり妥当



The image shows a terminal window with a dark background and a light-colored font. The window title is "Terminal". The user is logged in as "iu" on a machine named "bielinux" in the directory "/home/iu/Desktop/mac_share". The terminal shows the following commands and their outputs:

```
iu@bielinux[mac_share] pwd [11:19午前]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation_seq1.gff [11:19午前]
 2339  36213 481561 annotation_seq1.gff
iu@bielinux[mac_share] █ [11:19午前]
```

A red arrow with the number "1" points to the second command line.

W19-5: 中身を確認

①headと②tailでも確認。OKですね。ここまでで sequence1 のみの情報からなるgffファイルを作成できました。そして今やろうとしていることは、配列情報付きのgffファイルを作成してDNAPlotterで読み込めることを確認しようとしています。また、その解説を通じてLinuxコマンドのスキルアップを目指しています

```
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] wc annotation_seq1.gff
 2339  36213 481561 annotation_seq1.gff
iu@bielinux[mac_share] head -n 2 annotation_seq1.gff [11:19午前]
##gff-version 3
sequence1      Prodigal:2.6    CDS           101    1417    .      +      0
ID=LOCUS_00001;product=chromosomal replication initiator protein DnaA;inference=ab initio prediction:Prodigal:2.6,similar to AA sequence:INSD:BAP84581.1;locus_tag=LOCUS_00001;gene=dnaA
iu@bielinux[mac_share] tail -n 2 annotation_seq1.gff [11:22午前]
sequence1      Prodigal:2.6    CDS           2277458 2277592 .      -      0
ID=LOCUS_02252;product=50S ribosomal protein L34;inference=ab initio prediction:Prodigal:2.6,similar to AA sequence:INSD:AD000222.1;locus_tag=LOCUS_02252;gene=rpmH
sequence1      SignalP:4.1     sig_peptide   2277497 2277592 .      -
. ID=LOCUS_02252;note=predicted cleavage at residue 32;locus_tag=LOCUS_02252;gene=rpmH;inference=ab initio prediction:SignalP:4.1
iu@bielinux[mac_share] [11:22午前]
```

W19-6: 配列を結合

おさらい。annotation.gffは、①全部で6行からなるLH_complete.faを入力としてDFASTを実行した結果ファイルでした。そして最初の2行分がsequence1のFASTAファイルに相当します。やりたいのは、この部分を抽出してannotation_seq1.gffに追加保存すること

```
File Edit View Search Terminal Help
iu@bielinux[mac_share] pwd
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_complete.fa
-rwxrwxrwx 1 iu iu 2400619 12月 19 19:33 LH_complete.fa
iu@bielinux[mac_share] wc LH_complete.fa
6          6 2400619 LH_complete.fa
iu@bielinux[mac_share] █
```

[11:27午前]
[11:27午前]
[11:27午前]



W19-6: 配列を結合

①LH_complete.faの最初の2行分を抽出して annotation_seq1.gffに追加保存。②その後のwcコマンドで追加保存前の2,339行から2,341行と2行分だけ増えていることからうまくいっていると判断

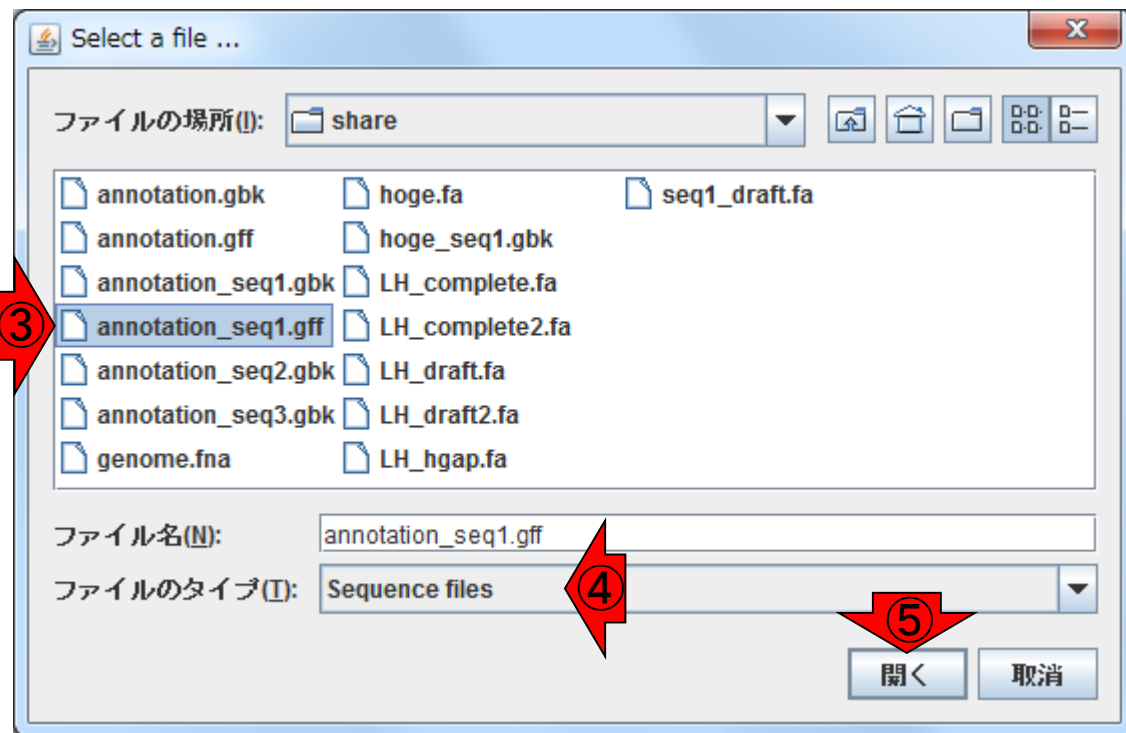
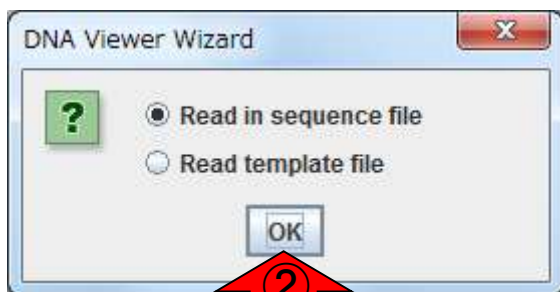
```
iu@bielinux[mac_share] pwd [11:27午前]
/home/iu/Desktop/mac_share
iu@bielinux[mac_share] ls -l LH_complete.fa [11:27午前]
-rwxrwxrwx 1 iu iu 2400619 12月 19 19:33 LH_complete.fa
iu@bielinux[mac_share] wc LH_complete.fa [11:27午前]
 6      6 2400619 LH_complete.fa
① iu@bielinux[mac_share] head -n 2 LH_complete.fa >> annotation_seq1.gff
iu@bielinux[mac_share] wc annotation_seq1.gff [11:34午前]
 2341   36215 2759557 annotation_seq1.gff
② iu@bielinux[mac_share] █ [11:34午前]
```


W19-7: DNAPlotter

①②DNAPlotterを起動して、作成した③
annotation_seq1.gffを、④ファイルのタイプ
はSequence filesのままでよいので、⑤開く

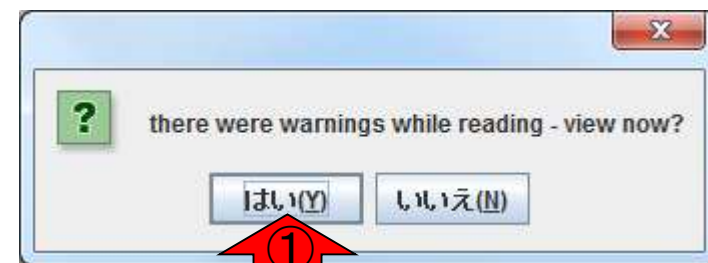
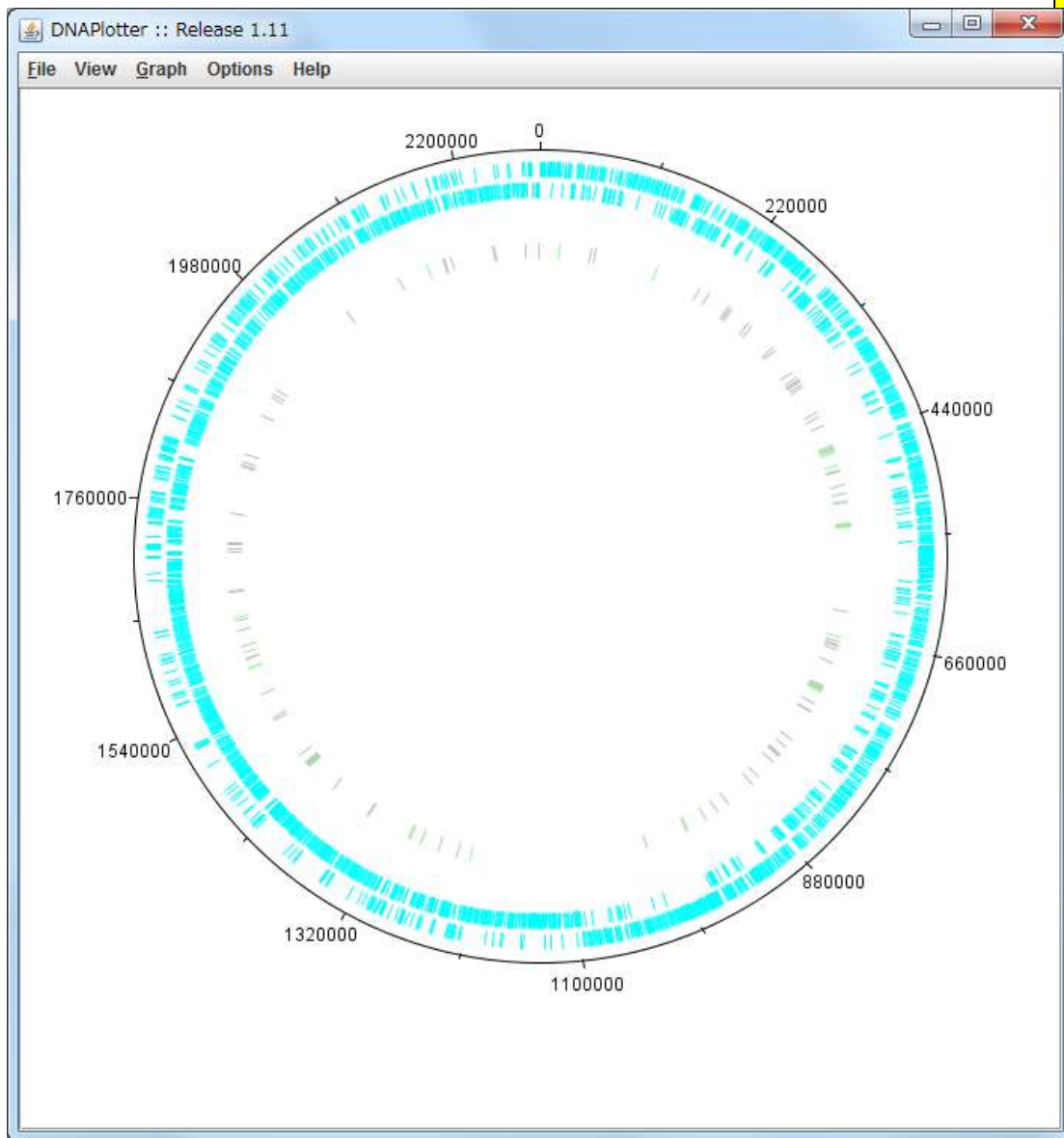


dnaplotter.jar



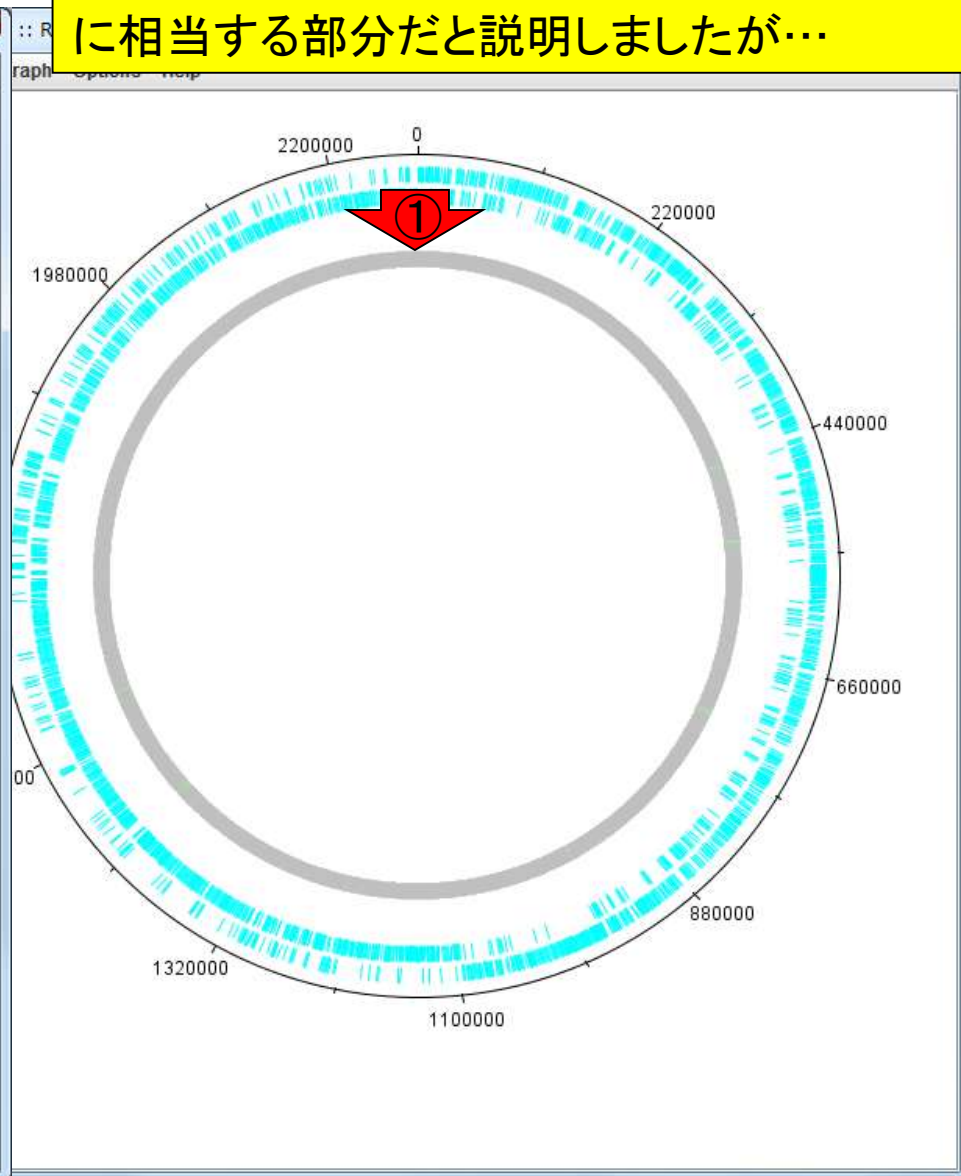
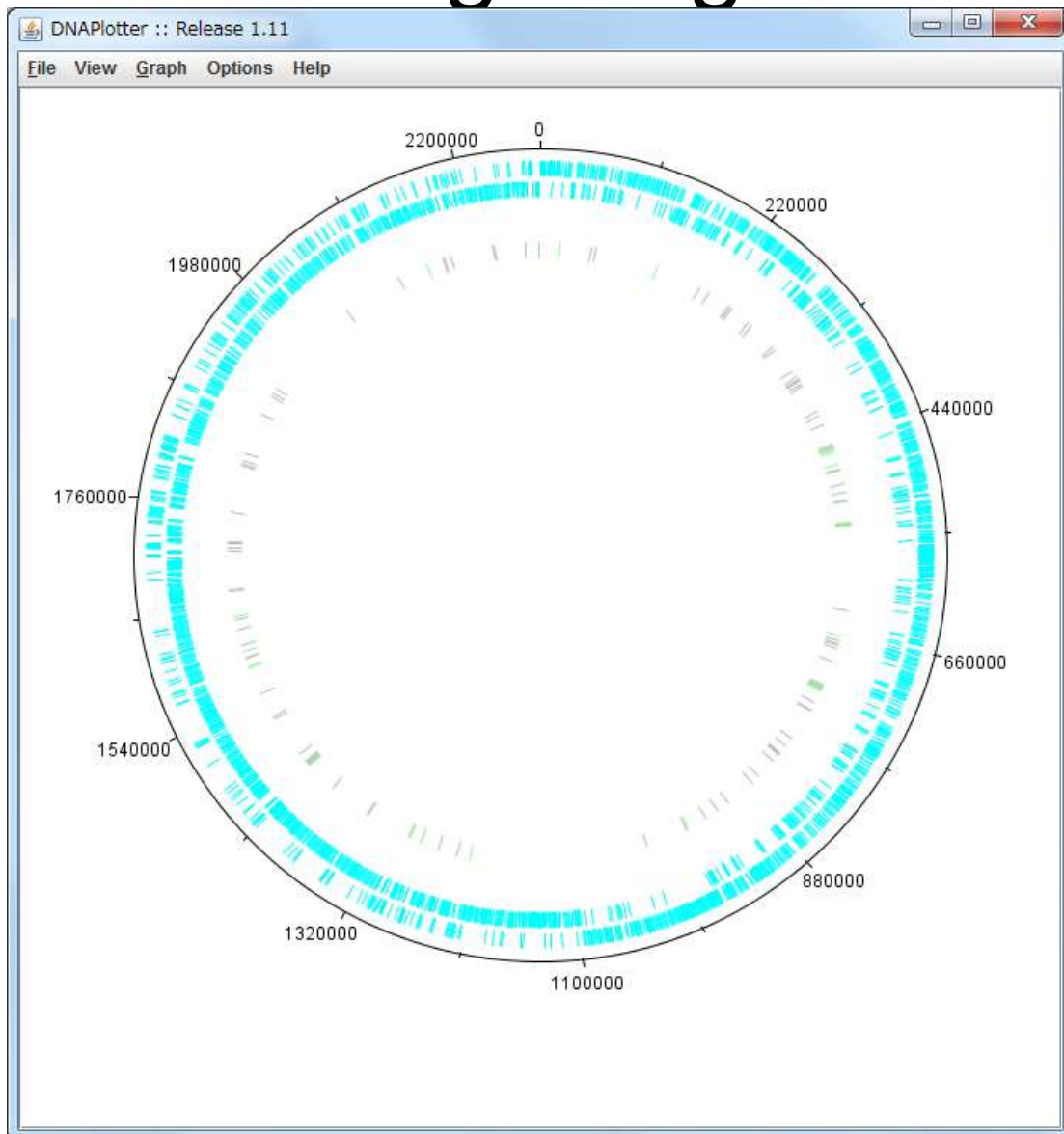
W19-7: DNAPlotter

①なんか多少警告は出るが気にしない。gbkファイルを読み込んで得られたプロット(W12-3)と若干の違いはありますが、配列情報付きのgffで無事読み込めたことがわかります。



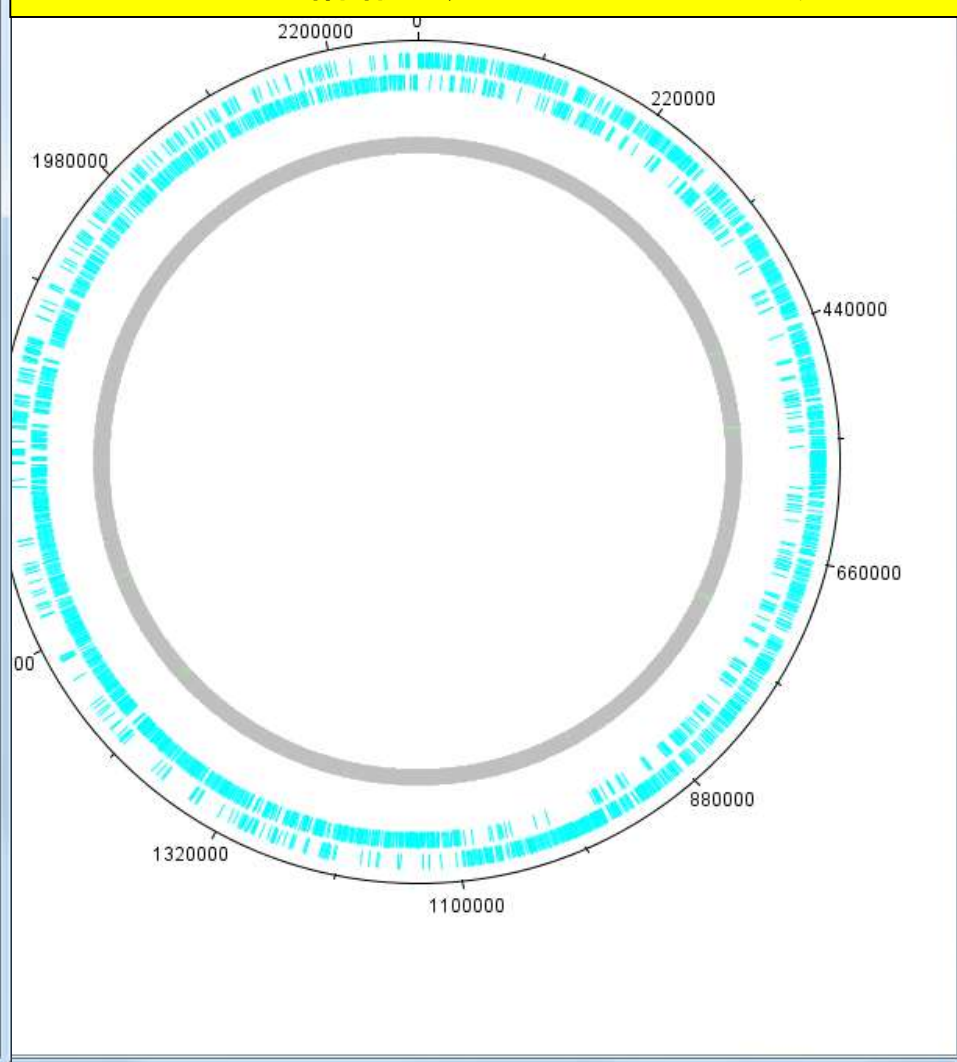
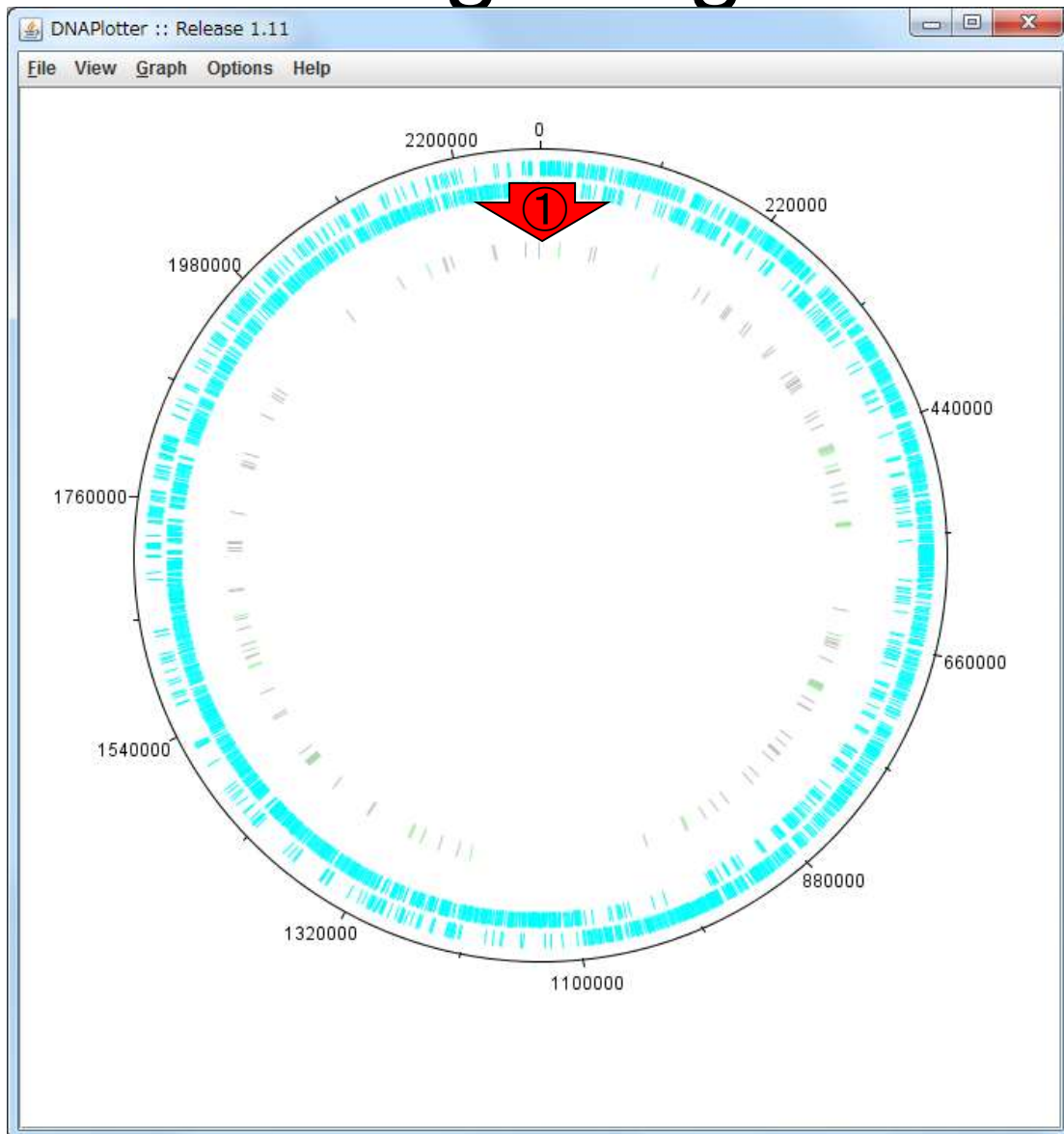
W19-8: gbkとgffの違い

gbkファイルを読み込んで得られたプロット(右)と比較すると、①の部分に違いがあることがわかります。これは、W12-10ではtRNAに相当する部分だと説明しましたが…



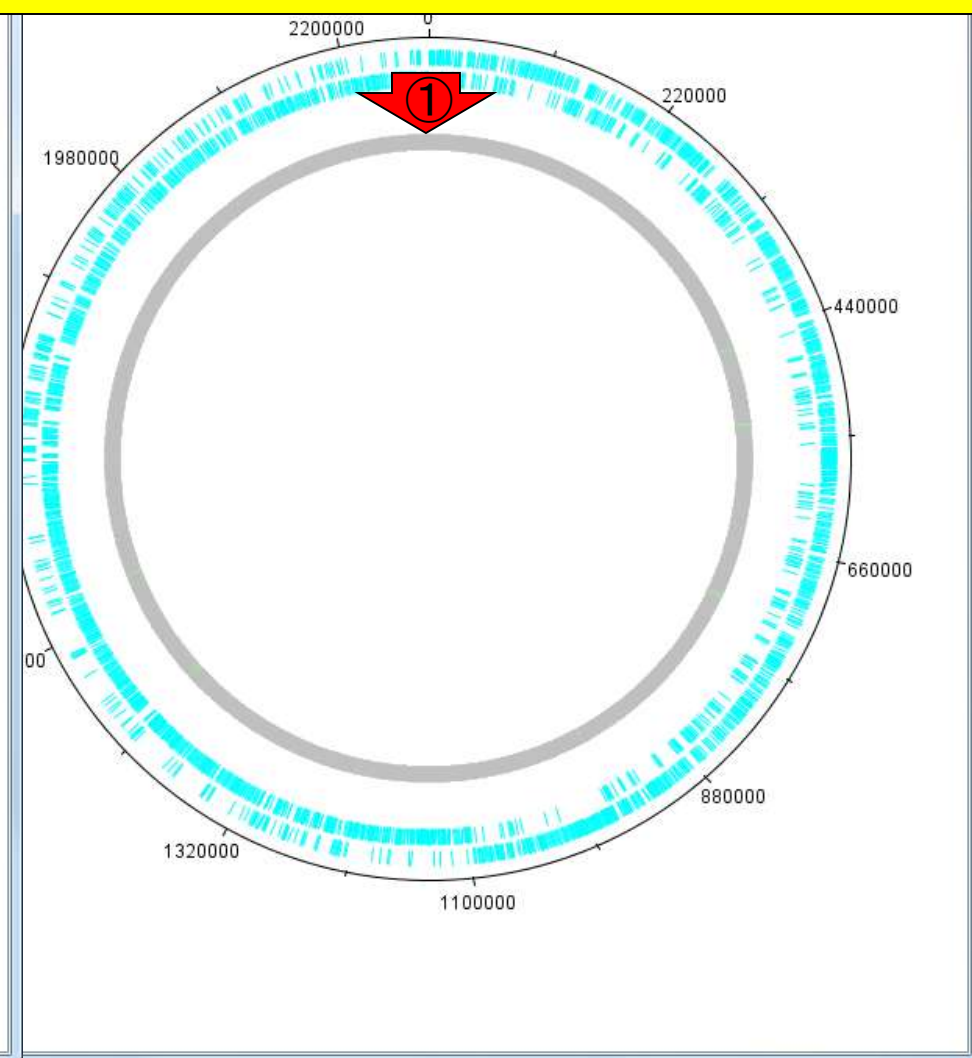
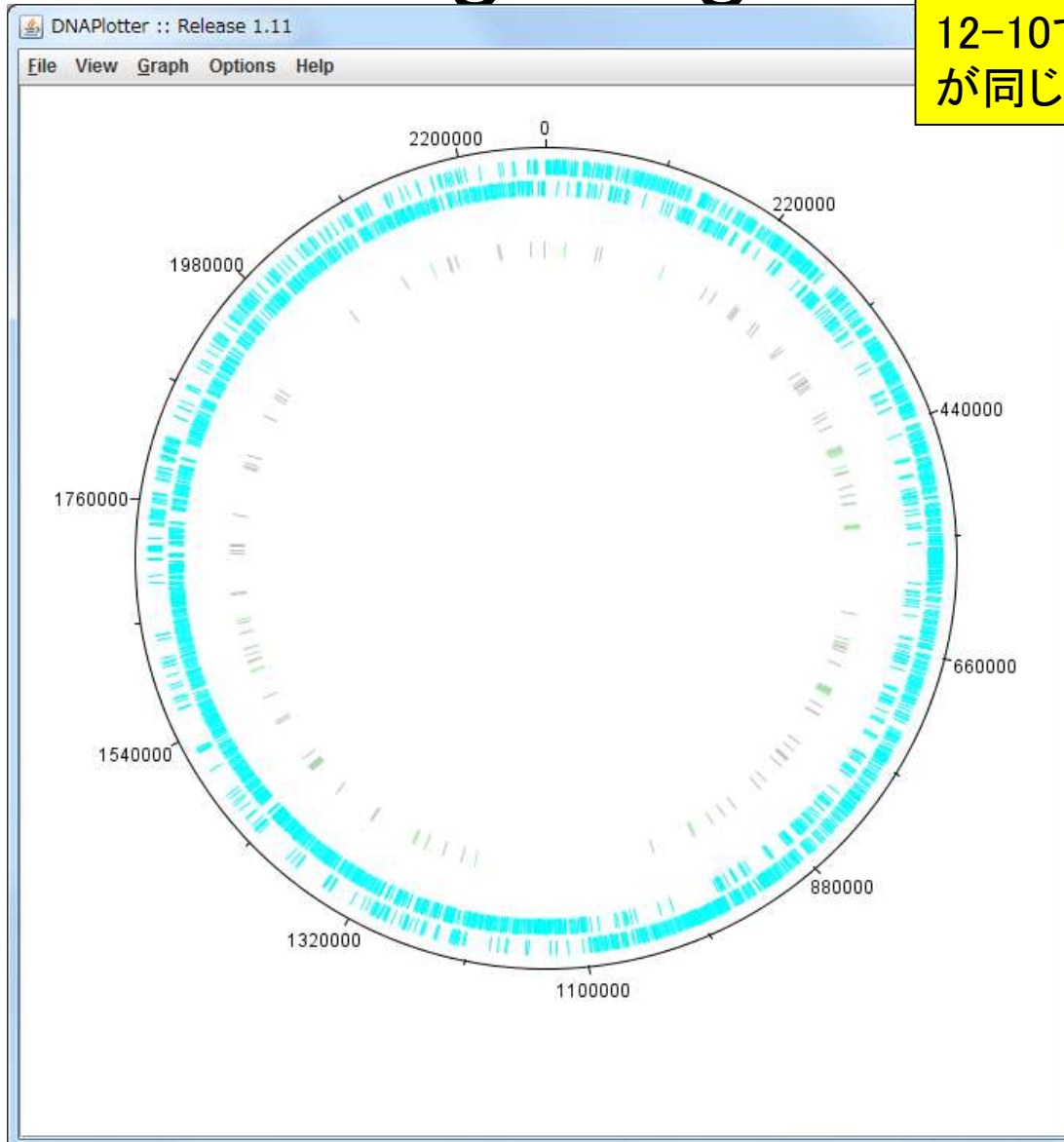
W19-8: gbkとgffの違い

配列情報付きのgffファイルを読み込んで得られたプロット(左)を眺めると、①rRNAが灰色、そしてtRNAが緑色で表示されているようです。つまり、①のエリアは、本当はtRNAだけではなくrRNAの情報も表示していたということ



W19-8: gbkとgffの

確かにW12-10で見られるようにDNAPlotterのウェブページ上でも、tRNAとrRNAが一体であるような印象を受けます。①が灰色ばかりになっていたのは、W12-9や12-10でも大まかにわかるように、gene featureとrRNAが同じ灰色で表現されていただけというオチ



W19-9: gene feature

```
File Edit View Search Terminal Help 12:34
1 LOCUS      sequence1      2277983 bp      DNA      B
1 CT 20-JAN-2017
2 DEFINITION Lactobacillus sp. strain unkown.
3 ACCESSION  sequence1
4 VERSION    sequence1
5 KEYWORDS   .
6 SOURCE     Lactobacillus sp
7 ORGANISM   Lactobacillus sp. Unclassified. .
8
9 COMMENT    Annotated using prokka 1.12-beta from
10           https://github.com/tseemann/prokka.
11 FEATURES   Location/Qualifiers
12     source  1..2277983
13           /strain="unkown"
14           /mol_type="genomic DNA"
15           /organism="Lactobacillus sp."
16     gene    101..1417
17           /locus_tag="LOCUS_00001"
18           /gene="dnaA"
19     CDS     101..1417
annotation_seq1.gbк
```



W12-10で見たrRNAの1つであるLOOC260_00331の情報を眺めるべく、00331で検索したところ。

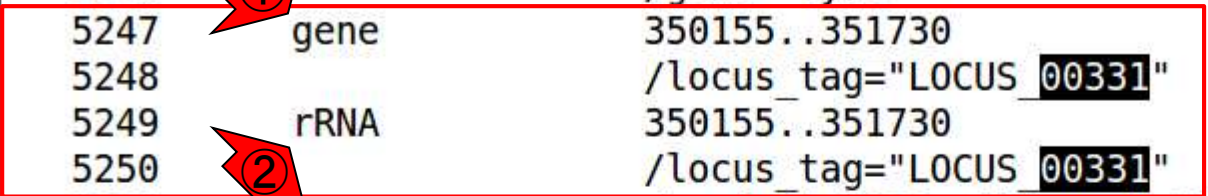
W19-9: gene feature

```
File Edit View Search Terminal Help 12:36
5245 KFVIVRRGKKKYFLVRINK"
5246 /gene="tyrS"
5247 gene 350155..351730
5248 /locus_tag="LOCUS_00331"
5249 rRNA 350155..351730
5250 /locus_tag="LOCUS_00331"
5251 /product="16S ribosomal RNA"
5252 gene 351956..354877
5253 /locus_tag="LOCUS_00332"
5254 rRNA 351956..354877
5255 /locus_tag="LOCUS_00332"
5256 /product="23S ribosomal RNA"
5257 gene 354977..355089
5258 /locus_tag="LOCUS_00333"
5259 rRNA 354977..355089
5260 /locus_tag="LOCUS_00333"
5261 /product="5S ribosomal RNA"
5262 gene 355287..356807
5263 /locus_tag="LOCUS_00334"
5264 /gene="opuCC"
```

gbkファイルを入力としてDNAPlotterを開いたときには、おそらく①のgene featureと②rRNA featureが同じ灰色として表現されていたために…

W19-9: gene feature

```
File Edit View Search Terminal Help 12:36
5245 KFVIVRRGKKKYFLVRINK"
5246 /gene="tyrS"
5247 gene 350155..351730
5248 /locus_tag="LOCUS_00331"
5249 rRNA 350155..351730
5250 /locus_tag="LOCUS_00331"
5251 /product="16S ribosomal RNA"
5252 gene 351956..354877
5253 /locus_tag="LOCUS_00332"
5254 rRNA 351956..354877
5255 /locus_tag="LOCUS_00332"
5256 /product="23S ribosomal RNA"
5257 gene 354977..355089
5258 /locus_tag="LOCUS_00333"
5259 rRNA 354977..355089
5260 /locus_tag="LOCUS_00333"
5261 /product="5S ribosomal RNA"
5262 gene 355287..356807
5263 /locus_tag="LOCUS_00334"
5264 /gene="opuCC"
```



W19-9: gene featu

①のような感じで灰色の輪っかのように見えてしまっていたのだろう。このような違いは、様々なファイル形式を自在に作成して読み込んでみないと分からない。我々はたまたまgbkとgffの違いの原因を特定することができたが、分からないまま数年が過ぎることもよくある

ID	Type	Coordinates	Product
5245		KFVIVRR	
5246		/gene="	
5247	gene	350155..3	
5248		/locus ta	
5249	rRNA	350155..3	
5250		/locus ta	
5251		/product=	
5252	gene	351956..3	
5253		/locus ta	
5254	rRNA	351956..3	
5255		/locus ta	
5256		/product=	
5257	gene	354977..3	
5258		/locus ta	
5259	rRNA	354977..3	
5260		/locus ta	
5261		/product=	
5262	gene	355287..3	
5263		/locus ta	
5264		/gene="op	

