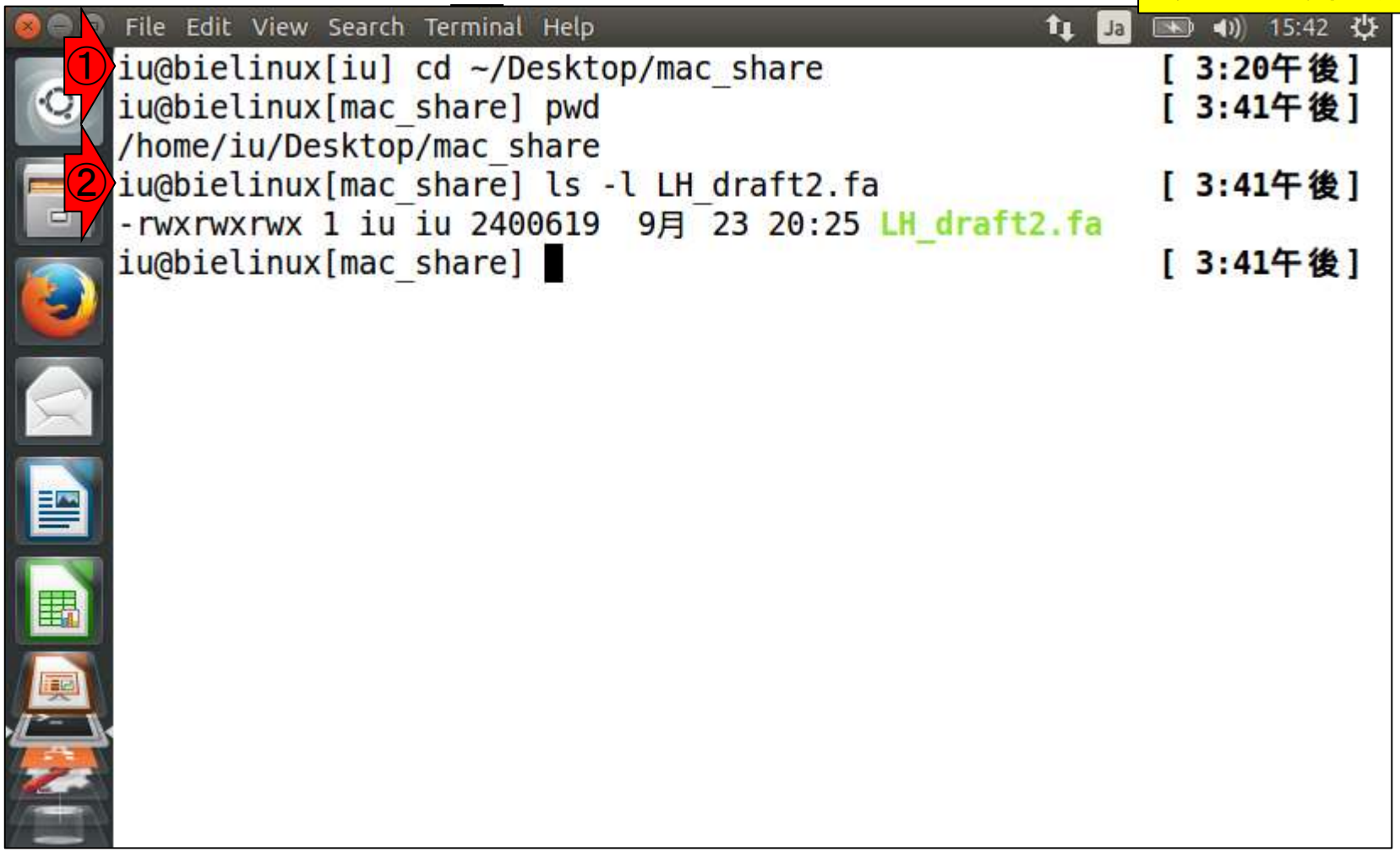
次世代シーケンサーデータの解析手法 第9回ゲノムアノテーションとその可視化、DDBJへの登録 ウェブ資料

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

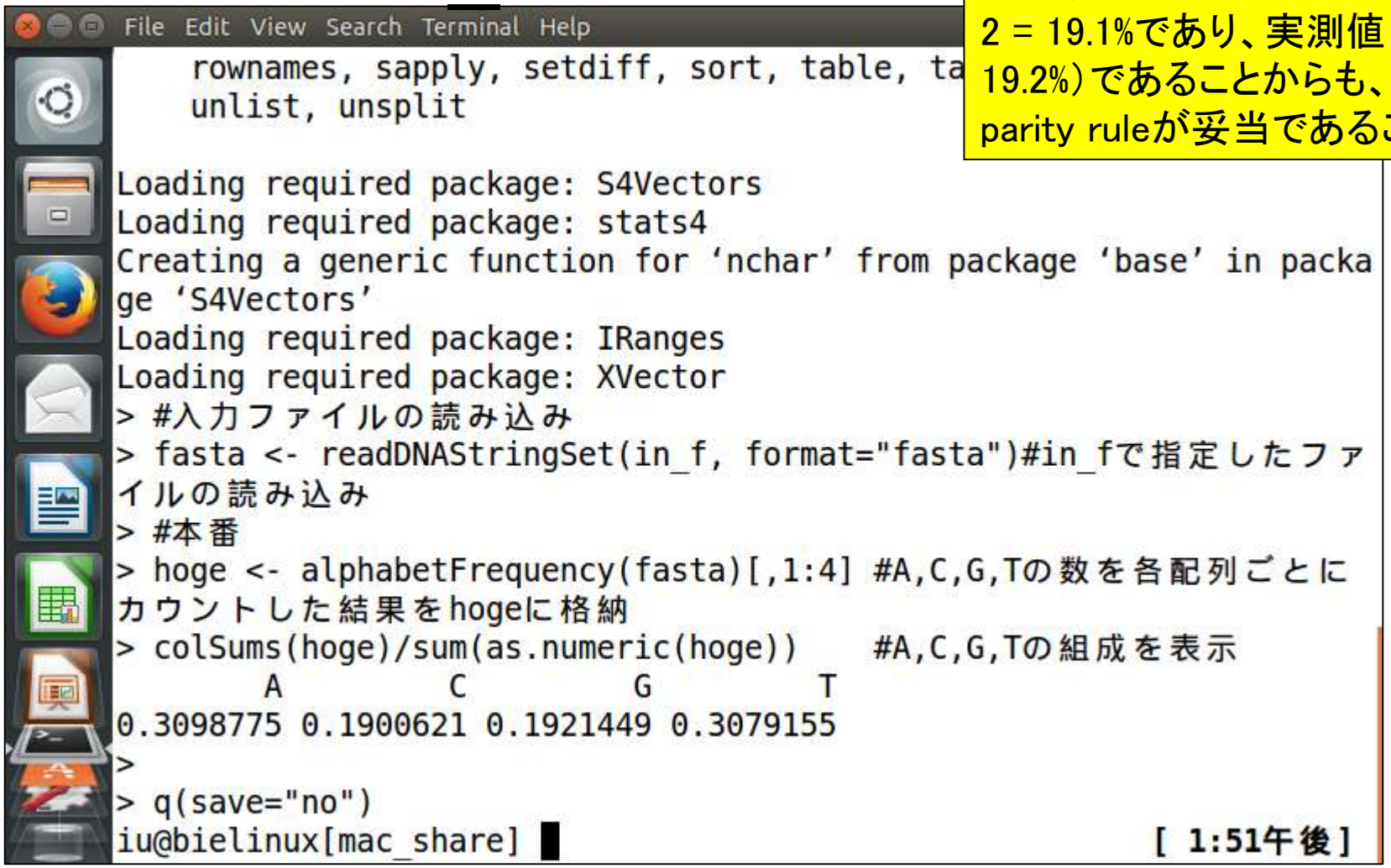
①共有フォルダ上に、②第9回で利用するファイル（LH_draft2.fa）があるという前提で進めていきます



Field et al., *Nat Biotechnol.*, **24**: 801–803, 2006

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%）であることからも、Chargaff's second parity ruleが妥当であることがわかる



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 ③
- Genome Fasta File: genome.fna
- Protein Fasta File: protein.faa
- CDS Fasta File: cds.fna
- RNA Fasta File: rna.fna
- Feature Table: features.tsv ②
- Genome Statistics: statistics.t

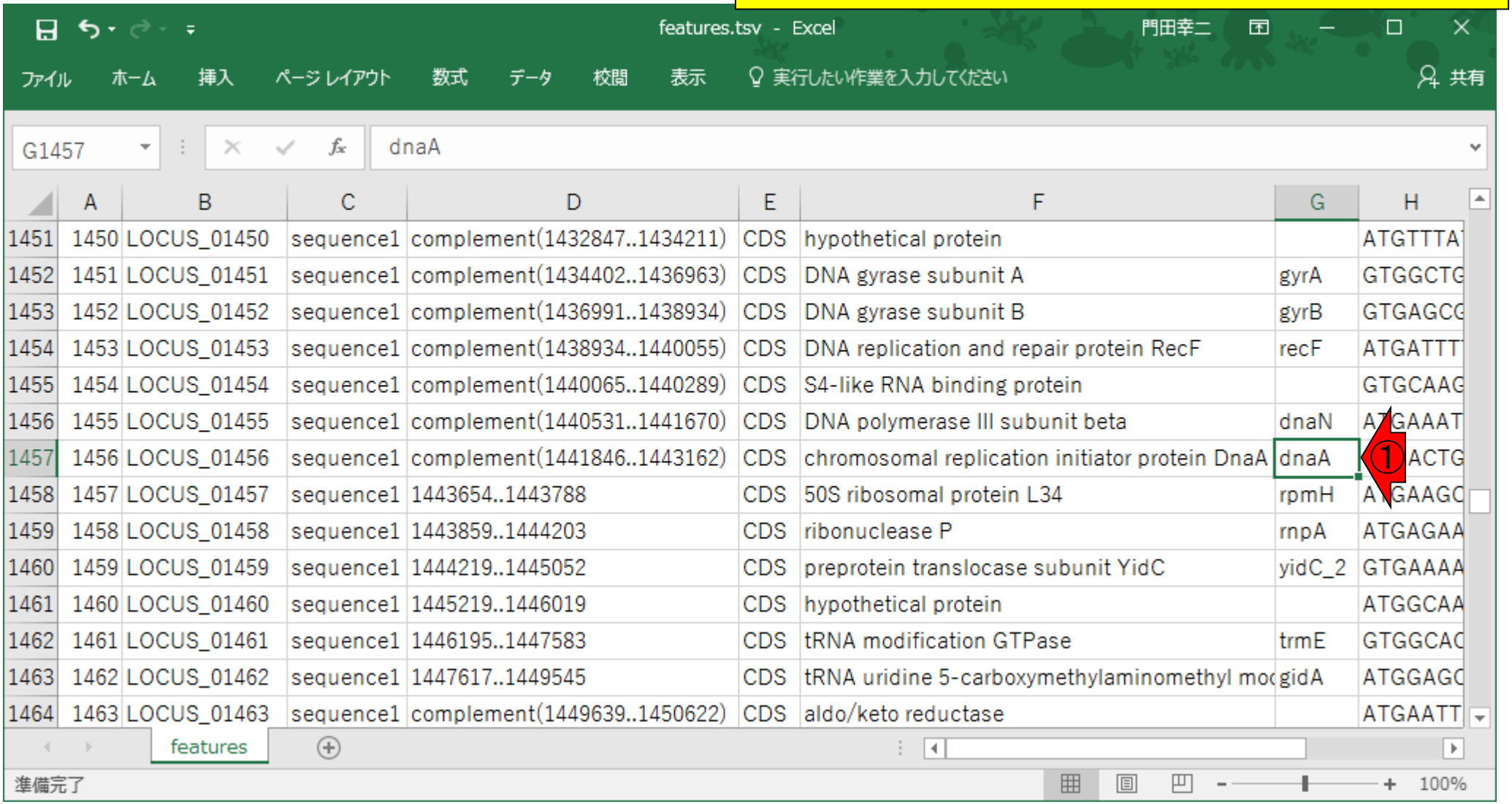
BLASTの実行と可視化

- dotter: Sonnhammer and Durbin, Gene, 1995

日本乳酸菌学会誌の DFAST: Tanizawa et al., *Biosci Microbiota Food Health*, **35**: 173–184, 2016

W2-1：features.tsv

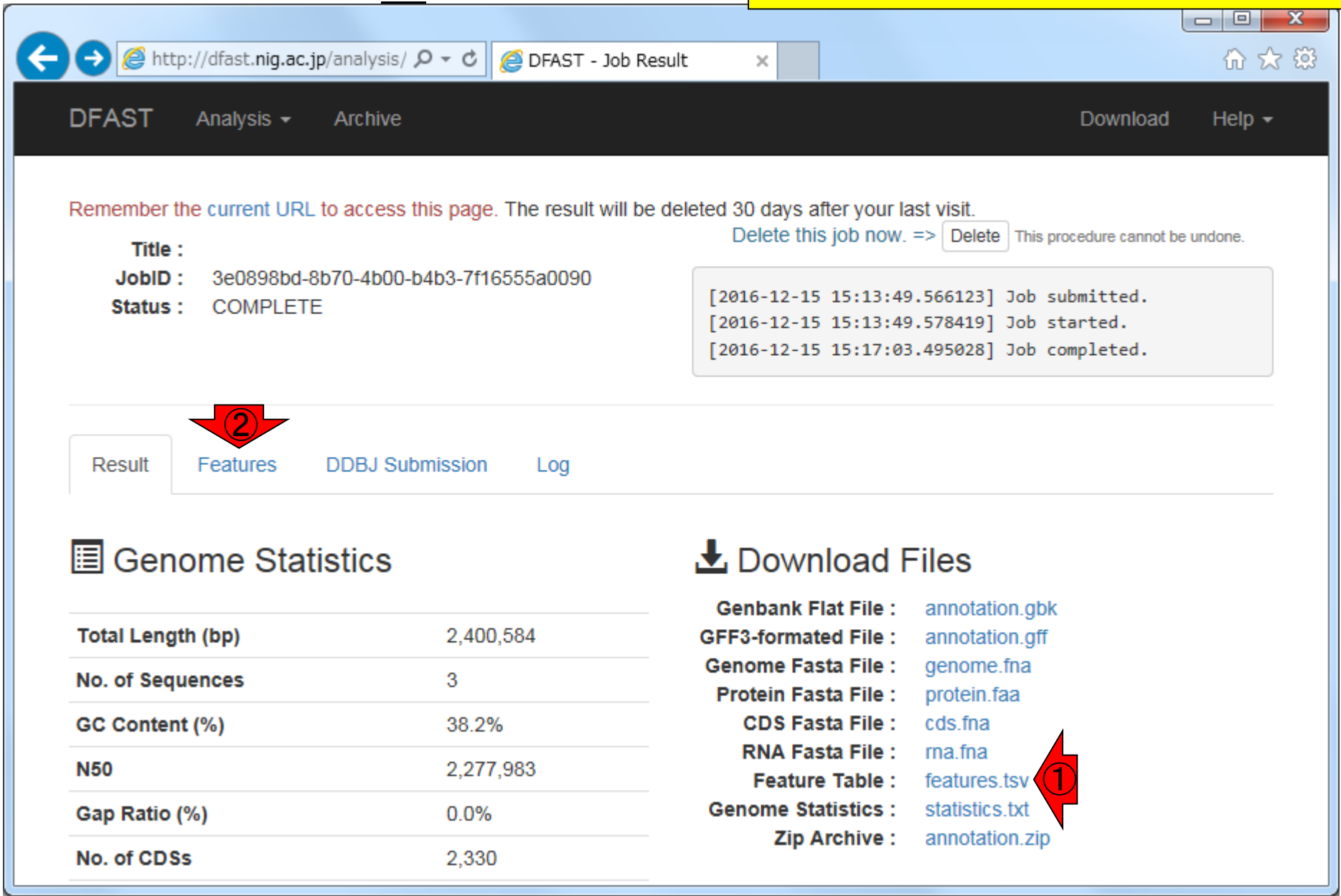
①dnaAは、sequence1の[1441846, 1443162 bp]領域にあった（ことを一応確認）。ただこれは4配列からなるLH_hgap.faを入力として実行した結果



	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	GTGAGCG
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	AGAAGC
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

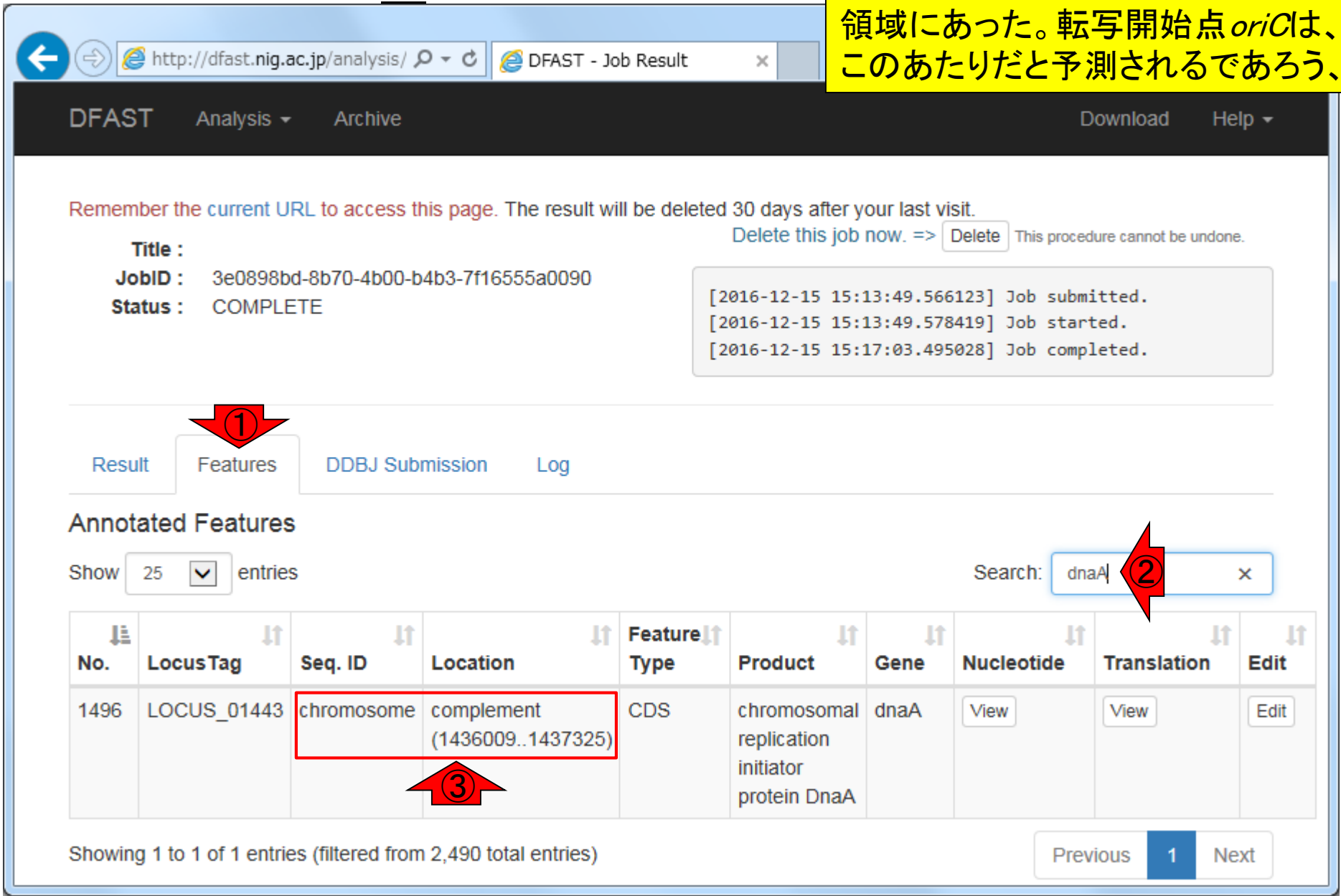
W2-2：LH_draft2.fa

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し、②Featuresタブで調べてもよい



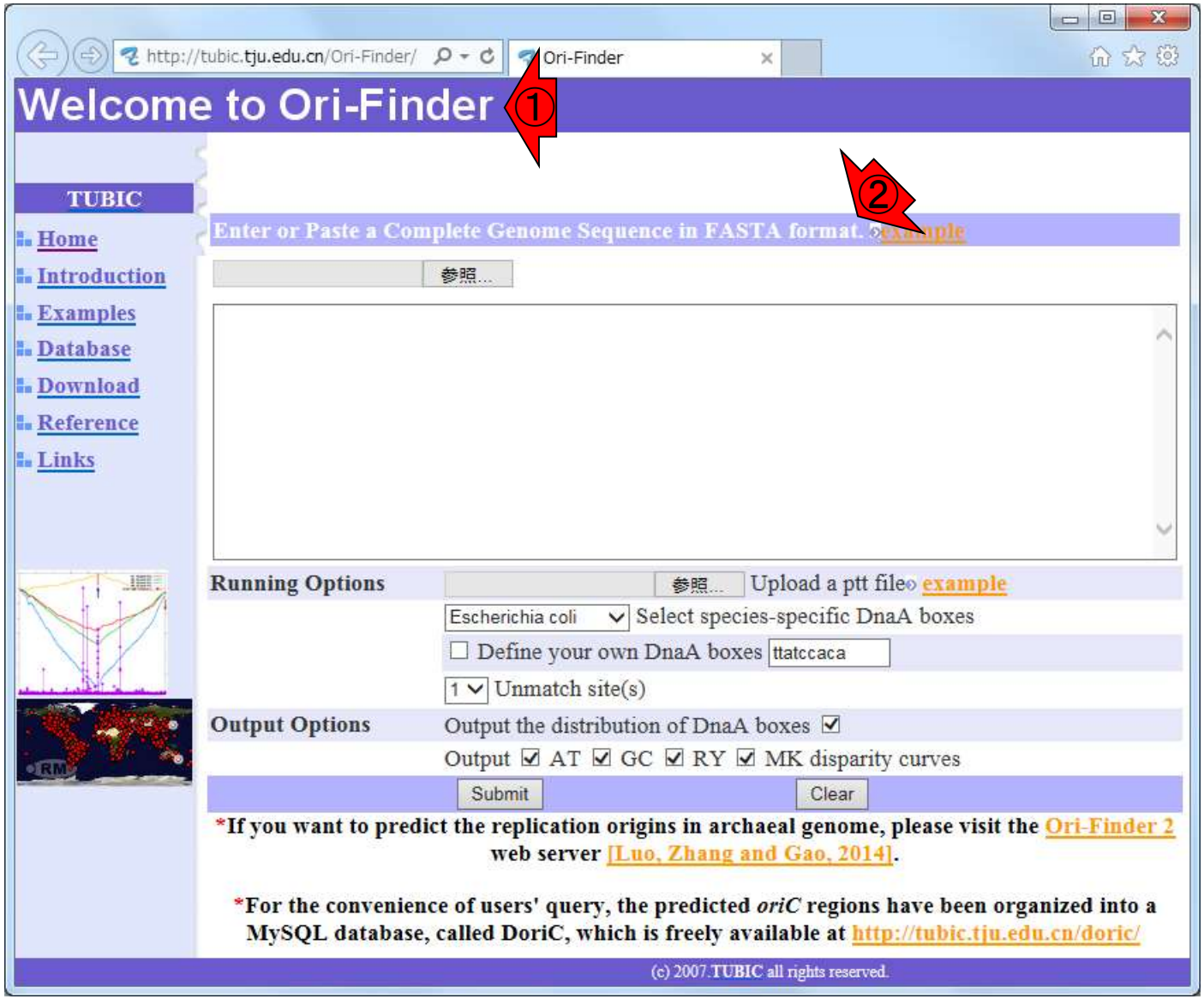
W2-2：LH_draft2.fa

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



W3-1：Ori-Finder

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



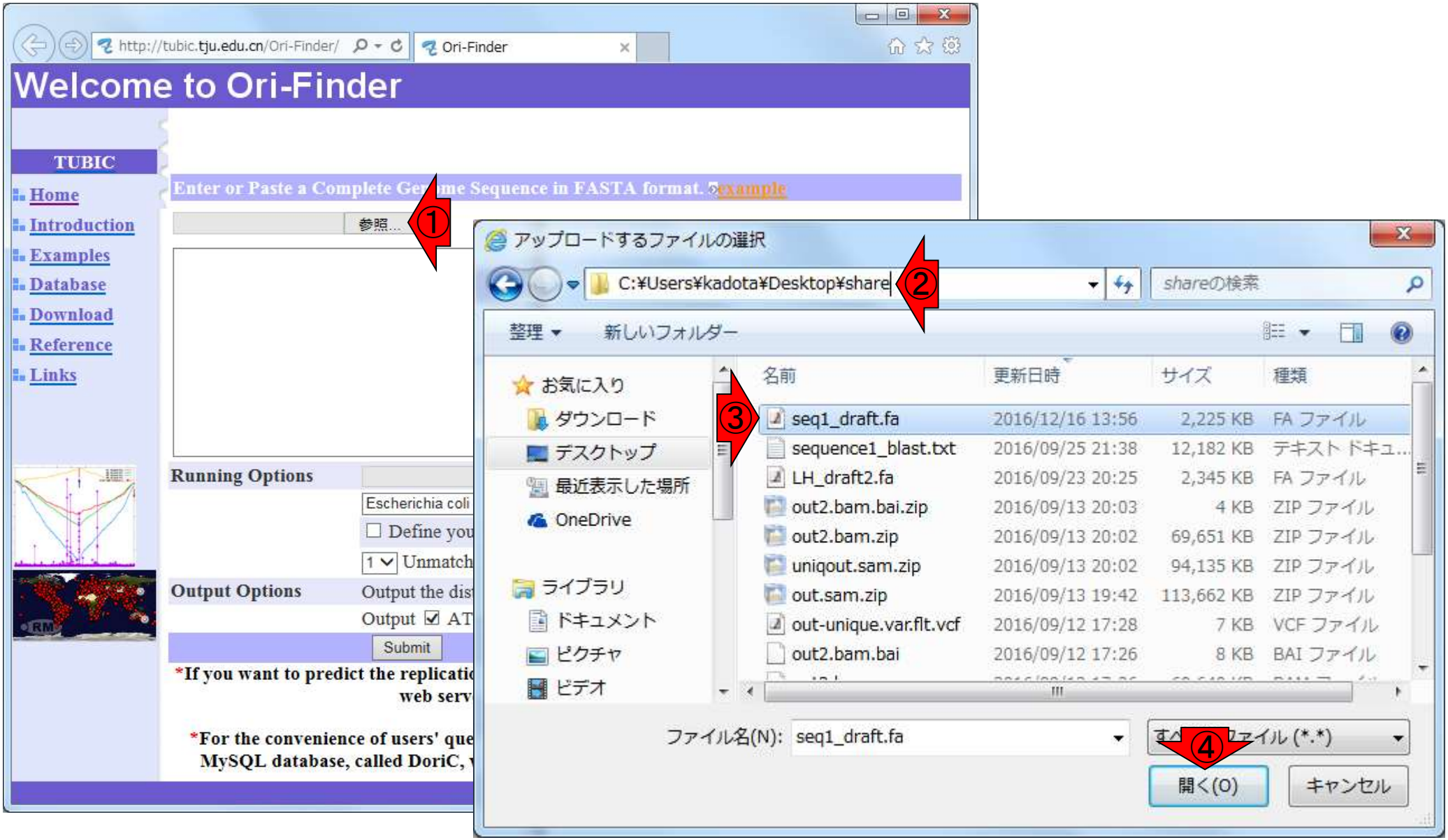
Gao and Zhang, *BMC Bioinformatics*, **9**: 79, 2008

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

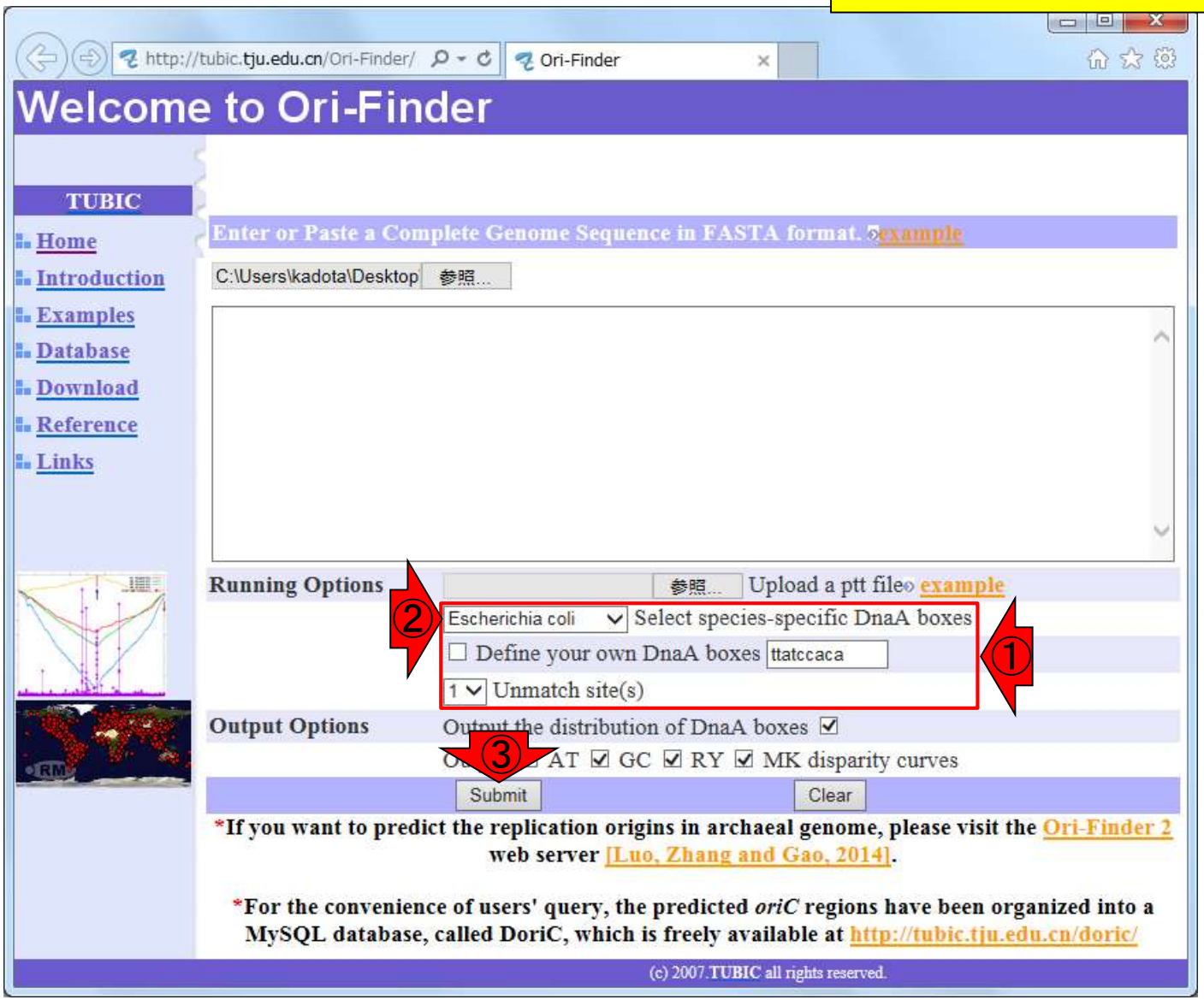
W3-3：Ori-Finder

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



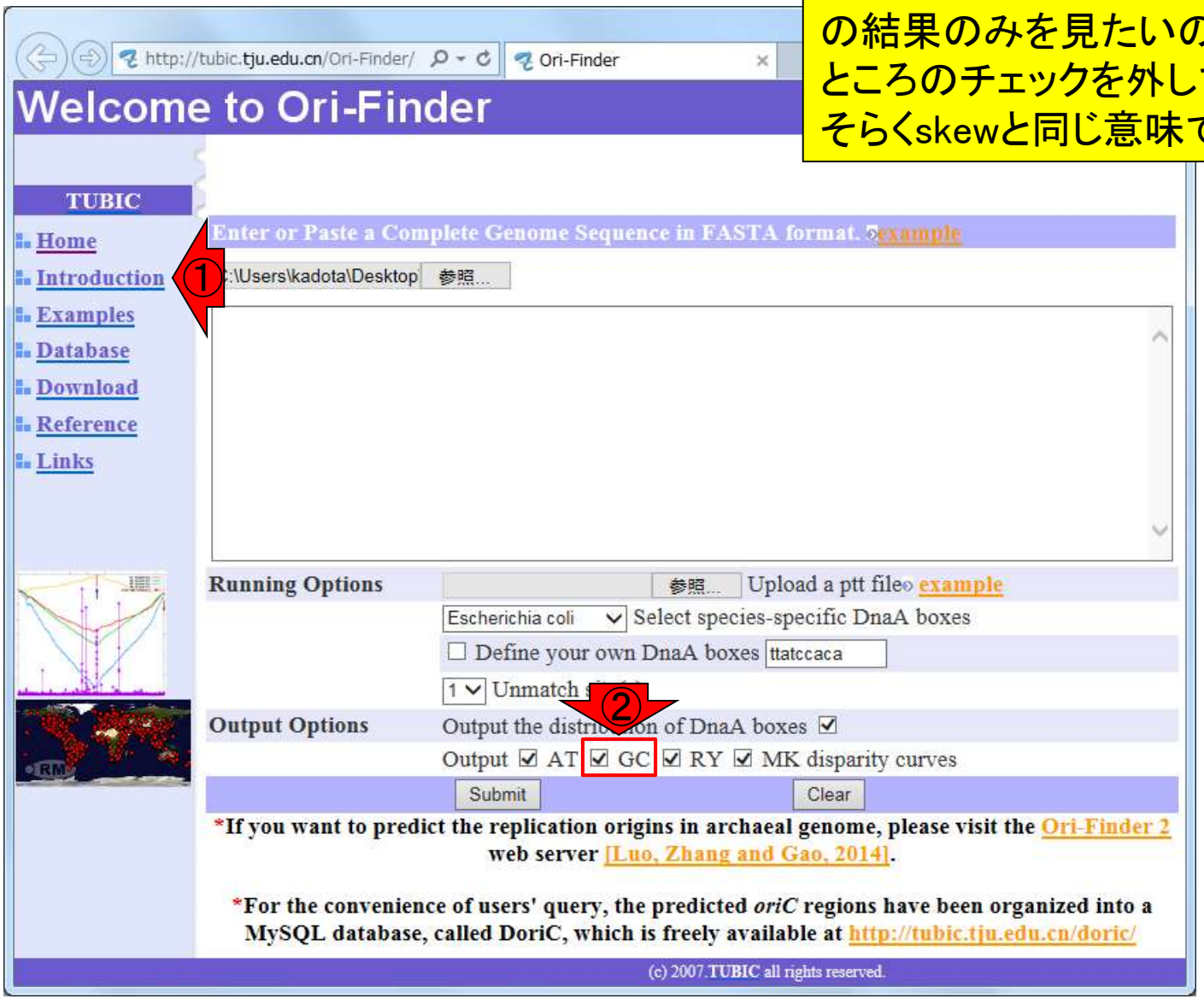
W3-3：Ori-Finder

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



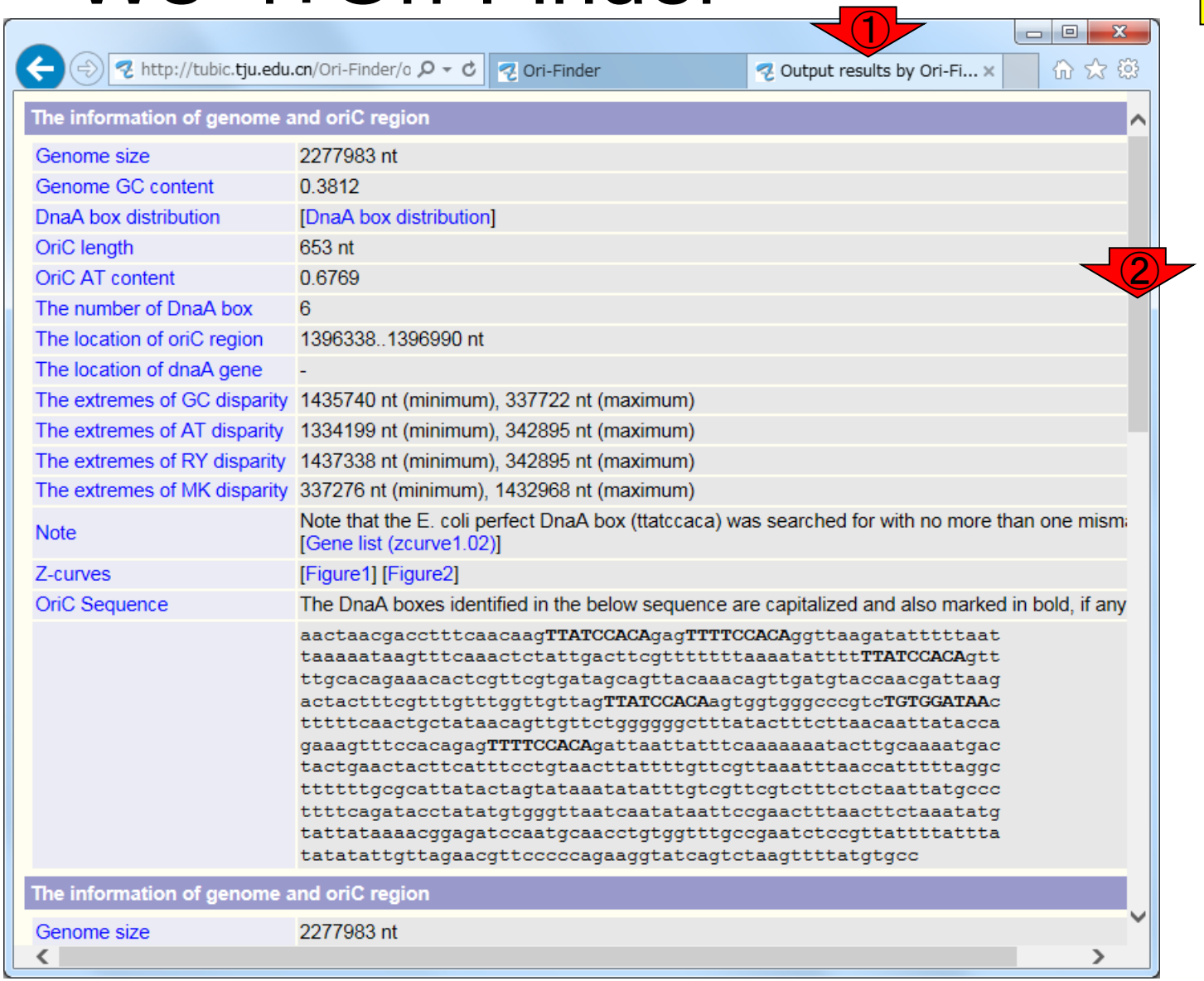
W3-3：Ori-Finder

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



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…



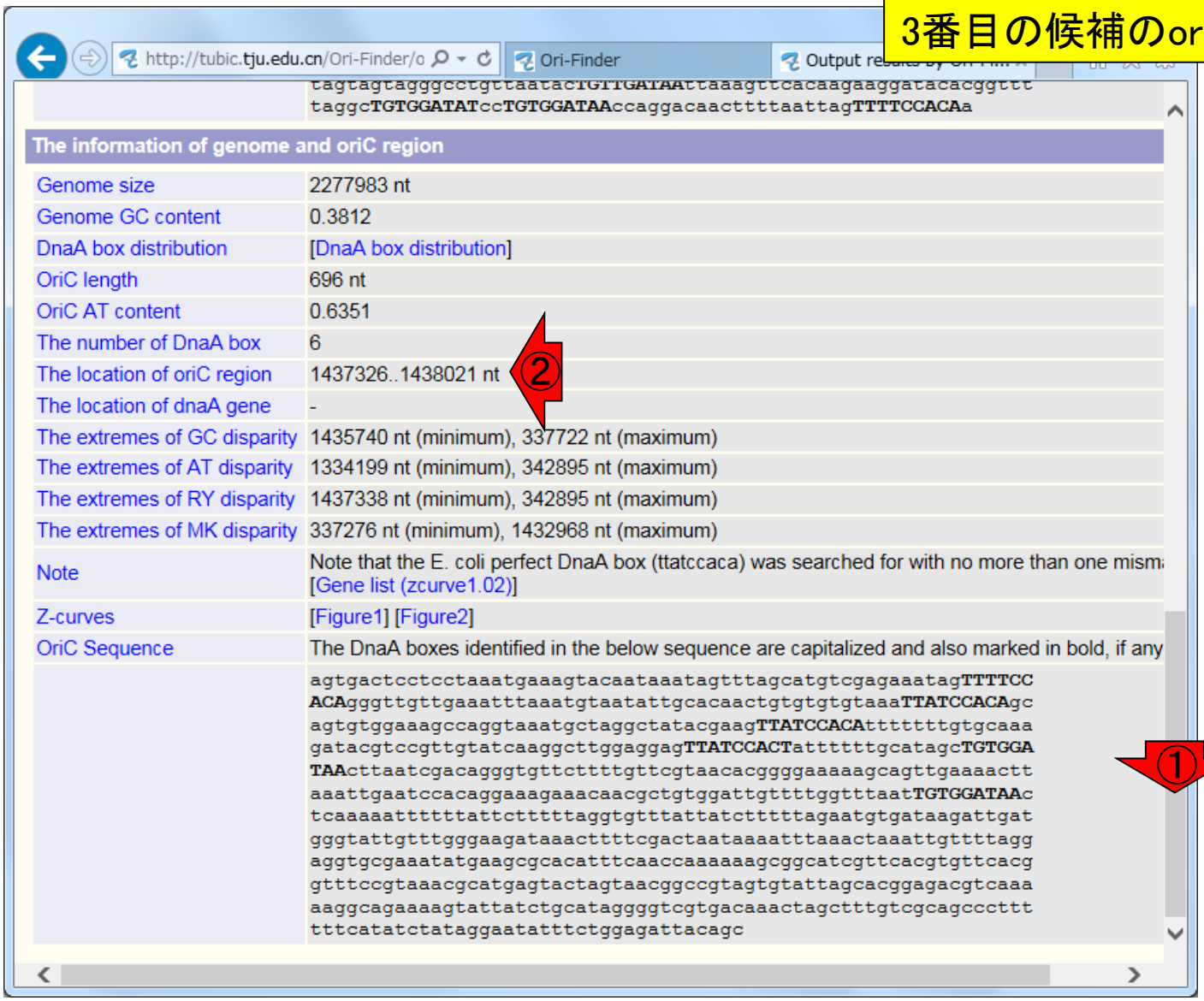
The information of genome and oriC region	
Genome size	2277983 nt
Genome GC content	0.3812
DnaA box distribution	[DnaA box distribution]
OriC length	653 nt
OriC AT content	0.6769
The number of DnaA box	6
The location of oriC region	1396338..1396990 nt
The location of dnaA gene	-
The extremes of GC disparity	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 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

aactaacgacctttcaacaag**TTATCCACA**gag**TTTTCCACA**ggttaagatatttttaat
taaaaataagtttcaaactctattgacttcgtttttttaaaatatttt**TTATCCACA**gtt
ttgcacagaaacactcgttcgtgatagcagttacaaacagttgatgtaccaacgattaag
actactttcgtttgtttggttgttag**TTATCCACA**agtggtgggcccgtc**TGTGGATAA**c
tttttcaactgctataacagttgttctgggggggctttatactttcttaacaattatacca
gaaagtttccacagag**TTTTCCACA**gattaattatttcaaaaaaatacttgcaaaatgac
tactgaactacttcatttcctgtaacttattttgttcgttaaatttaaccatttttaggc
ttttttgcgcattatactagtataaatatatttgtcgttcgtctttctctaattatgccc
ttttcagatacctatatgtgggttaatcaatataattccgaactttaacttctaaatatg
tattataaaacggagatccaatgcaacctgtggtttgccgaatctccgttatttttattta
tatatattgttagaacgttcccccagaaggtatcagtctaagttttatgtgcc

The information of genome and oriC region	
Genome size	2277983 nt

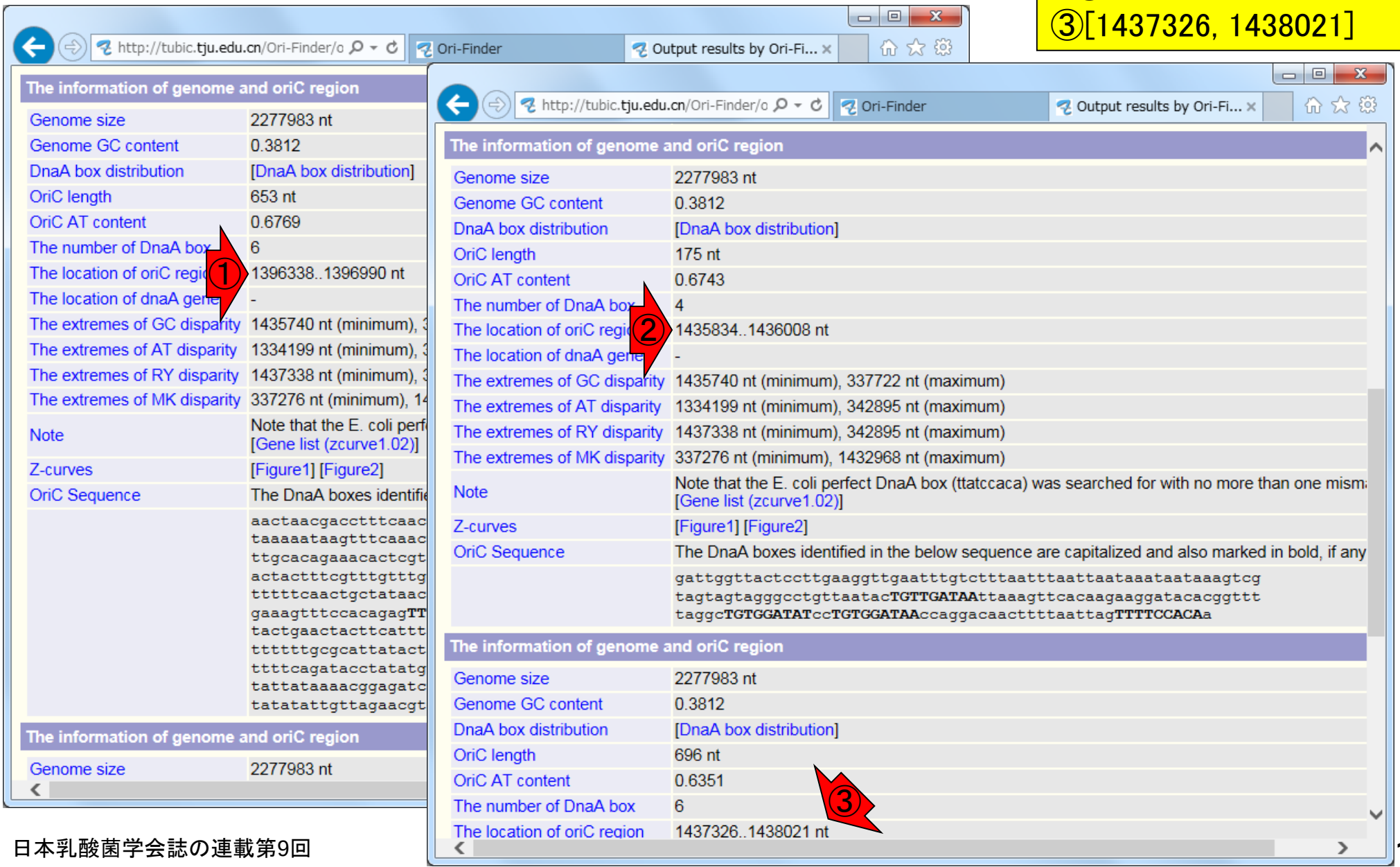
W3-4：Ori-Finder

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]



W3-4：Ori-Finder

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



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

Ori-Finder

Output results by Ori-Fi...

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), 3
The extremes of AT disparity	1334199 nt (minimum), 3
The extremes of RY disparity	1437338 nt (minimum), 3
The extremes of MK disparity	337276 nt (minimum), 14
Note	Note that the E. coli perf [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identifie

aactaacgacctttcaac
taaaaataagtttcaaac
ttgcacagaaacactcgt
actactttcgtttgtttg
tttttcaactgctataac
gaaagtttccacagagTT
tactgaactacttcattt
ttttttgcgcattatact
ttttcagatacctatatg
tattataaaacggagatc
tatatattgttagaacgt

The information of genome and oriC region	
Genome size	2277983 nt

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

Ori-Finder

Output results by Ori-Fi...

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 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

gattggttactccttgaaggttgaatttgtctttaatttaattaataaataataaagtcg
tagtagtagggcctgttaatac**TGTTGATAA**ttaaagttcacaagaaggatacacggttt
taggc**TGTGGATAT**cc**TGTGGATAA**ccaggacaacttttaattag**TTTTCCACA**a

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

W3-4：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

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

Ori-Finder

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), 3
The extremes of AT disparity	1334199 nt (minimum), 3
The extremes of RY disparity	1437338 nt (minimum), 3
The extremes of MK disparity	337276 nt (minimum), 14
Note	Note that the E. coli perf [Gene list (zcurve1.02)]
Z-curves	[Figure1] [Figure2]
OriC Sequence	The DnaA boxes identifie

aactaacgacctttcaac
taaaaataagtttcaaac
ttgcacagaaacactcgt
actactttcgtttgtttg
tttttcaactgctataac
gaaagtttccacagagTT
tactgaactacttcattt
ttttttgcgcattatact
ttttcagatacctatatg
tattataaaacggagatc
tatatattgttagaacgt

The information of genome and oriC region	
Genome size	2277983 nt

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

Ori-Finder

Output results by Ori-Fi...

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 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

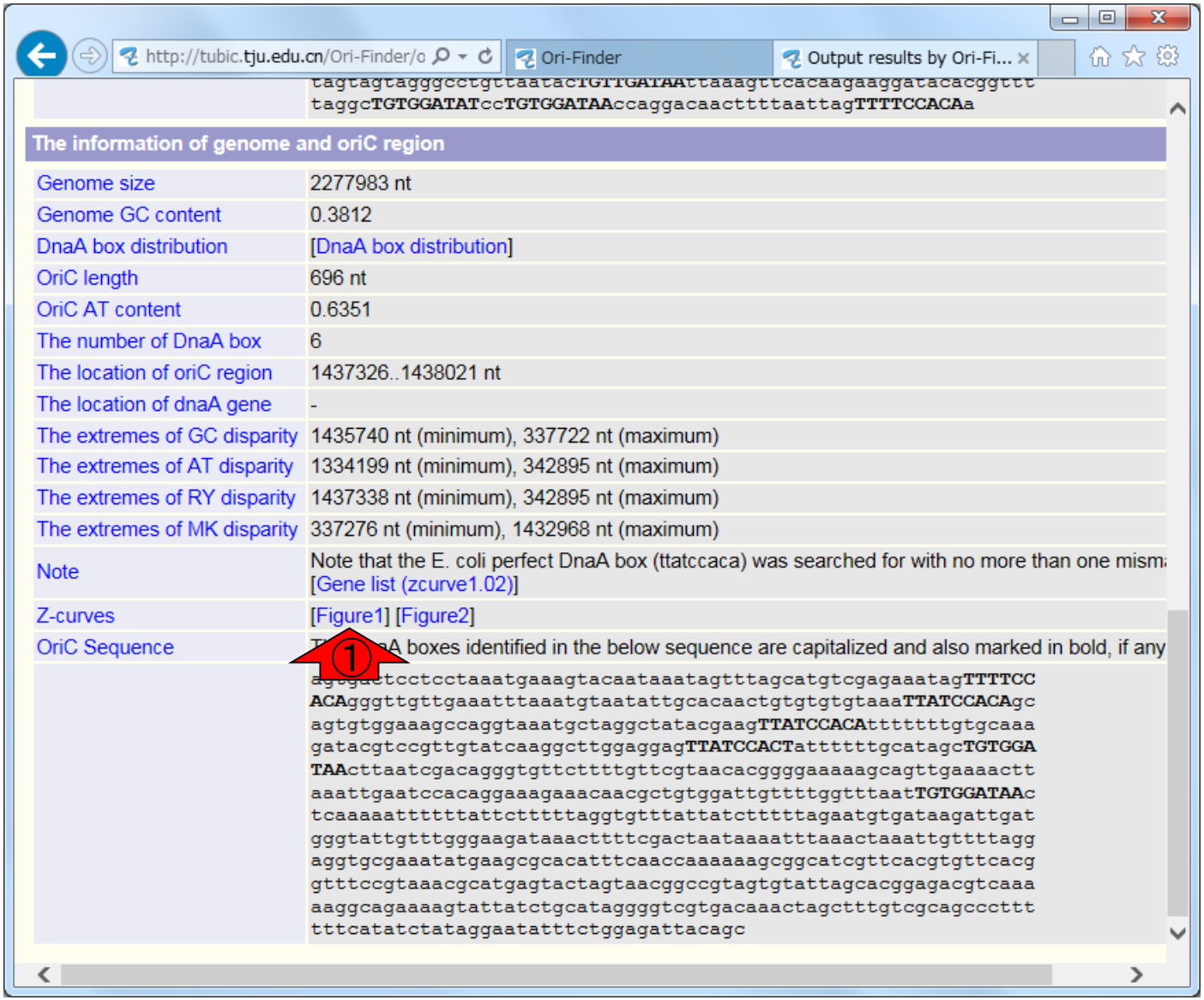
gattggttactccttgaaggttgaatttgtctttaatttaattaataaataataaagtcg
tagtagtagggcctgttaatac**TGTTGATAA**ttaaagttcacaagaaggatacacggttt
taggc**TGTGGATAT**cc**TGTGGATAA**ccaggacaacttttaattag**TTTTCCACA**a

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

③

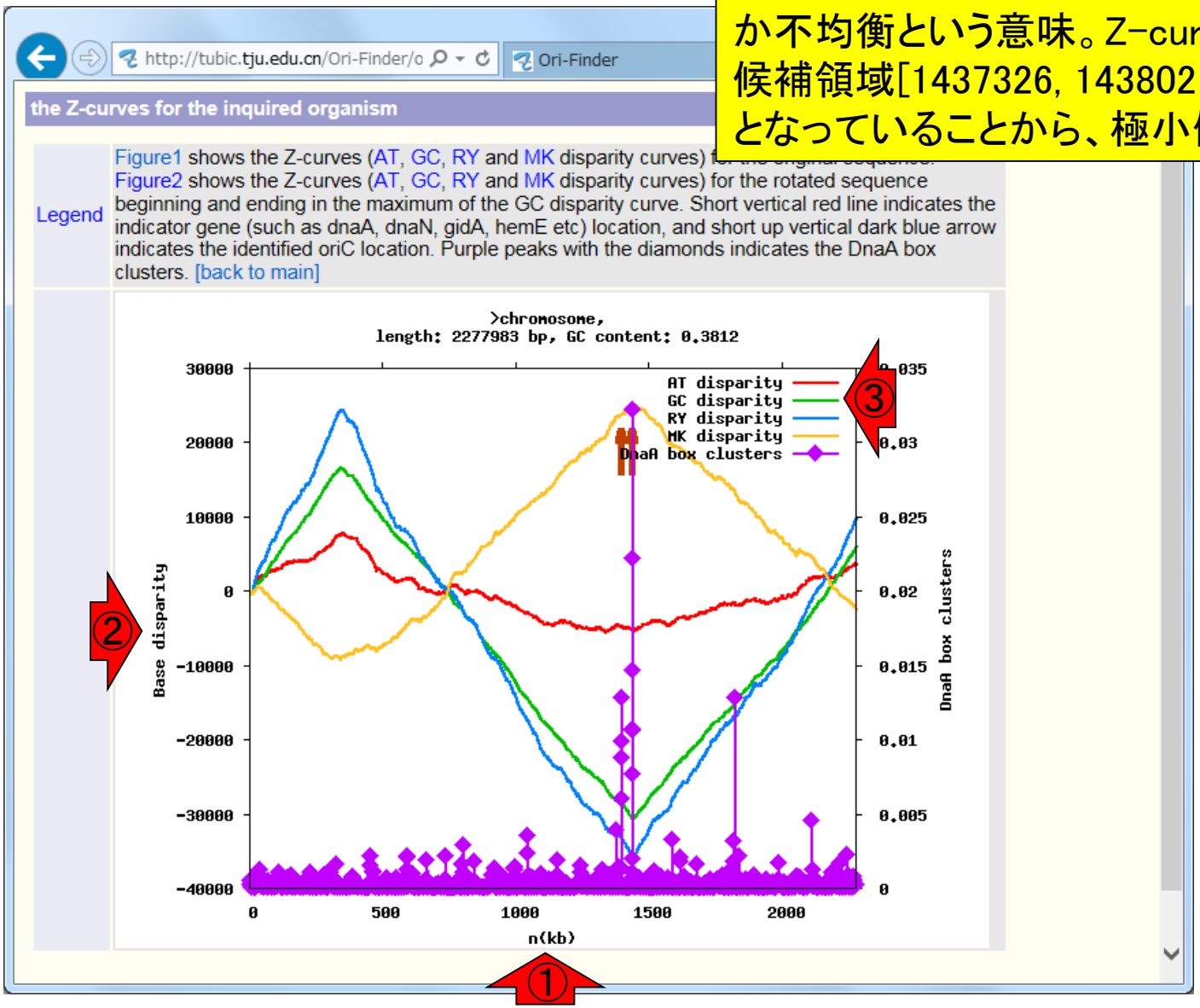
①一応Z-curve法で得られた図を見てみる

W3-5：Z-curve



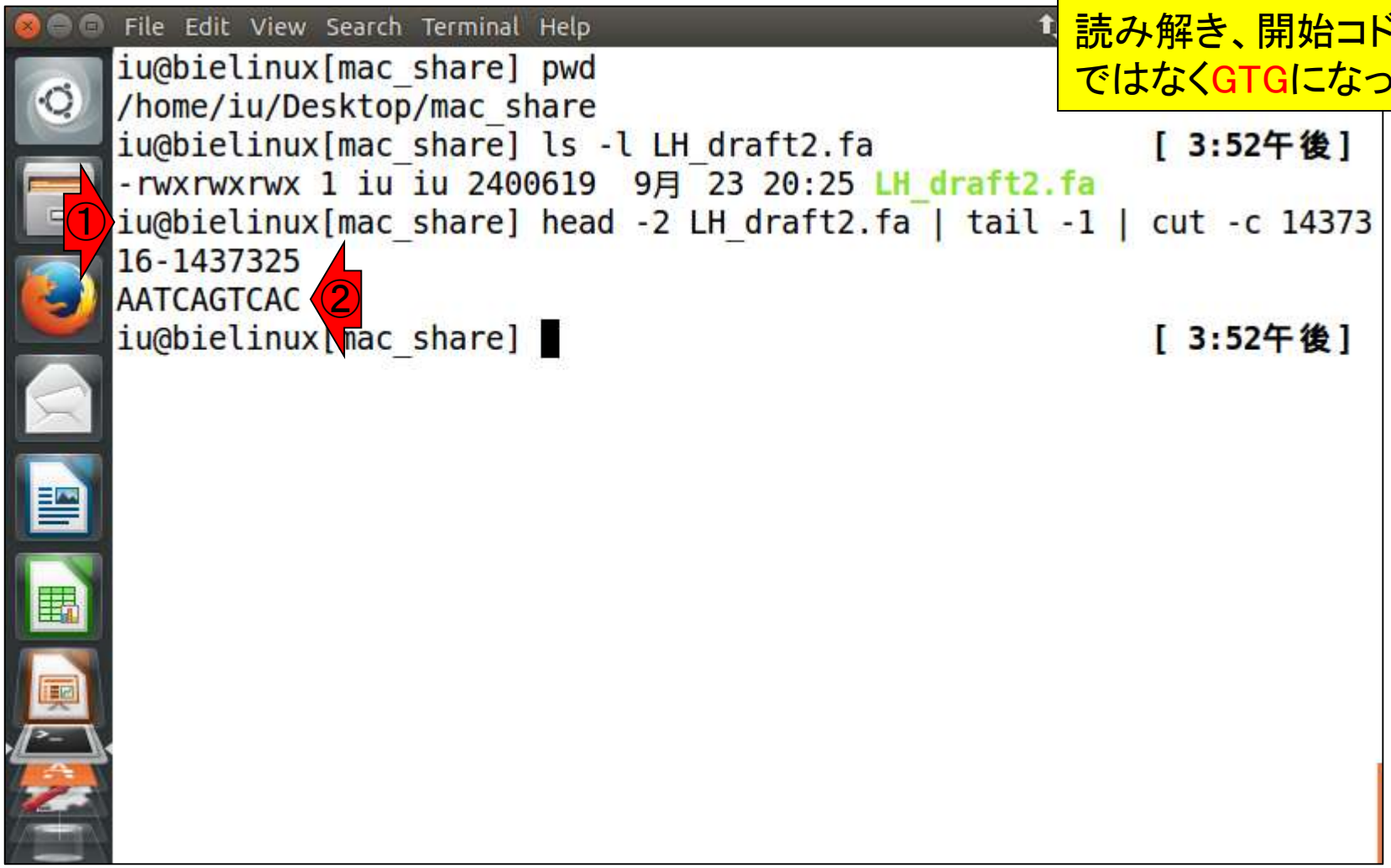
W3-5：Z-curve

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



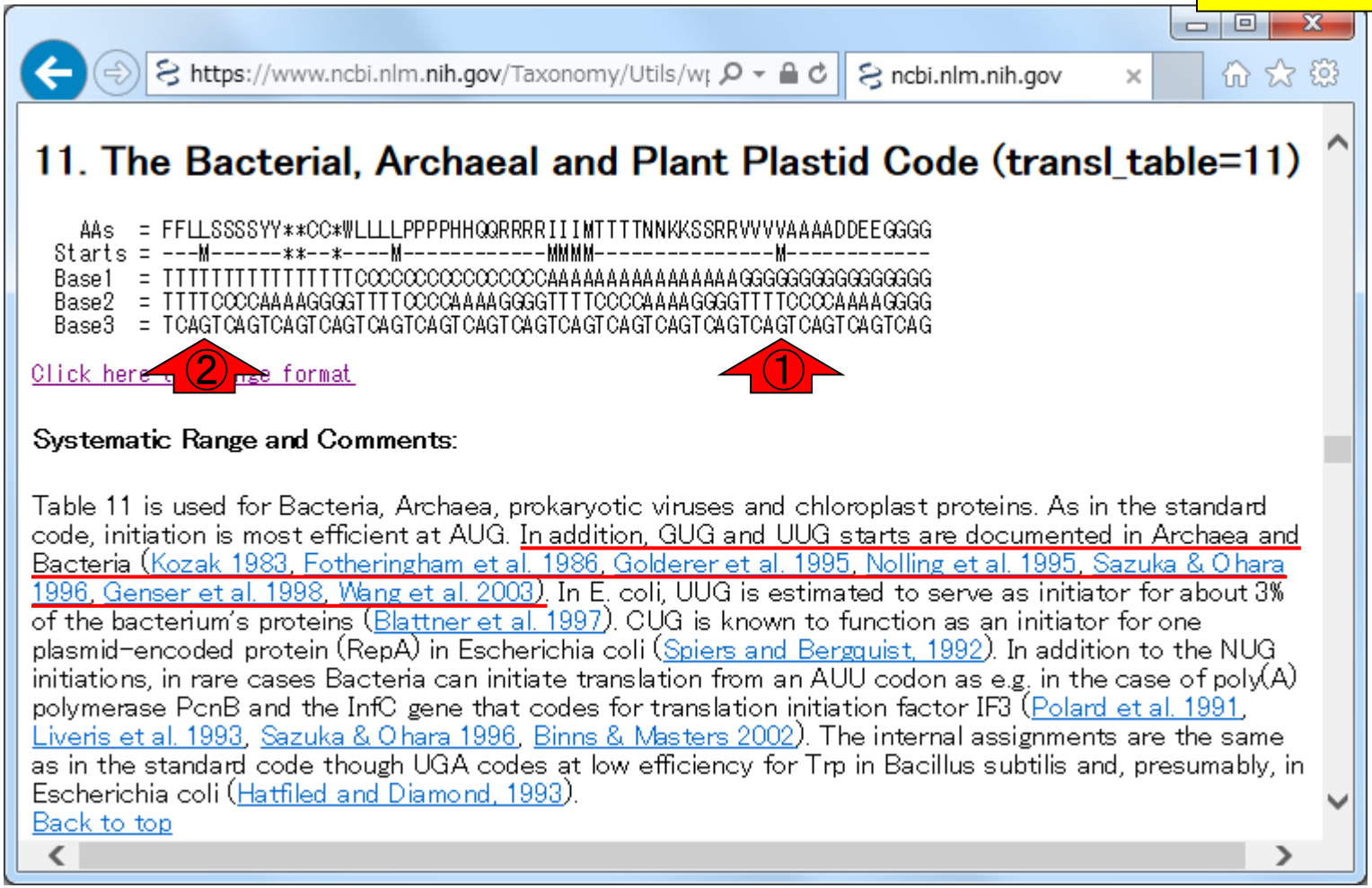
W4-1：*dnaA*の開始コドン

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②AATCAGTCAC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



W4-2：バクテリアのコドン表

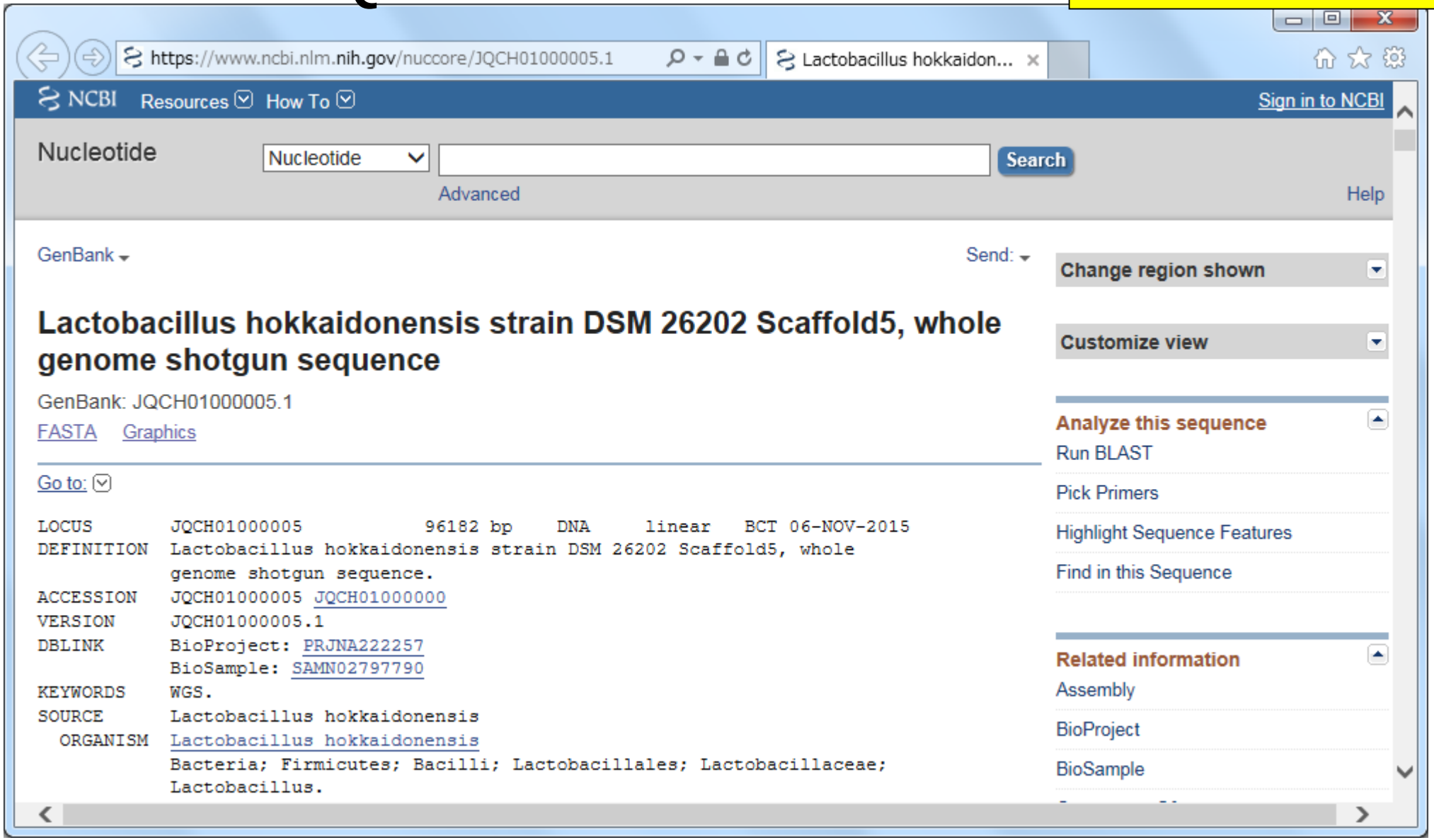
バクテリアのコドン表を眺めると、
開始コドンが①GTG（や②TTG）で
あったとしても間違いではない



https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG11

W4-3：JQCH01000005.1

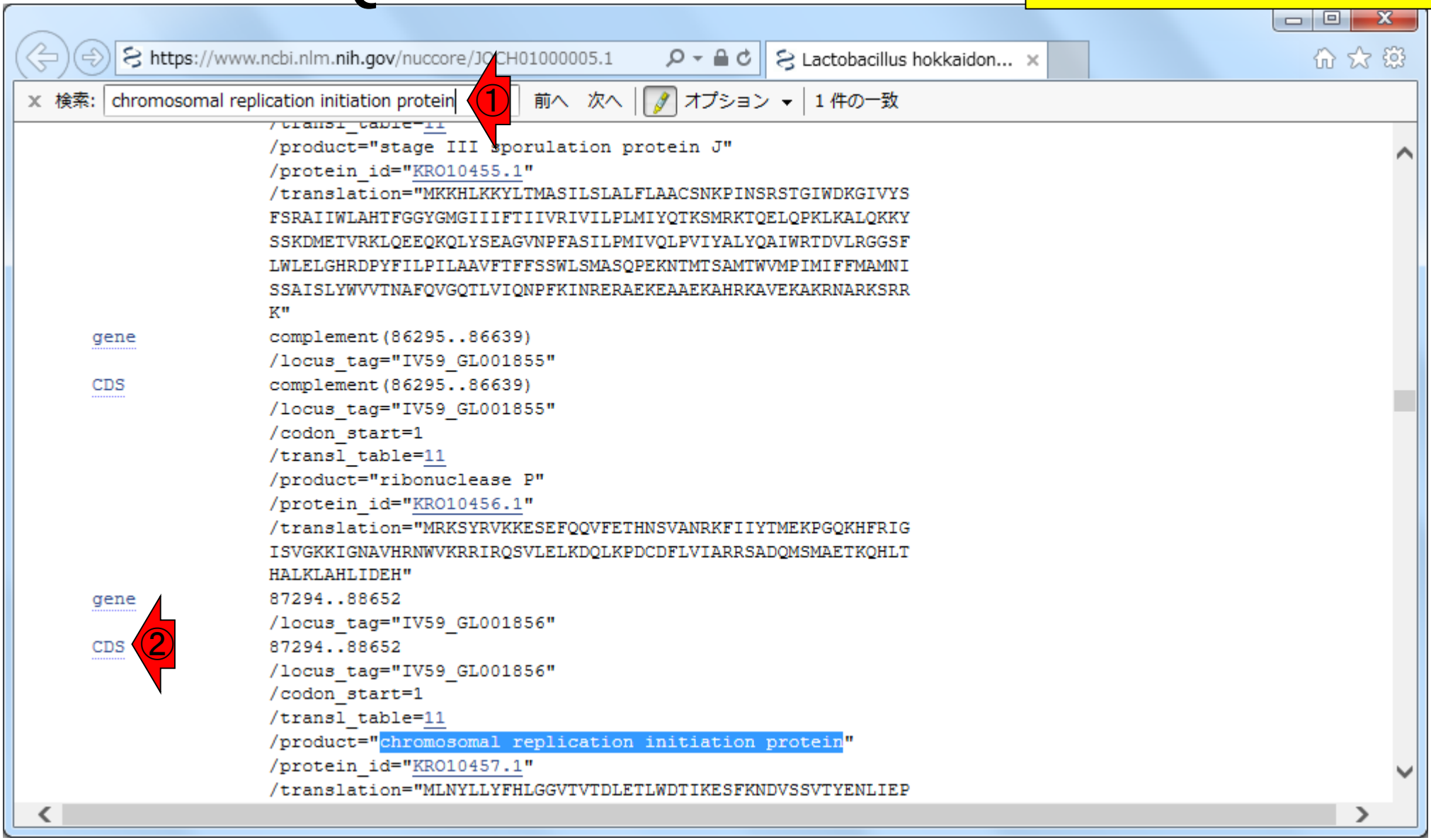
他の研究グループから出された*L. hokkaidonensis*の基準株のゲノム配列



https://www.ncbi.nlm.nih.gov/nuccore/JQCH01000005.1

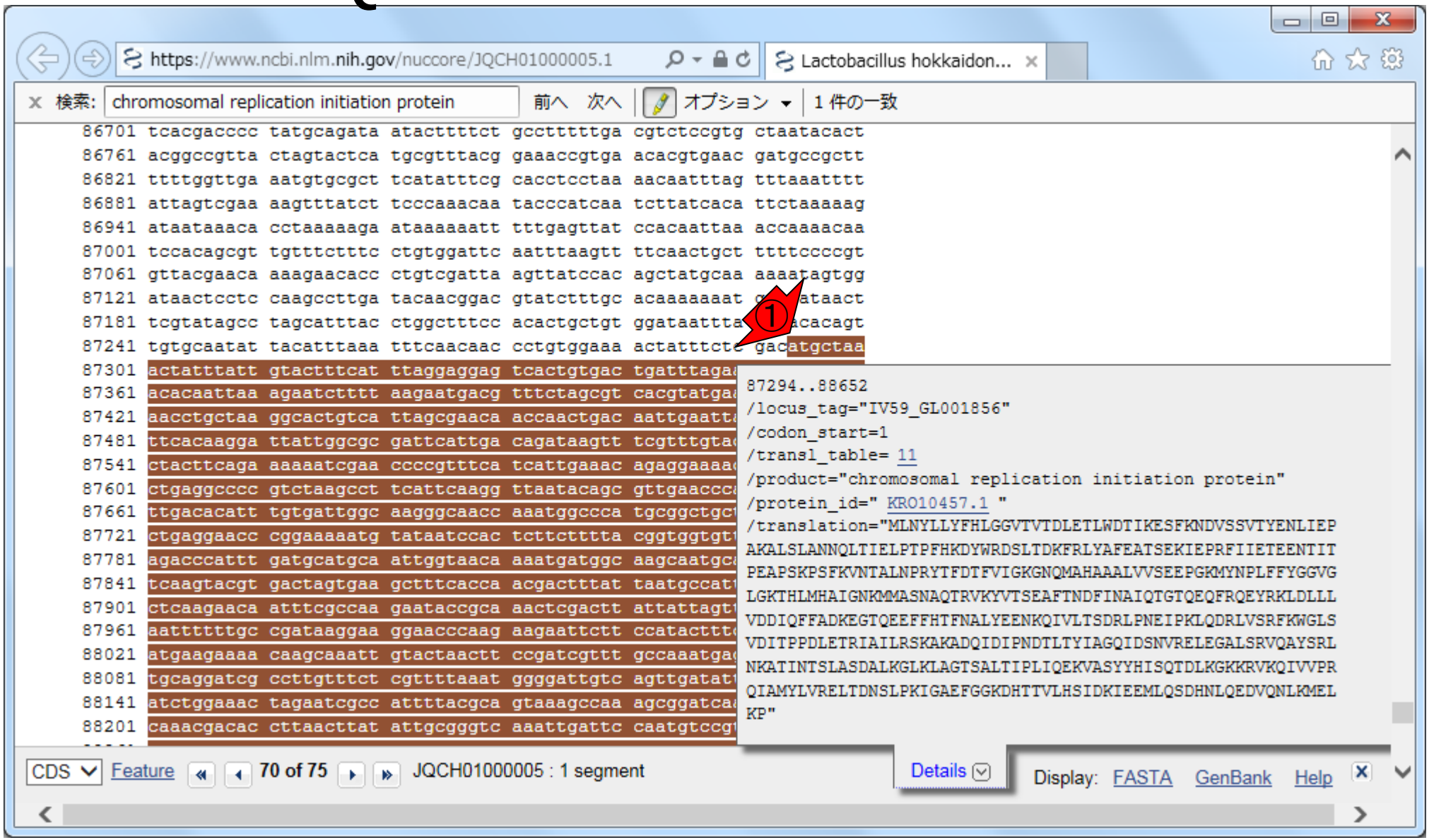
W4-3：JQCH01000005.1

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



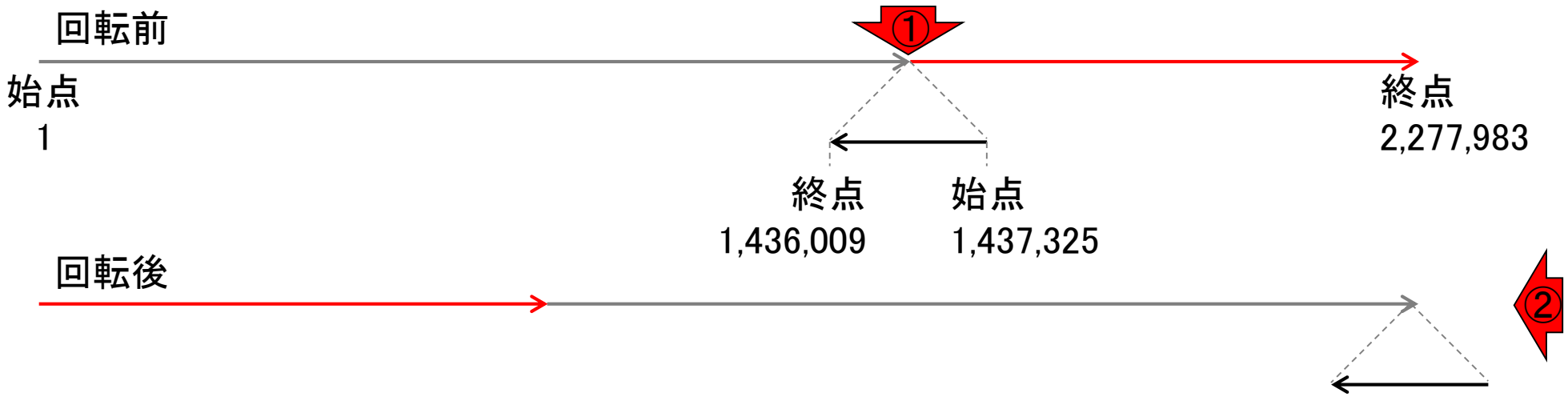
W4-3：JQCH01000005.1

①確かにATGから始まっていることがわかる



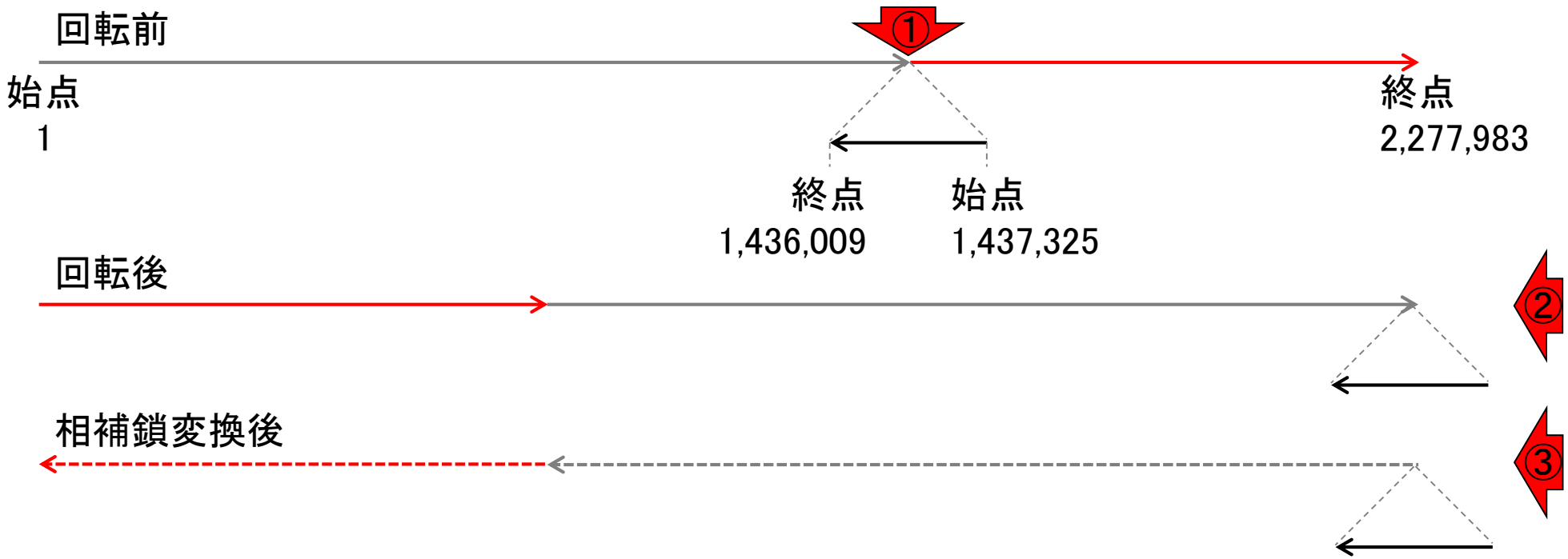
W4-4：回転前後

染色体配列の回転前後の関係（イメージ図）。①DFAST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）



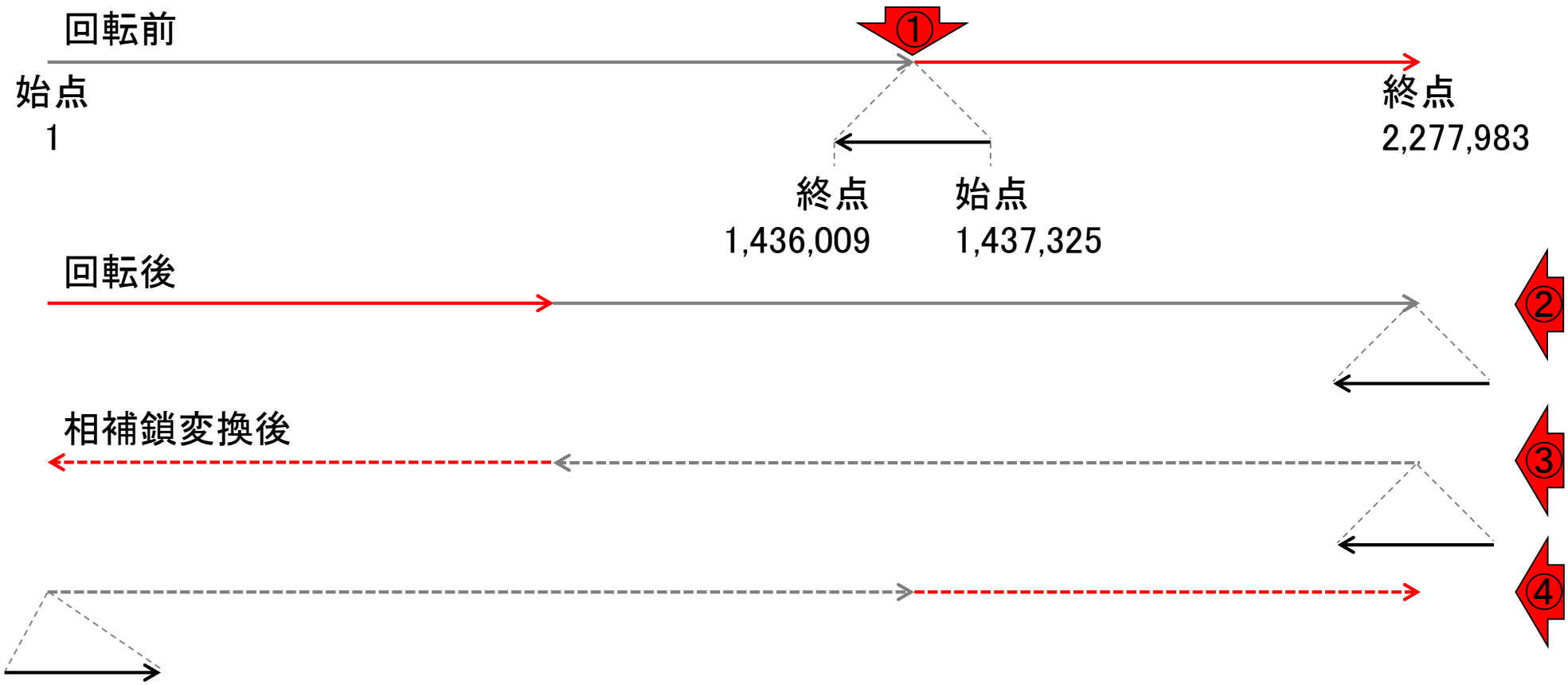
W4-4：相補鎖変換後

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



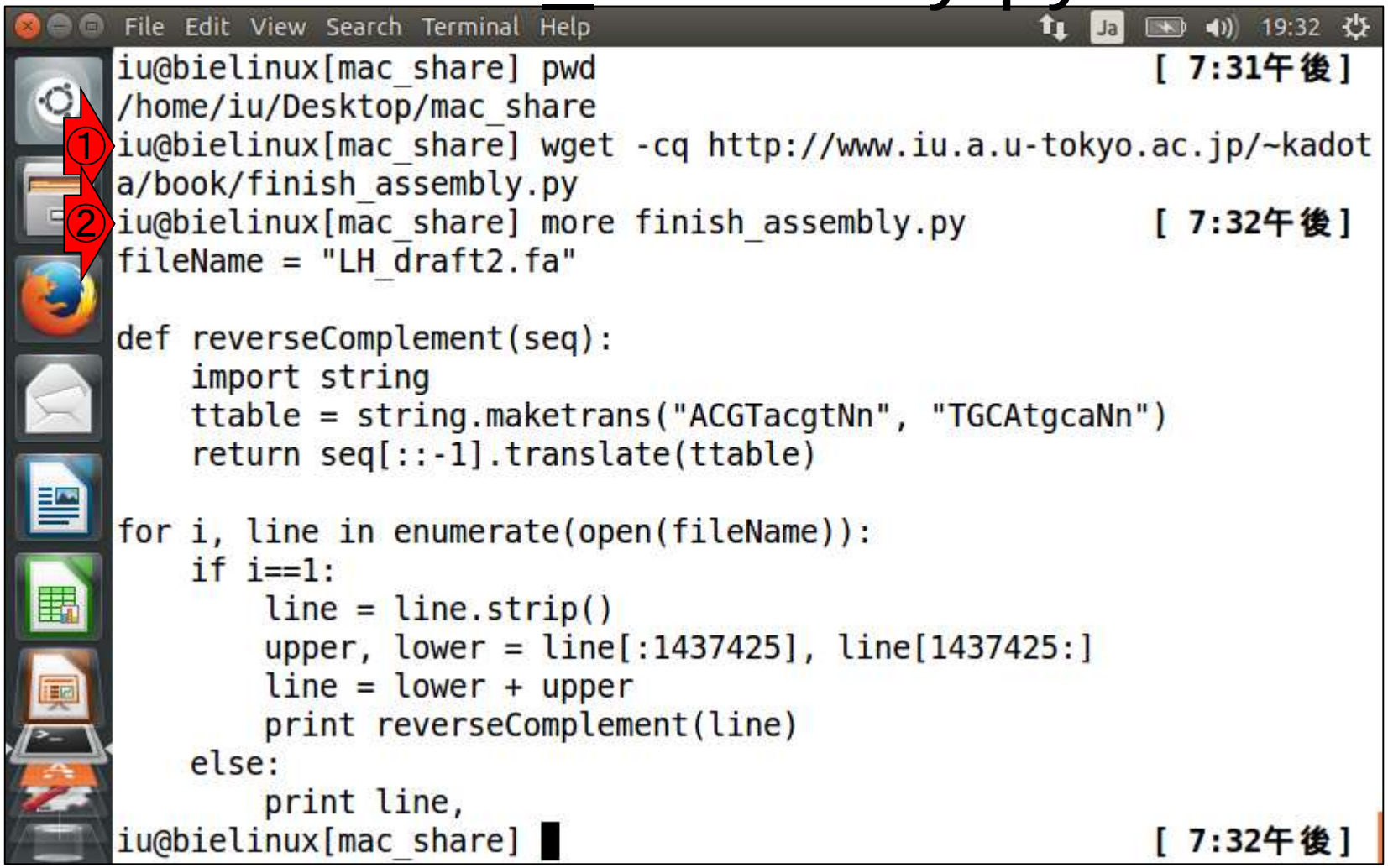
W4-4：相補鎖変換後

④は③を逆向きにしただけで、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）



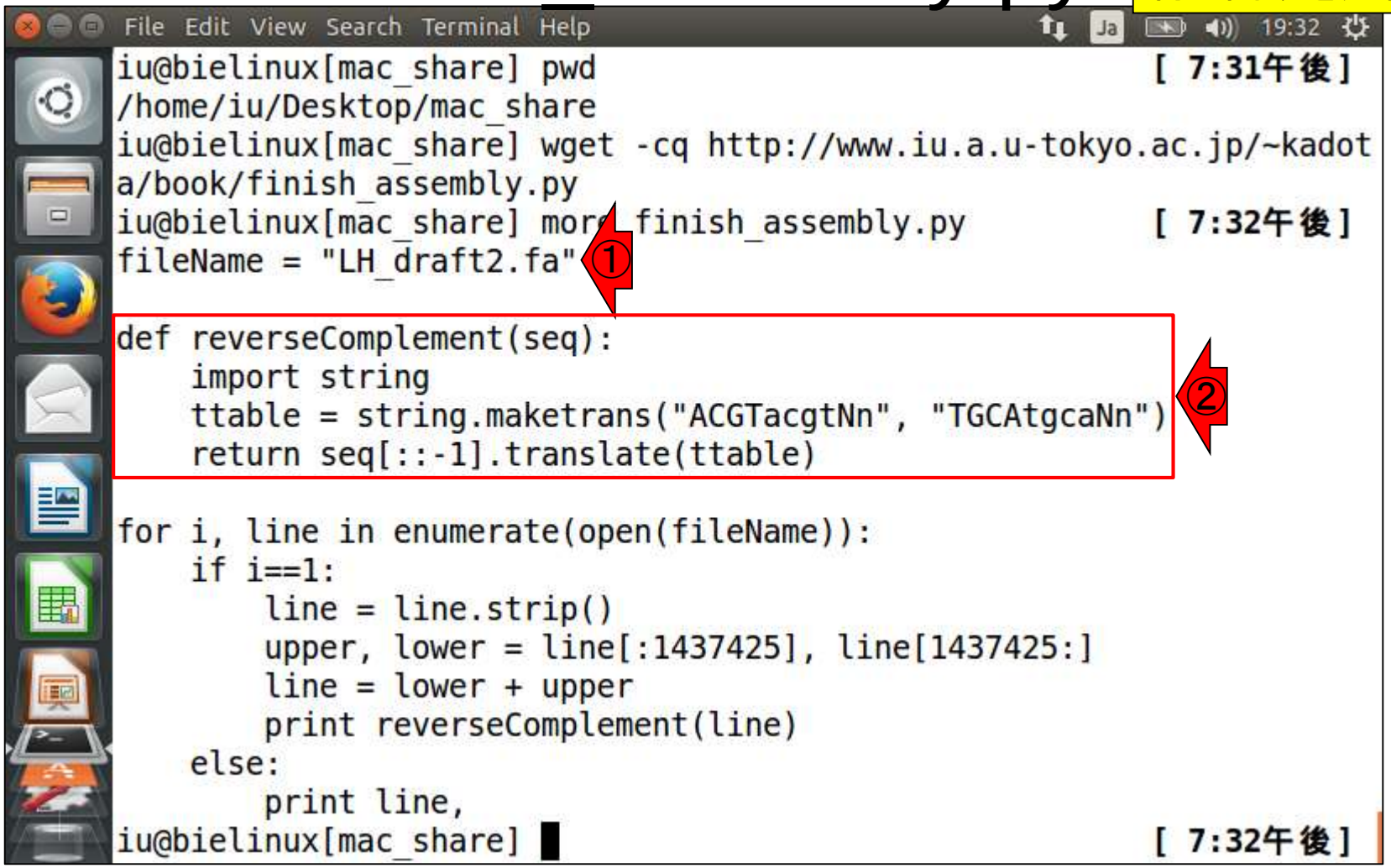
finish_assembly.pyの①wgetと②more

W4-5：finish_assembly.py



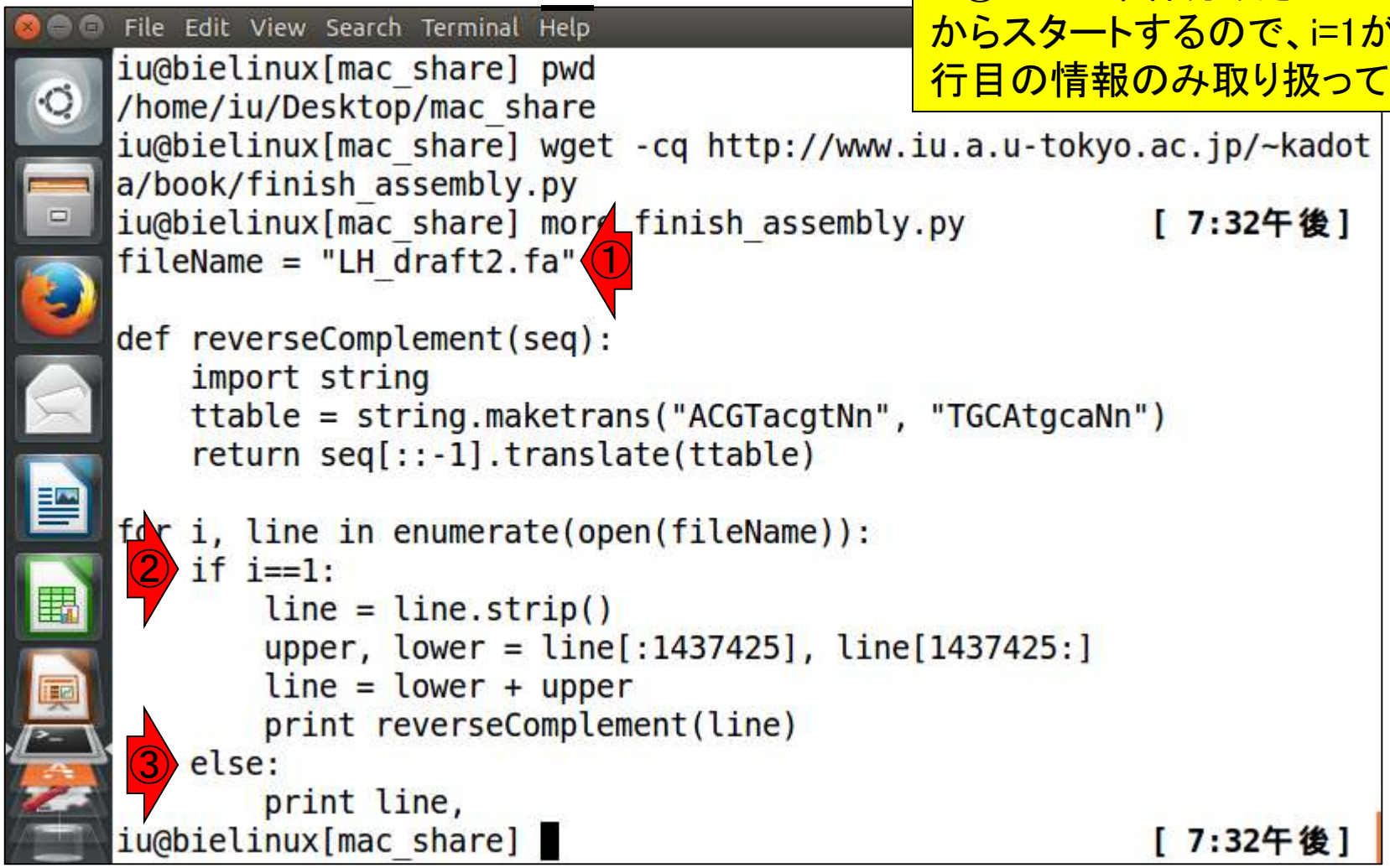
W4-5：finish_assembly.py

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



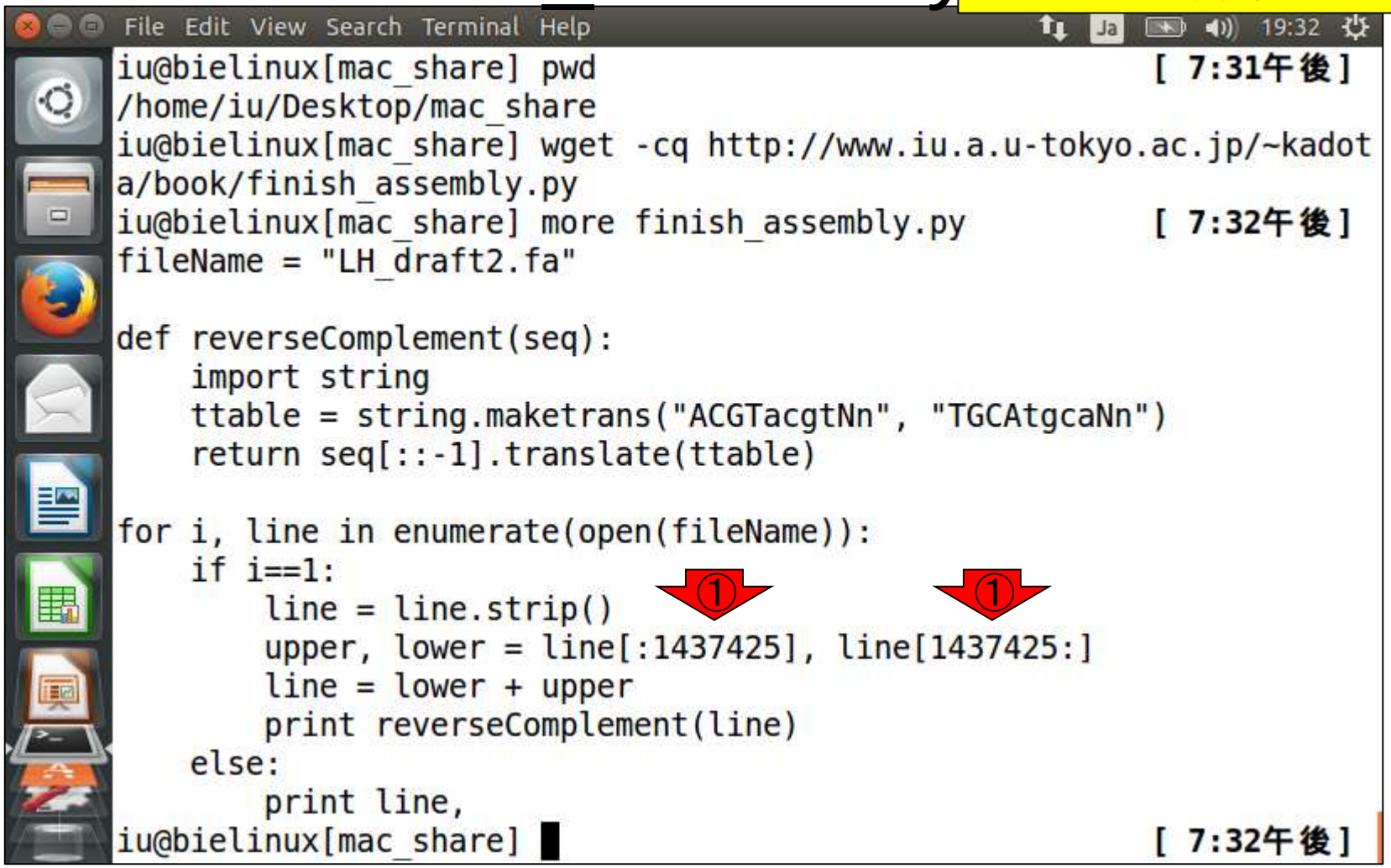
W4-5：finish_assemb

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



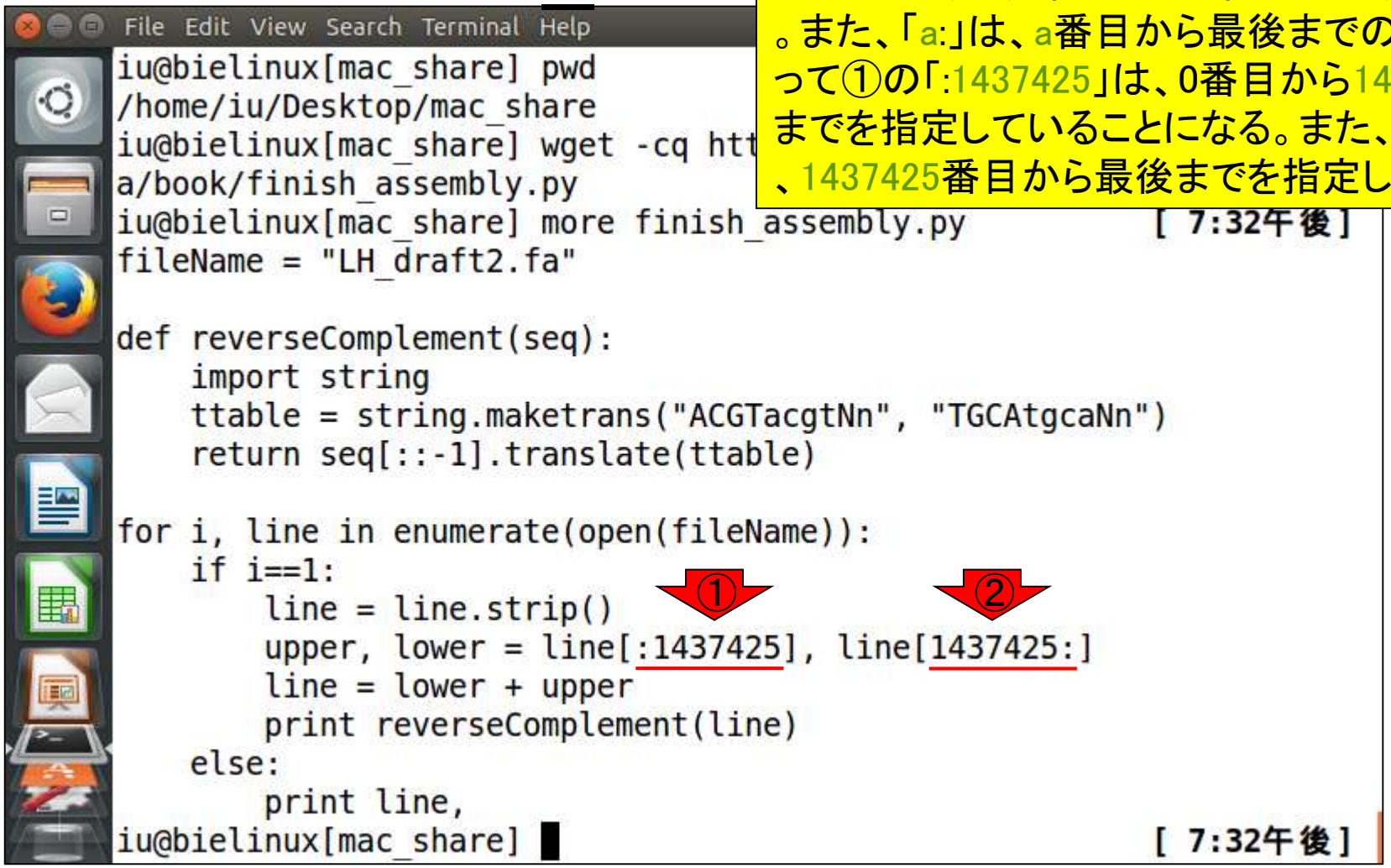
W4-5：finish_assembly

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



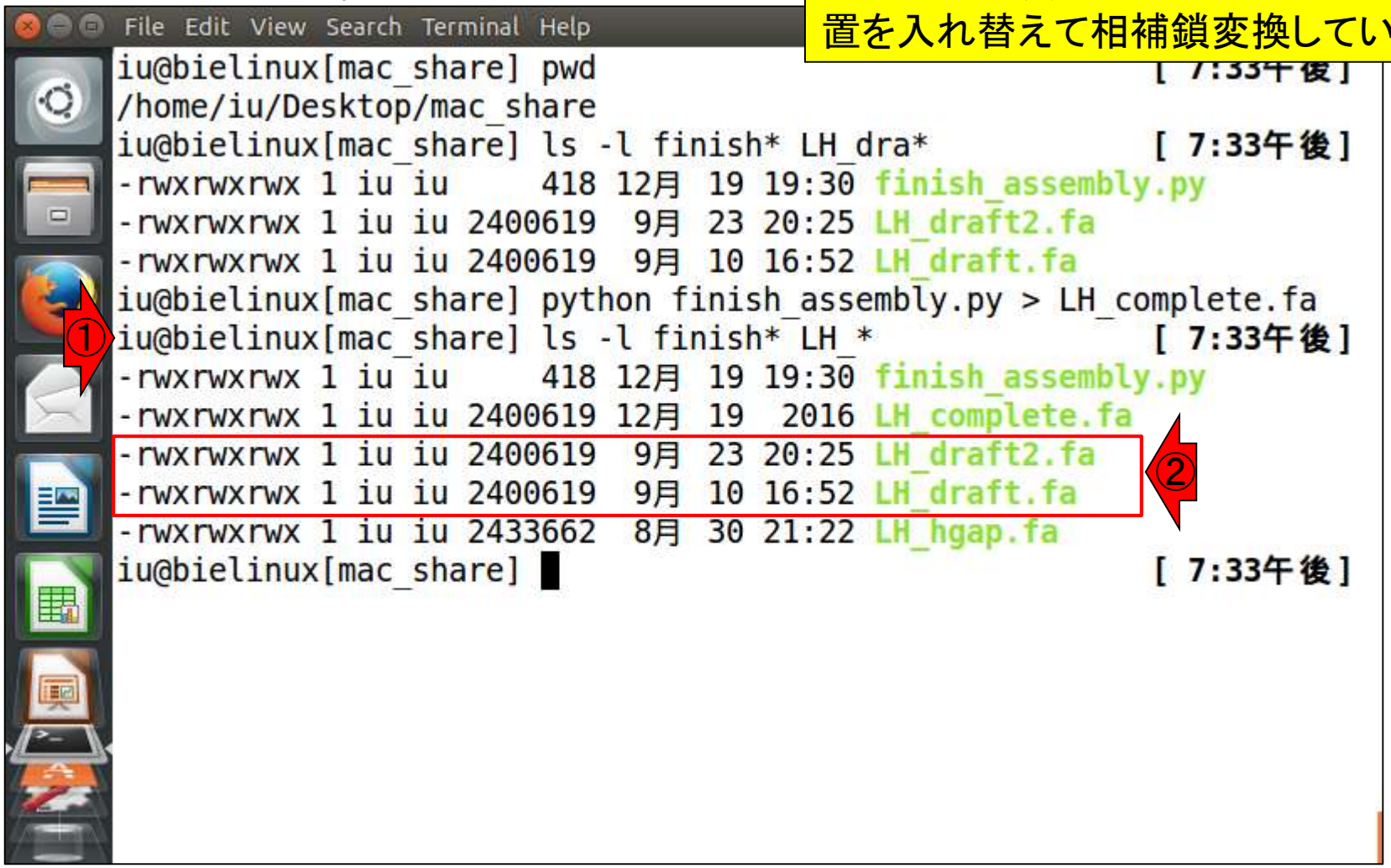
W4-5：finish_asse

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番目から最後までの指定となる。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



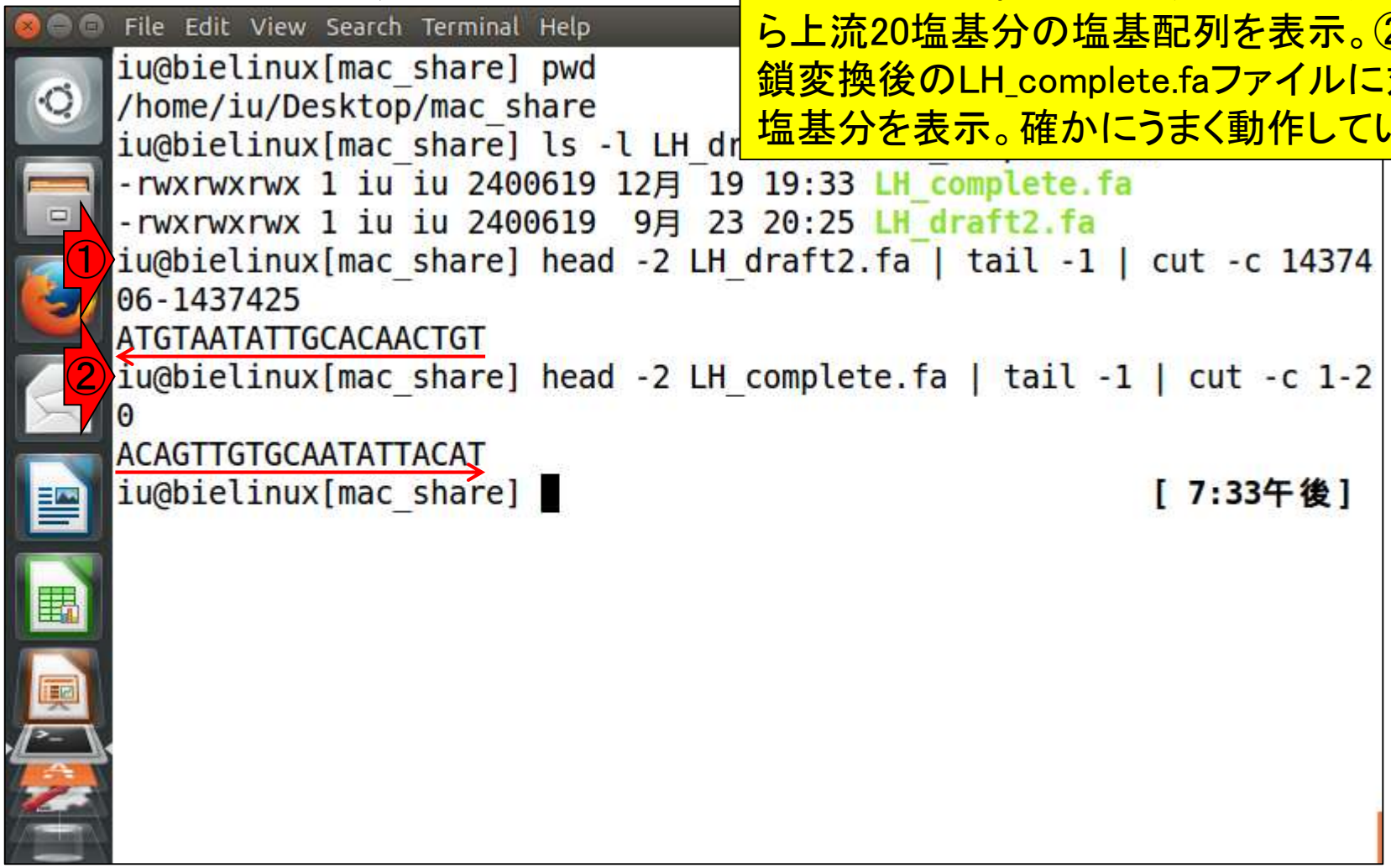
W4-6：実行

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



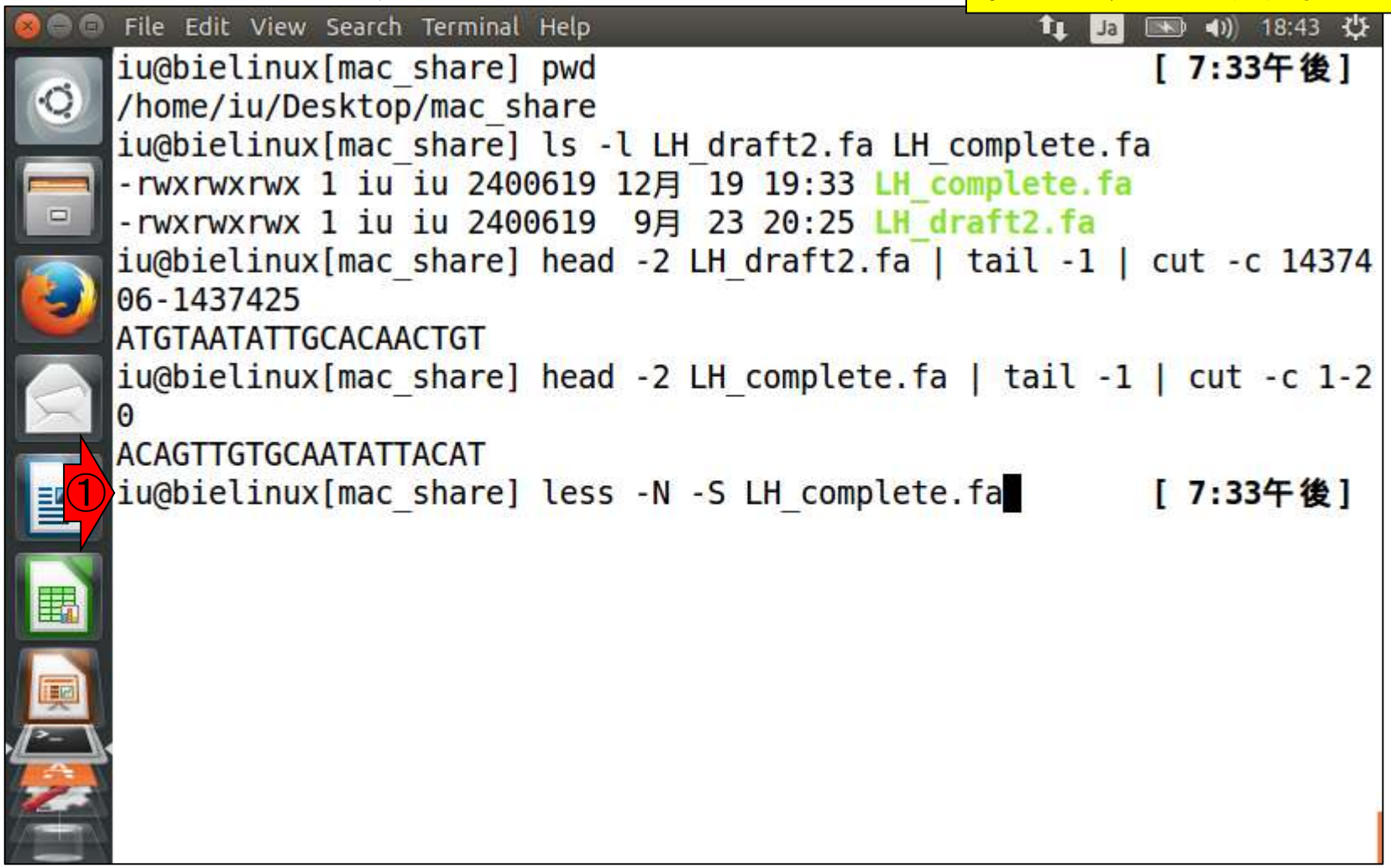
W4-7：動作確認

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



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

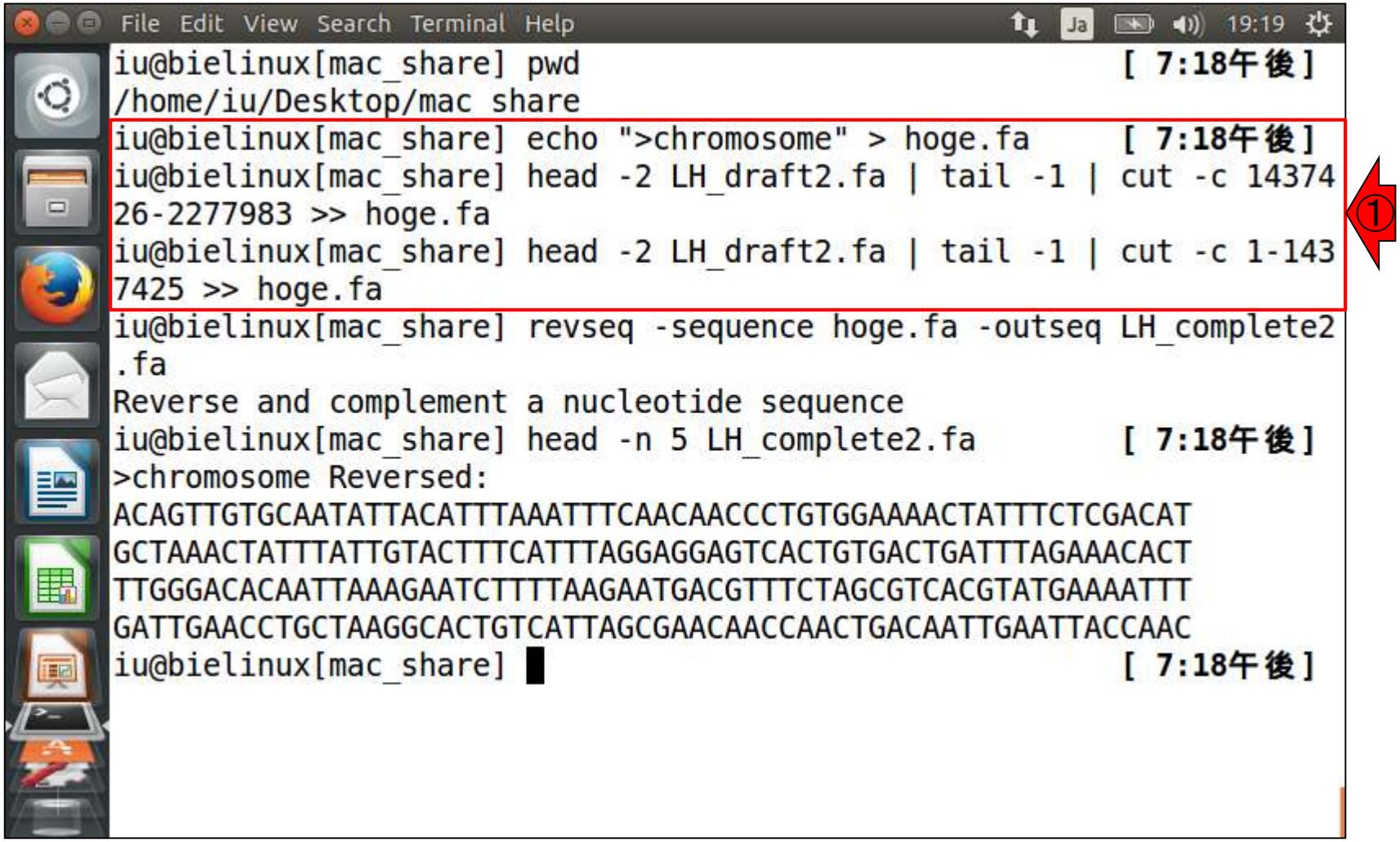


LH_complete.faは全部で6行であることを確認。qで抜ける

W4-7：動作確認

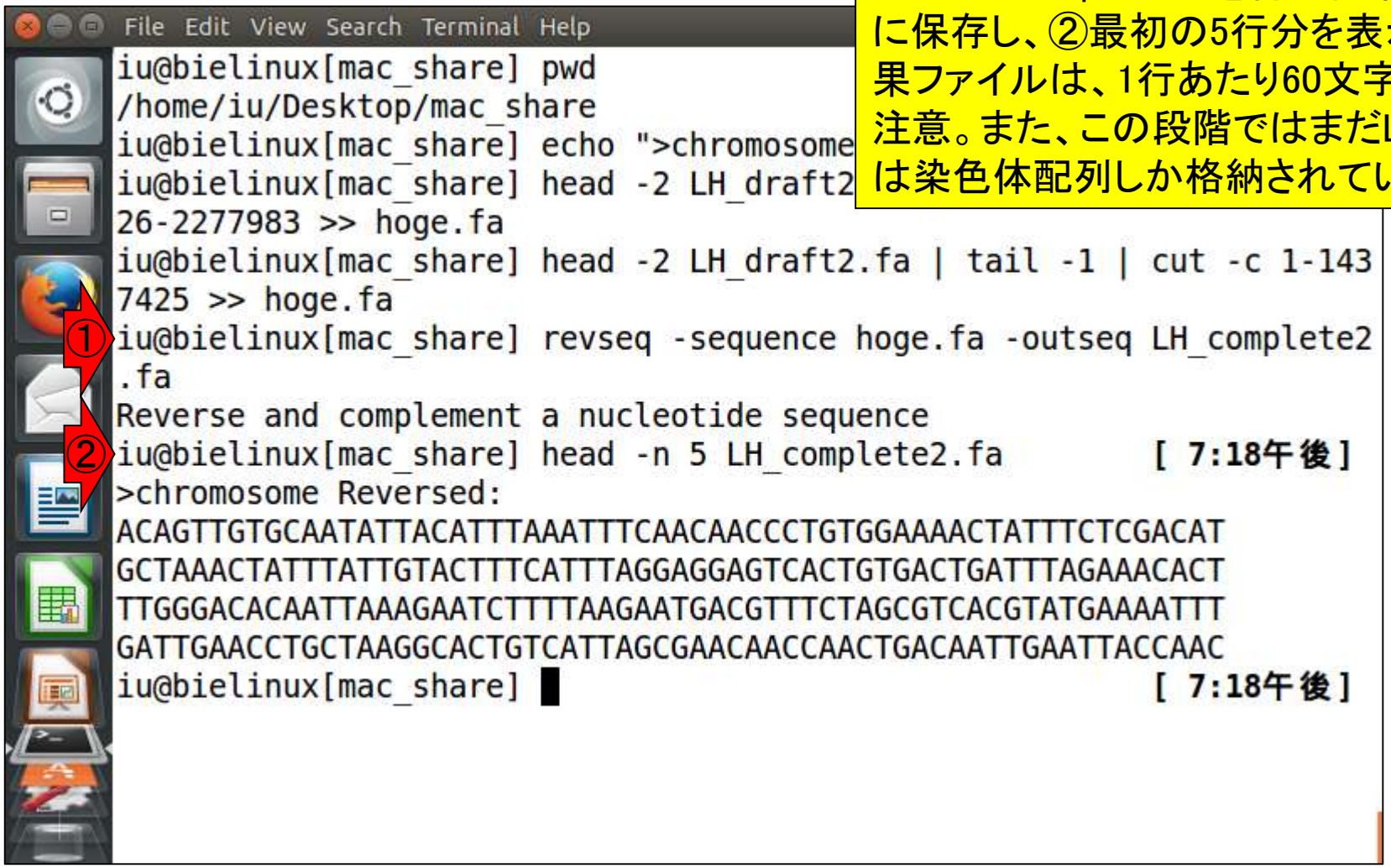
W4-8：もう一つの手段

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



W4-8：もう一つの手段

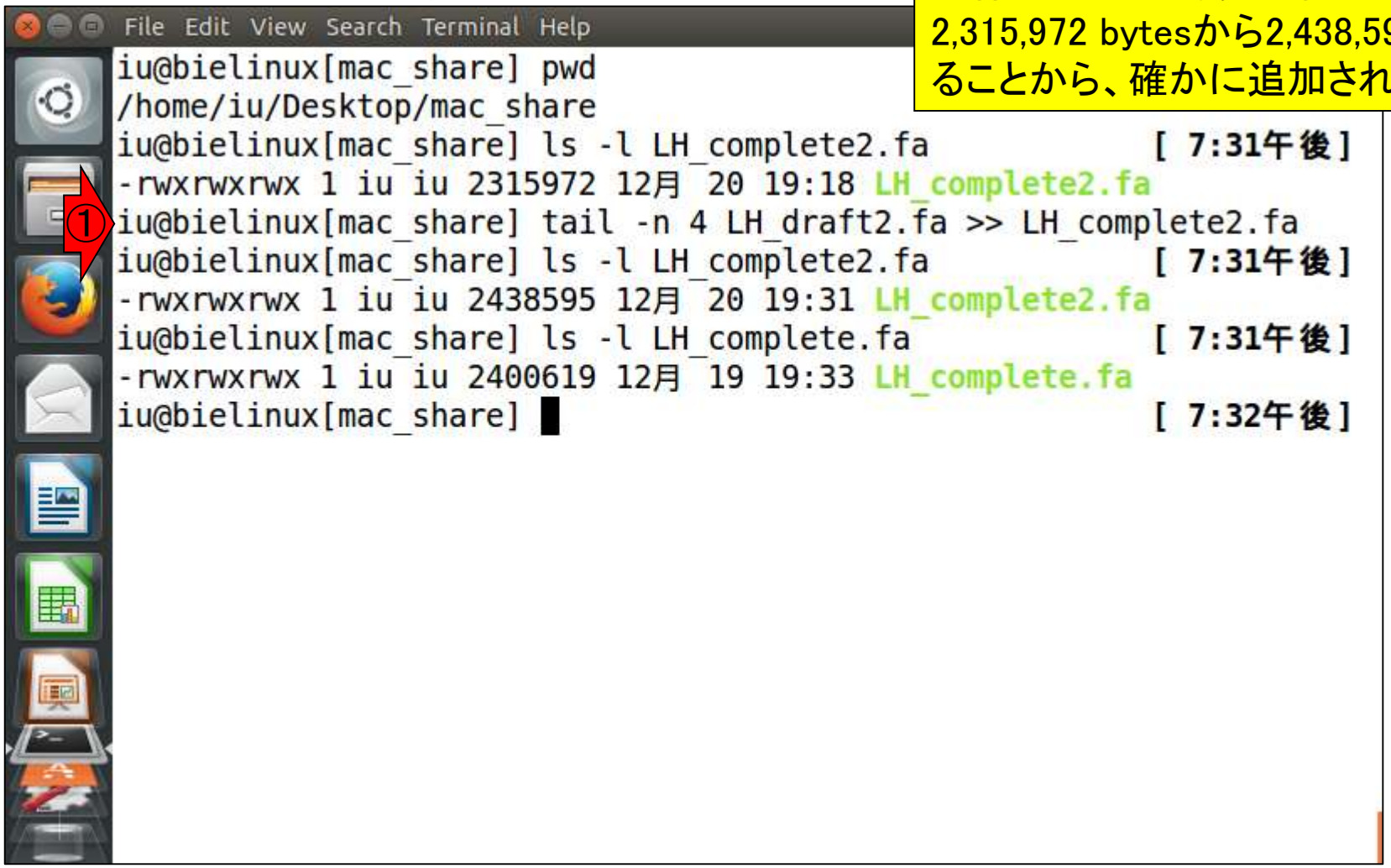
次に、Bio-Linuxにプレインストールされている、①EMBOSSのrevseqプログラムを用いてhoge.faのreverse complementを得た結果をLH_complete2.faに保存し、②最初の5行分を表示。revseq実行結果ファイルは、1行あたり60文字になっている点に注意。また、この段階ではまだLH_complete2.faには染色体配列しか格納されていない点にも注意



EMBOSS：Rice et al., *Trends Genet*, **16**: 276–277, 2000

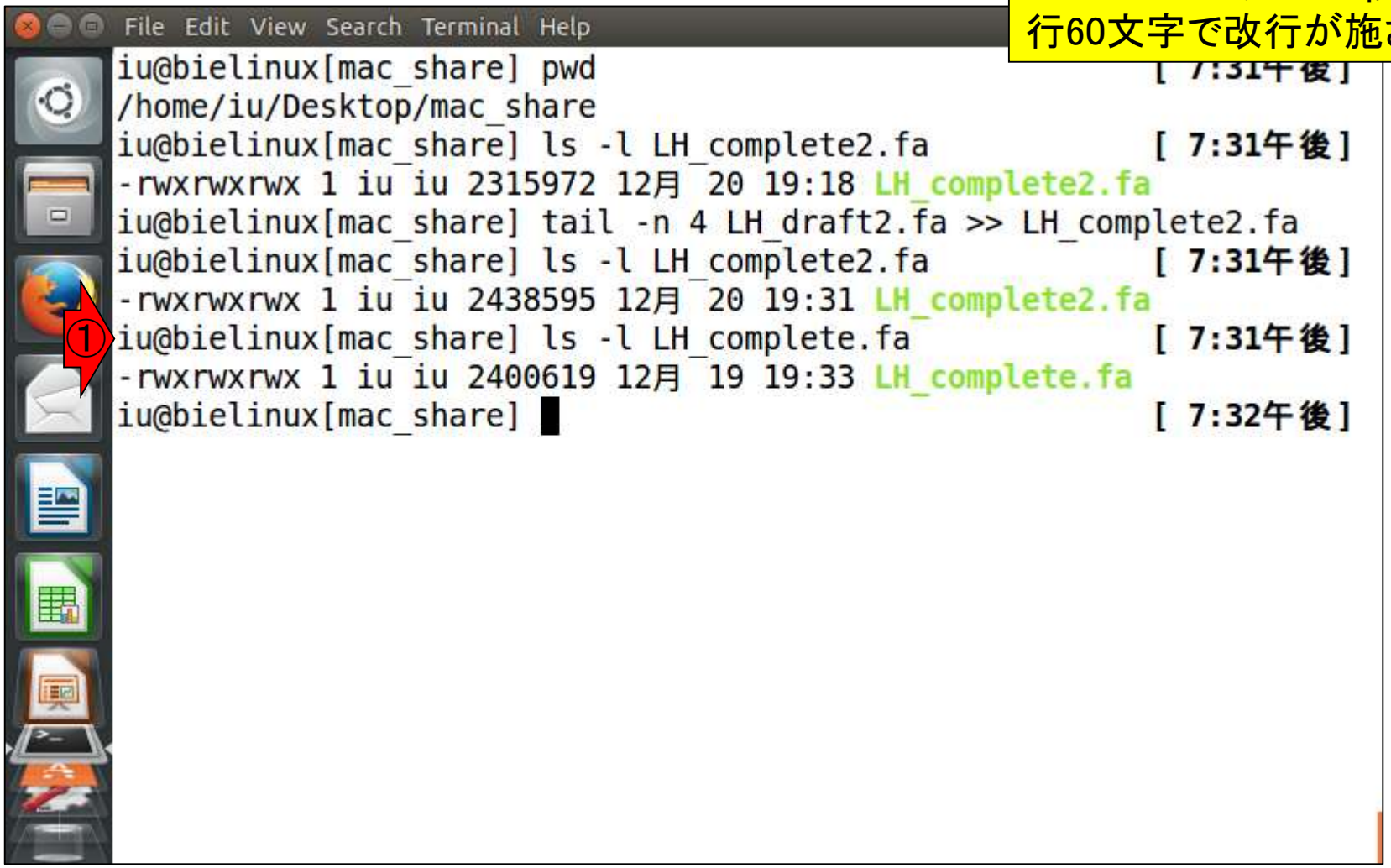
W4-8：もう一つの手段

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



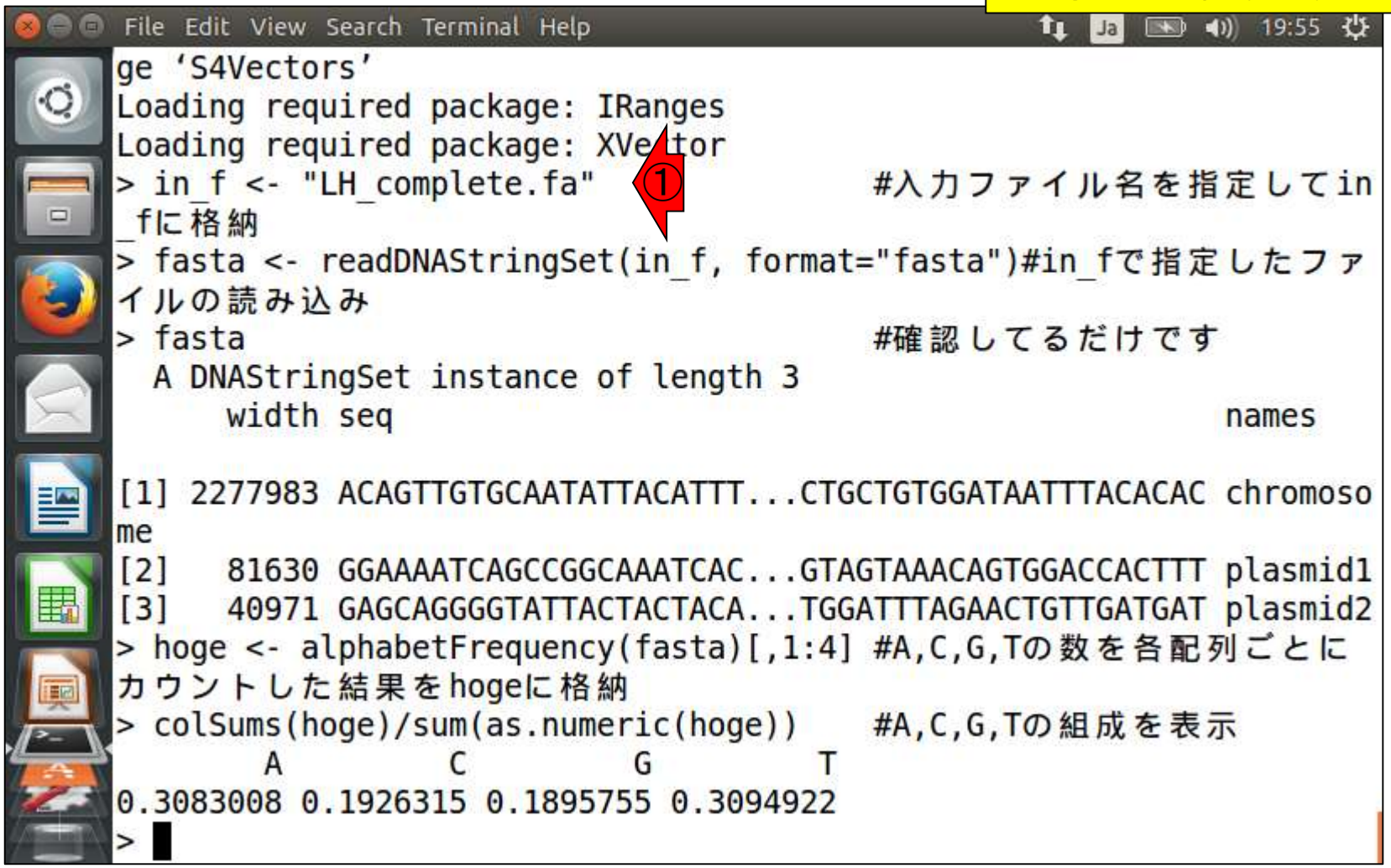
W4-8：もう一つの手段

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



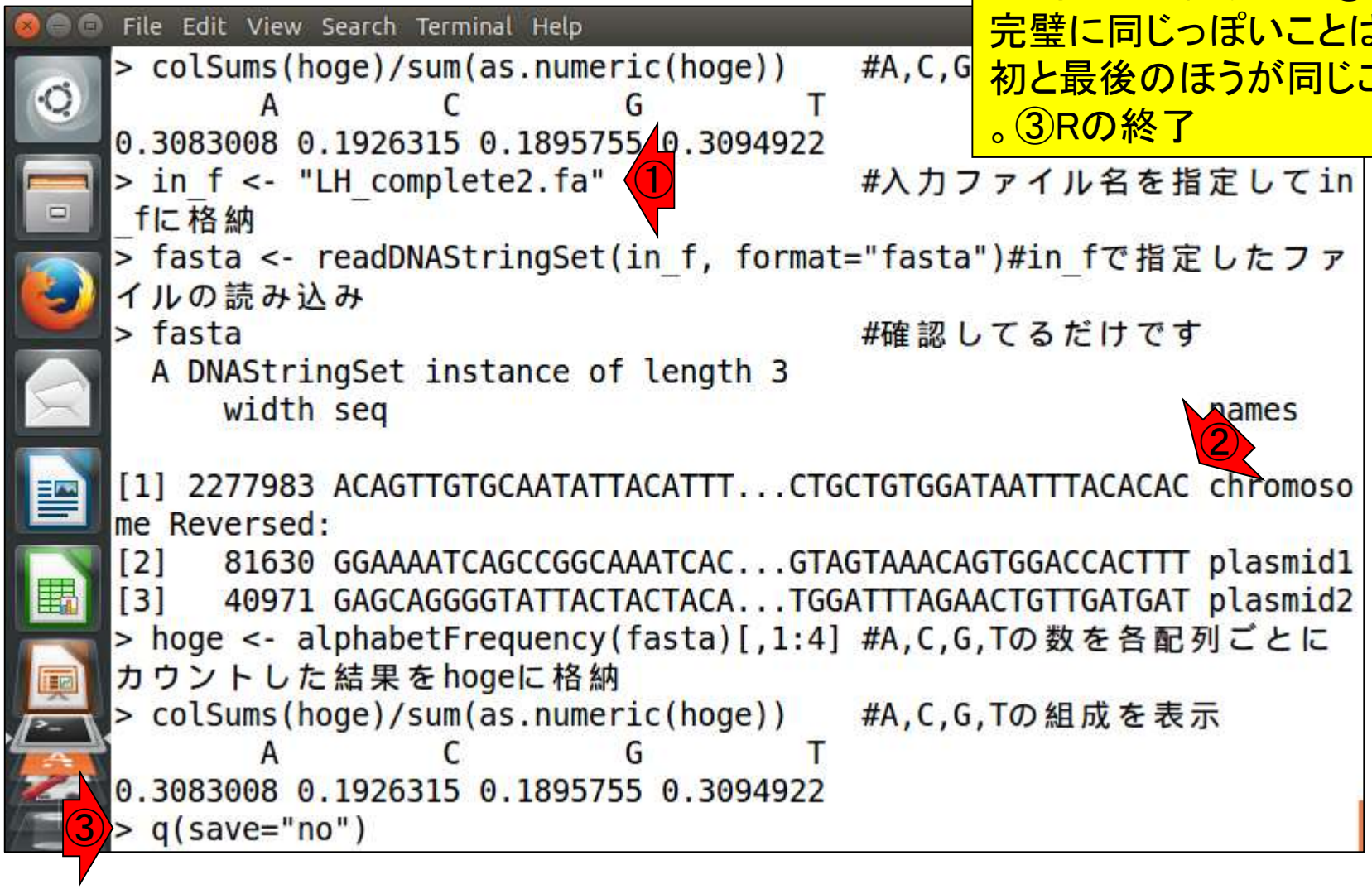
W4-9:しつこく確認

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



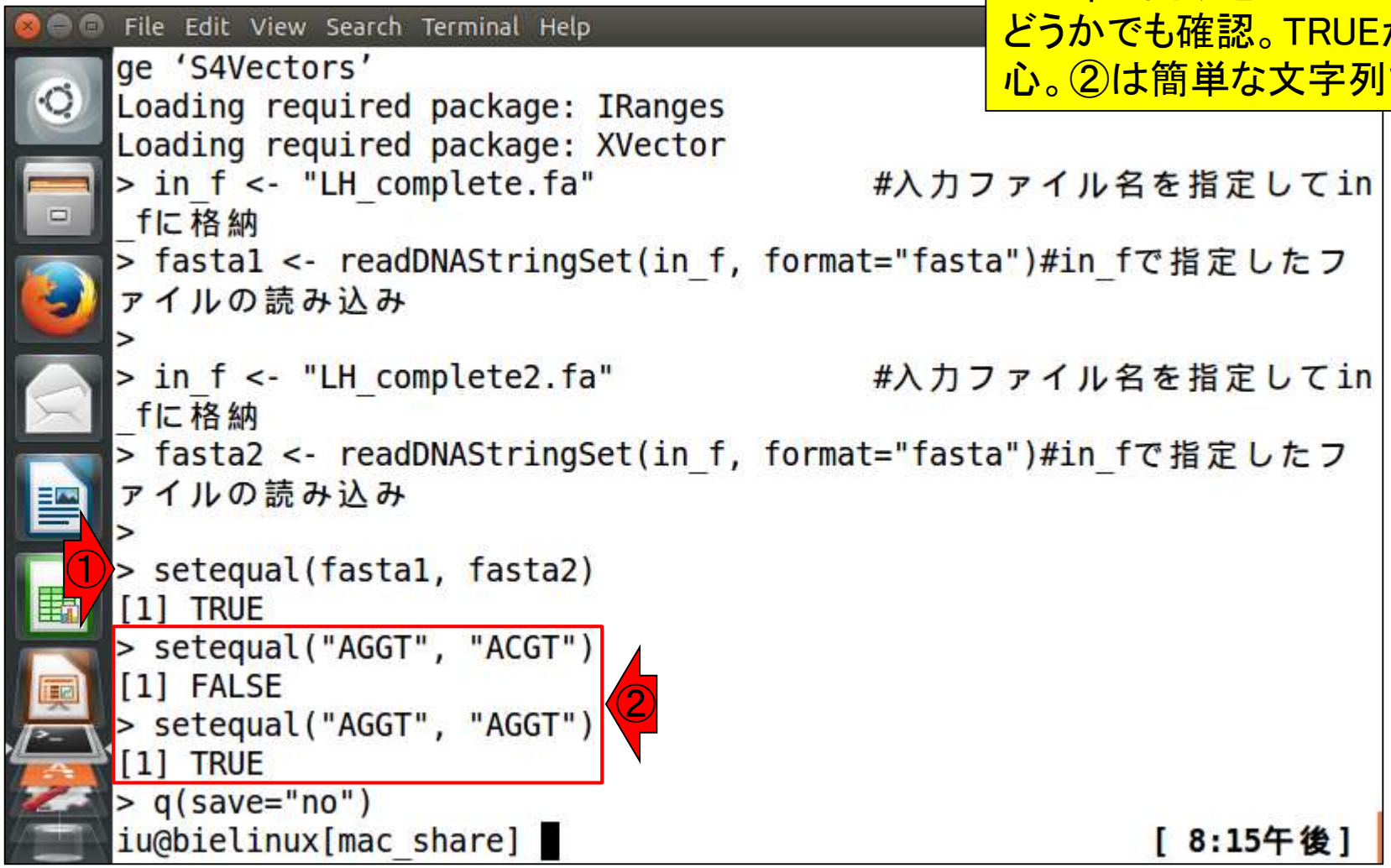
W4-9:しつこく確認

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も推測可能。③Rの終了



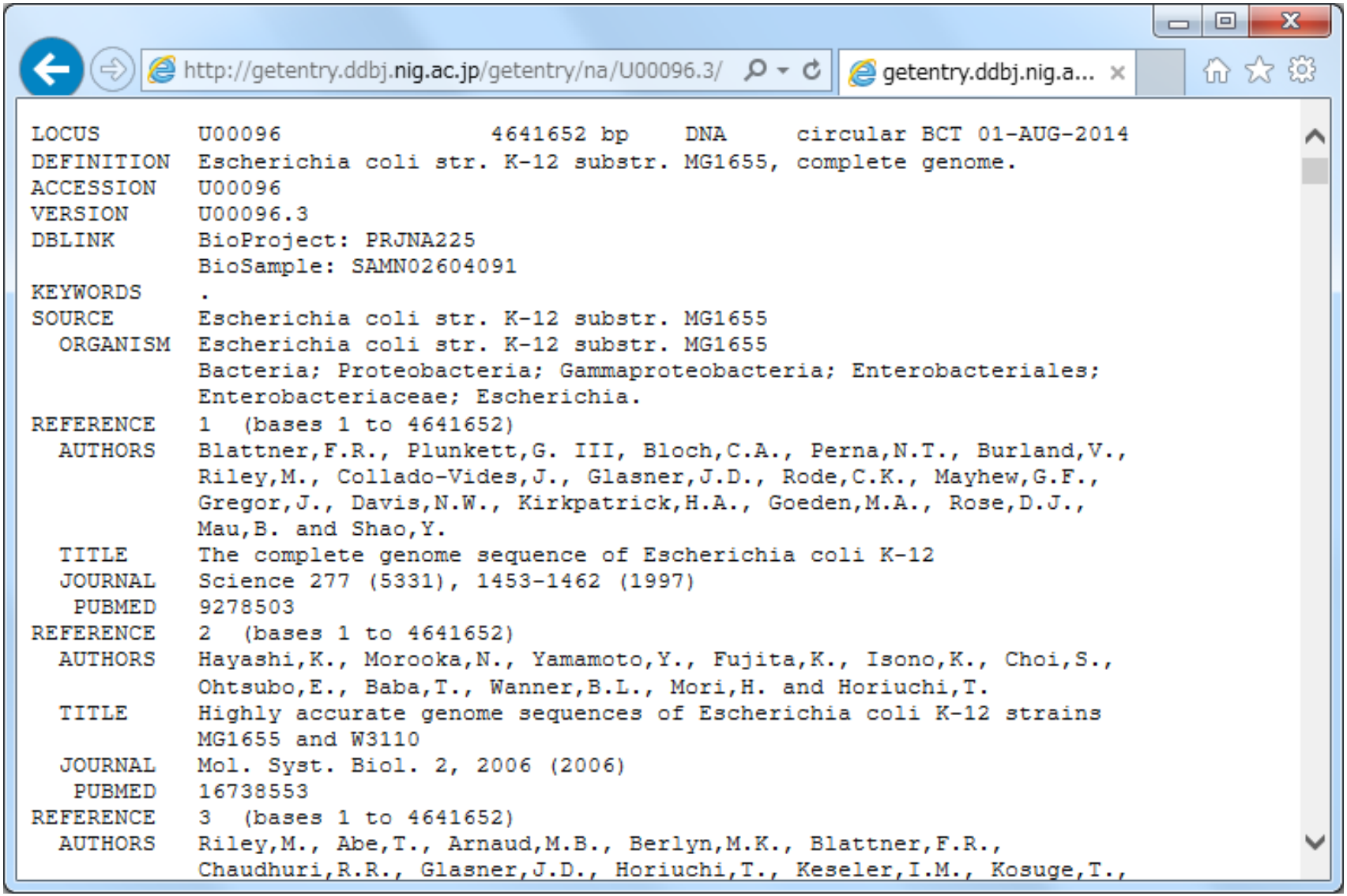
W4-10：ねちっこく確認

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



W5-1：U00096

entryの具体例1。大腸菌 K-12株の染色体配列の登録内容です




```
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.,
```

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

W5-1：U00096

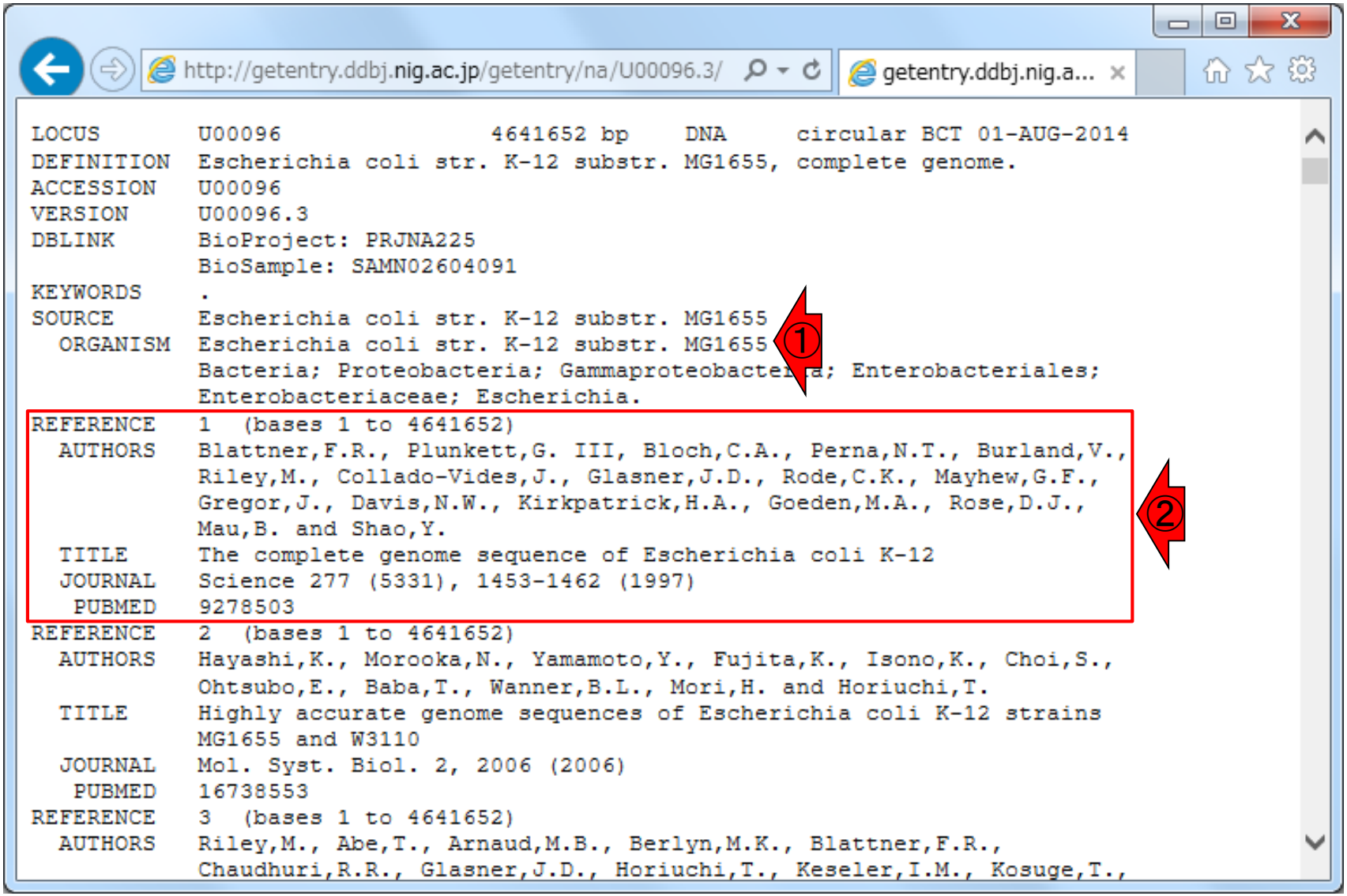
①U00096はこのentryの識別子（アクセッション番号）、②ピリオドの後の3はバージョン番号。これは最初の登録から2回更新されていることを意味する。登録直後はU00096.1、1回目の更新後にU00096.2、そして2回目の更新後にU00096.3となります

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

```
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.,
```

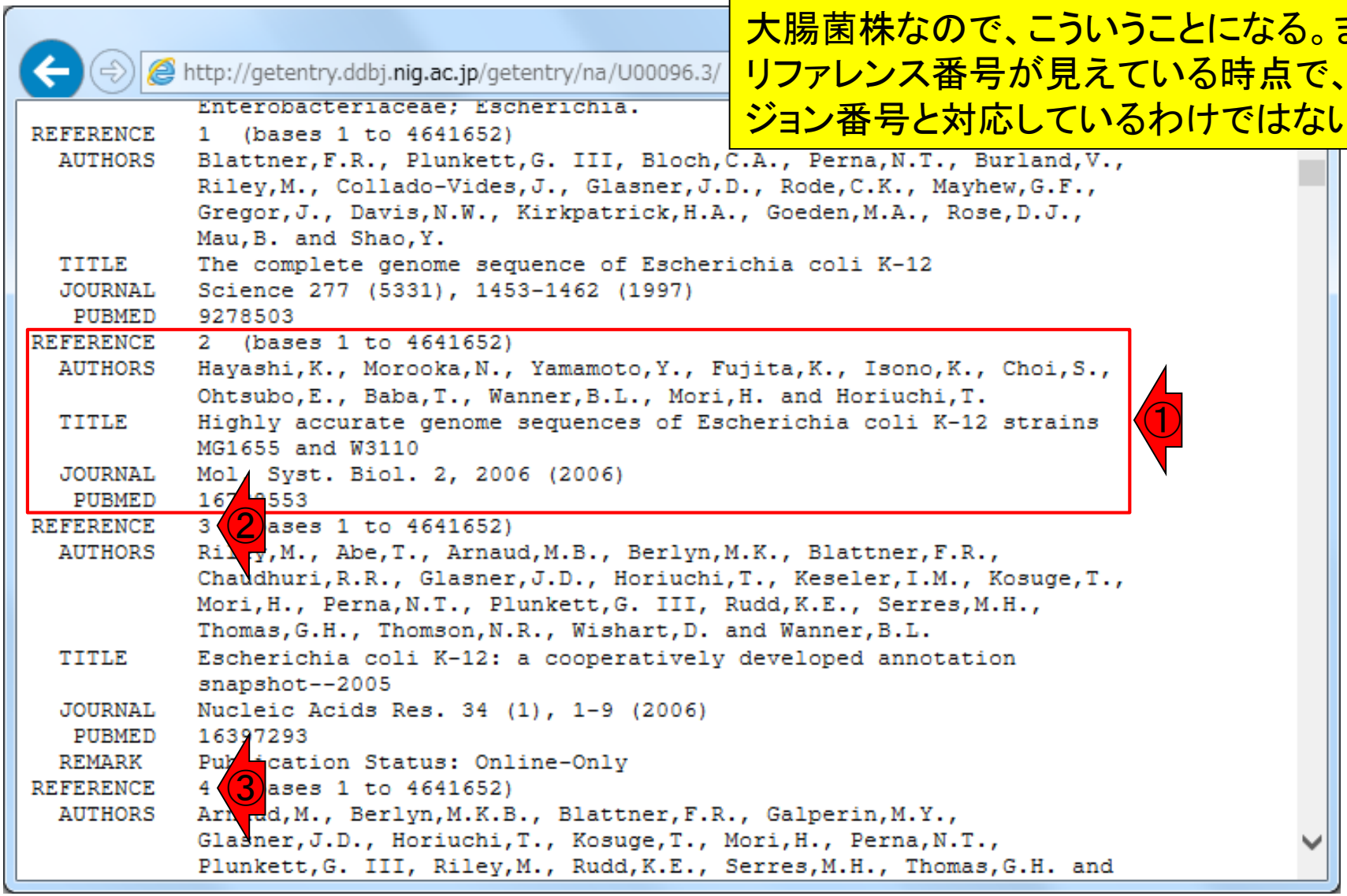
W5-1：U00096

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



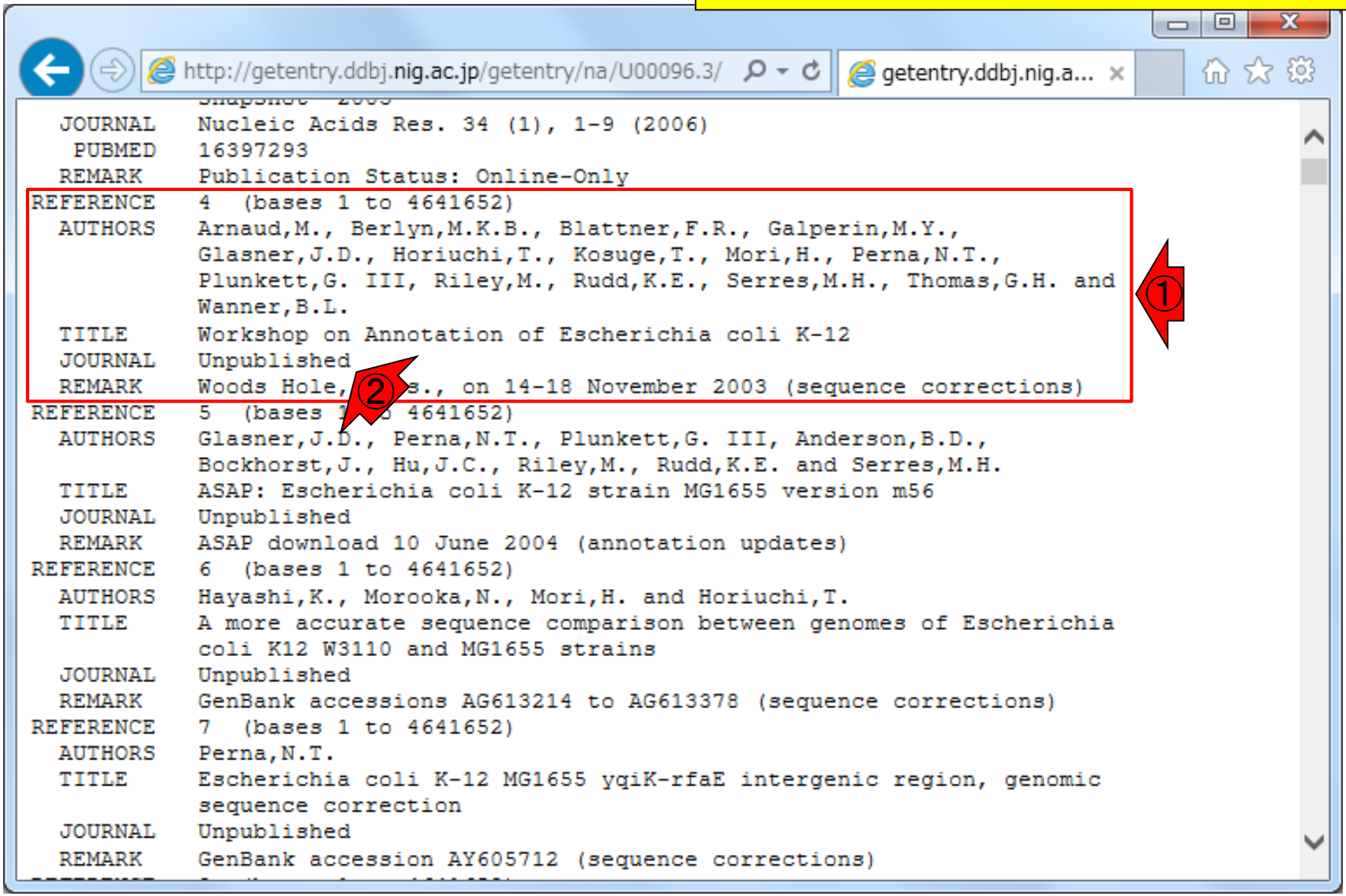
W5-1：U00096

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



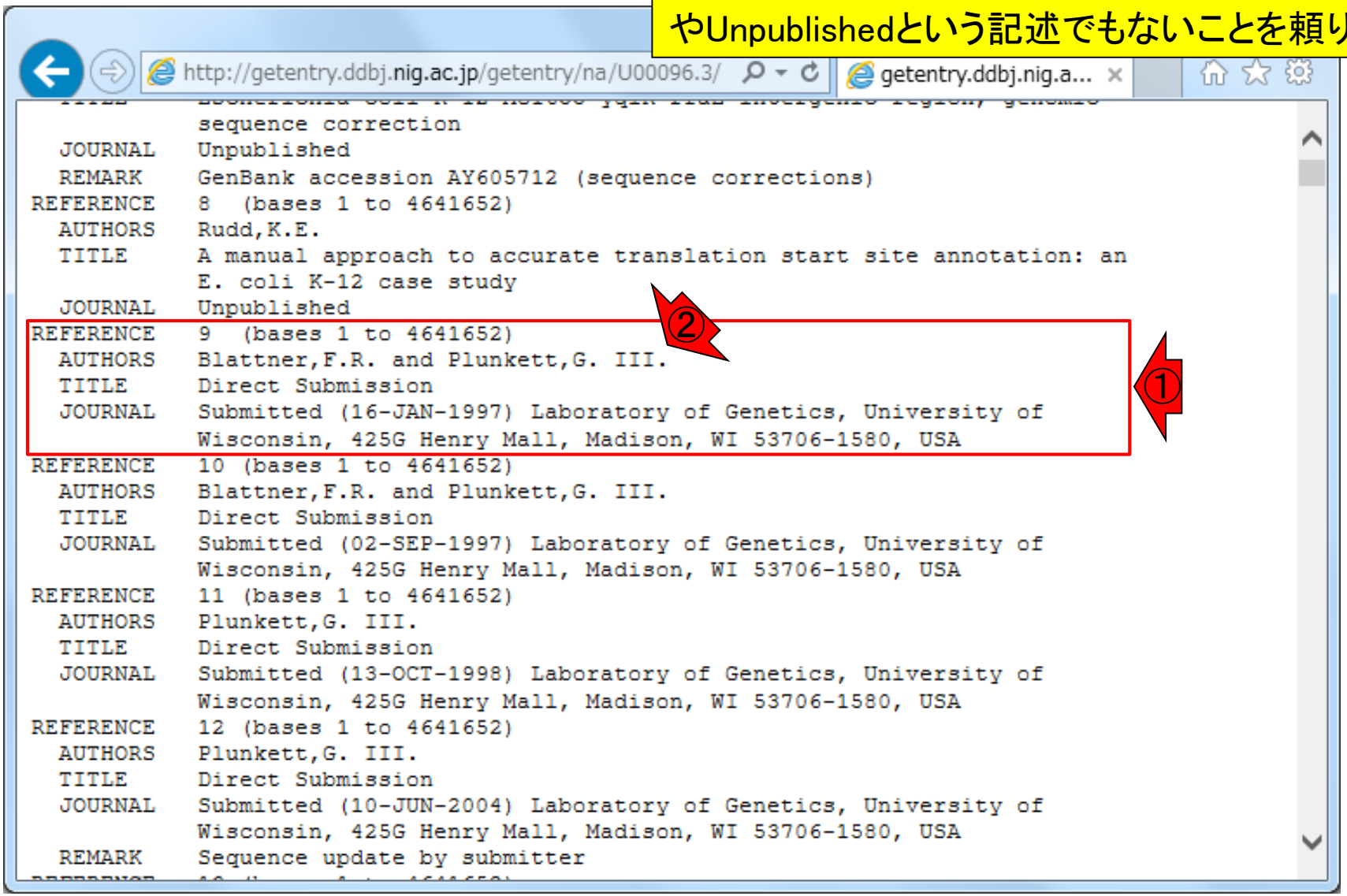
W5-1：U00096

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）



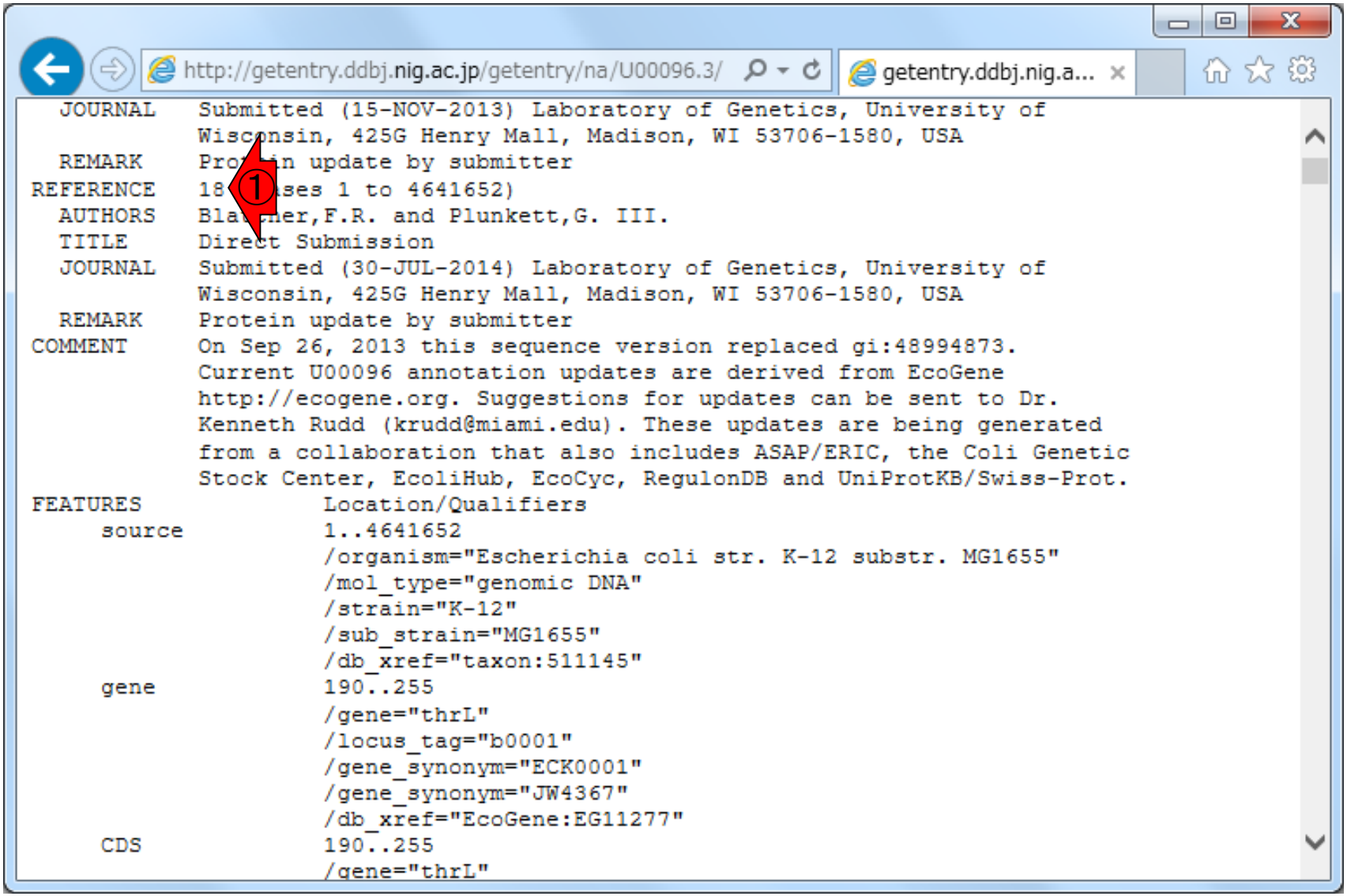
W5-2：U00096

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のところが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



W5-2：U00096

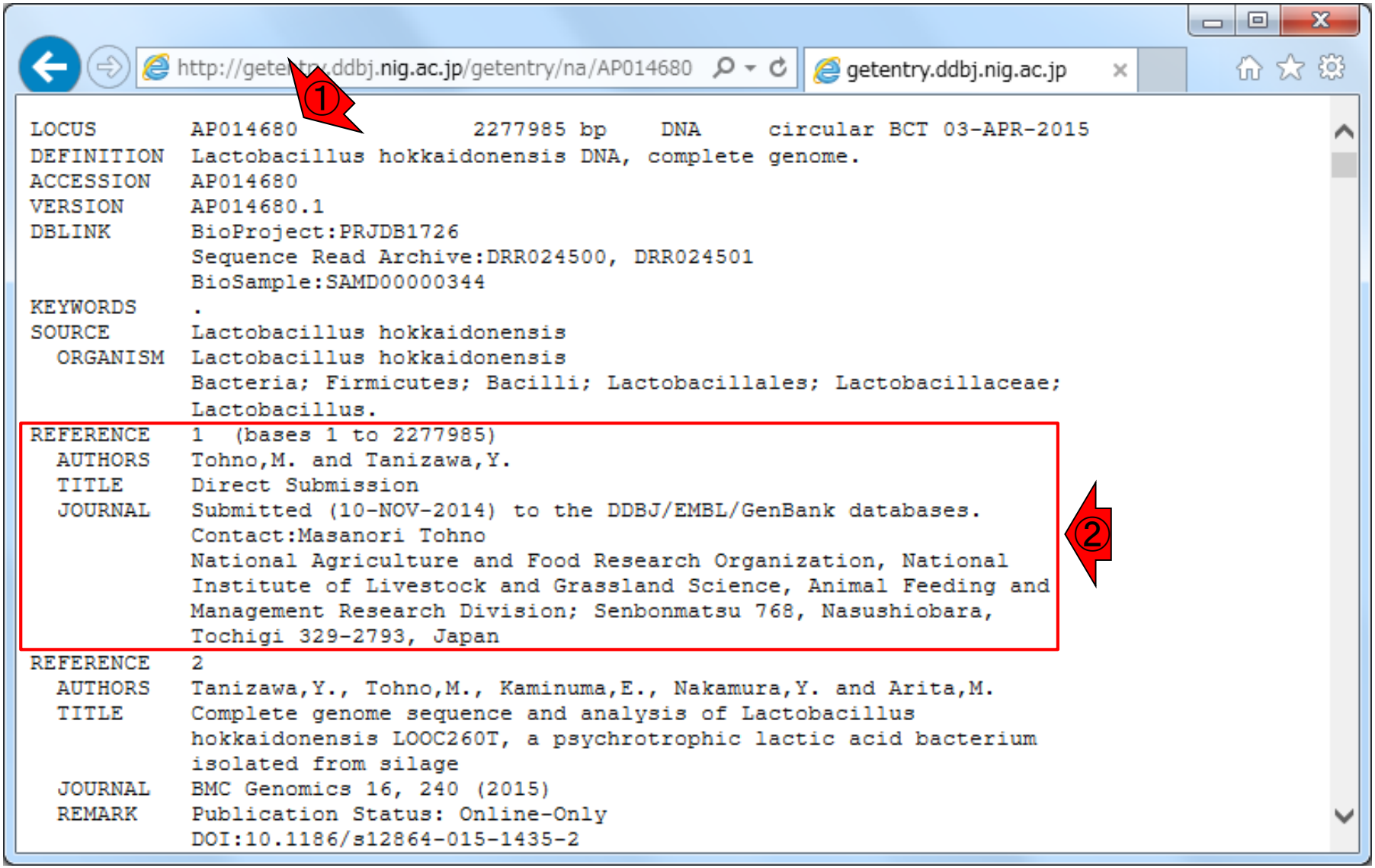
この場合は①REFERENCE 18まであるが、これは相当多いほうです




```
  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"
```

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

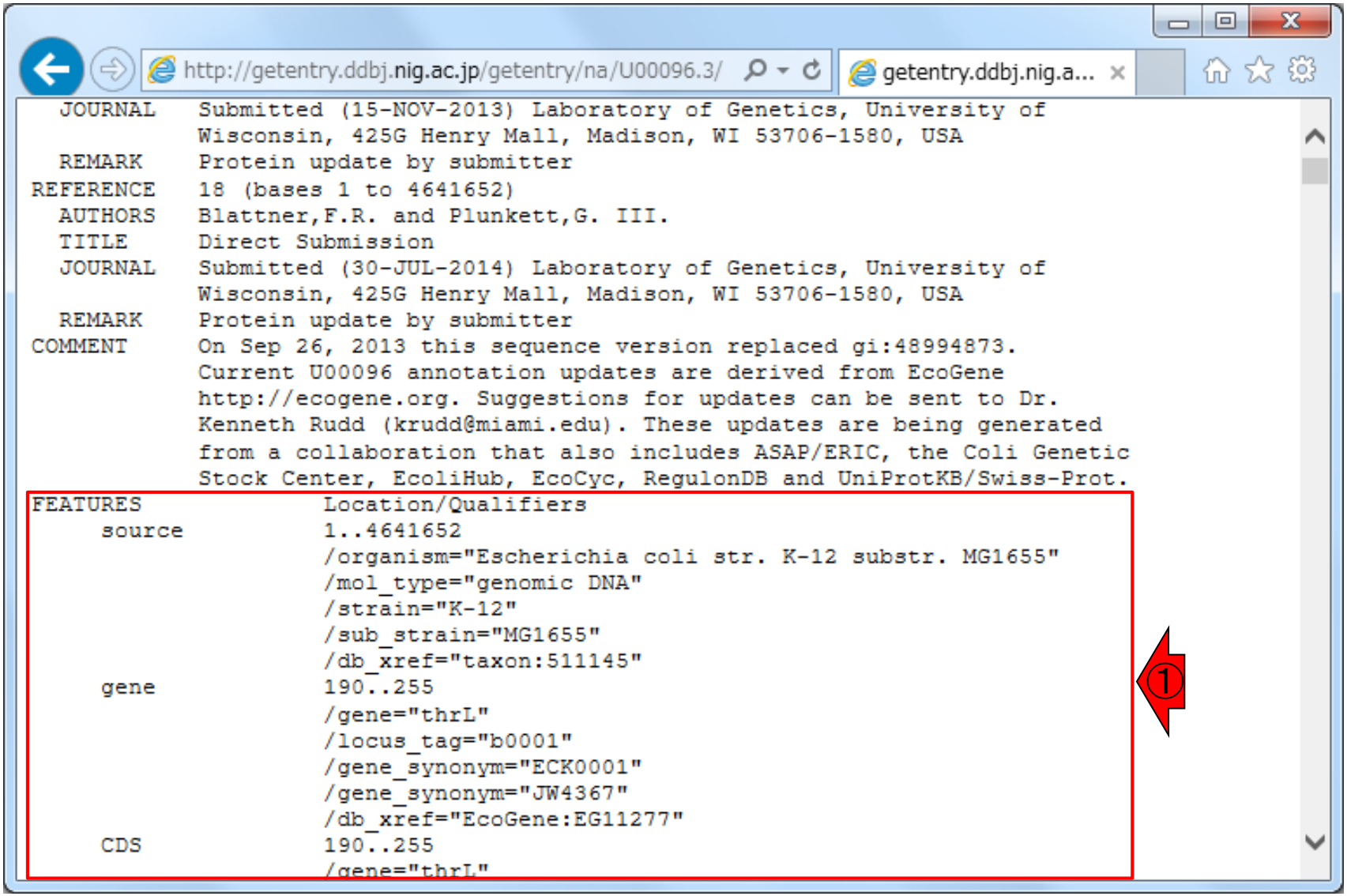
詳細についてはこちら

W5-2-2：REFERENCE 1

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

W5-3：FEATURES

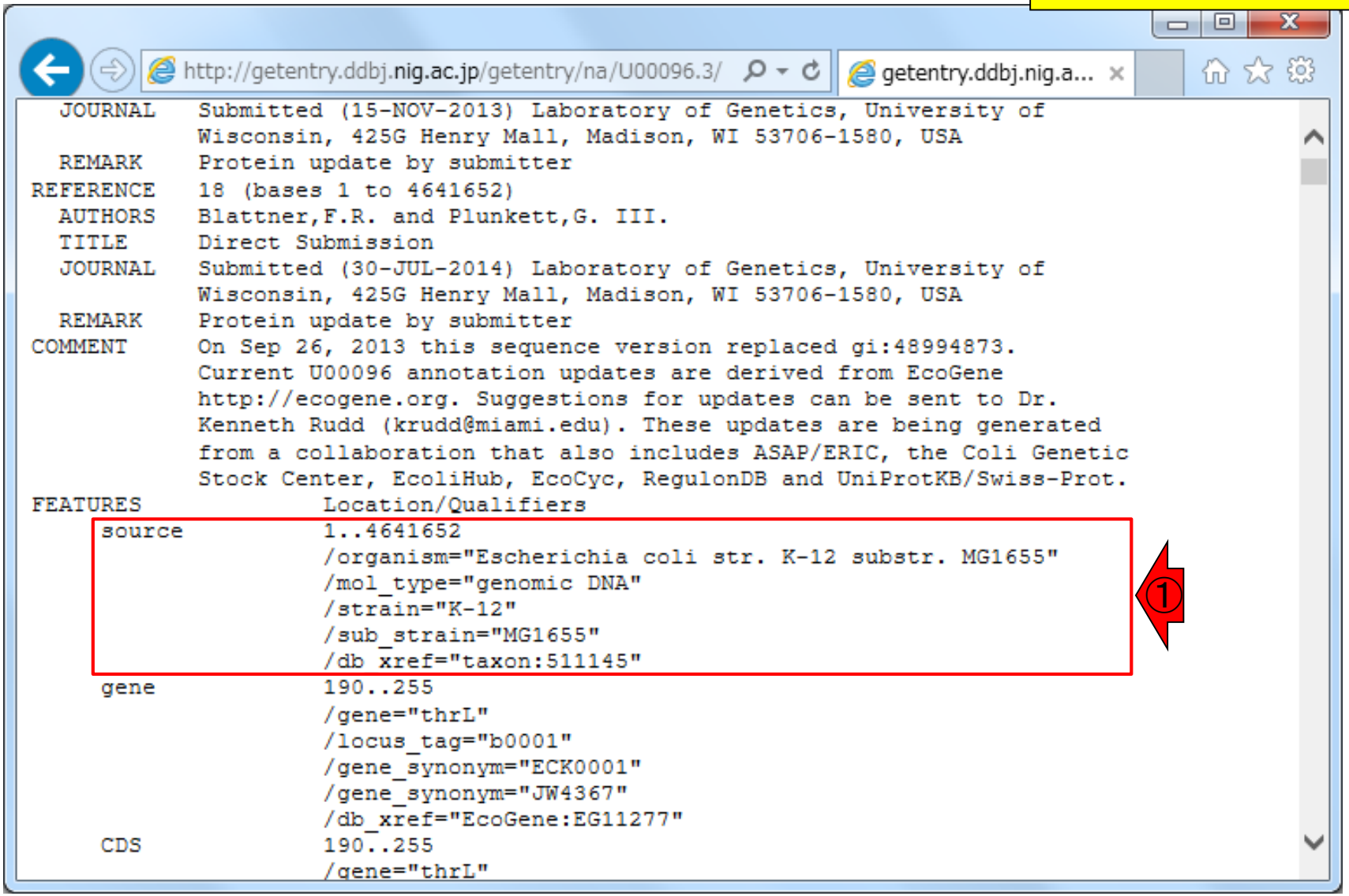
再びU00096に戻り、、、続いて①FEATURESの階層



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

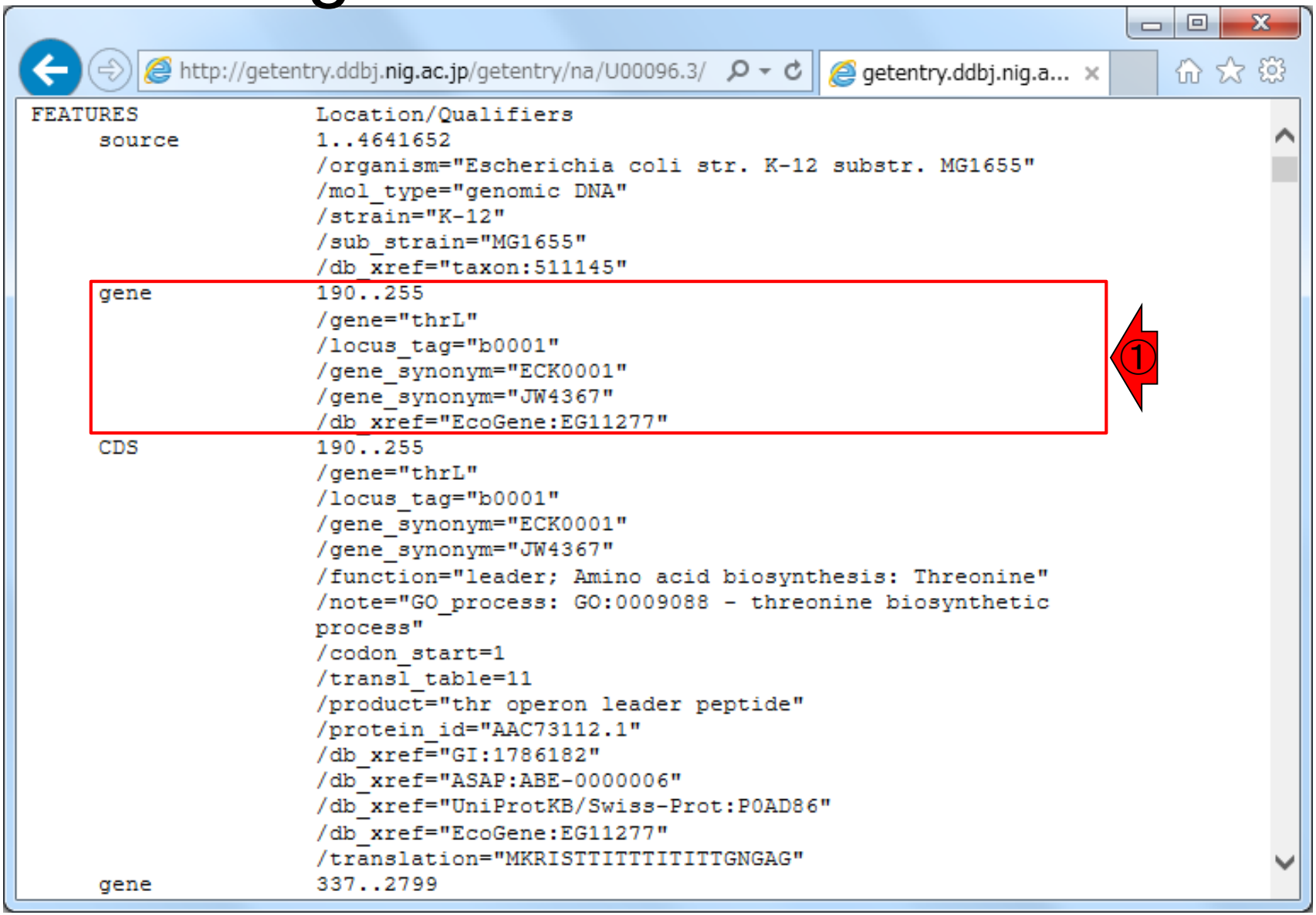
W5-4：source feature

①source 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



W5-5：gene feature

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

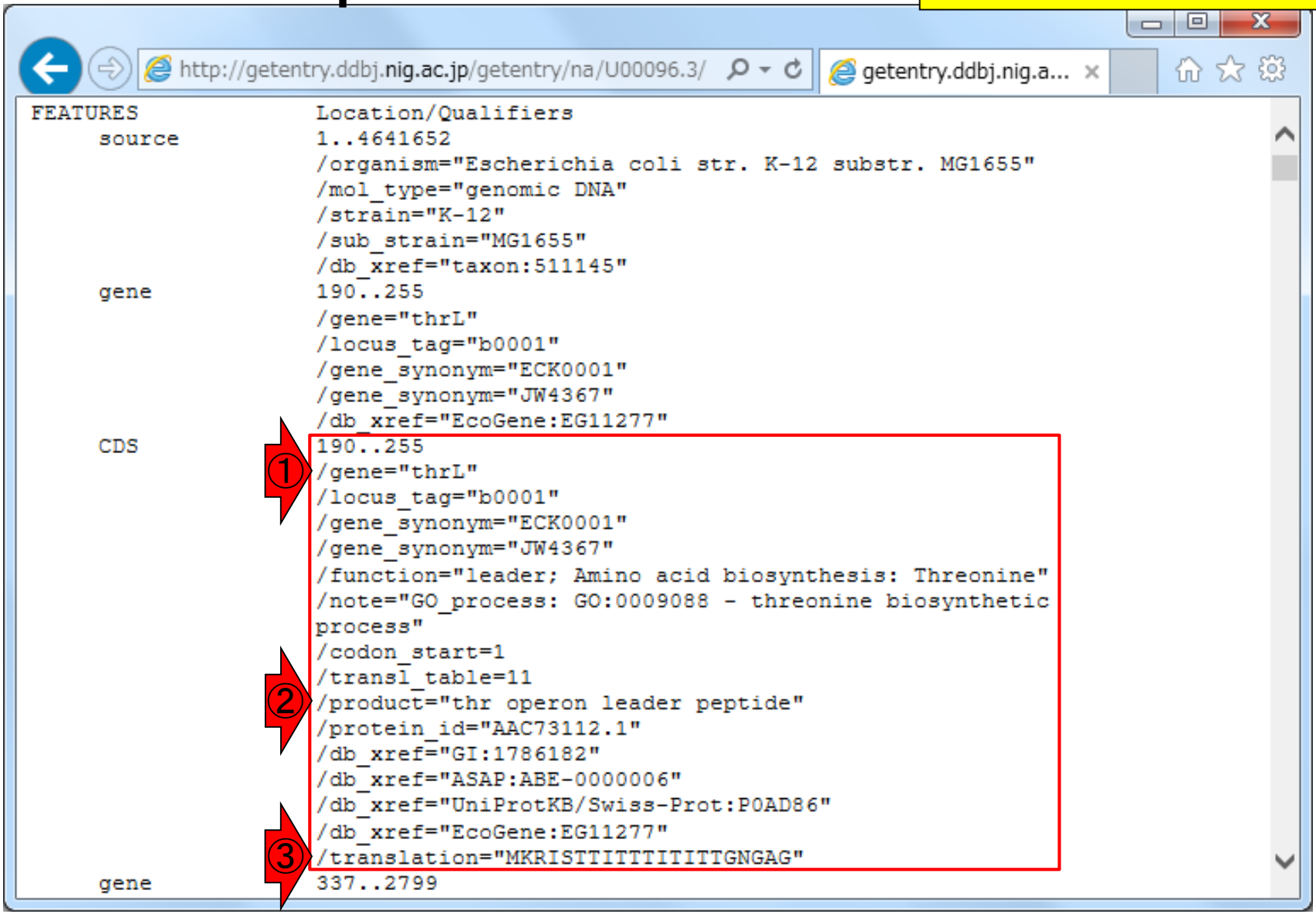


W5-5：CDS feature

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

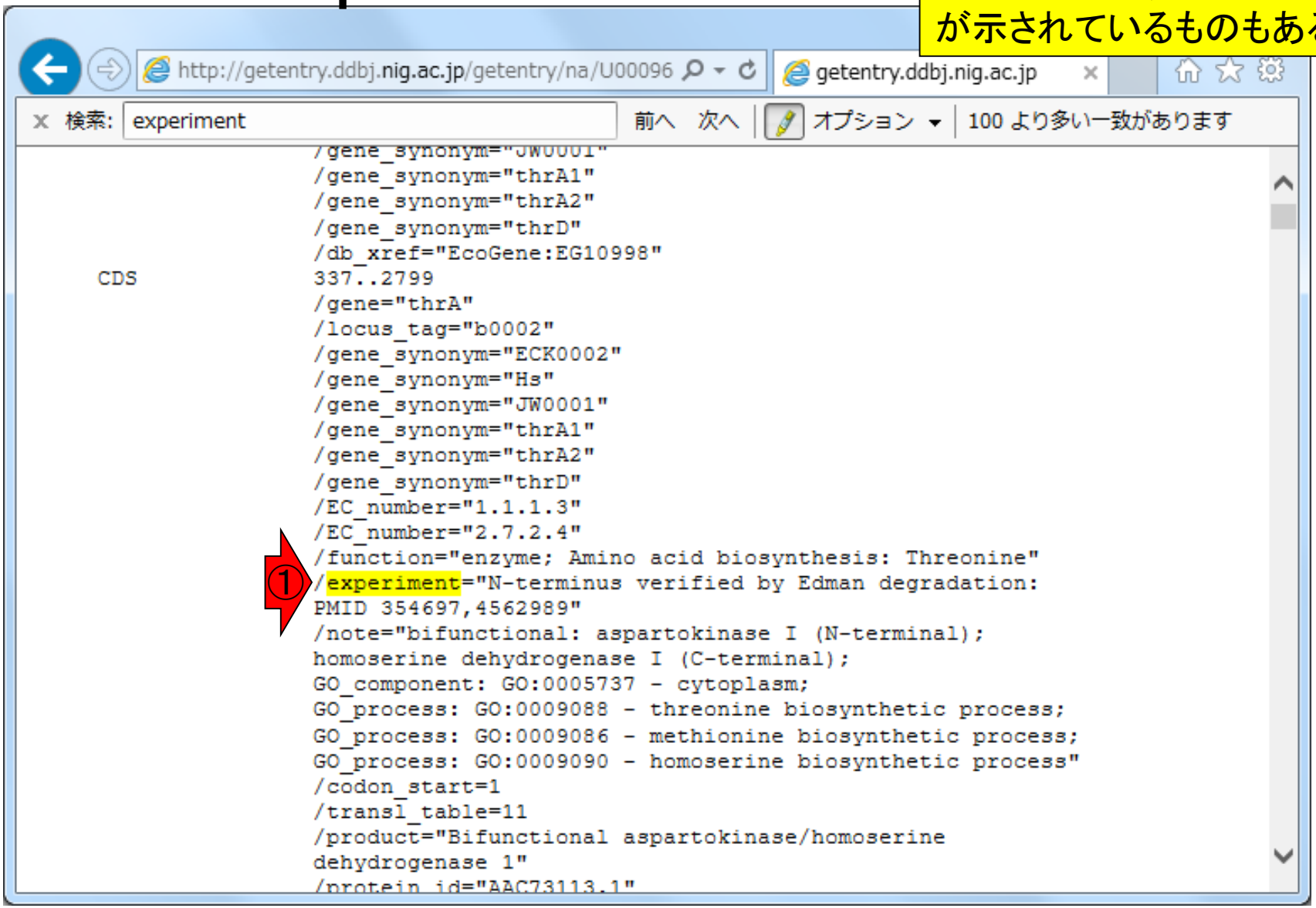
W5-6：qualifier

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



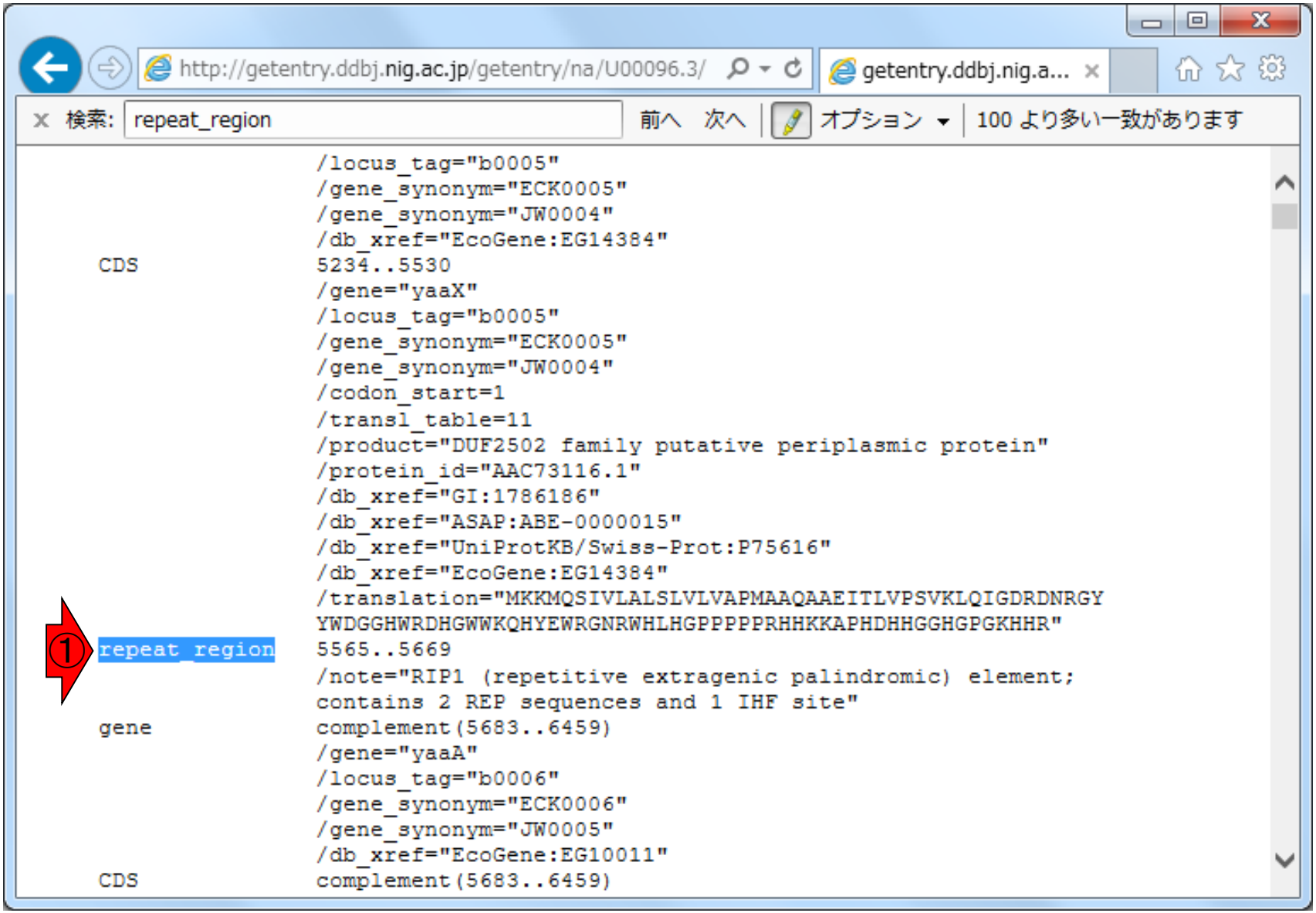
W5-7 : qualifier

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



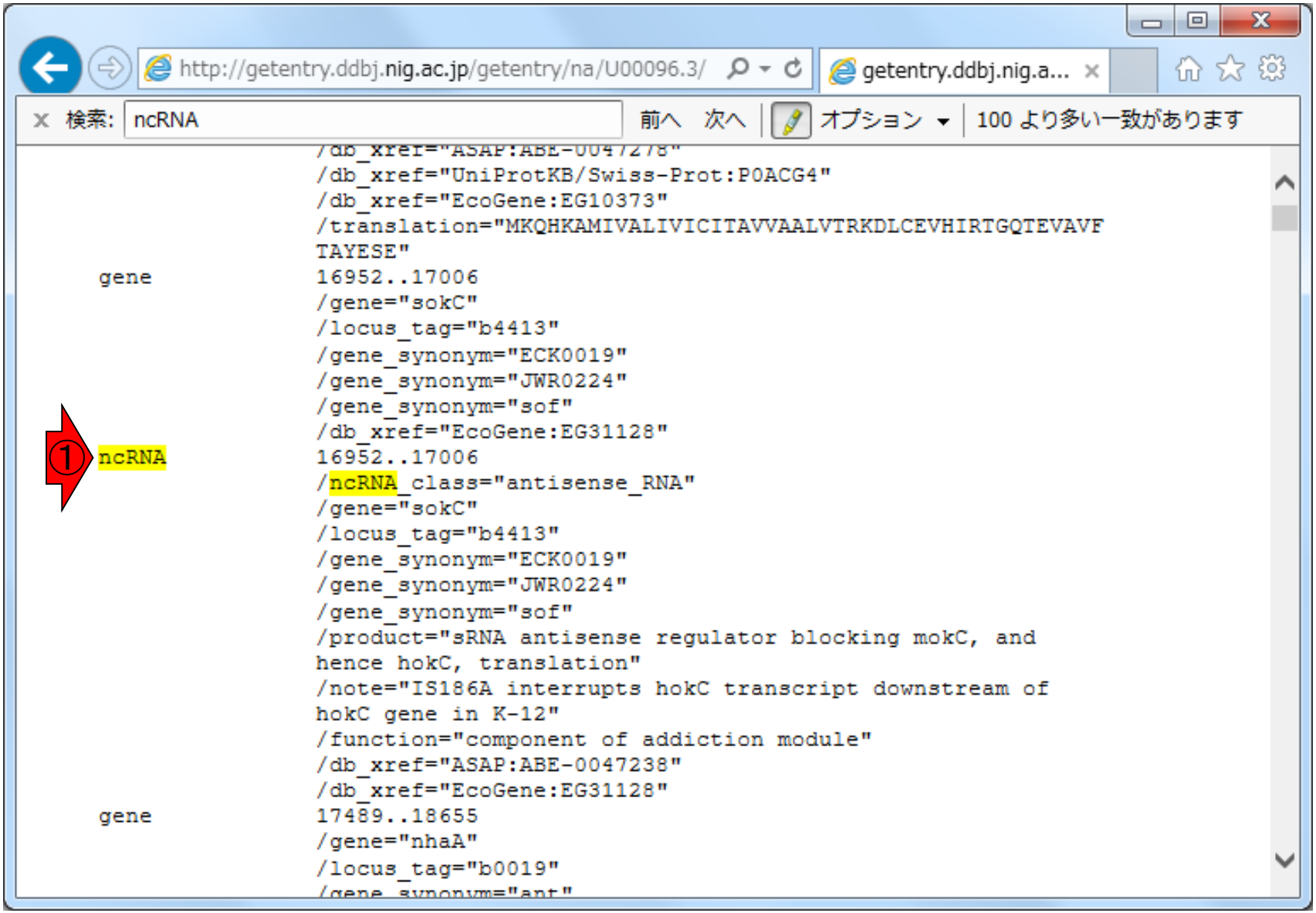
W5-8：featureの種類

gene featureやCDS feature以外にも、①repeat_region featureや…



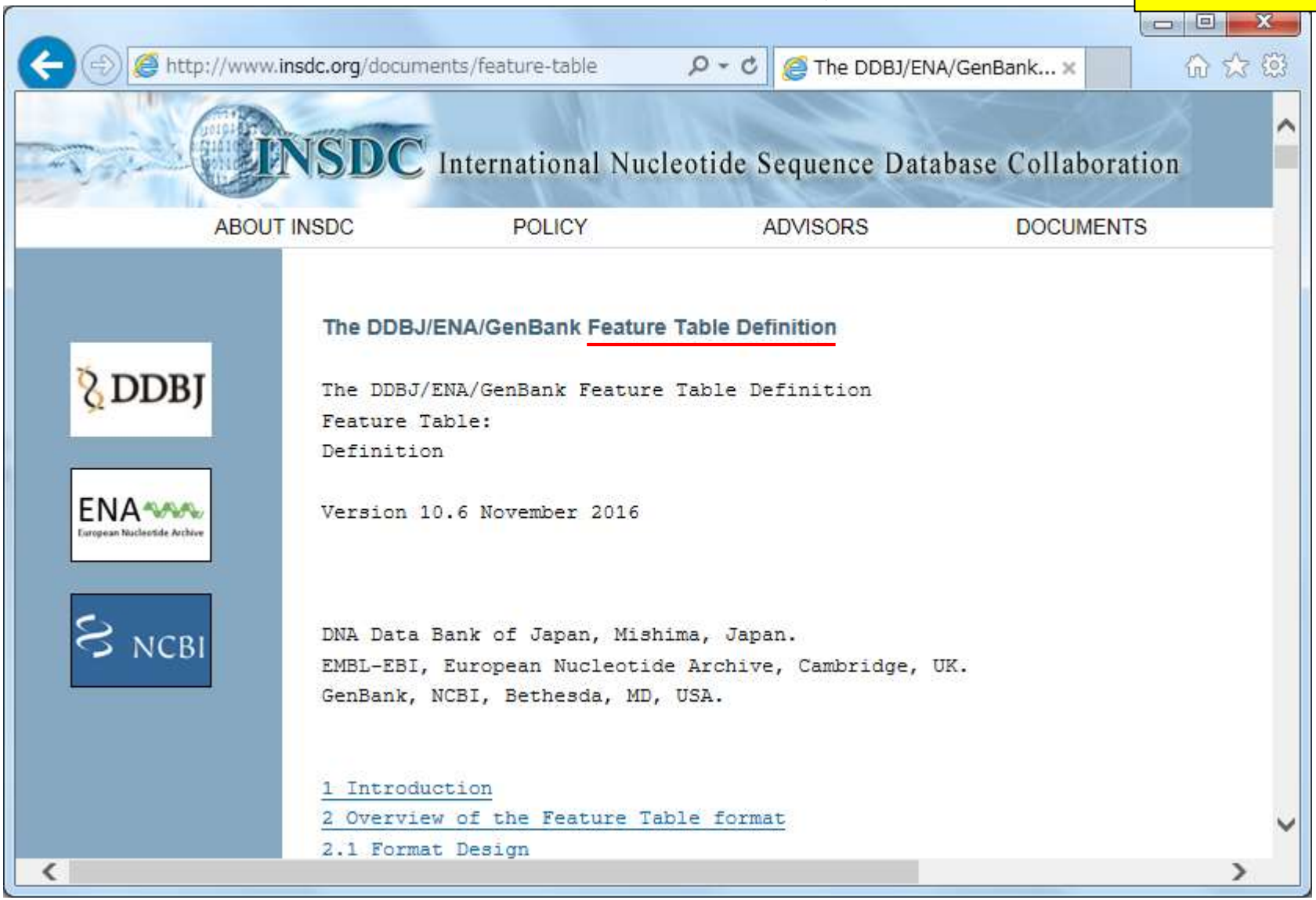
W5-9：featureの種類

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



W5-10：featureの記載方法

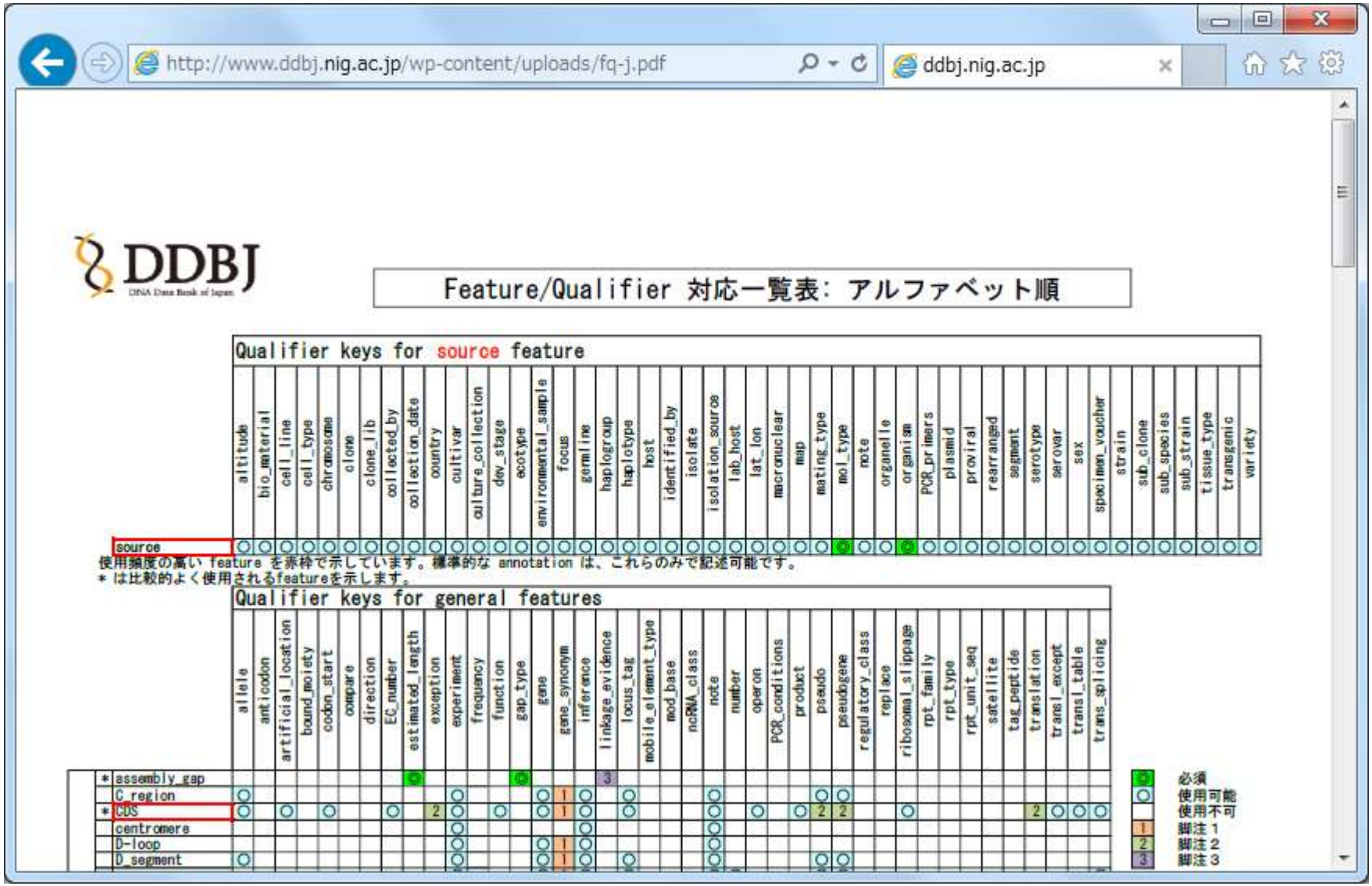
各featureの記載方法は、INSDCのFeature Table Definitionで定められている



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

W5-11：対応表

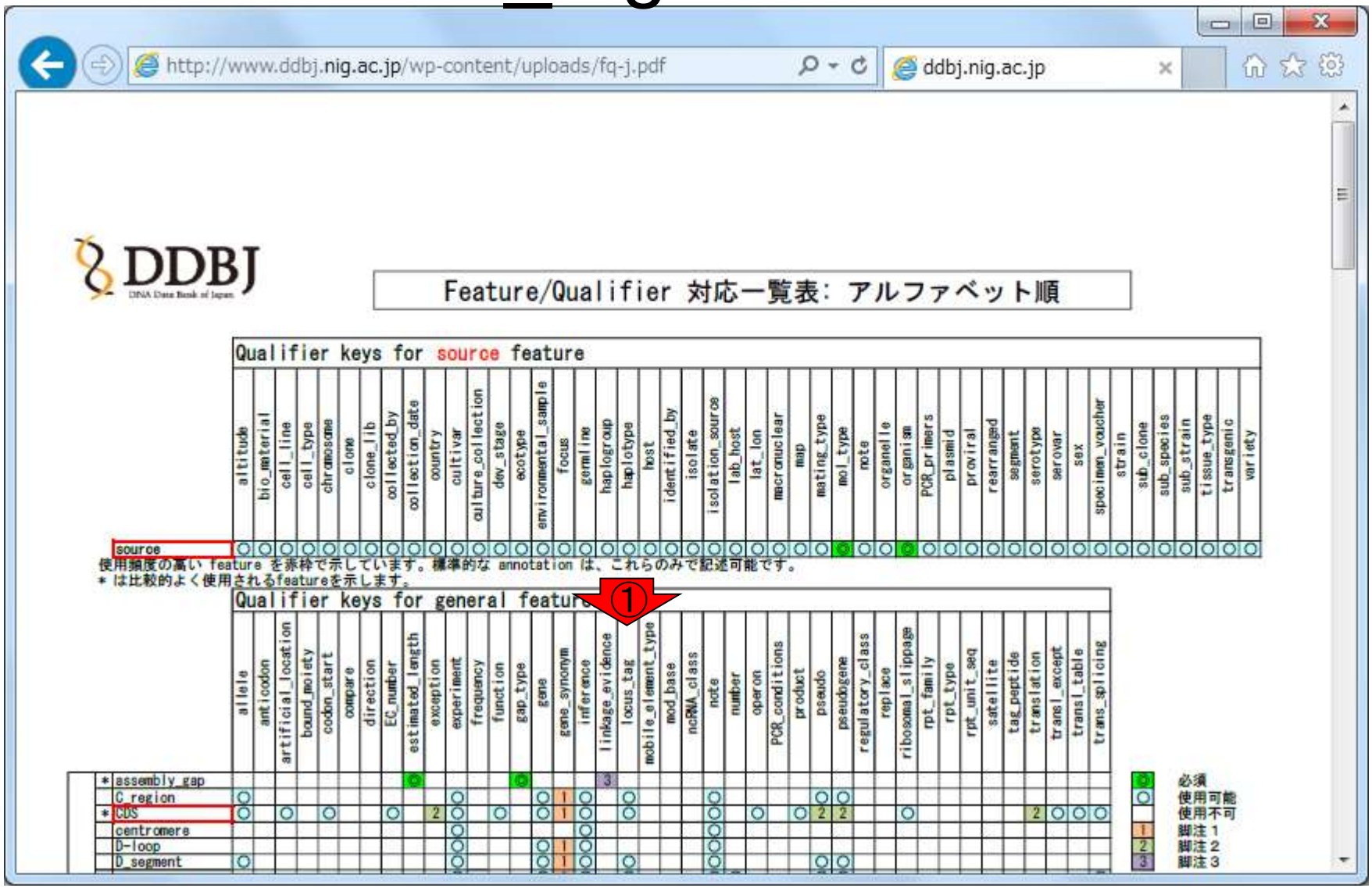
DDBJが公開している、featureごとに利用可能なqualifierの対応表



http://www.ddbj.nig.ac.jp/wp-content/uploads/fq-j.pdf

W6-1：locus_tag

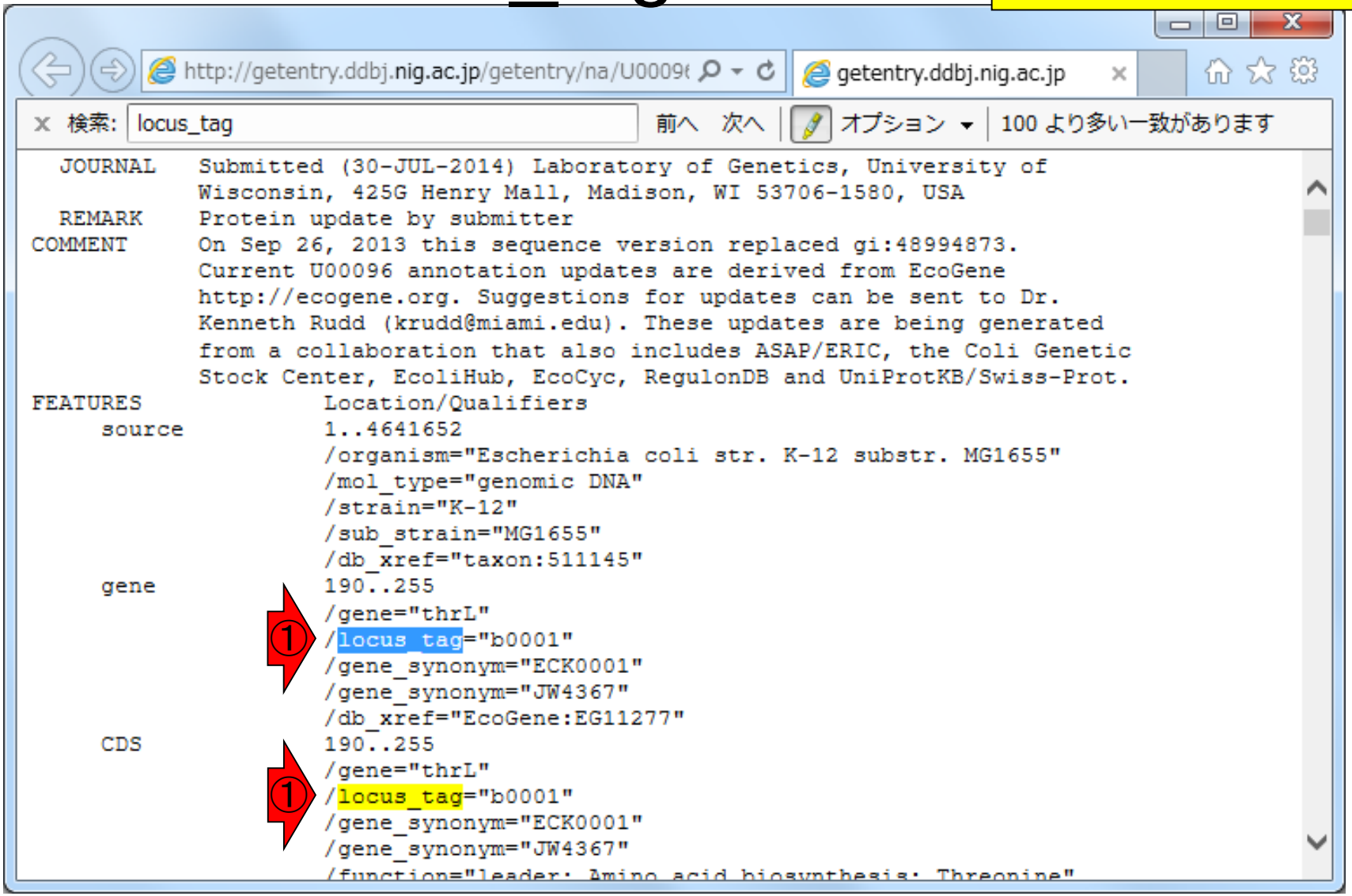
①locus_tagというqualifierは、アノテーションされた遺伝子領域を指し示すための一意な識別子



http://www.ddbj.nig.ac.jp/wp-content/uploads/fq-j.pdf

W6-2：locus_tag

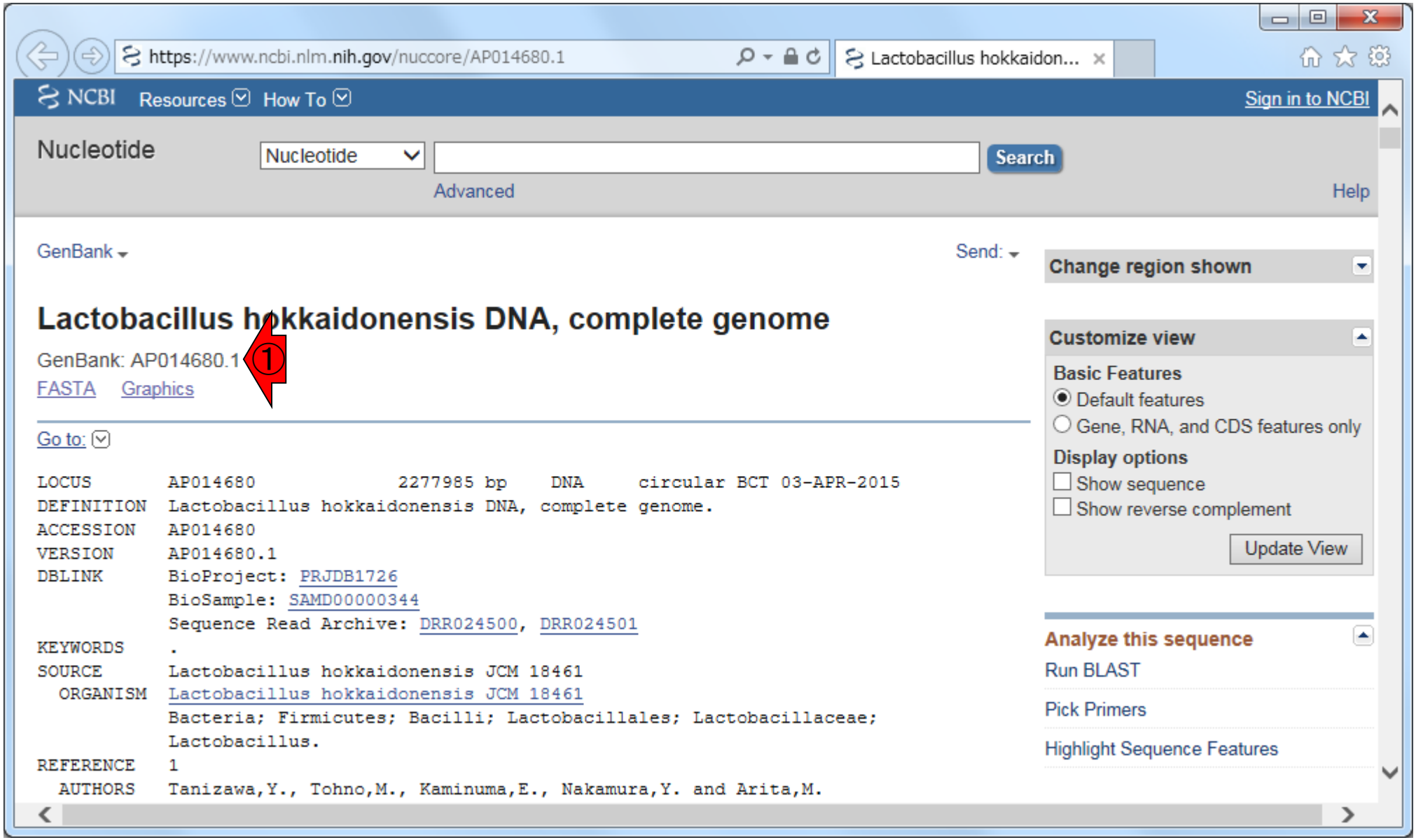
①U00096(W5-3と同じものです)上で見られるlocus_tag情報は昔の記述法であり、現在のものとは異なるので注意



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



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、①LOOC260_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

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

```
  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;
```

① ① ①

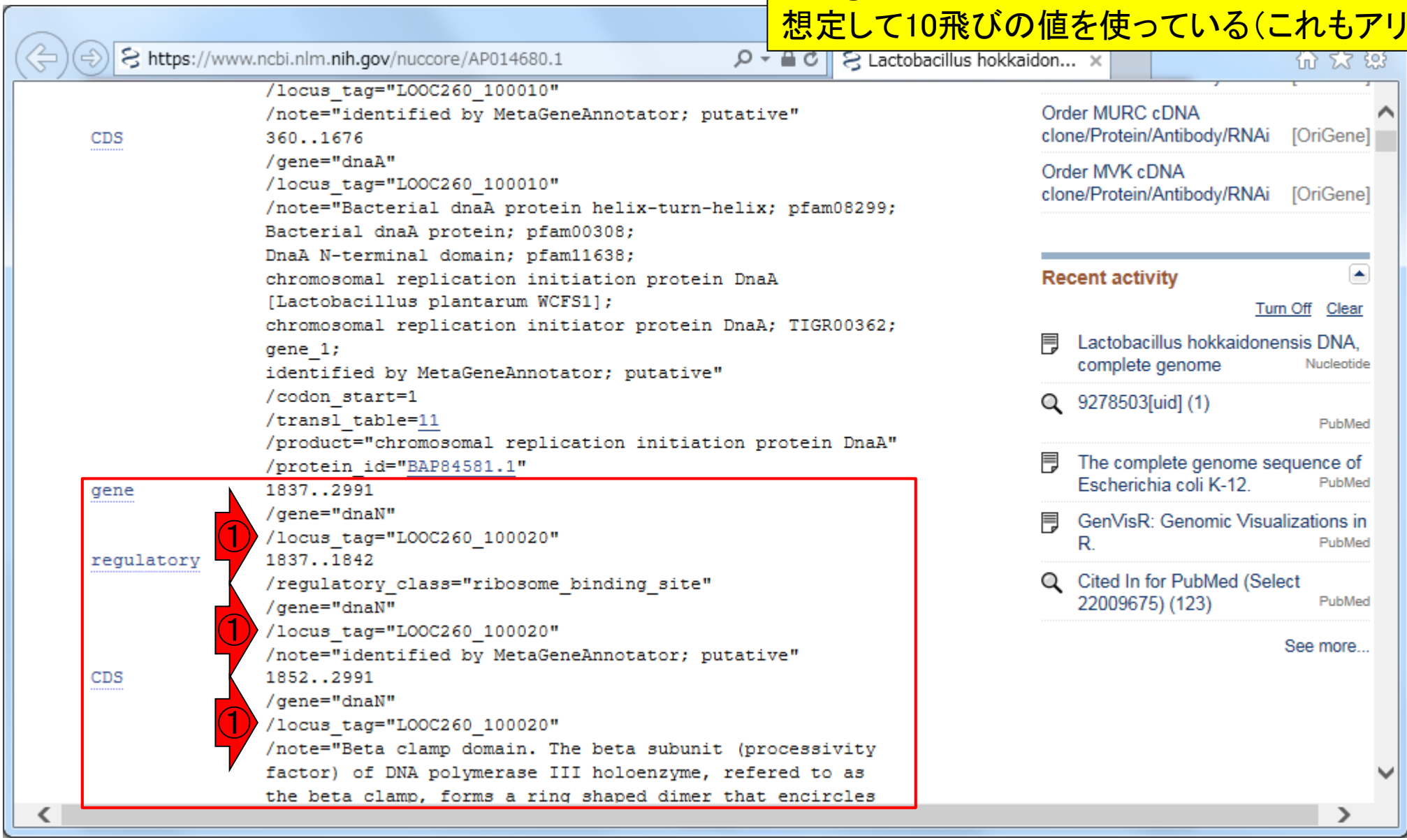
Genome
Identical RefSeq
Protein
PubMed
PubMed (Weighted)
Taxonomy

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）

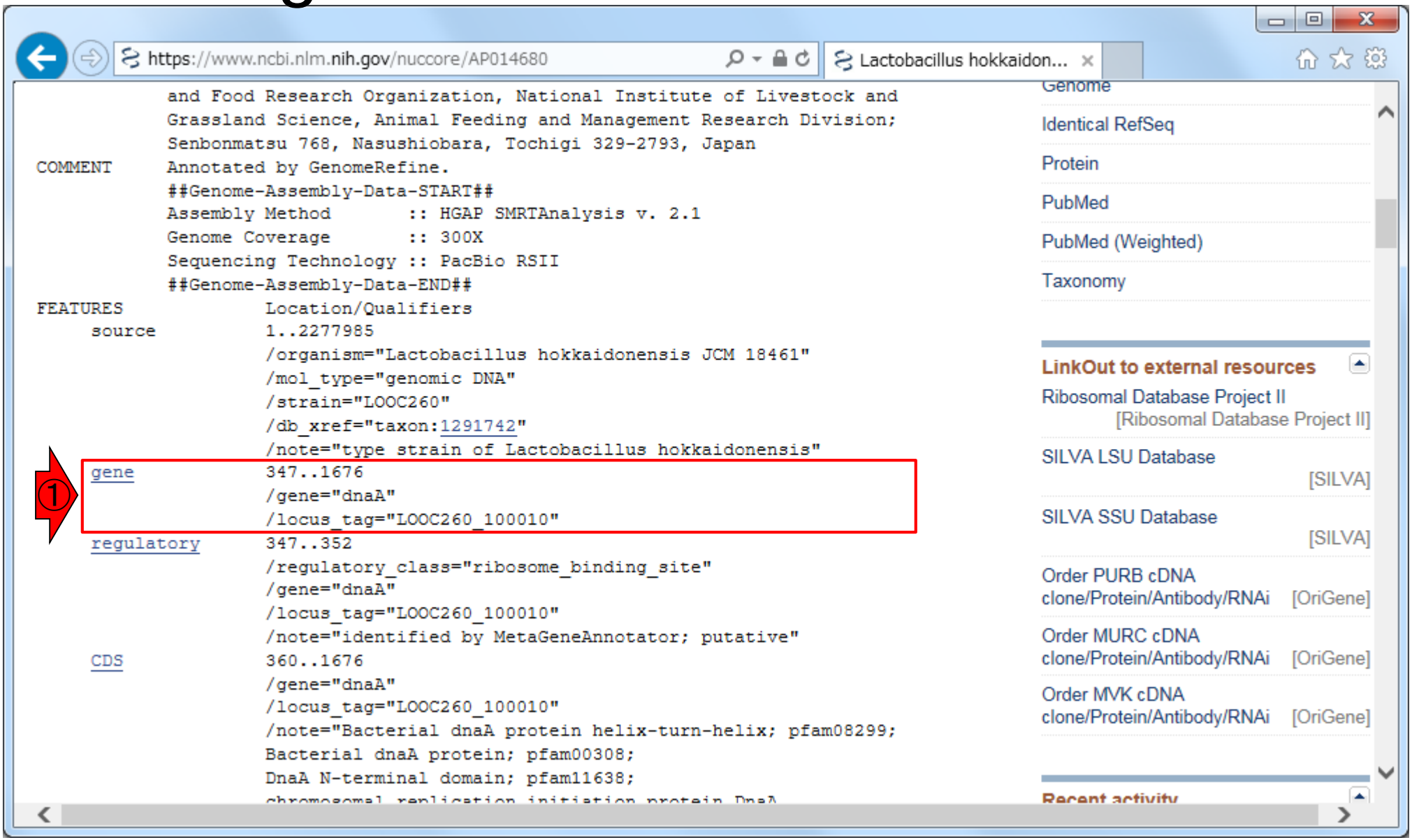


W6-5：locus_tag公式情報

DDBJ上で公開されている
locus_tag qualifierに関する情報

W6-6：gene feature

GenBank形式では、①gene featureが存在するが…



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

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

getentry.ddbj.nig.ac.jp

```
  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"
```

EMBL形式でも、①gene featureは存在しない

W6-6：gene feature

http://www.ebi.ac.uk/ena/data/view/AP014680&display=text

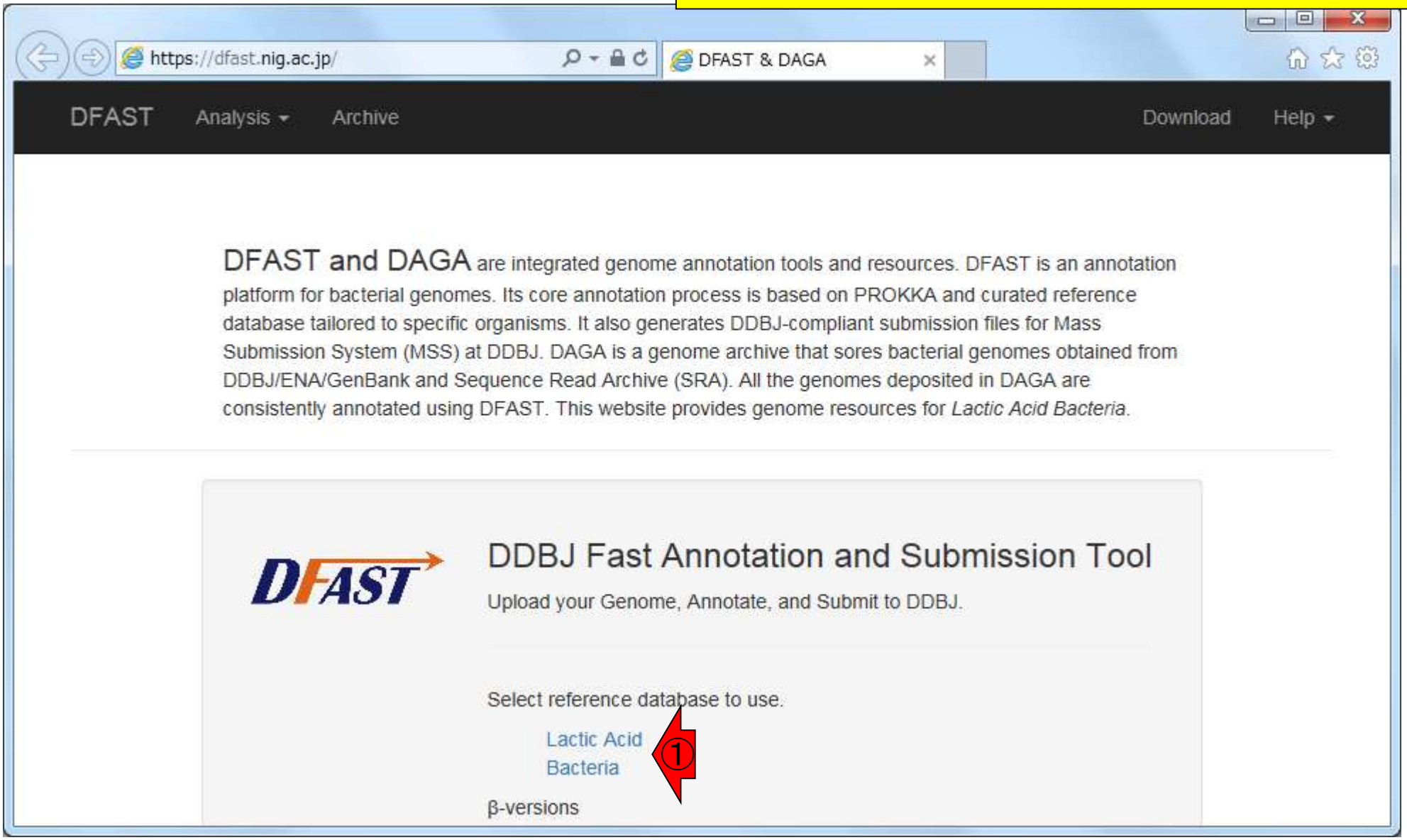
ebi.ac.uk

```
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"
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]"
```

①

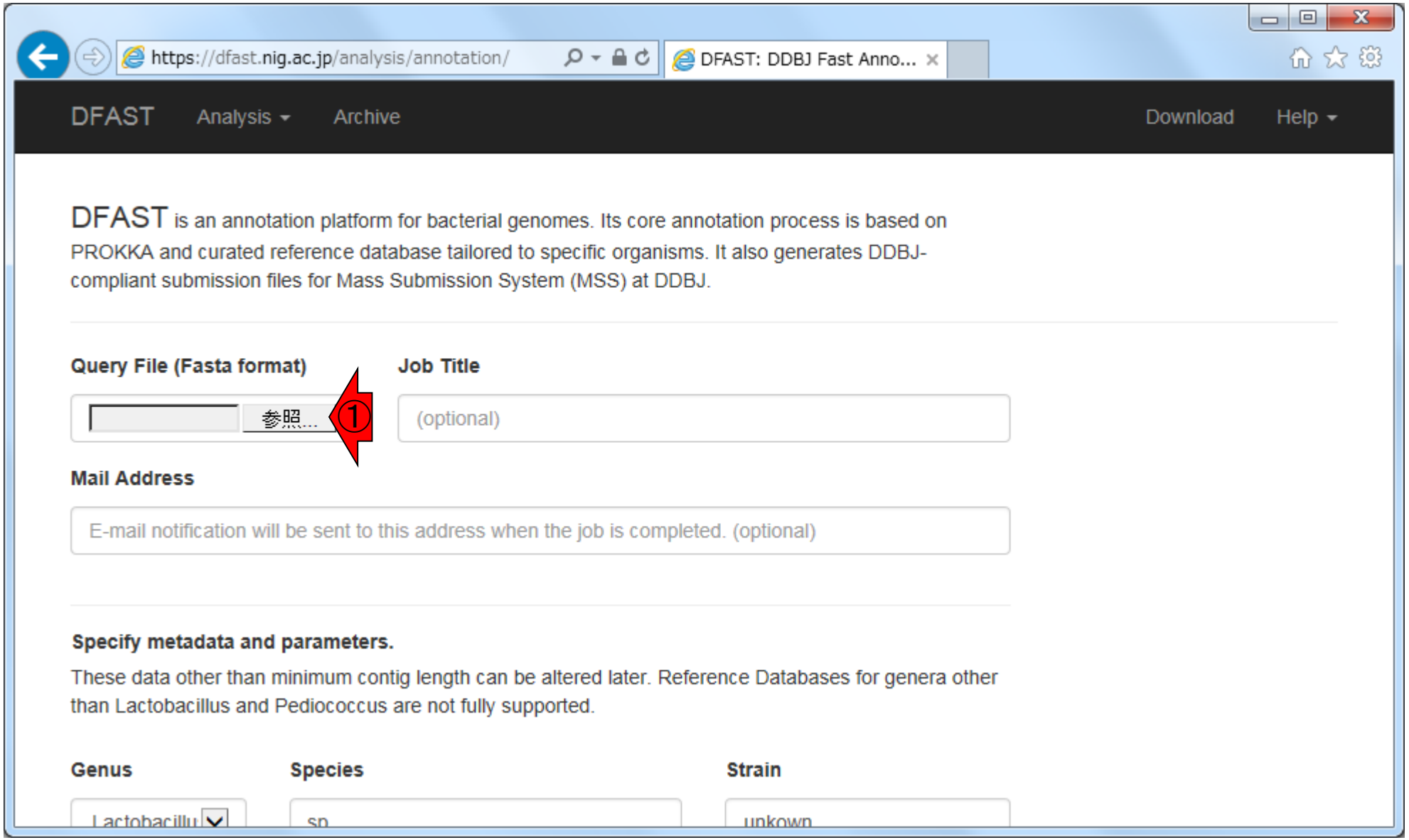
W7-1：DFAST

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



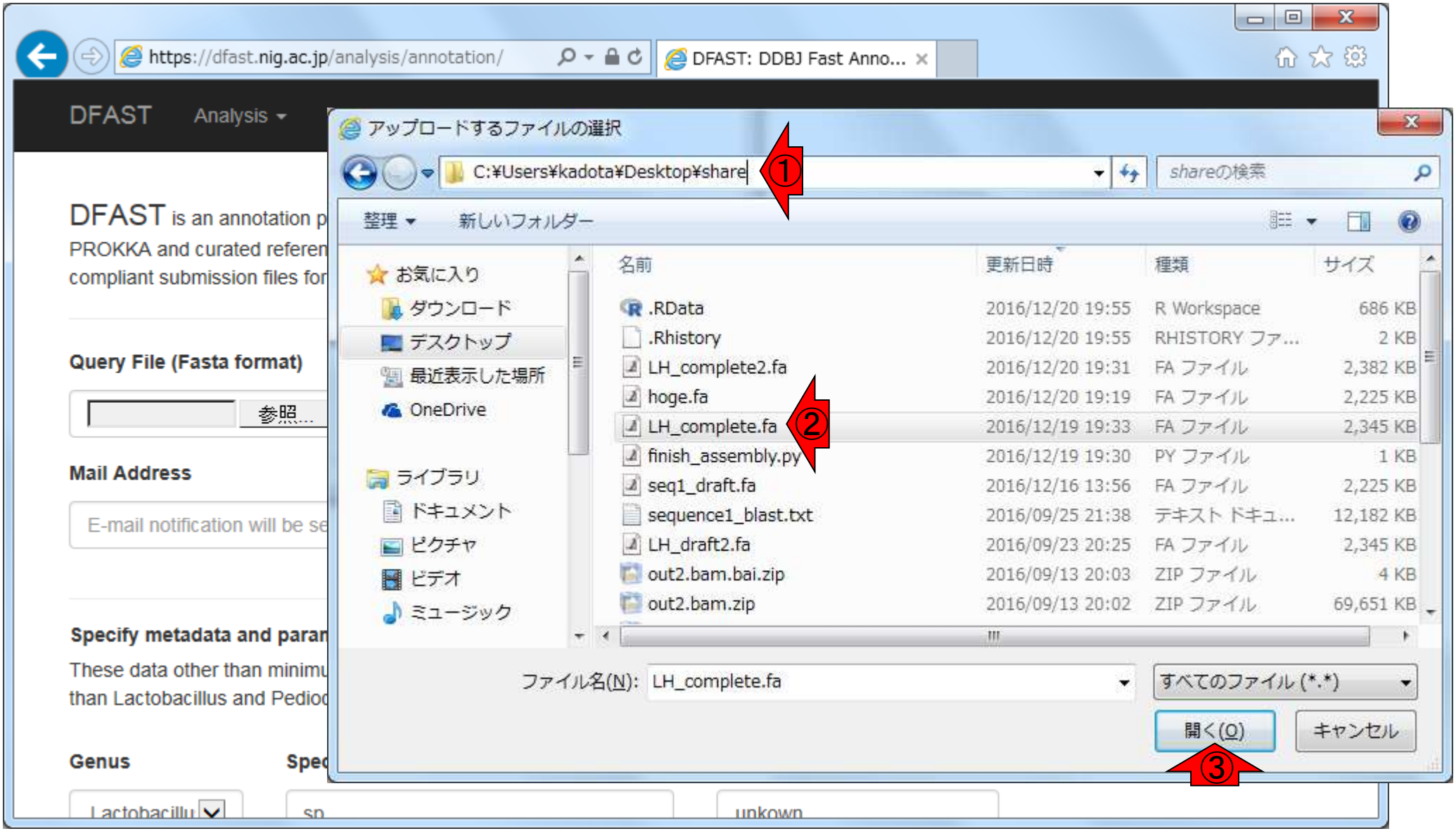
①参照、入力ファイル（LH_complete.fa；W4-6）を指定

W7-1：DFAST



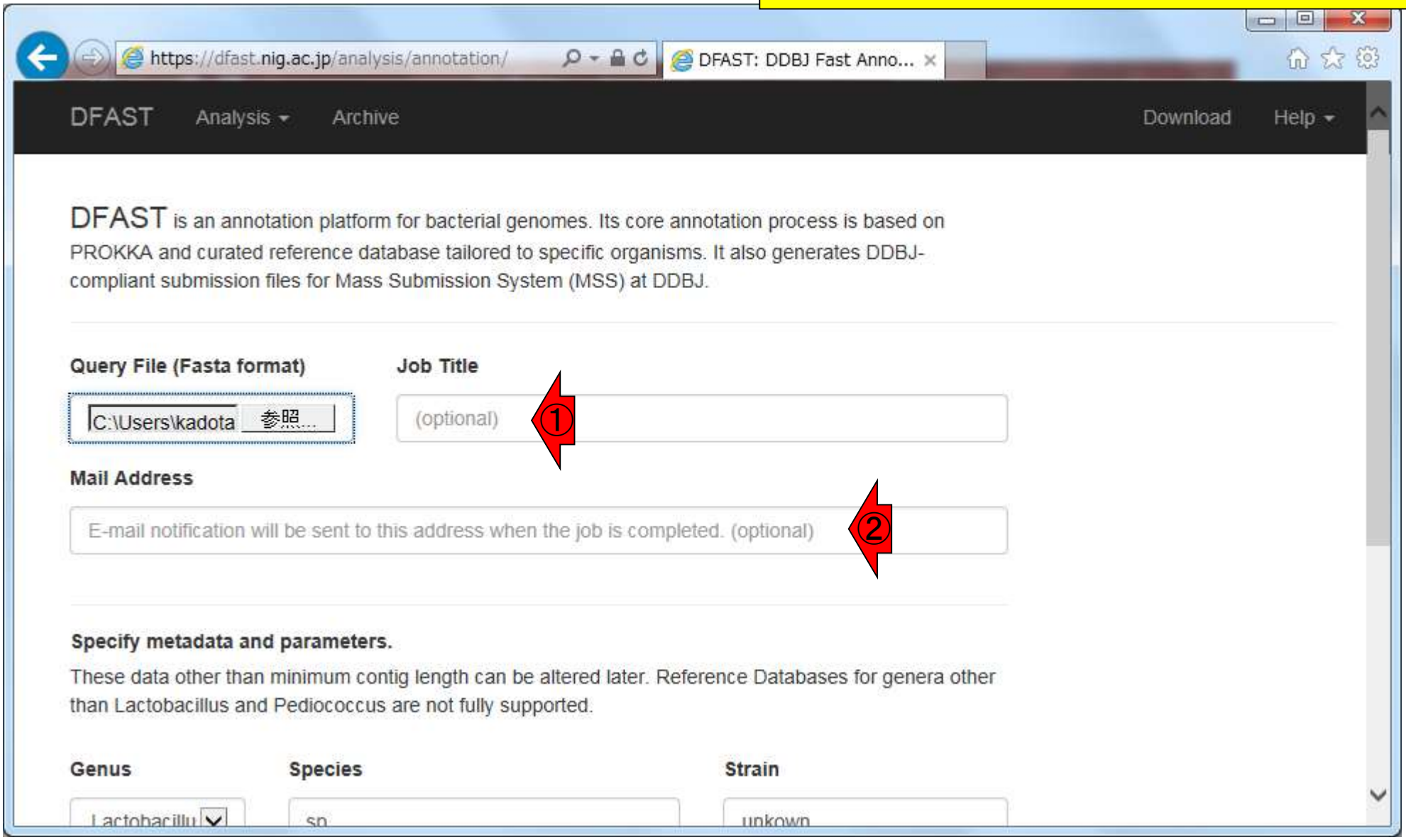
W7-1：DFAST

①共有フォルダ上にある、②LH_complete.faを選んで、③開く



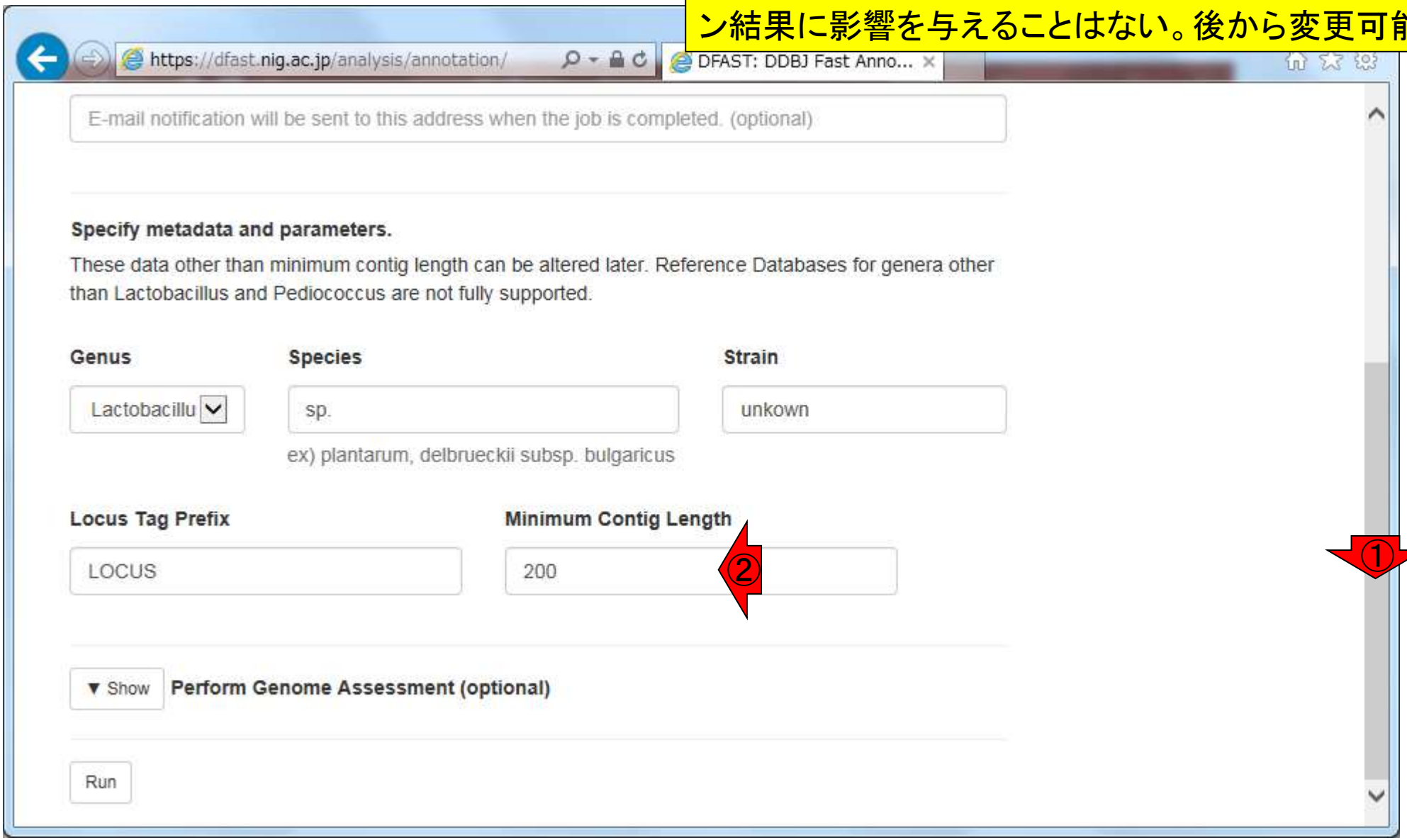
W7-1：DFAST

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



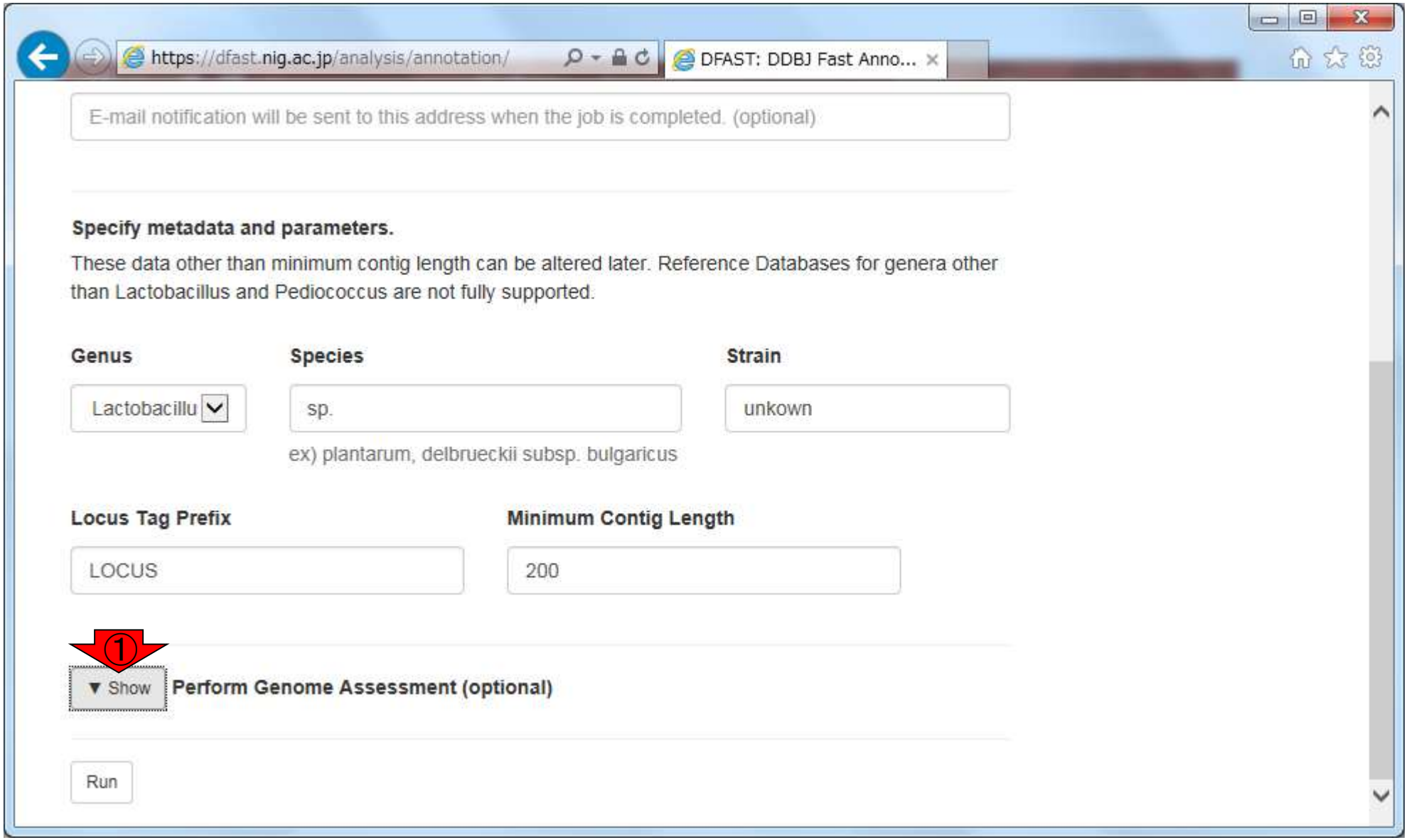
W7-1：DFAST

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



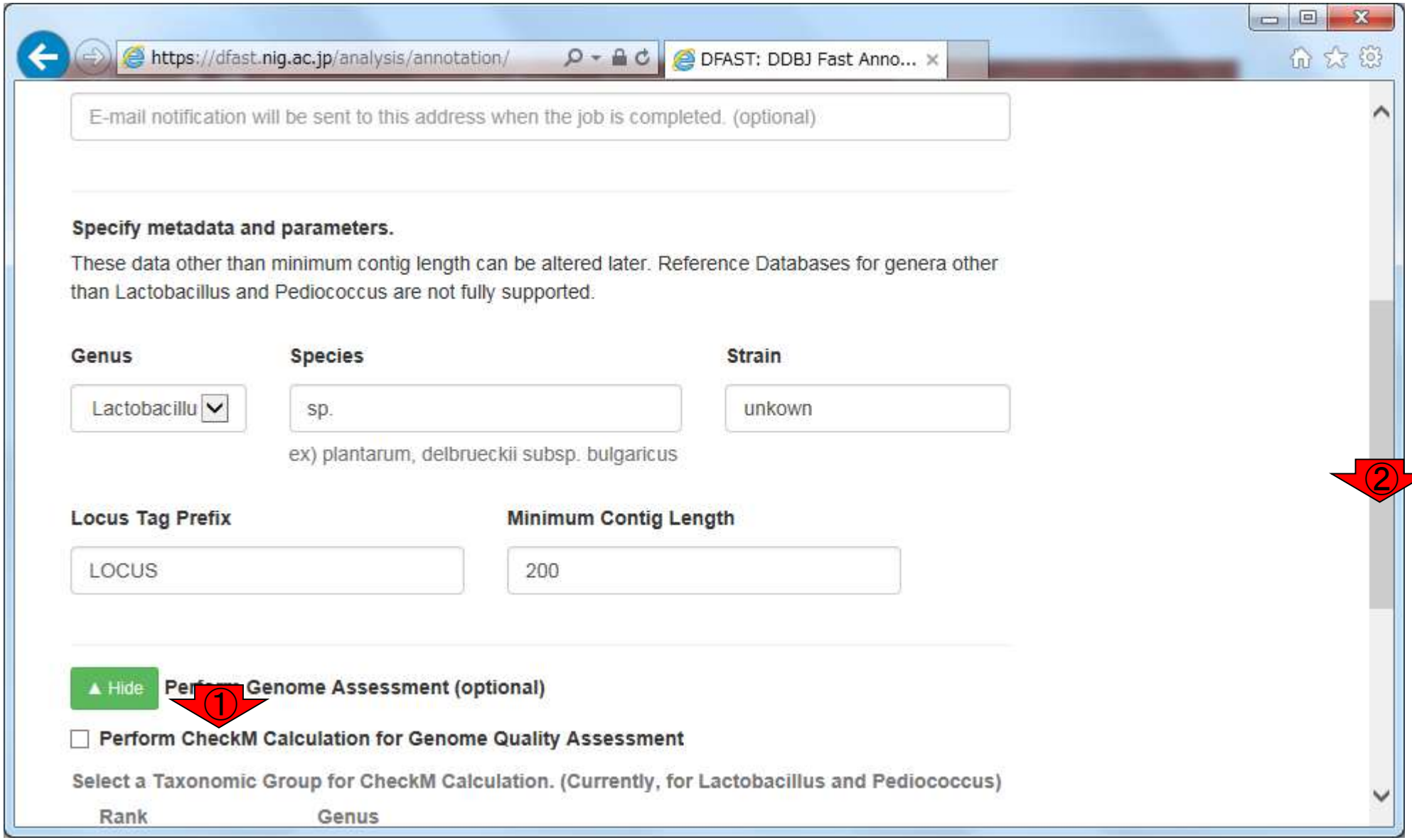
①Show

W7-2：CheckMとANI



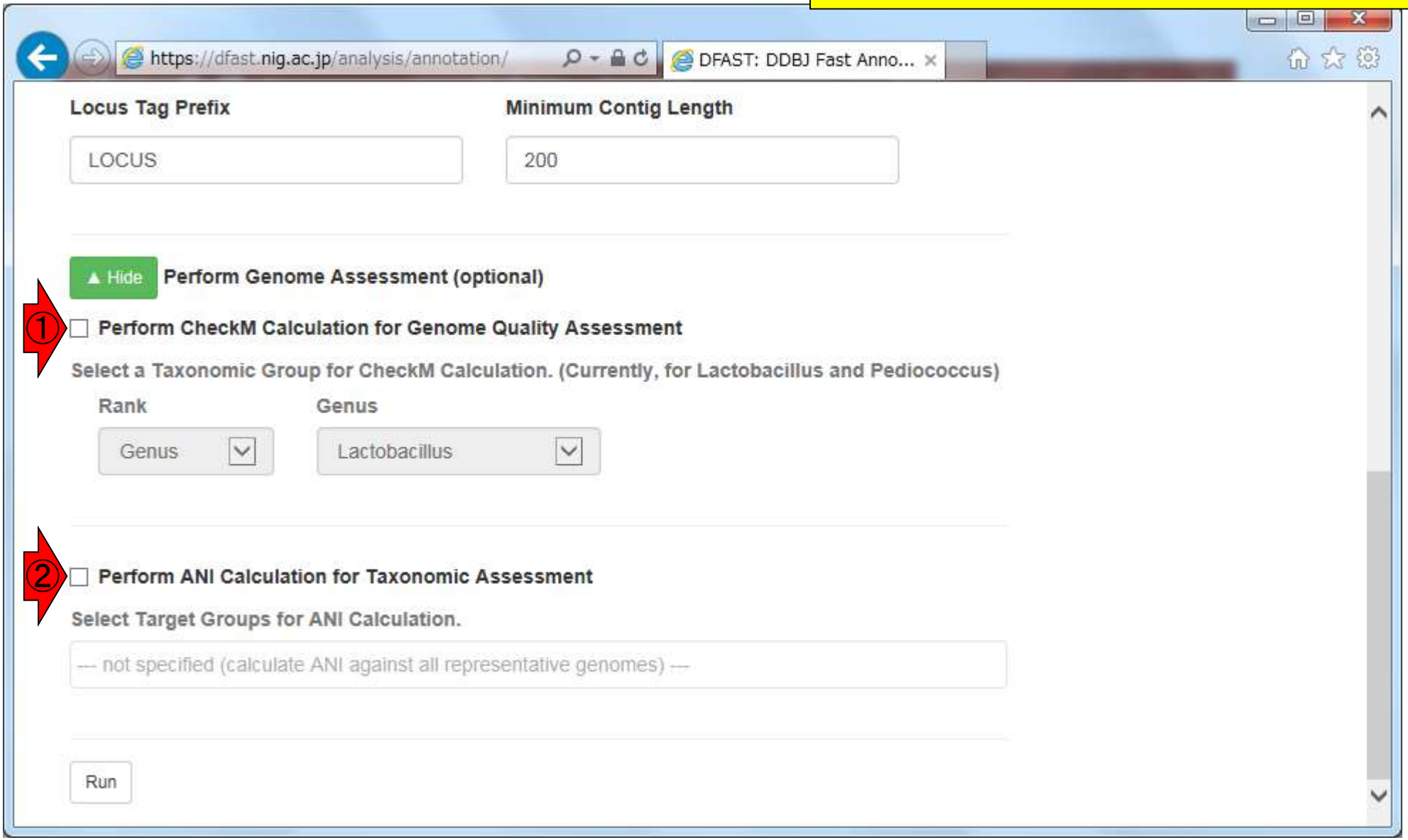
①CheckMが見られる。②ページ下部に移動

W7-2：CheckMとANI



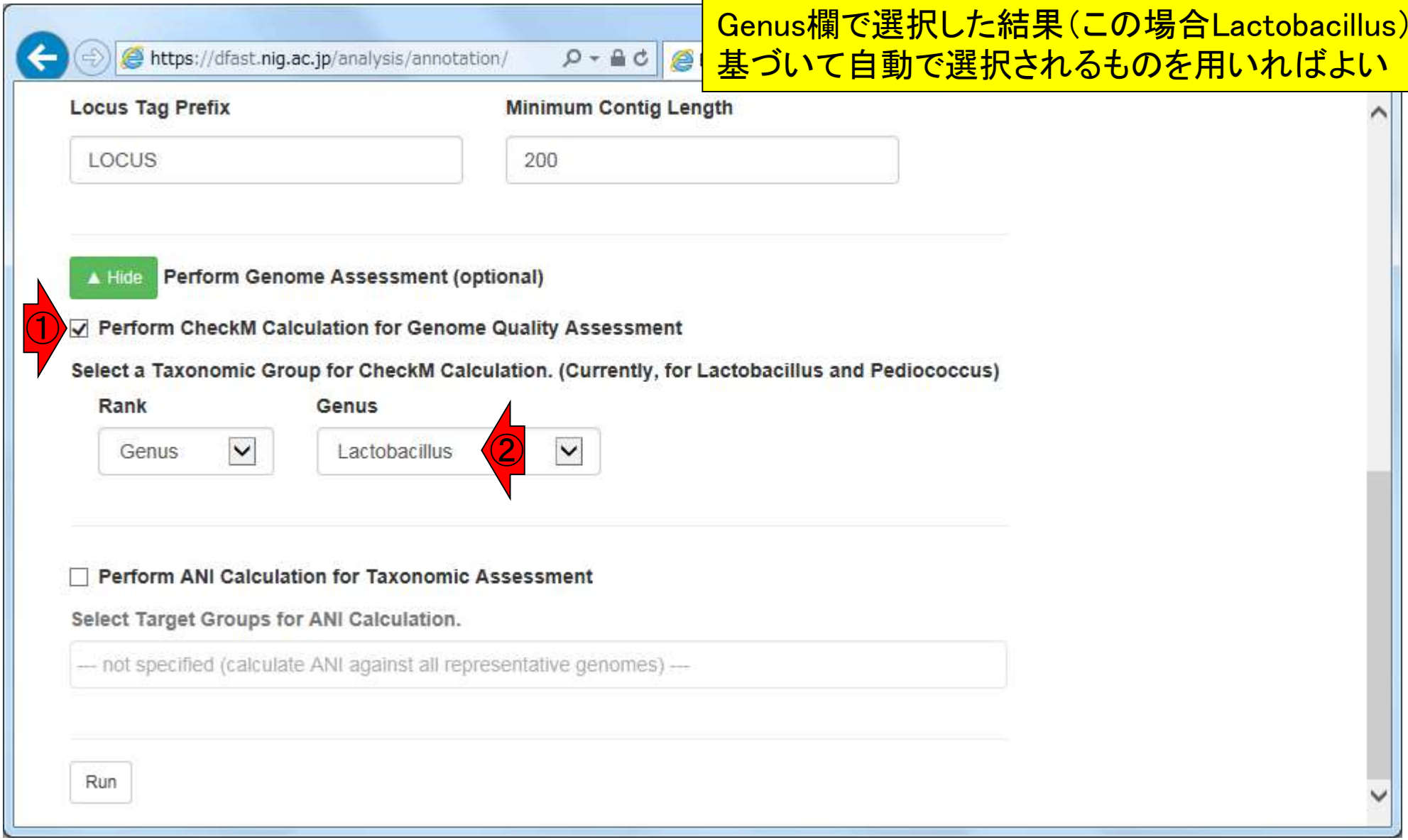
W7-2：CheckMとANI

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



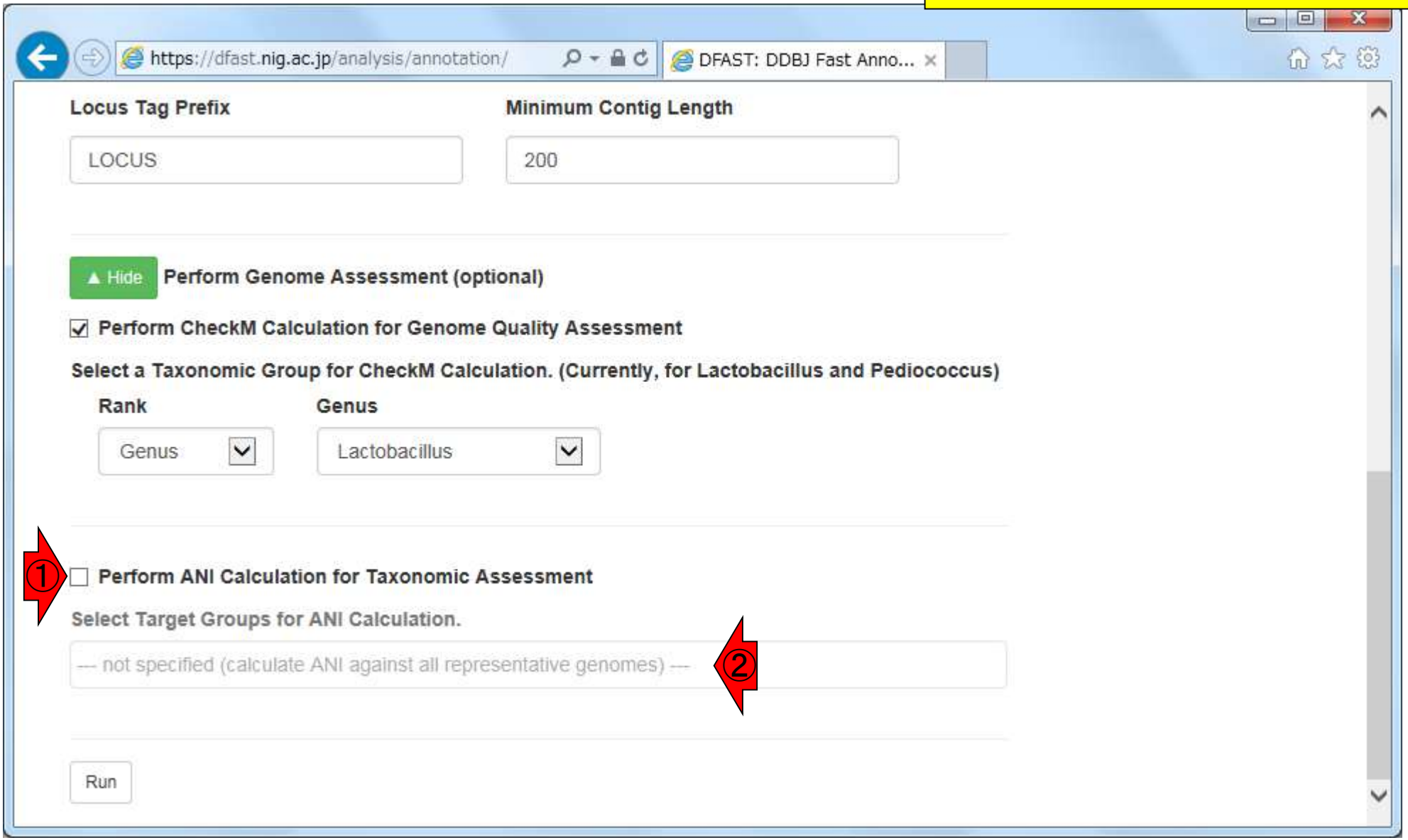
W7-3：CheckM

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)のマーカー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



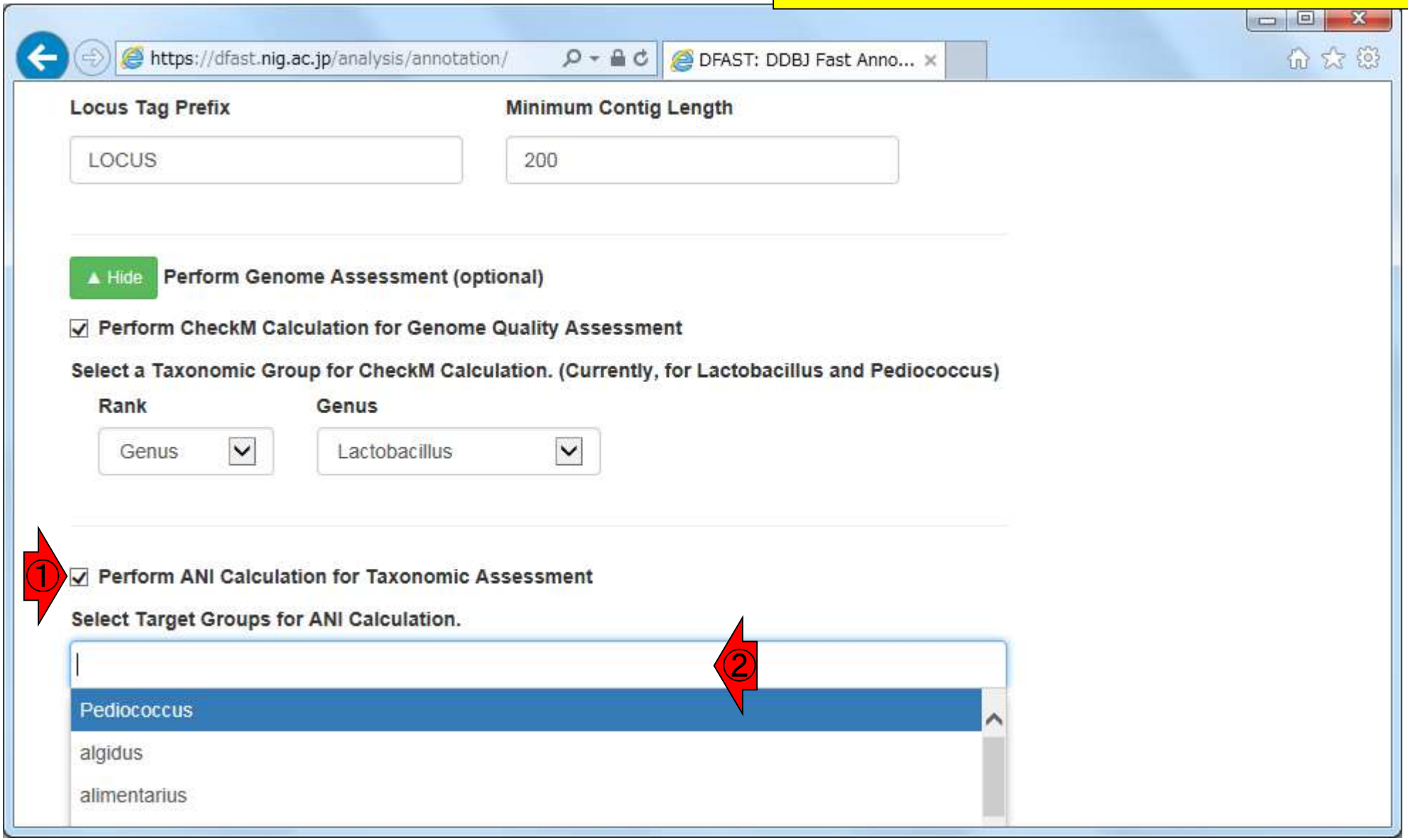
W7-4 : ANI

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）



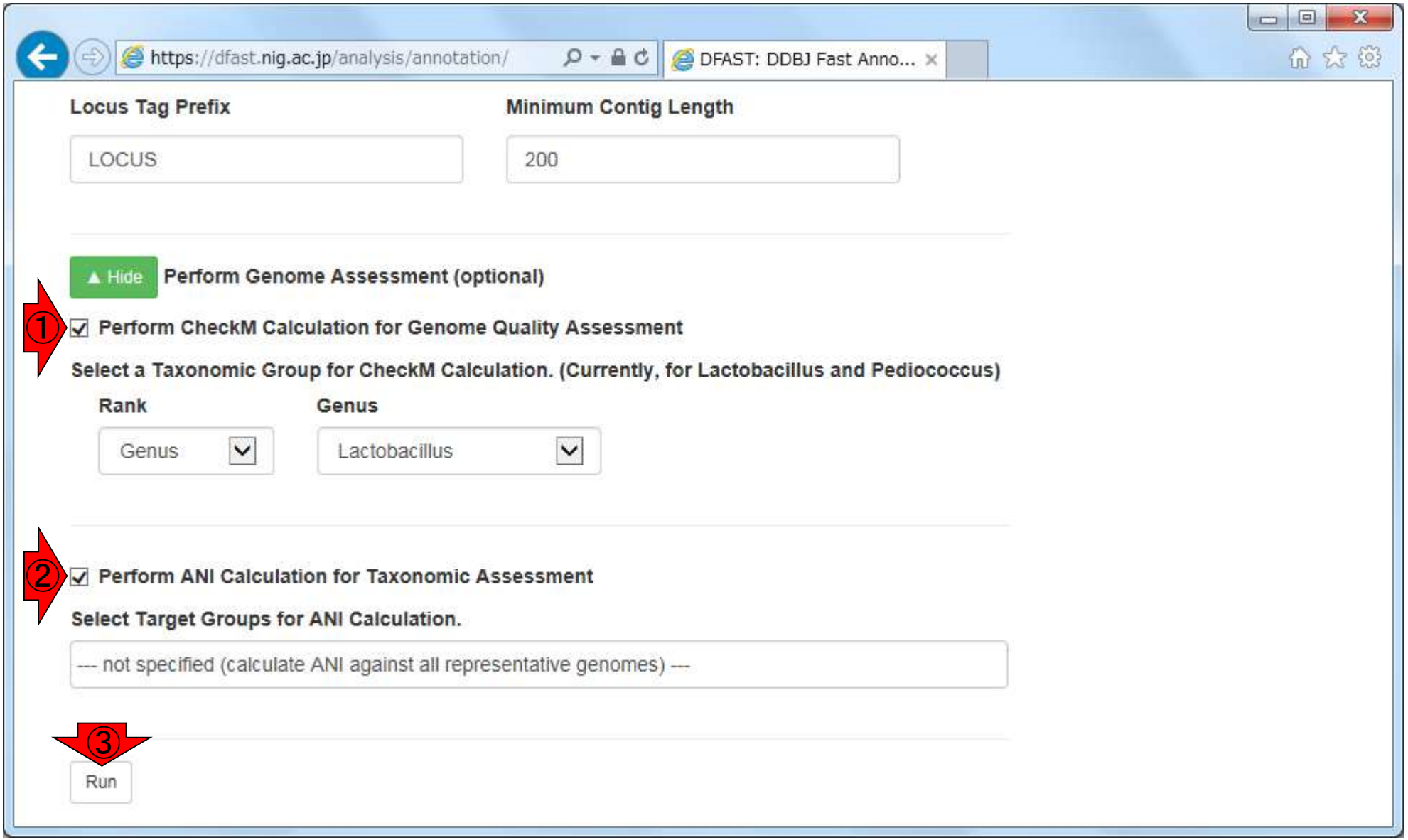
W7-4：ANI

（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よい



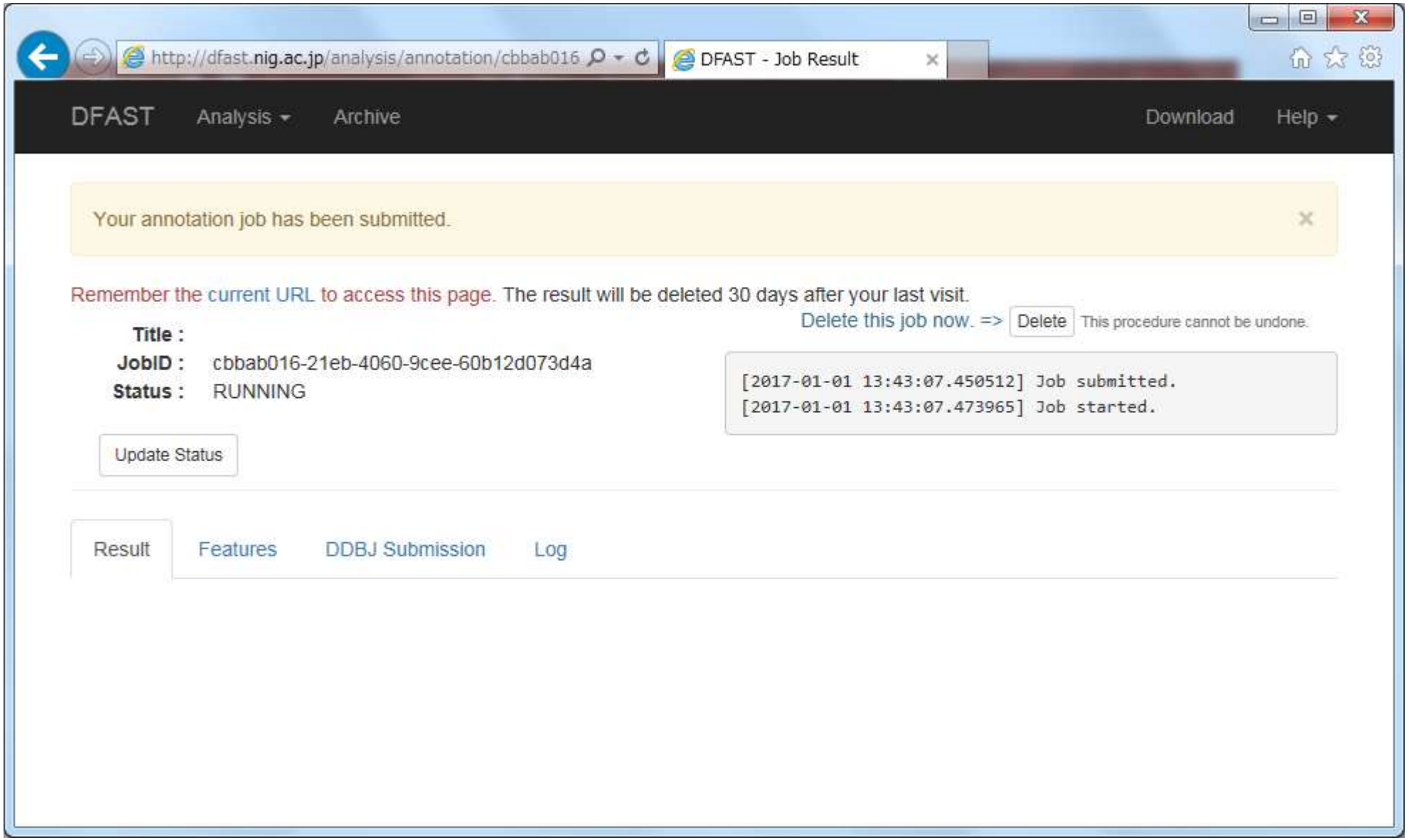
①CheckMと②ANIを計算するようにして、③Run

W7-5：実行



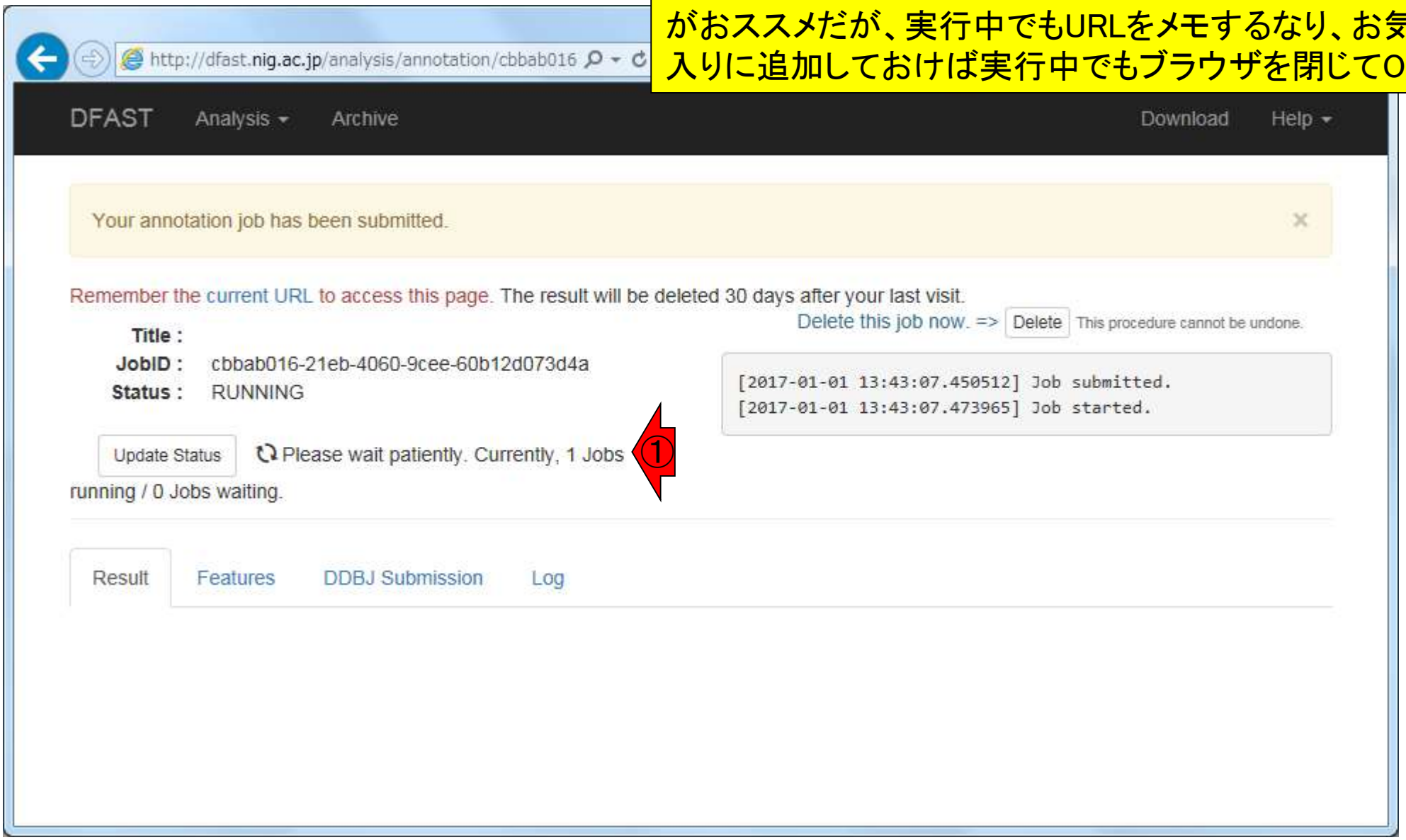
実行ボタン押した直後の状態

W7-5：実行



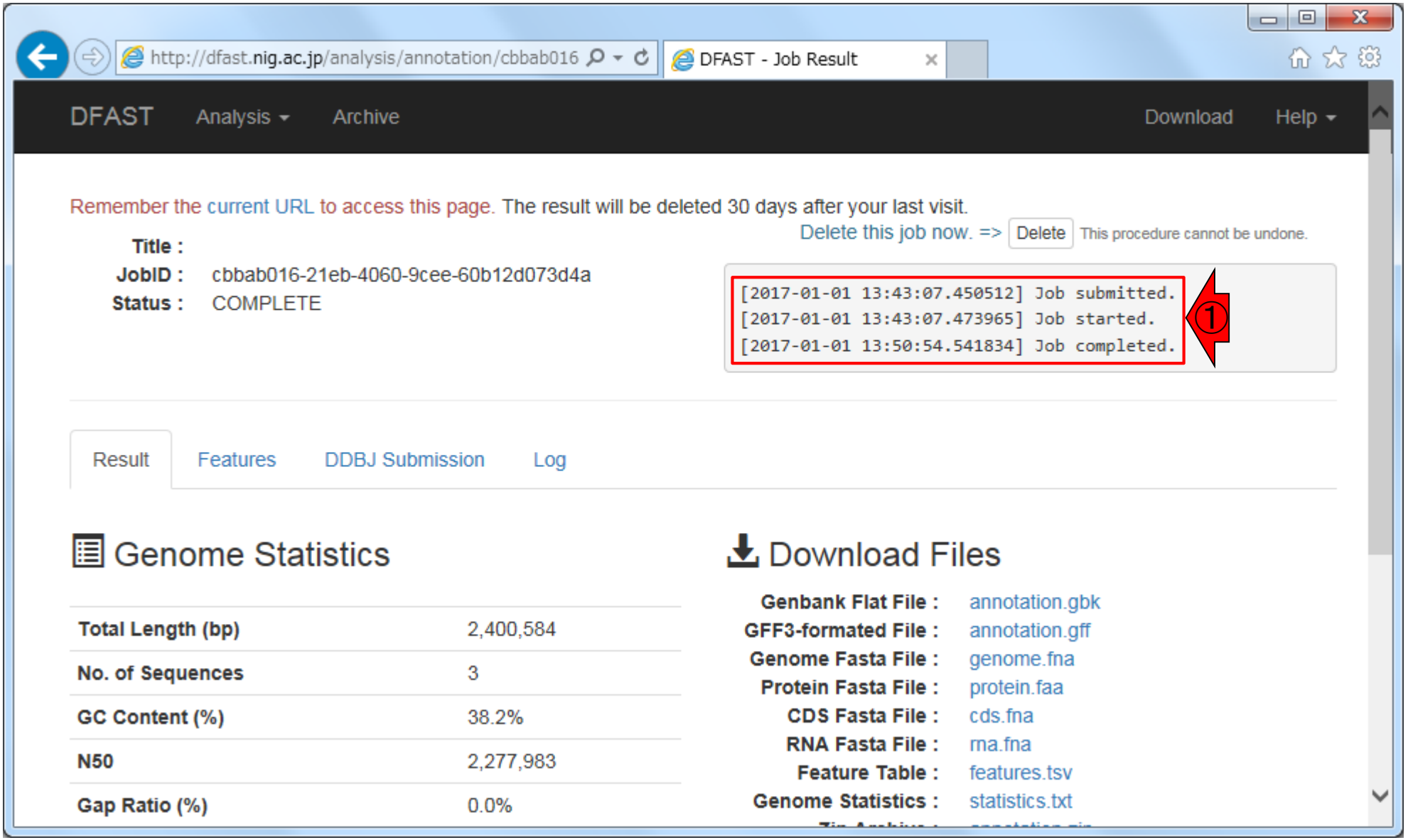
W7-5：実行中…

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おススメだが、実行中でもURLをメモするなり、お気に入りに追加しておけば実行中でもブラウザを閉じてOK



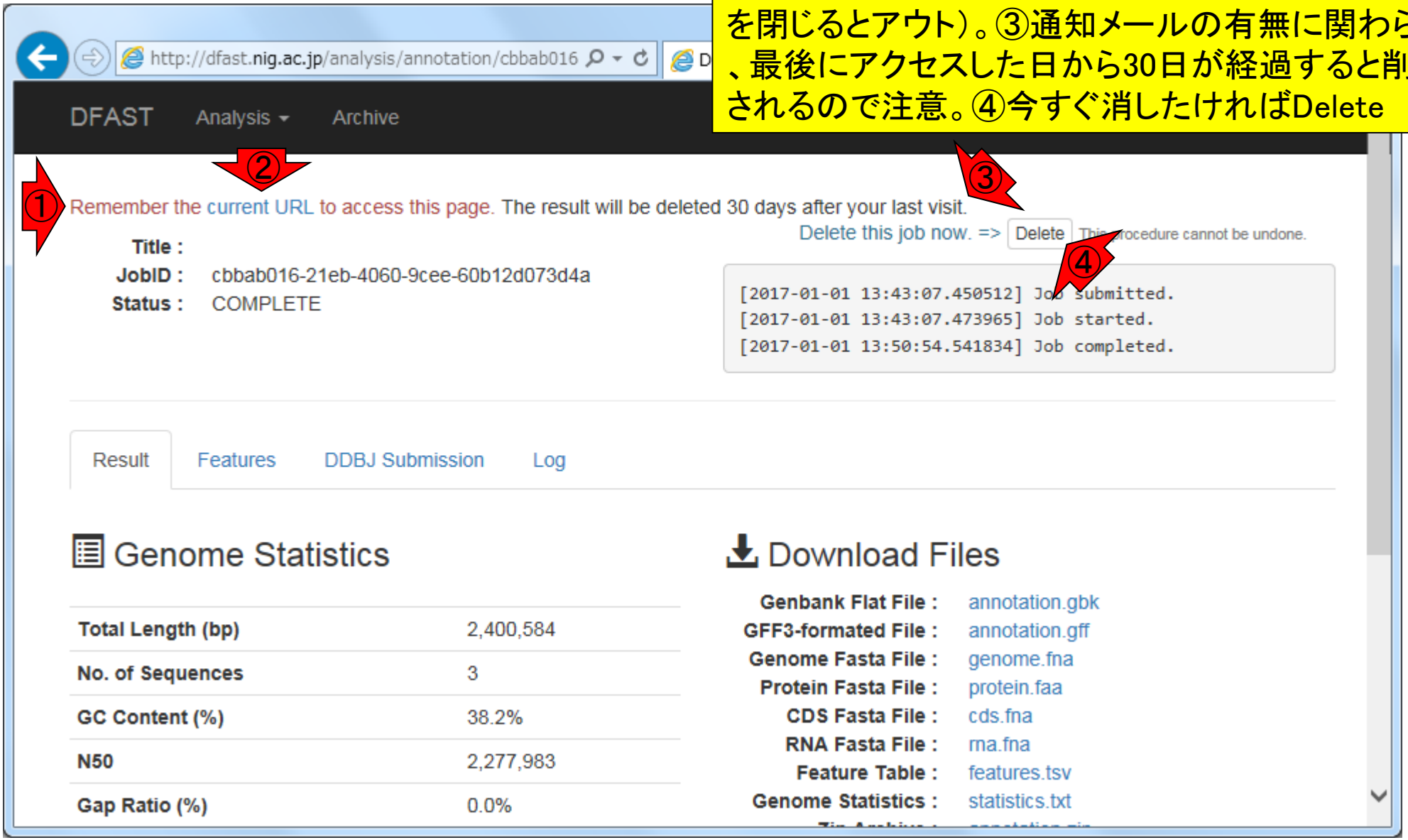
W7-6：実行結果

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



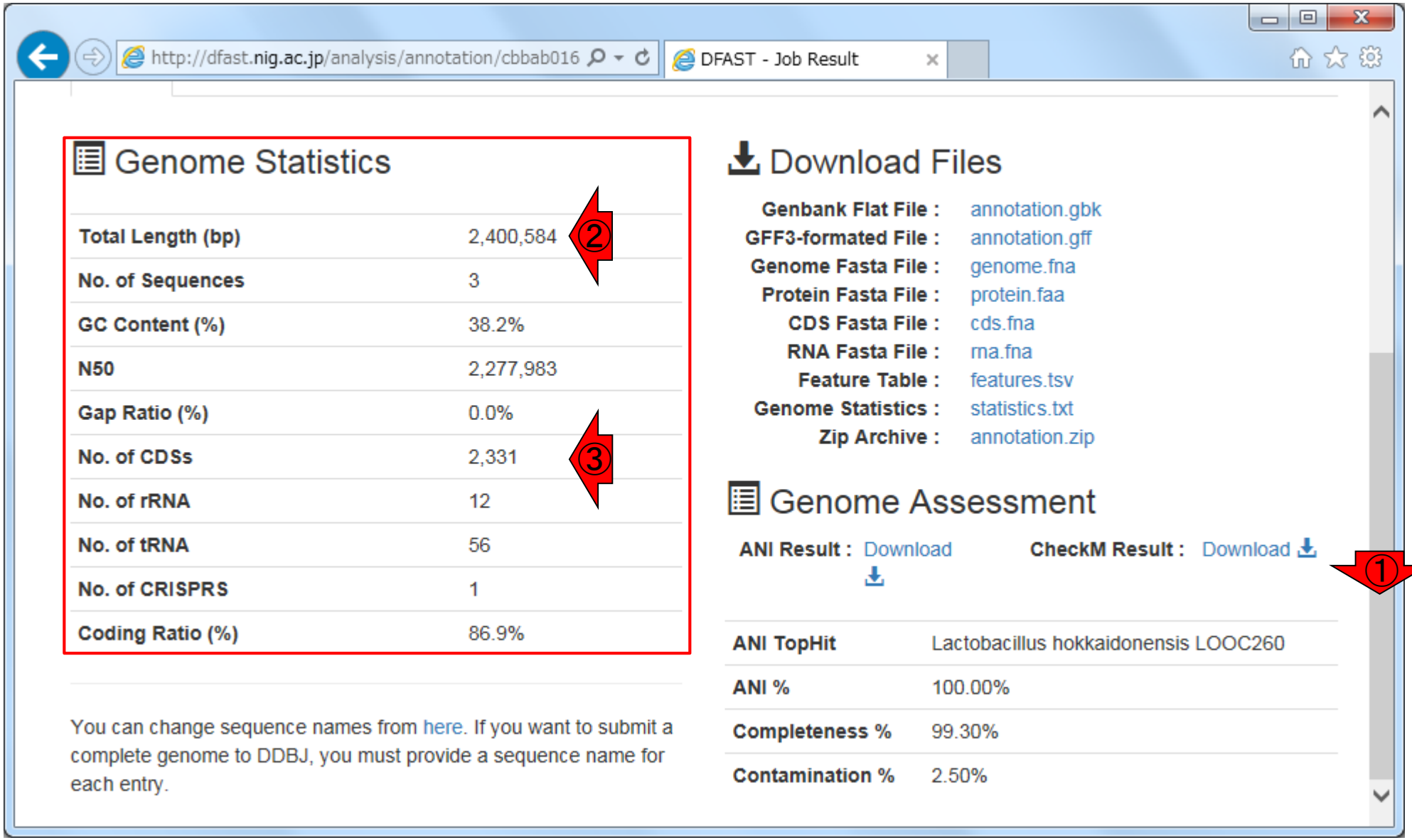
W7-6：実行結果

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



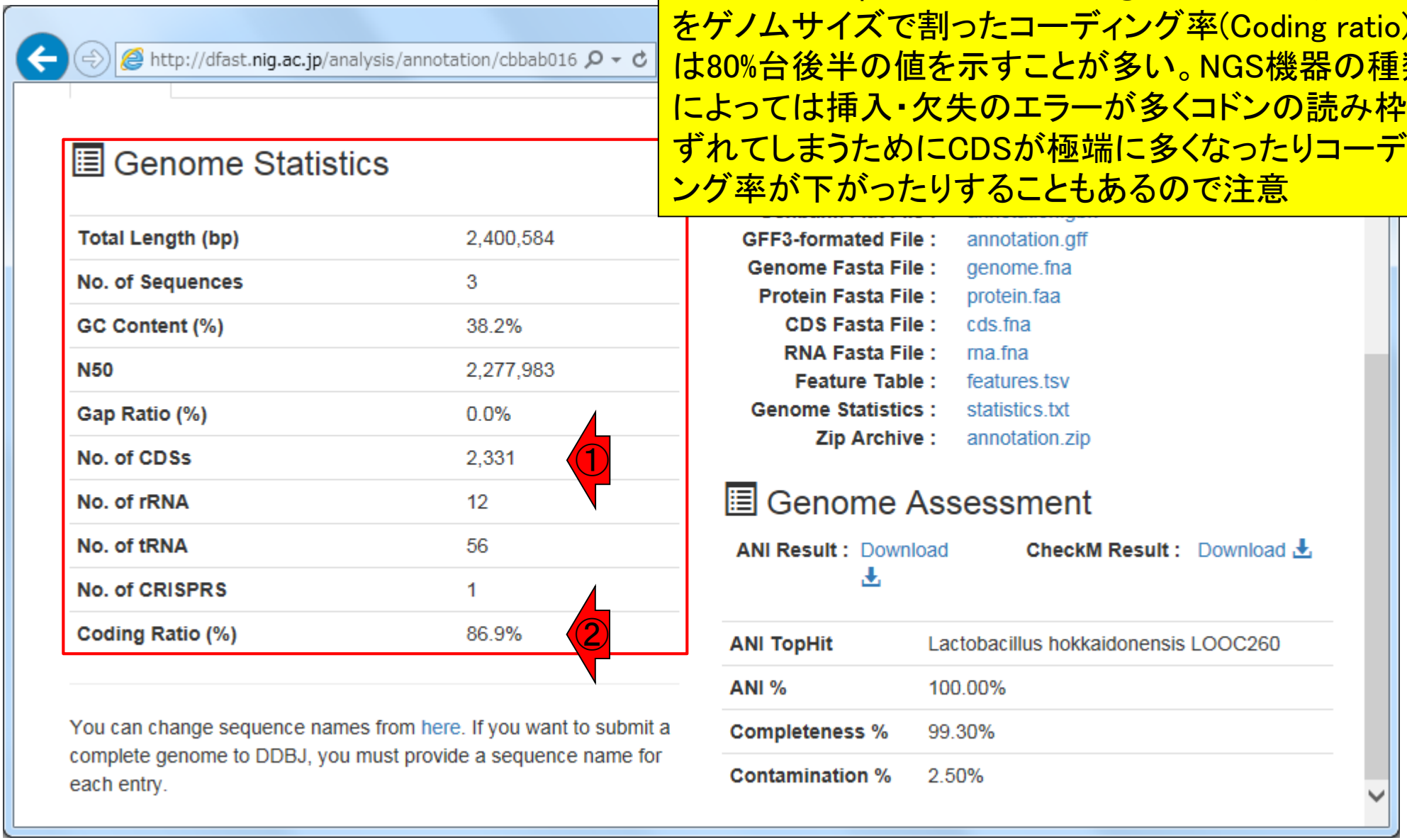
W7-7：基本統計量

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



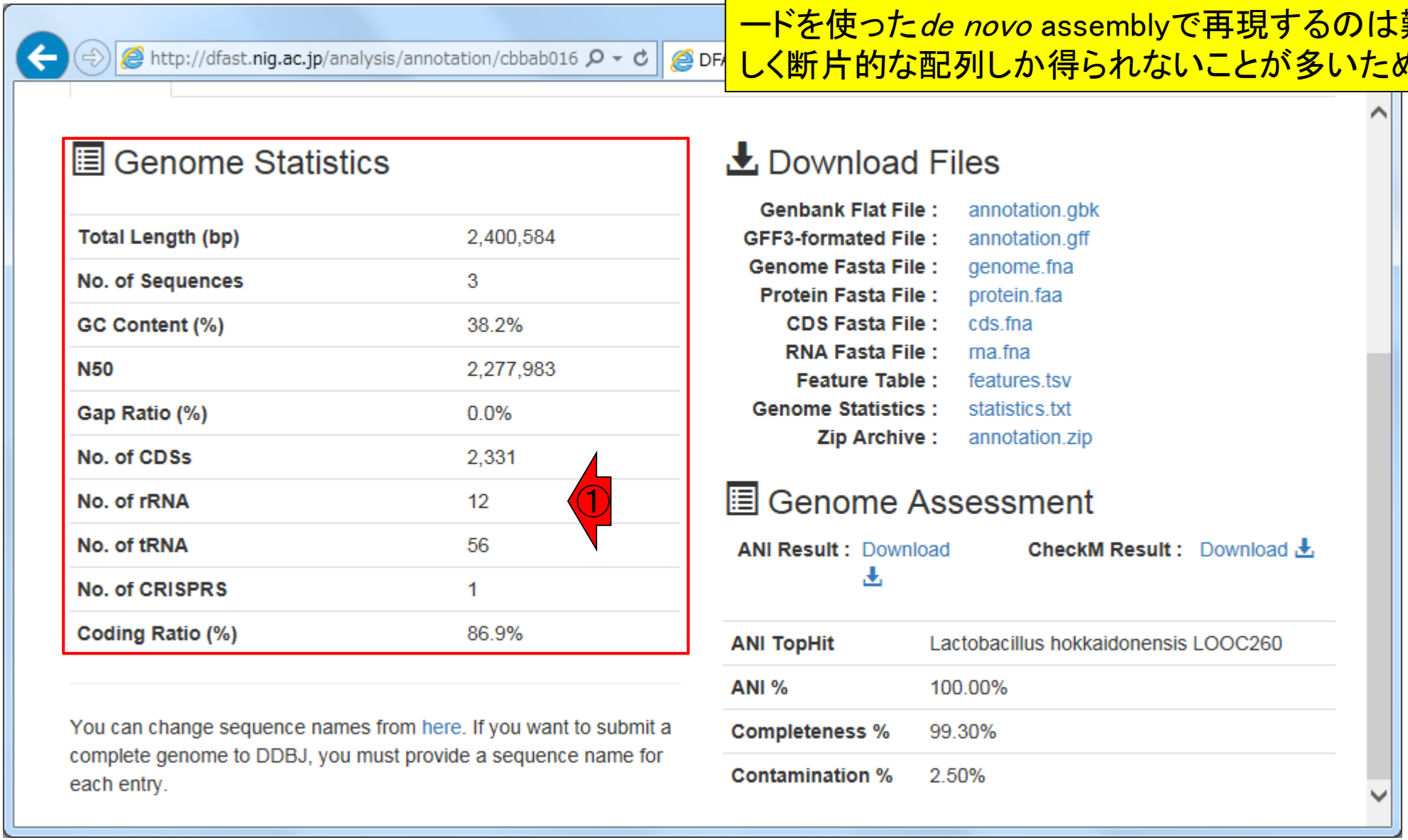
W7-7：基本統計量

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



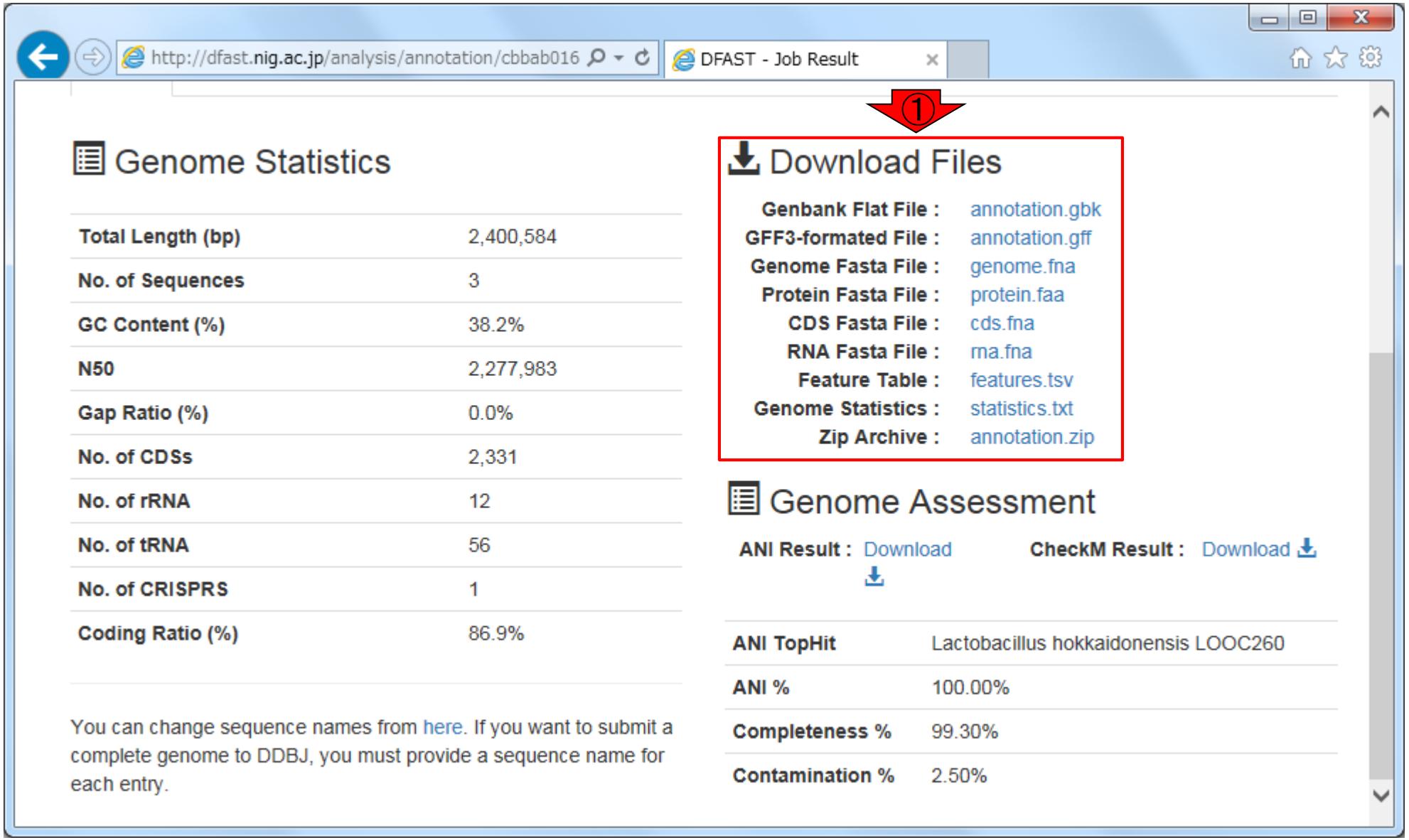
W7-7：基本統計量

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



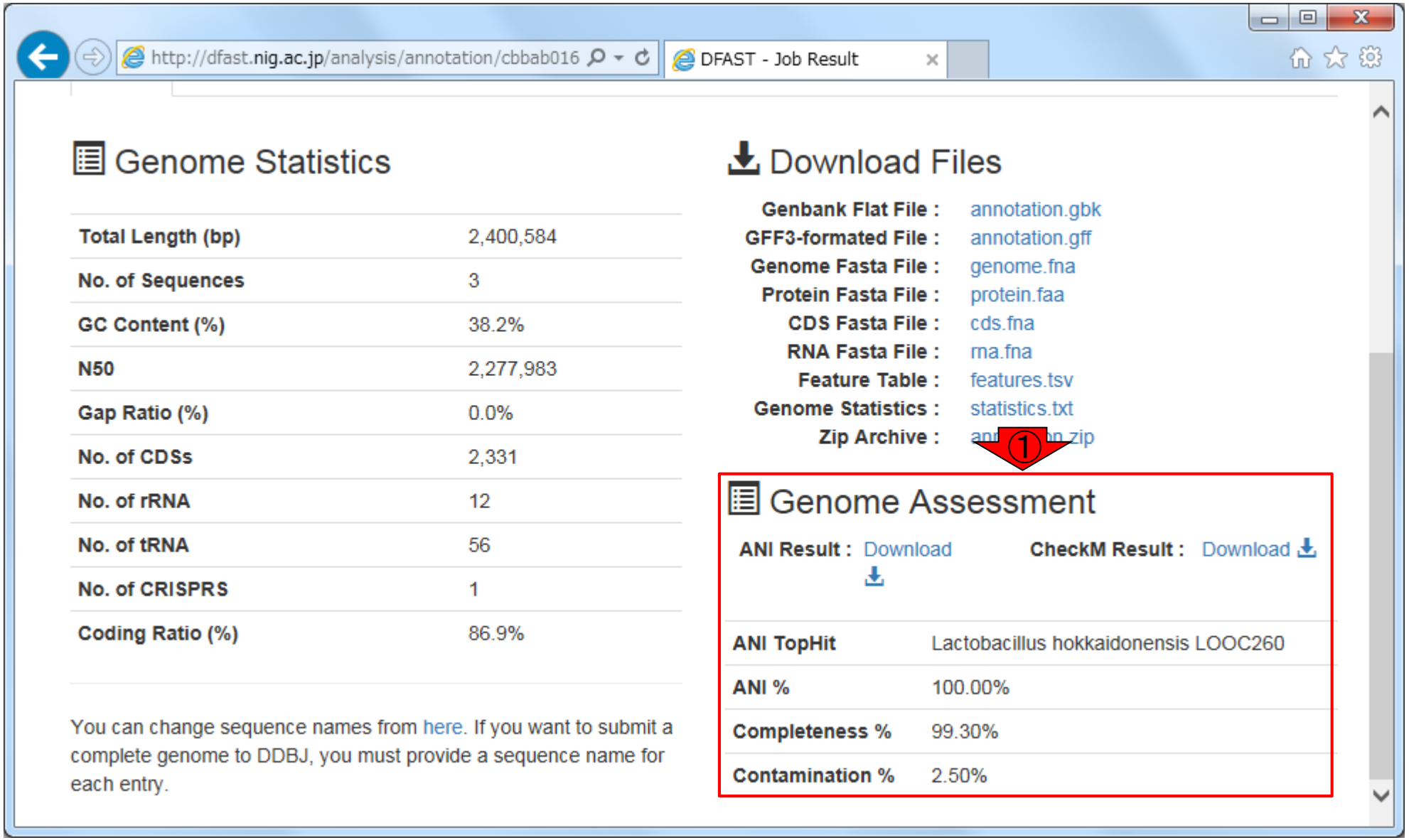
①様々なファイルをダウンロード可能

W7-8：ダウンロード



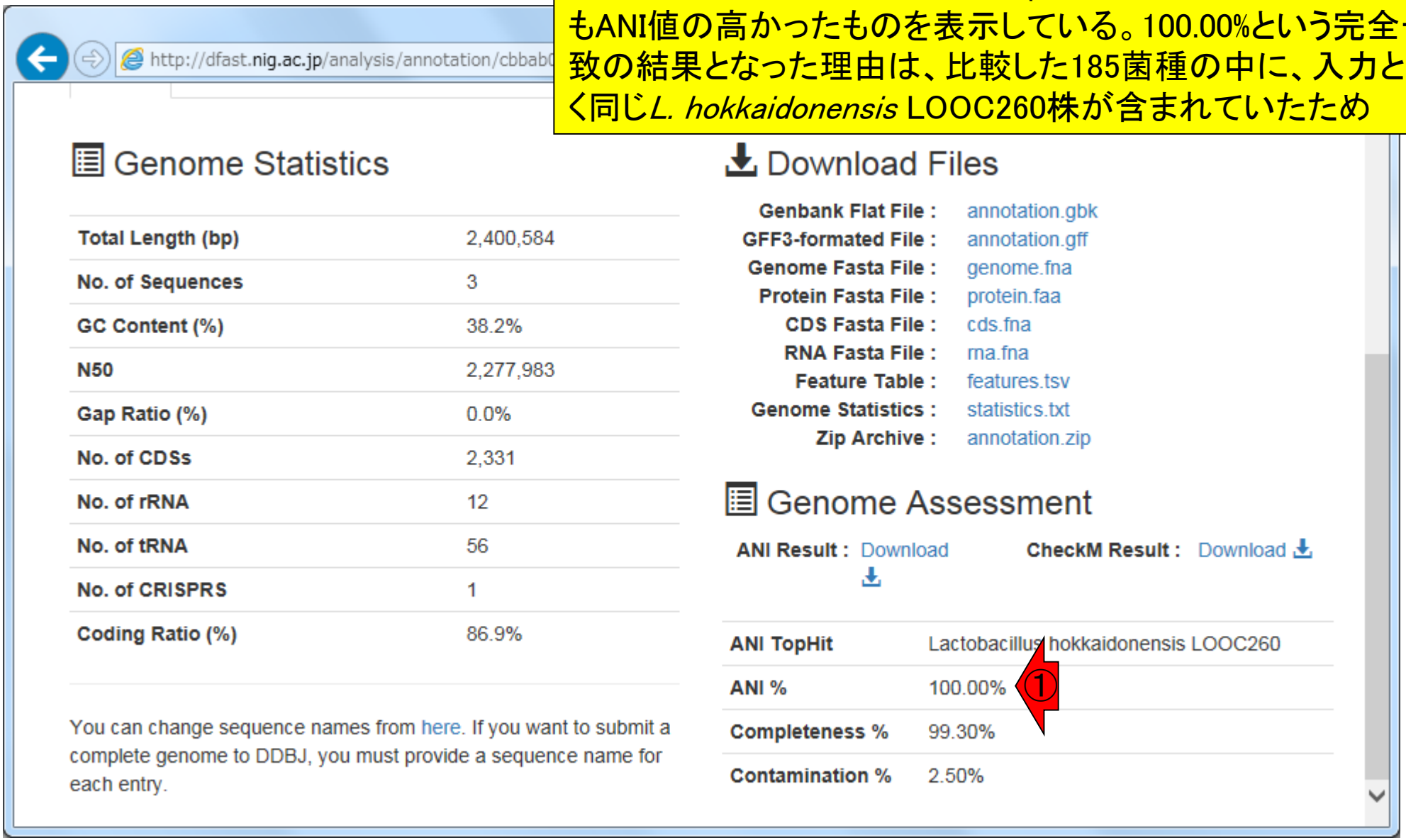
W7-9：CheckMとANI

（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



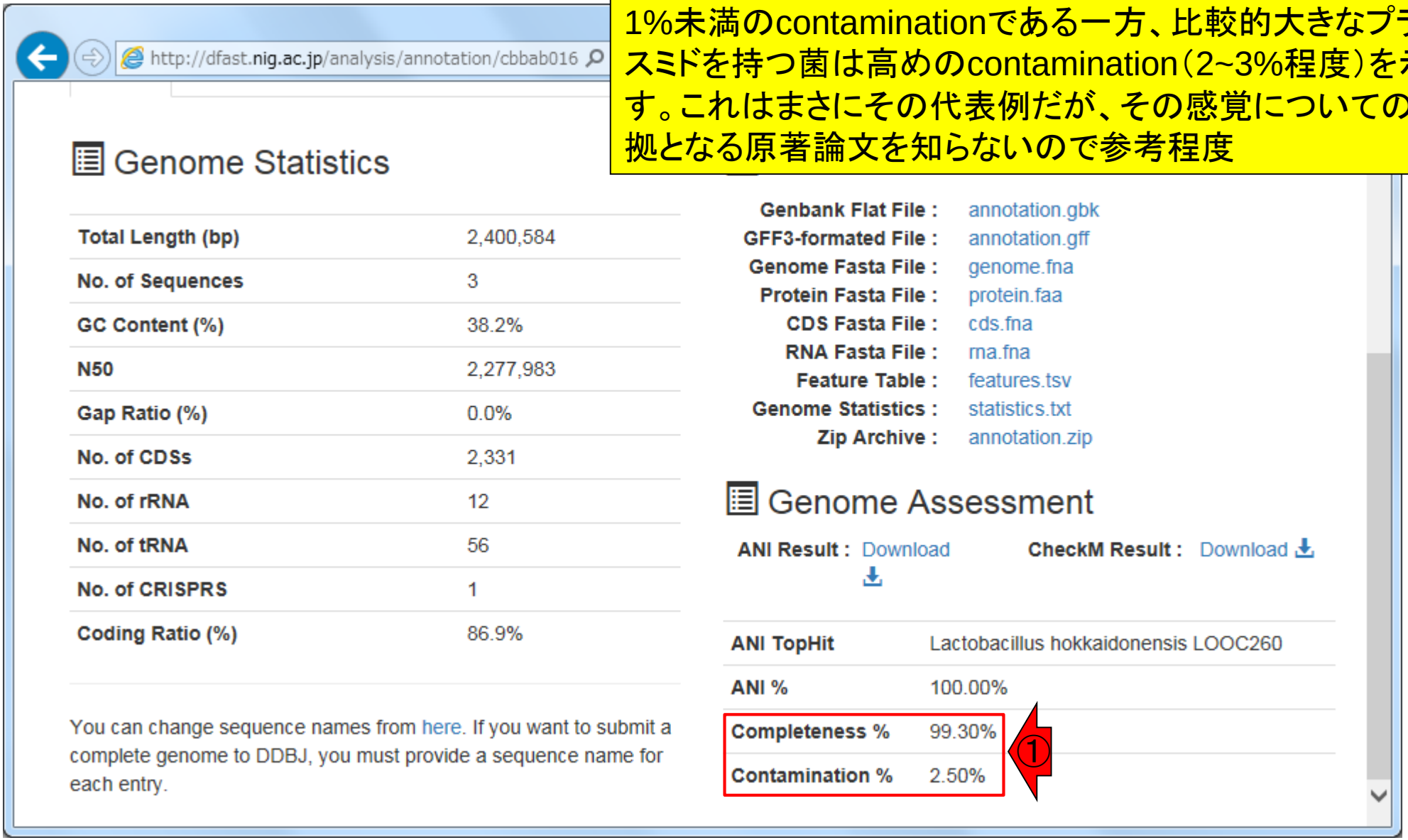
W7-10：ANI

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



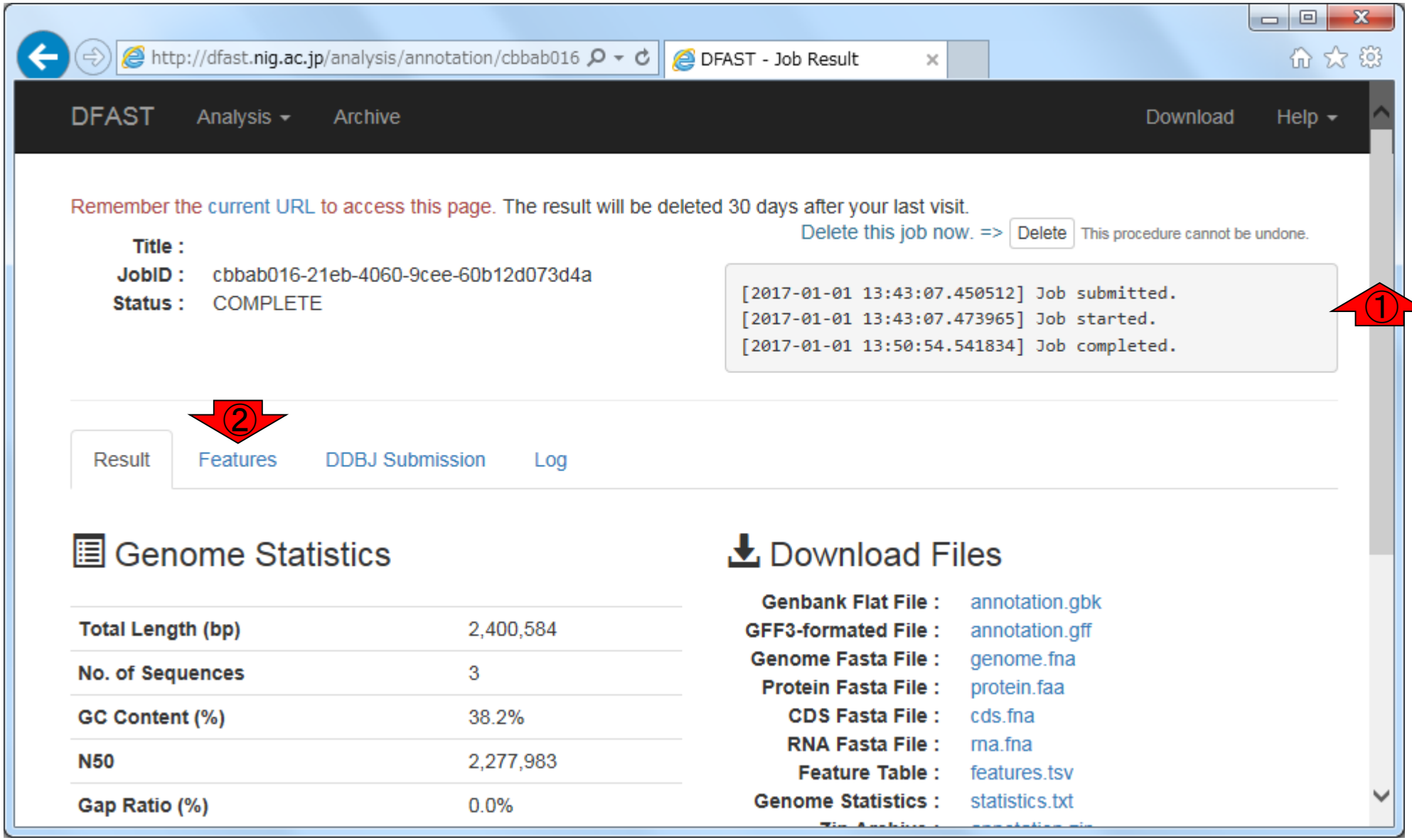
W7-11：CheckM

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



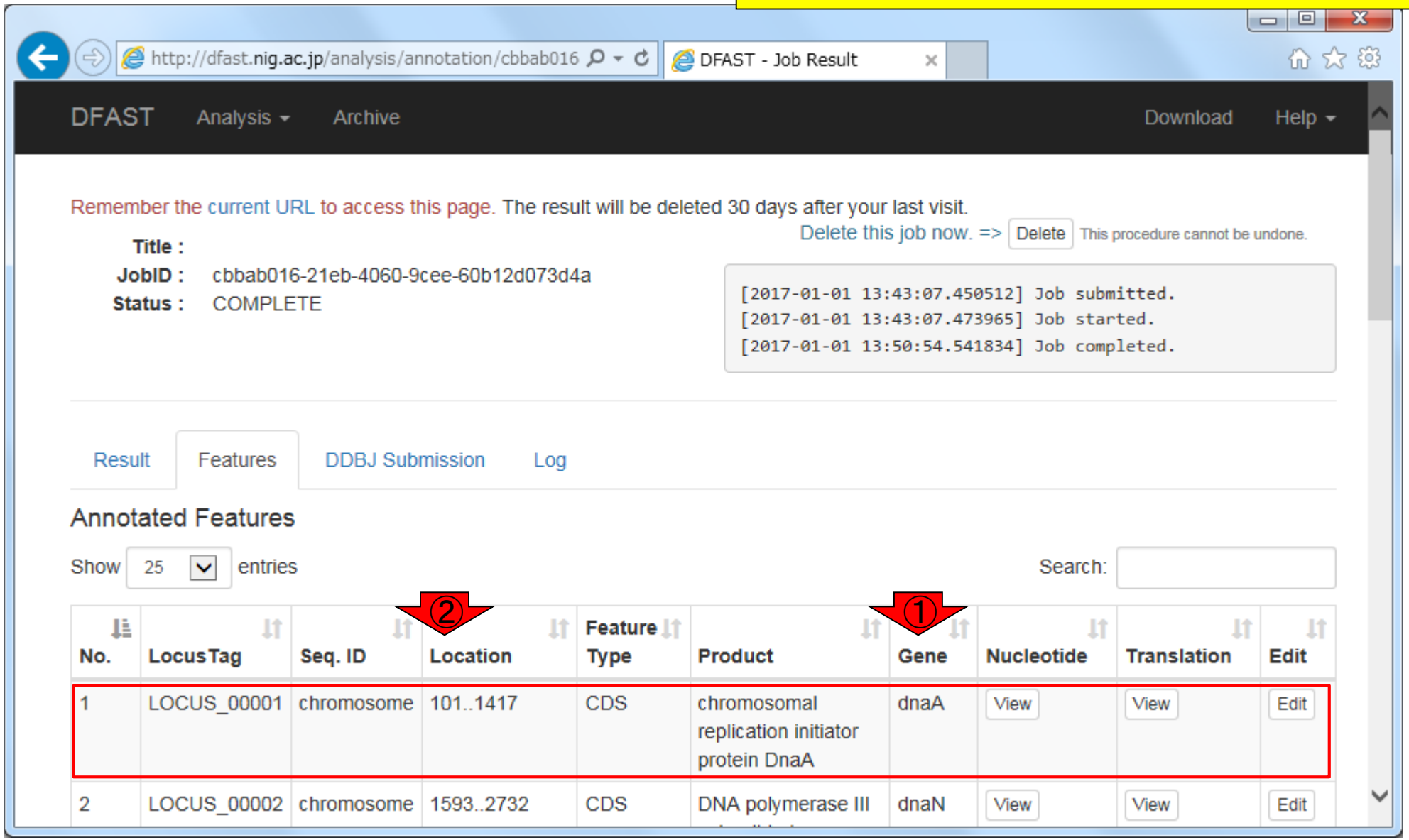
W8-1：Features

①ページ上部に再び移動し、②Featuresタブをクリック



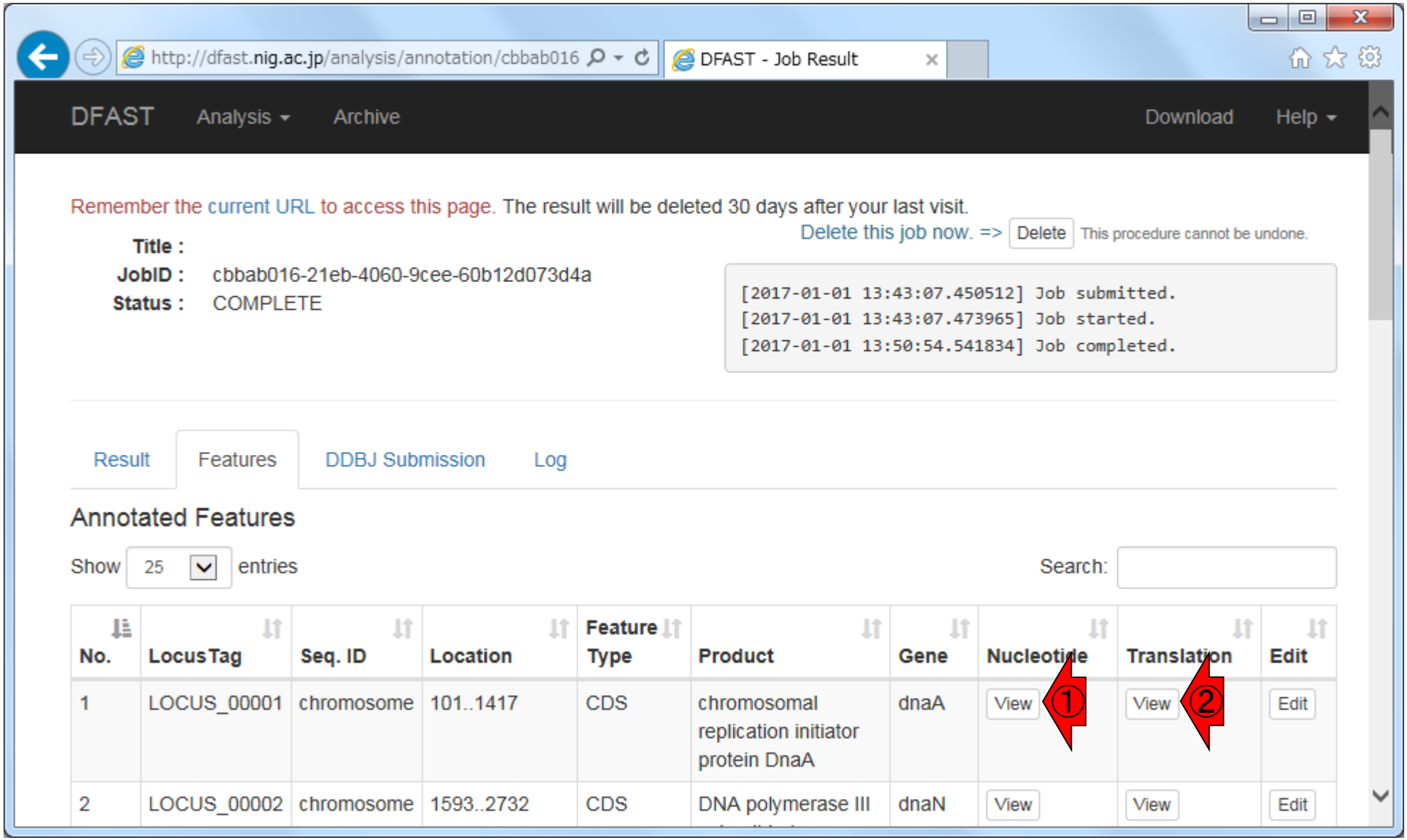
W8-2：dnaA

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、そうなって当たり前（W4）



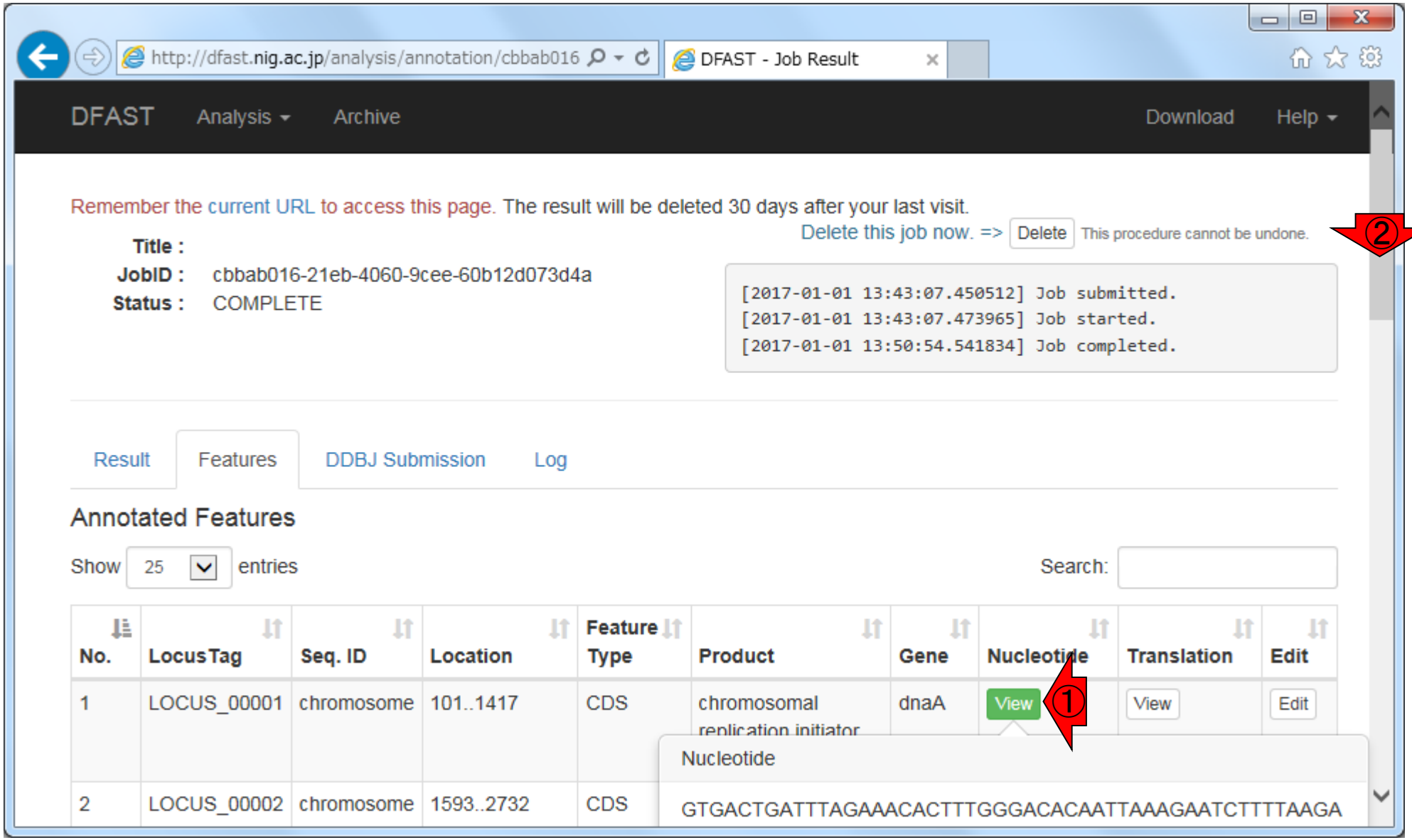
W8-3：配列情報取得

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



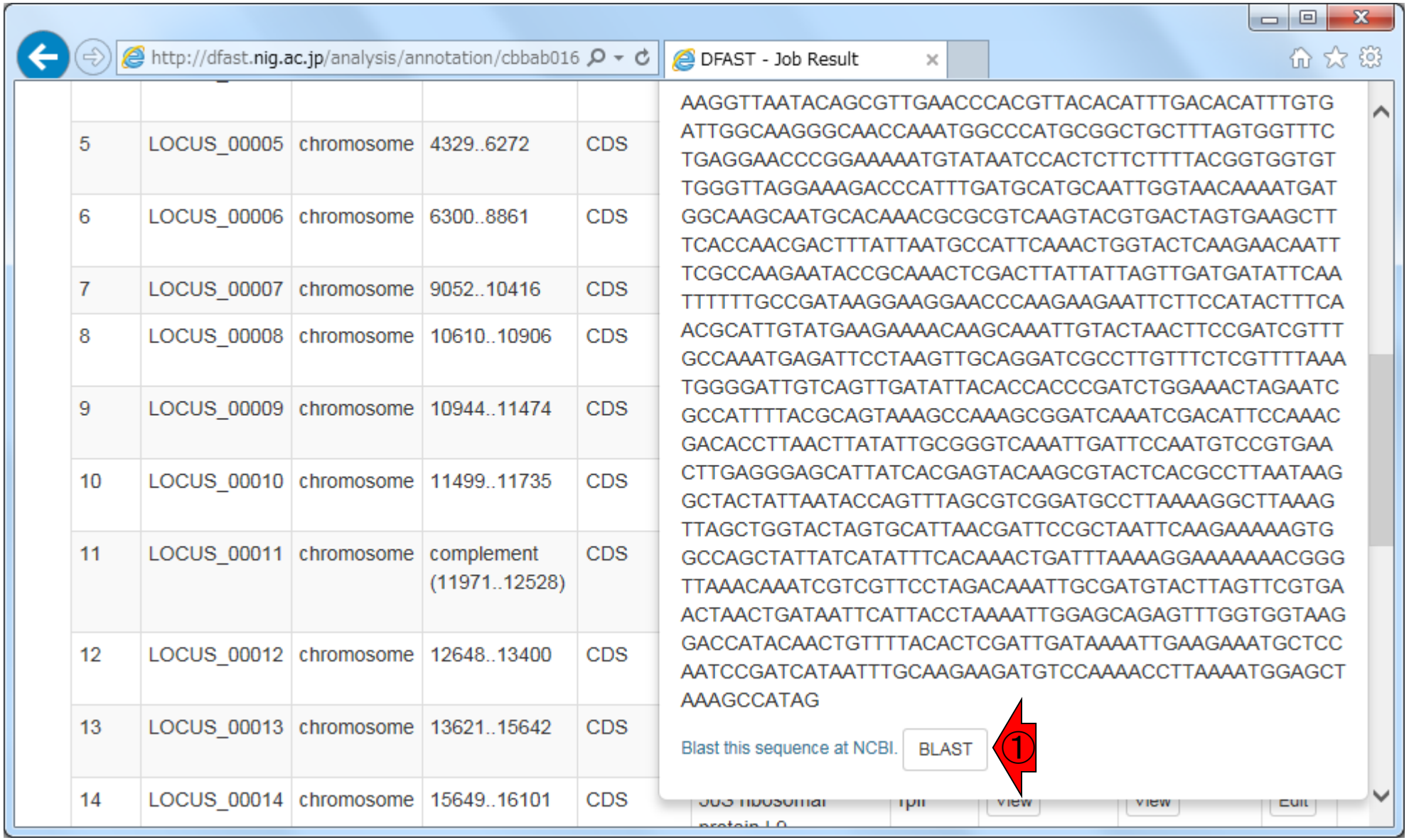
W8-4：Nucleotide

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



W8-4：Nucleotide

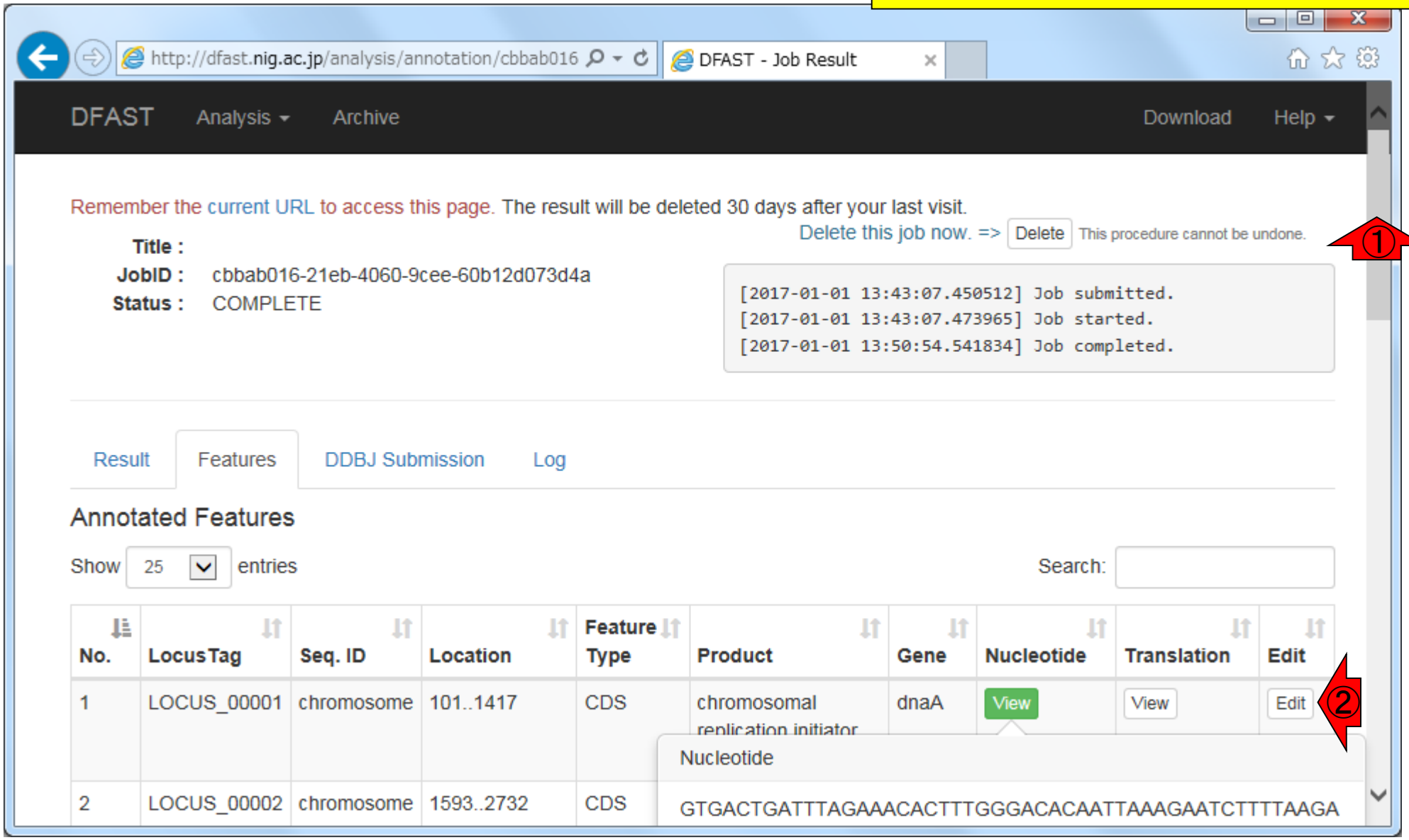
①この塩基配列を入力として、NCBI BLASTを実行することができる


http://dfast.nig.ac.jp/analysis/annotation/cbbab016
DFAST - Job Result
AAGGTTAATACAGCGTTGAACCCACGTTACACATTTGACACATTTGTG
ATTGGCAAGGGCAACCAAATGGCCCATGCGGCTGCTTTAGTGGTTTC
TGAGGAACCCGGAAAAATGTATAATCCACTCTTCTTTTACGGTGGTGT
TGGGTTAGGAAAGACCCATTTGATGCATGCAATTGGTAACAAAATGAT
GGCAAGCAATGCACAAACGCGCGTCAAGTACGTGACTAGTGAAGCTT
TCACCAACGACTTTATTAATGCCATTCAAACTGGTACTCAAGAACAATT
TCGCCAAGAATACCGCAAACTCGACTTATTATTAGTTGATGATATTCAA
TTTTTTGCCGATAAGGAAGGAACCCAAGAAGAATTCTTCCATACTTTCA
ACGCATTGTATGAAGAAAACAAGCAAATTGTACTAACTTCCGATCGTTT
GCCAAATGAGATTCCTAAGTTGCAGGATCGCCTTGTTTCTCGTTTTAAA
TGGGGATTGTCAGTTGATATTACACCACCCGATCTGGAAACTAGAATC
GCCATTTTACGCAGTAAAGCCAAAGCGGATCAAATCGACATTCCAAAC
GACACCTTAACTTATATTGCGGGTCAAATTGATTCCAATGTCCGTGAA
CTTGAGGGAGCATTATCACGAGTACAAGCGTACTCACGCCTTAATAAG
GCTACTATTAATACCAGTTTAGCGTCGGATGCCTTAAAAGGCTTAAAG
TTAGCTGGTACTAGTGCATTAACGATTCCGCTAATTCAAGAAAAAGTG
GCCAGCTATTATCATATTTCACAAACTGATTTAAAAGGAAAAAAACGGG
TTAAACAAATCGTCGTTCCTAGACAAATTGCGATGTACTTAGTTCGTGA
ACTAACTGATAATTCATTACCTAAAATTGGAGCAGAGTTTGGTGGTAAG
GACCATACAACTGTTTTACACTCGATTGATAAAATTGAAGAAATGCTCC
AATCCGATCATAATTTGCAAGAAGATGTCCAAAACCTTAAAATGGAGCT
AAAGCCATAG
Blast this sequence at NCBI.
BLAST
①

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

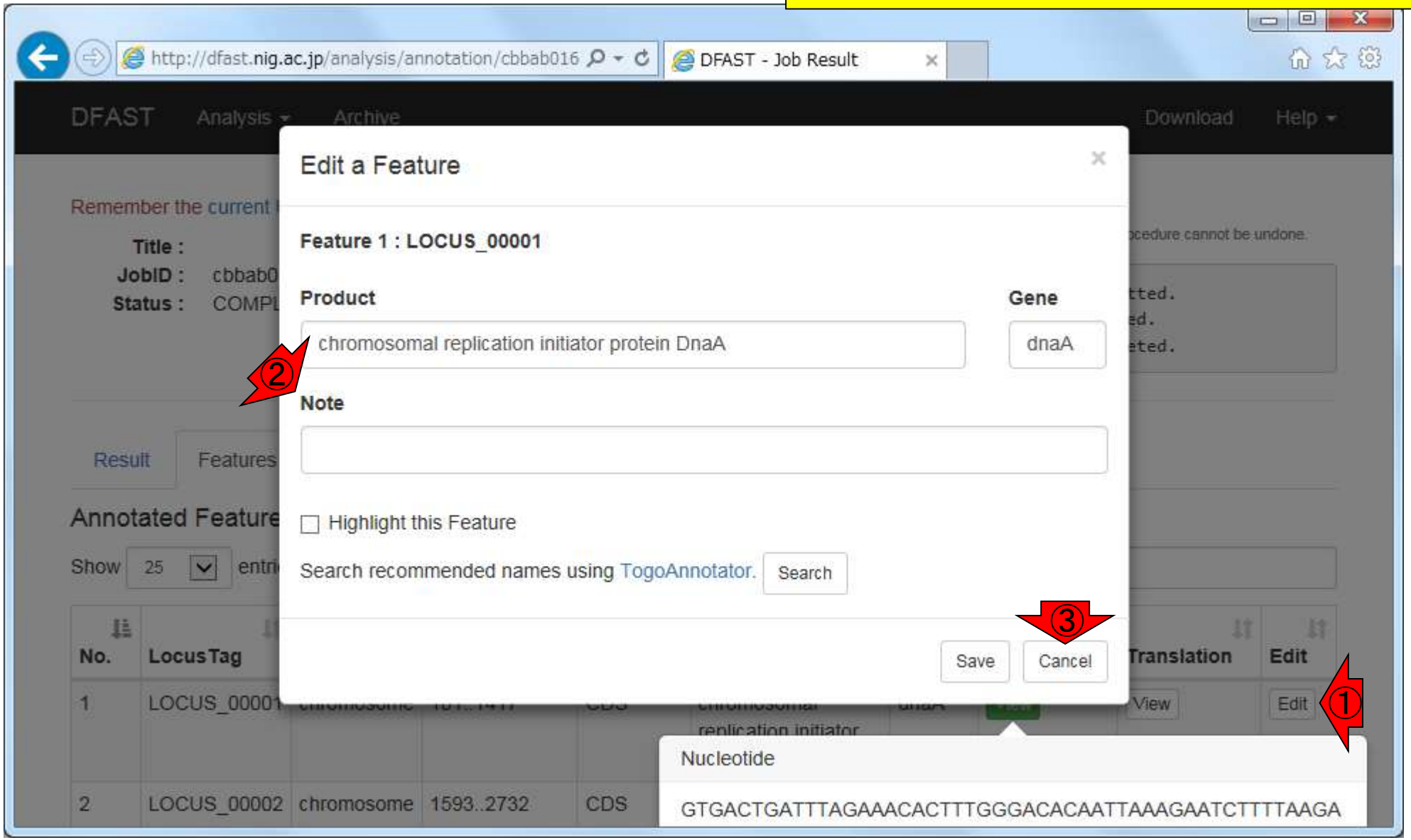
W8-5 : Edit

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



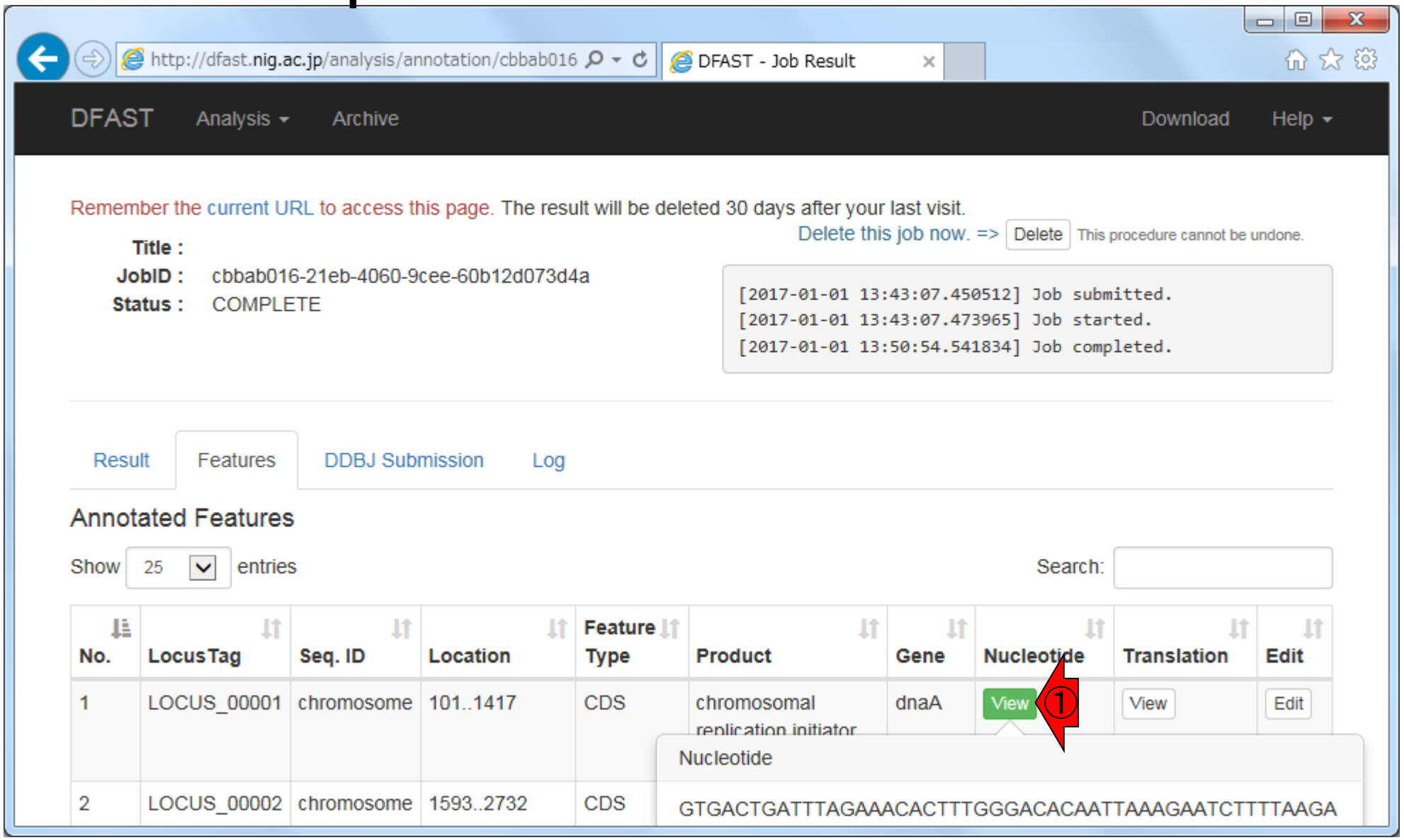
W8-5 : Edit

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



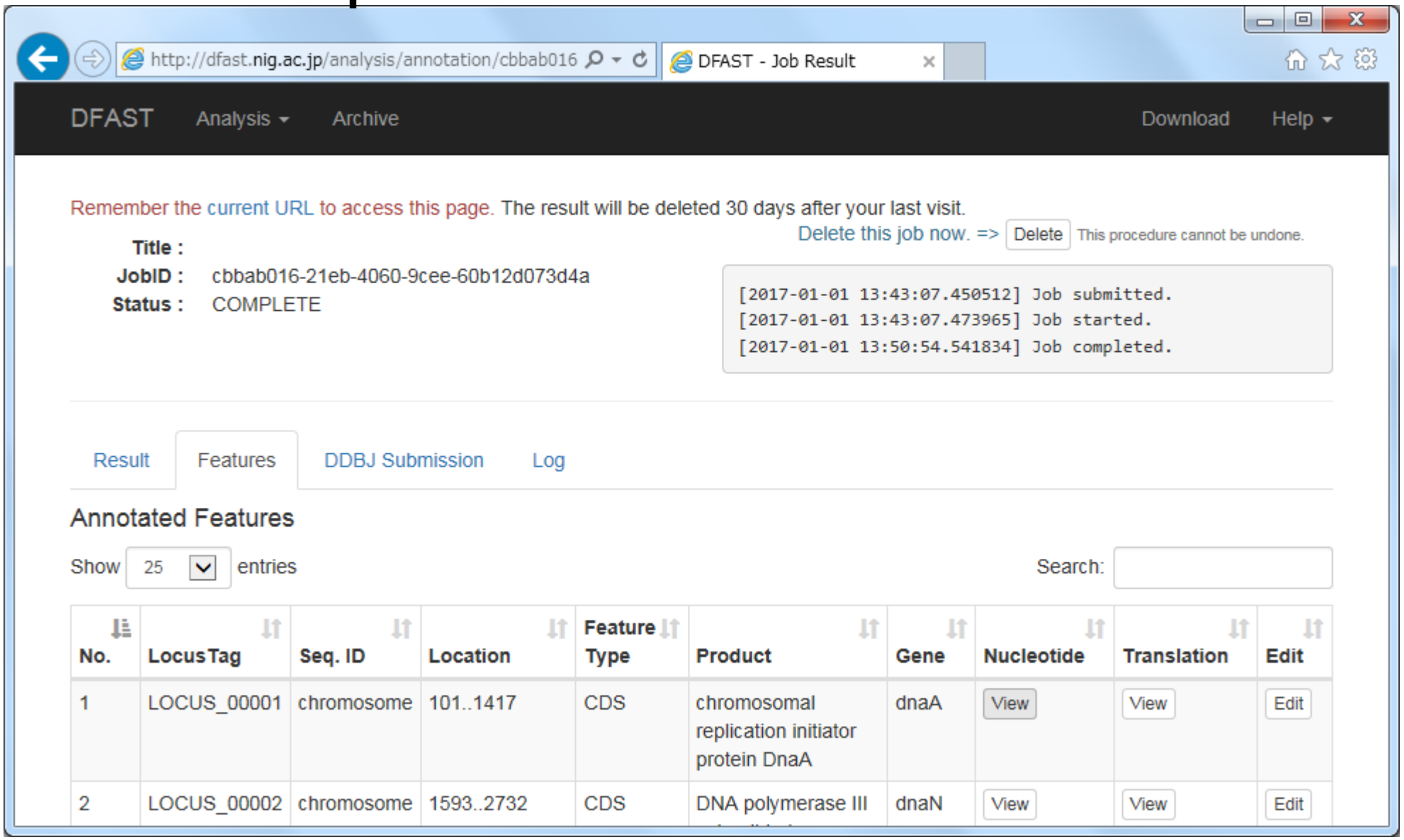
W8-6：Tips

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



ほらね

W8-6：Tips

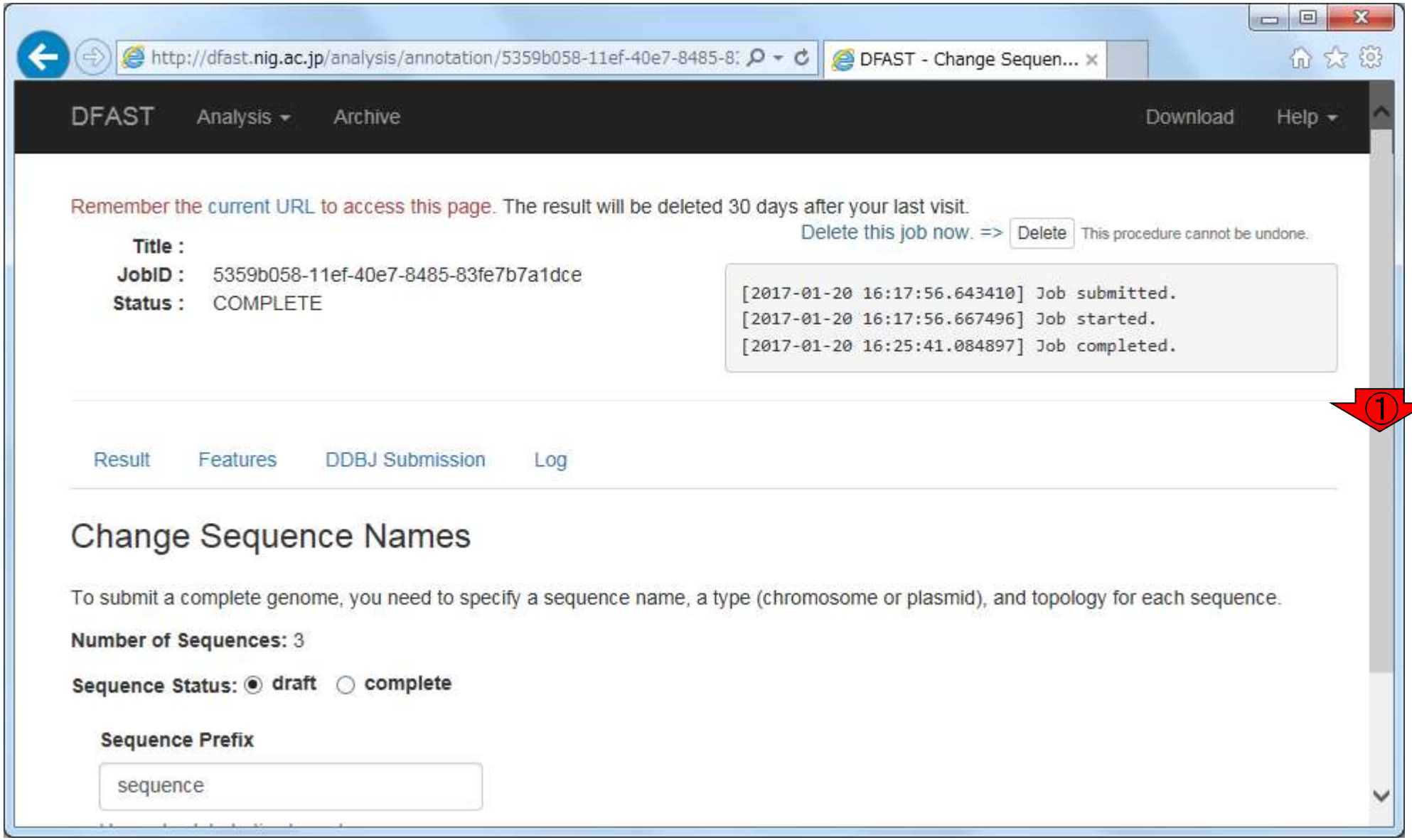

http://dfast.nig.ac.jp/analysis/annotation/cbbab016
DFAST - Job Result
DFAST
Analysis
Archive
Download
Help
Remember the current URL to access this page. The result will be deleted 30 days after your last visit.
Delete this job now. => Delete This procedure cannot be undone.
Title :
JobID : cbbab016-21eb-4060-9cee-60b12d073d4a
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

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

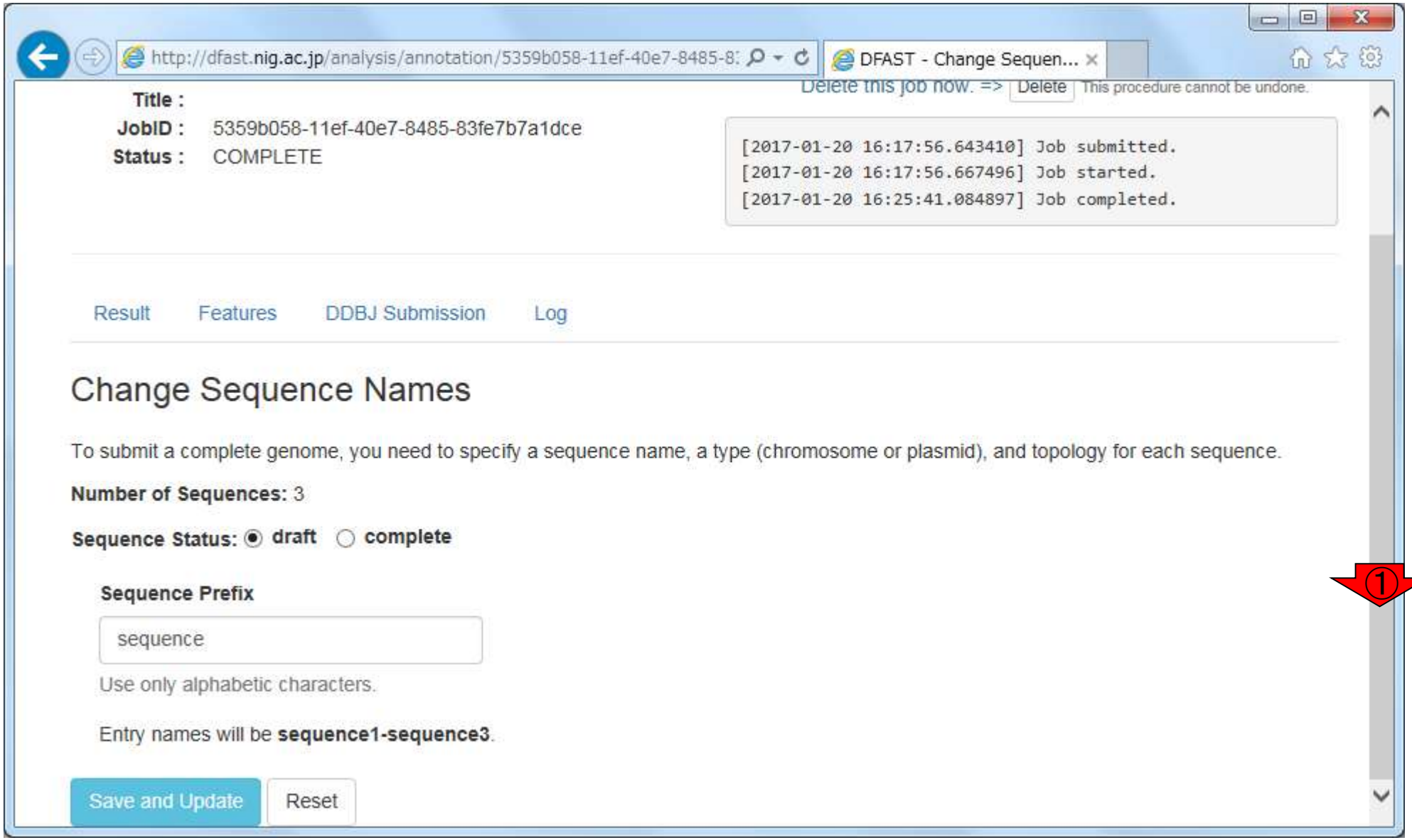
W9-1：DDBJ Submission

「こちらのページ」のリンク先。①ページ下部まで移動



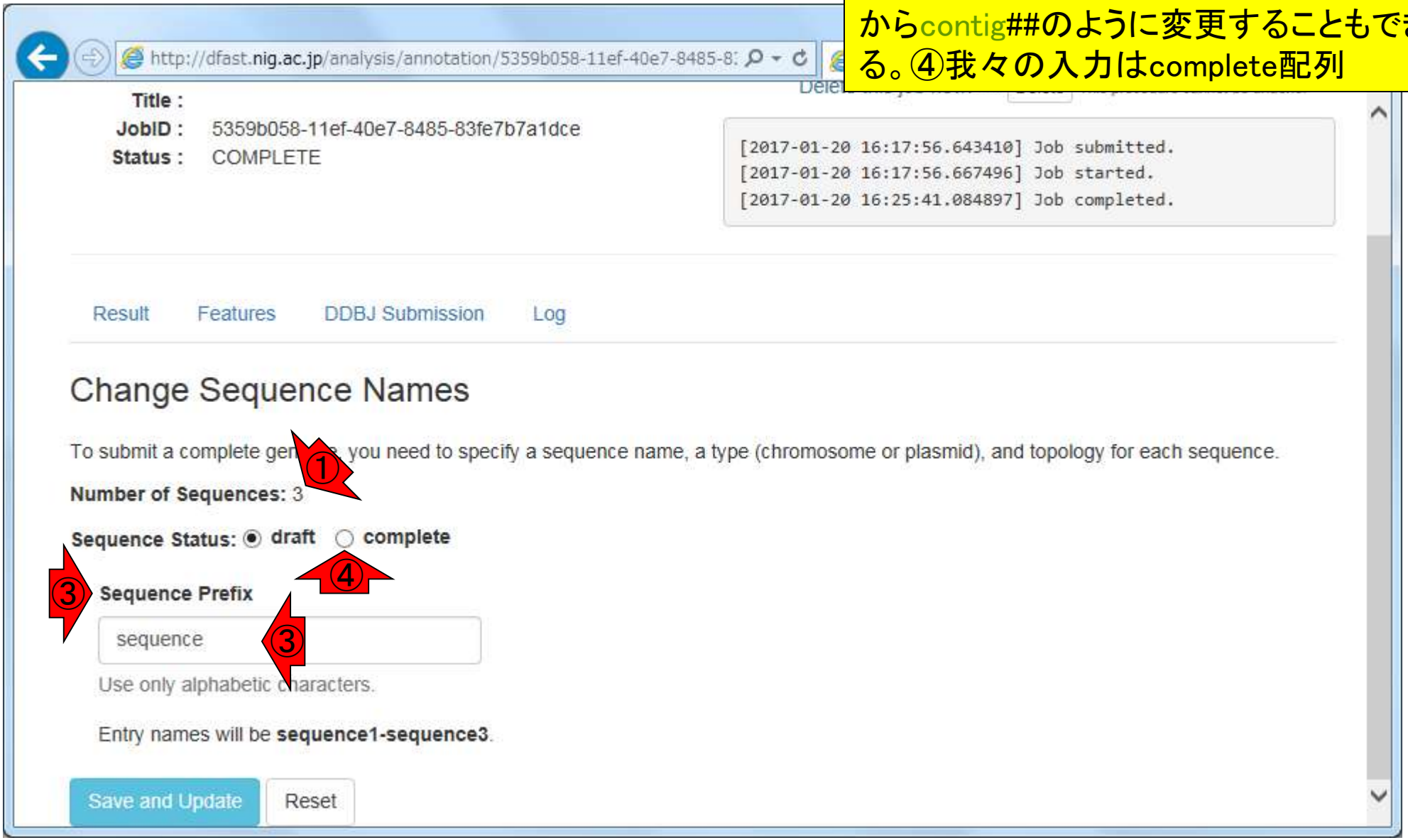
①ページ下部まで移動した状態

W9-1：DDBJ Submission



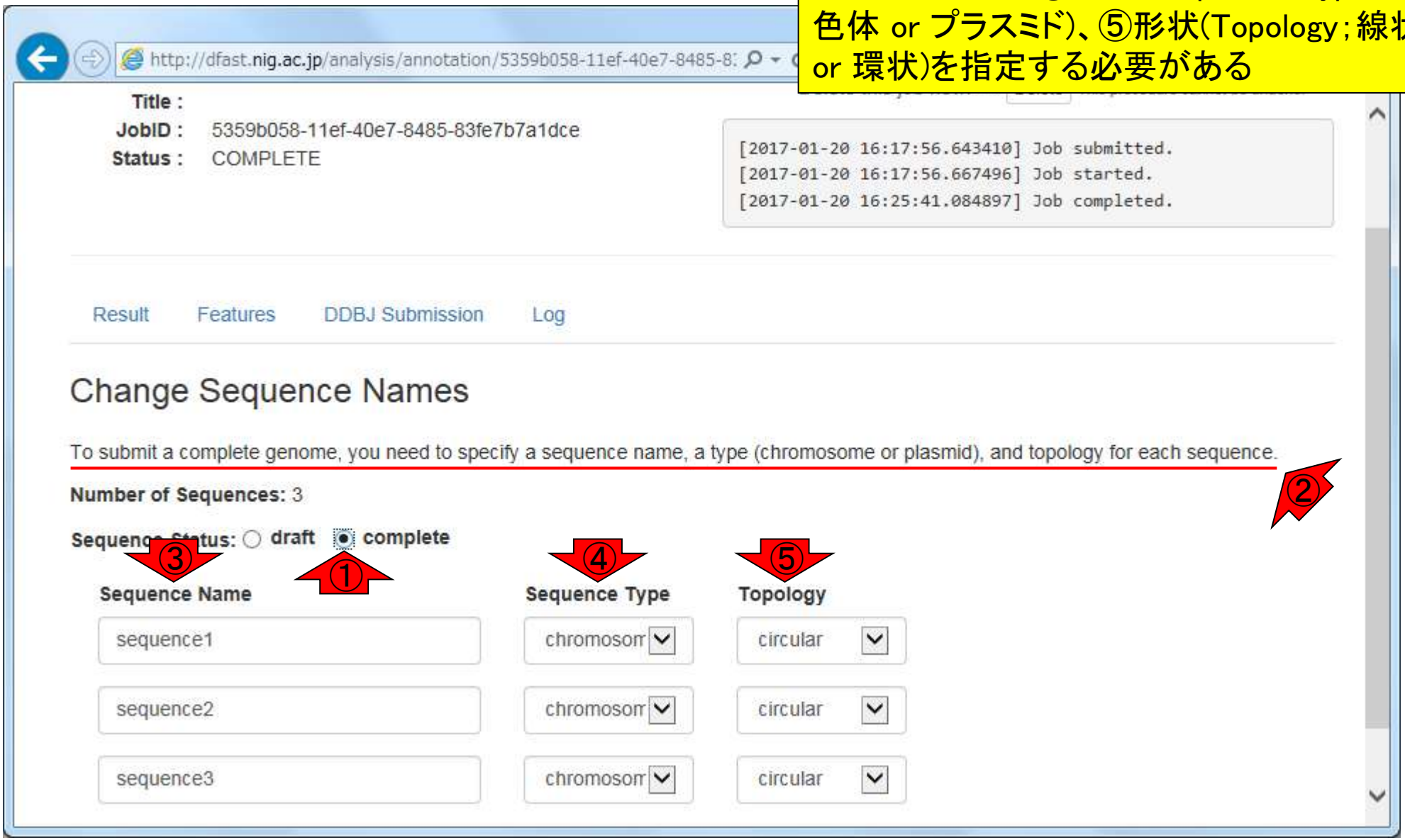
W9-2：配列名の変更

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



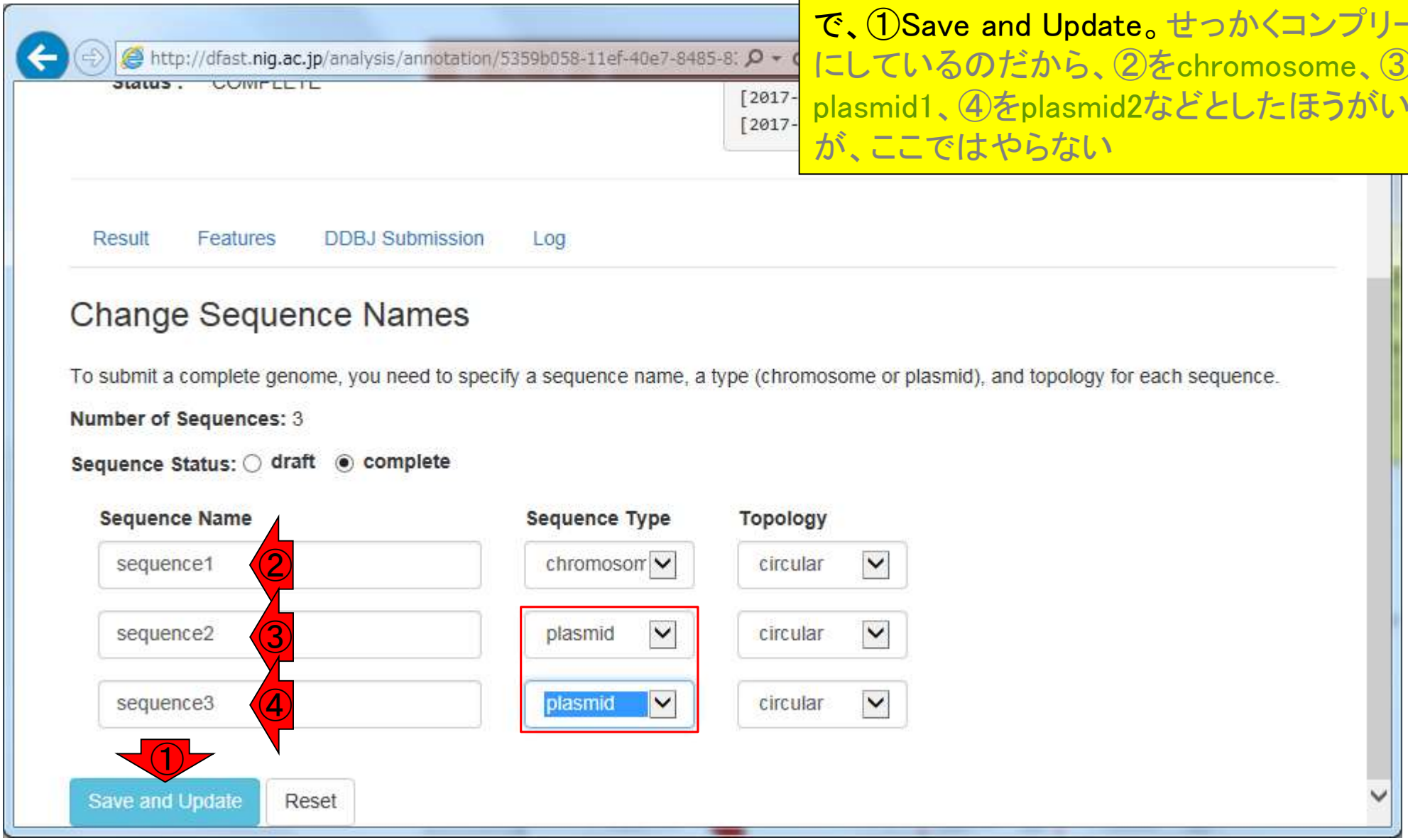
W9-2：配列名の変更

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



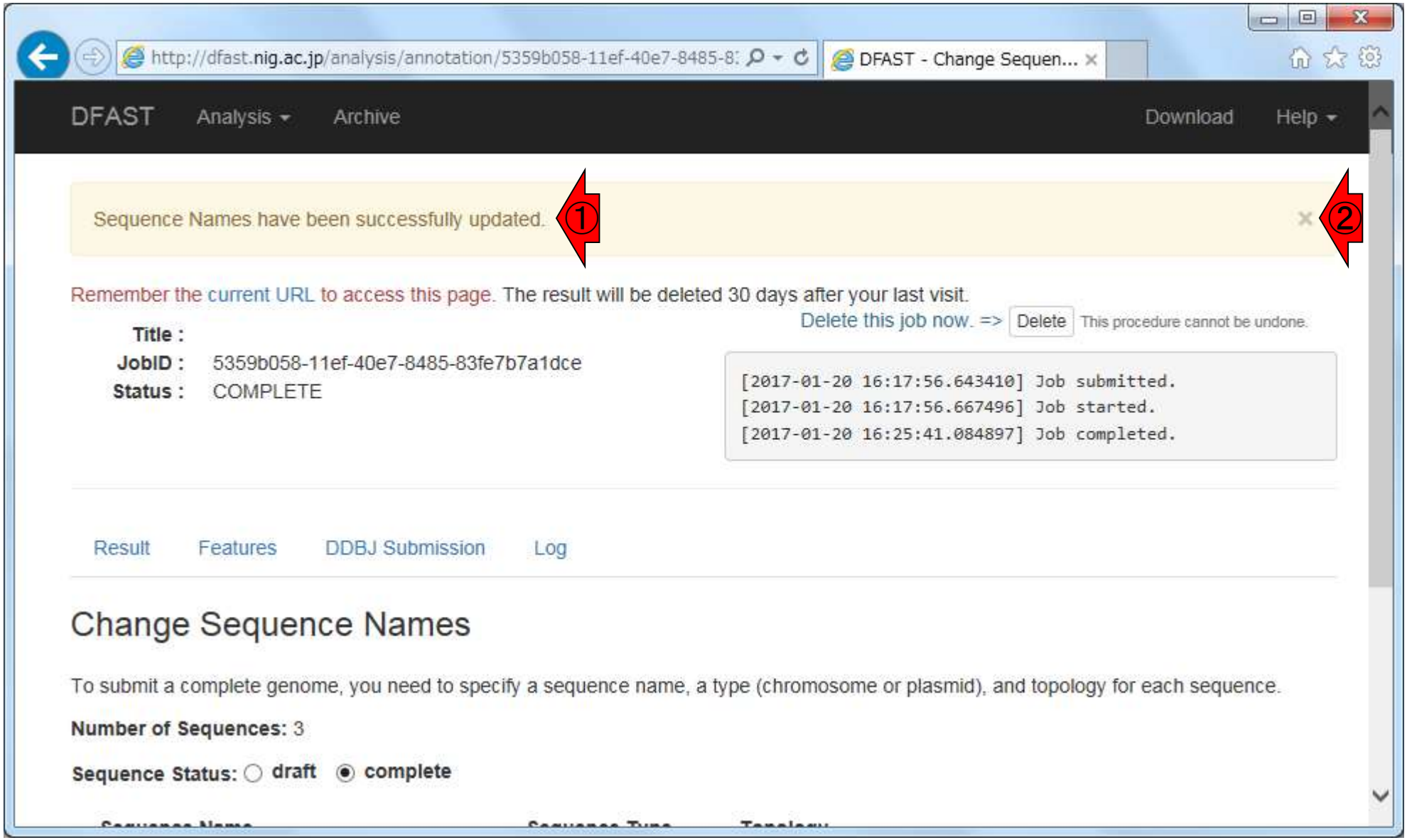
W9-2：配列名の変更

（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



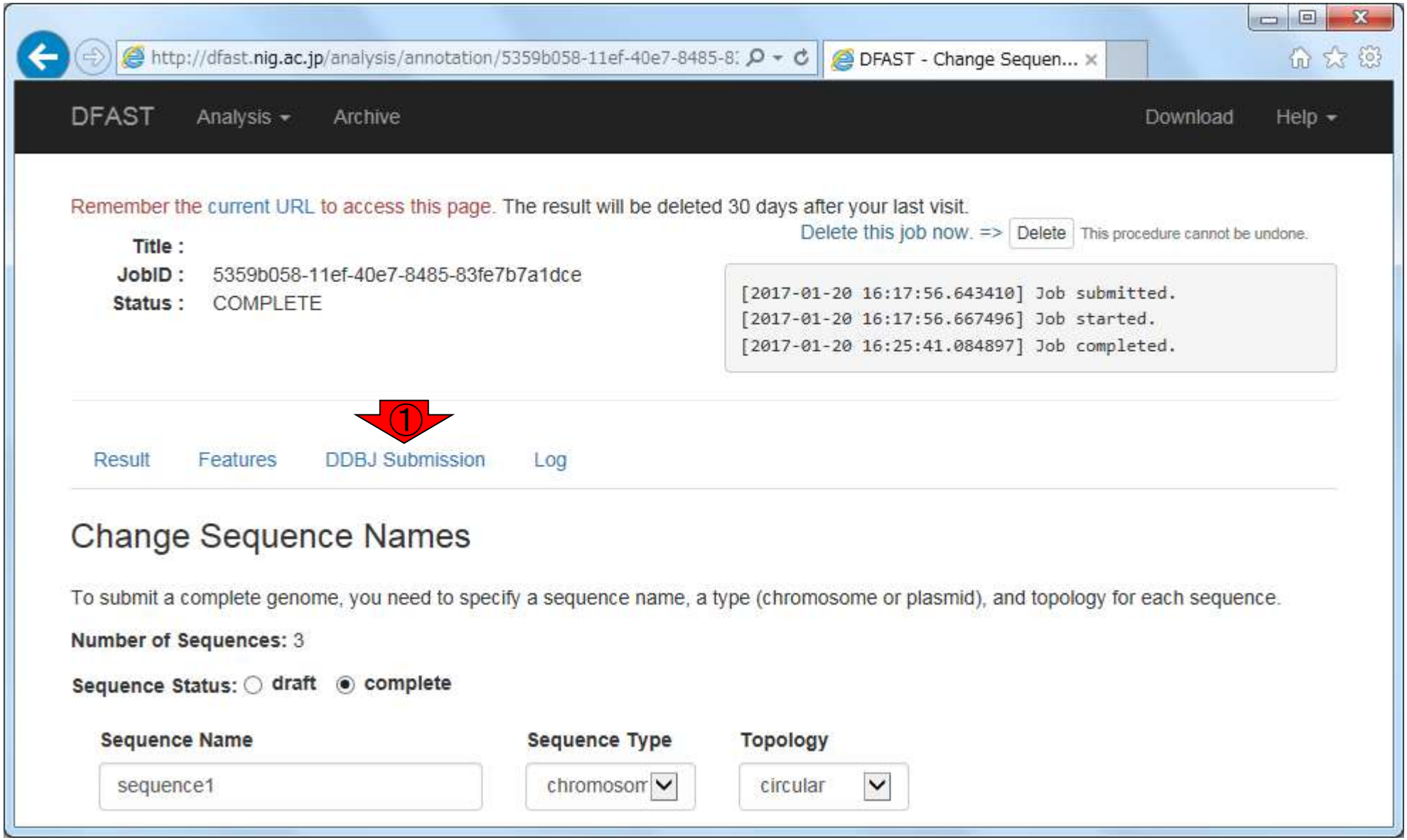
①正しく反映されたようだ。②×

W9-2：配列名の変更



①DDBJ Submissionタブをクリック

W9-2：配列名の変更

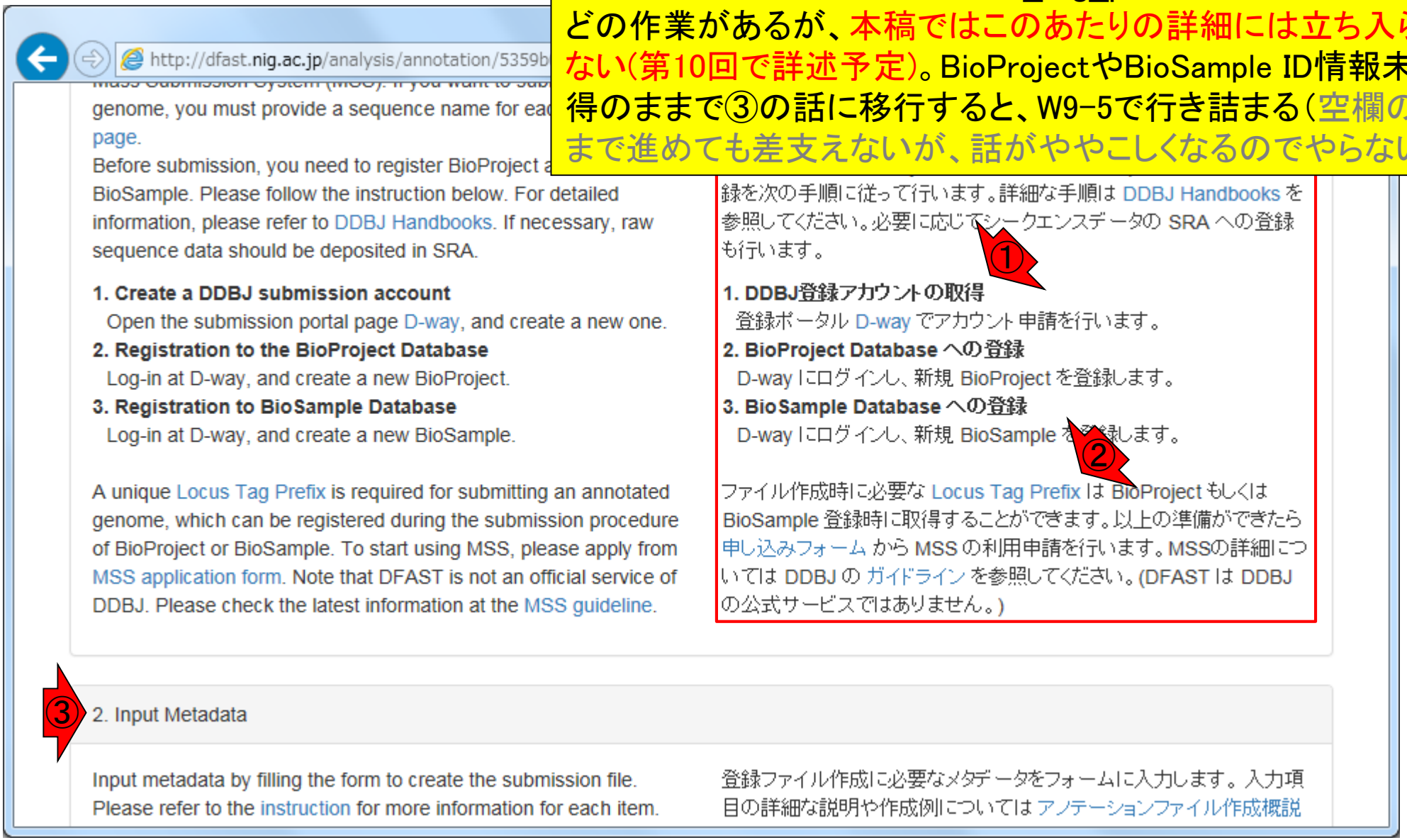


①ページ下部に移動し、、、

W9-3：アカウント取得など

W9-3：アカウン

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情報未取得のままで③の話に移行すると、W9-5で行き詰まる(空欄のまま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)



W9-4：Input Metadata

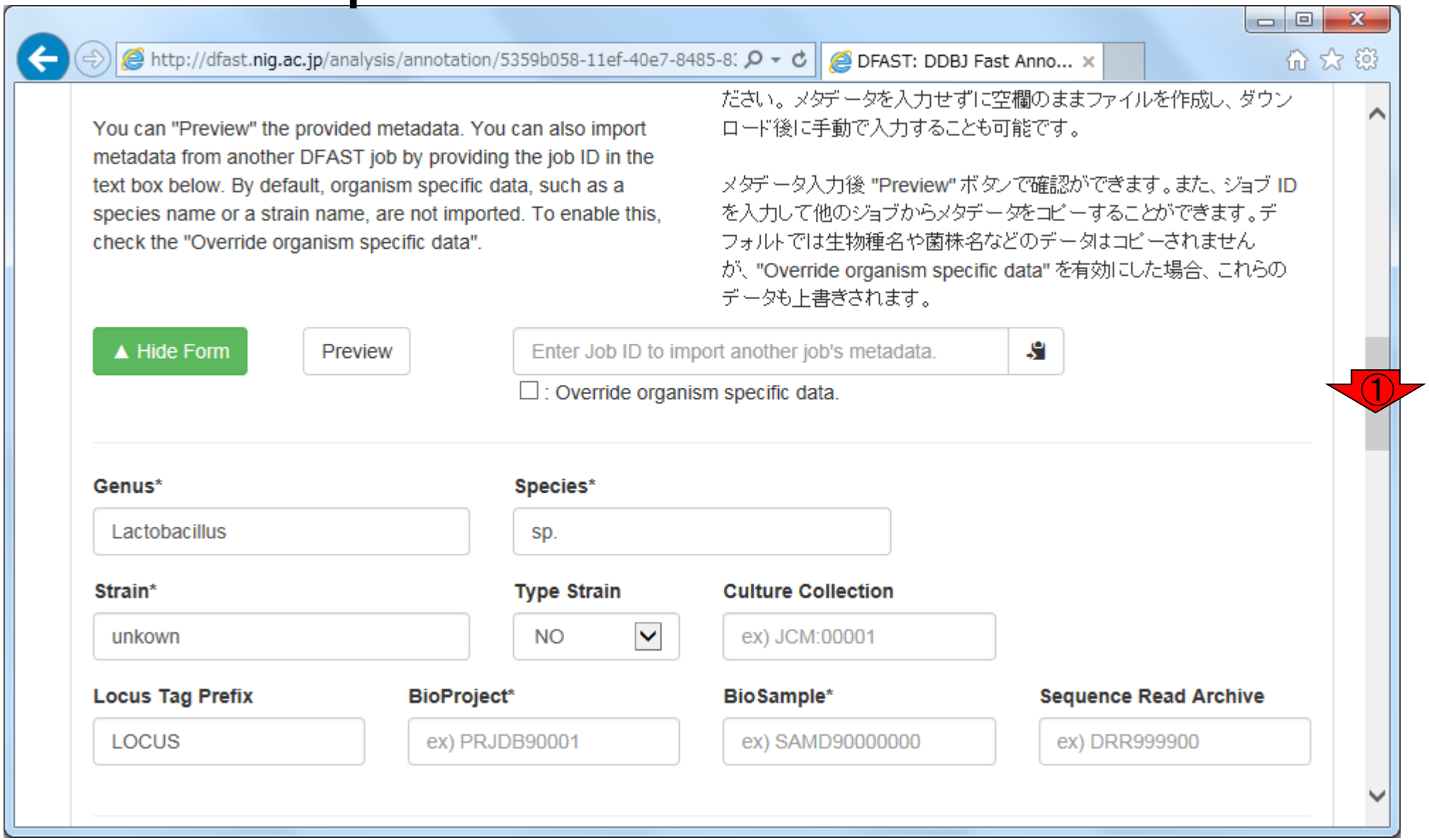
①「2. Input Metadata」の
②Show Formをクリック

W9-4：Input Metadata

入力フォームが下のほうに見えているっぽいので、、、

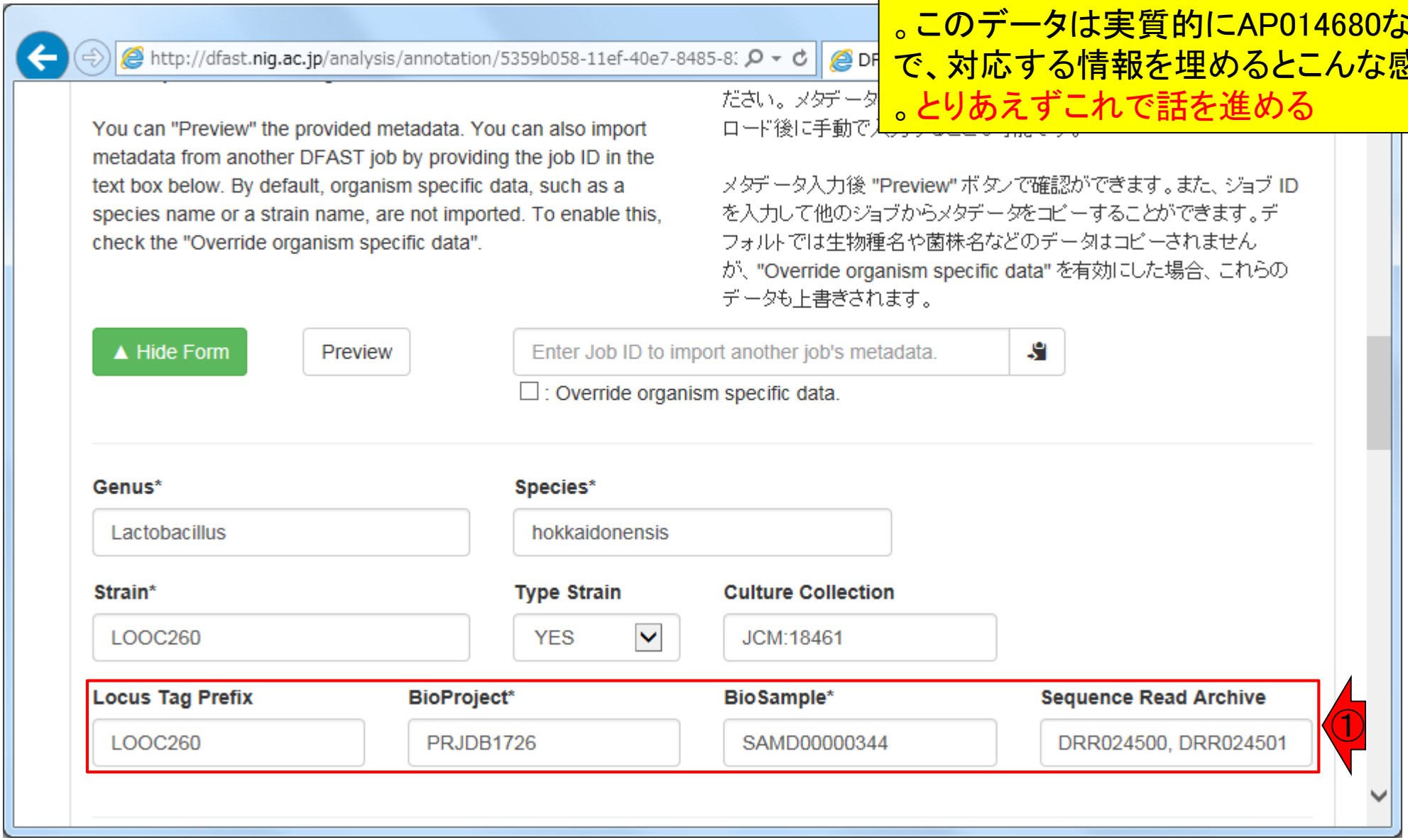
①少しページ下部に移動したところ

W9-4：Input Metadata



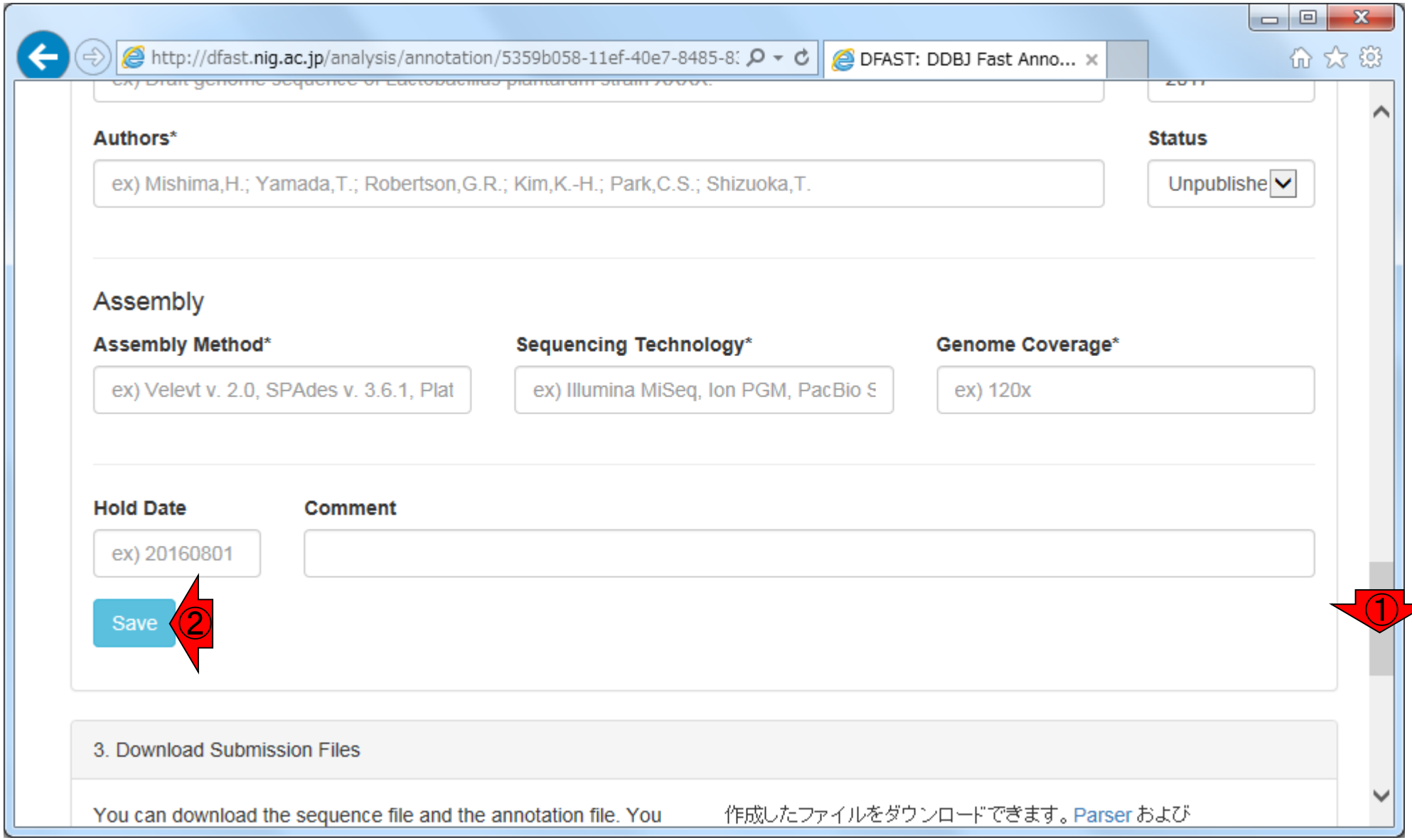
W9-5：情報未取得だと…

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



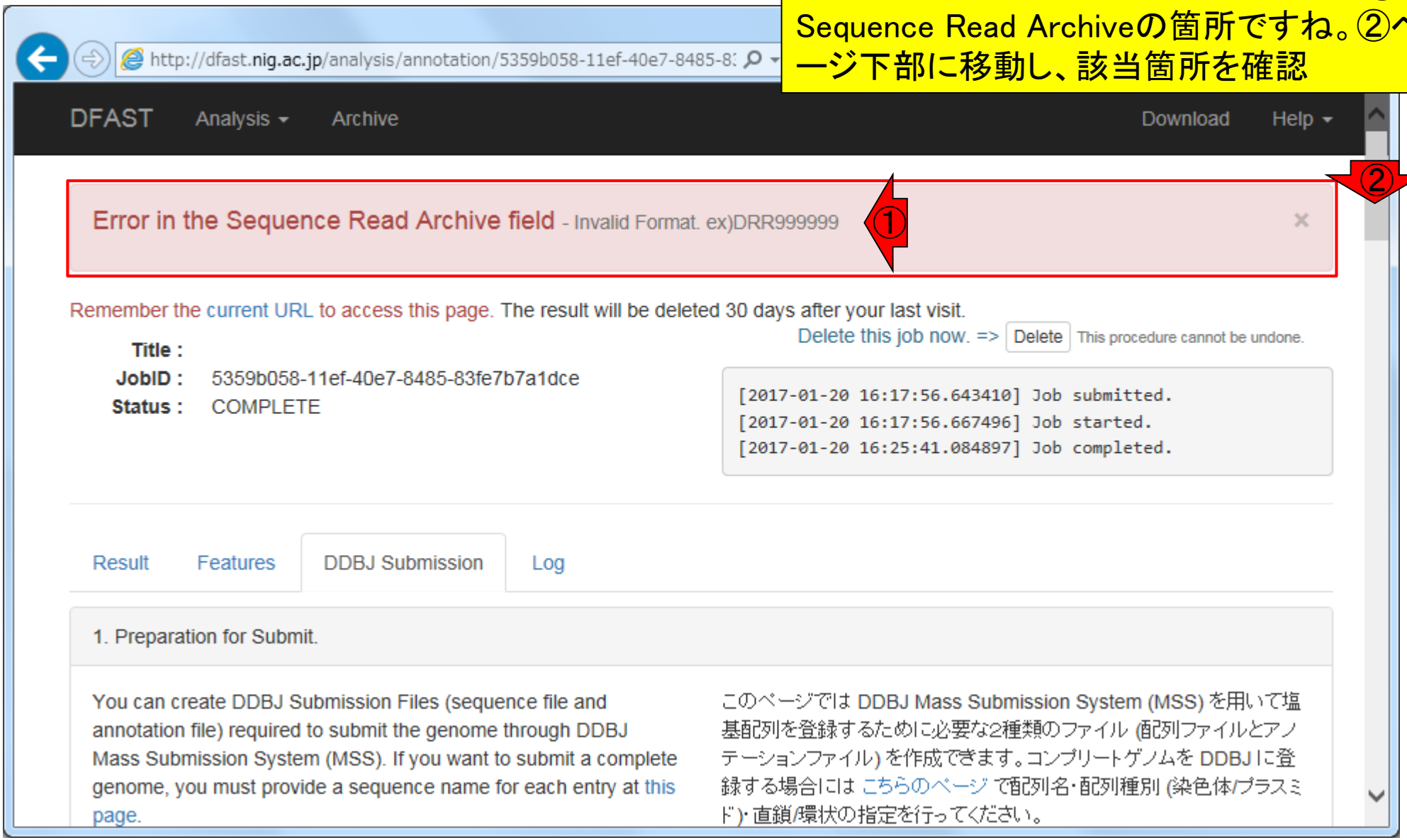
W10-1：適宜Save

①ページ下部に移動して、
適宜②Saveしておきましょう



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。①Sequence 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

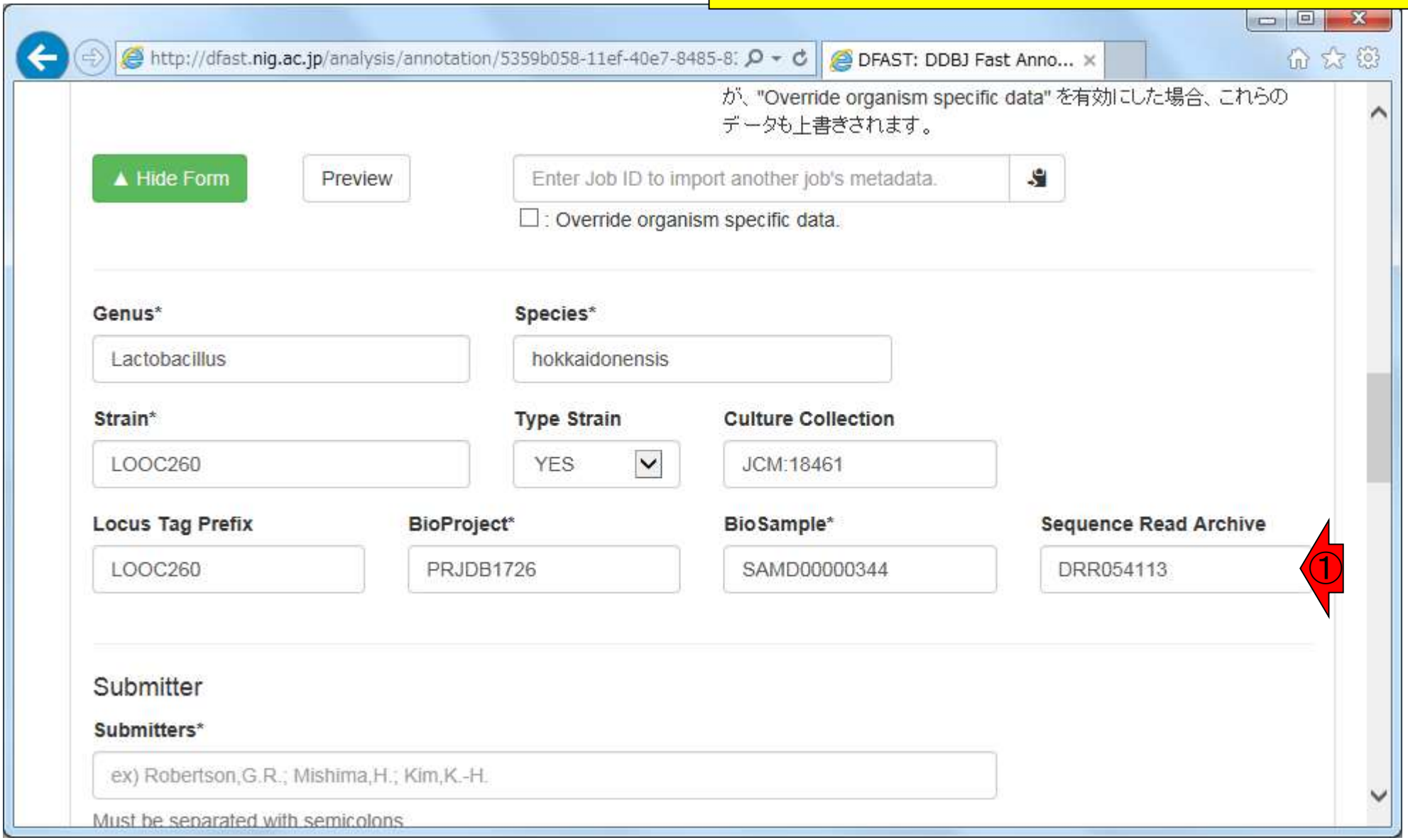


W10-2：Error

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

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

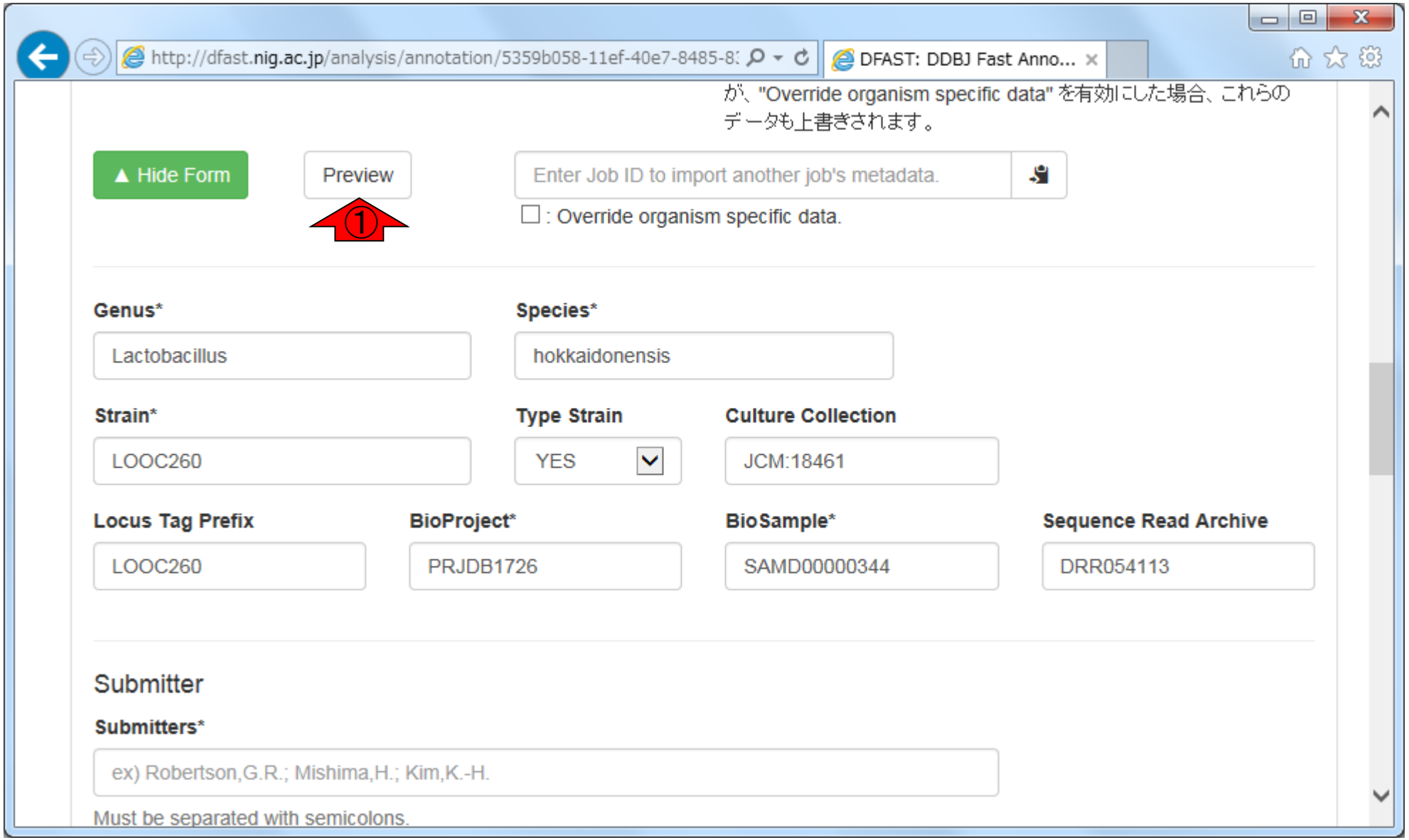


無事エラーが消えた!

W10-3：Error対処

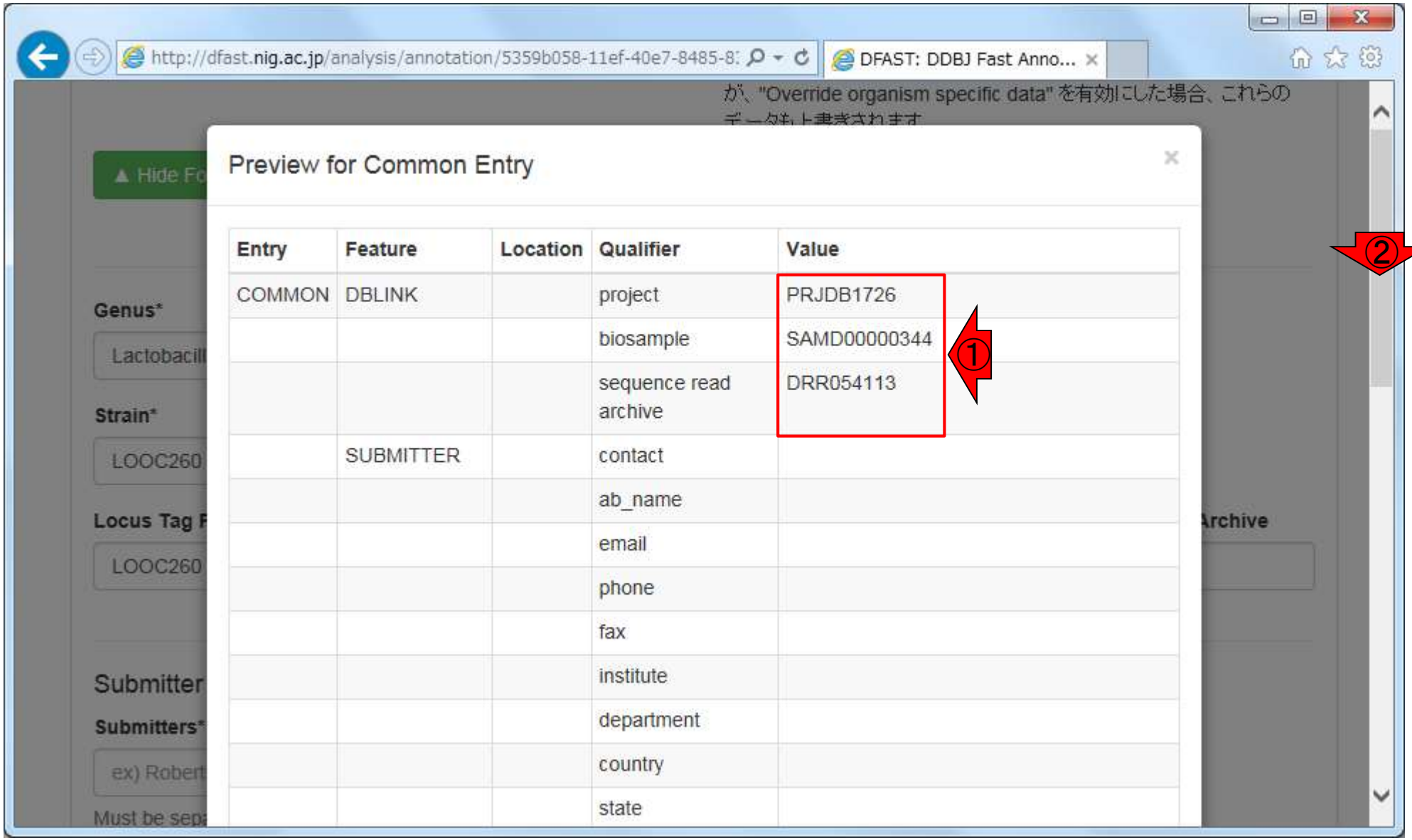
W10-4：Preview

①Previewボタンを押して、
これまでの入力内容を確認



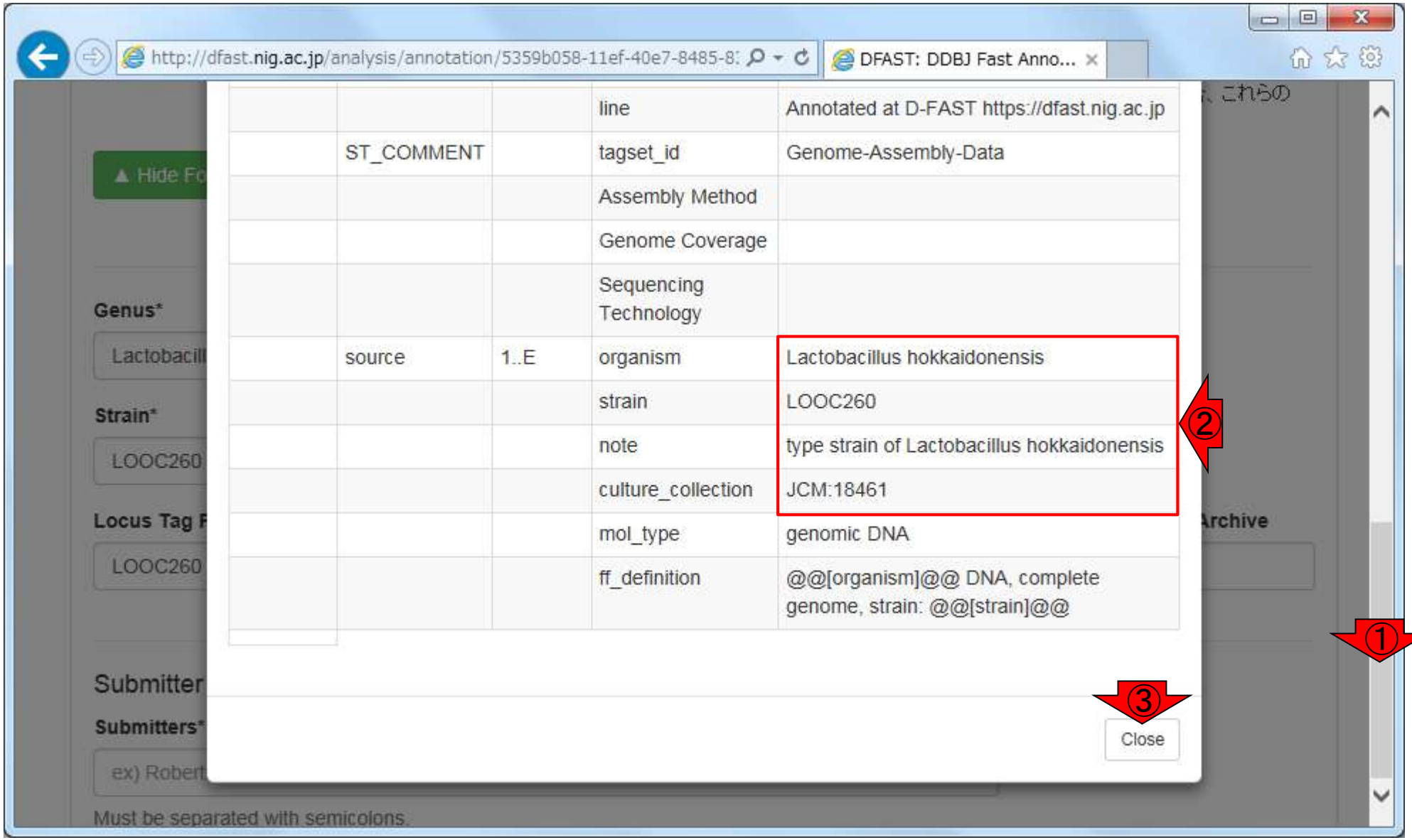
W10-4：Preview

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



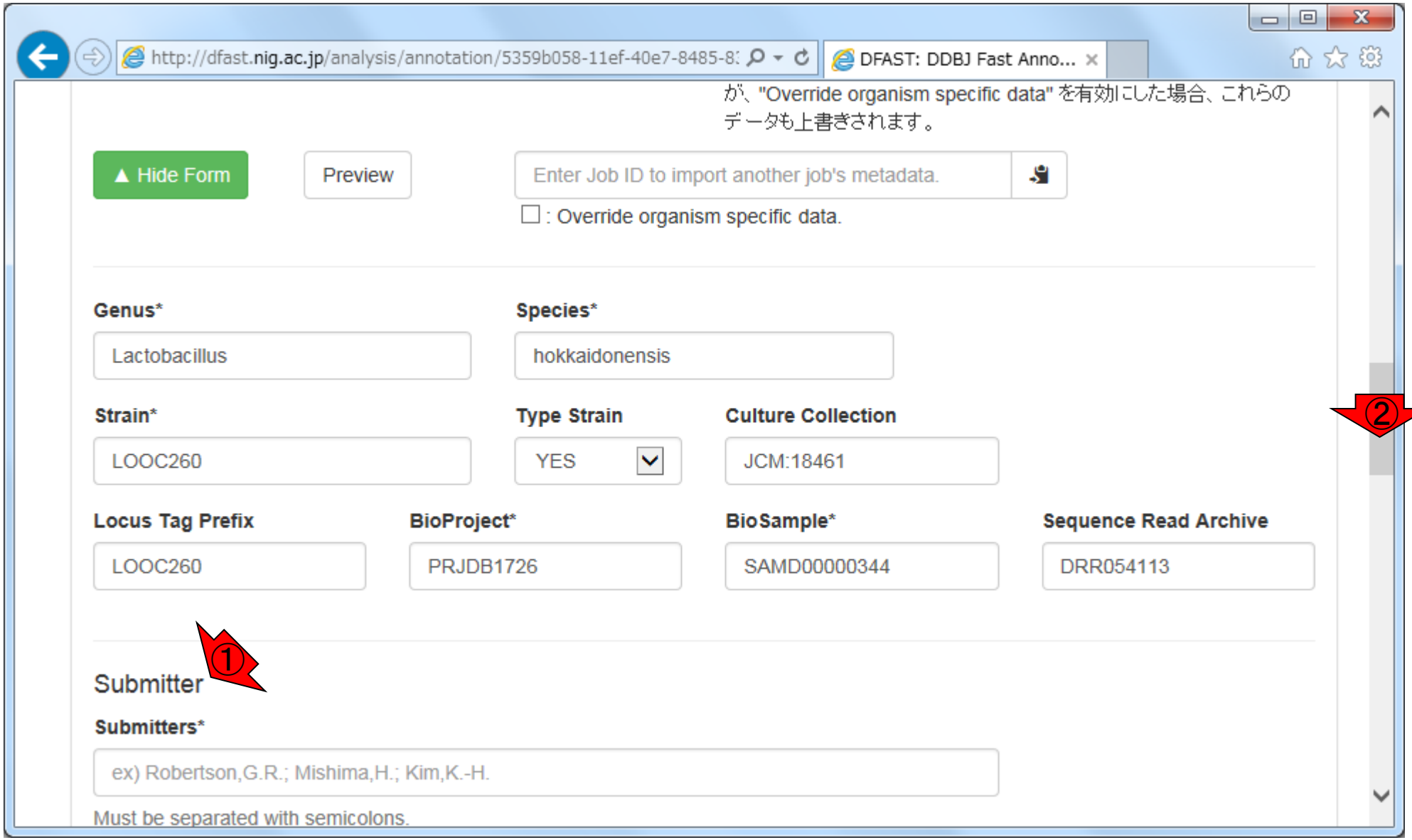
W10-4：Preview

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



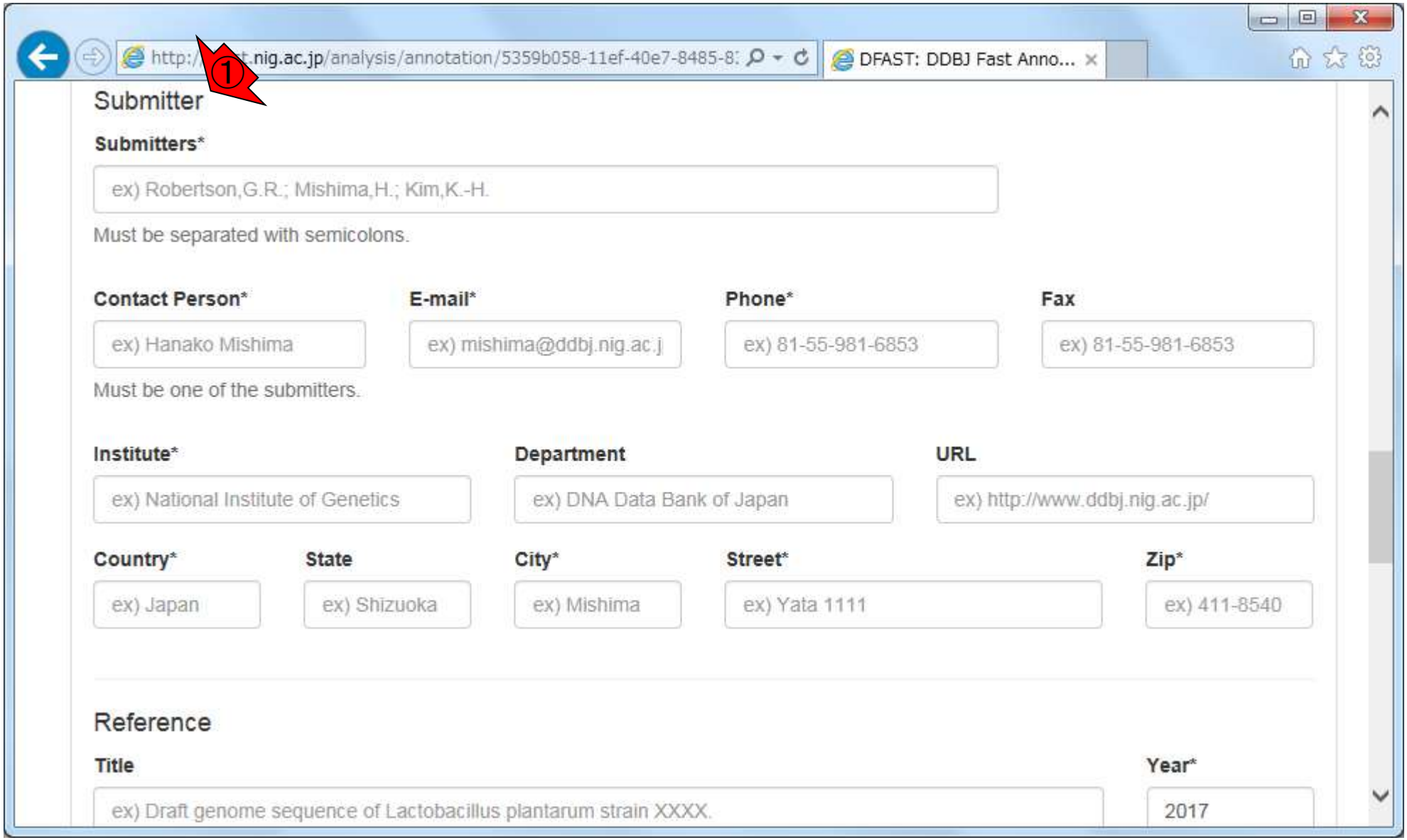
W10-5：Submitter

①次はSubmitter情報。②ちょっとページ下部に移動



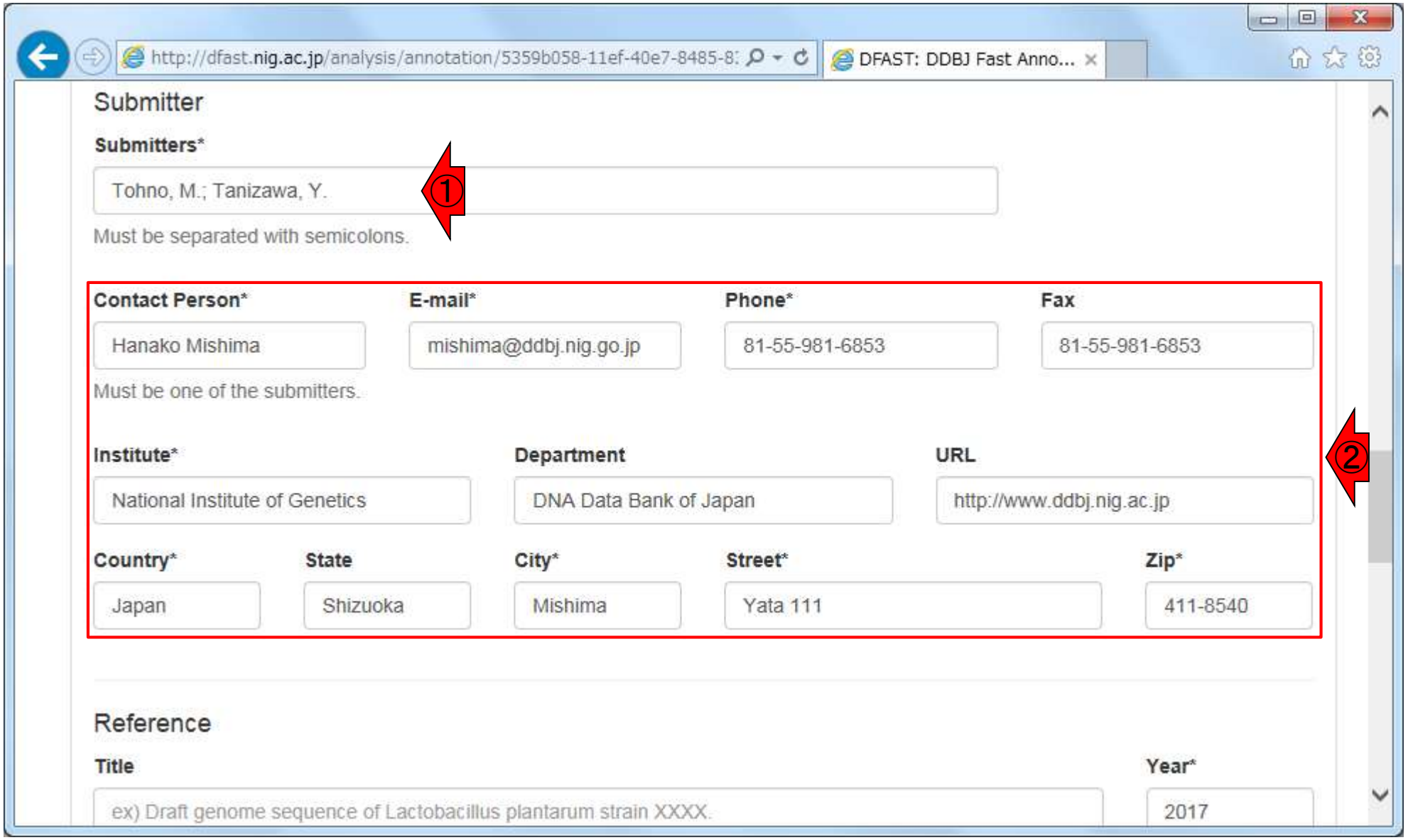
W10-5：Submitter

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



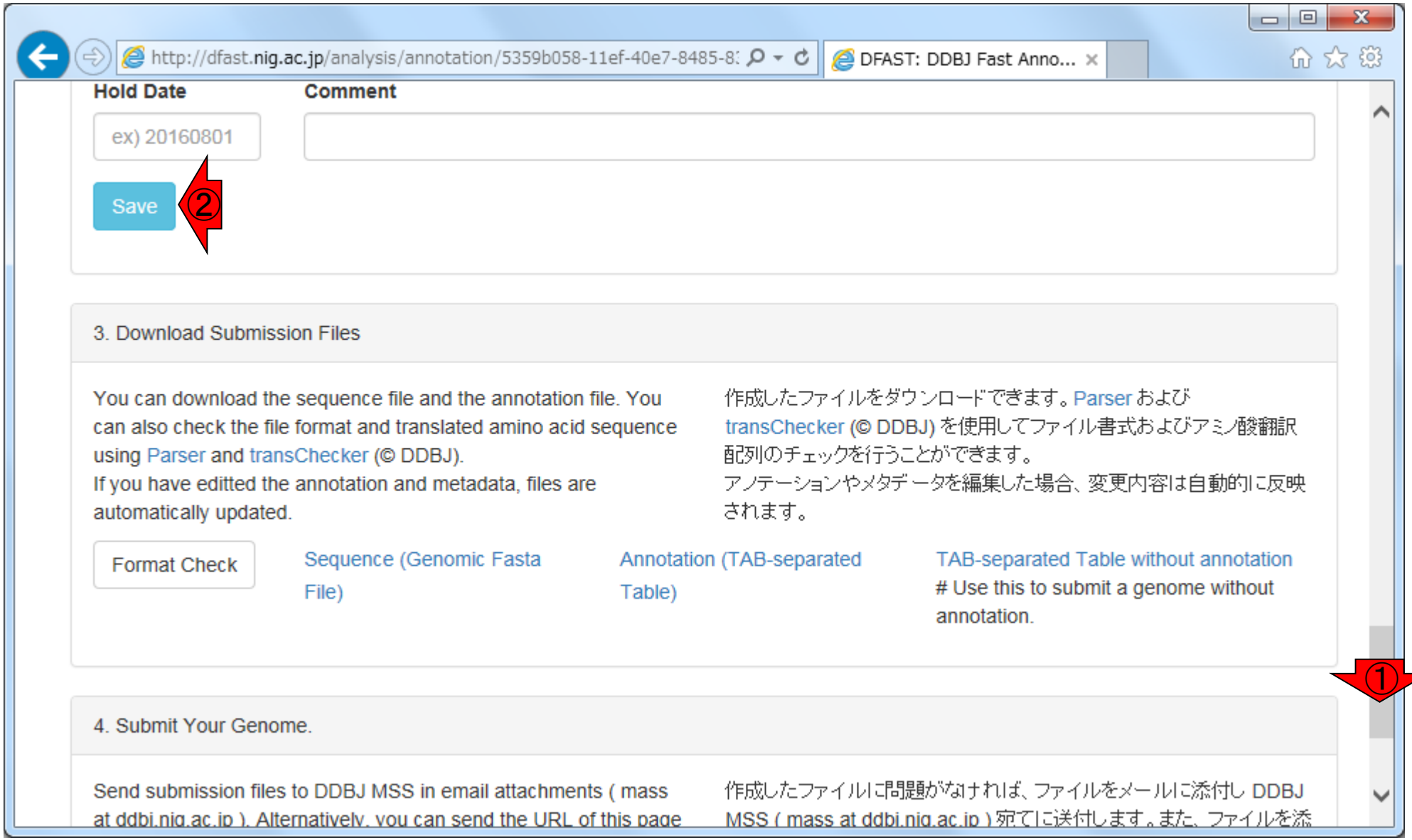
W10-5：Submitter

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



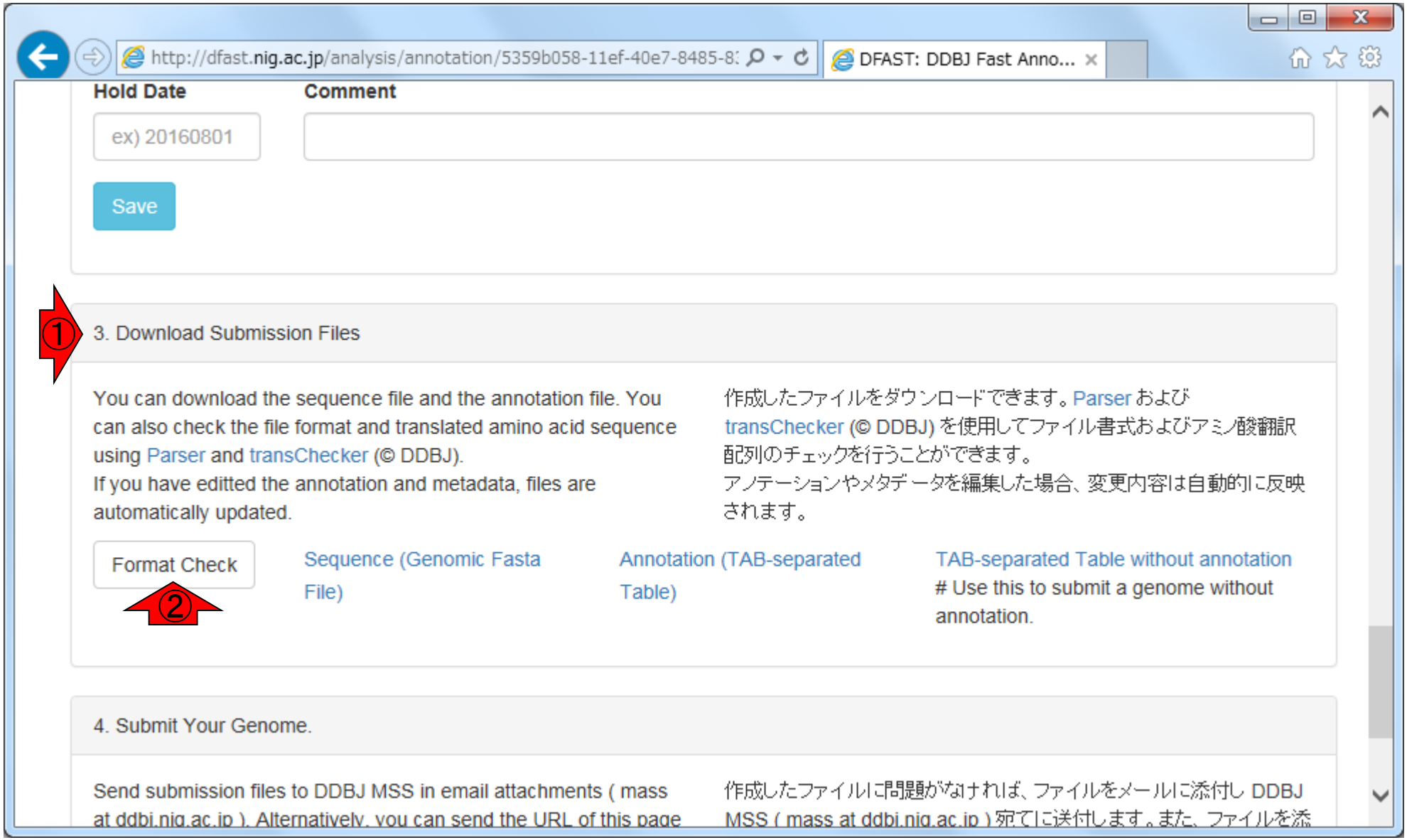
W10-5：Submitter

①このあたりまで移動。
②適宜Saveしましょう



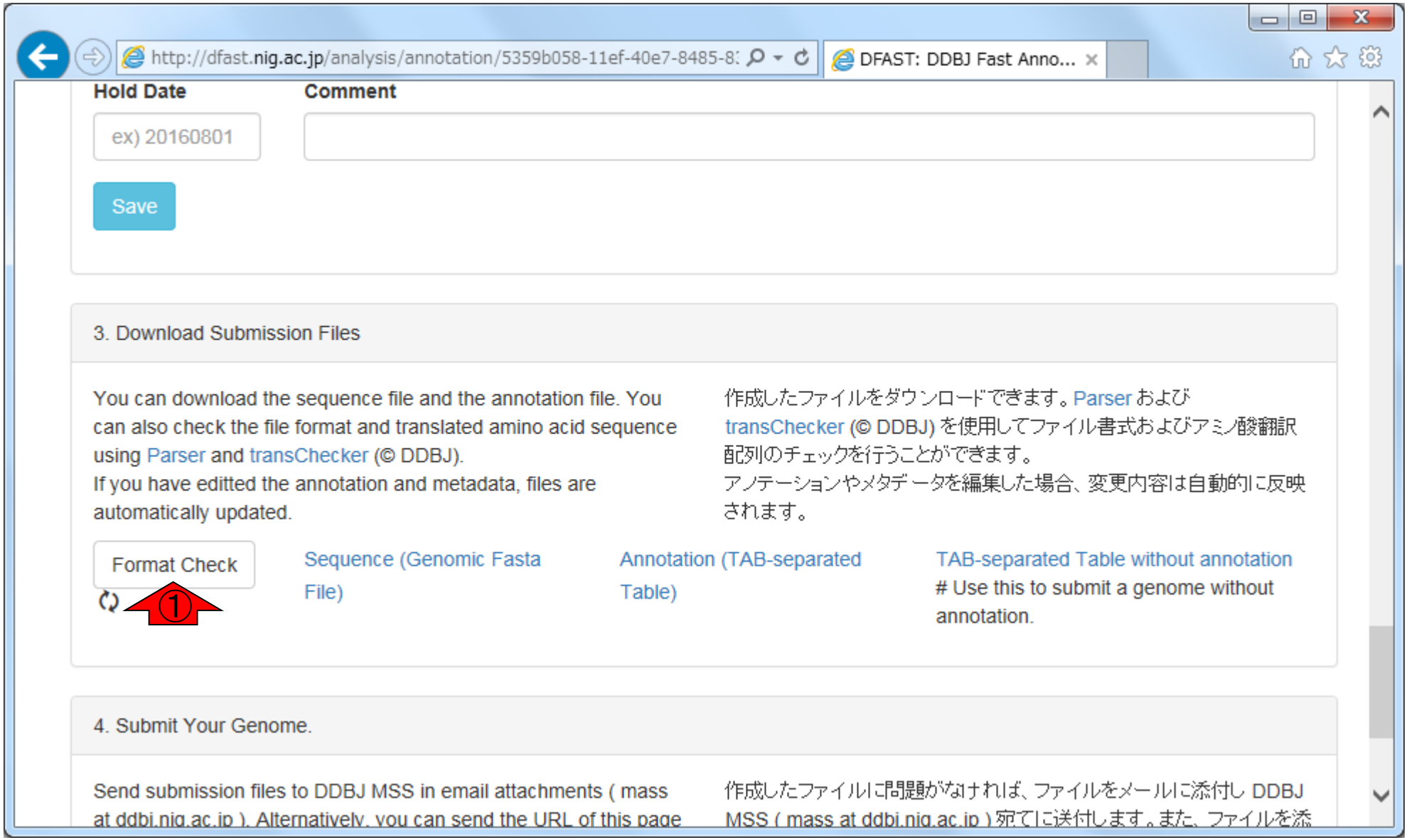
W10-6：Format Check

①「3. Download Submission Files」の、②Format Check



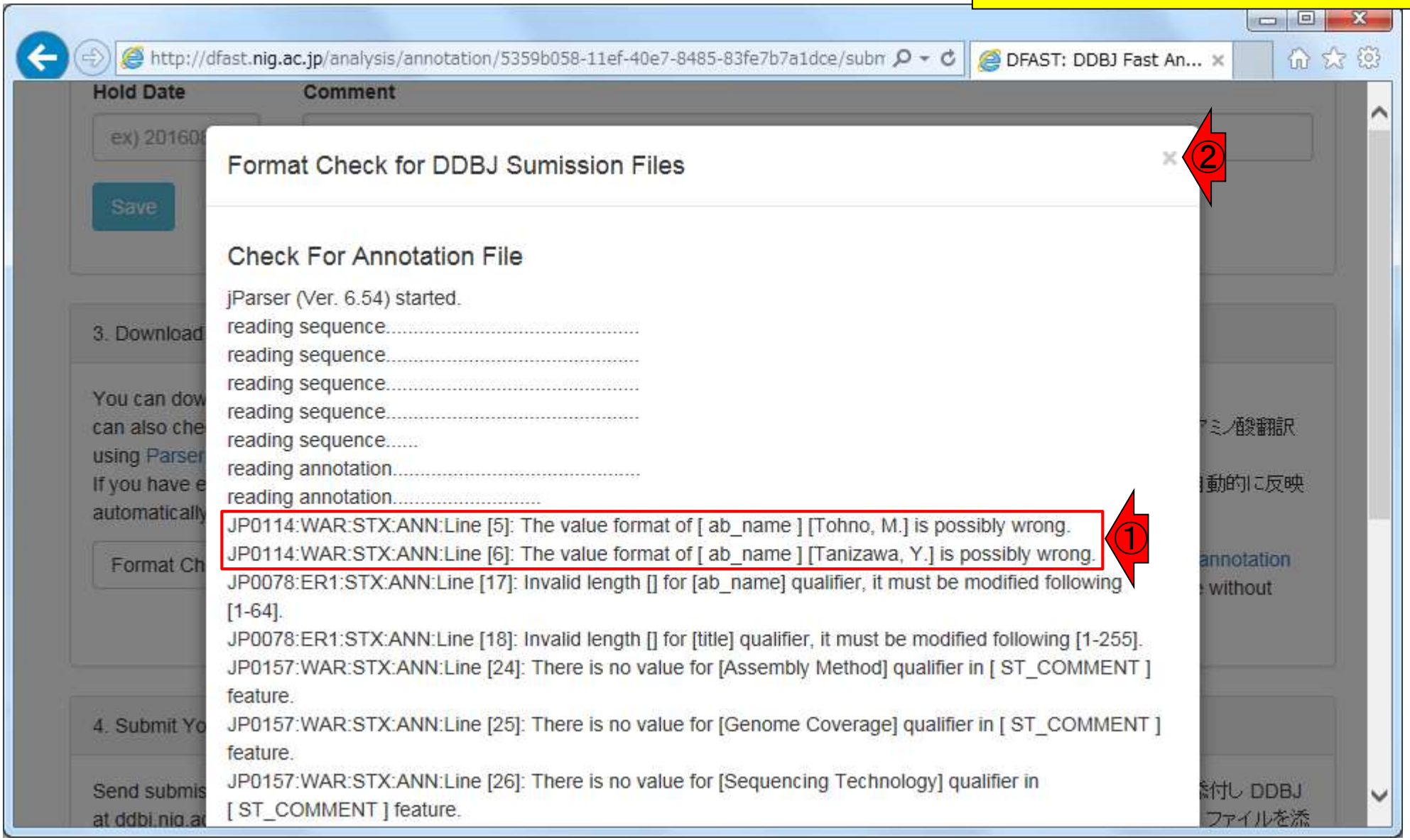
①Format Checkボタンを押した直後の状態

W10-6：Format Check



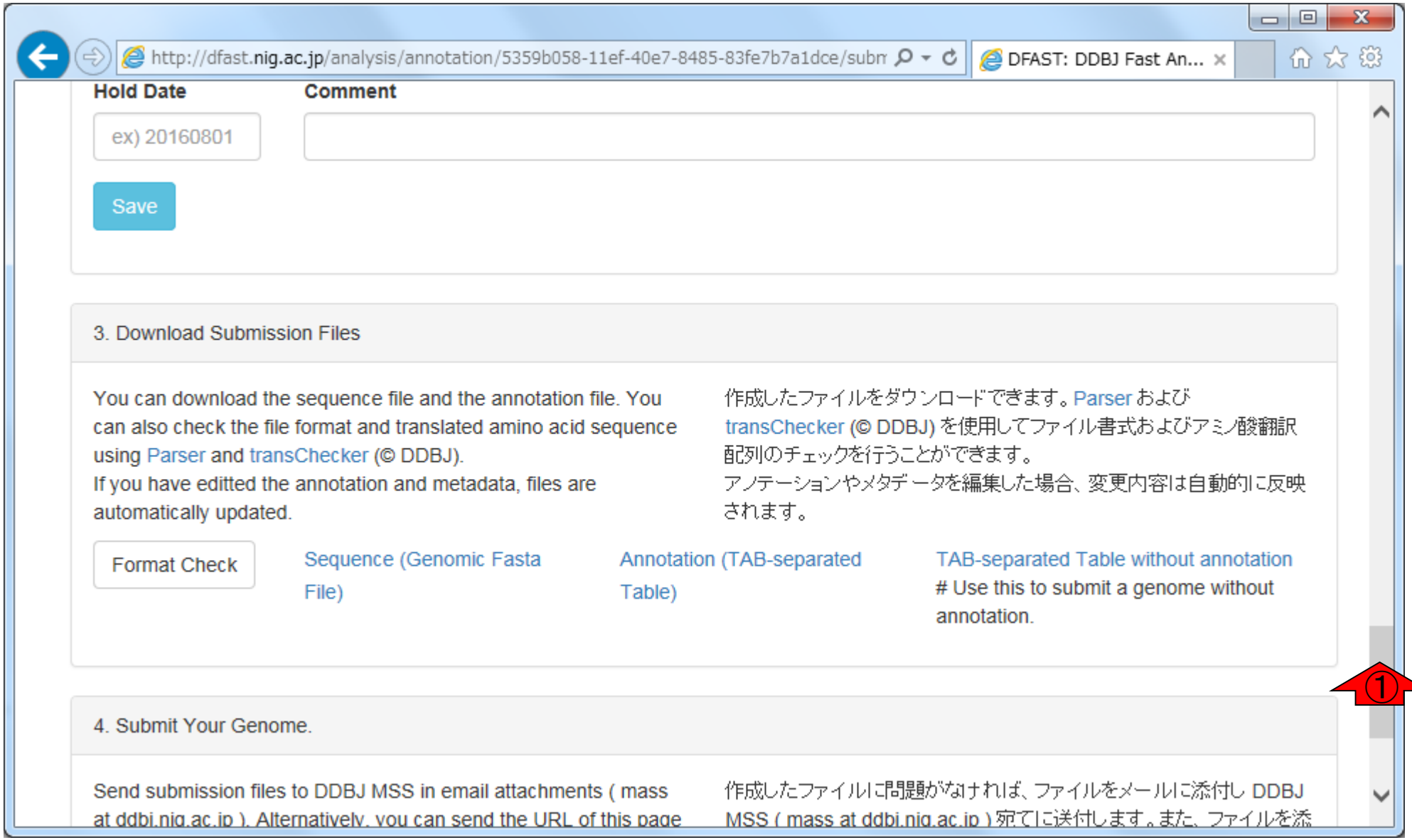
W10-6：Format Check

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



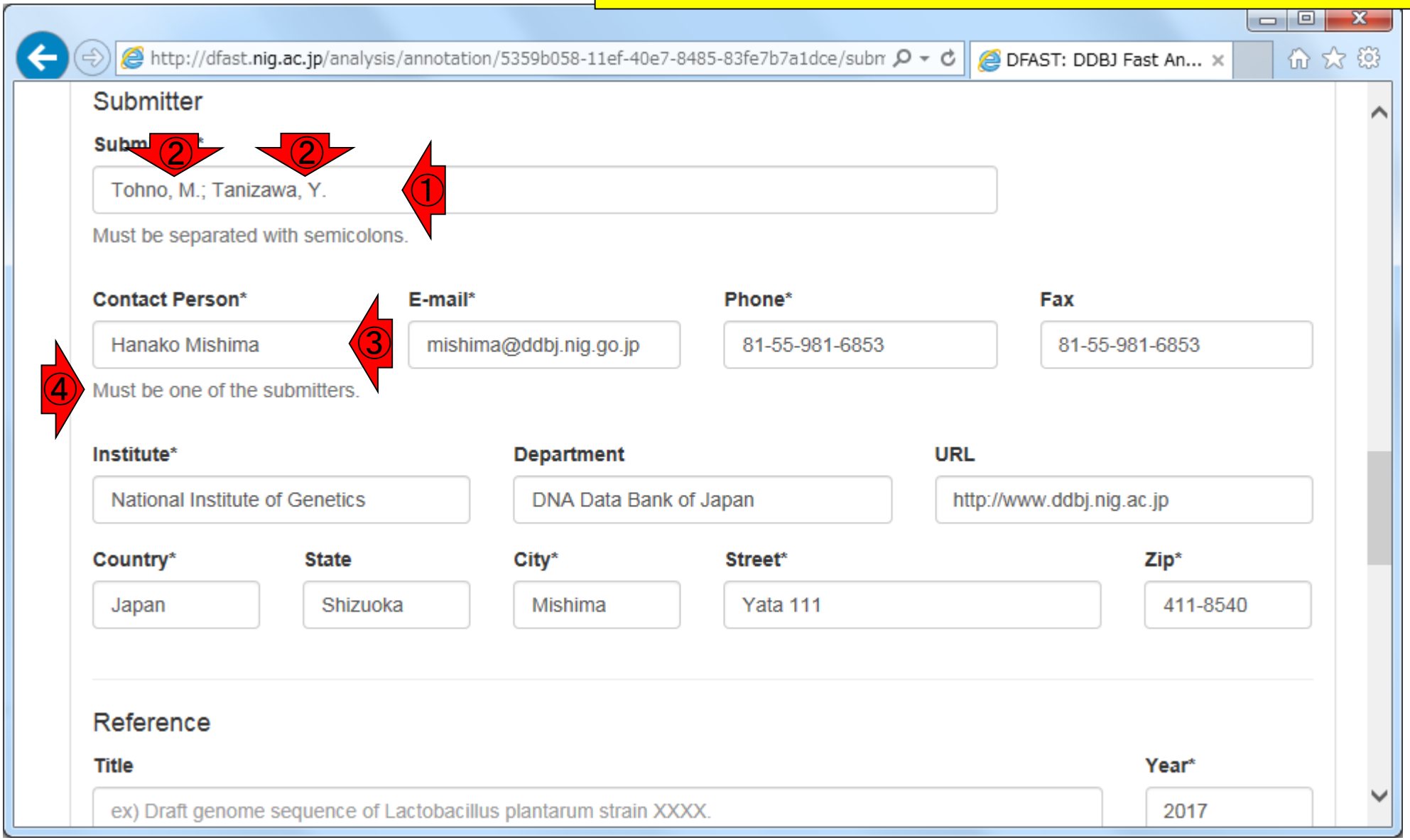
W10-7：修正

Submitters情報のあたりを見るべく、①ページ上部に移動



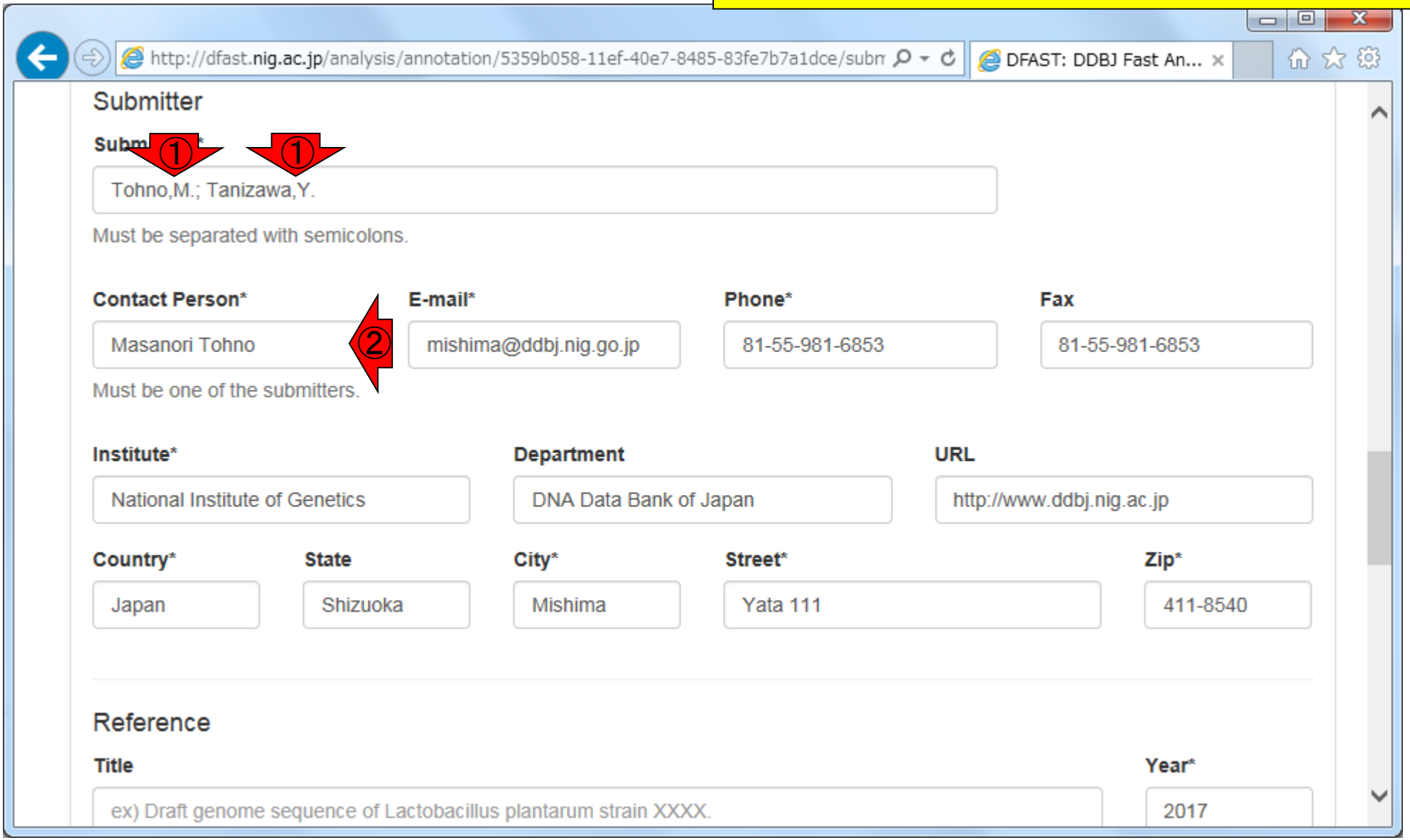
W10-7：修正

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



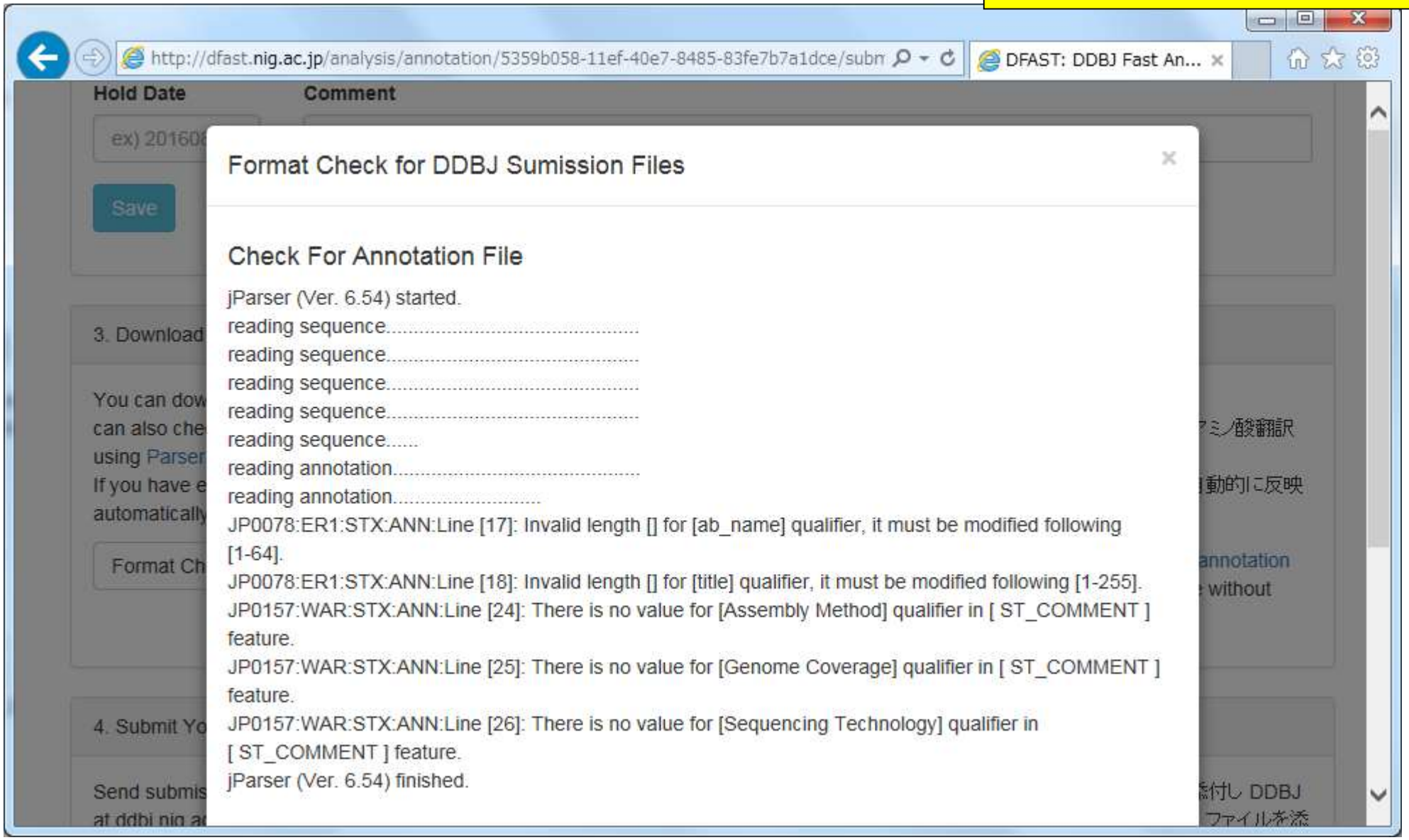
W10-7：修正

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…



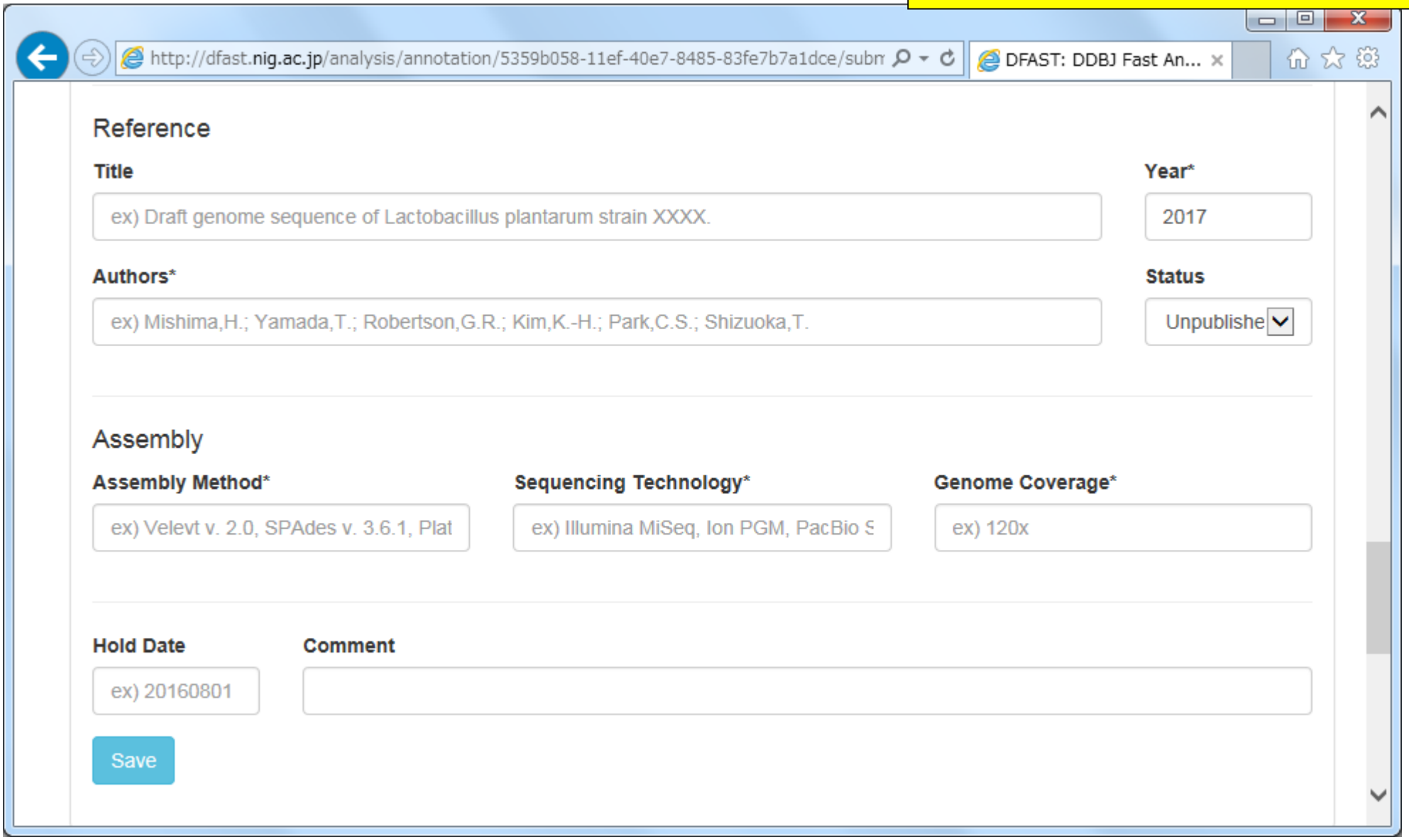
W10-8：Format Check

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



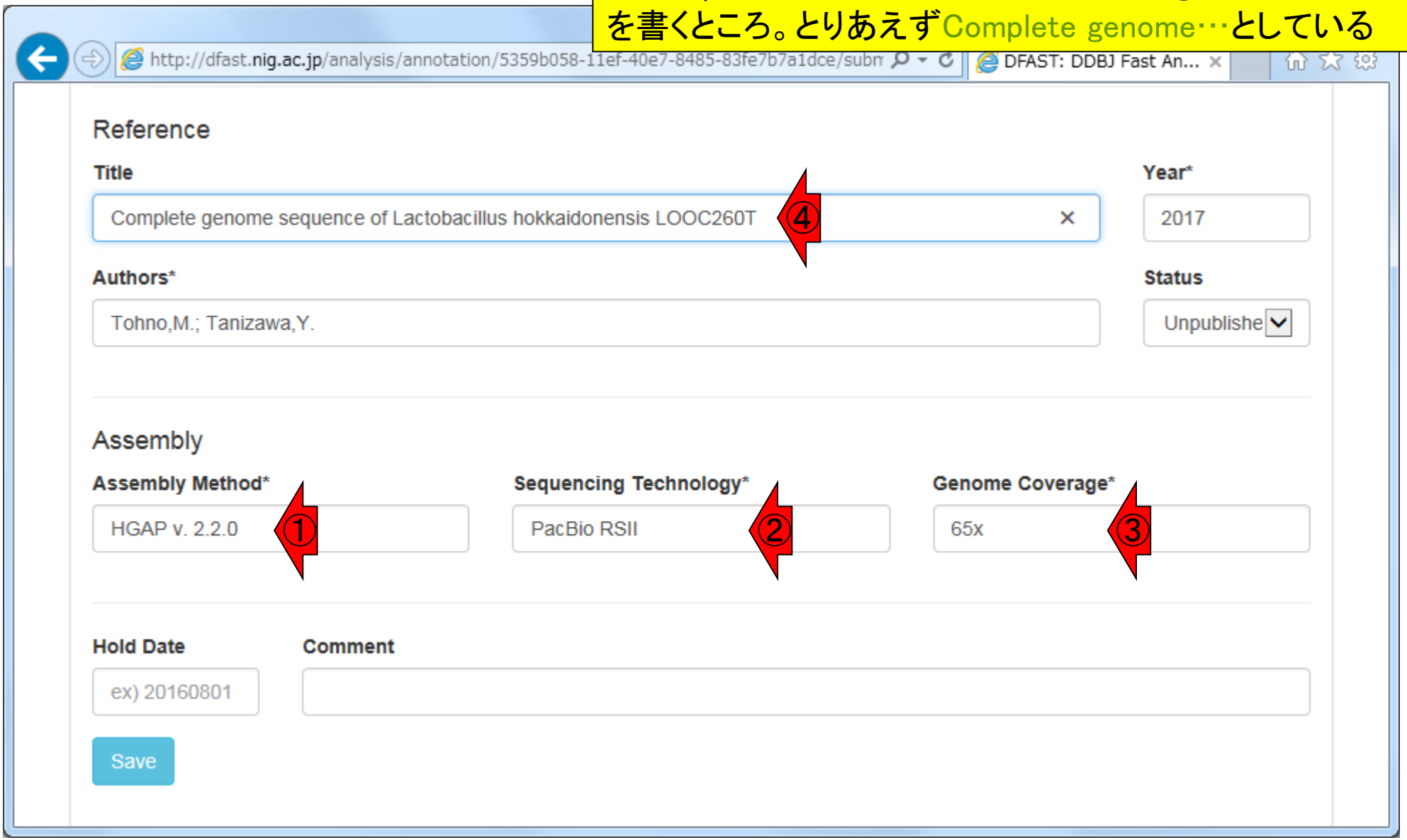
W10-9：修正

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



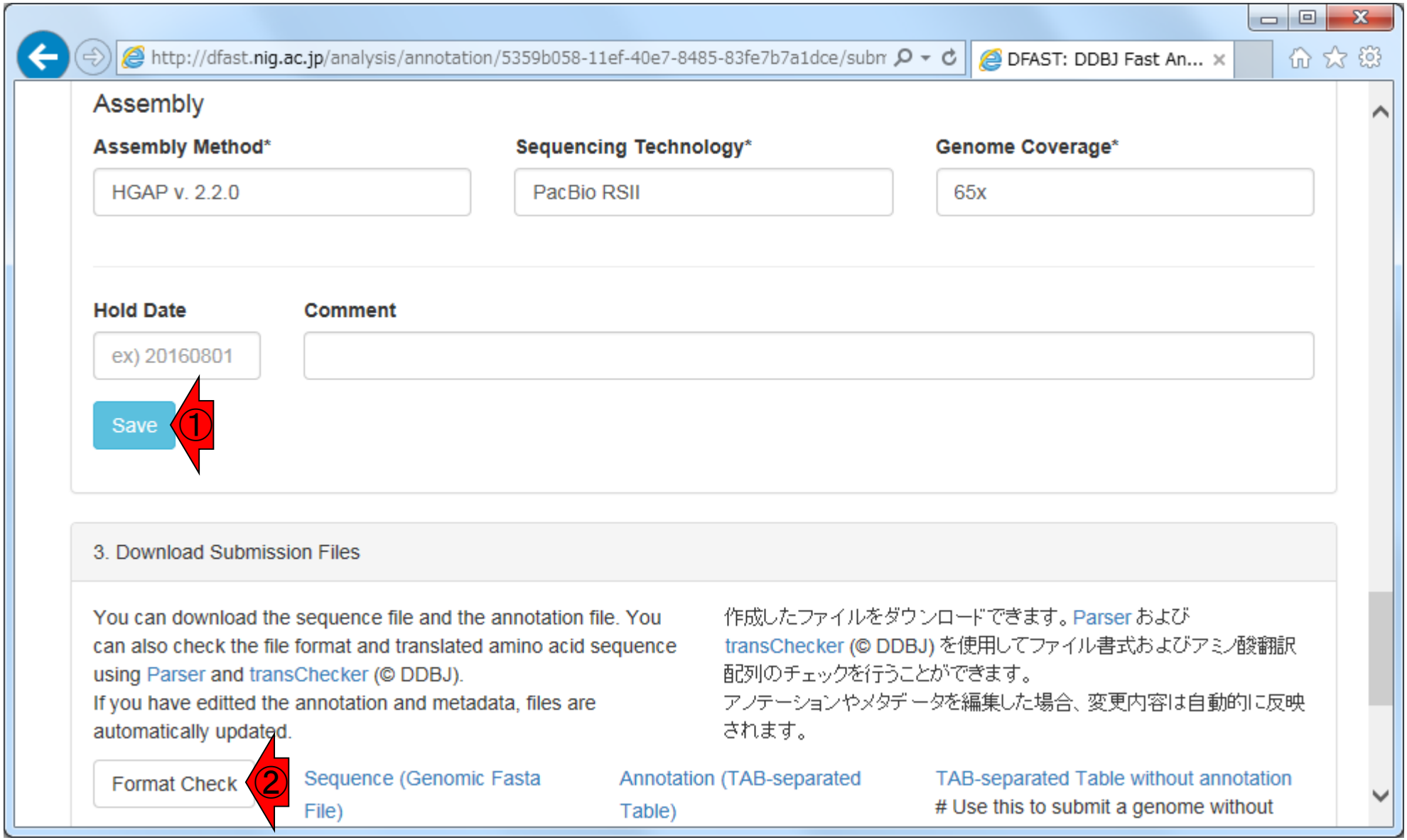
W10-9：修正

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



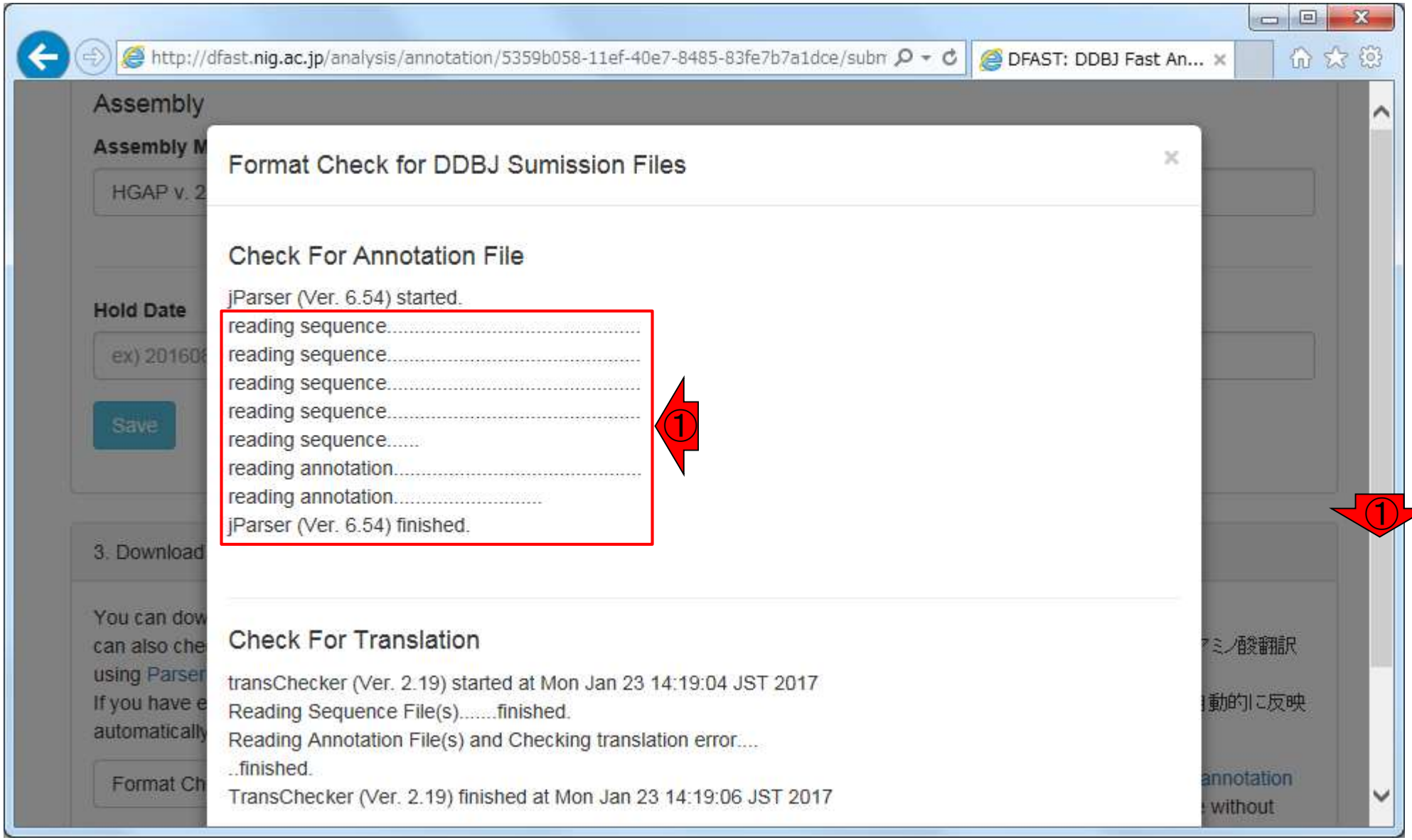
①Saveして②Format Check

W10-10：Format Check



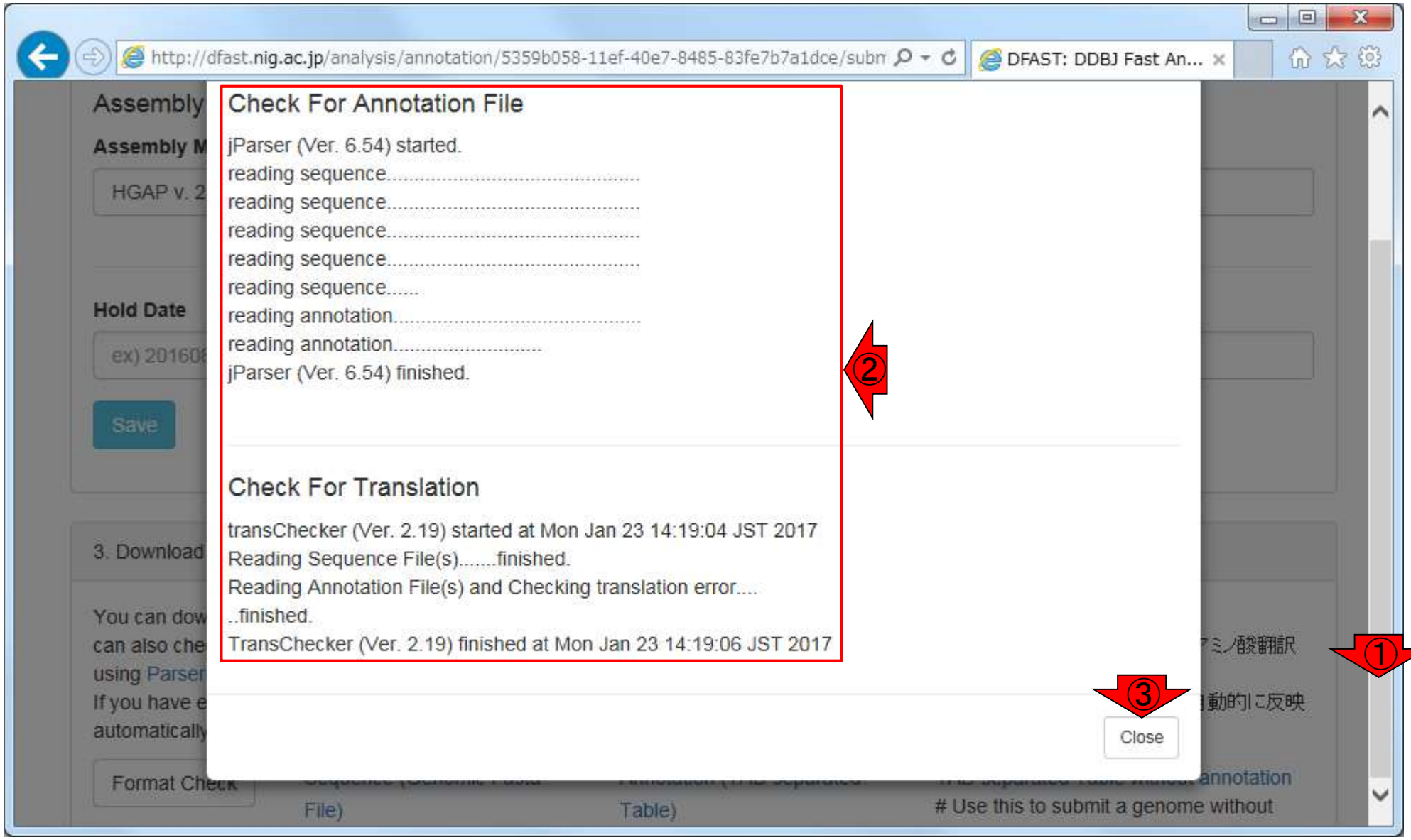
W10-10：Format Check

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



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③Close

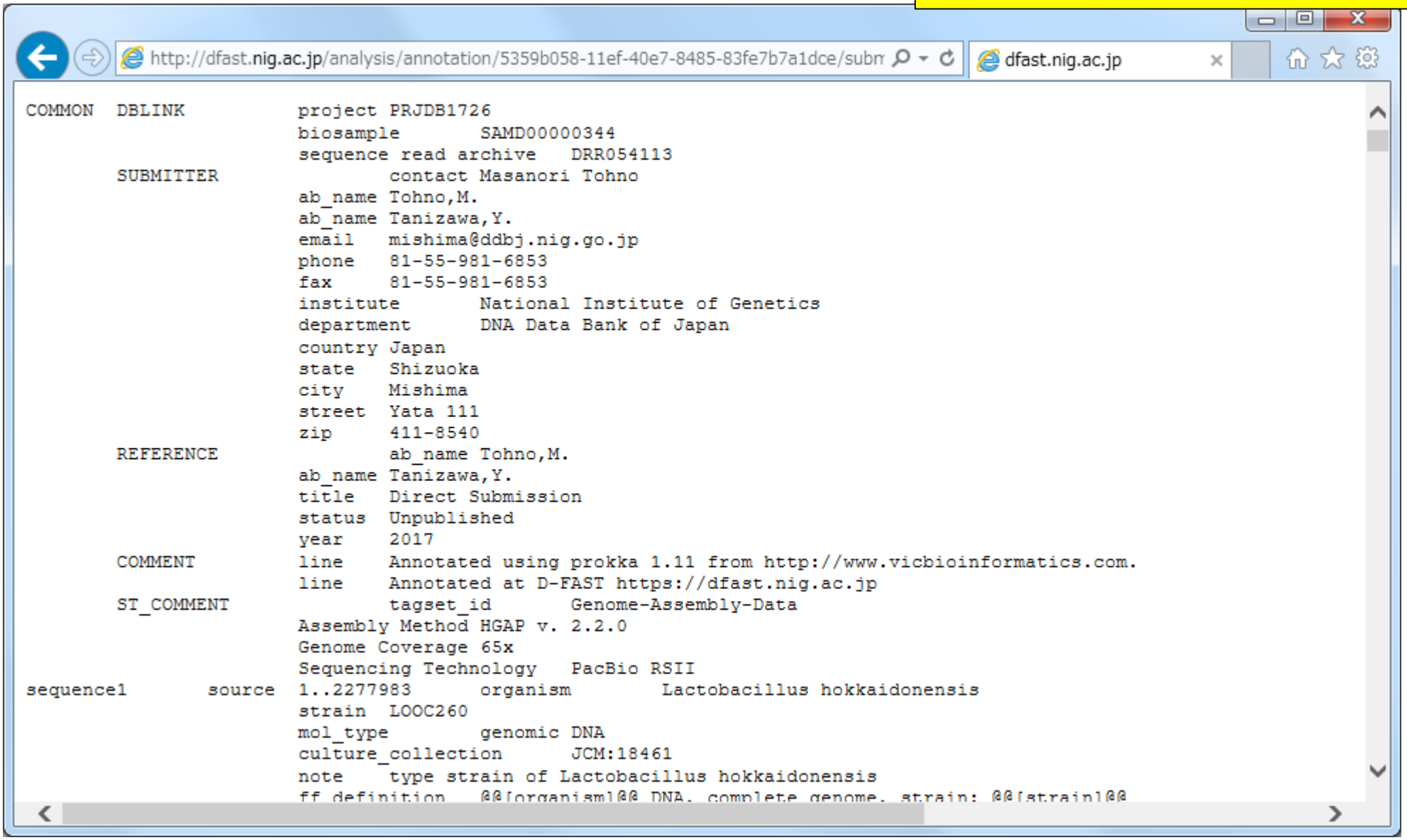


W10-11：とりあえず完成

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

W10-12：Annotation

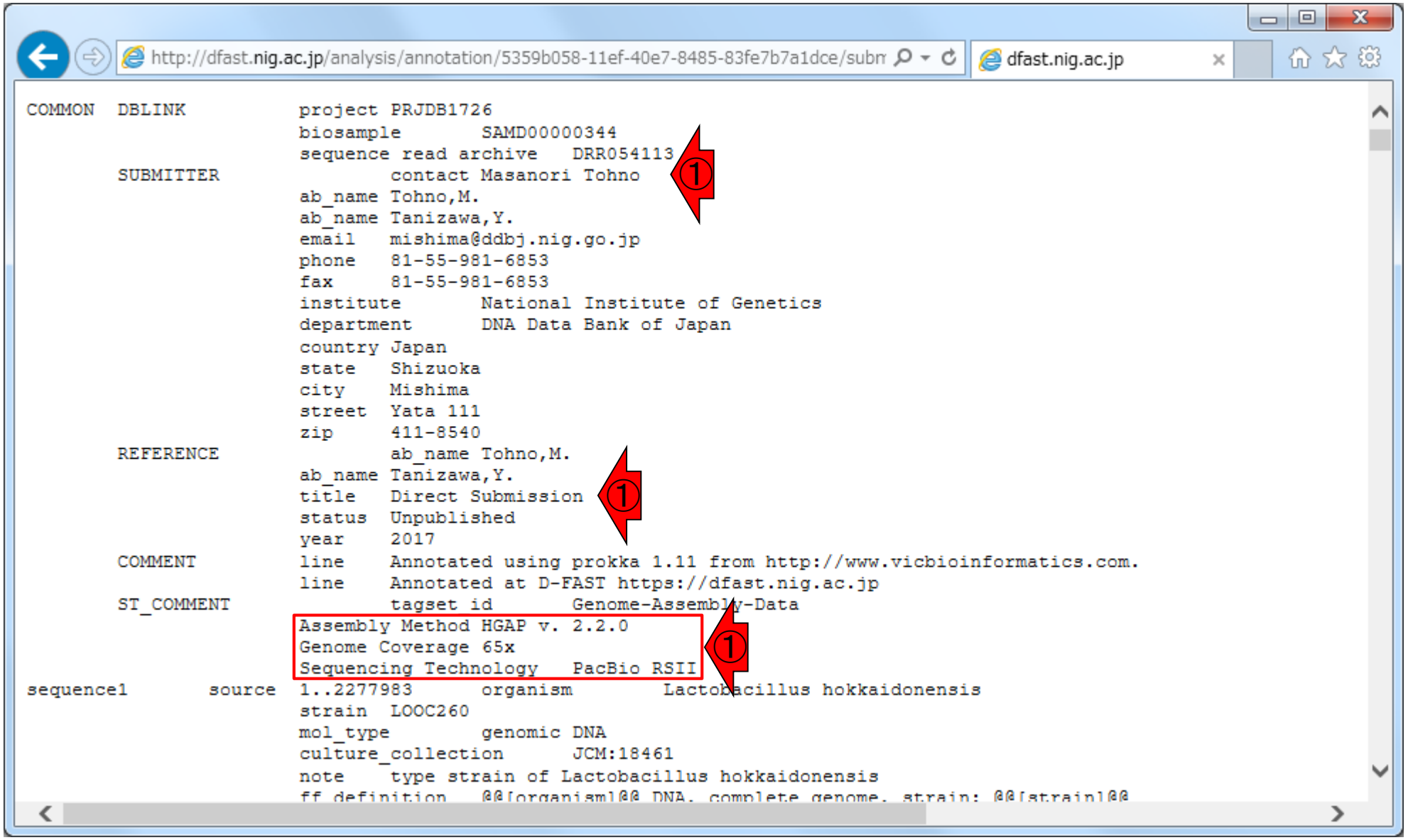
DDBJ形式ファイルが作成できていることがわかります。AP014680の情報とかなり似ていることがわかります



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

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安心



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

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

dfast.nig.ac.jp

```
COMMON   DBLINK          project PRJDB1726
                         biosample       SAMD00000344
                         sequence read archive   DRR054113
         SUBMITTER               contact Masanori Tohno
                         ab_name Tohno,M.
                         ab_name Tanizawa,Y.
                         email   mishima@ddbj.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
sequence1        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]@@
```

①

①DNAPlotterのトップページ

W11-1：DNAPlotter

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

DNAPlotter | Sanger I...

wellcome trust sanger institute

SCIENCE

DNAPlotter ①

Overview

Download

Learn

License

Overview

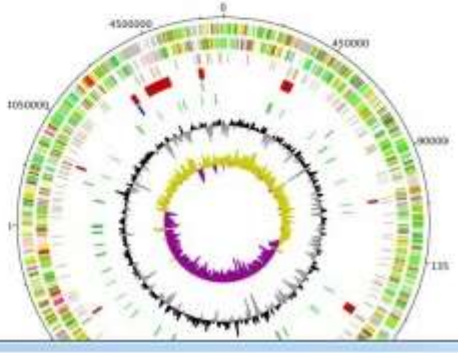
DNAPlotter can be used to generate images of circular and linear DNA maps to display regions and features of interest. The images can be inserted into a document or printed out directly. As this uses Artemis it can read in the common file formats EMBL, GenBank and GFF3.

Download and Installation

Tool type

- Annotation

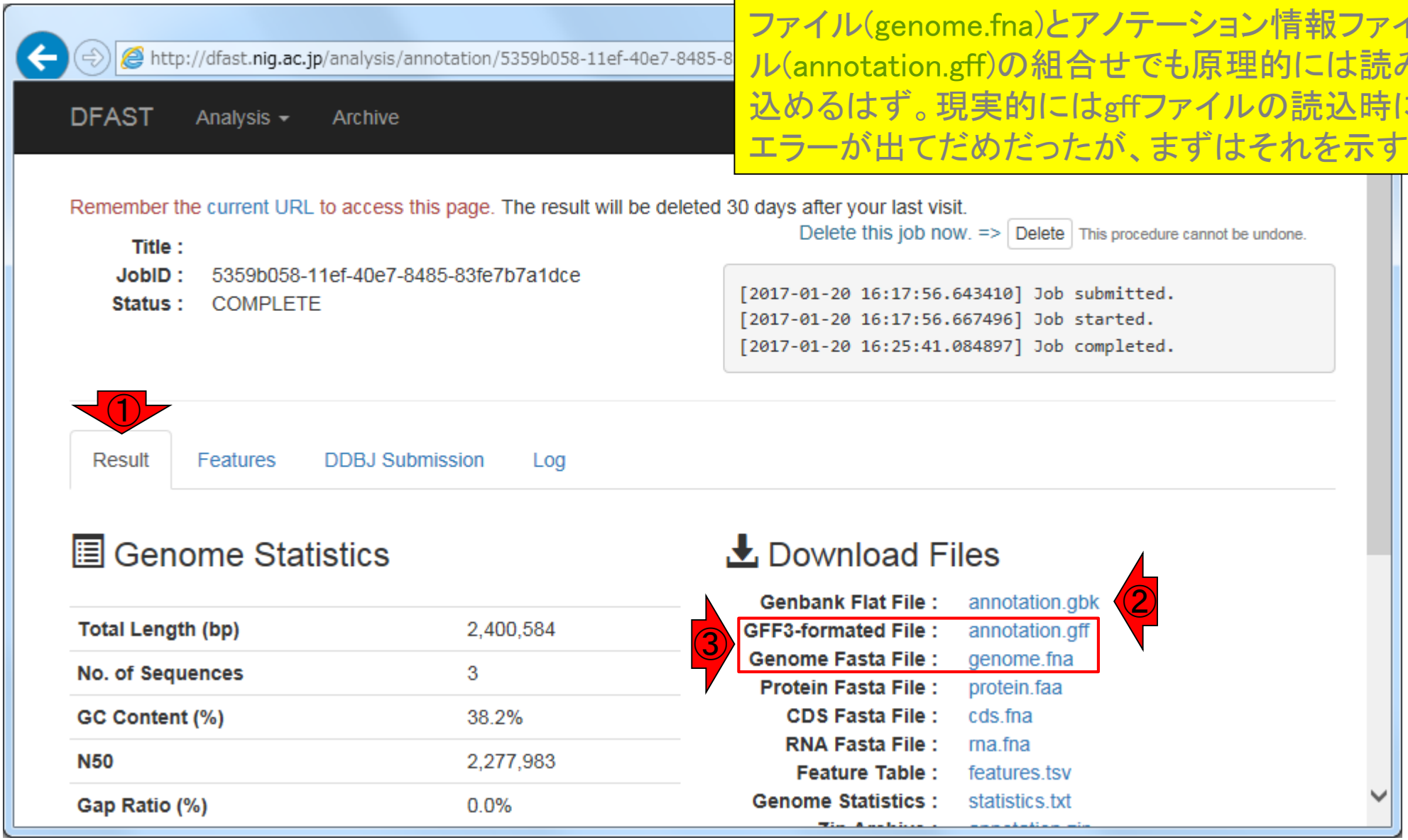
Screenshots



Carver et al., *Bioinformatics*, **25**: 119–120, 2009

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



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

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

DNAPlotter | Sanger I...

wellcome trust sanger institute

①

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.

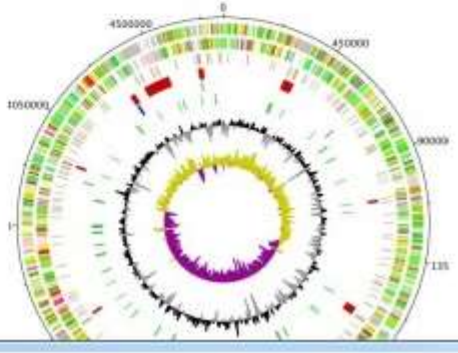
②

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

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

Overview

Download

Learn

License

Contact

Related

Publications

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

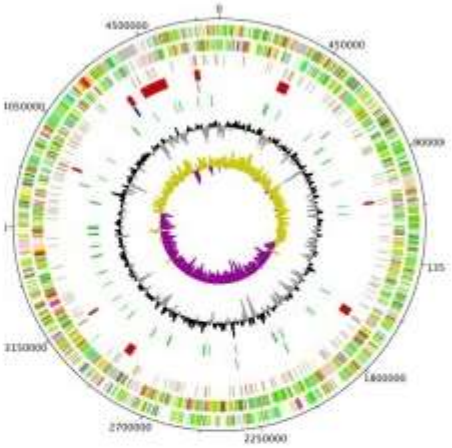
①

②

It can also be launched with Java Web Start:

Launch ③

For further download and installation instructions,



Example showing *S. typhi* genome

Related Tools

- Artemis

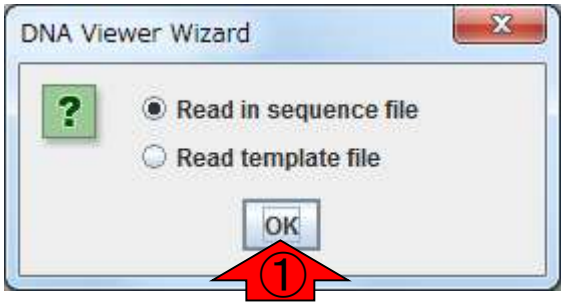
A free genome browser and annotation tool that allows visualisation of sequence features,

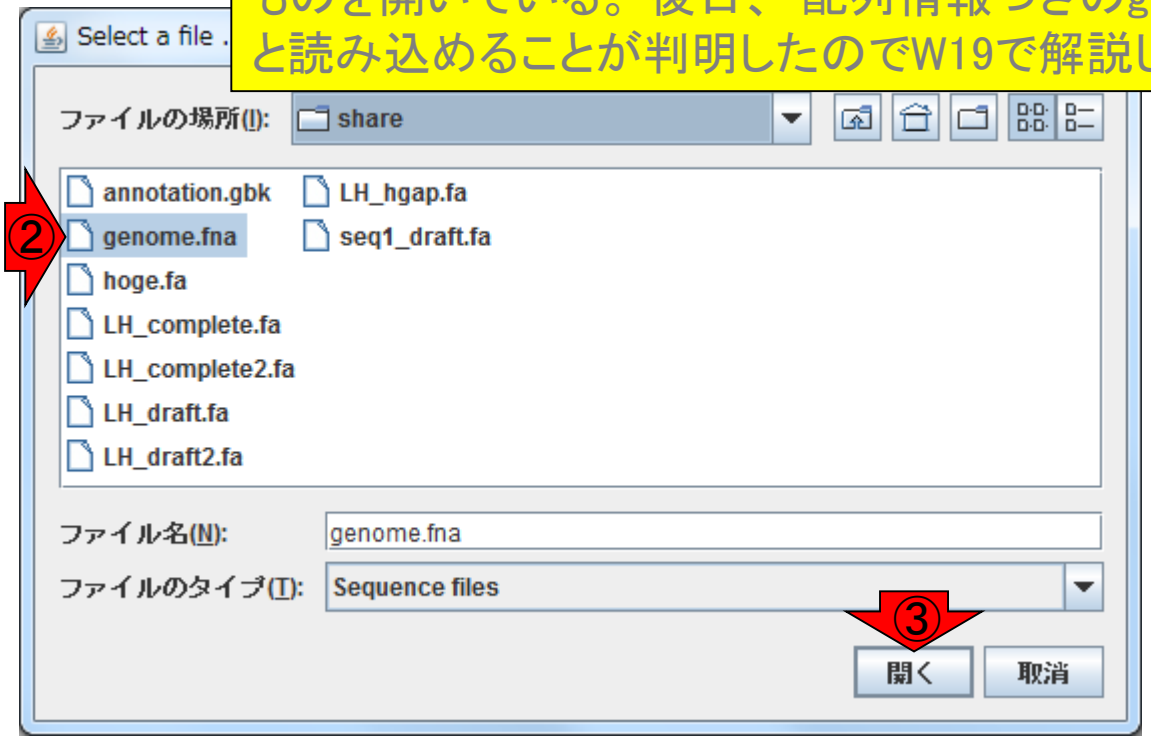
W11-4：ダウンロードと実行

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

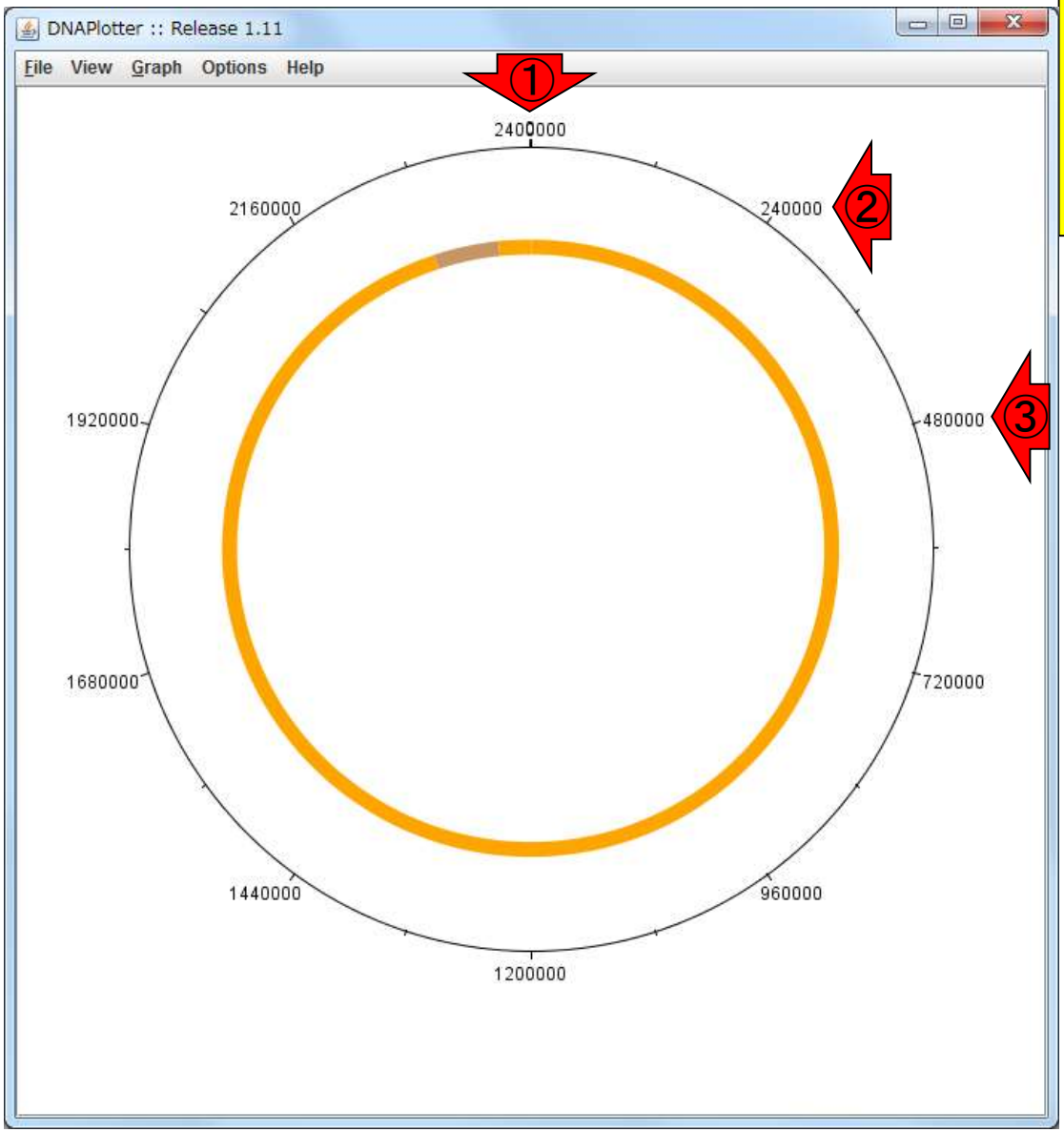
W11-4：実行

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、”配列情報つきの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





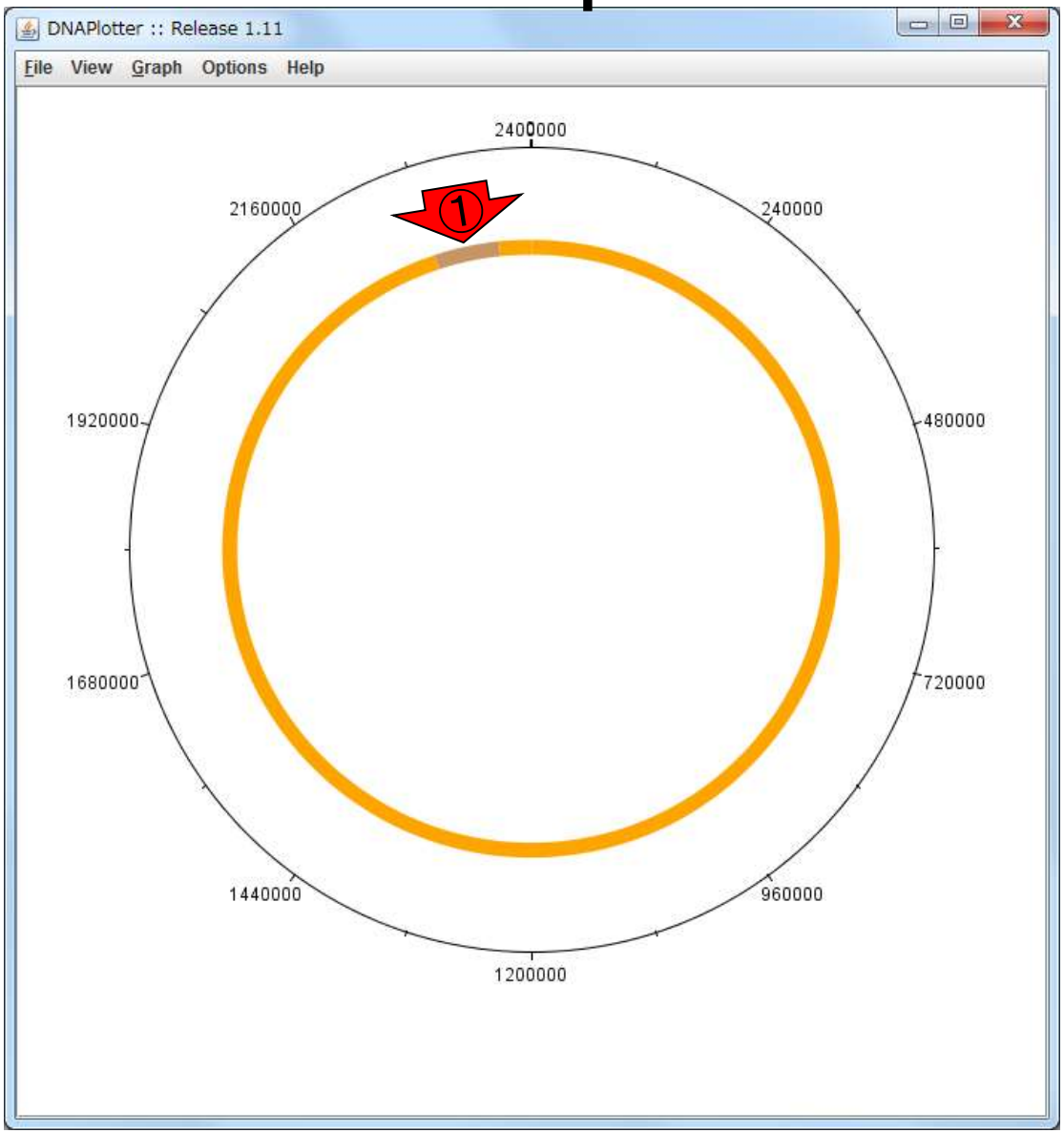
W11-5：ゲノム読込後



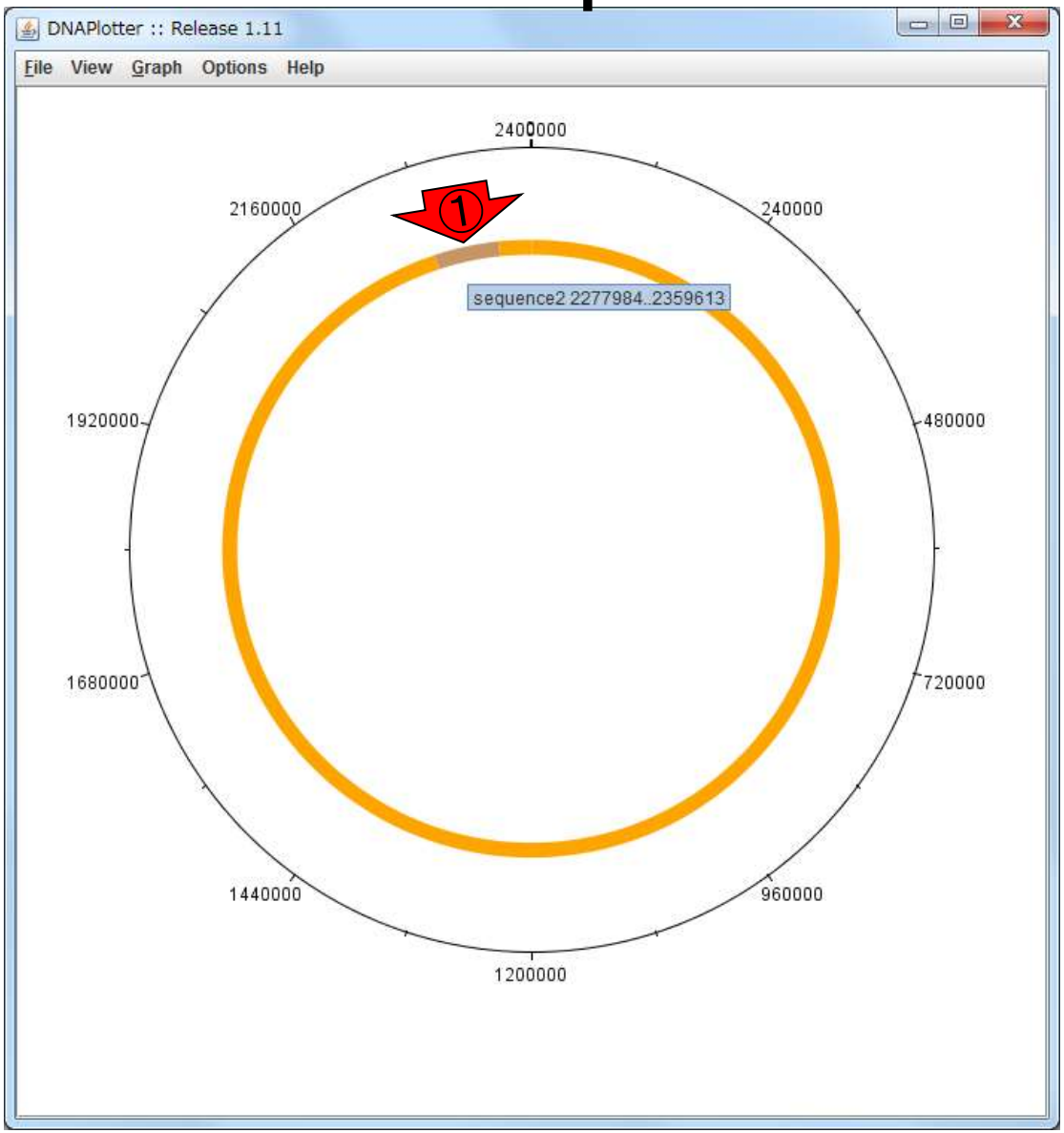
すぐにこんな感じのものが描画されます。このファイルは、計3配列からなり、ゲノムサイズのトータルは2,400,584 bpでした(W7-7)。それゆえ①の場所が2,400,584に近い2400000という値になっているのだろうと判断。10等分した位置を示しているようなので、②が240000、③が480000のような感じになっているのだろう

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…



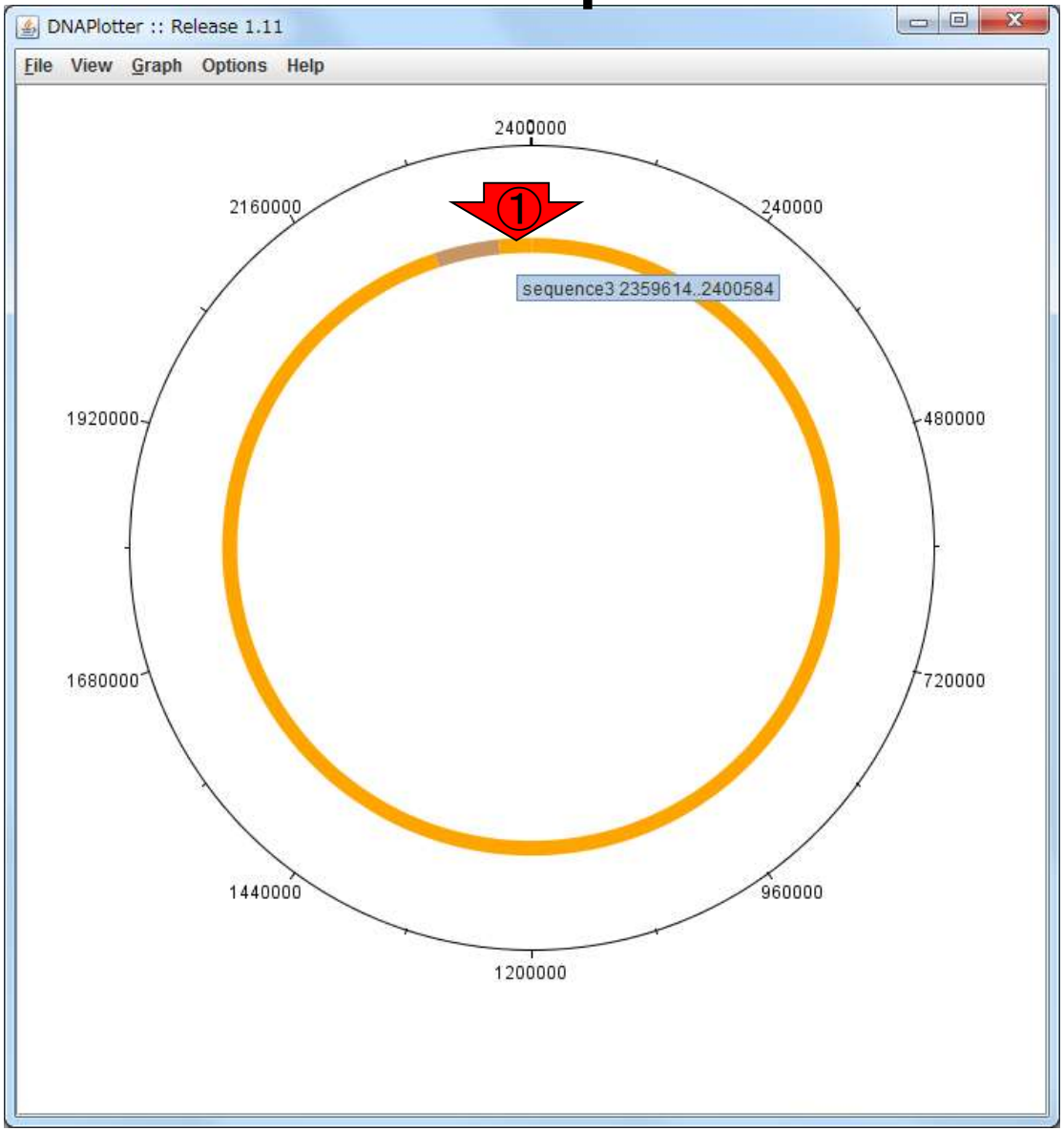
W11-6：sequence2の領域



「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

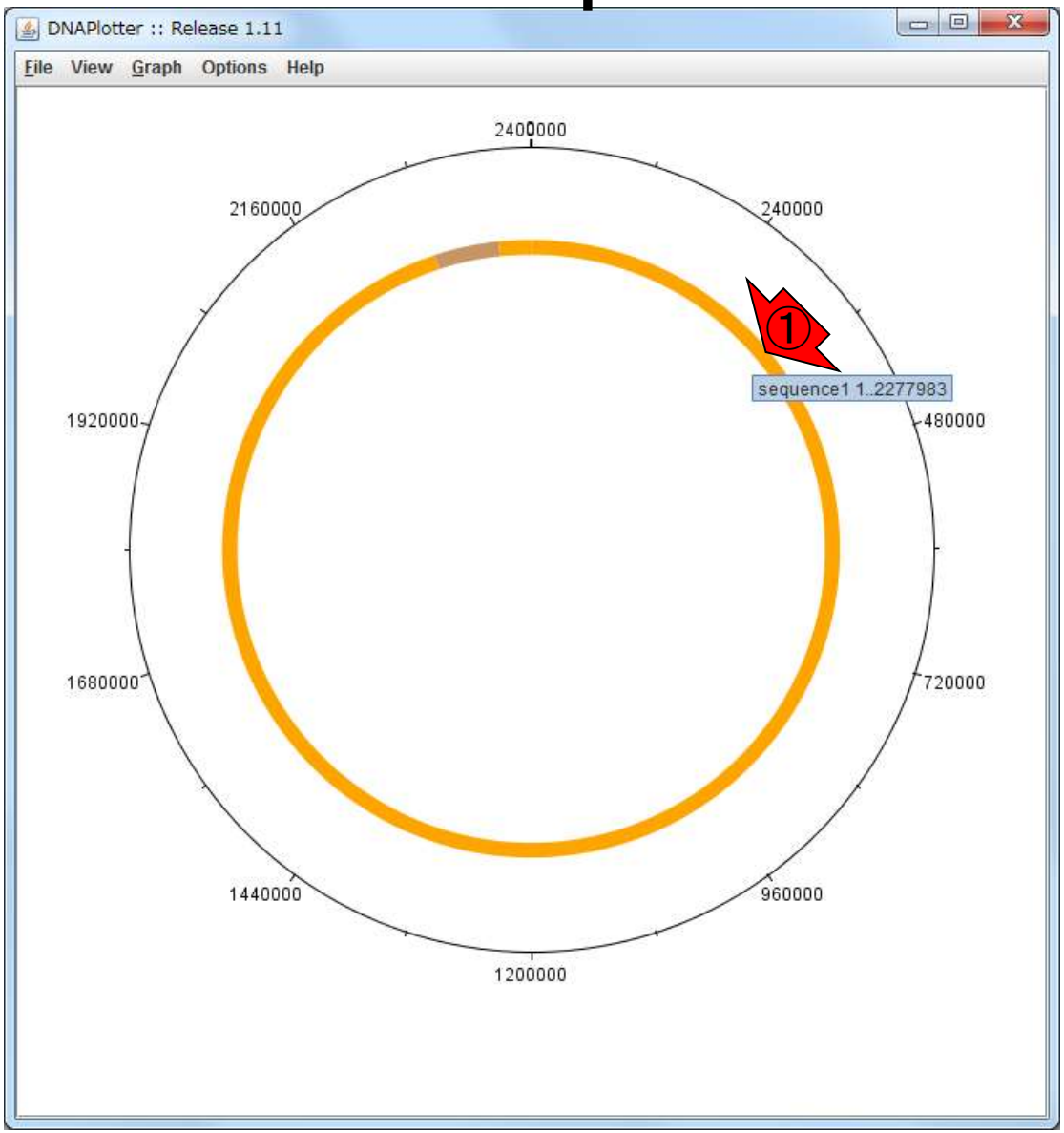
W11-6：sequence3の領域

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



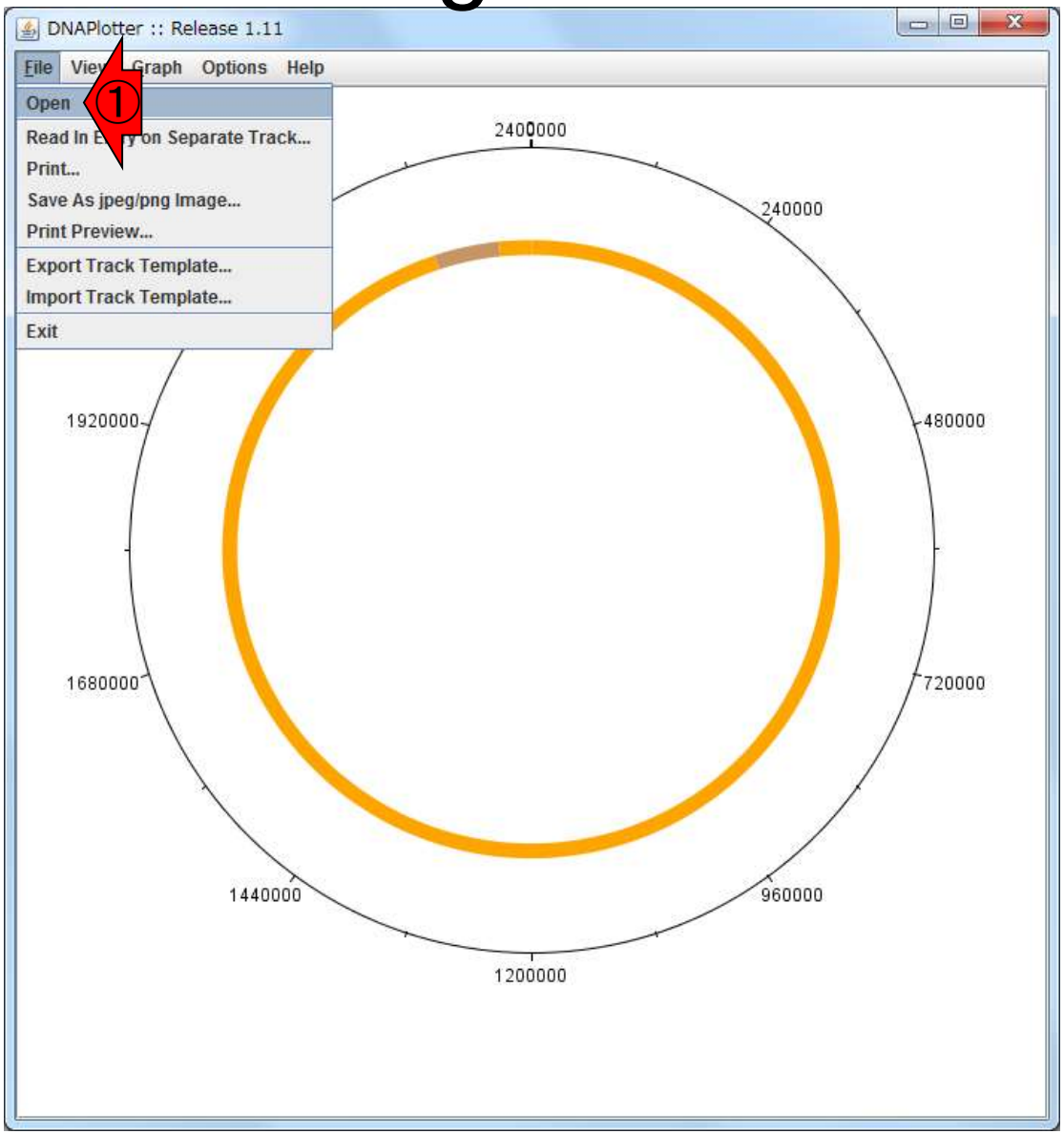
W11-6：sequence1の領域

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



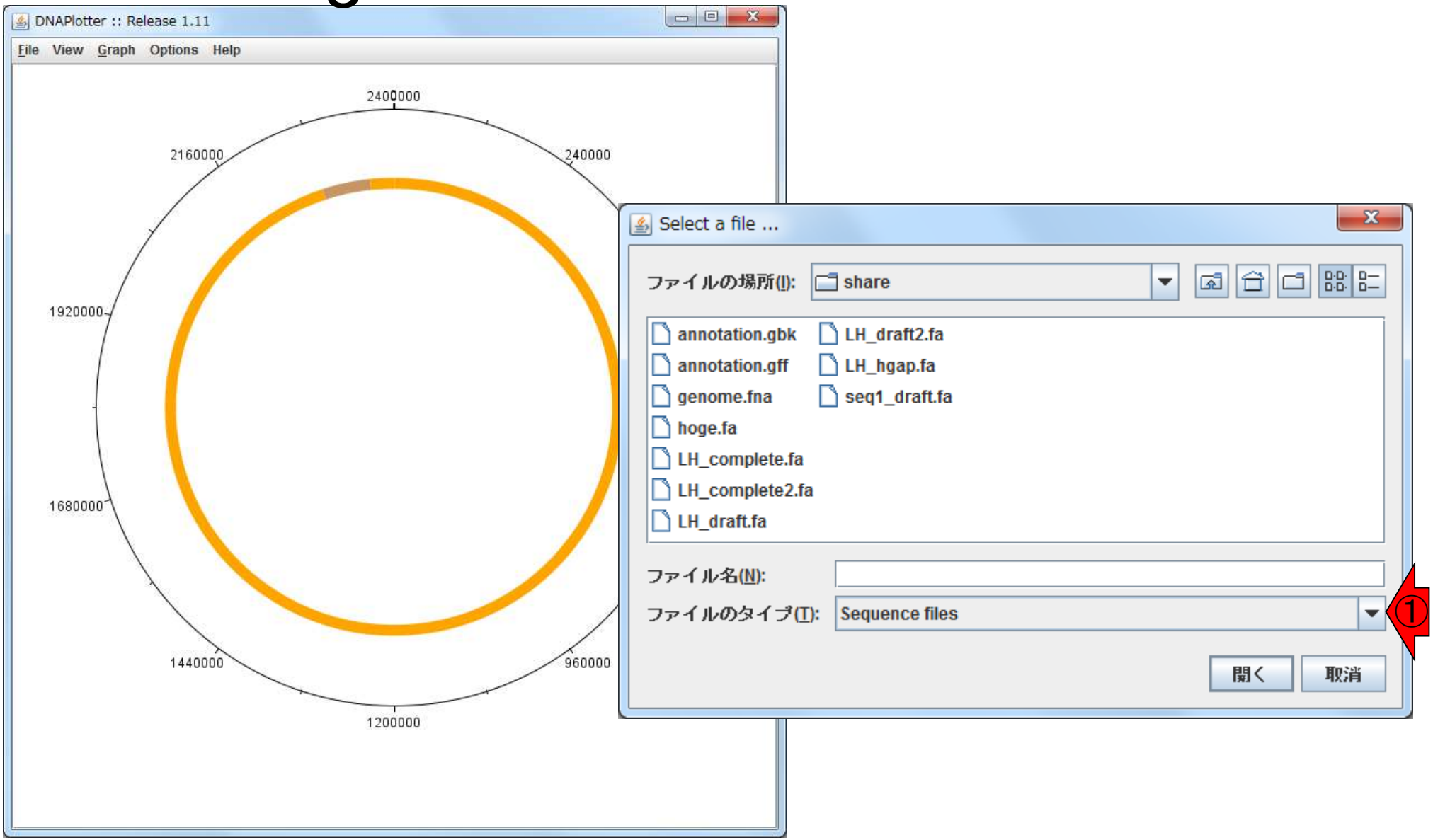
W12-1：gffファイル読込

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



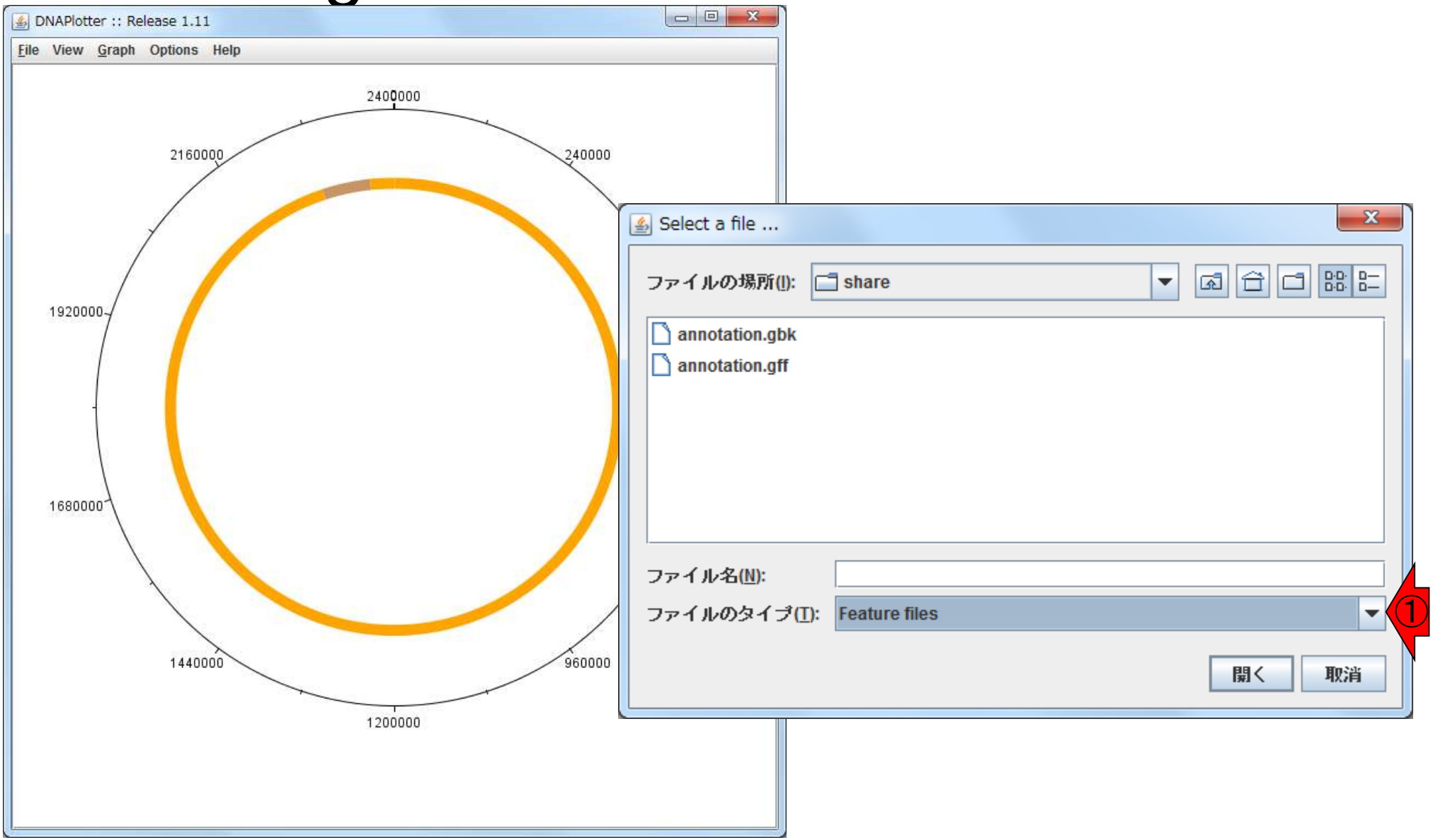
W12-1：gffファイル読込

ファイルのタイプを①
Sequence filesから…



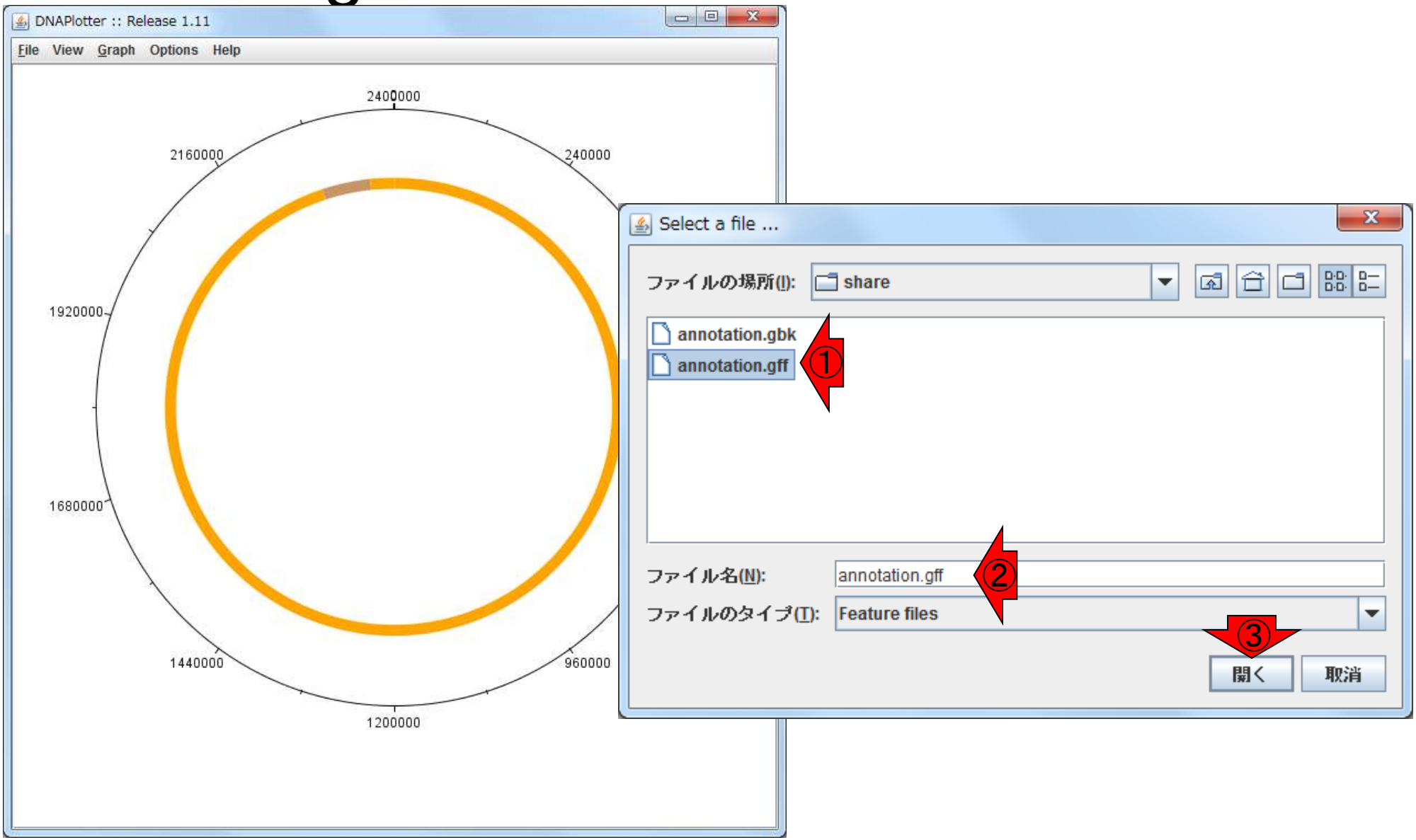
W12-1：gffファイル読込

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



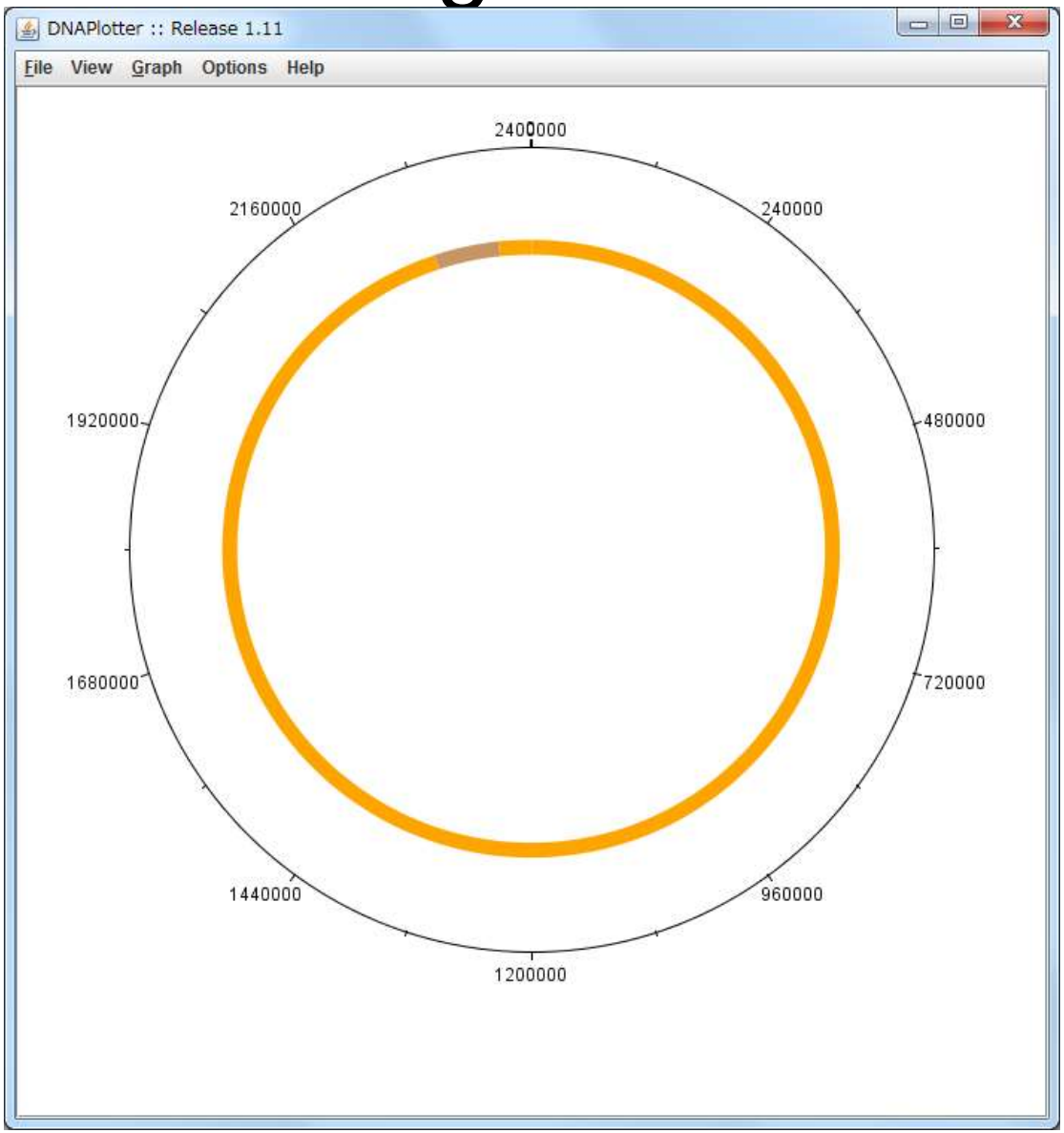
W12-1：gffファイル読込

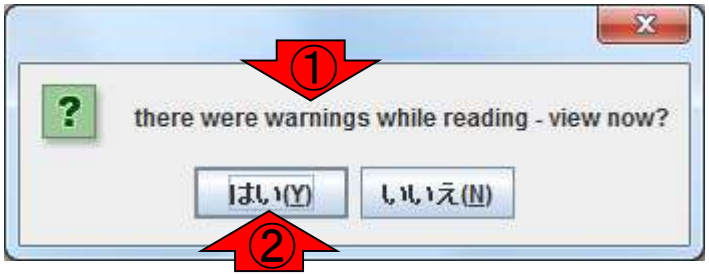
①annotation.gffを選択して、②のところに表示させてから、③開く



①警告がでます…。②はい、を押すと…

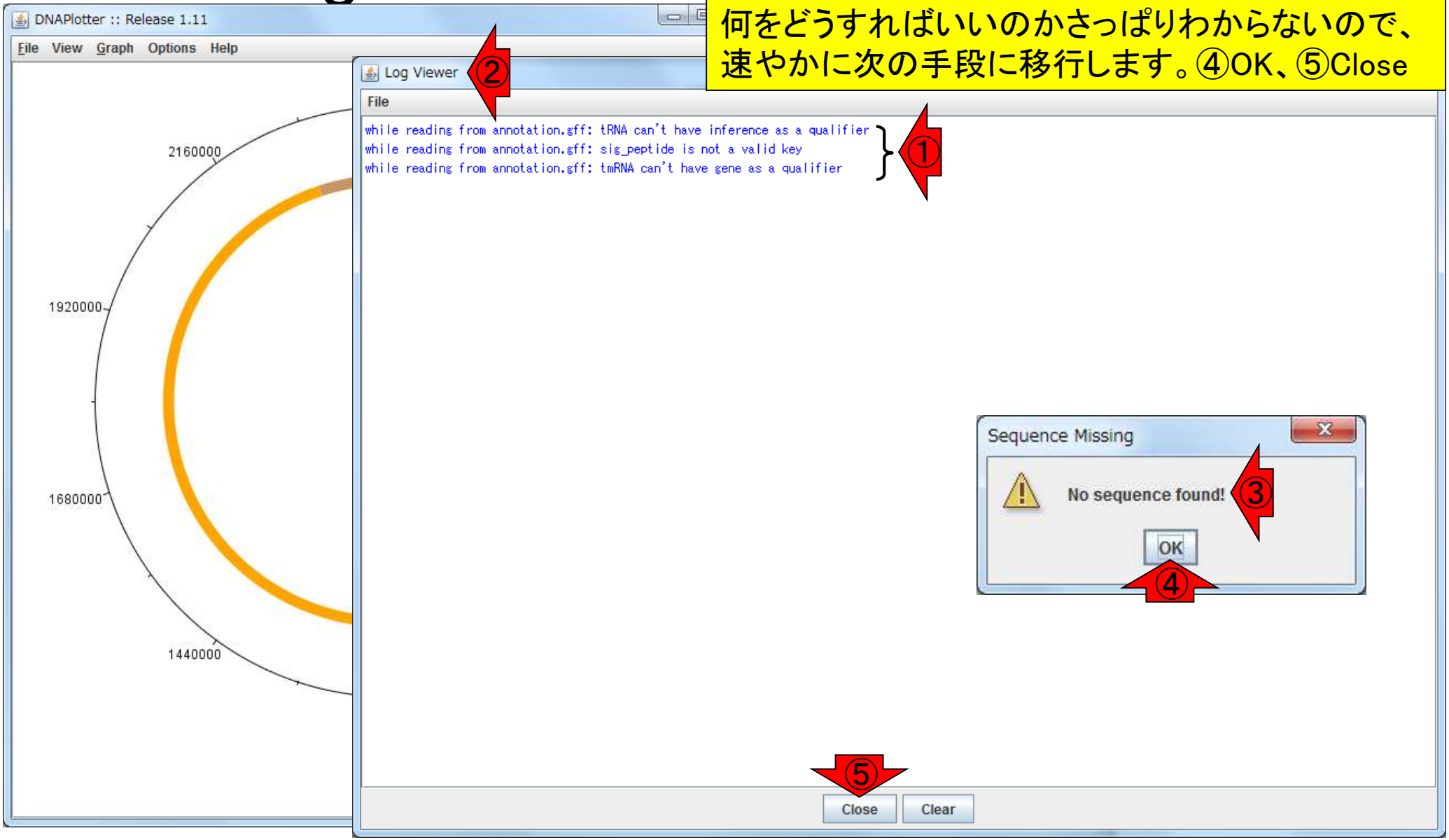
W12-1：gffファイル読込



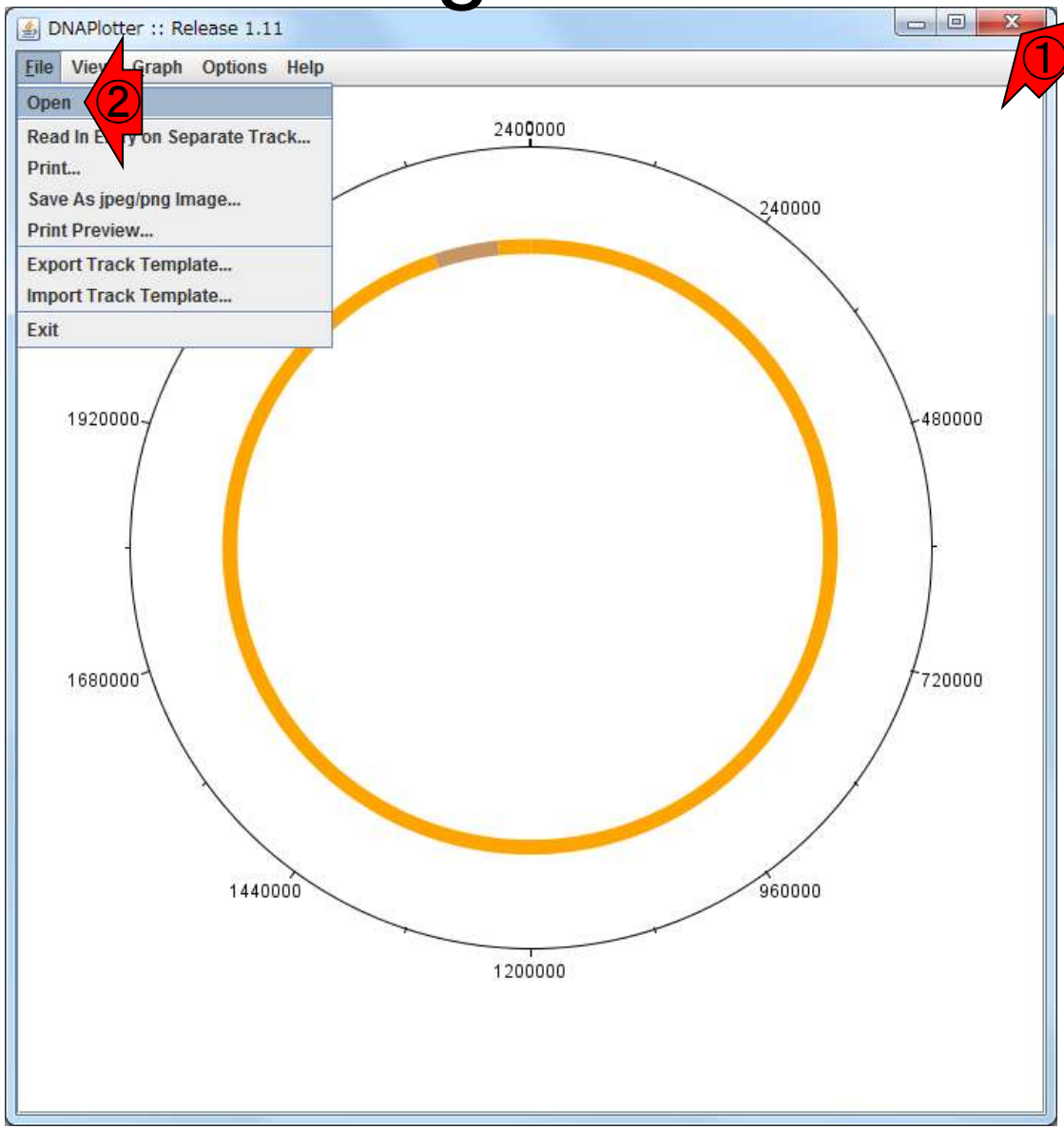


W12-1：gffファイル読

①なんだか、can't などのネガティブなメッセージが②Log Viewerにでて、③No sequence found!となってしまいます。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



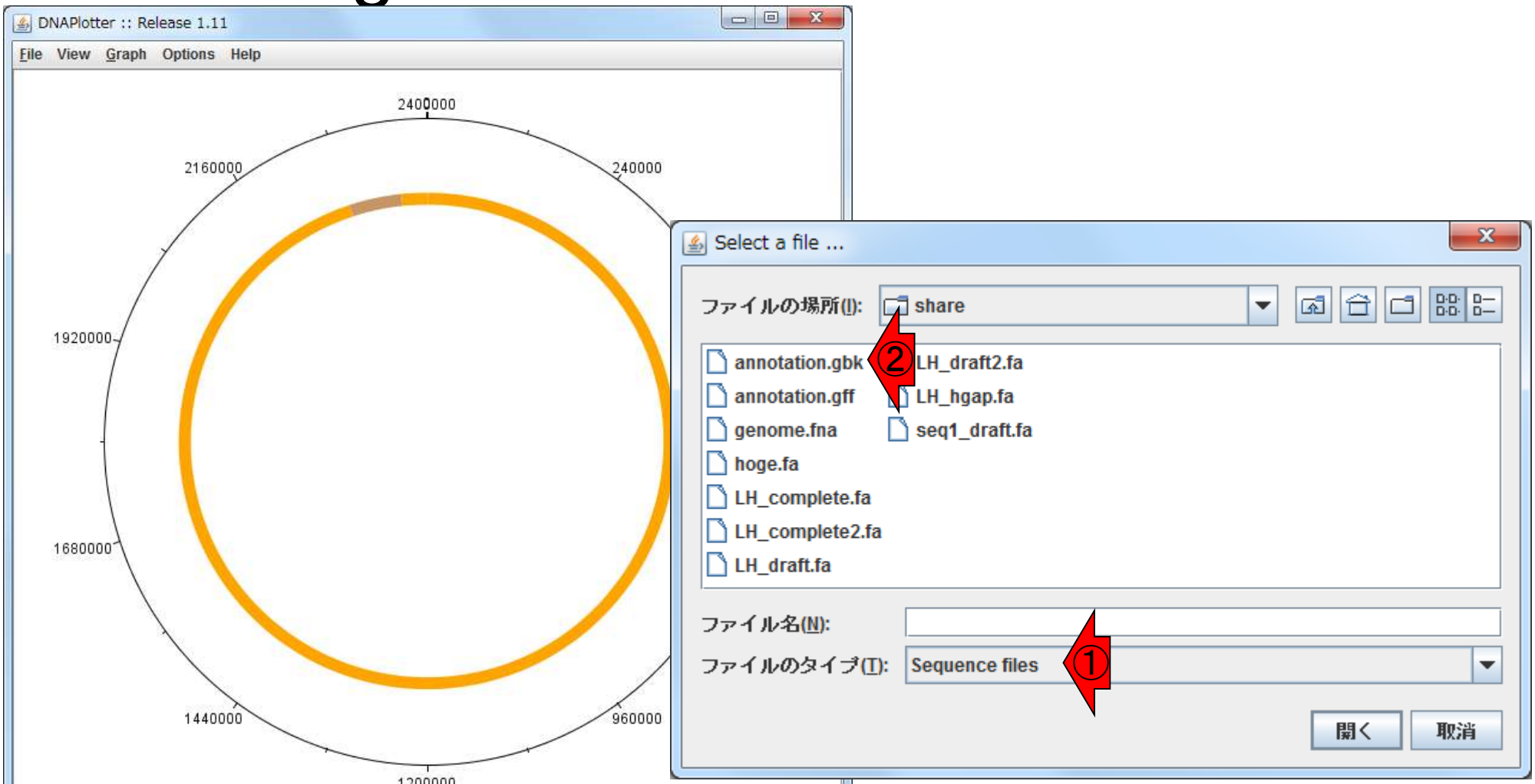
W12-2：gbkファイル読込



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②Openでもよい。ここでは後者で説明

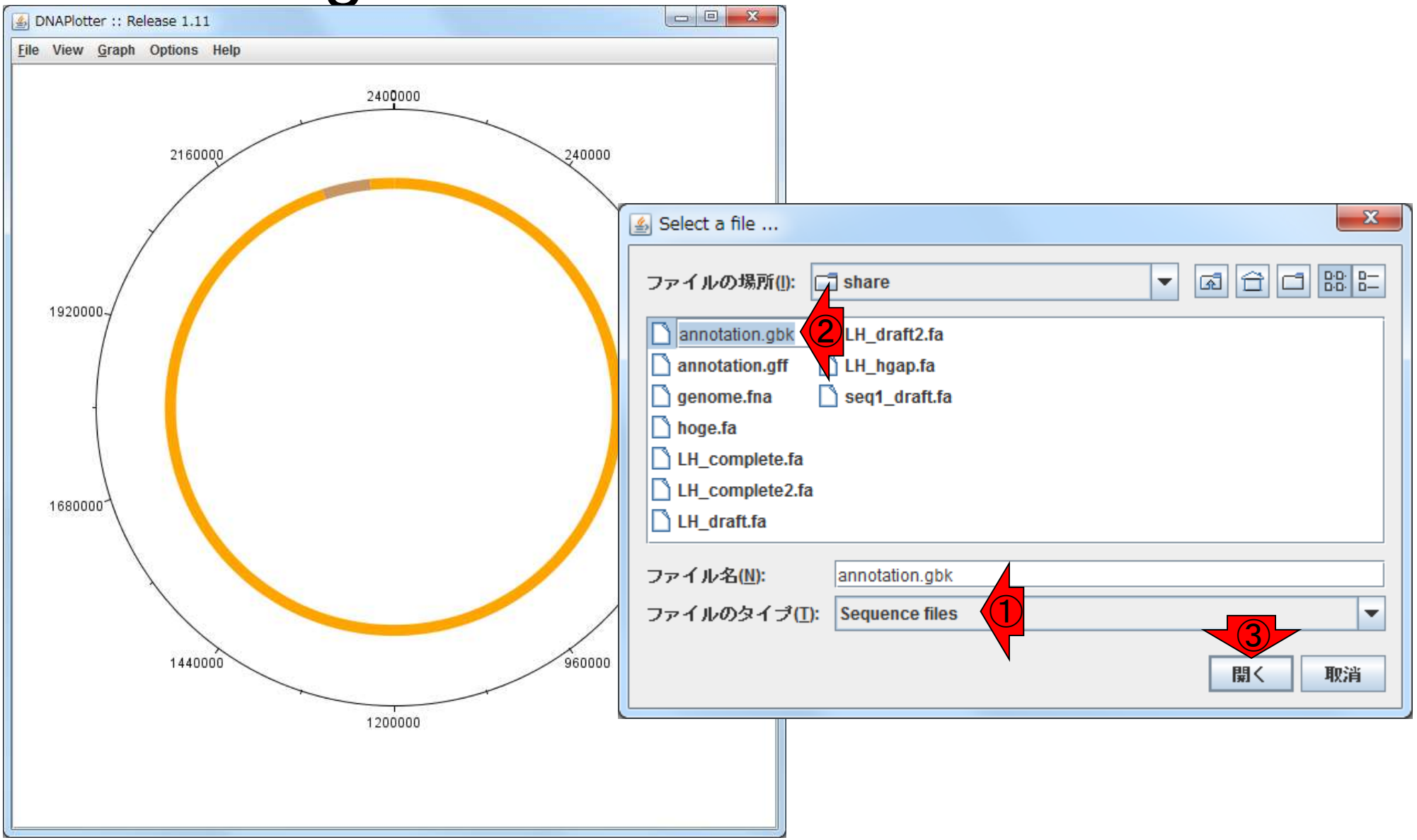
W12-2：gbkファイル読込

①ファイルのタイプがSequence filesのままでもよいので、



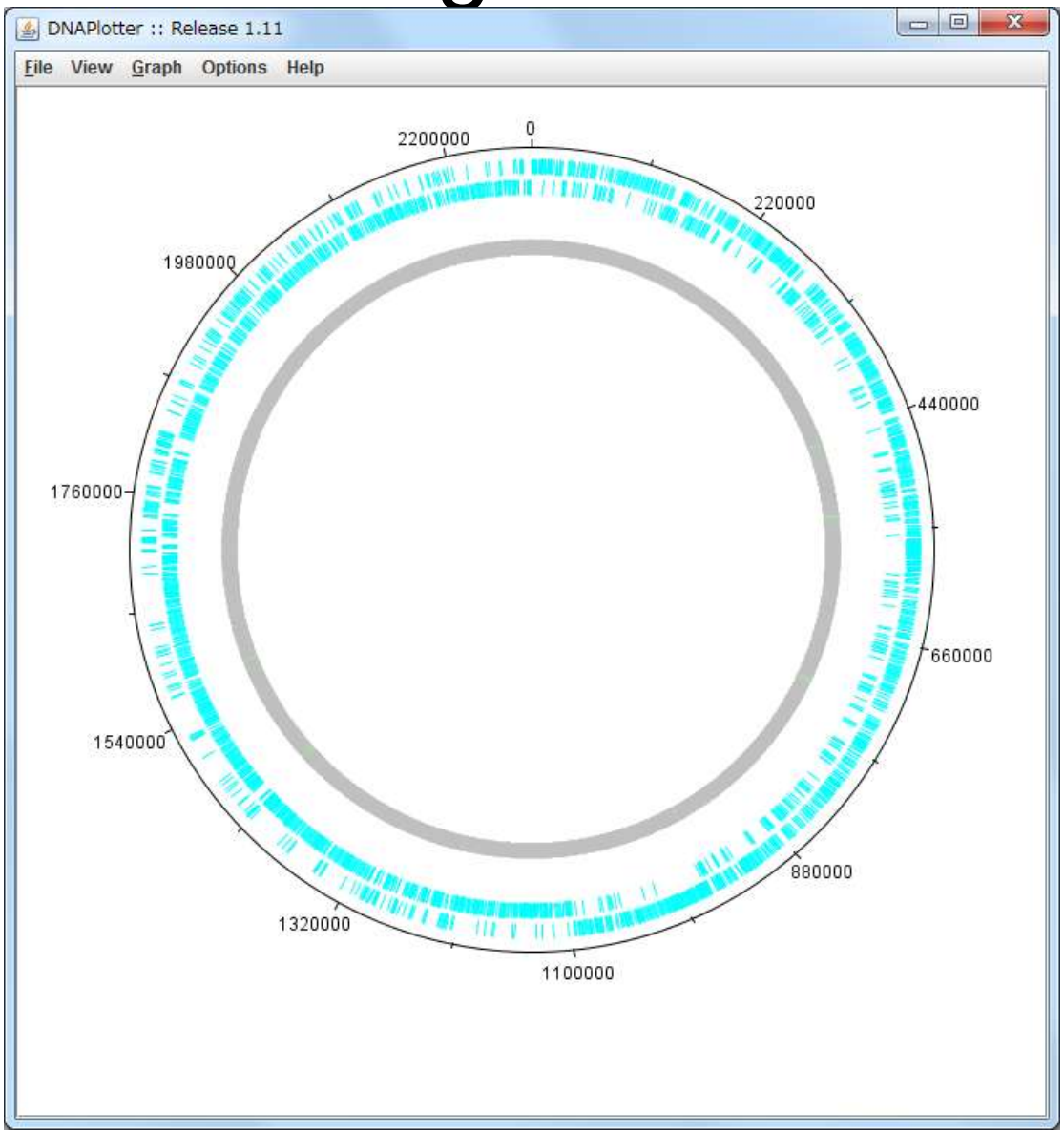
②annotation.gbkを選択して、③開く

W12-2：gbkファイル読込

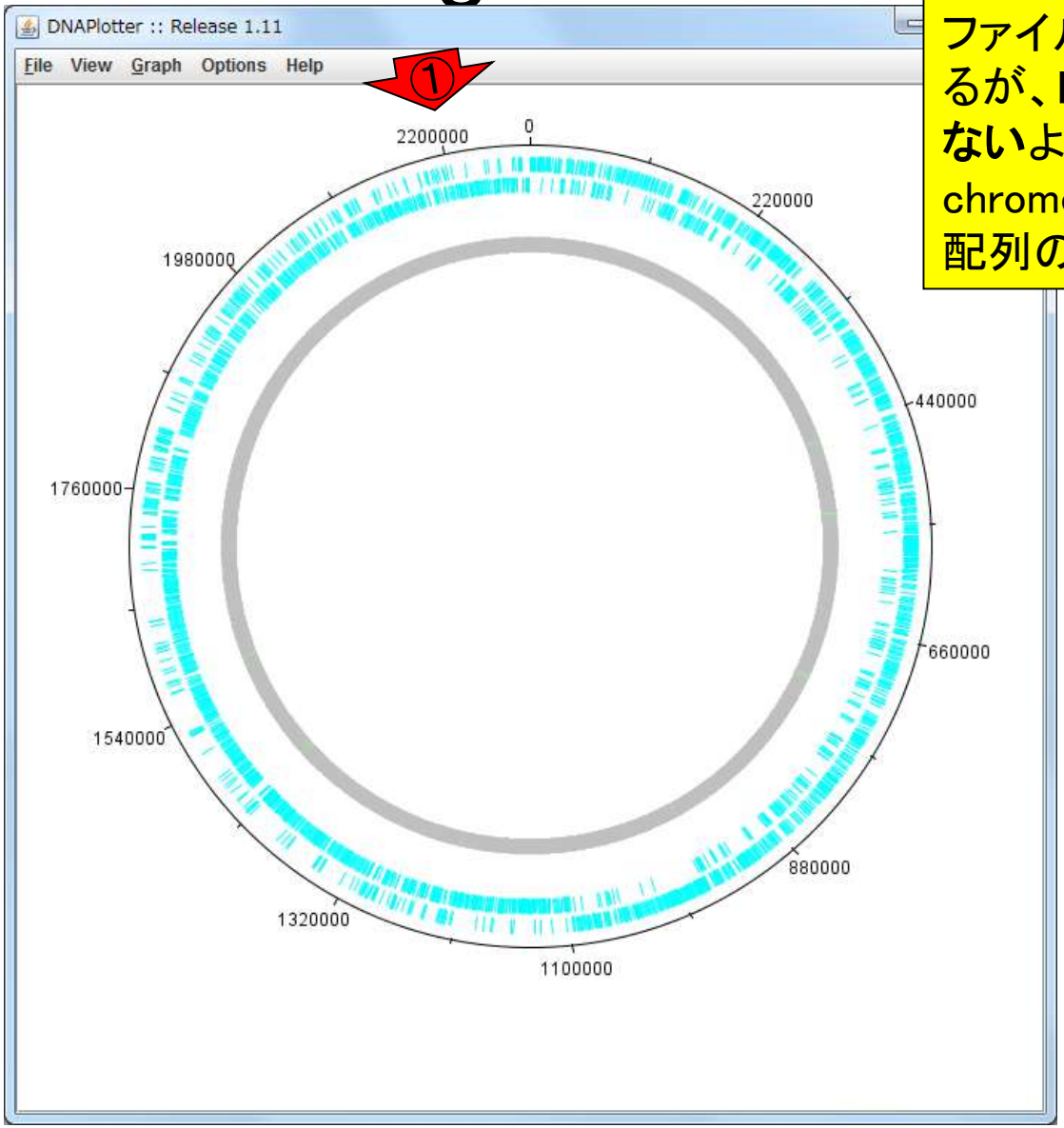


こんな感じになります

W12-3：gbk読込後



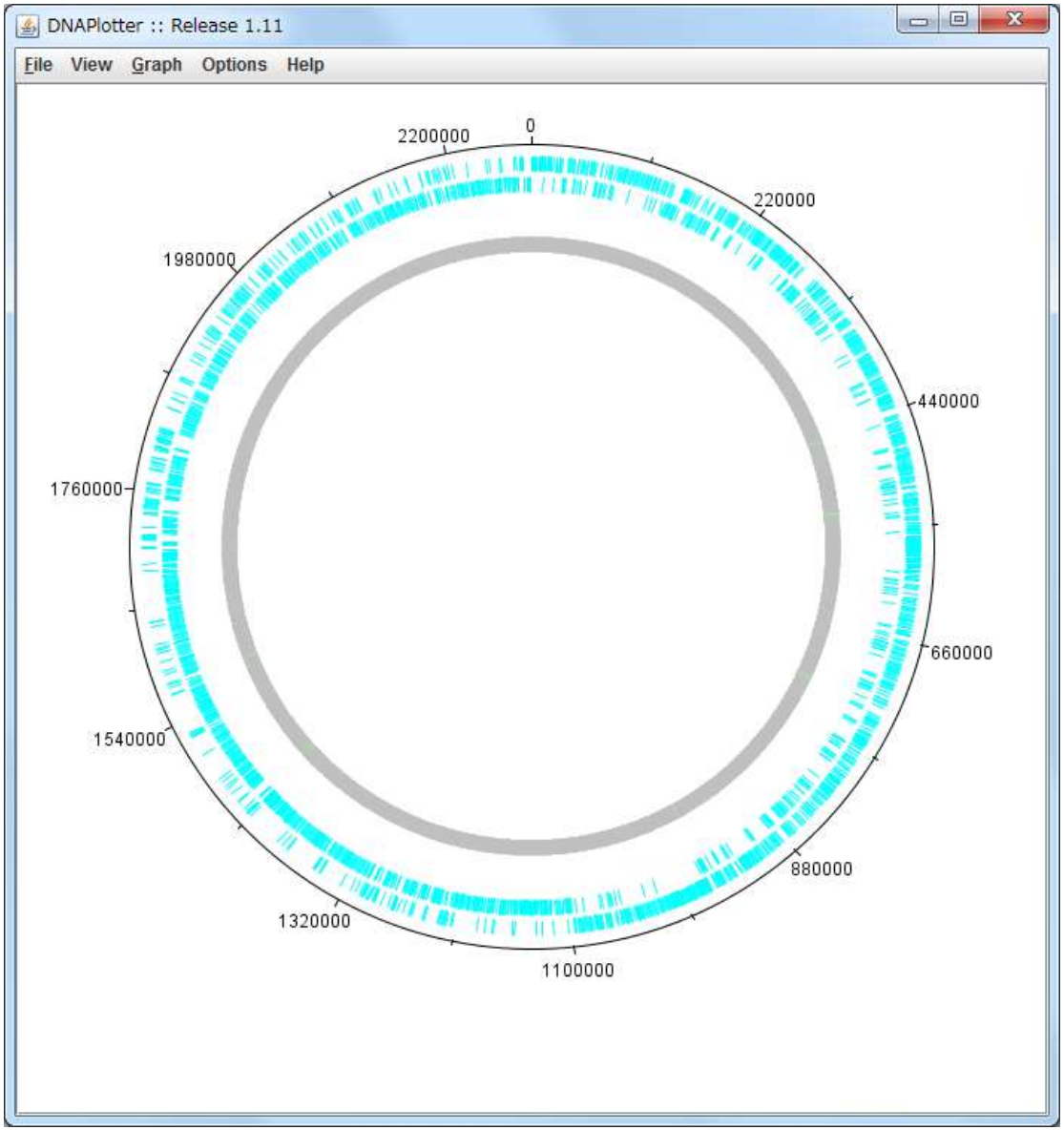
W12-3：gbk読込後



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ね。**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

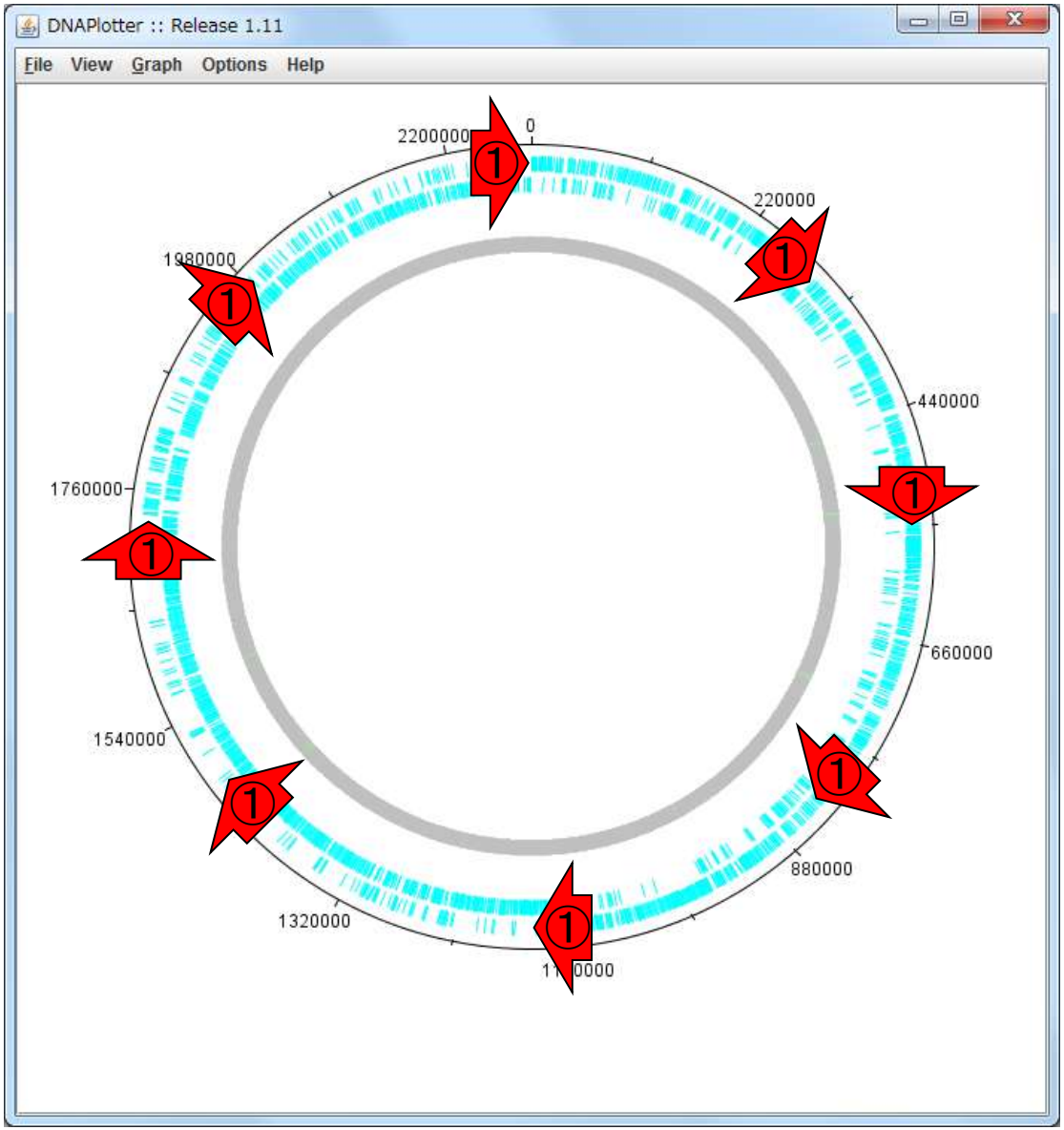
W12-4：解説

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



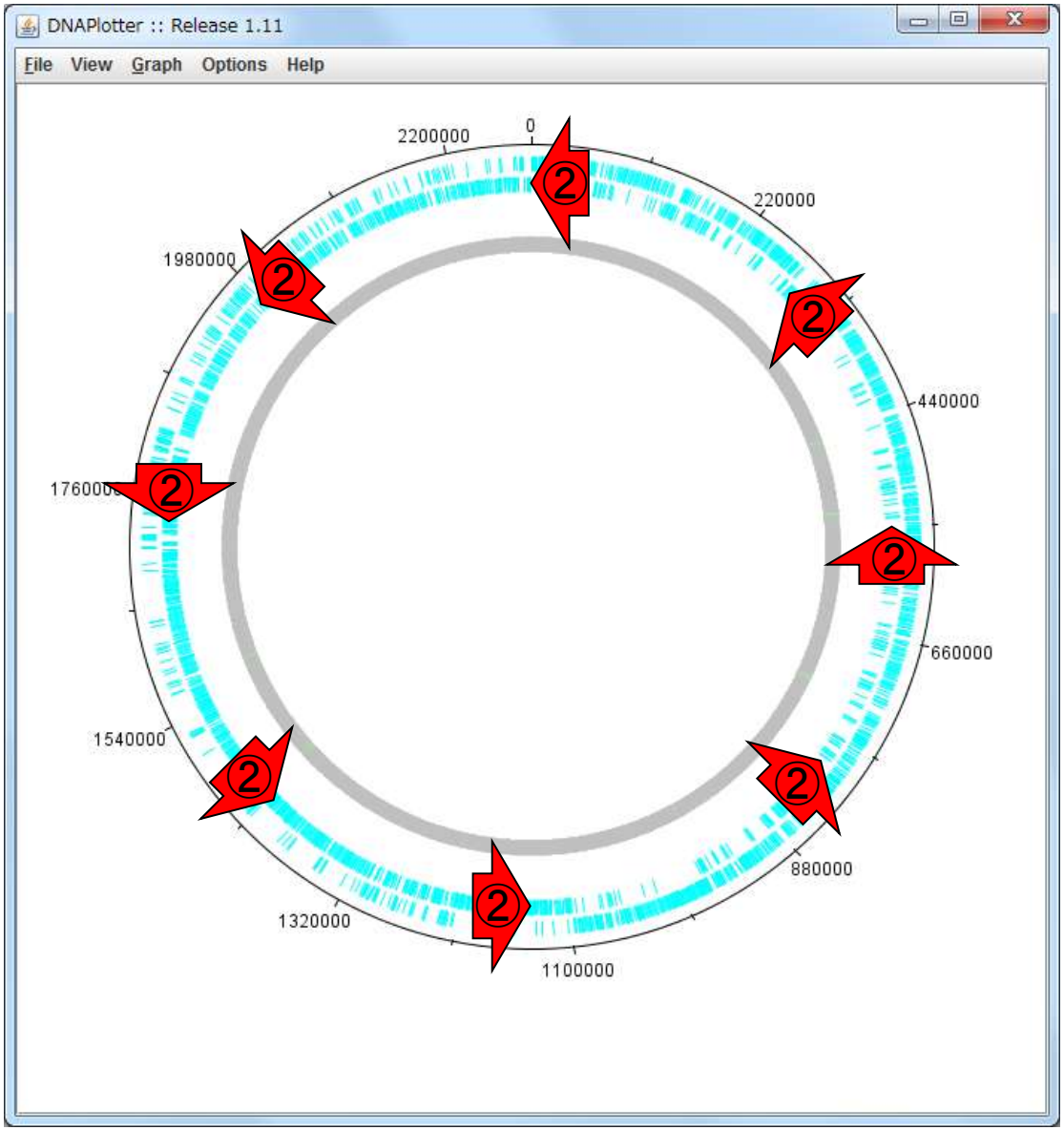
①外側の輪っかは、順鎖上のCDSを表します

W12-4：解説

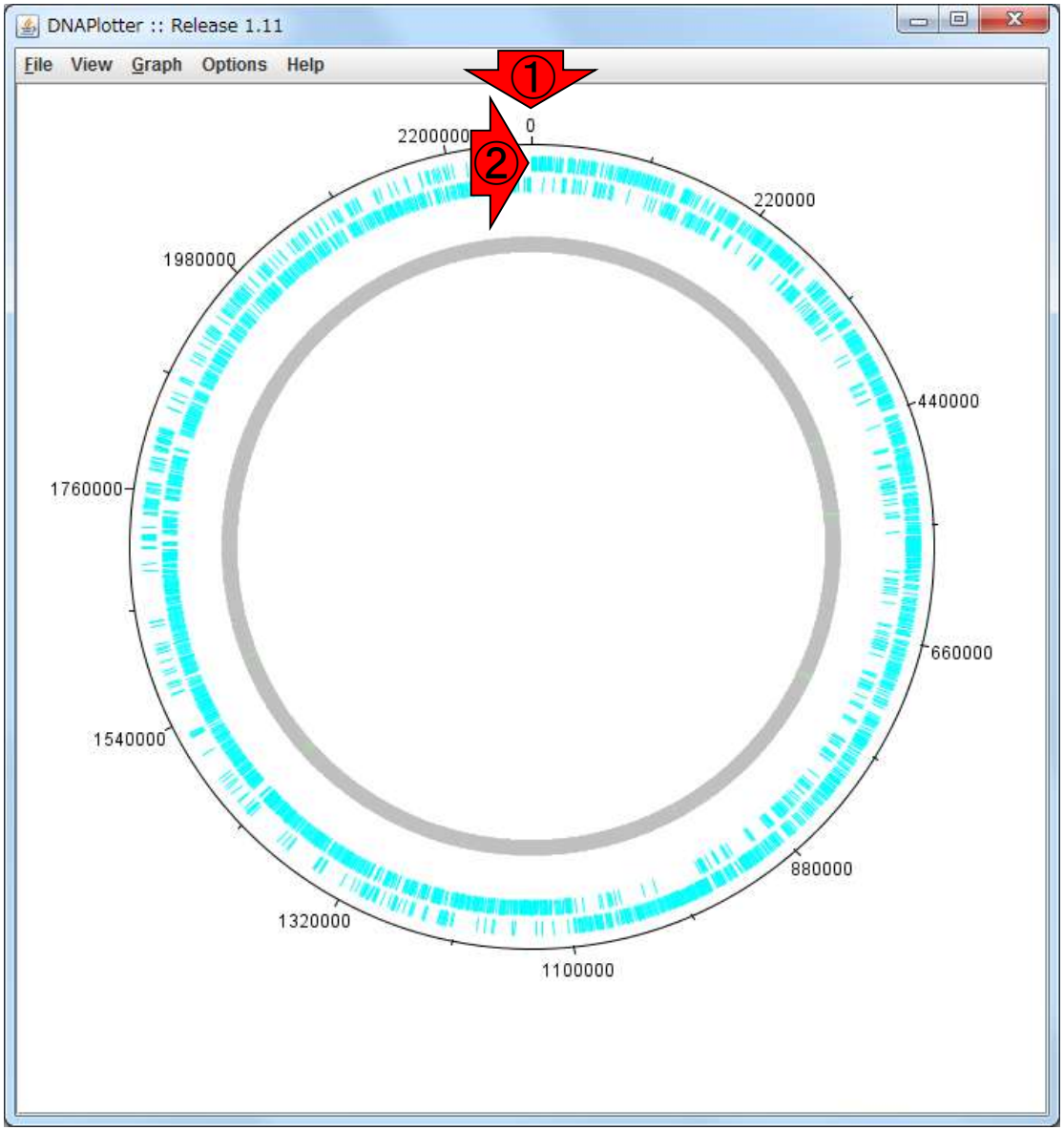


②内側の輪っかは、相補鎖上のCDSを表します

W12-4：解説



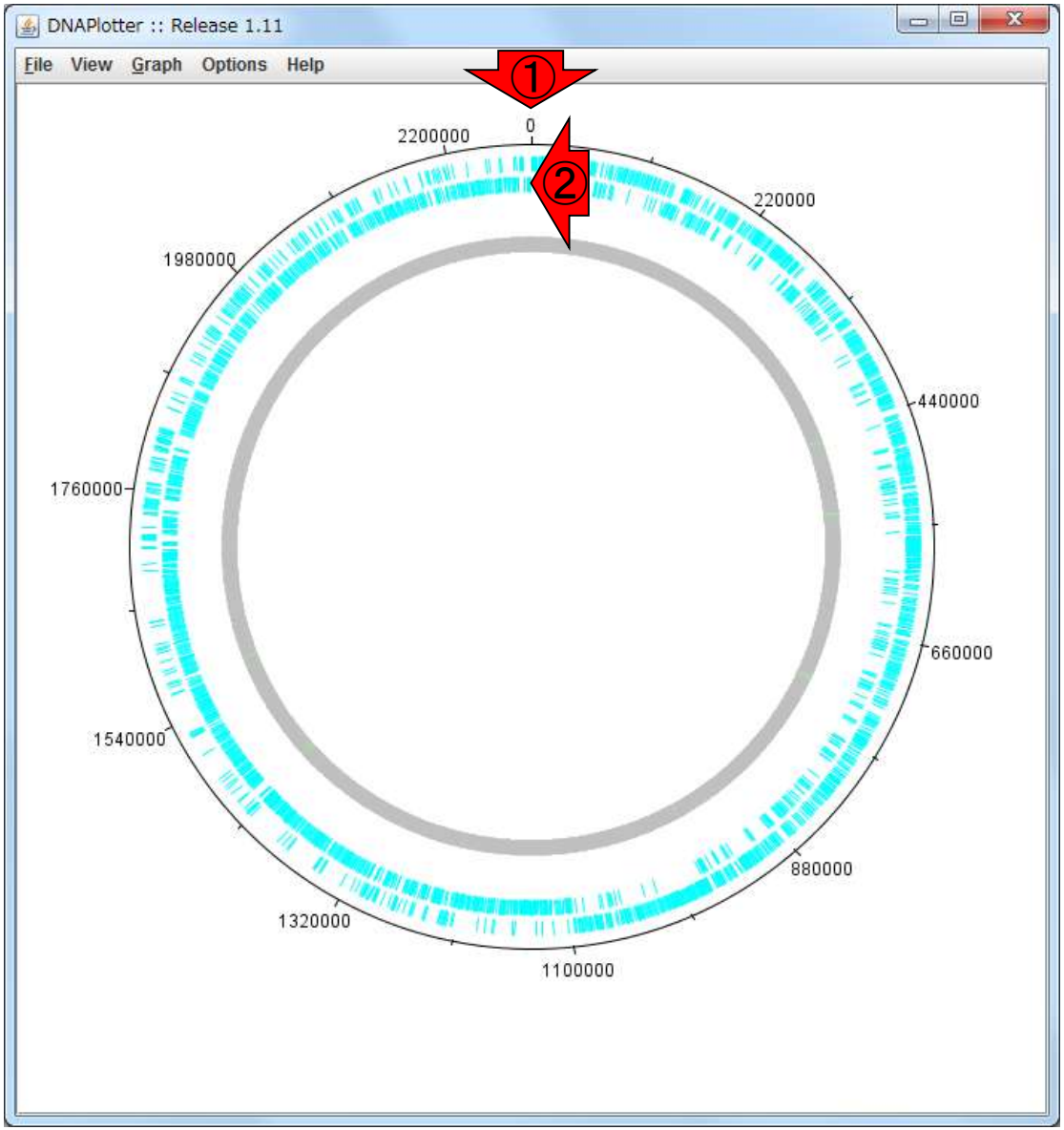
W12-4：解説



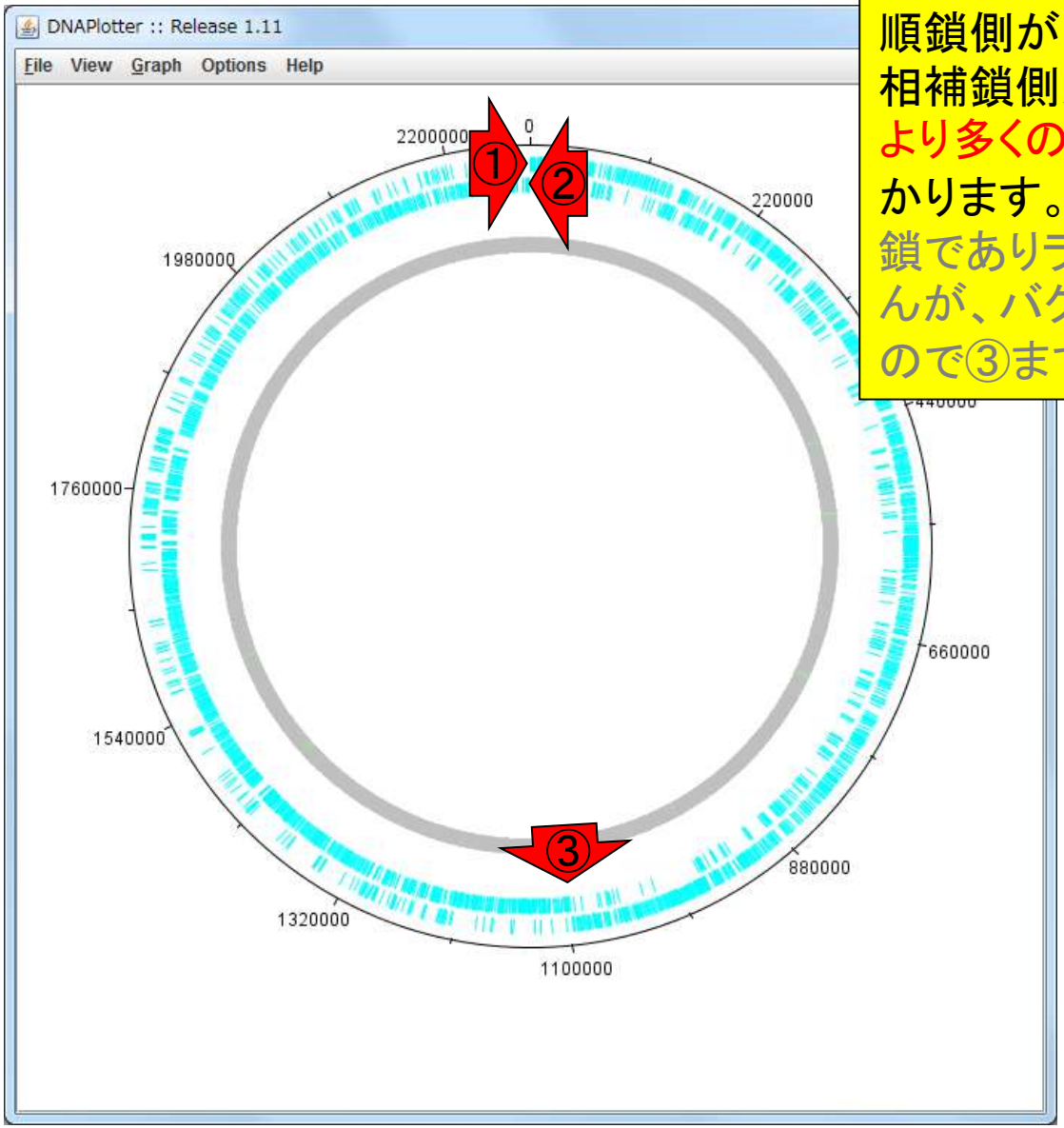
この配列は複製開始点が①の場所になるように回転させたものでした(W4)。したがって、複製開始点が開いて、②時計回りに複製が進む場合は順鎖がリーディング鎖

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



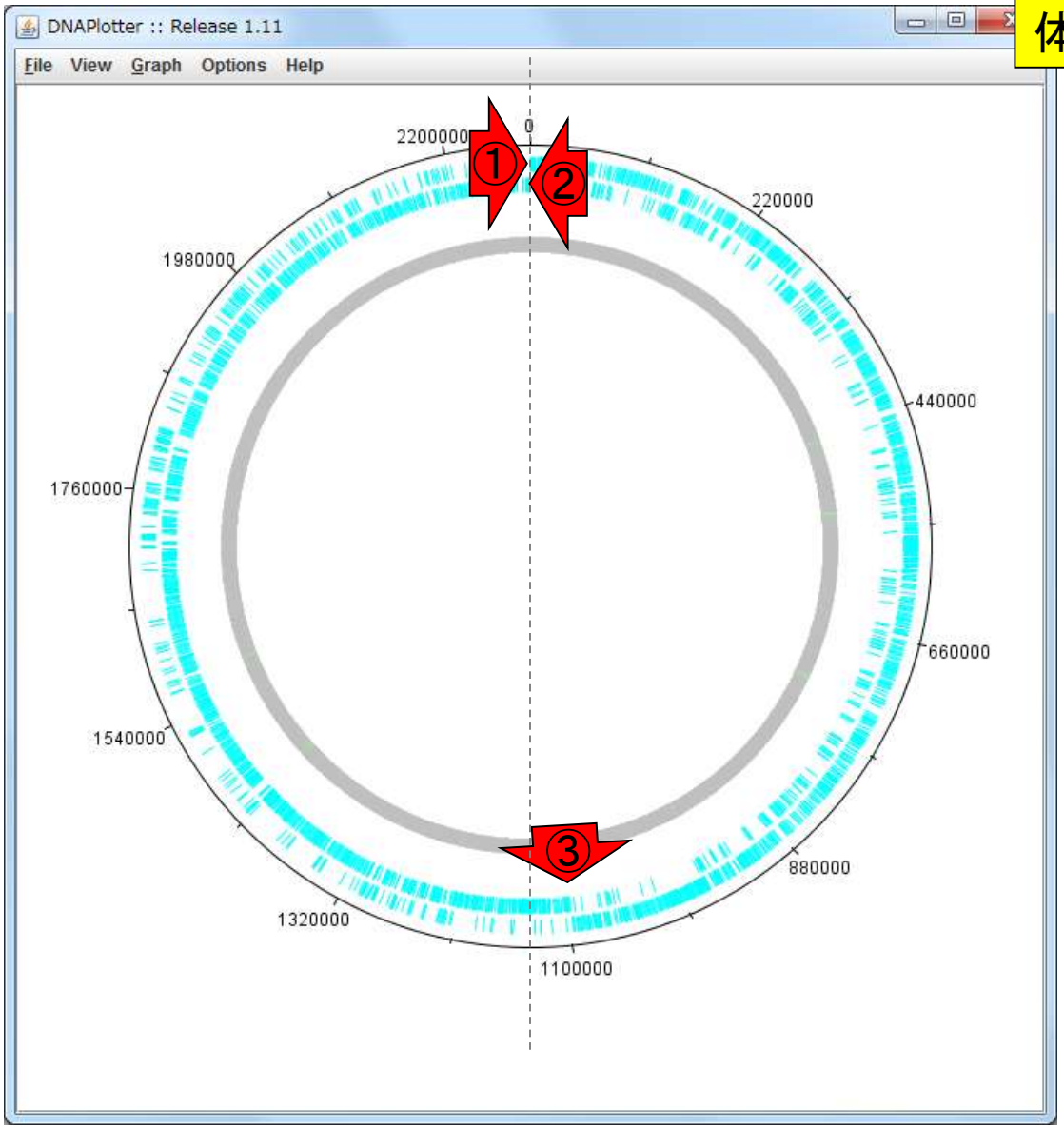
W12-4：解説

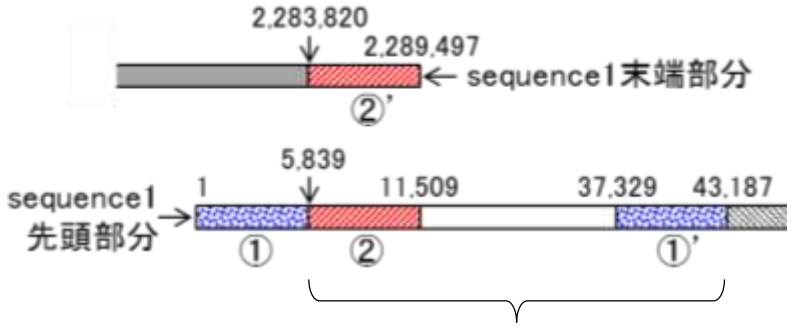


実際の複製では、①時計回りと②反時計回りの複製が同時に進むので、③のあたりが分かれ目になります。③までを区切りとして考えると、①時計回りに複製が進む順鎖側がリーディング鎖、②反時計回りに複製が進む相補鎖側がリーディング鎖となります。リーディング鎖により多くの遺伝子(CDS)が存在する傾向がはっきりとわかります。最後までぐるっと一周すれば全部リーディング鎖でありラギング鎖じゃないかと思われるかもしれませんが、バクテリアの場合は複製開始点が1か所のみなので③までを区切りとして考えるのでよいと思います

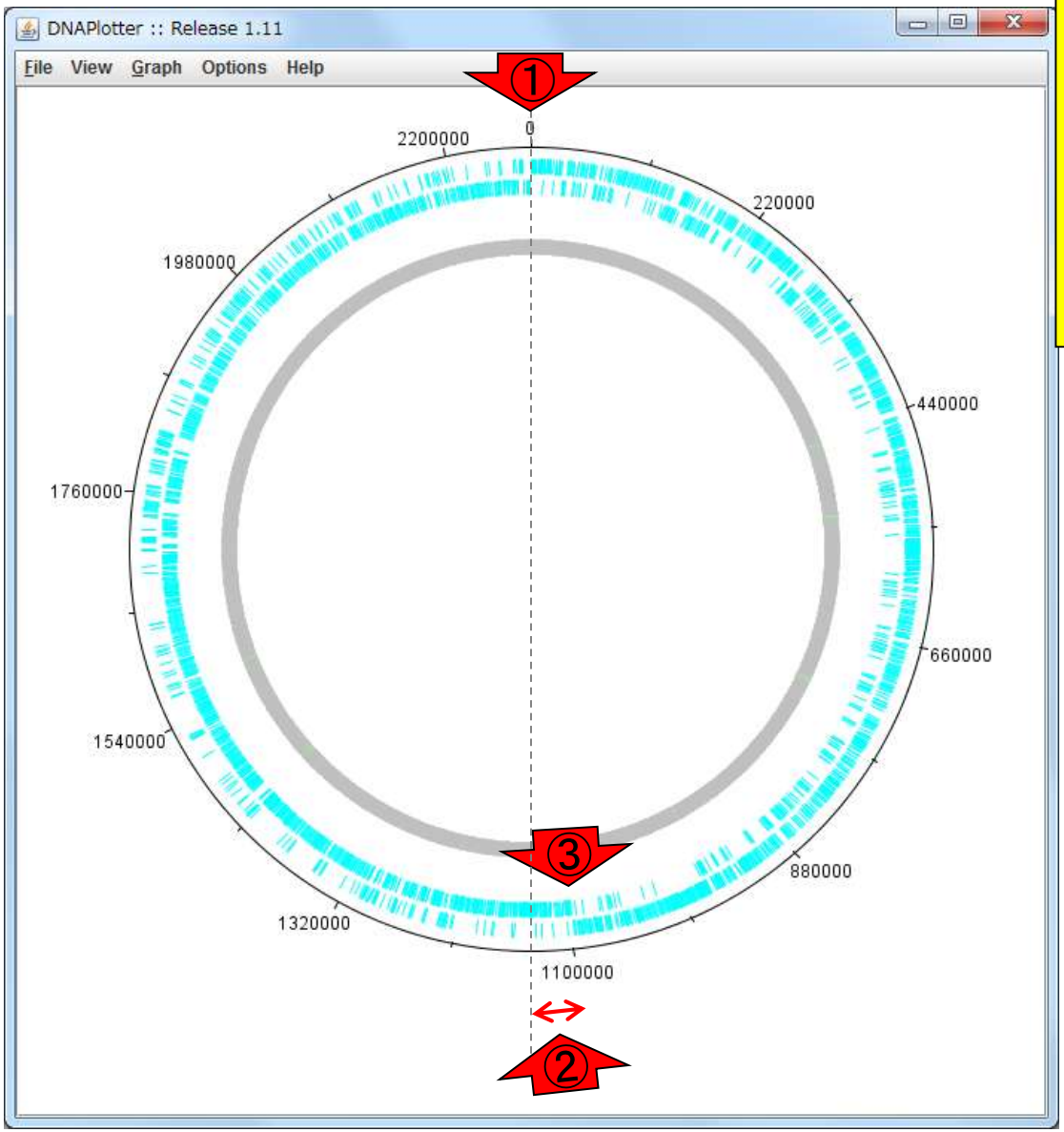
W12-5：ずれの理由

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

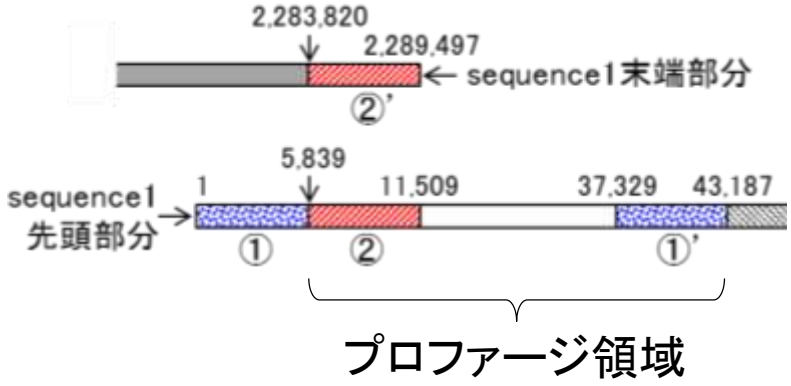




W12-5：ずれの理由

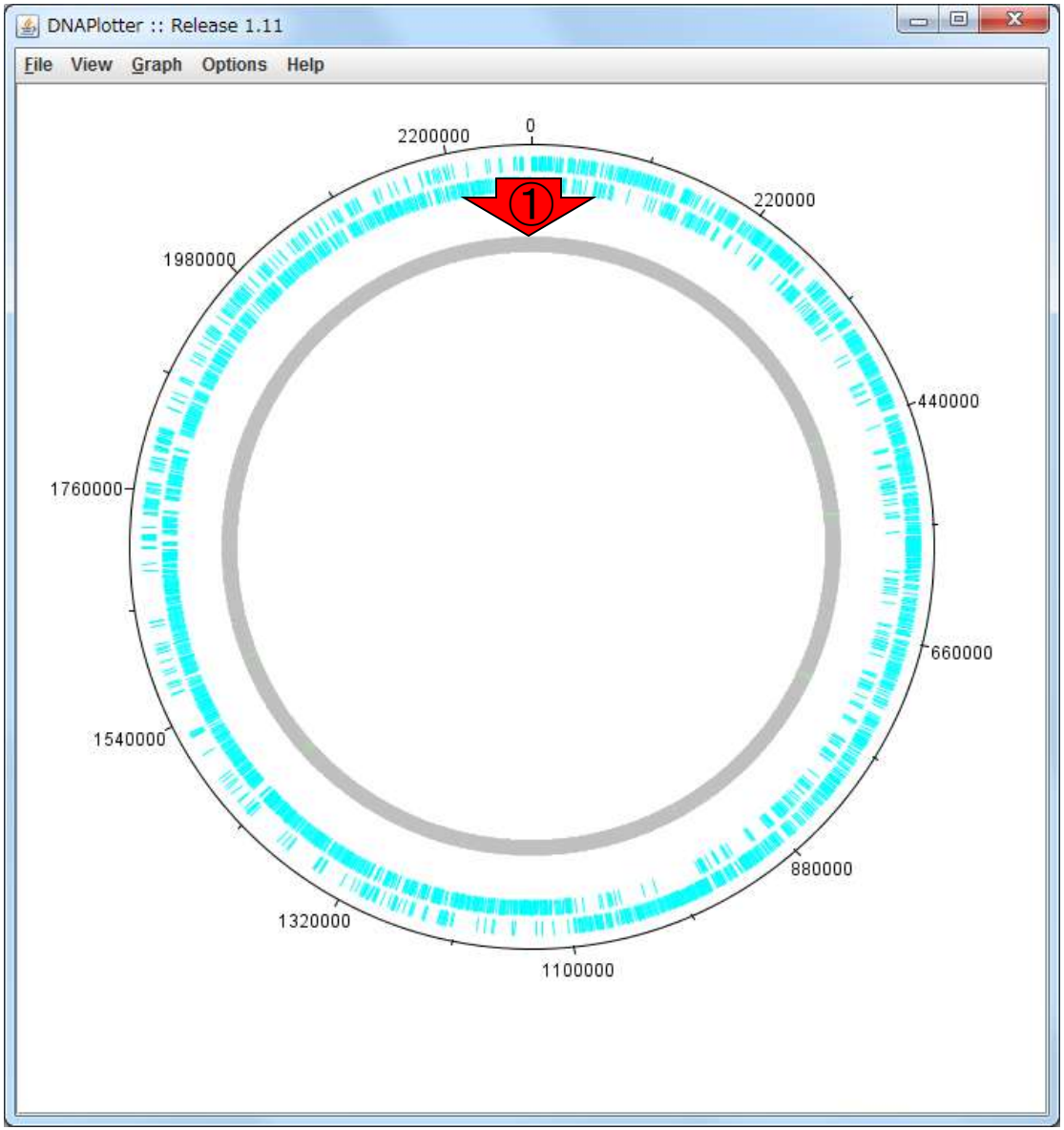


①sequence1の配列長は2,277,983 bpなので、②の長さが40,000 bp弱。プロファージ領域が(43,187 − 5,839 + 1) = 37,349 bp程度なので、③程度の若干のずれは妥当と当初判断しましたが、それだと③が現在の時計の長針で28分ごろの位置ではなく32分ごろの位置にならないとオカシイのではないかと思うに至っております。ですから、プロファージ領域のみが原因ではないと思われます

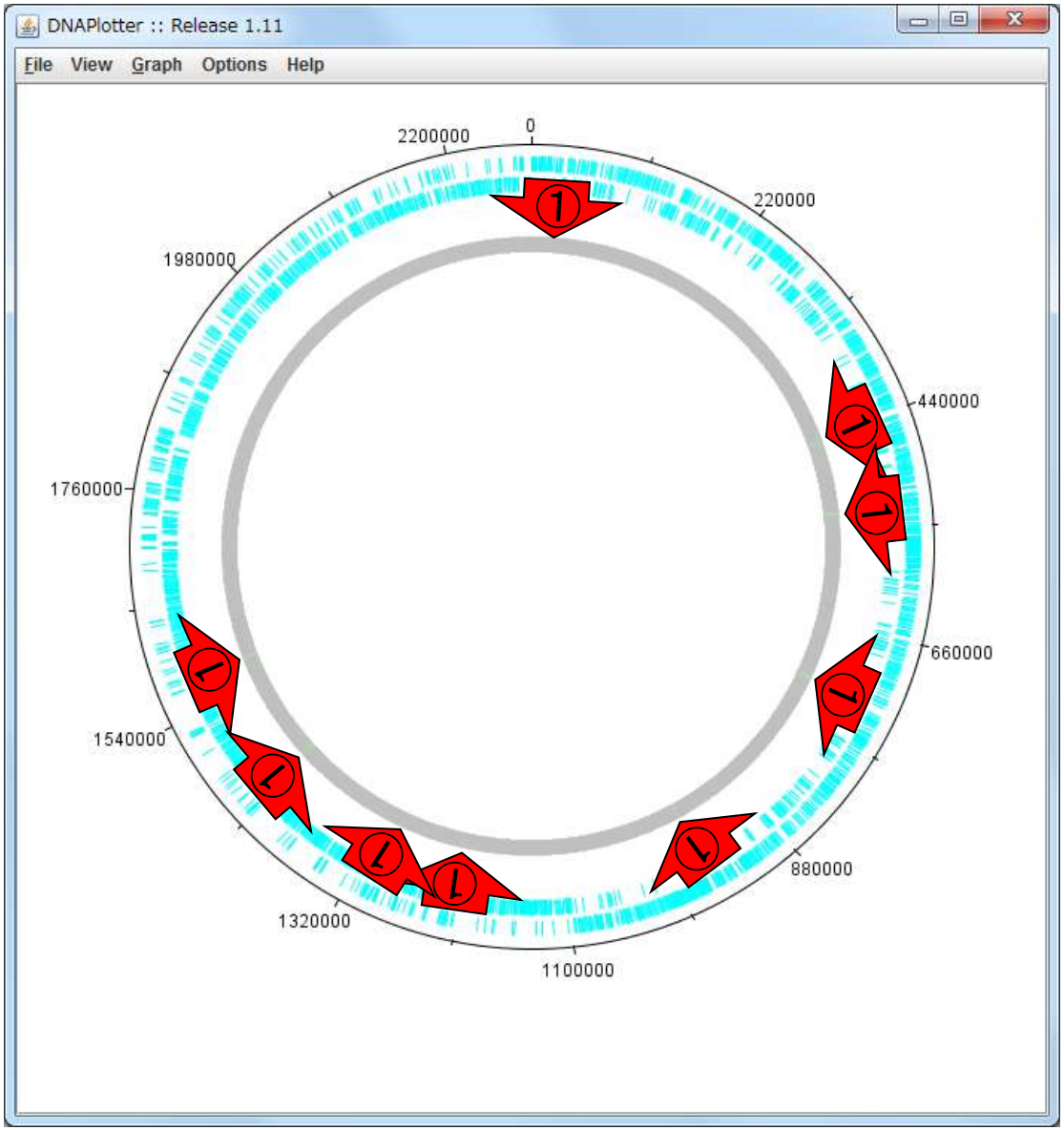


W12-6：tRNA

①灰色の輪っかは、tRNAの位置情報を表します



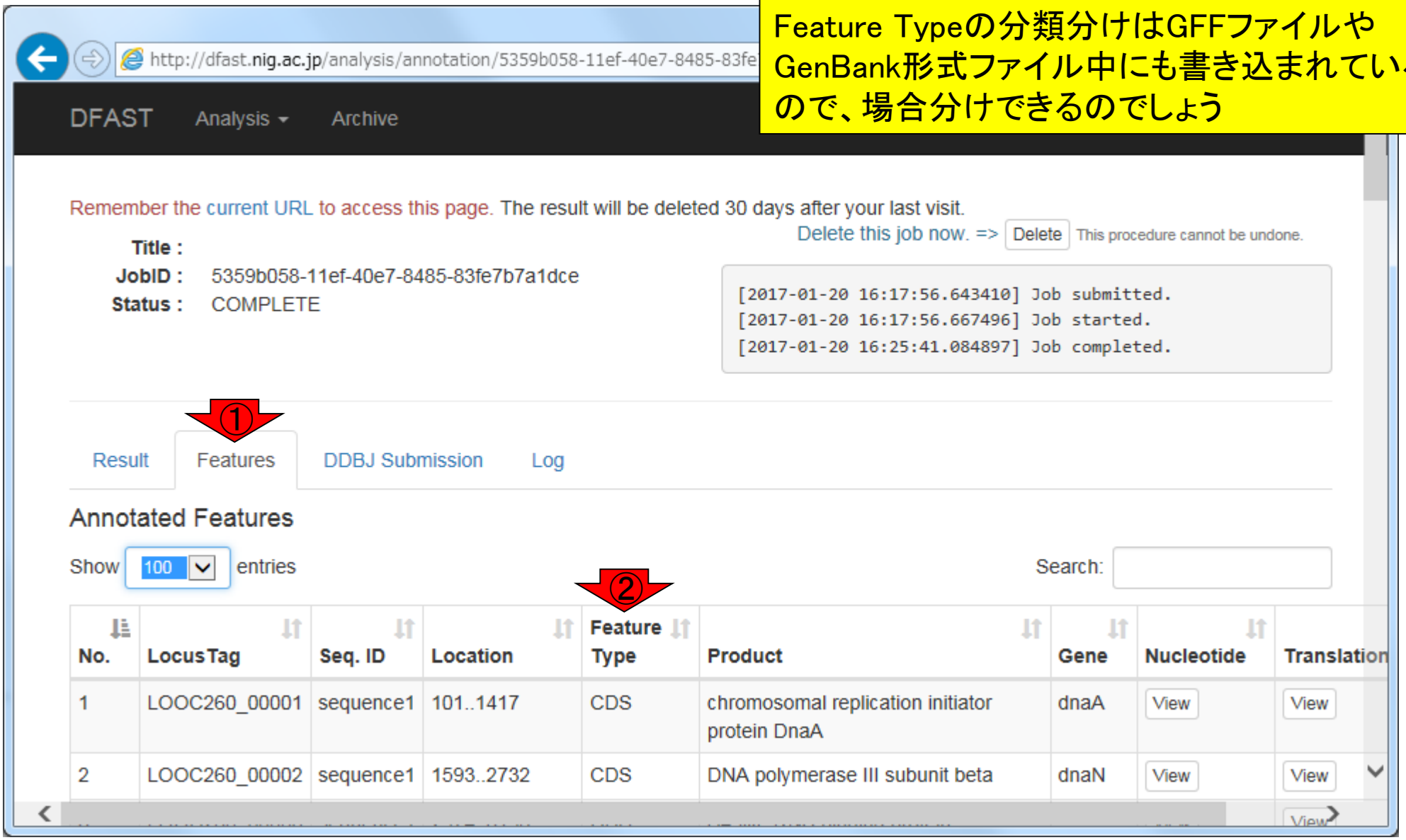
W12-6：tRNA



よーく見ると、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

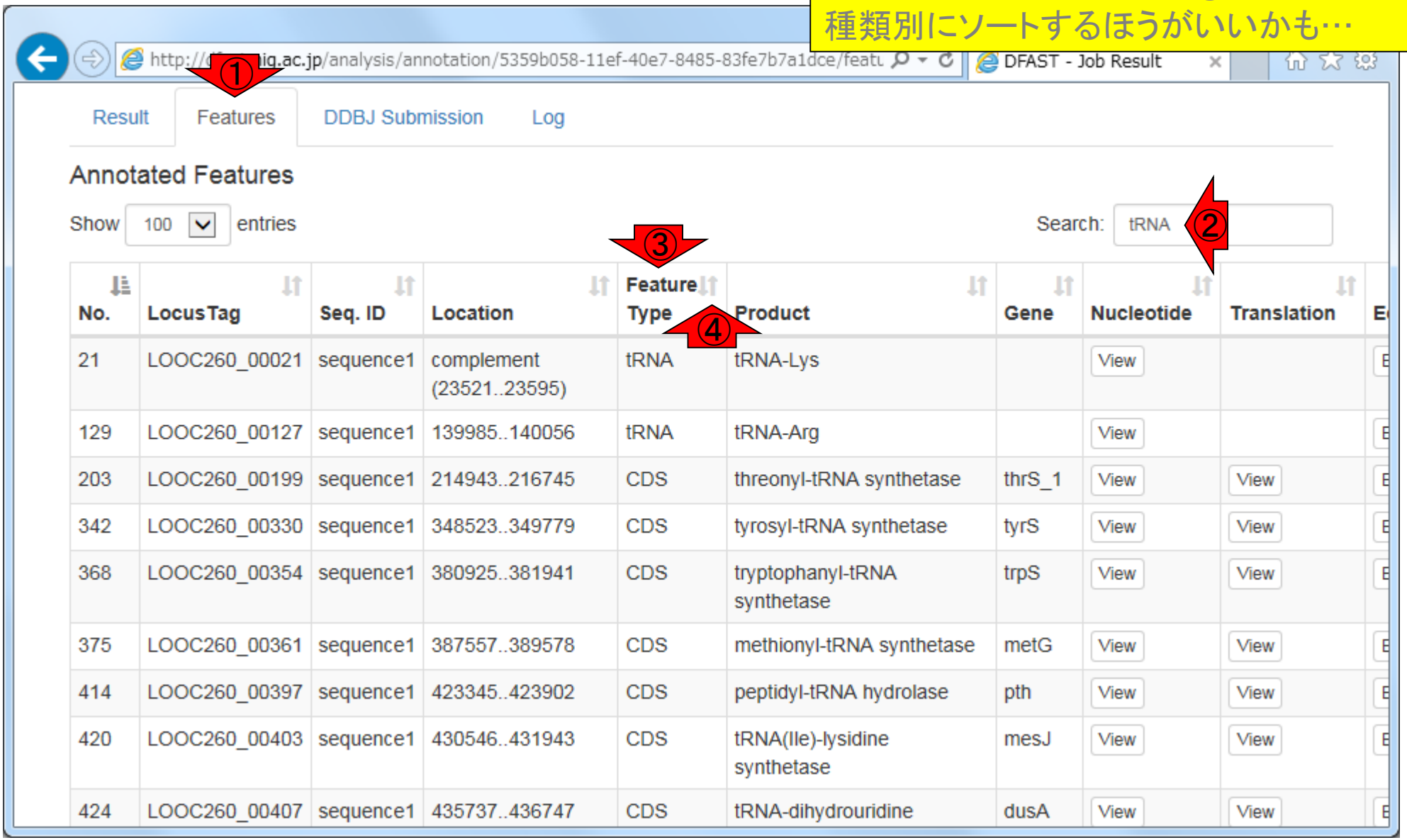
W12-7：DFAST

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ょう



W12-7：DFAST

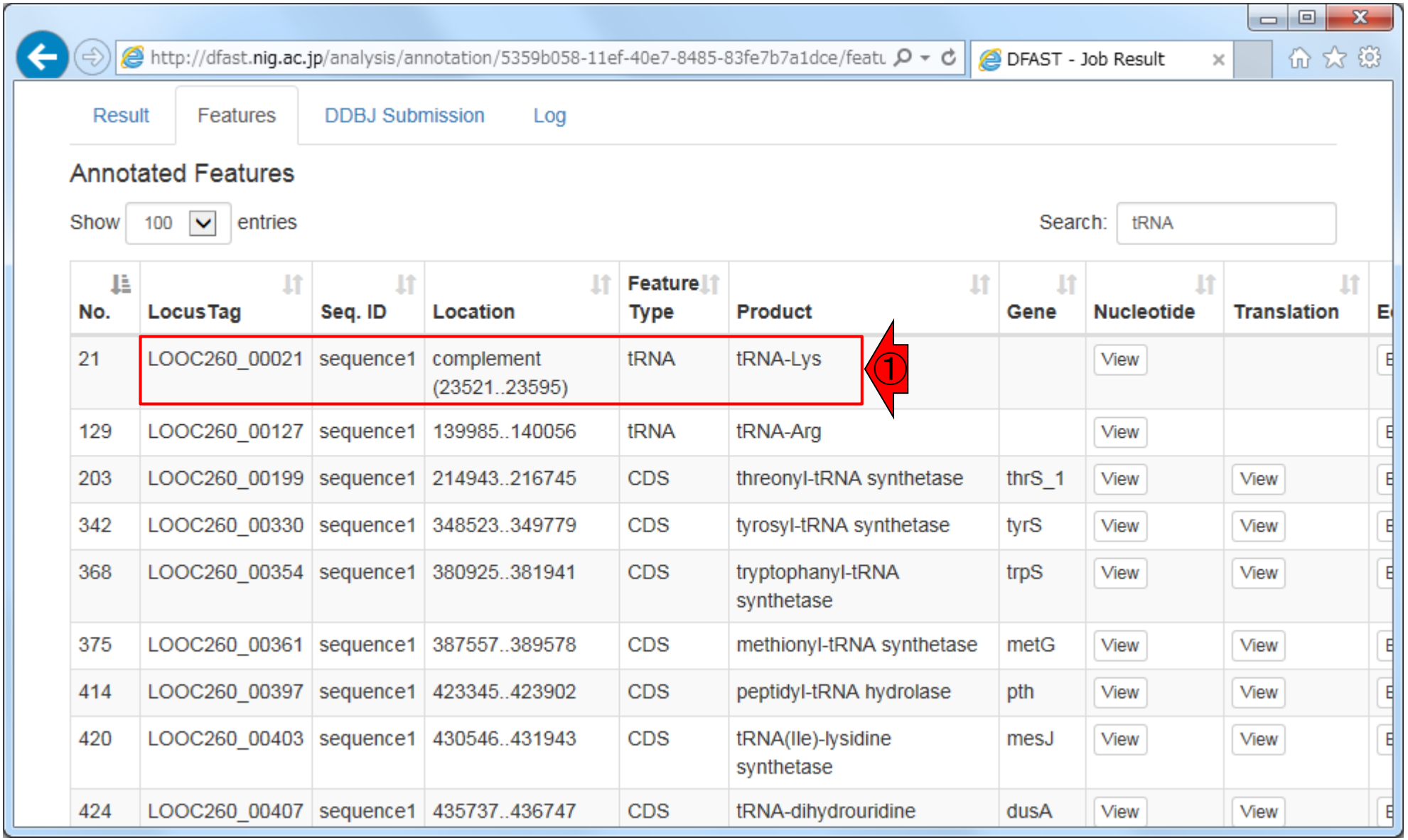
DFASTの①Featureタブ内で、②tRNAなどで検索すると、③Feature TypeがtRNAとなっているものがみられます。④のあたりとかで種類別にソートするほうがいいかも…



No.	Locus Tag	Seq. ID	Location	Feature Type	Product	Gene	Nucleotide	Translation	E
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

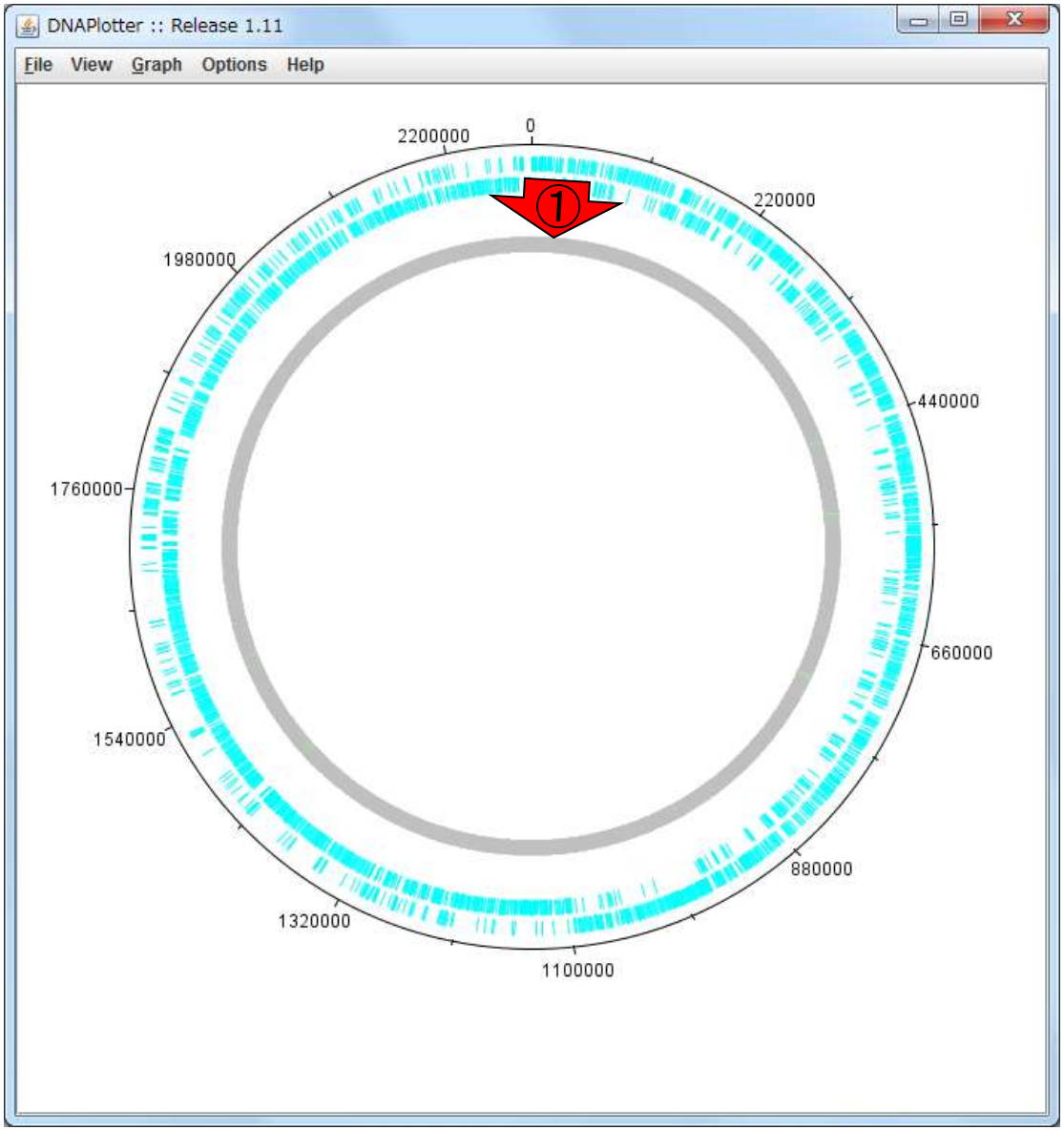
例えば①のtRNAは23500番目の塩基あたりに相当するので…



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

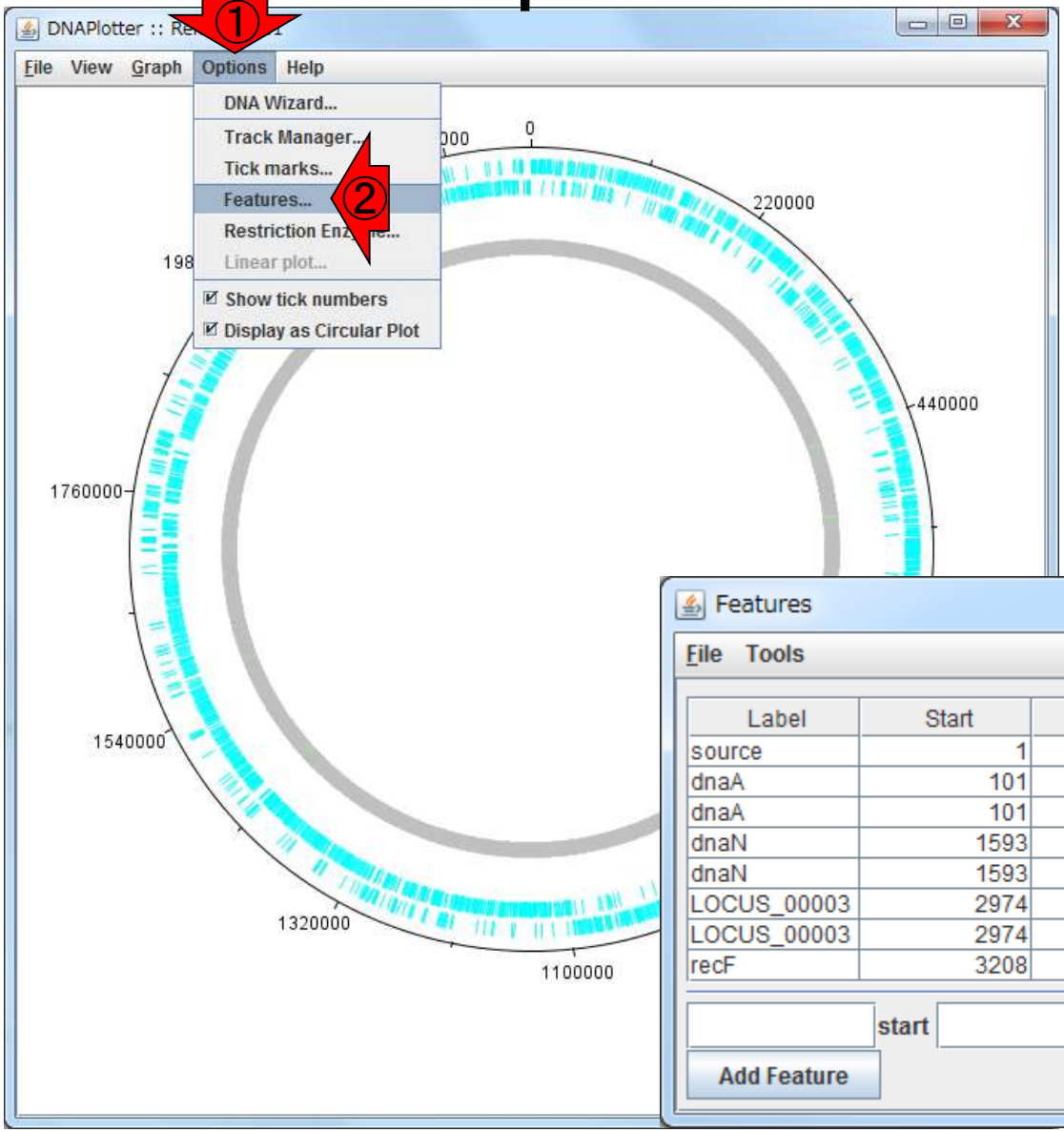
①の緑色の線に相当するのでしょう

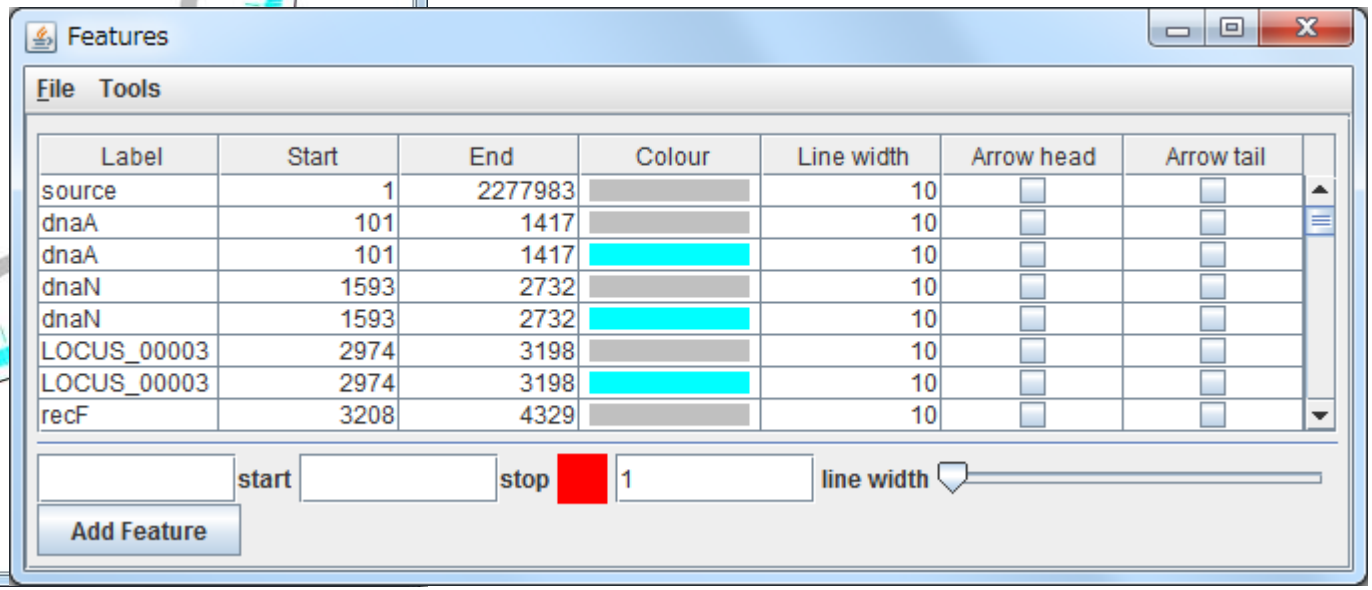
W12-8：23500番目あたり



W12-9：Options

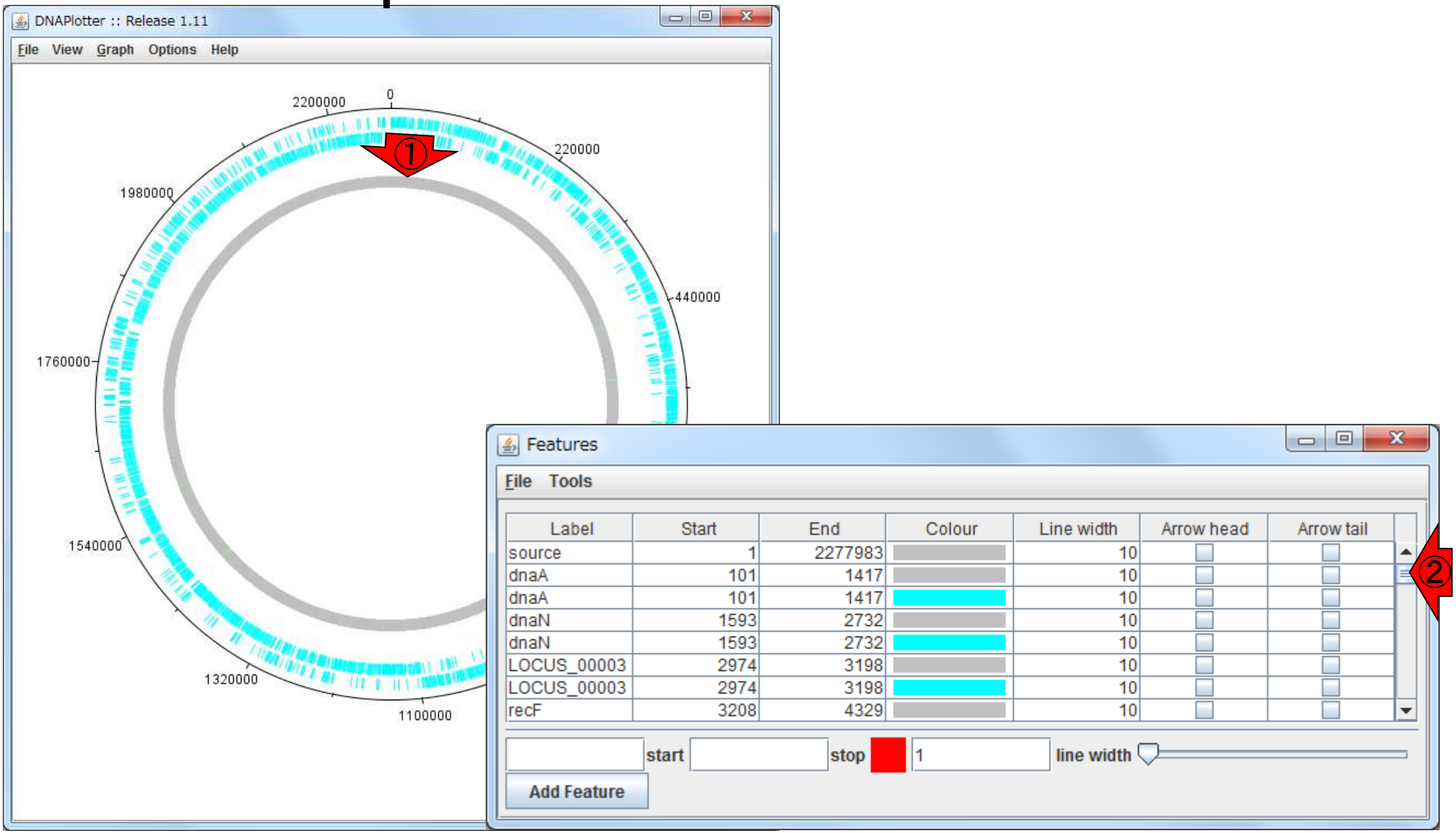
①Optionsの②Featuresで、より詳細なFeatures（CDS or tRNA)の情報をみることができます。どれが水色で表示されていて、どれが緑色で表示されているかなどです





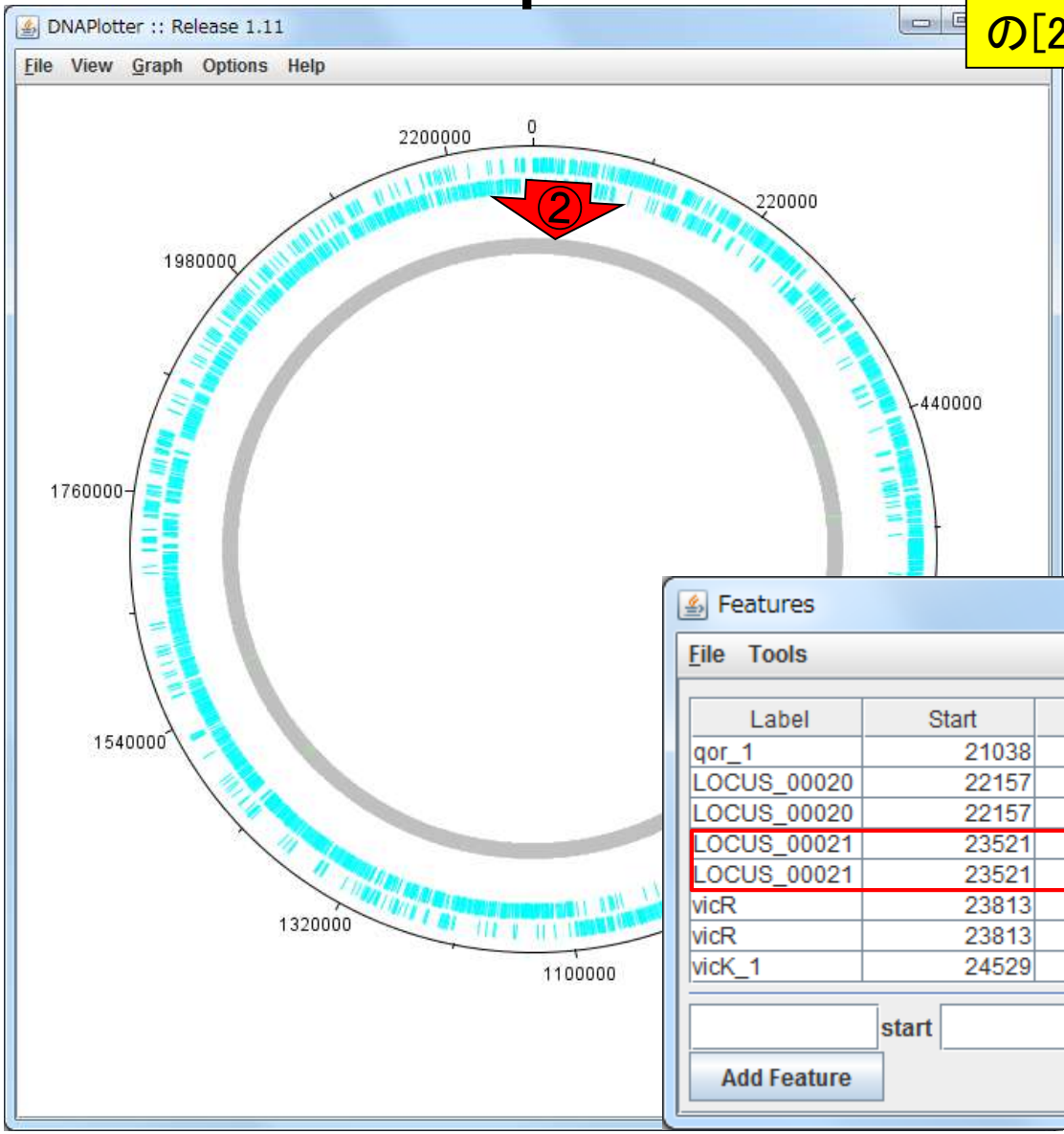
W12-9：Options

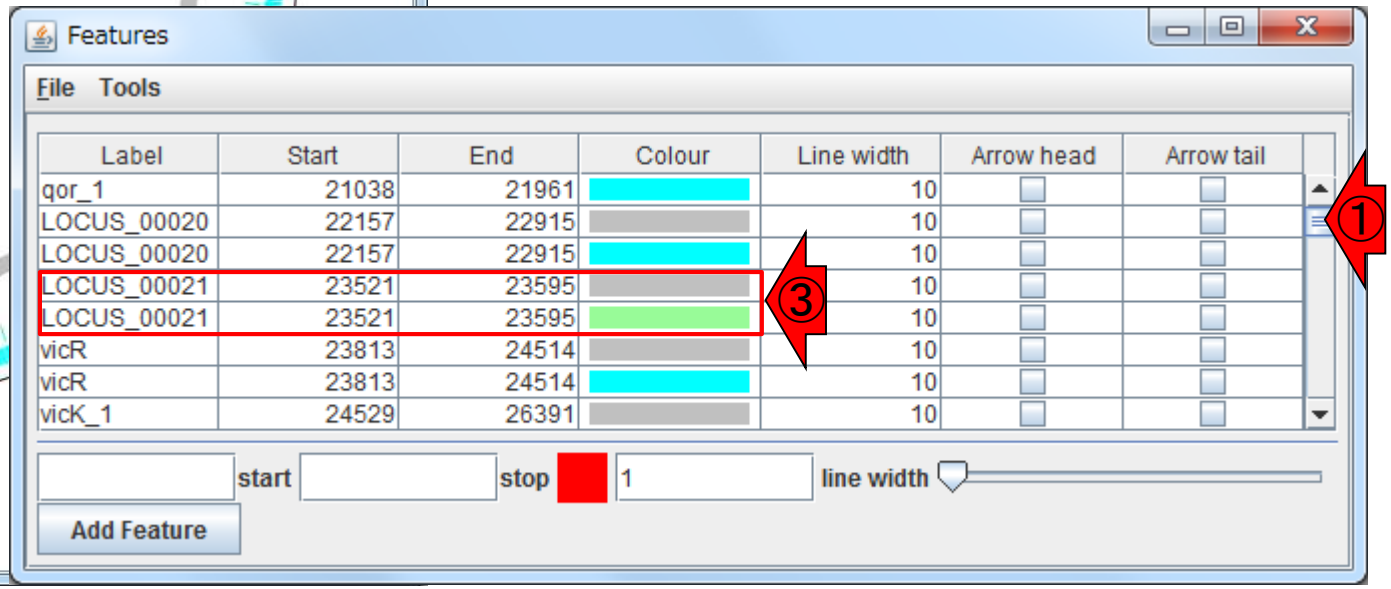
①緑色に見えている23500番目あたりを実際に調べてみる。②のカーソルを下に移動



W12-9：Options

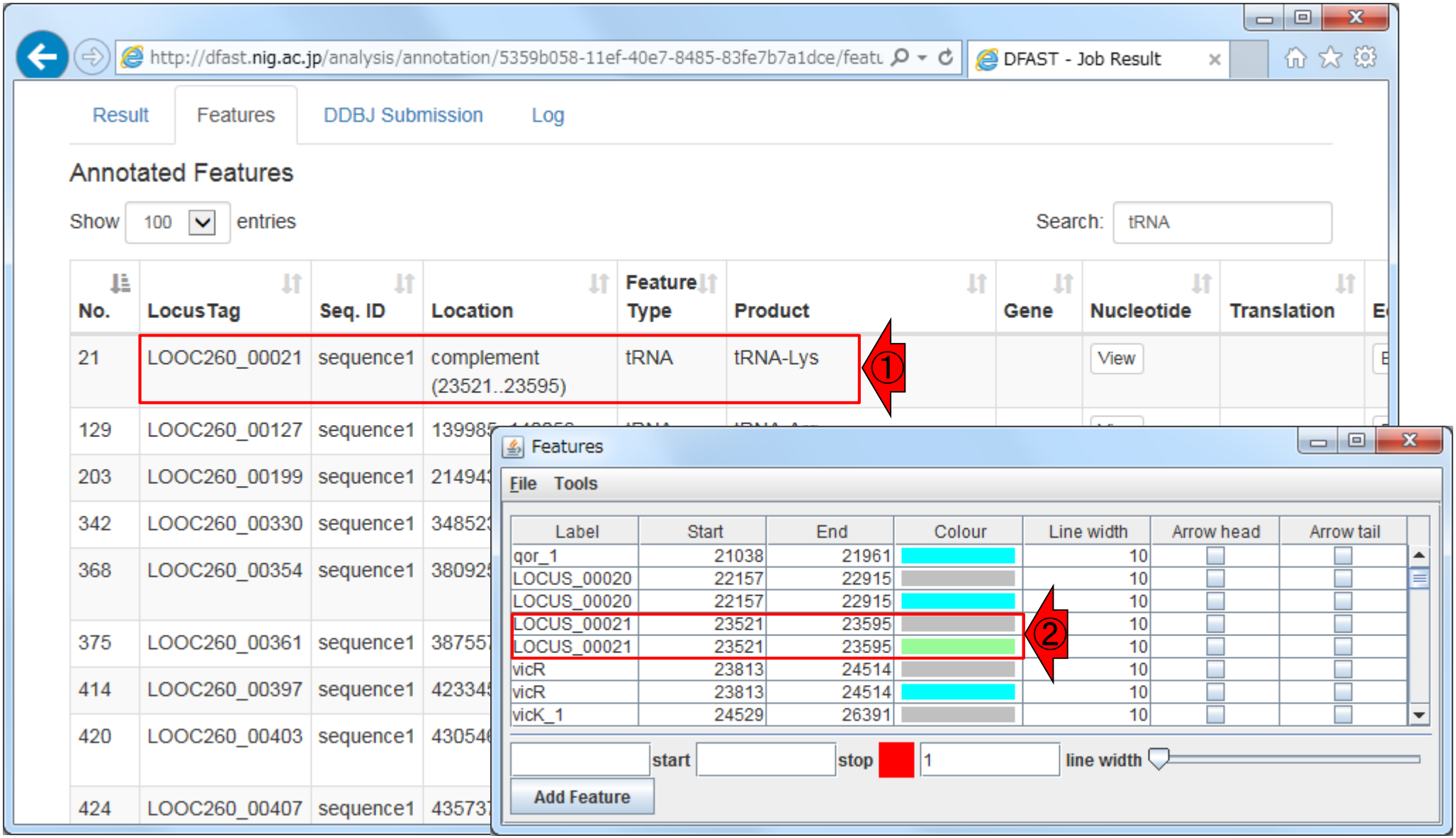
①カーソルを少し下に移動させ、だいたい23500番目あたりでColourが緑色になっているところを探した結果。②の緑色は、③赤枠で示すLOCUS_00021の[23521, 23595] が見えているのだろう





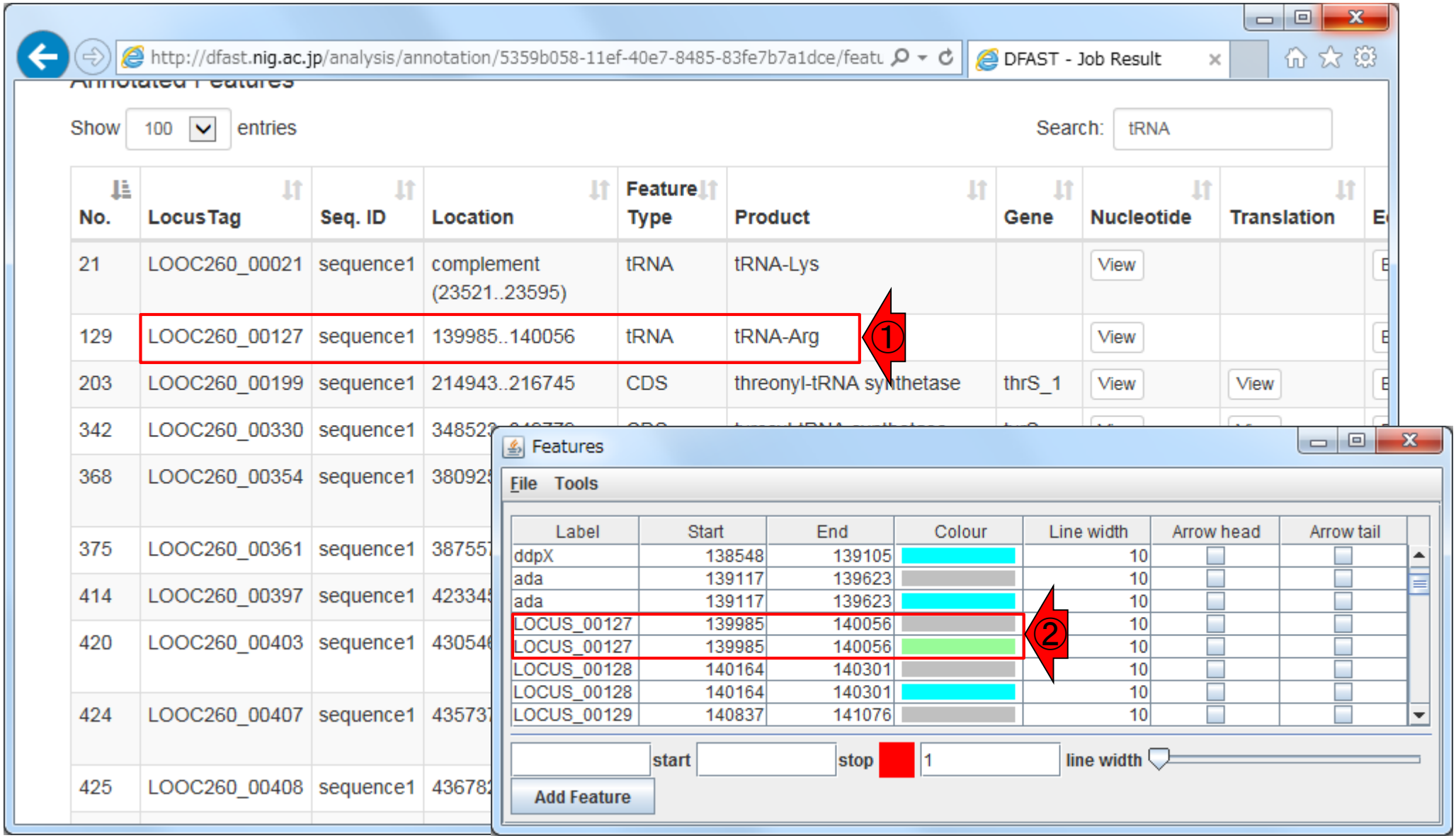
妥当ですね

W12-9：DFASTで確認



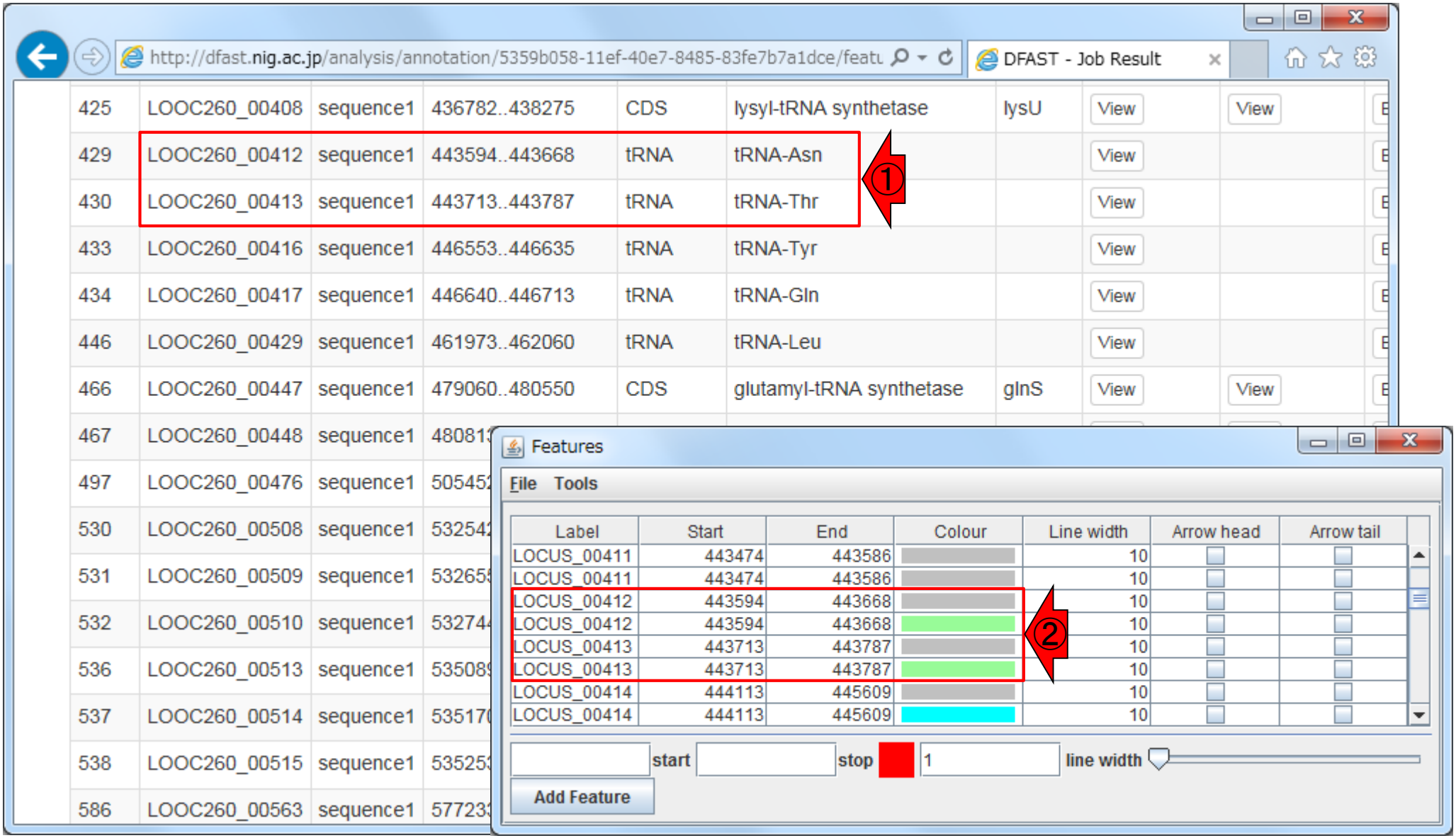
これも妥当です

W12-9：DFASTで確認



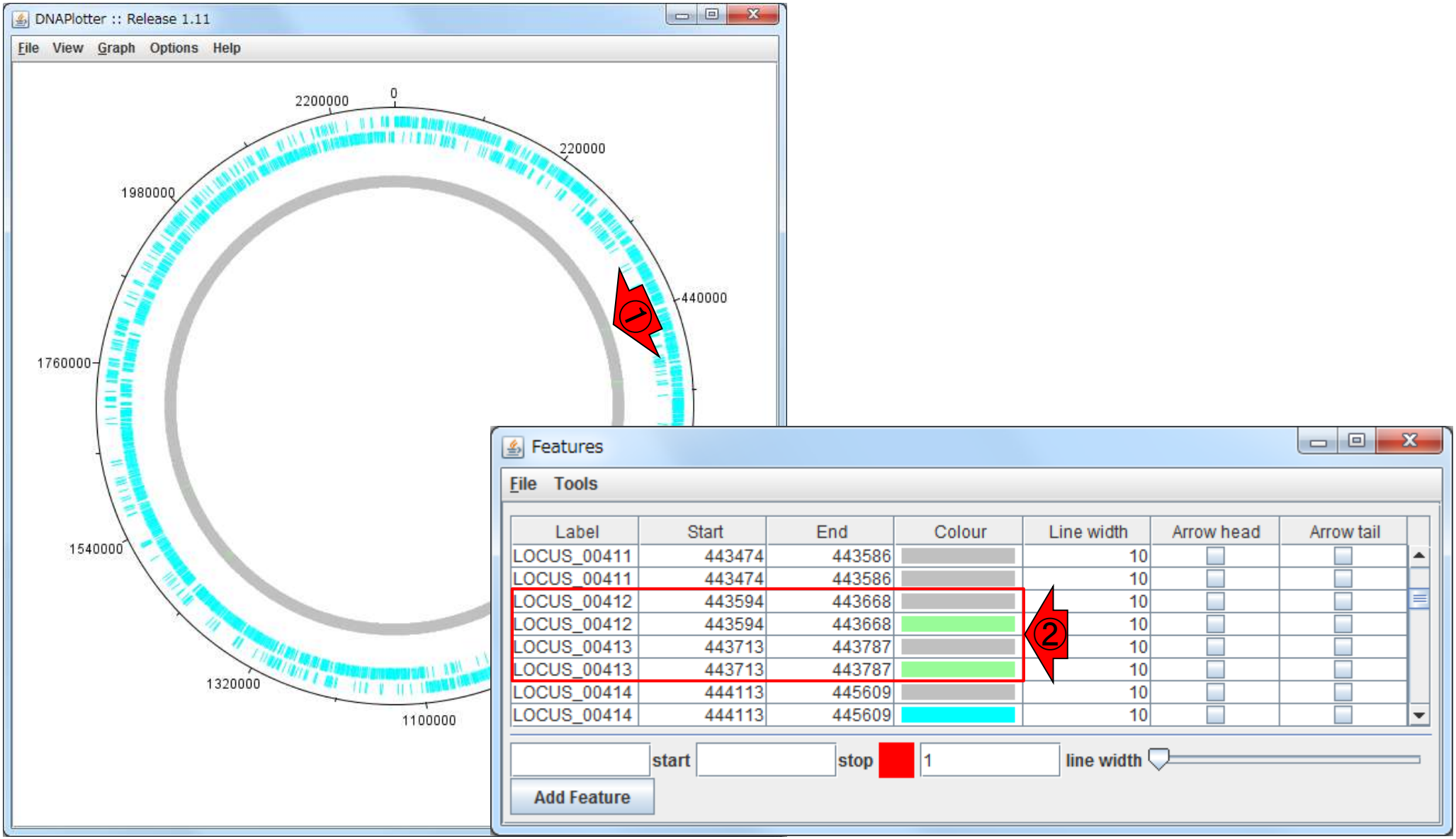
完璧です

W12-9：DFASTで確認



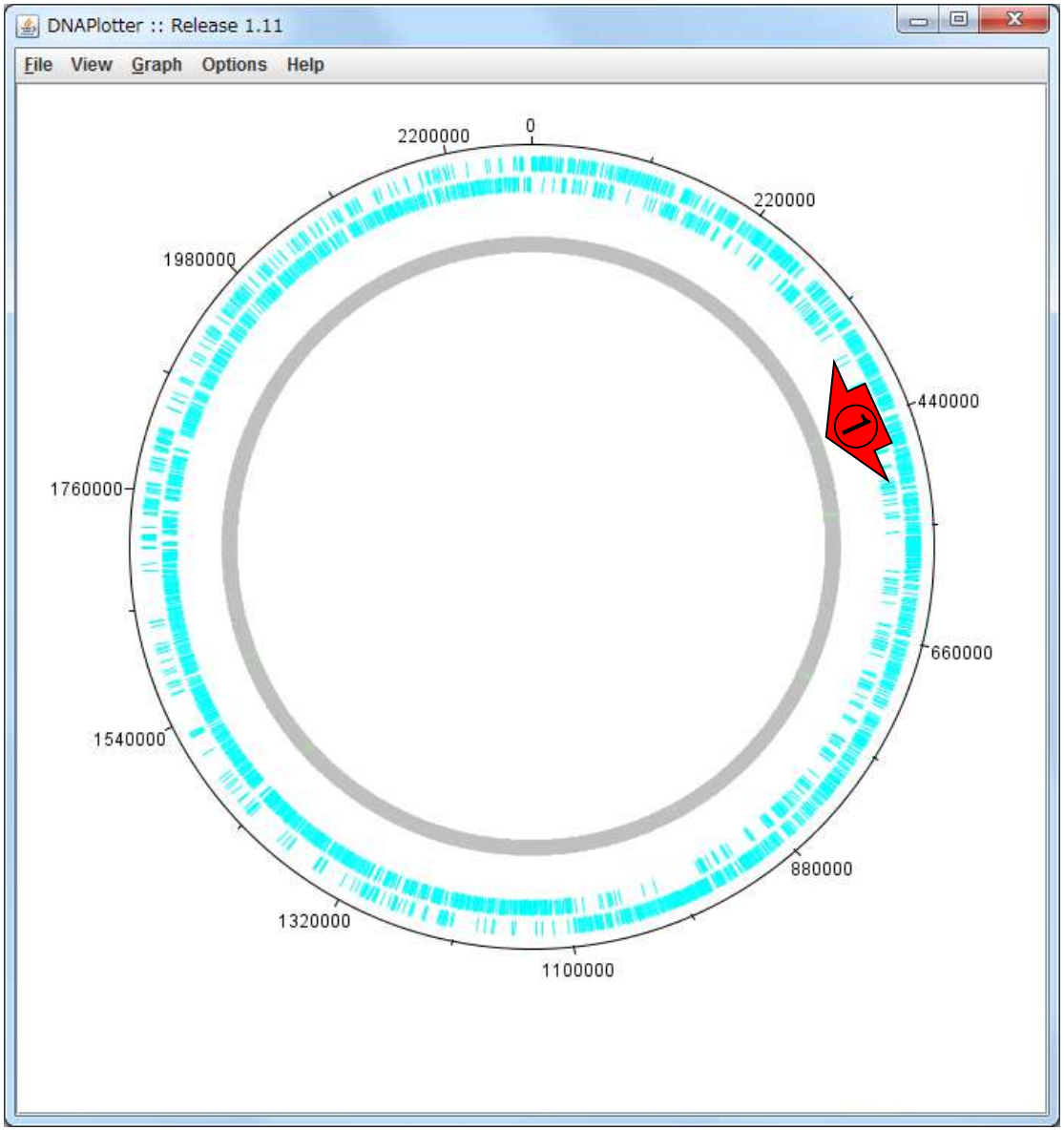
完璧です

W12-9：DFASTで確認



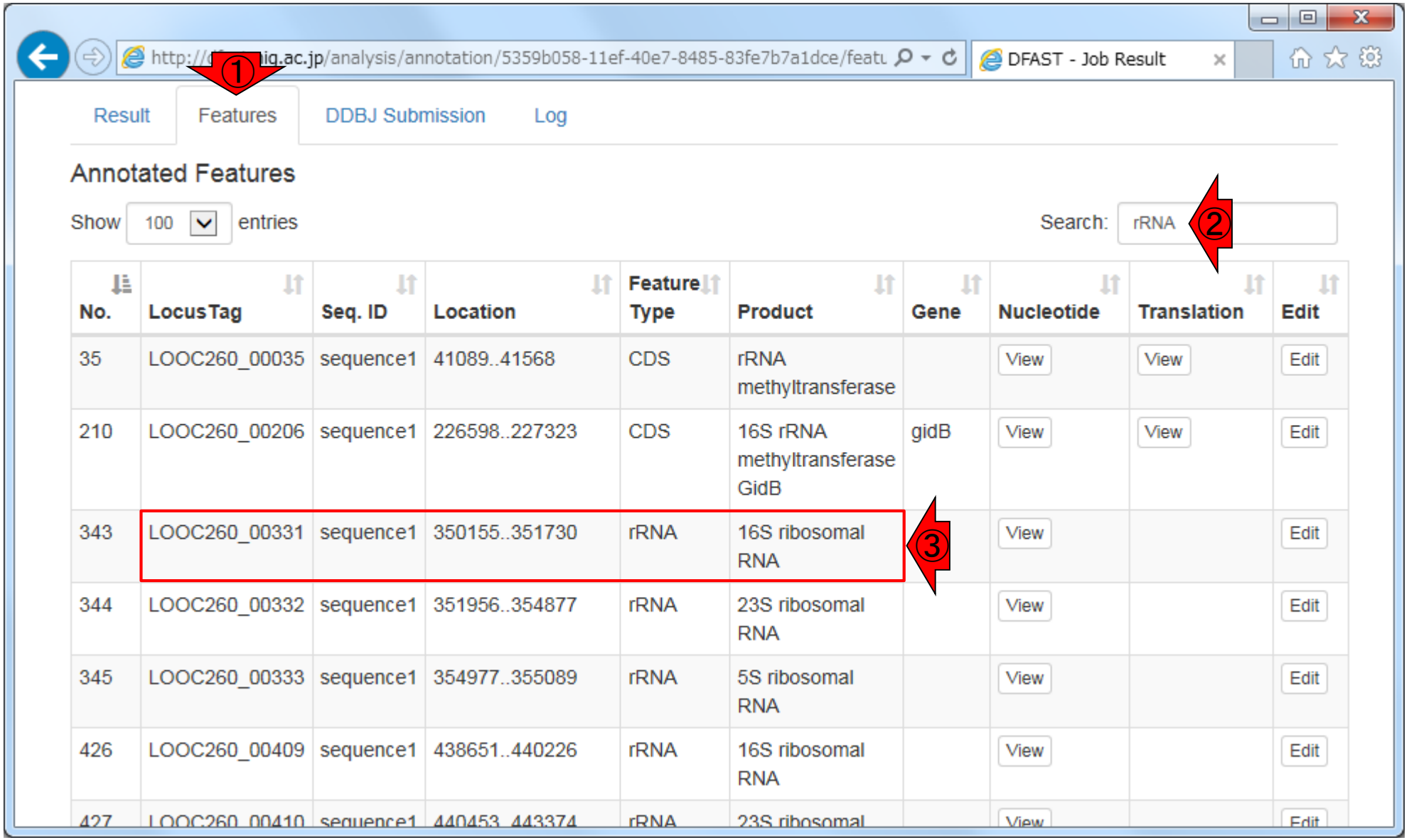
W12-10：ちなみに…

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だけでのようです



W12-10：ちなみに…

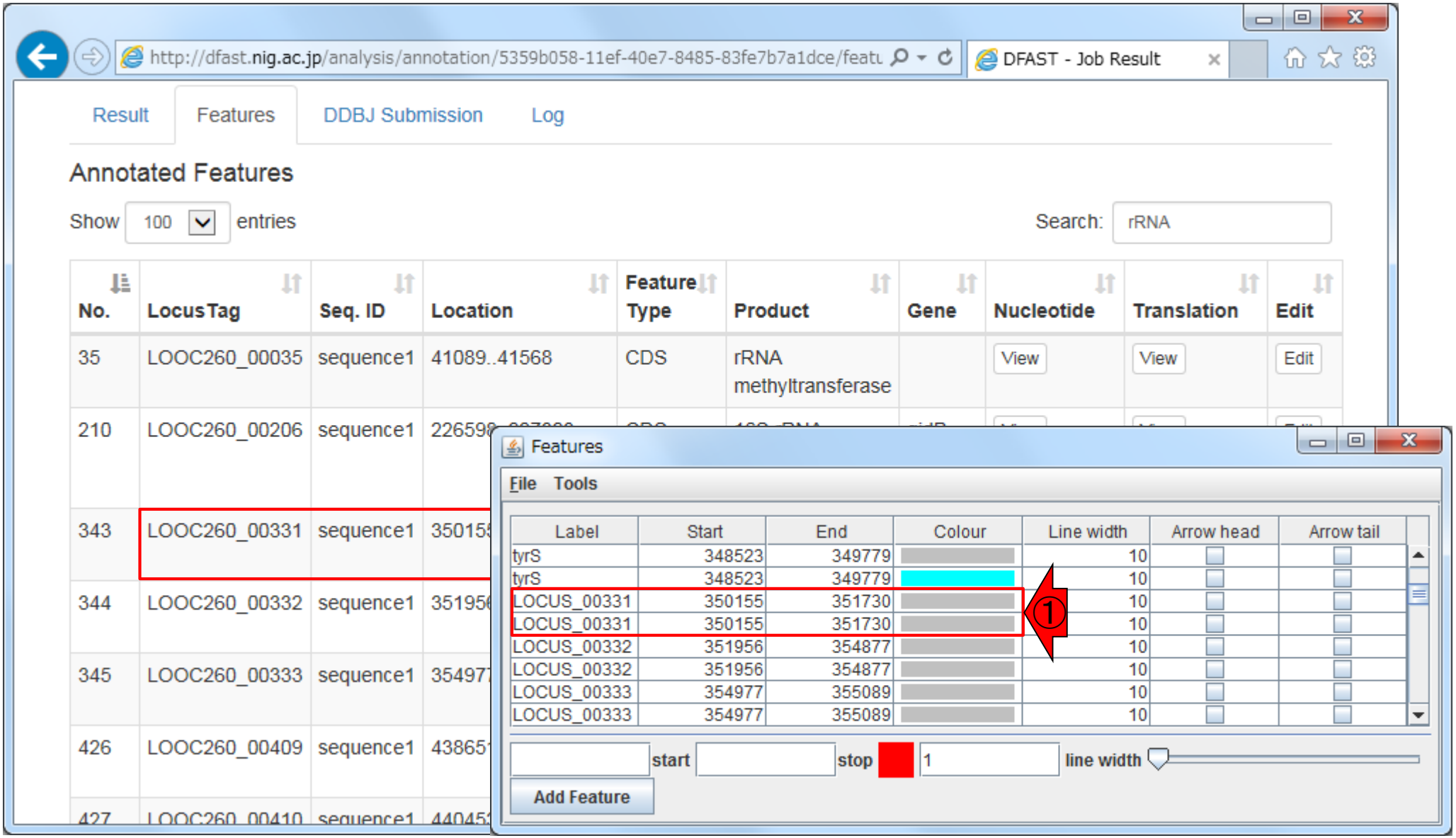
例えば、DFASTの①Featureタブ上で、②rRNAで検索して得られる、③は…



No.	Locus Tag	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		View		Edit

①確かに緑色にはなっていません

W12-10：ちなみに…



W12-10：ちなみに…

DNAPlotterのページのちょっと下のほうに…

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

DNAPlotter | Sanger I...

wellcome trust sanger institute

SCIENCE

DNAPlotter

Overview

Download

Learn

License

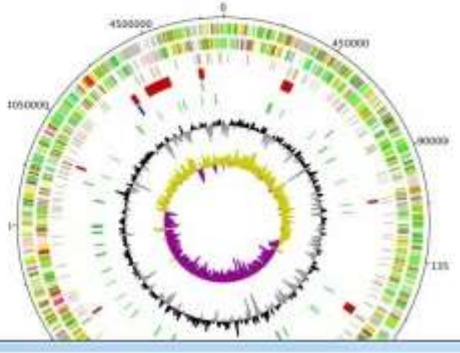
Overview

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.

Tool type

- Annotation

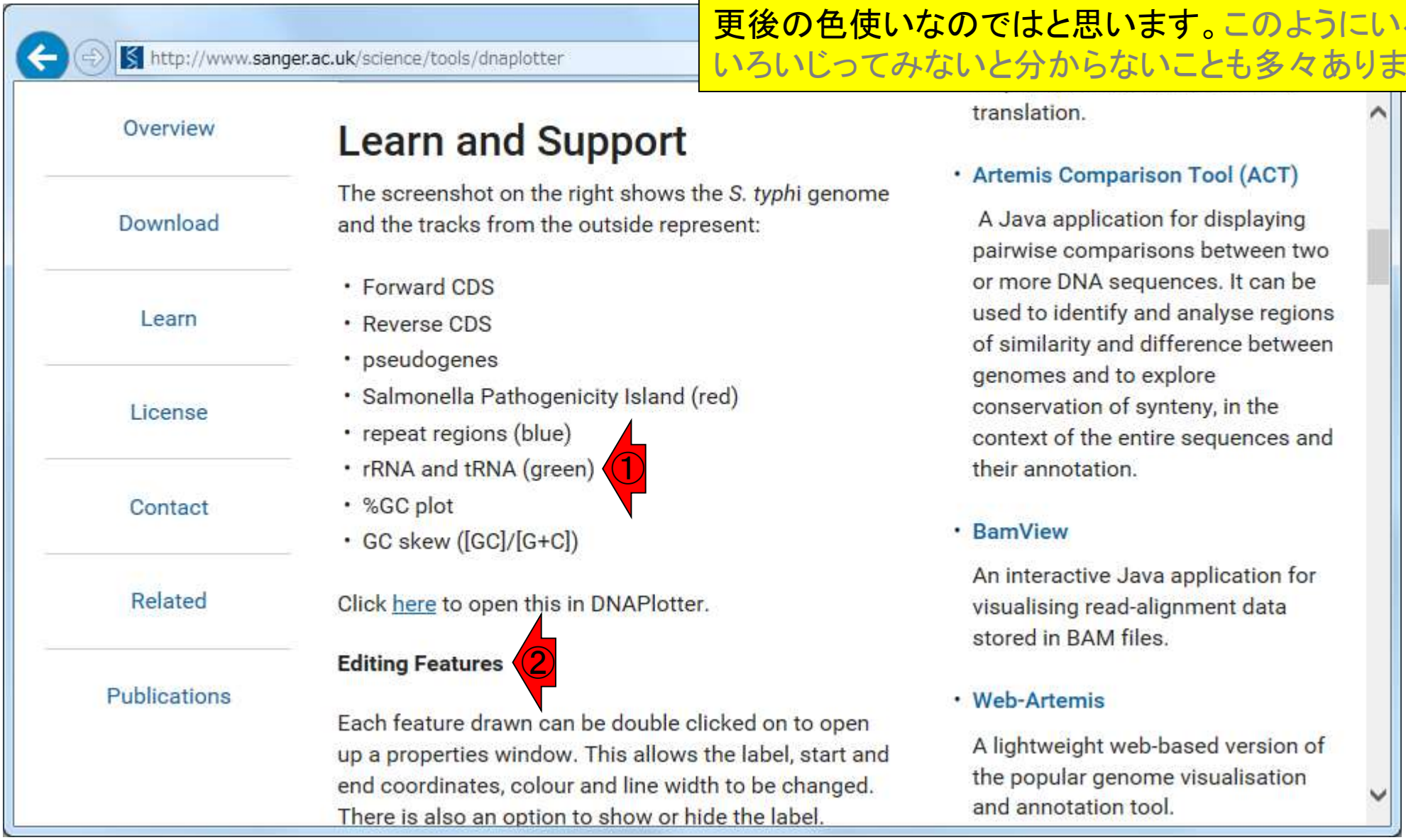
Screenshots



Download and Installation

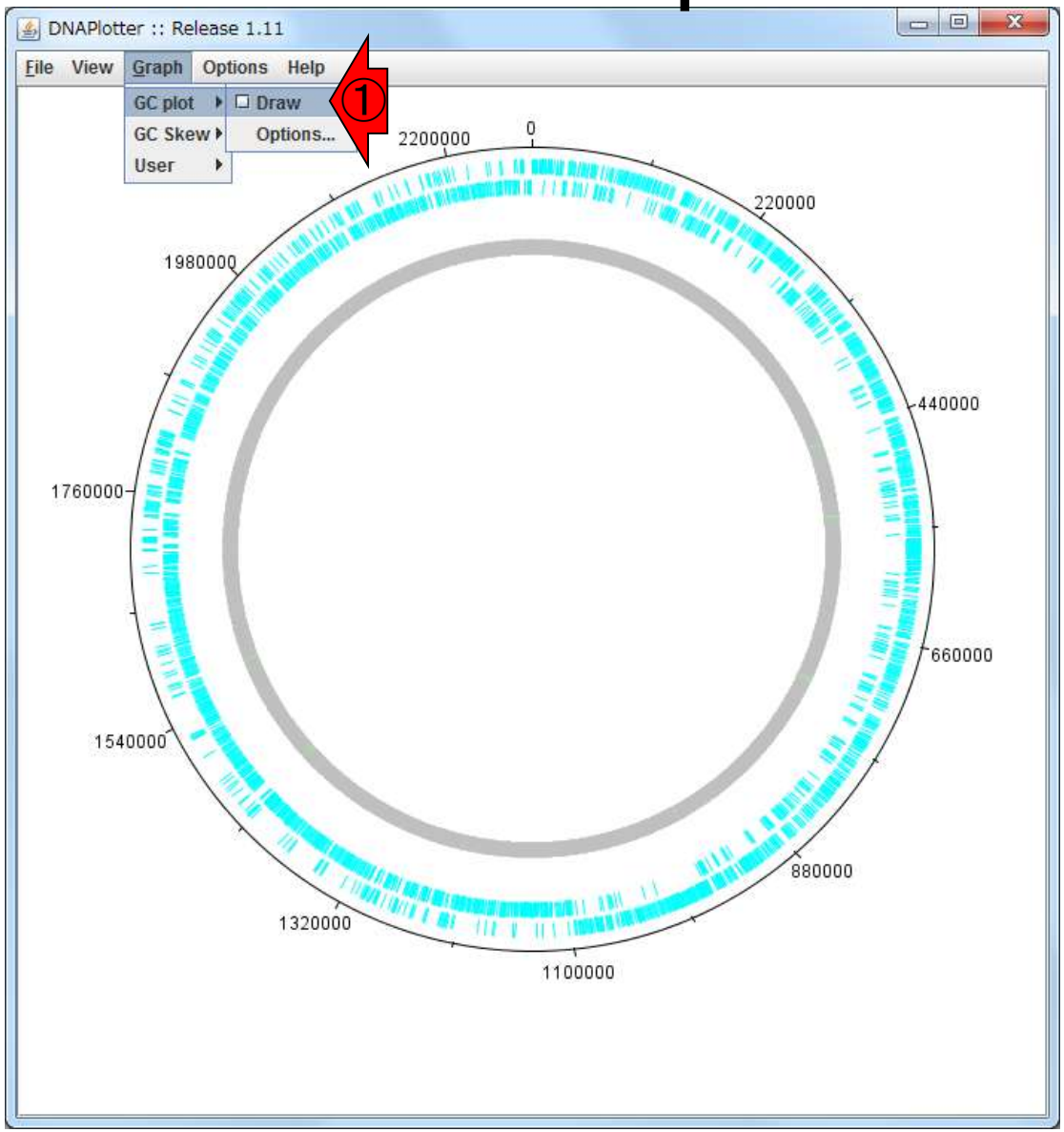
W12-10：ちなみに..

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

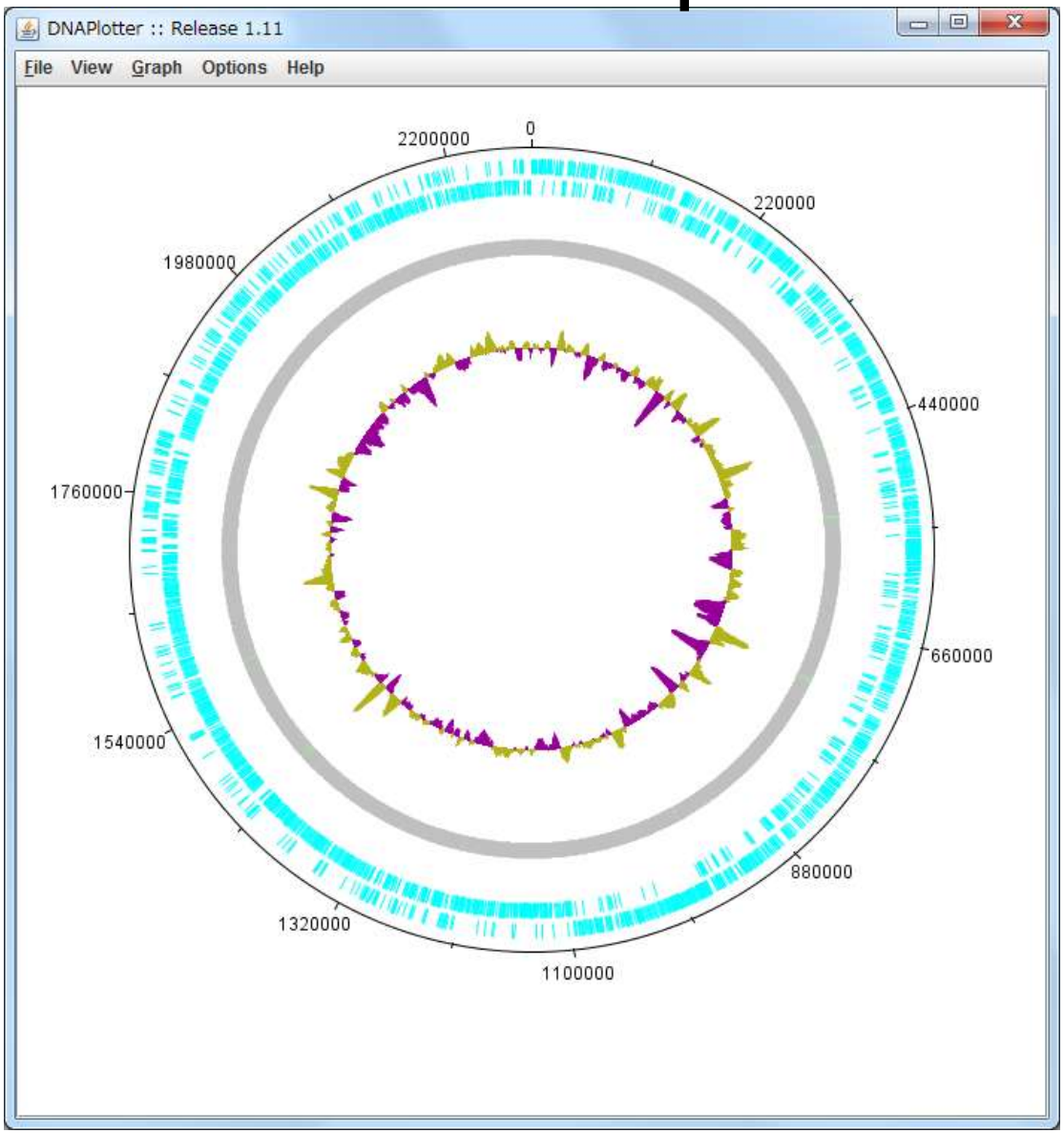


W12-11：GC plot

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

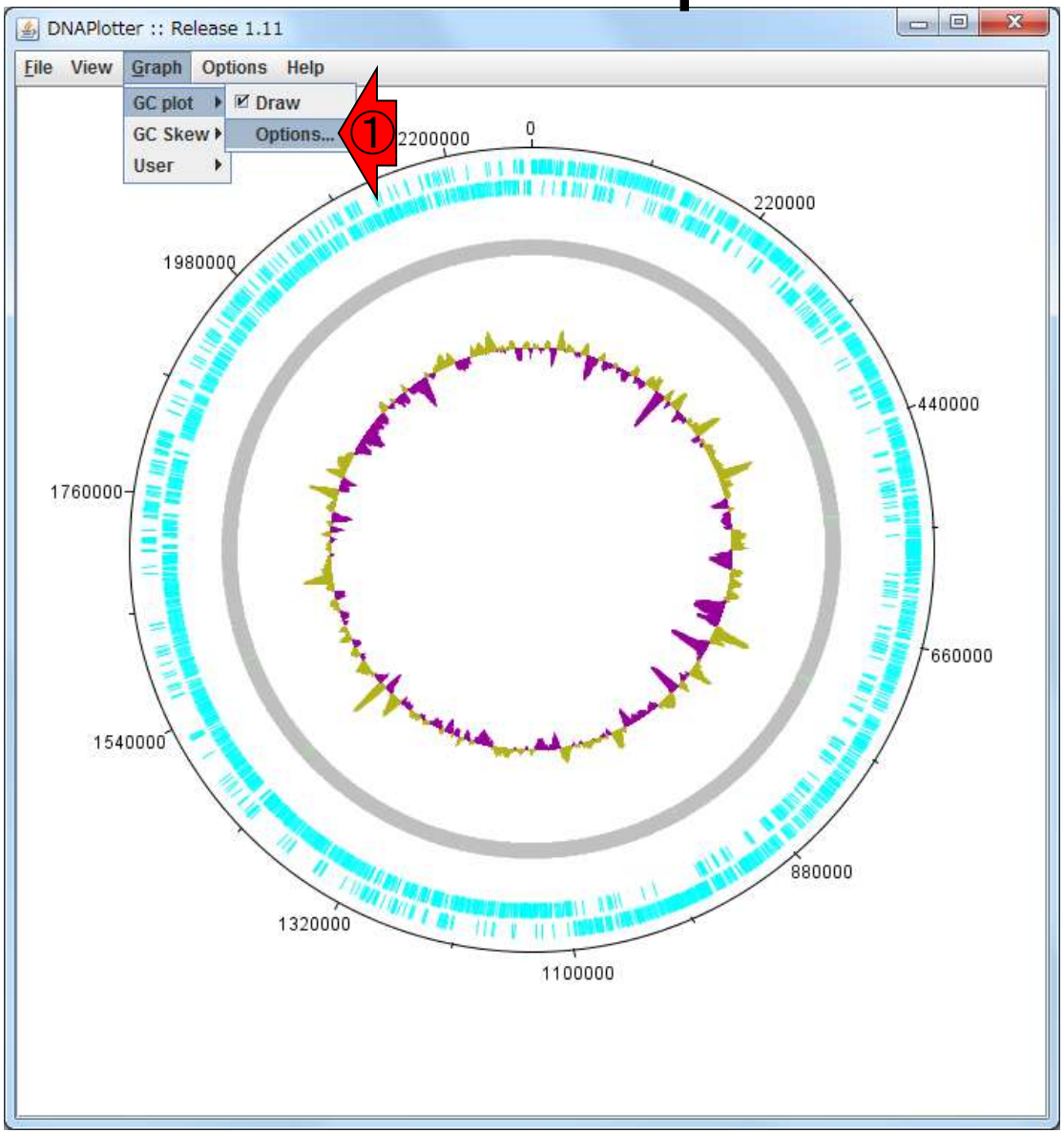


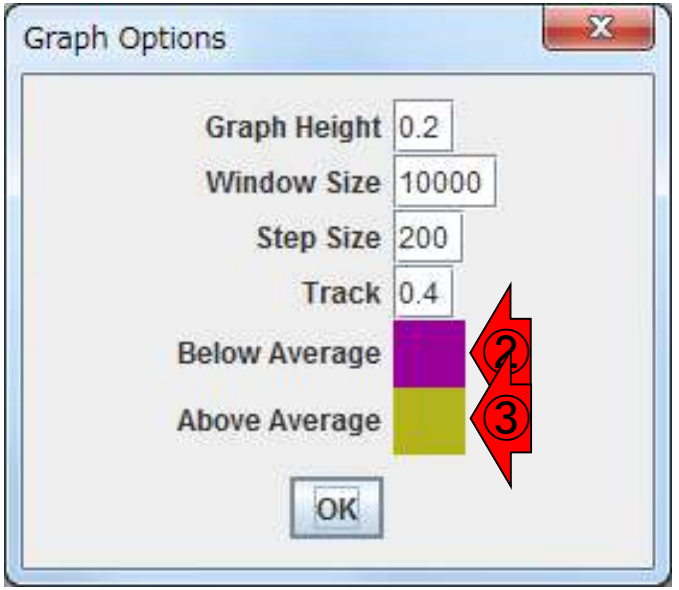
W12-11：GC plot



W12-11：GC plot

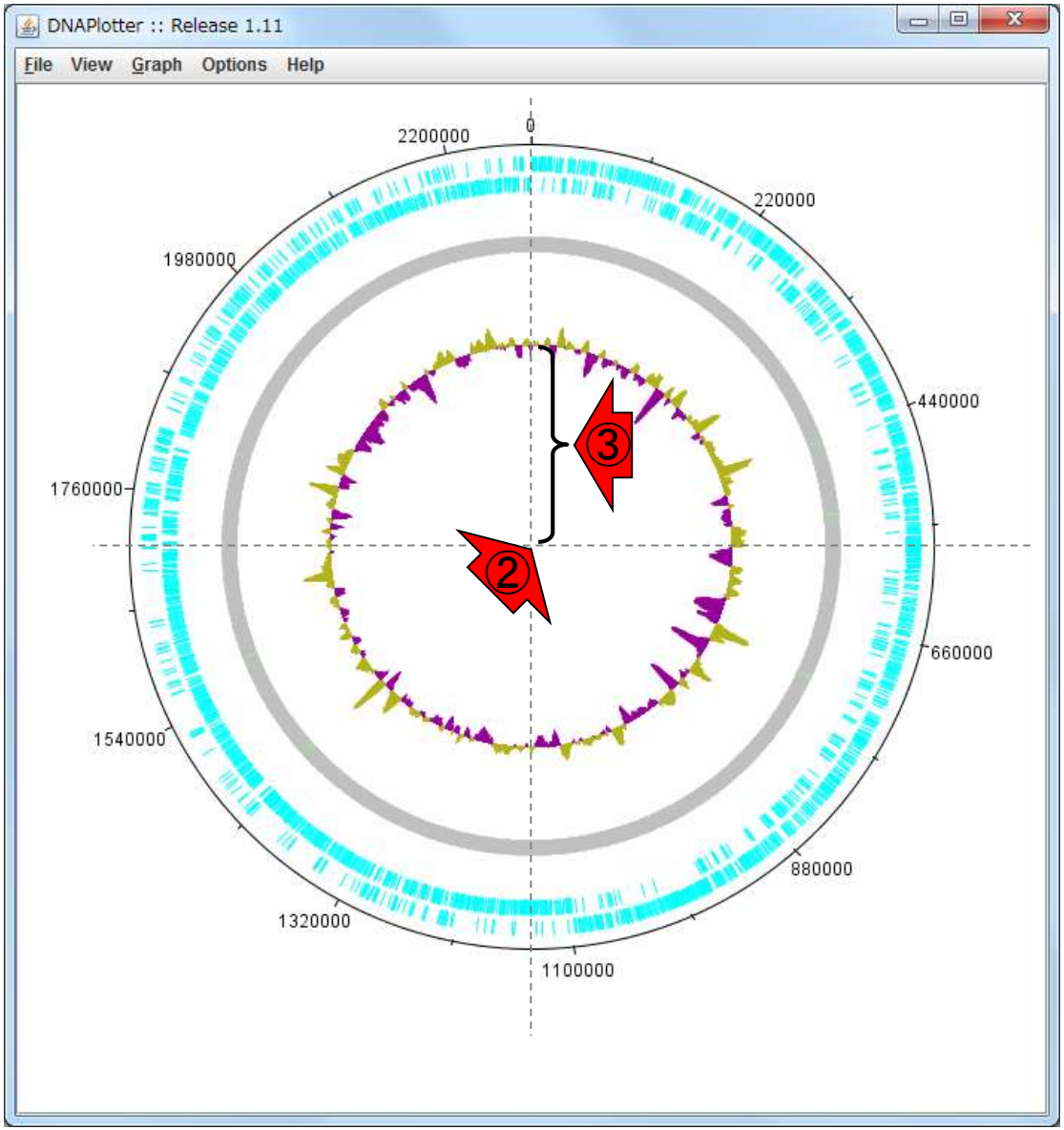
①Graph – GC plot – Options…で、色使いの意味を知ることができます。②GC含量が平均よりも低い領域は赤紫色?!、③平均よりも高い領域は黄土色?!で表されていることがわかります

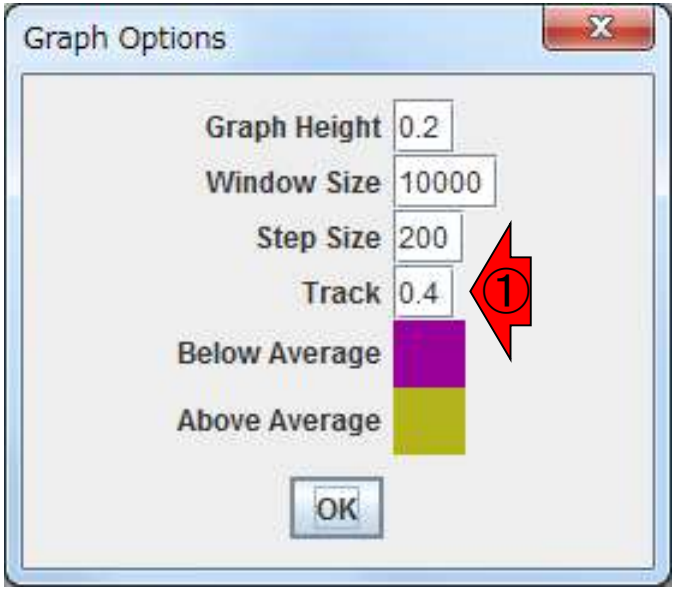




W12-12：Track

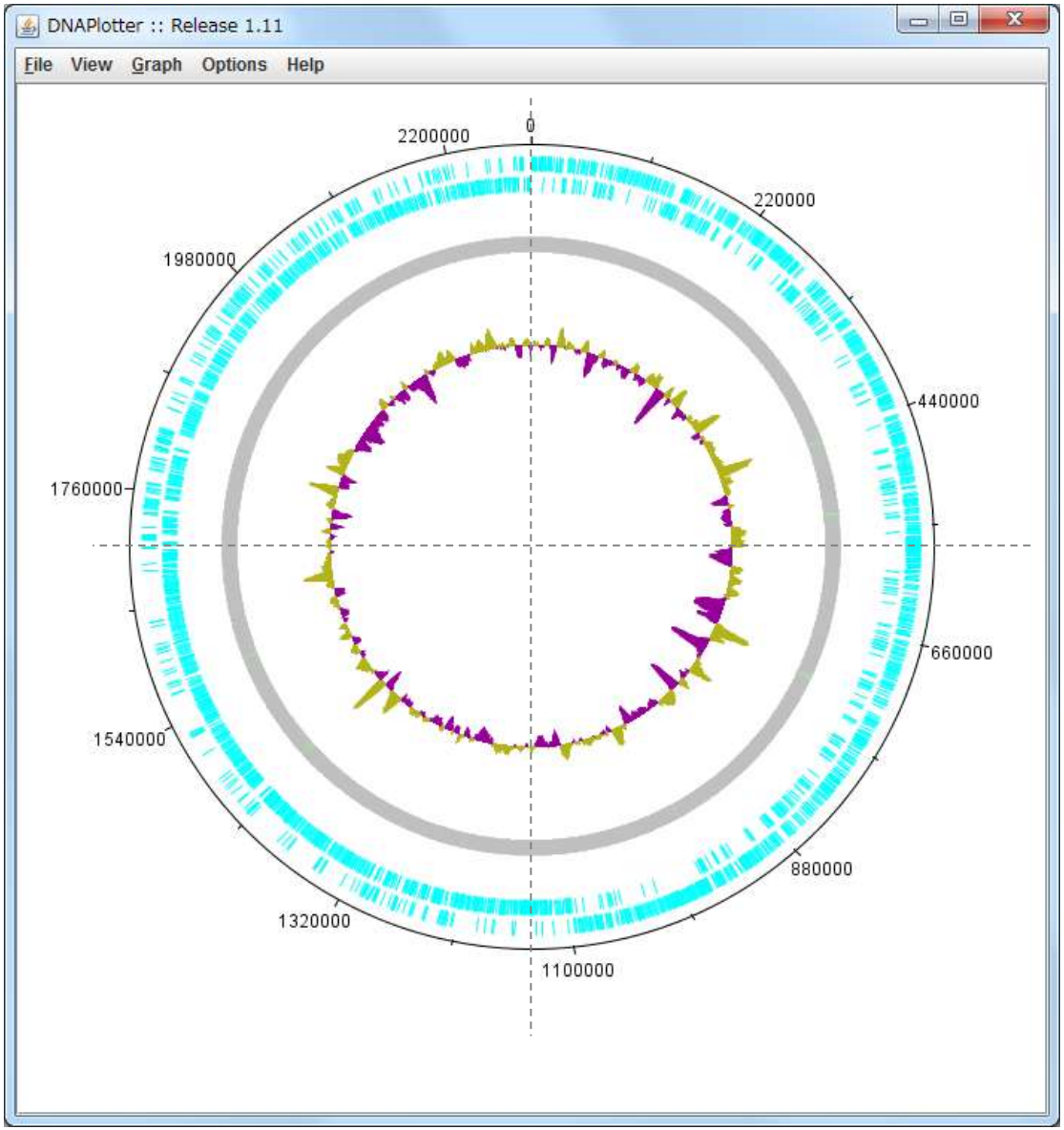
①Track = 0.4というのは、②中心からの③相対的な距離と解釈すればよい

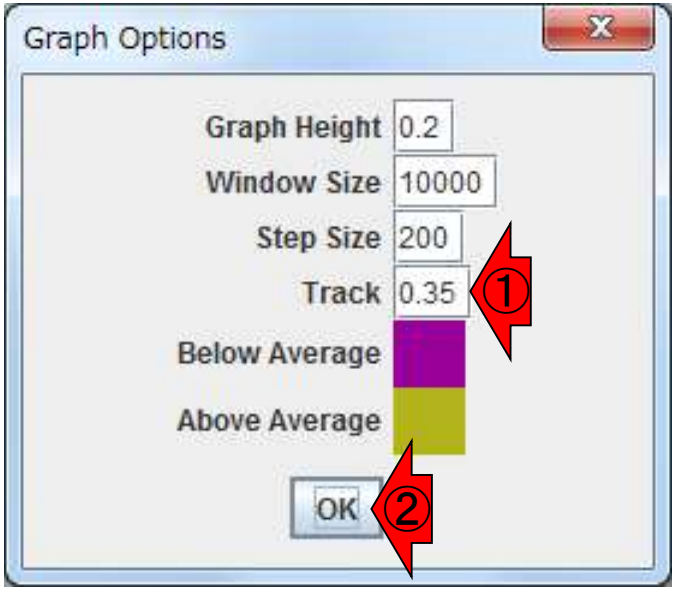




例えば①Track = 0.35として、②OKを押すと…

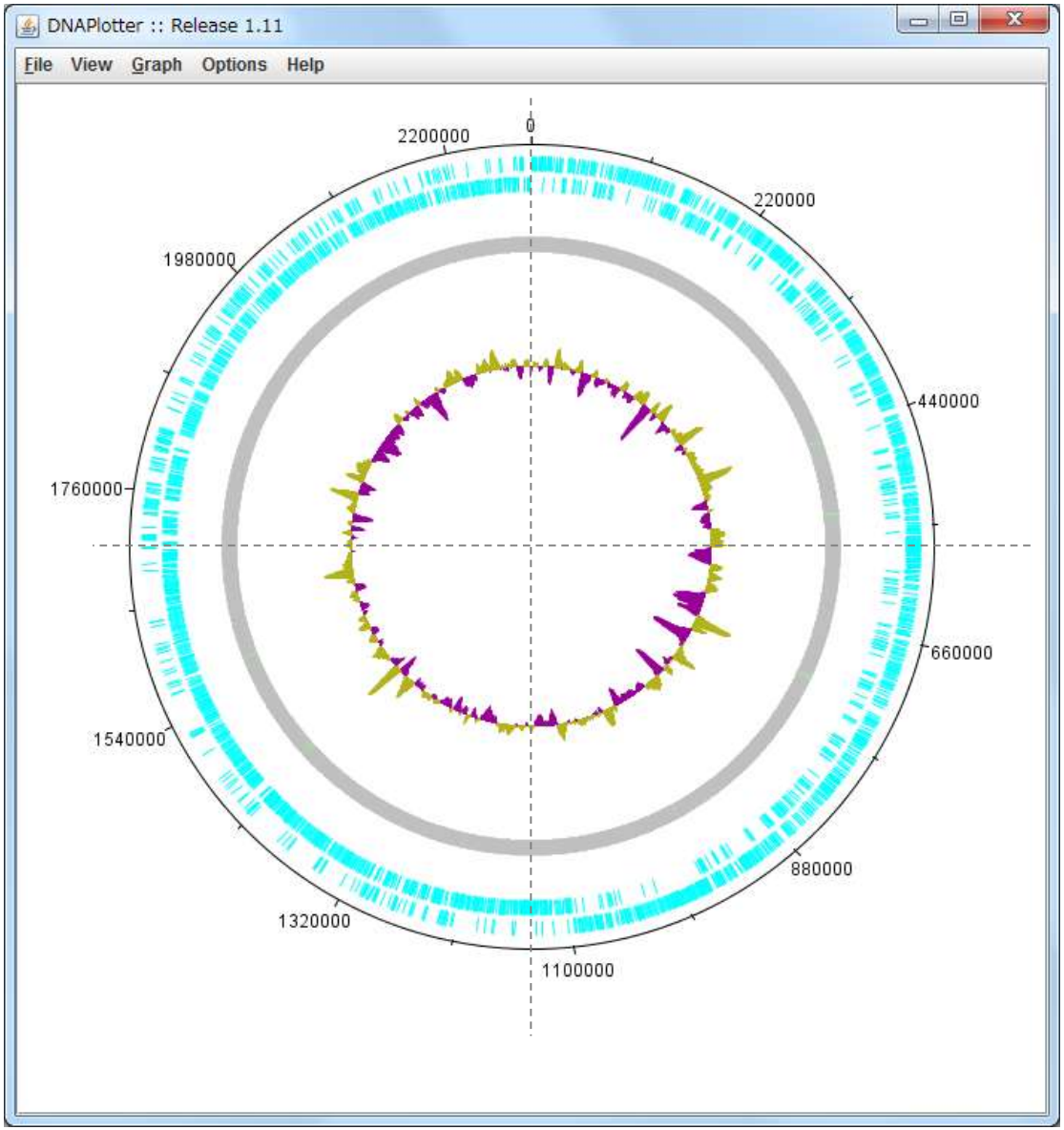
W12-12：Track





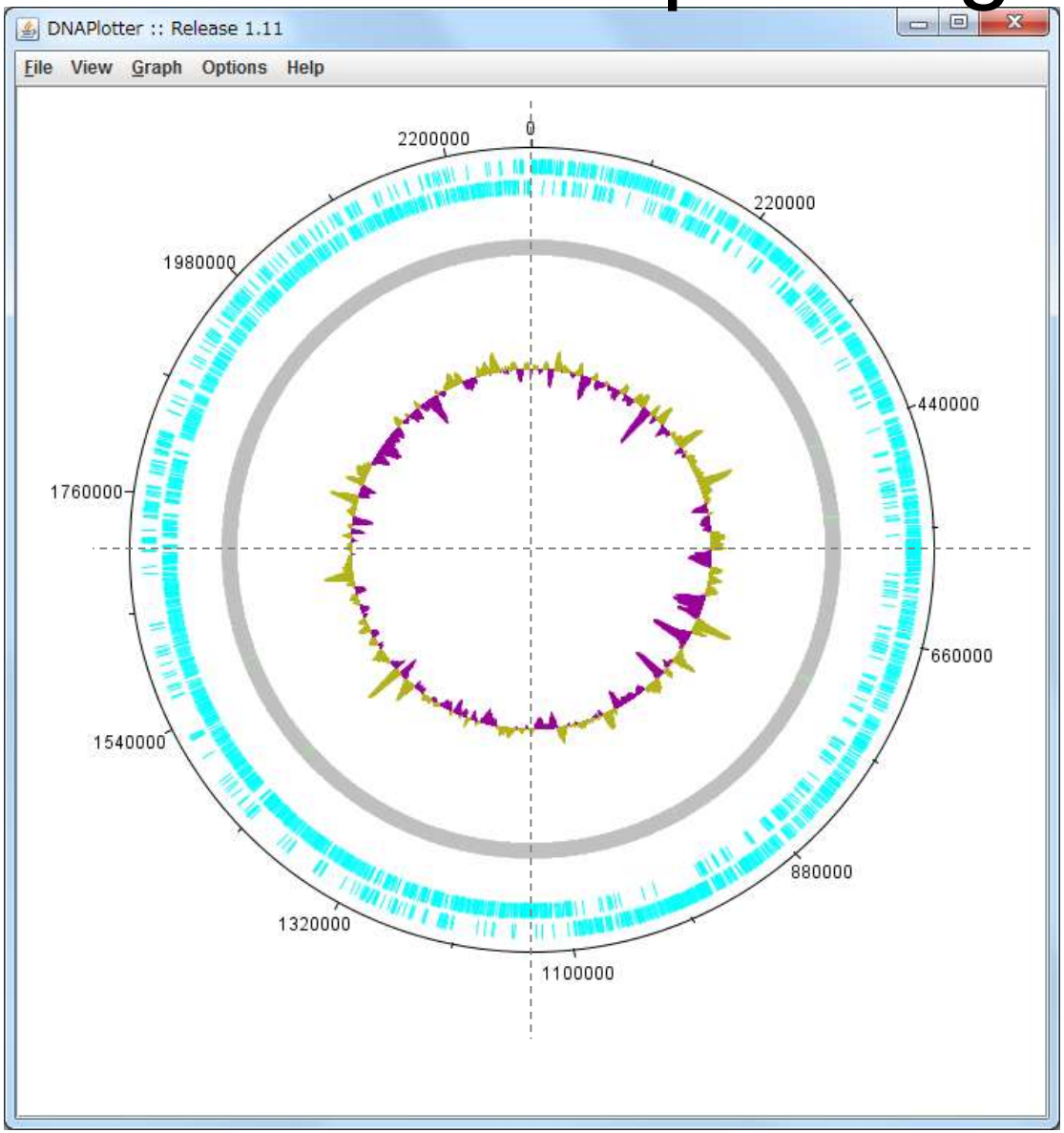
全体として小さくなります

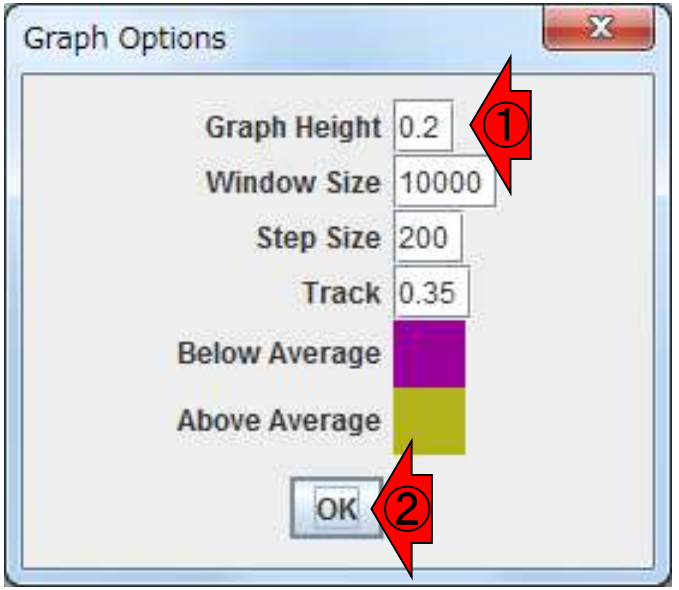
W12-12：Track



W12-13：Graph Height

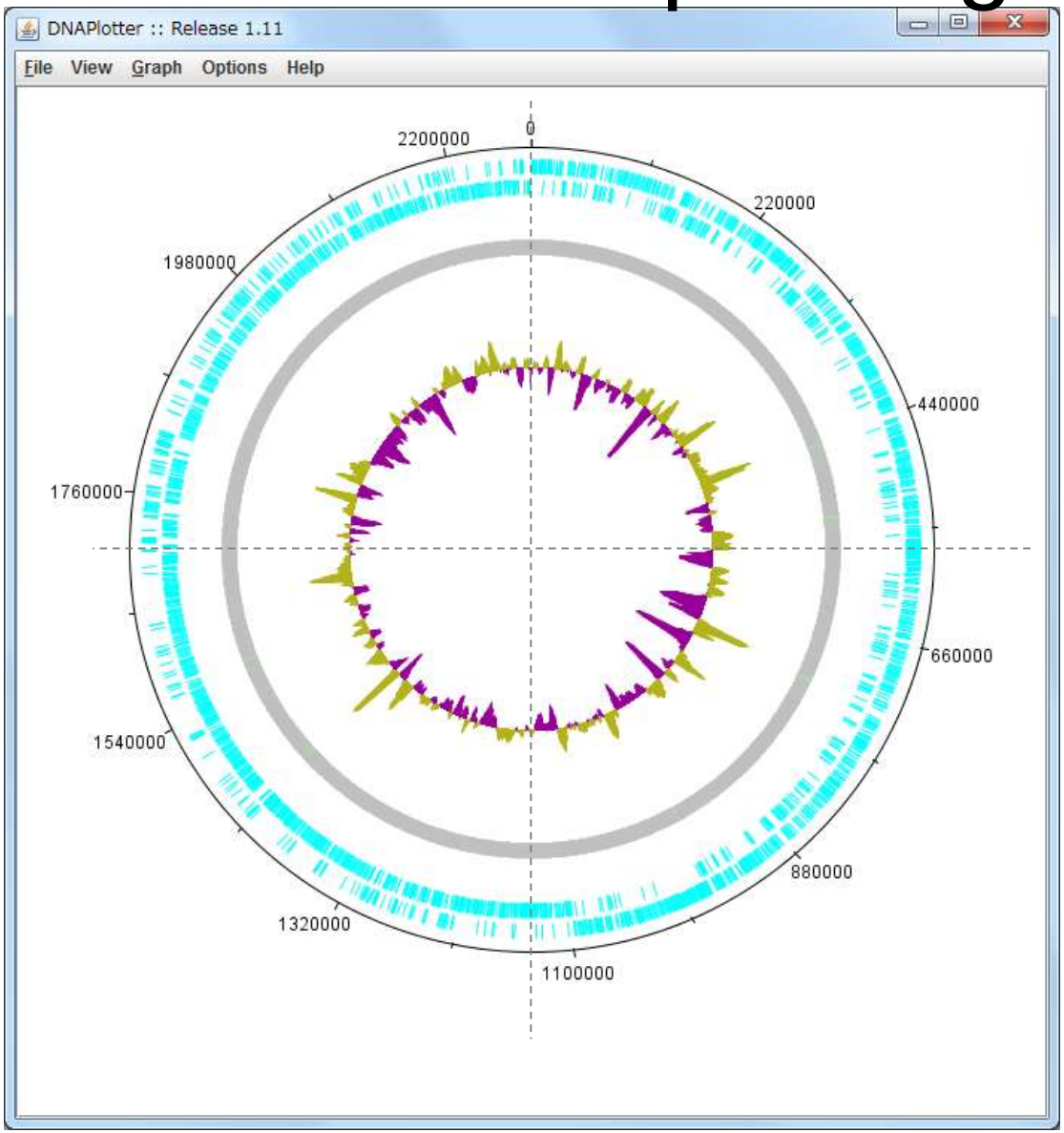
①Graph Heightは高さの調整するところなのでしょう。例えば①Graph Height = 0.3にすると、0.2のときに比べて上下に伸びる感じに見えるのだろうと妄想しながらいじるのです

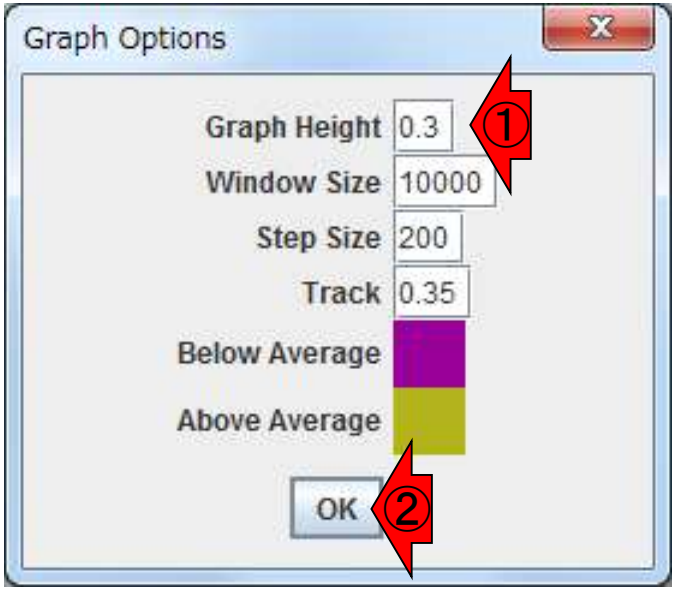




①Graph Height = 0.3にした結果

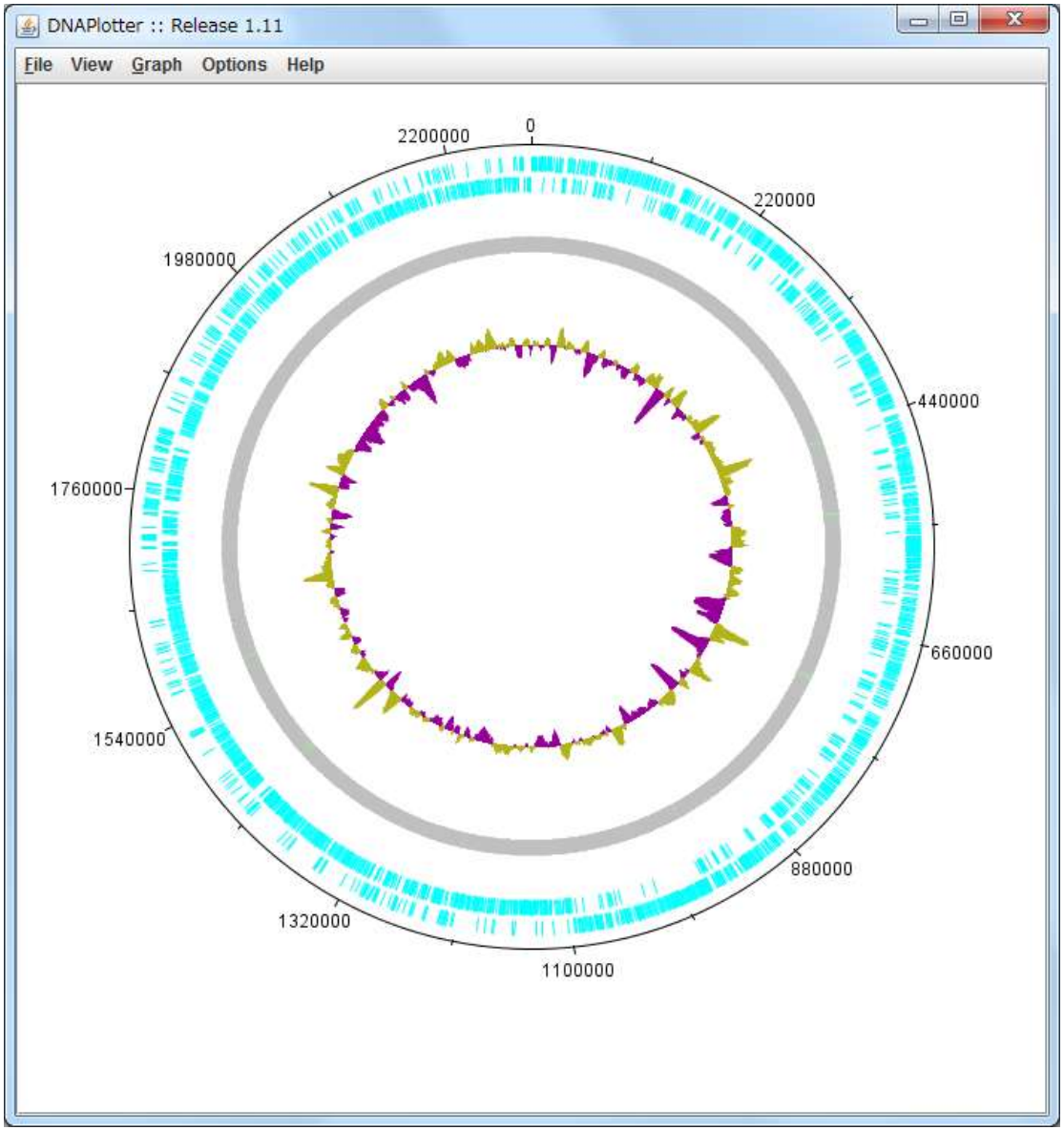
W12-13：Graph Height

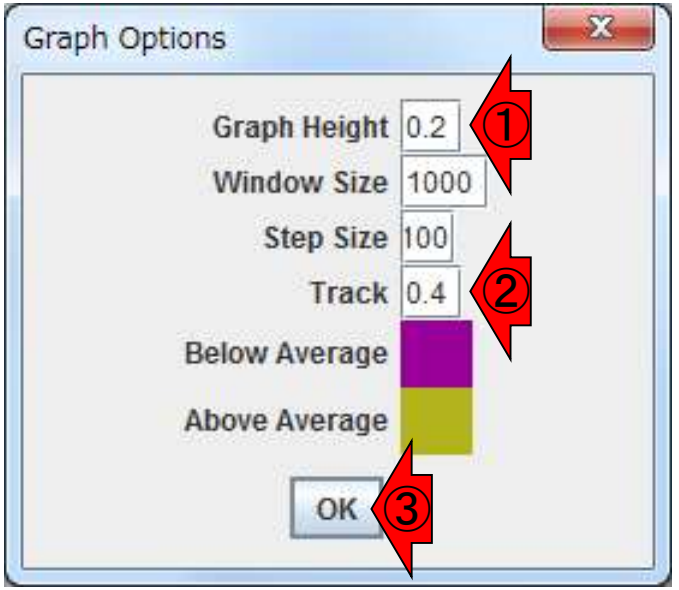




W12-14：デフォルトに戻す

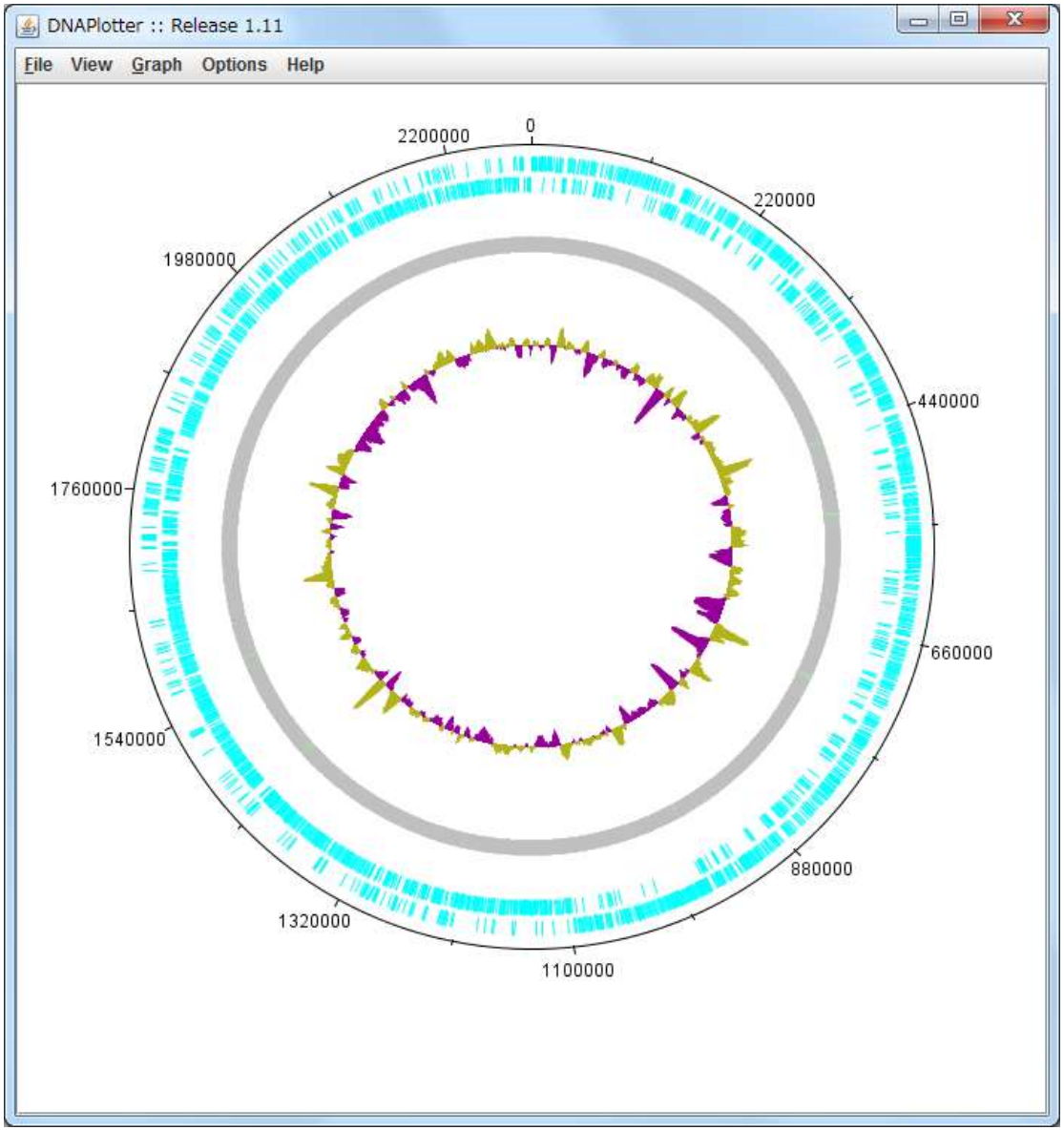
一旦デフォルトの①0.2と②0.4に戻して、③OK

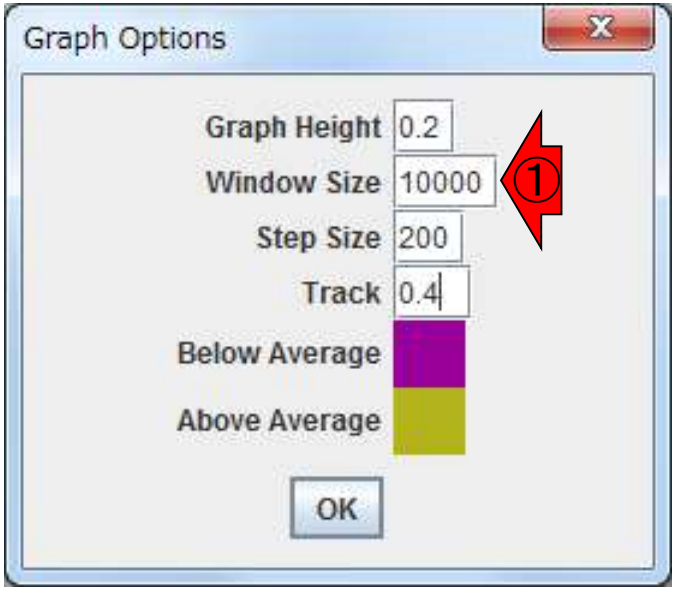




W12-15：Window Size

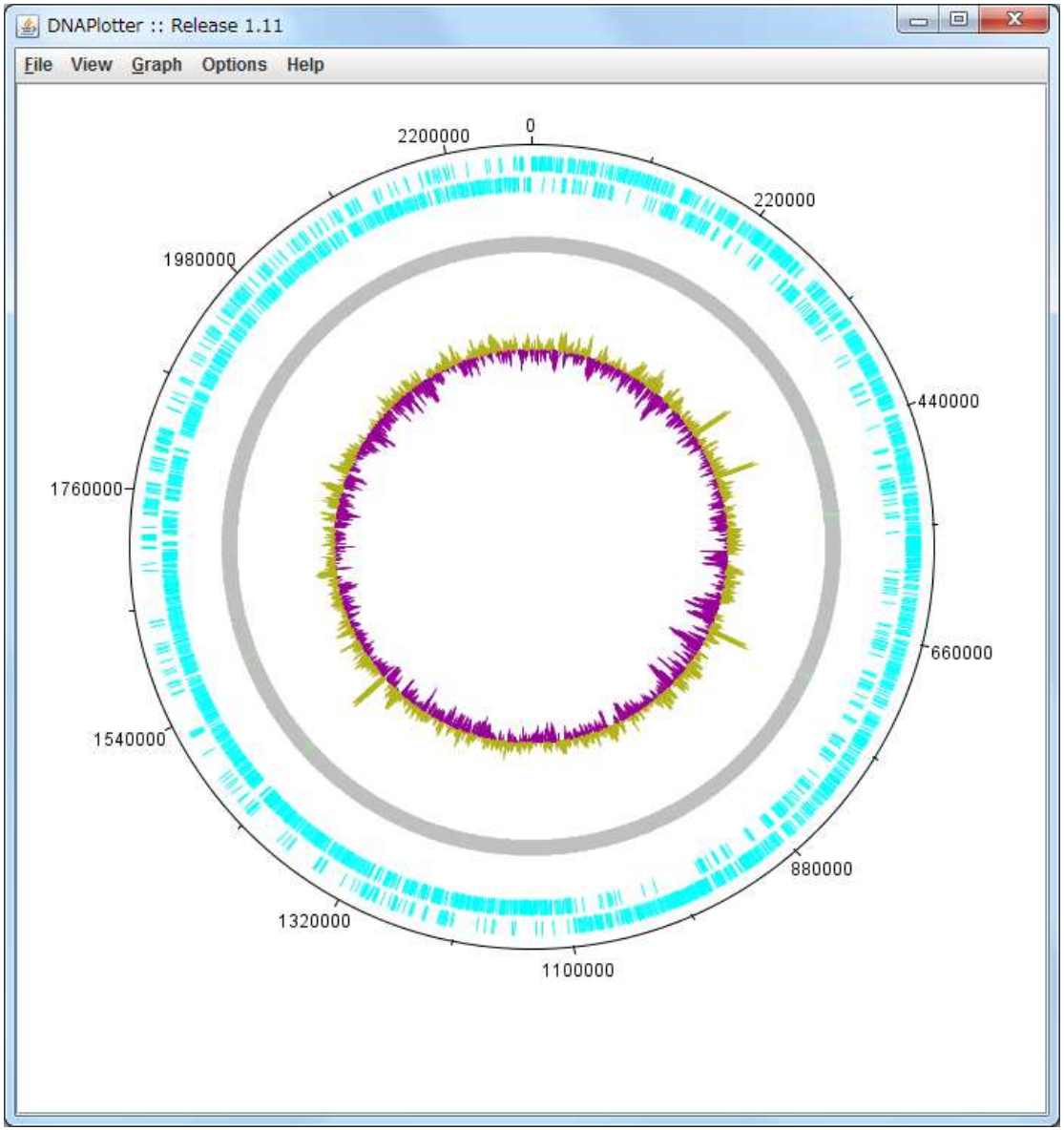
GC plotの描画目的は、どの領域がGC含量が高いかな～など、GC含量分布の全体像を把握すること。①Window Sizeは、Window（つまり幅）を10000塩基とって、その範囲の平均のGC含量を計算せよという指定

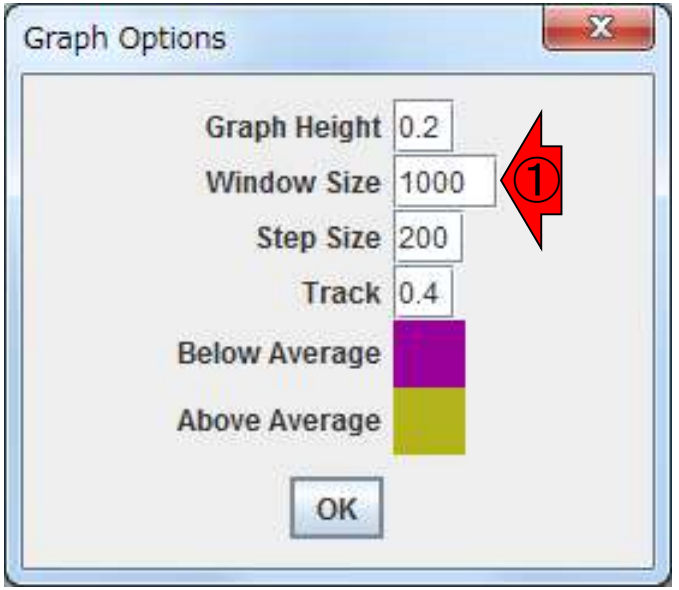




W12-15：Window Size

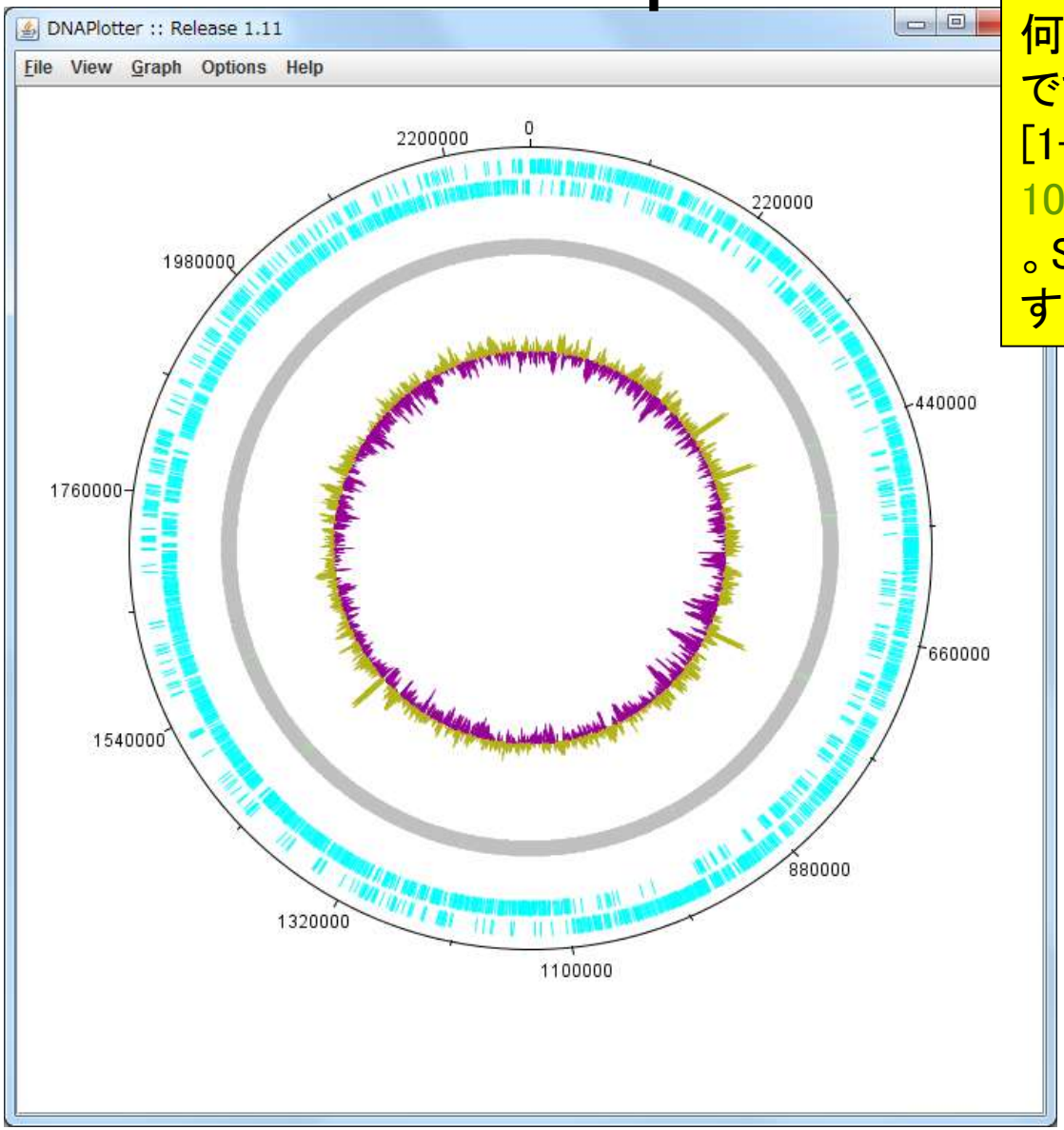
なので、例えば①Window Size = 1000にすると、その範囲が1/10になるので、得られるプロットの解像度が上がります。つまりなだらかさの度合いが減少する、という意味です

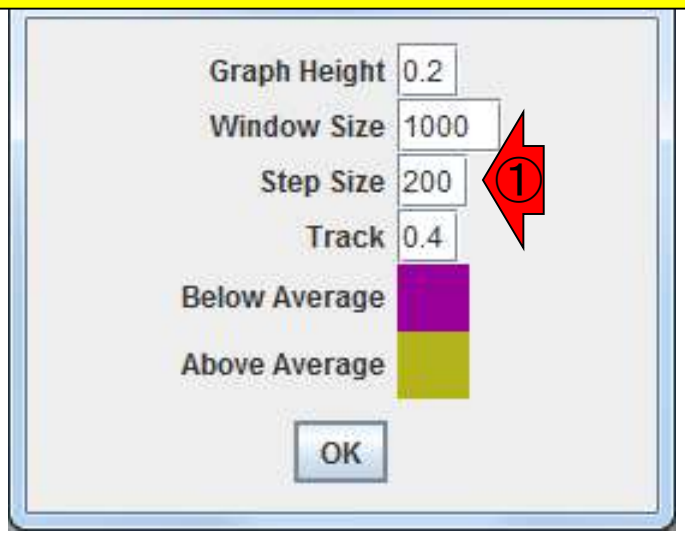




W12-16：Step Size

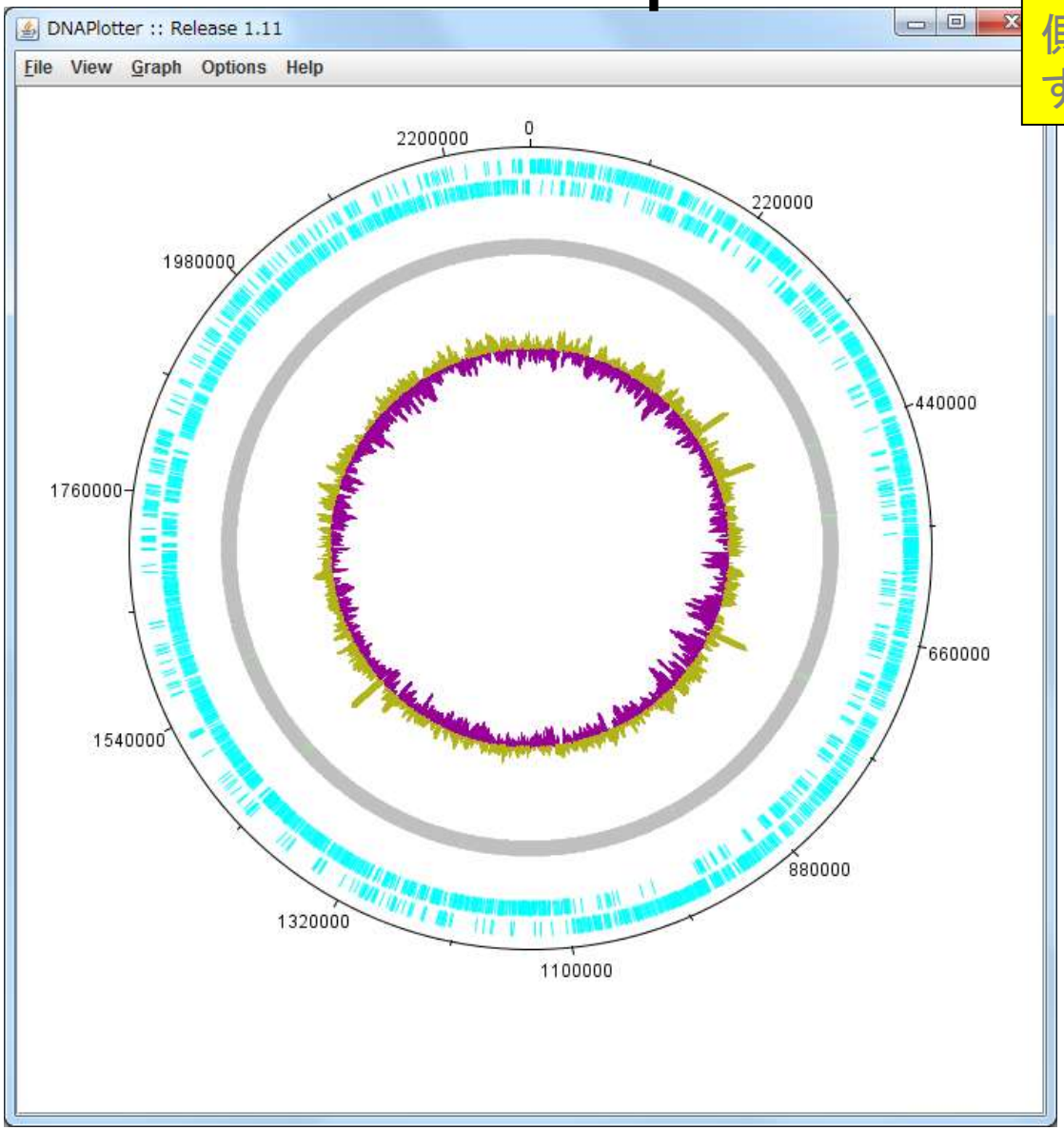
Step Sizeは、どれだけ計算をさぼるか、といった意味合いのパラメータです。具体的には、例えば最初に領域[1, 1000]ででGC含量を計算し、次に何塩基分とばした位置から計算するかということです。Step Size = 200だと、次に計算する領域が[1+200, 1000+200]、そしてその次が[1+200+200, 1000+200+200]の領域を計算するという意味です。Step Sizeを小さくするほど正確な計算になりますが、その分だけ時間がかかります

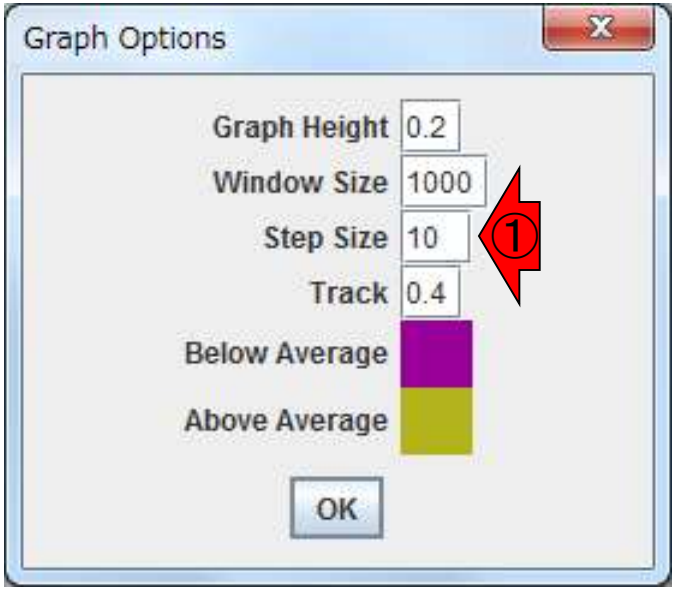




W12-16：Step Size

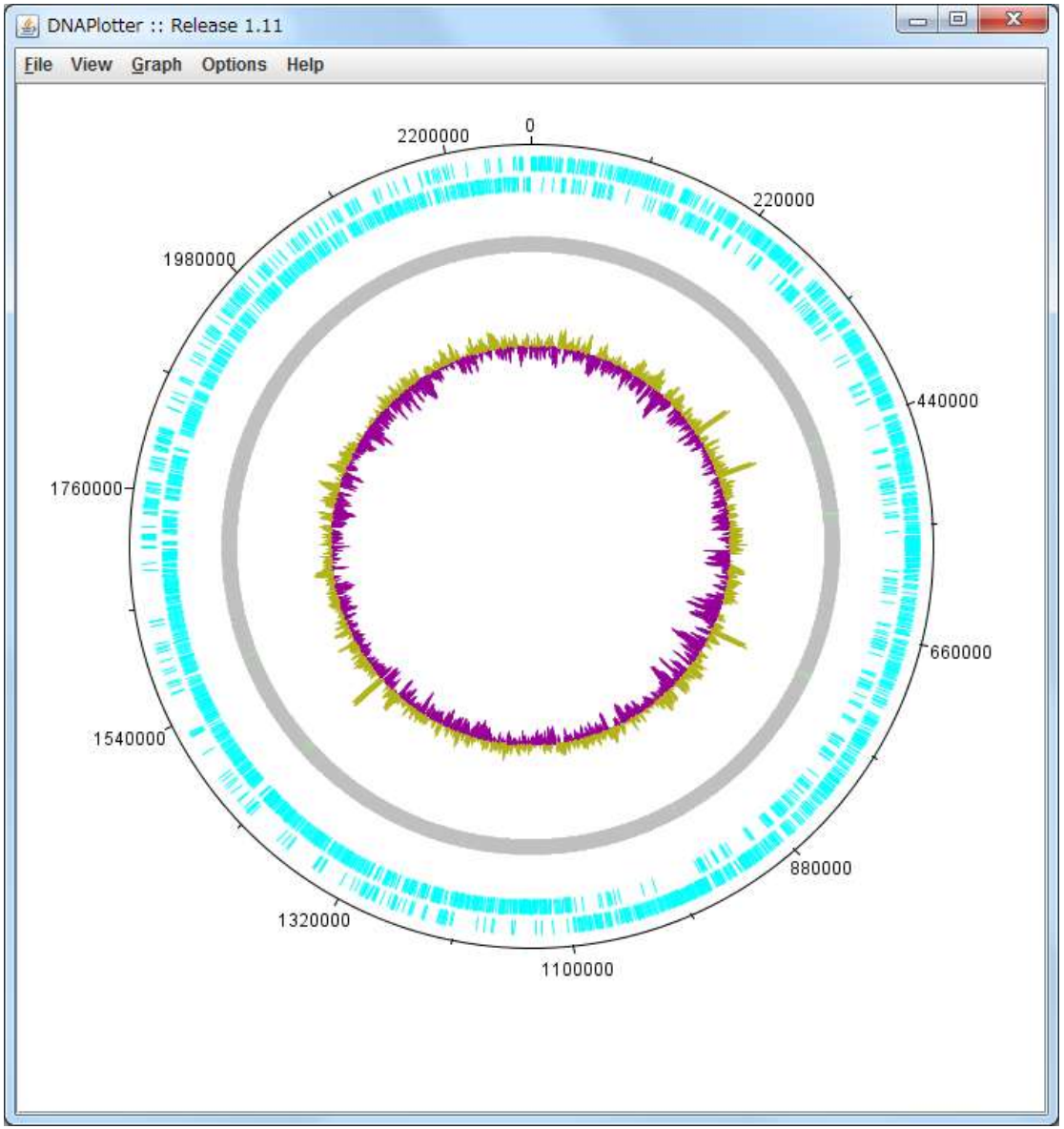
例えばStep Sizeを200から①10に変更するということは、単純に計算時間が200/10 = 20倍かかることを意味します。このあたりはプログラム側のデフォルトに任せるのが一般的だと思いますが、気になるヒト用に解説したまでです

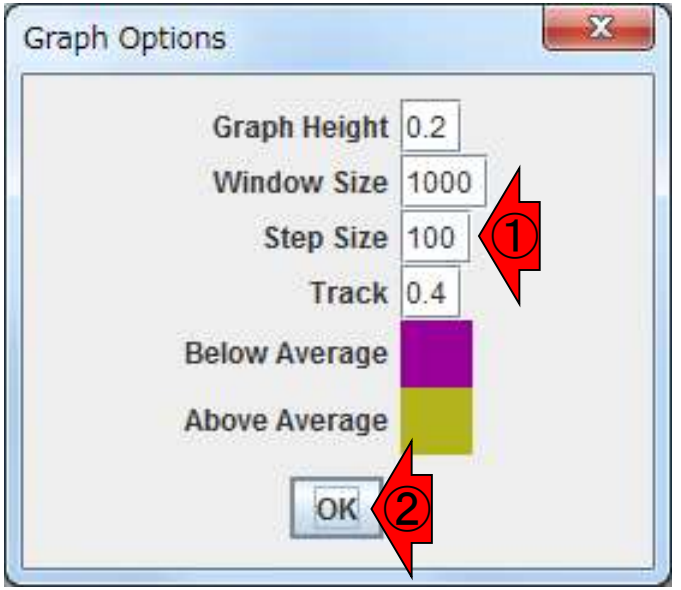




W12-17：デフォルトに戻す

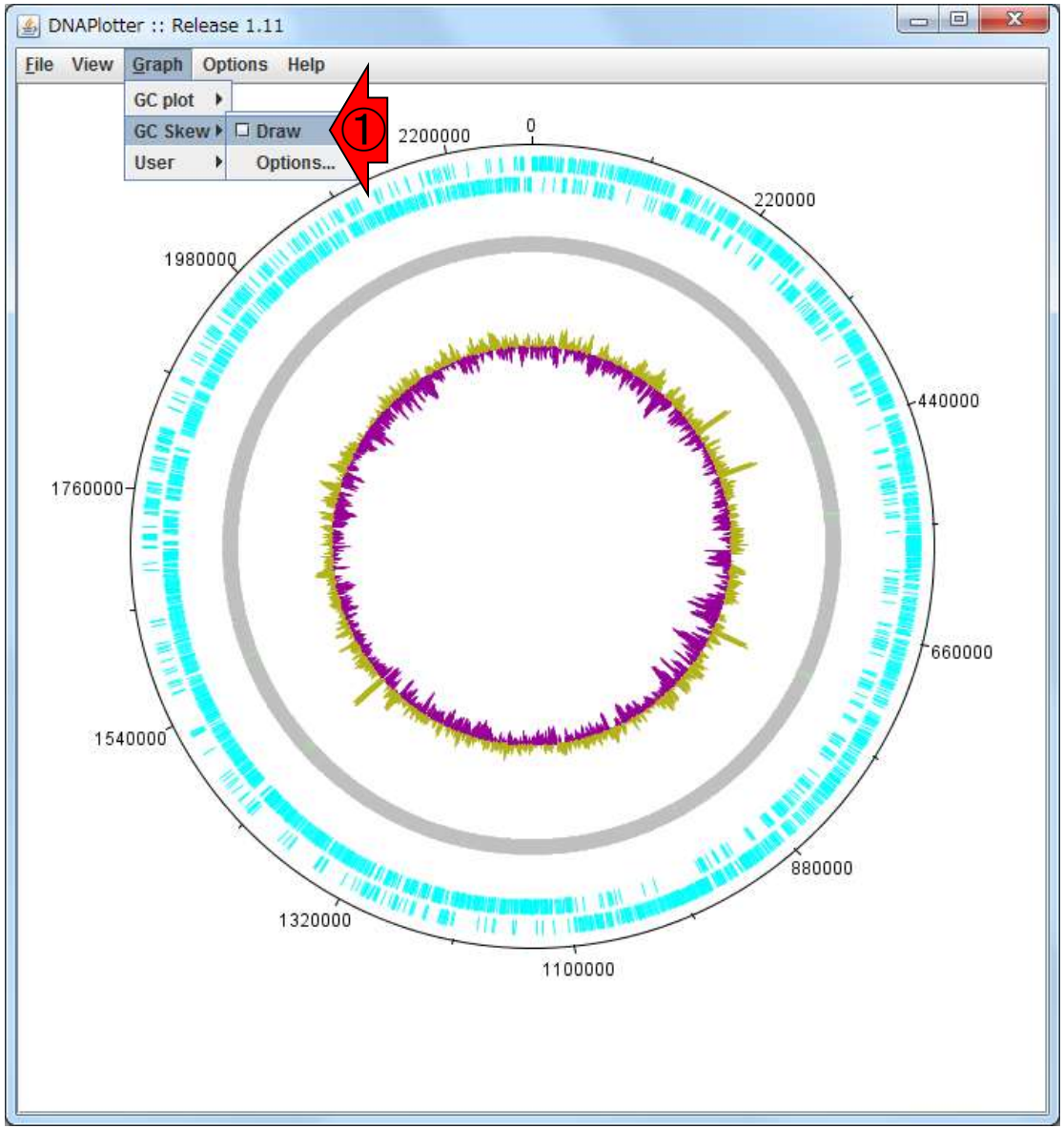
一旦デフォルトの①100に戻して、②OK



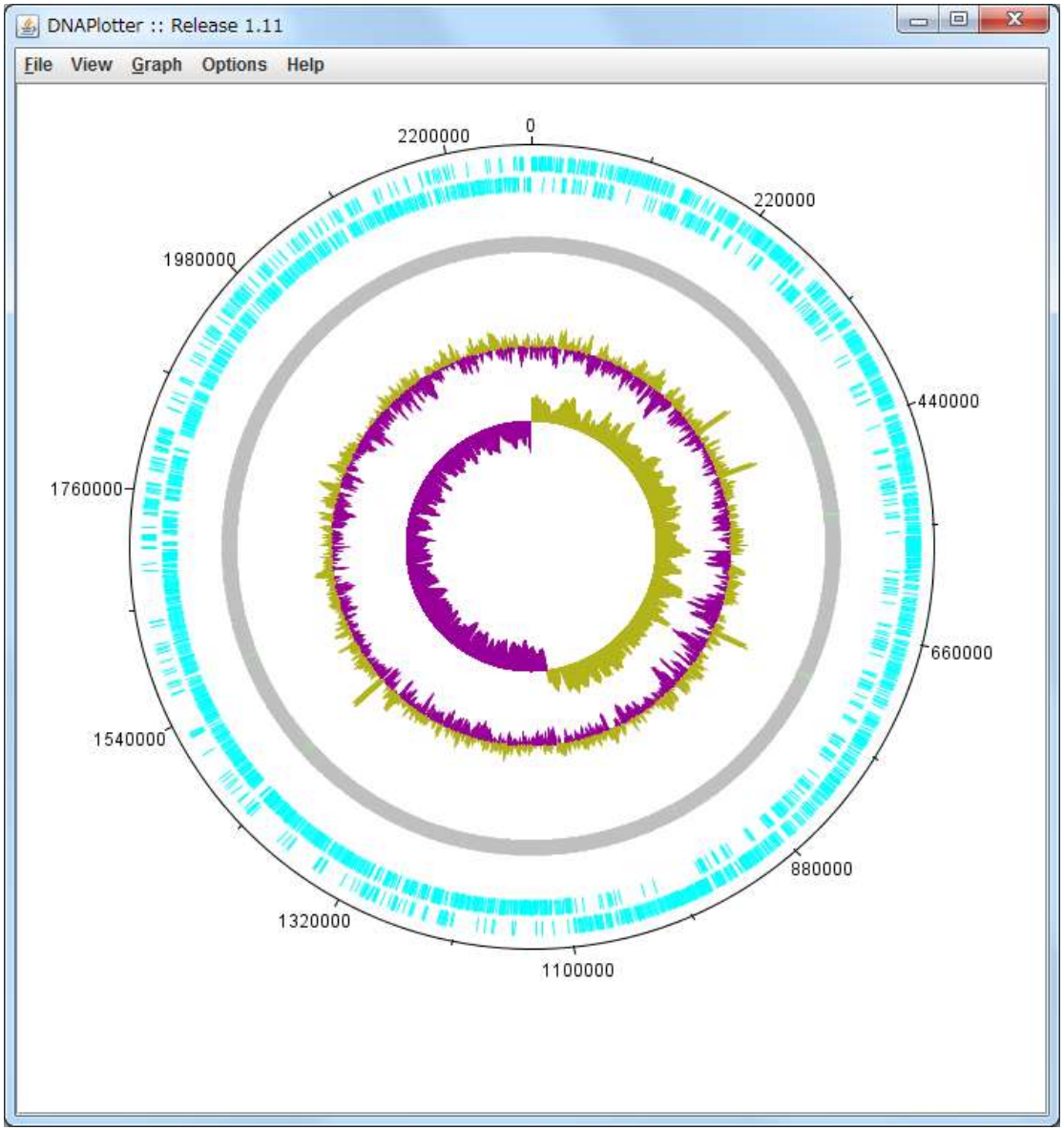


W12-18：GC skew

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

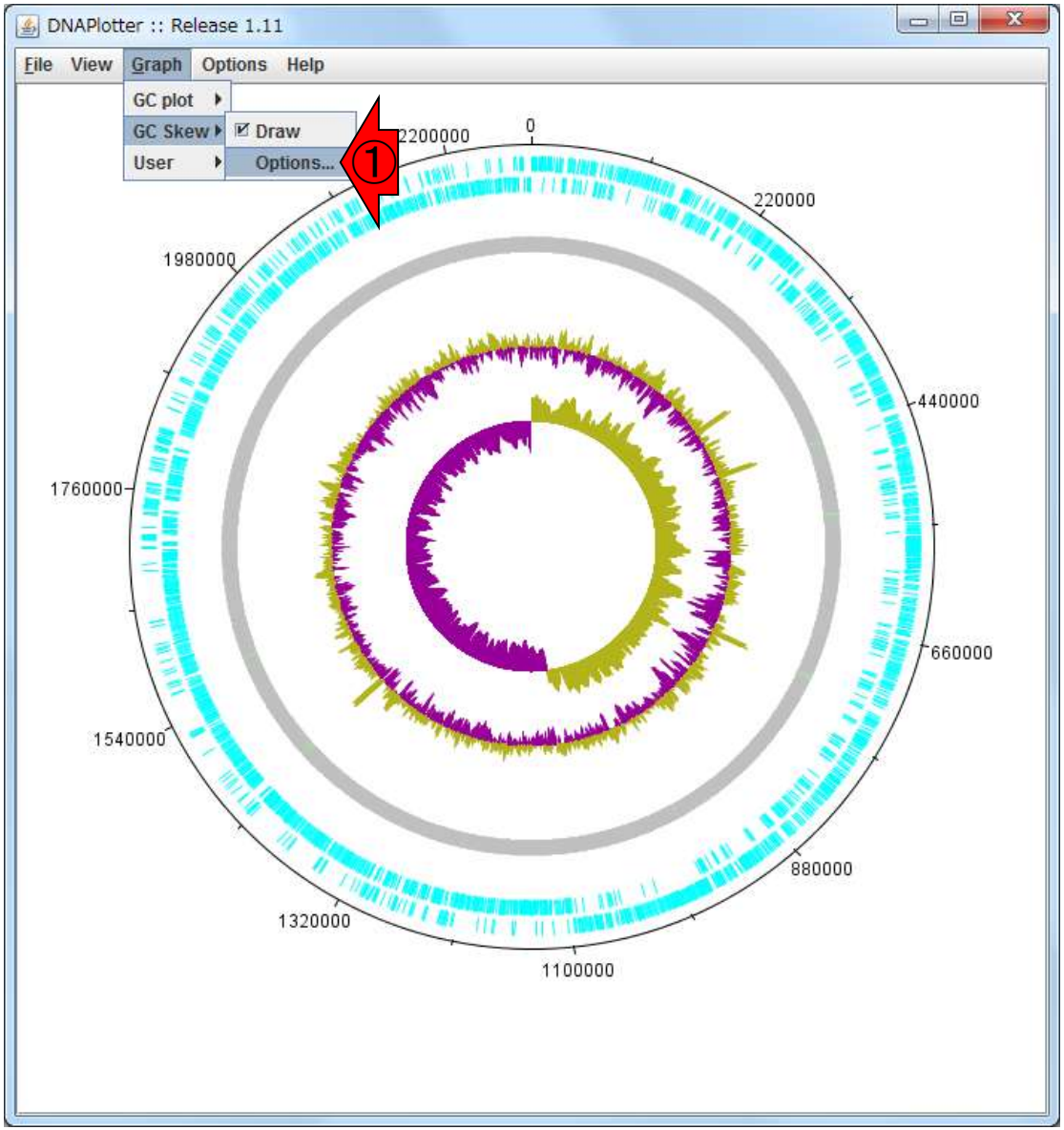


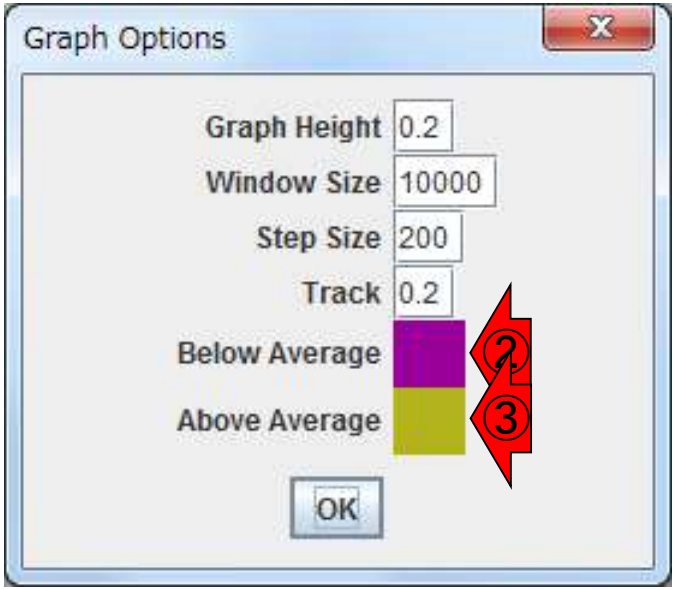
W12-18：GC skew



W12-18：GC skew

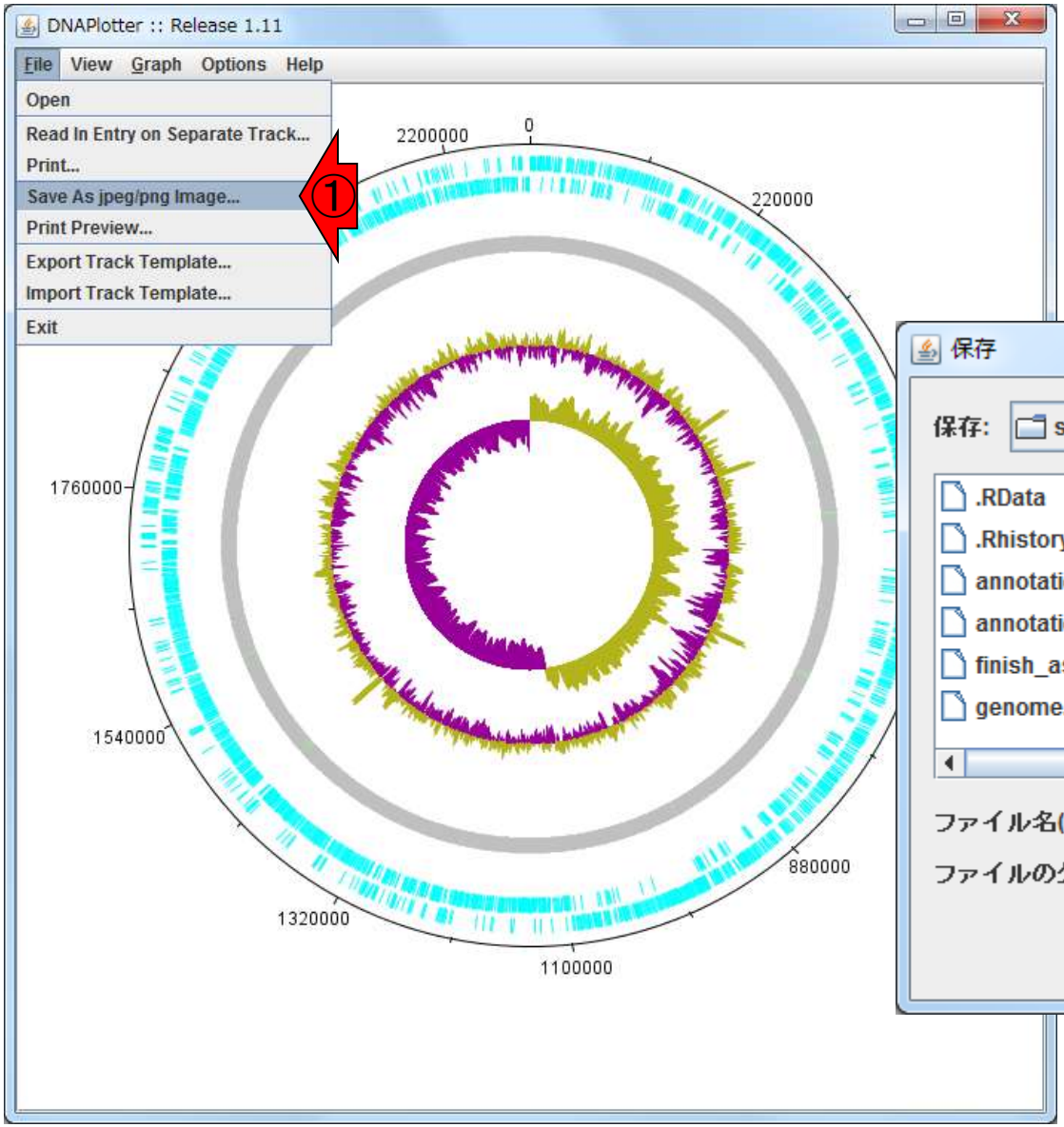
①Graph – GC skew – Options…で、色使いの意味を知ることができます。②GC skewが平均よりも低い領域は赤紫色?!、③平均よりも高い領域は黄土色?!で表されていることがわかります

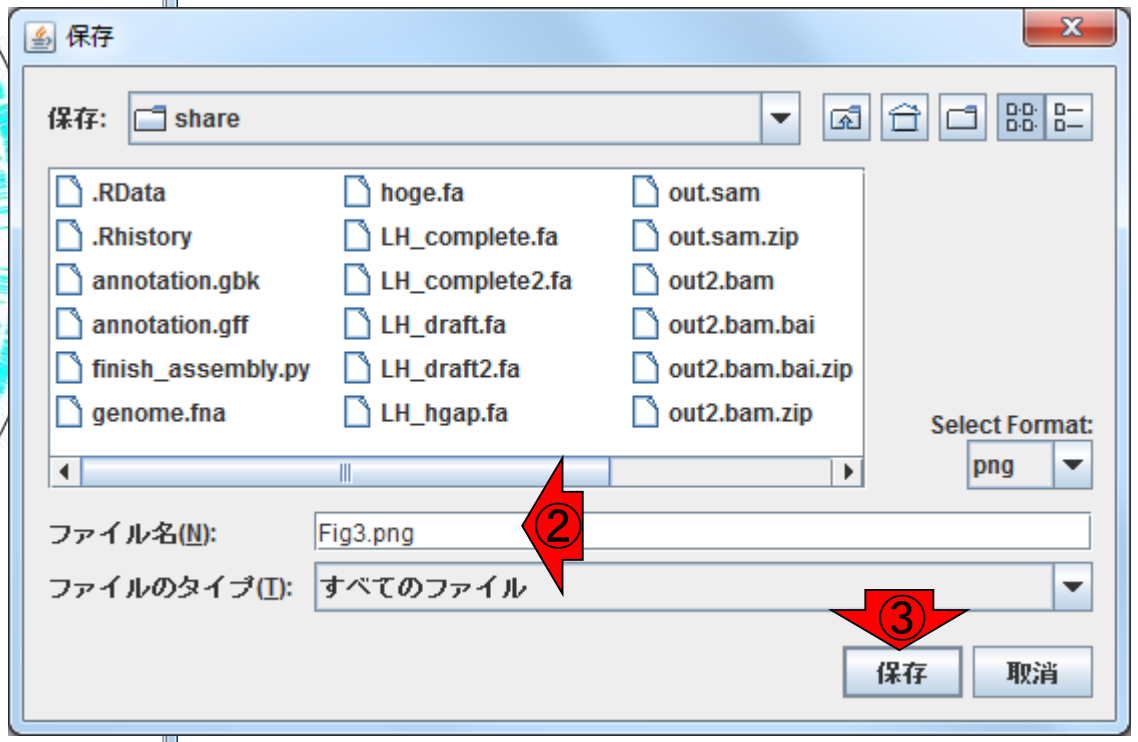




W12-19：保存

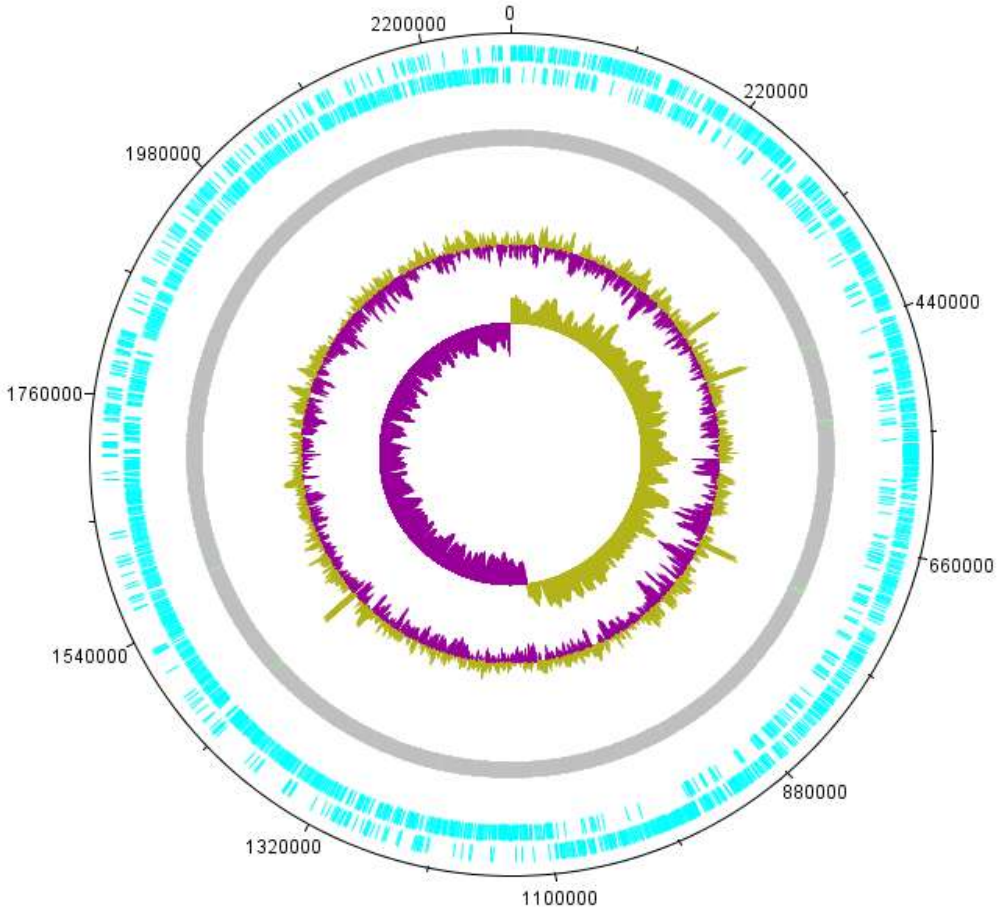
①File – Save As jpeg/png Image...で保存できます。ここでは②Fig3.pngとして保存しました。原稿中のFig. 3と同じものです



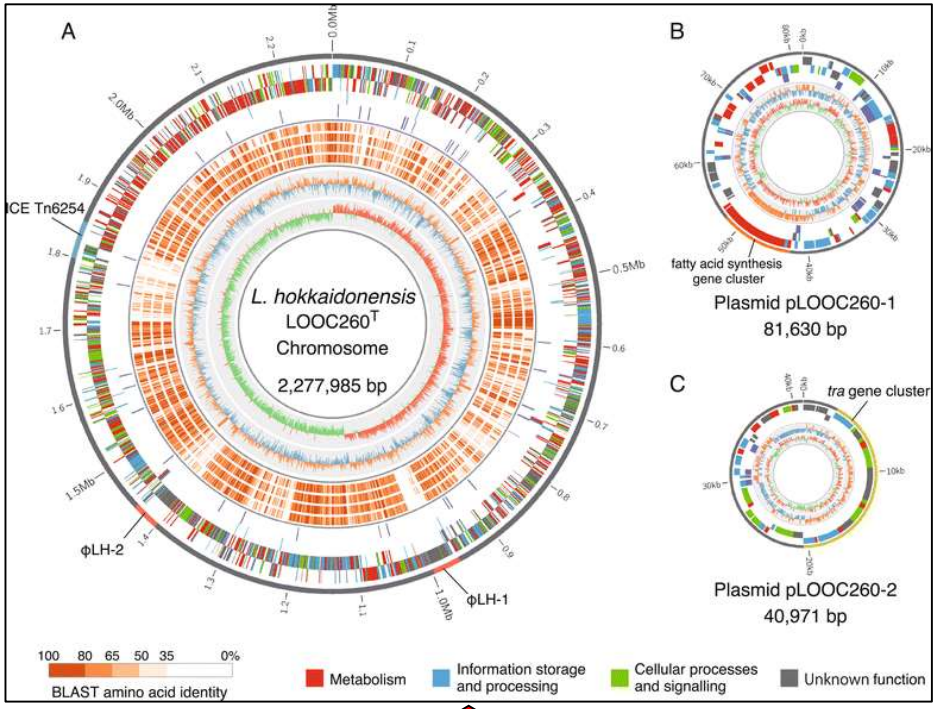


W12-20：Fig. 3

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



①

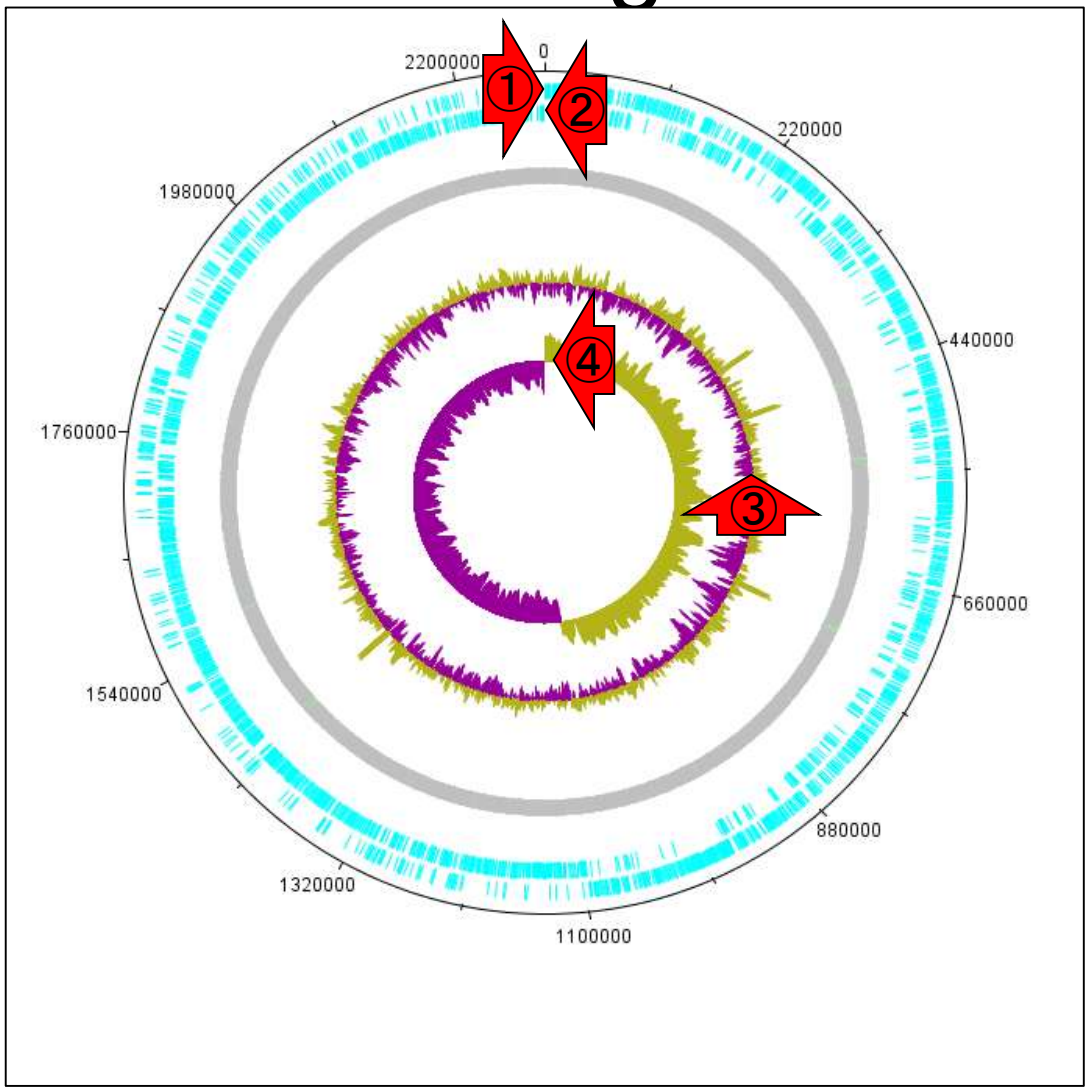


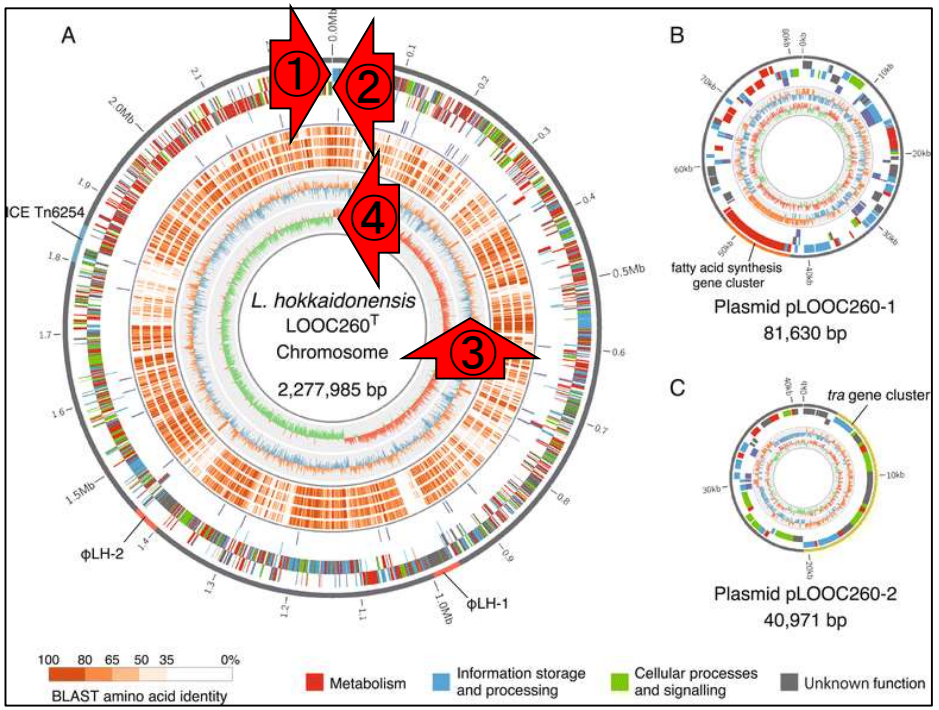
②

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

W12-20：Fig. 3

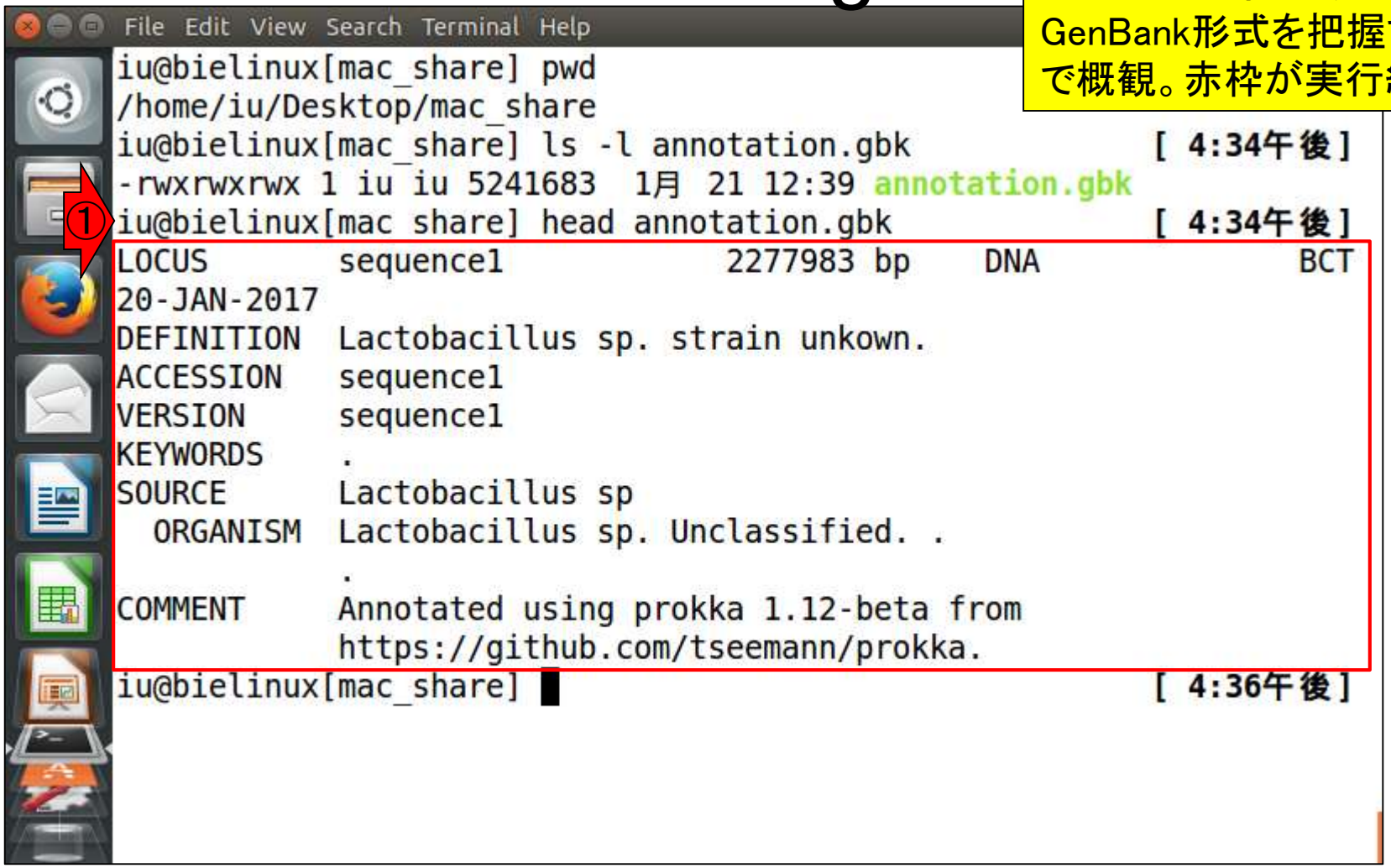
具体的には、①②リーディング鎖上に多くのCDSが存在する傾向、③GC plot、④GC skewの分布などです





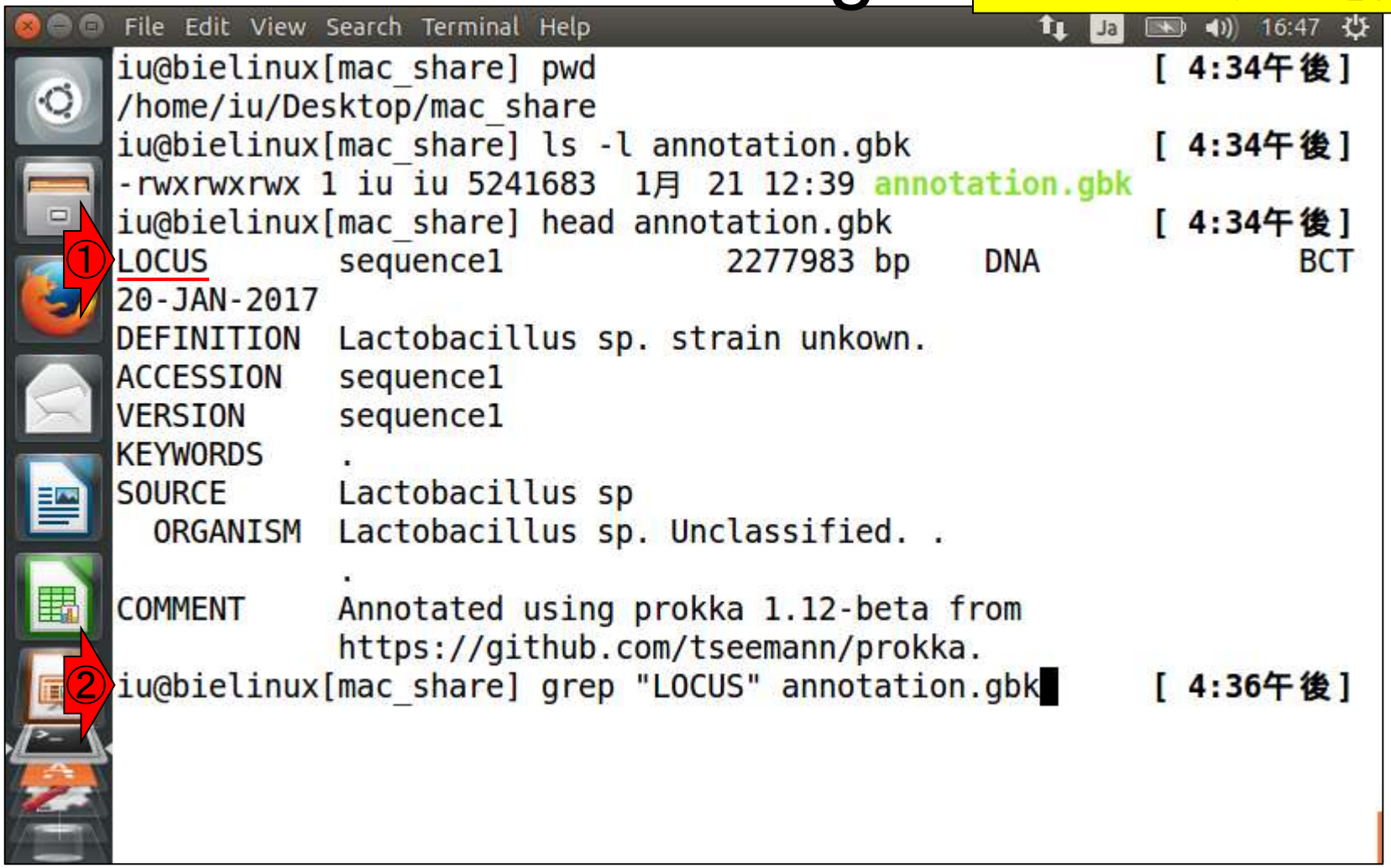
W13-1：annotation.gbk

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



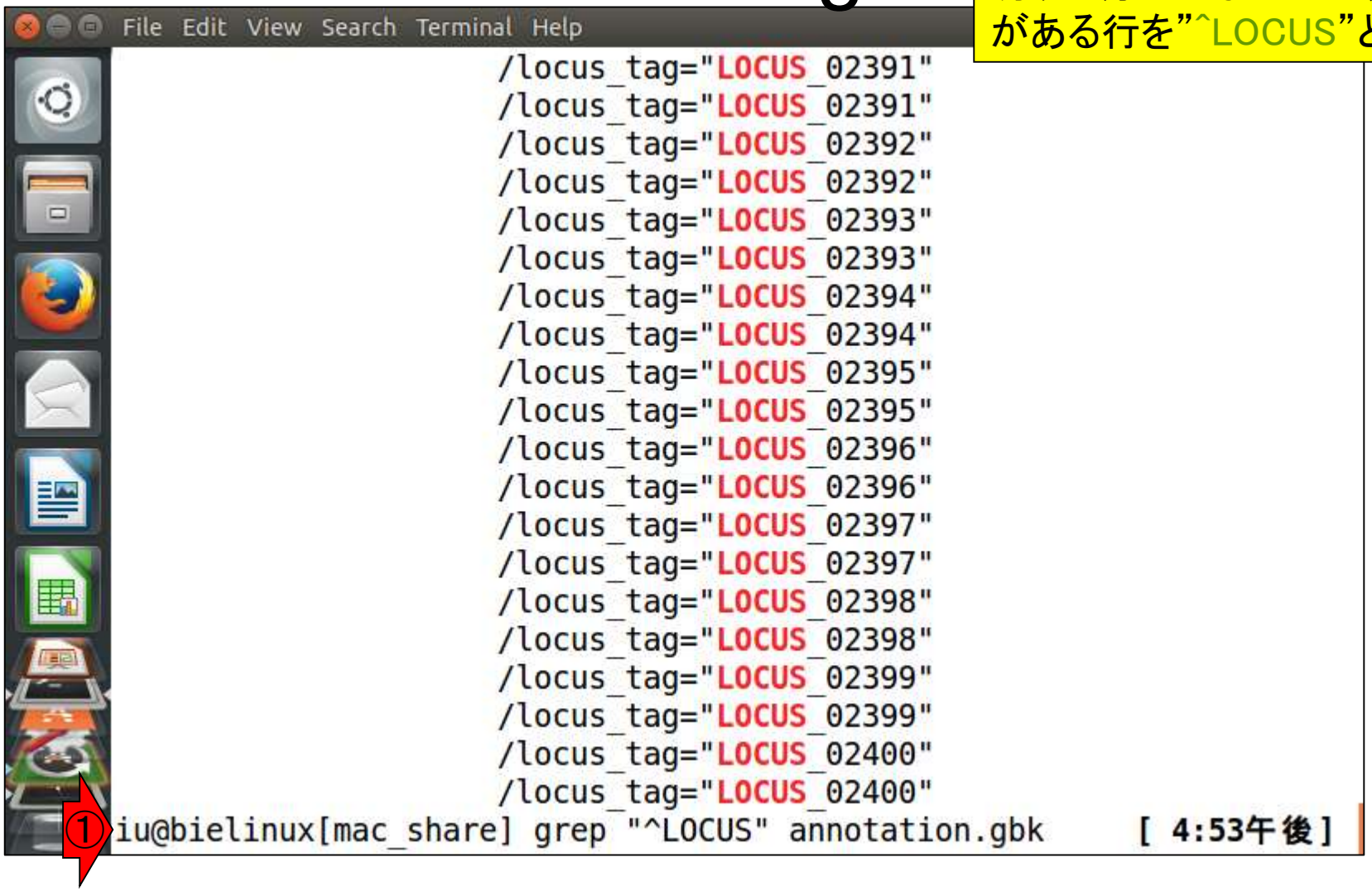
W13-1：annotation.gbk

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



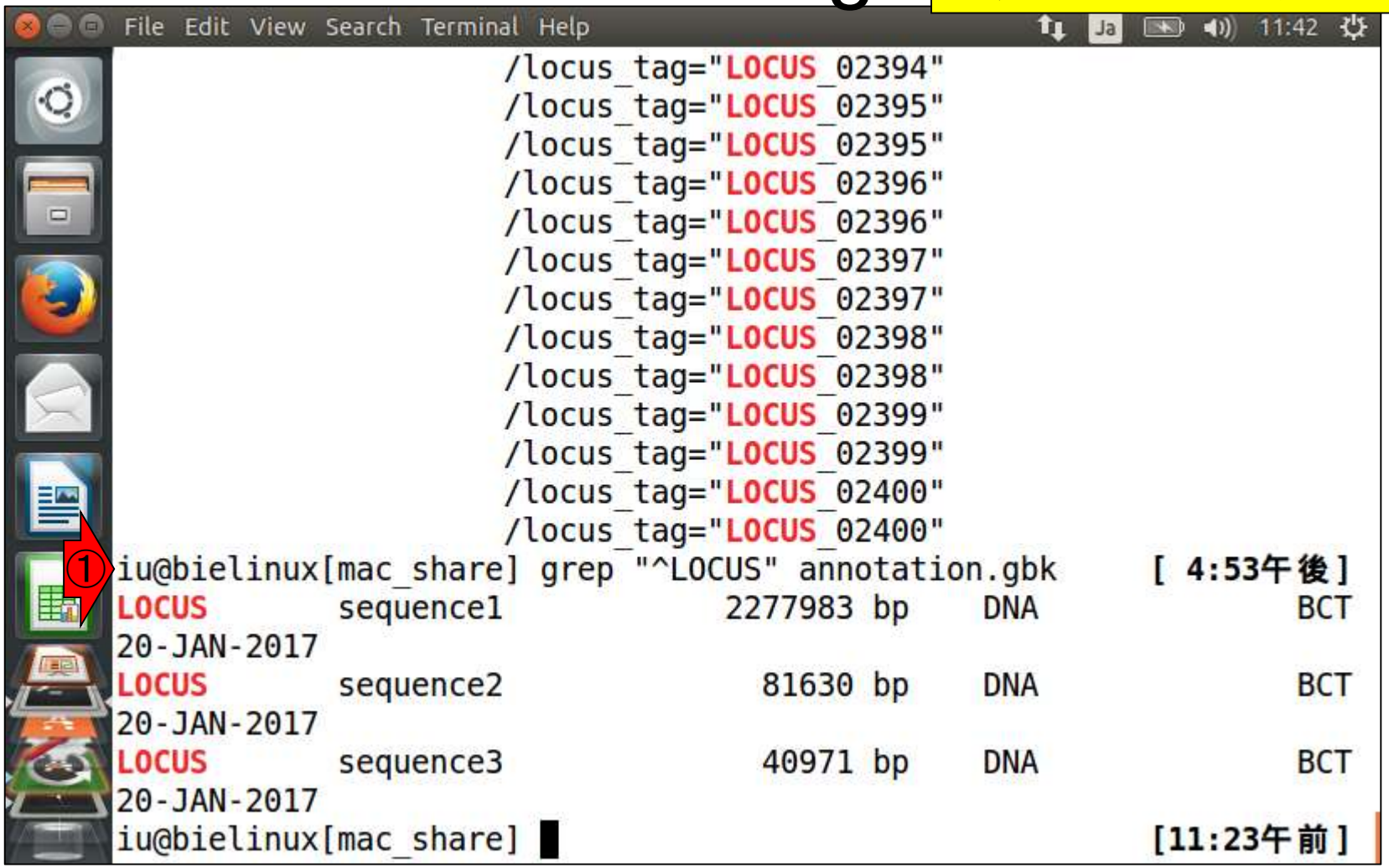
W13-1：annotation.gbk

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



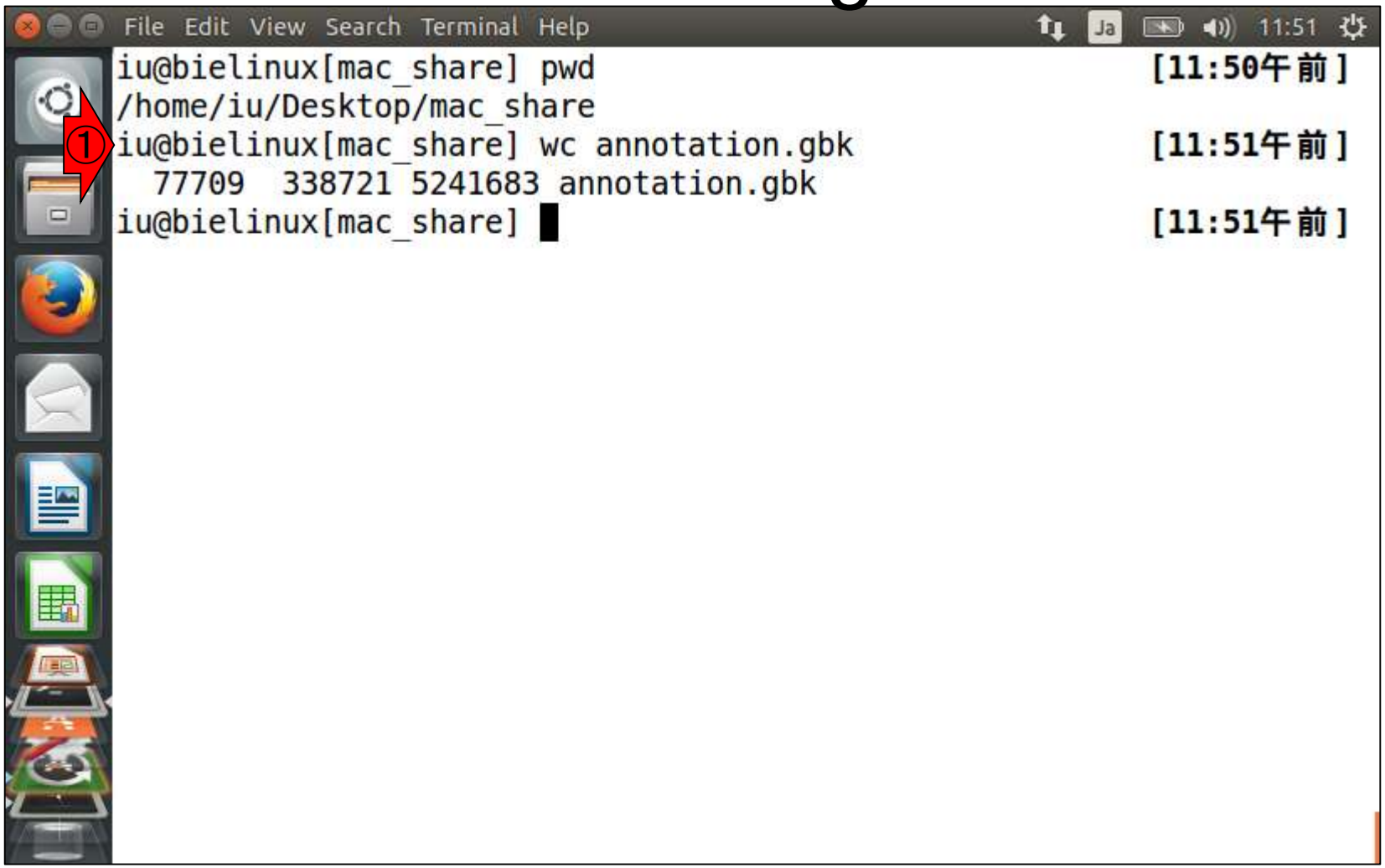
W13-1：annotation.gb

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



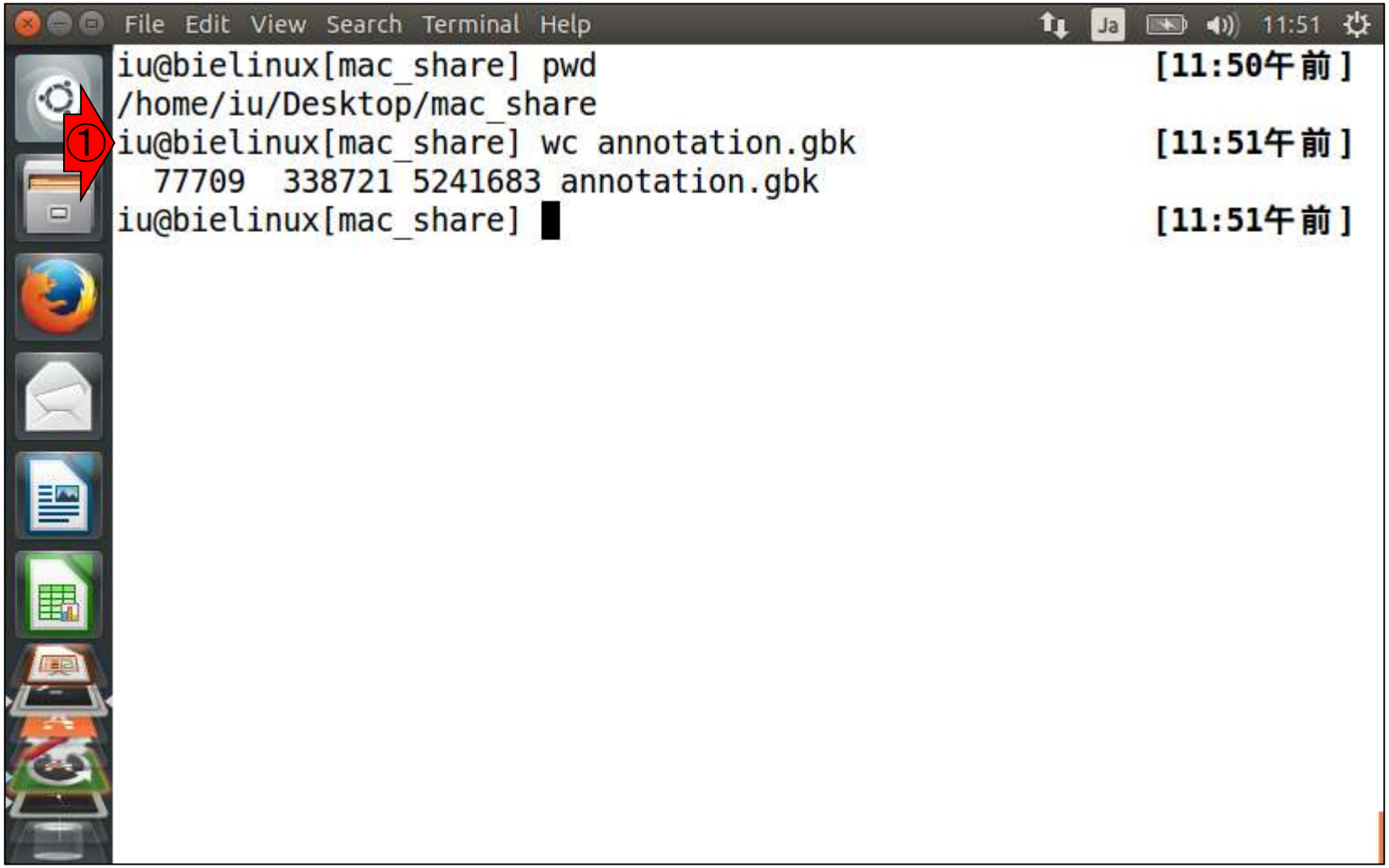
W13-2：annotation.gbk

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



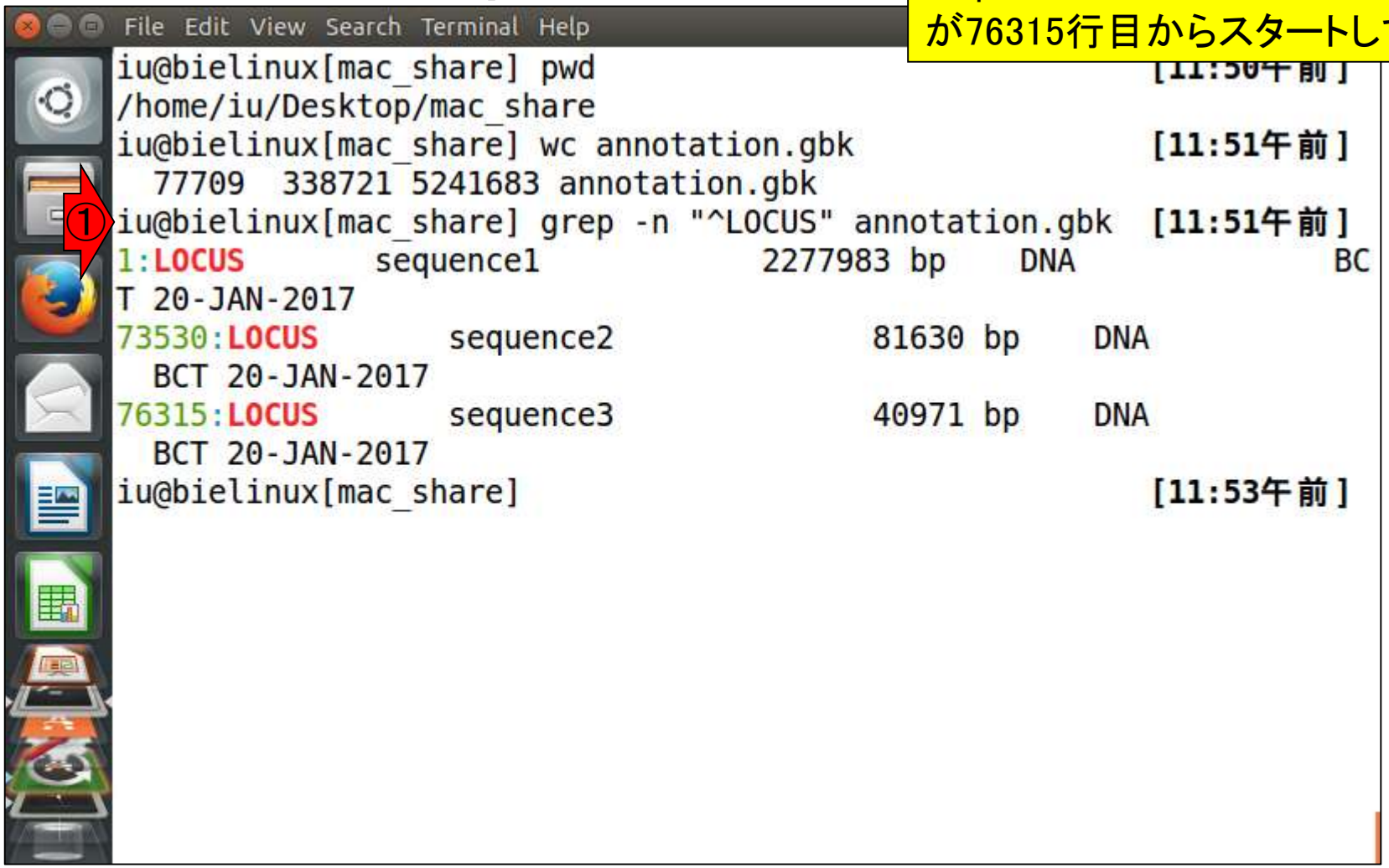
W13-2：全体像を把握

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



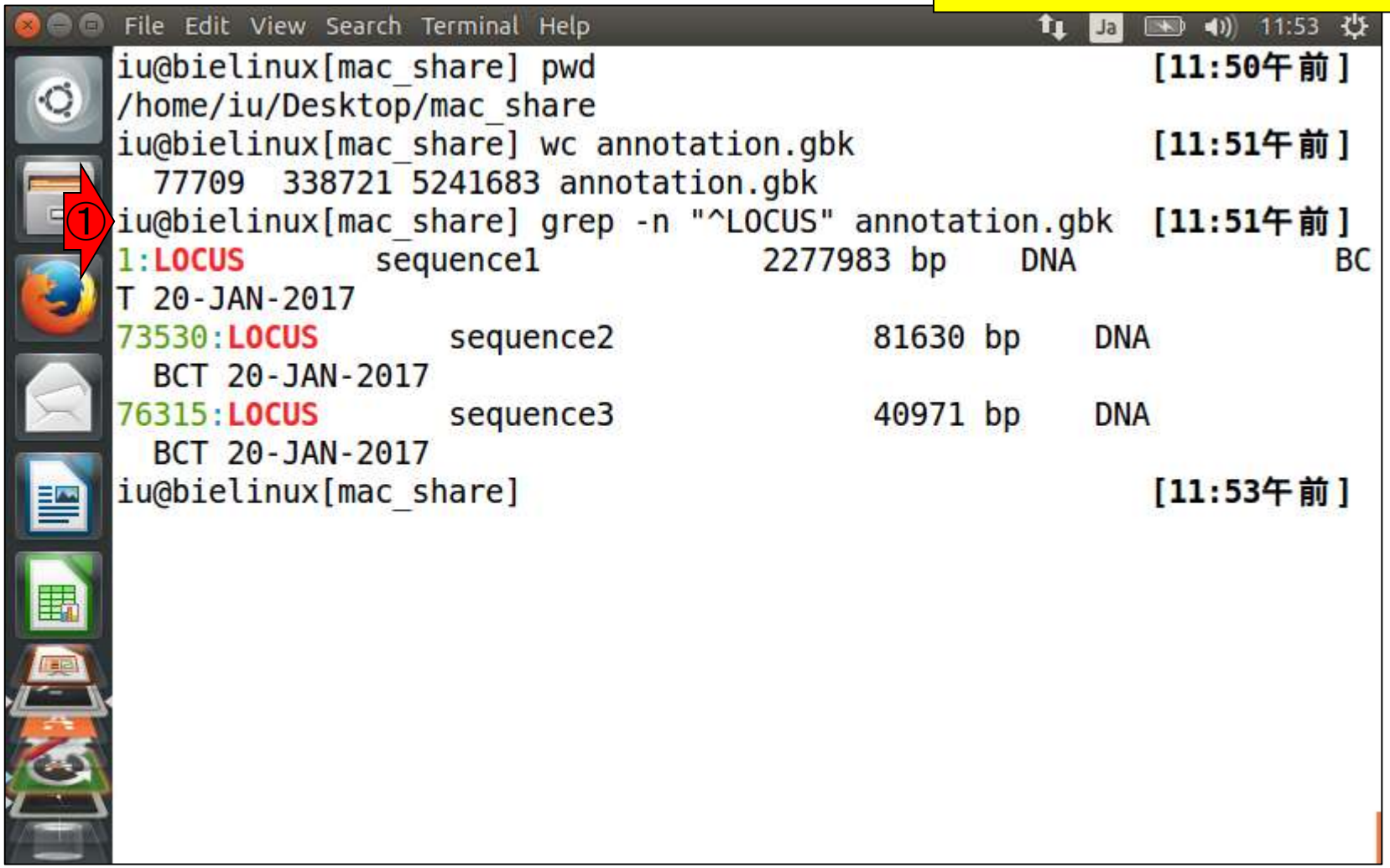
W13-2：全体像を把握

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



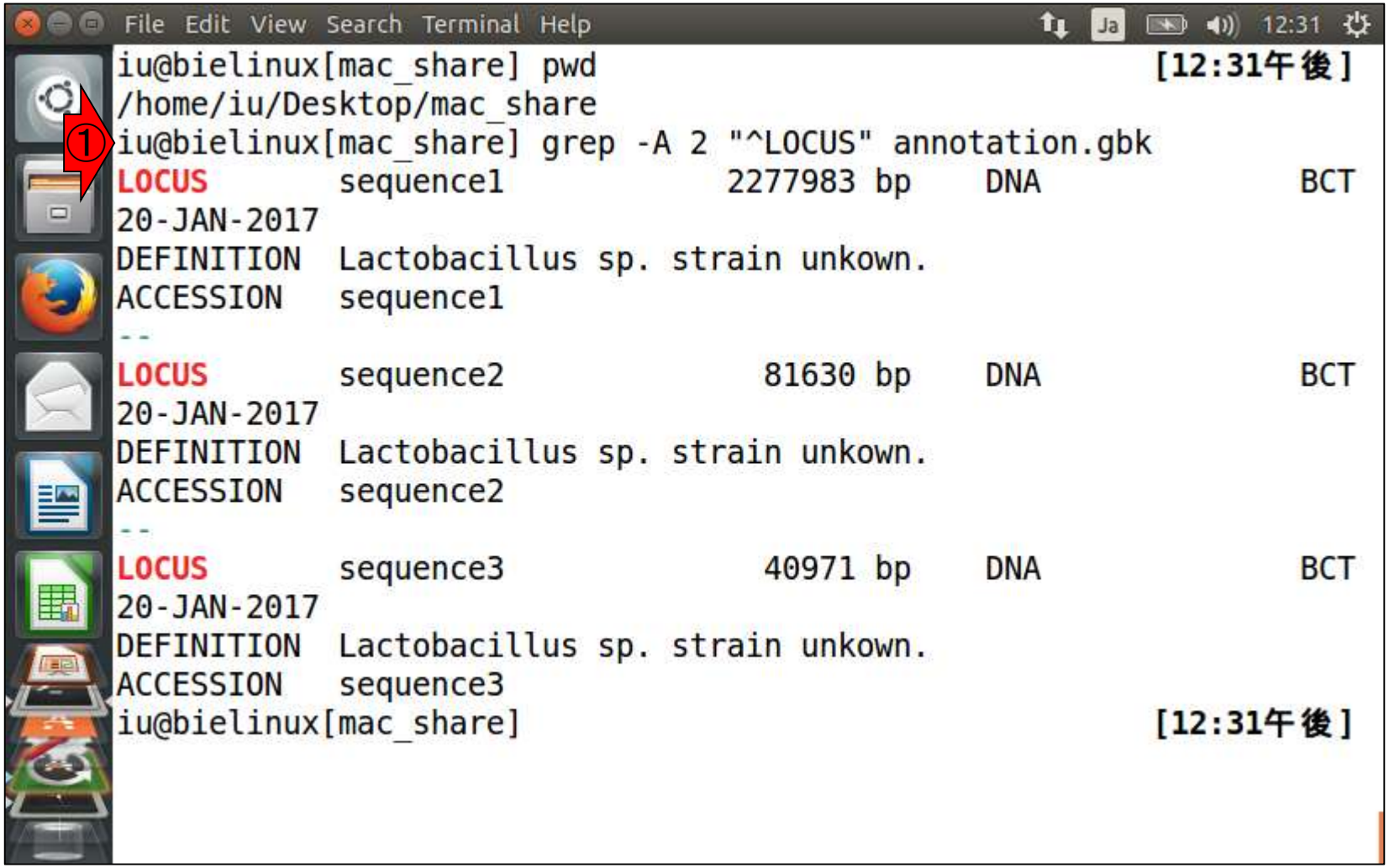
W13-2：全体像を把握

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)



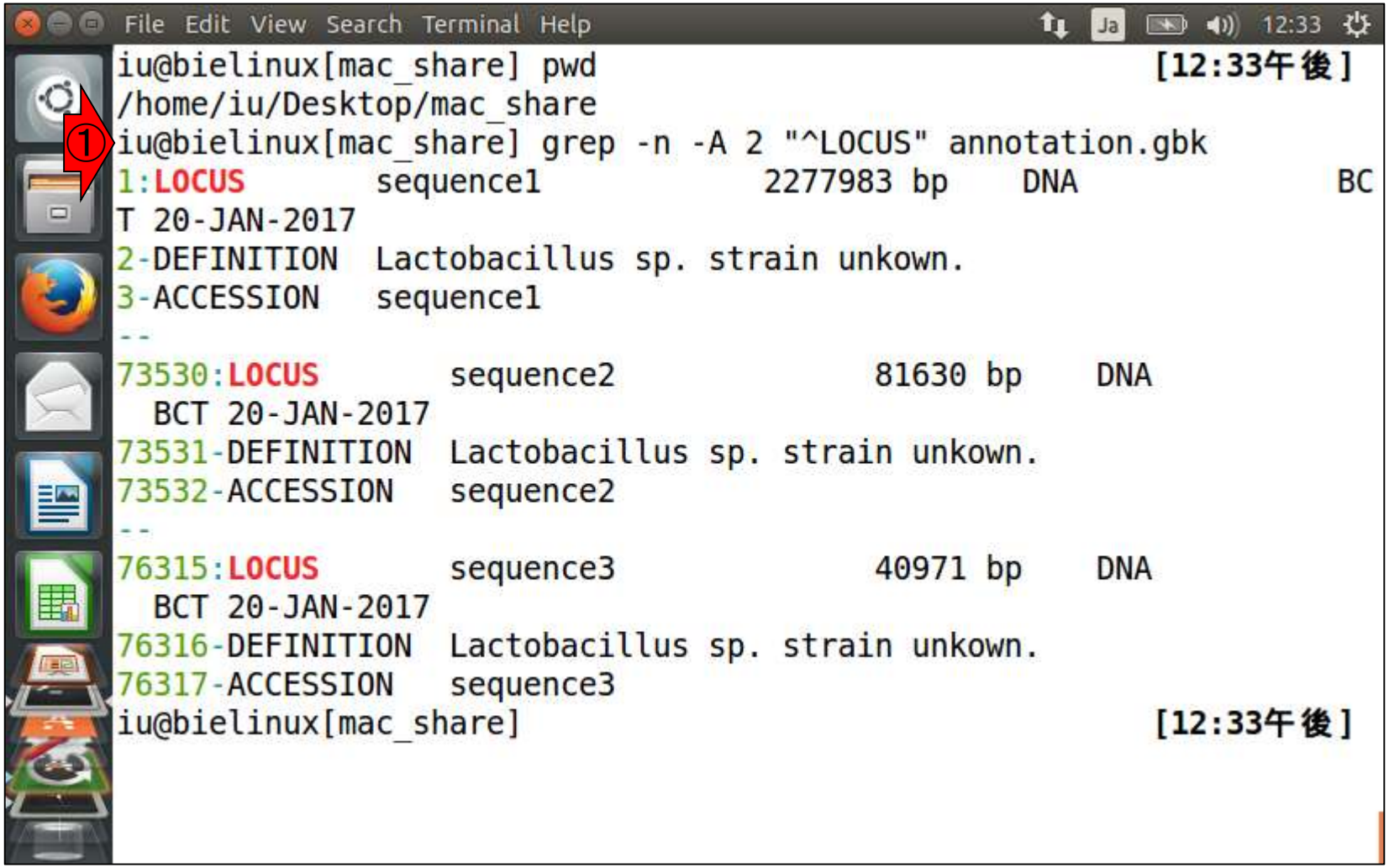
W13-3：さらに把握

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)。



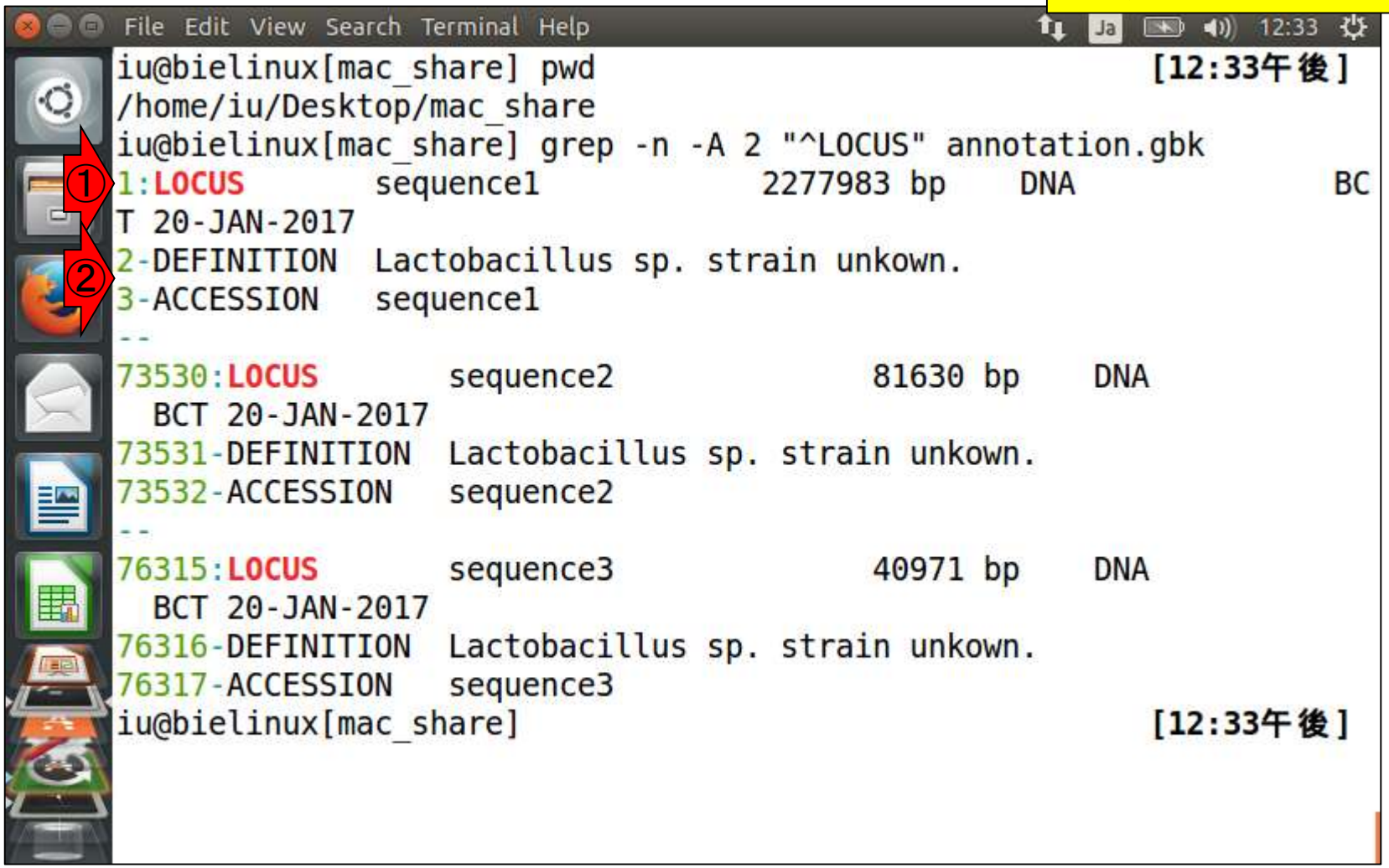
①grep −n −A 2として、行番号も表示させている

W13-3：さらに把握



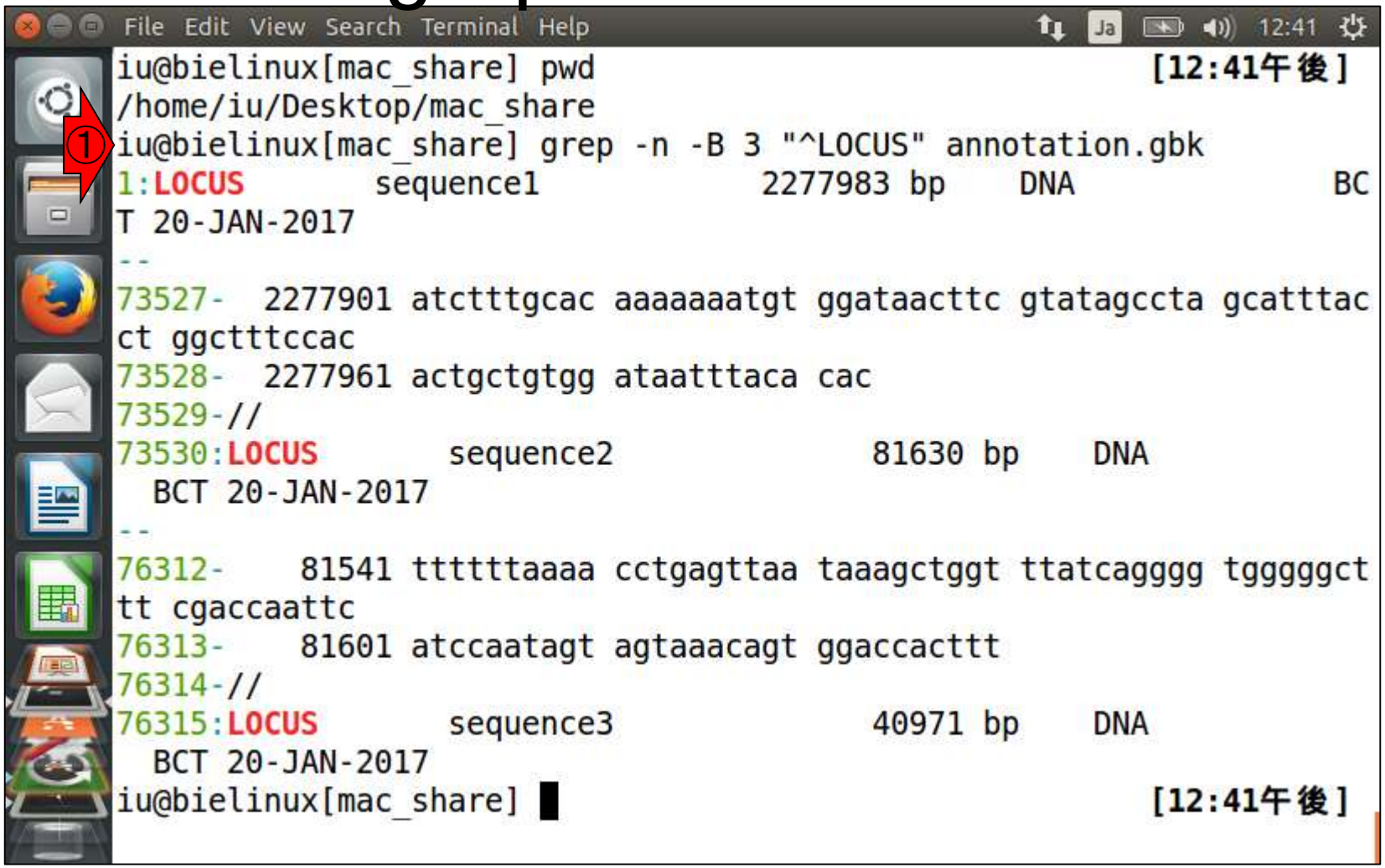
W13-3：さらに把握

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



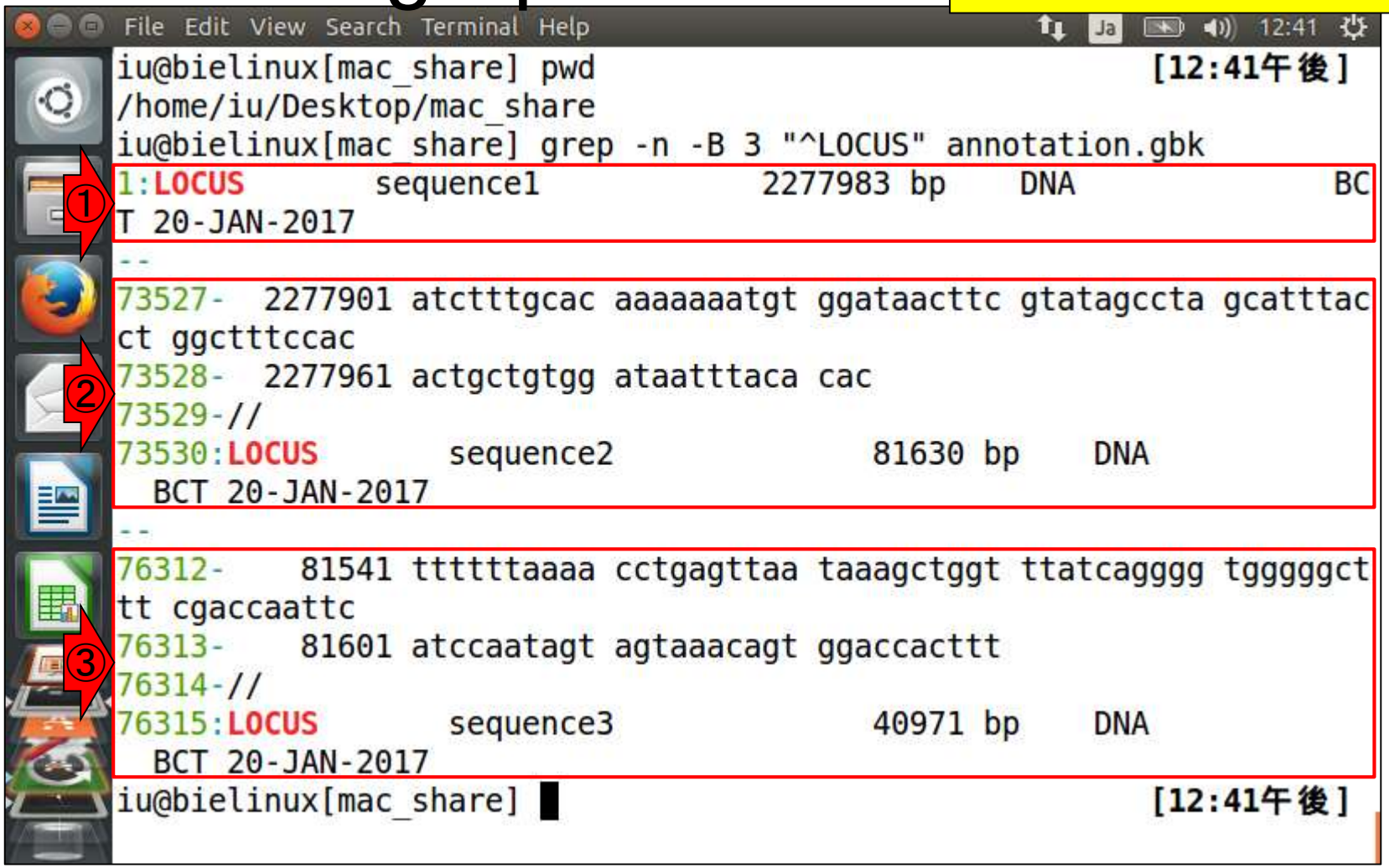
W13-4：grep -B

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



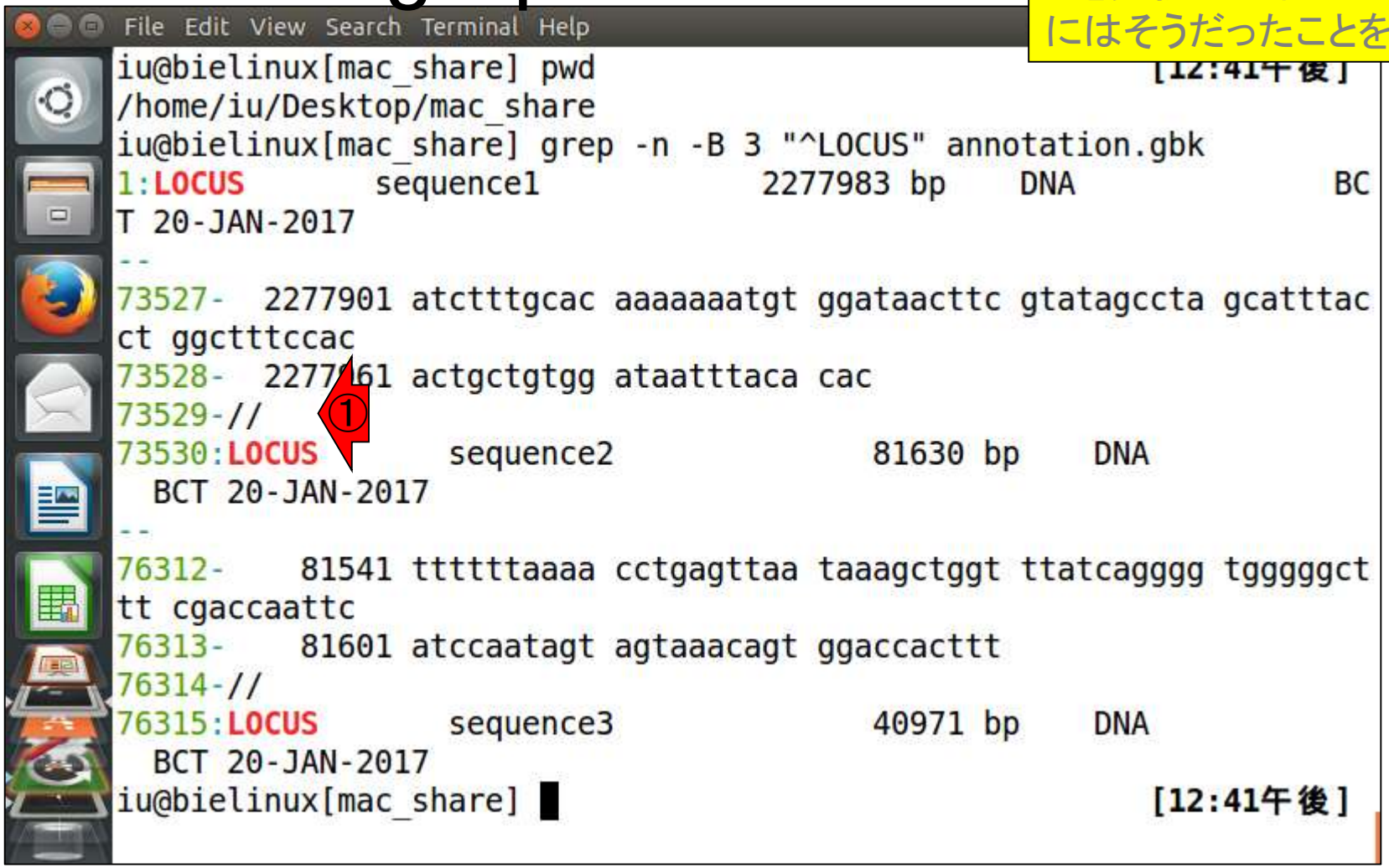
W13-4：grep -B

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



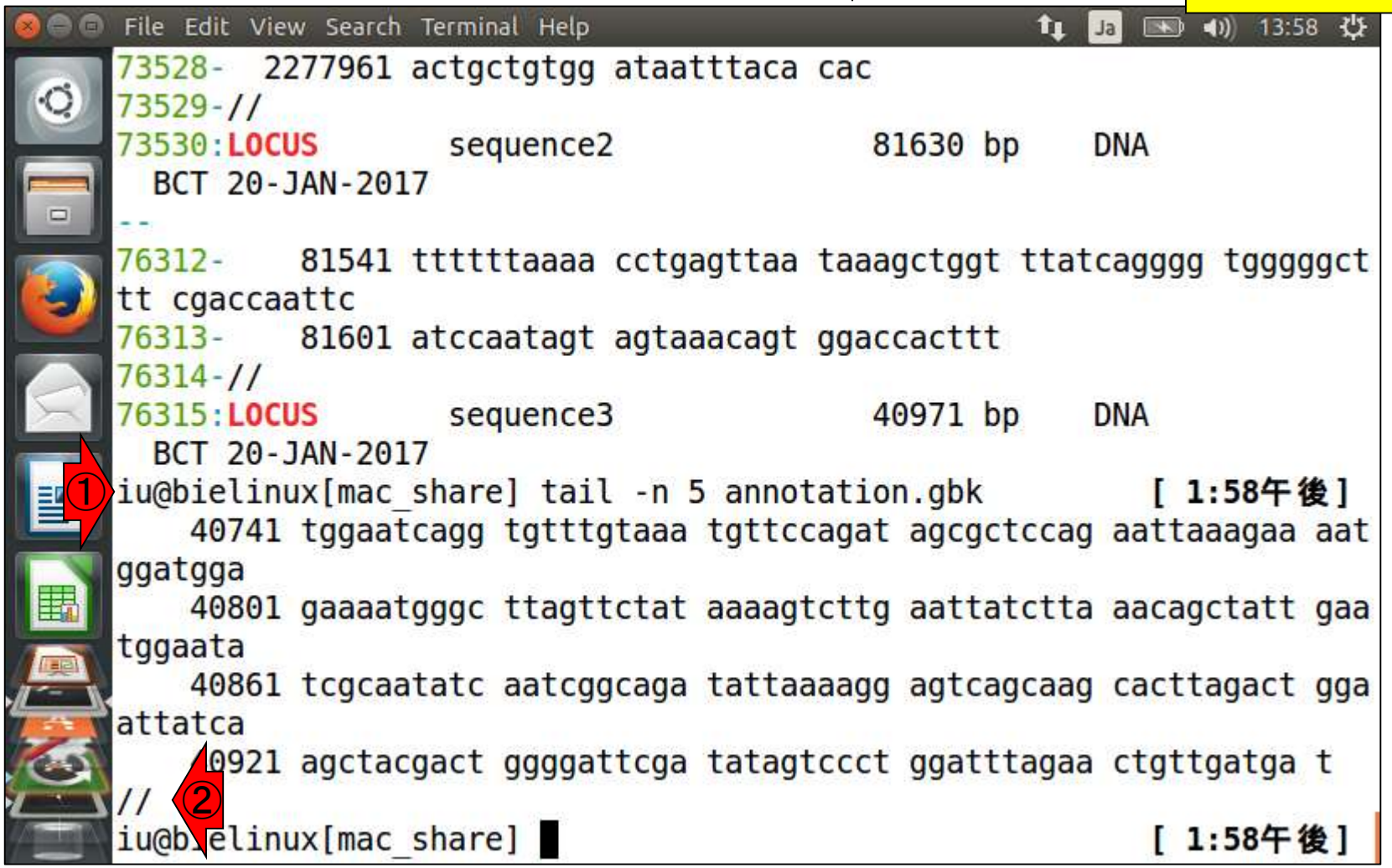
W13-4：grep -B

①//が各配列の最後を表すのでしょう。
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



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

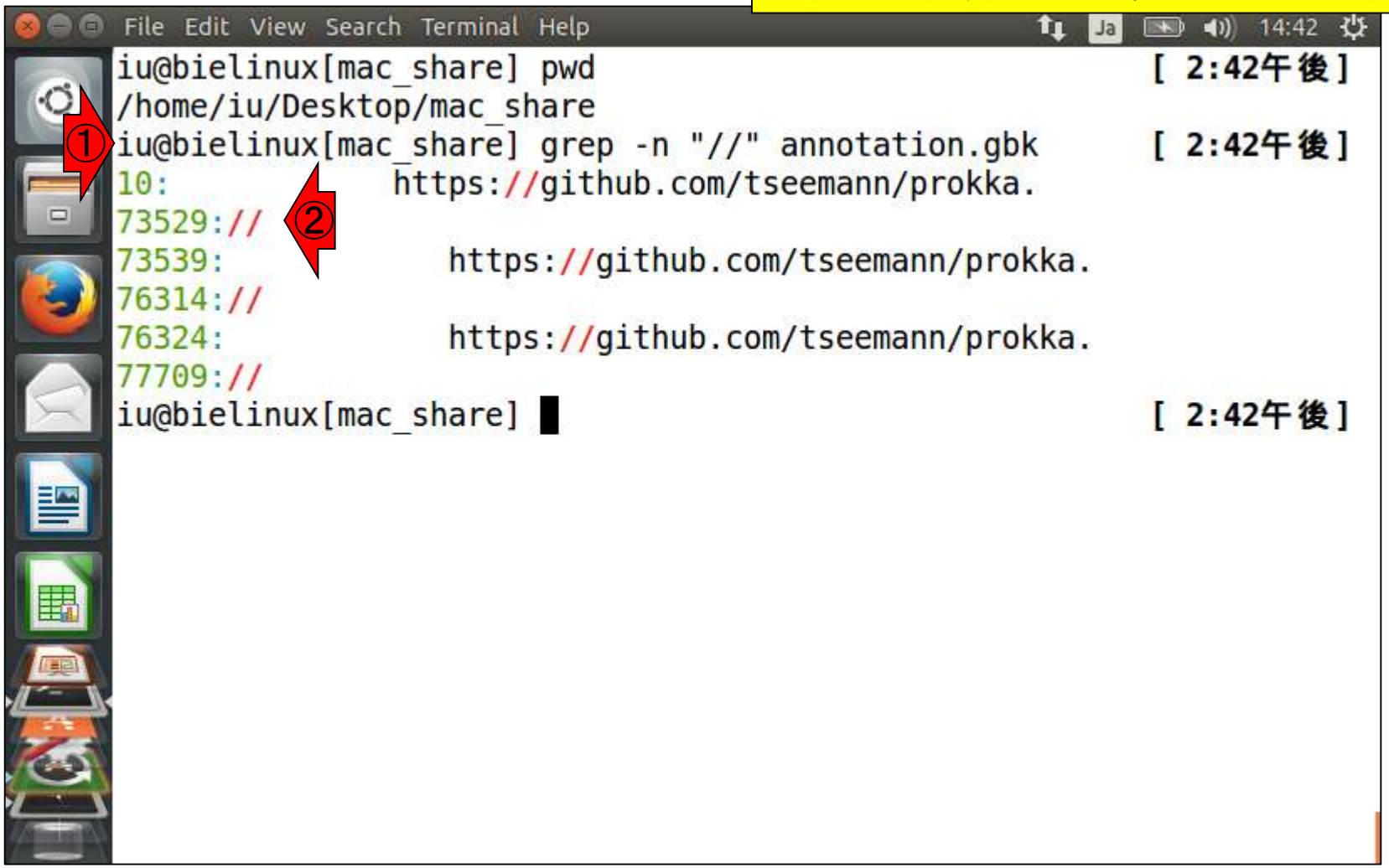


W14-1：awk

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

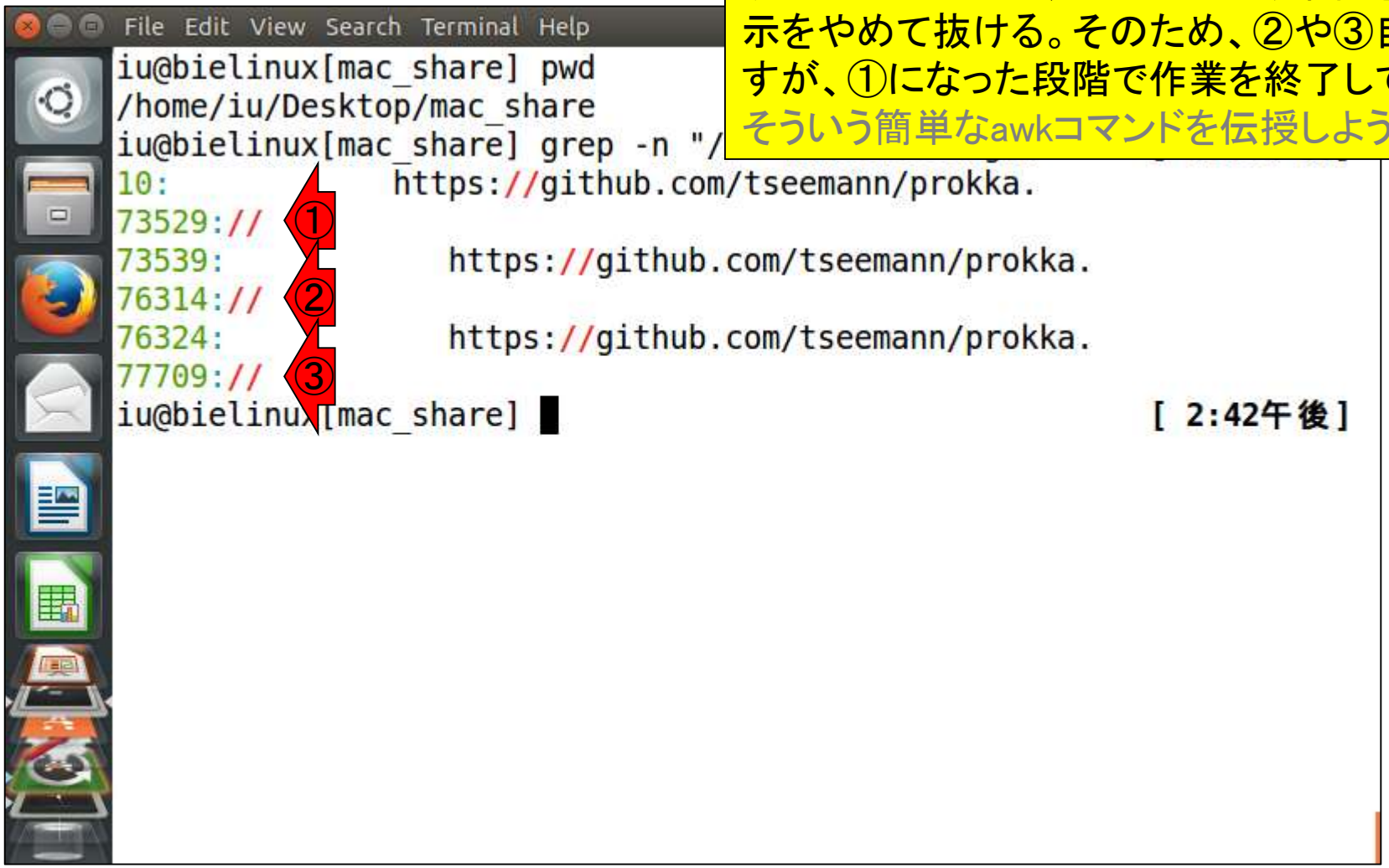
W14-1：awk

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ことを行う



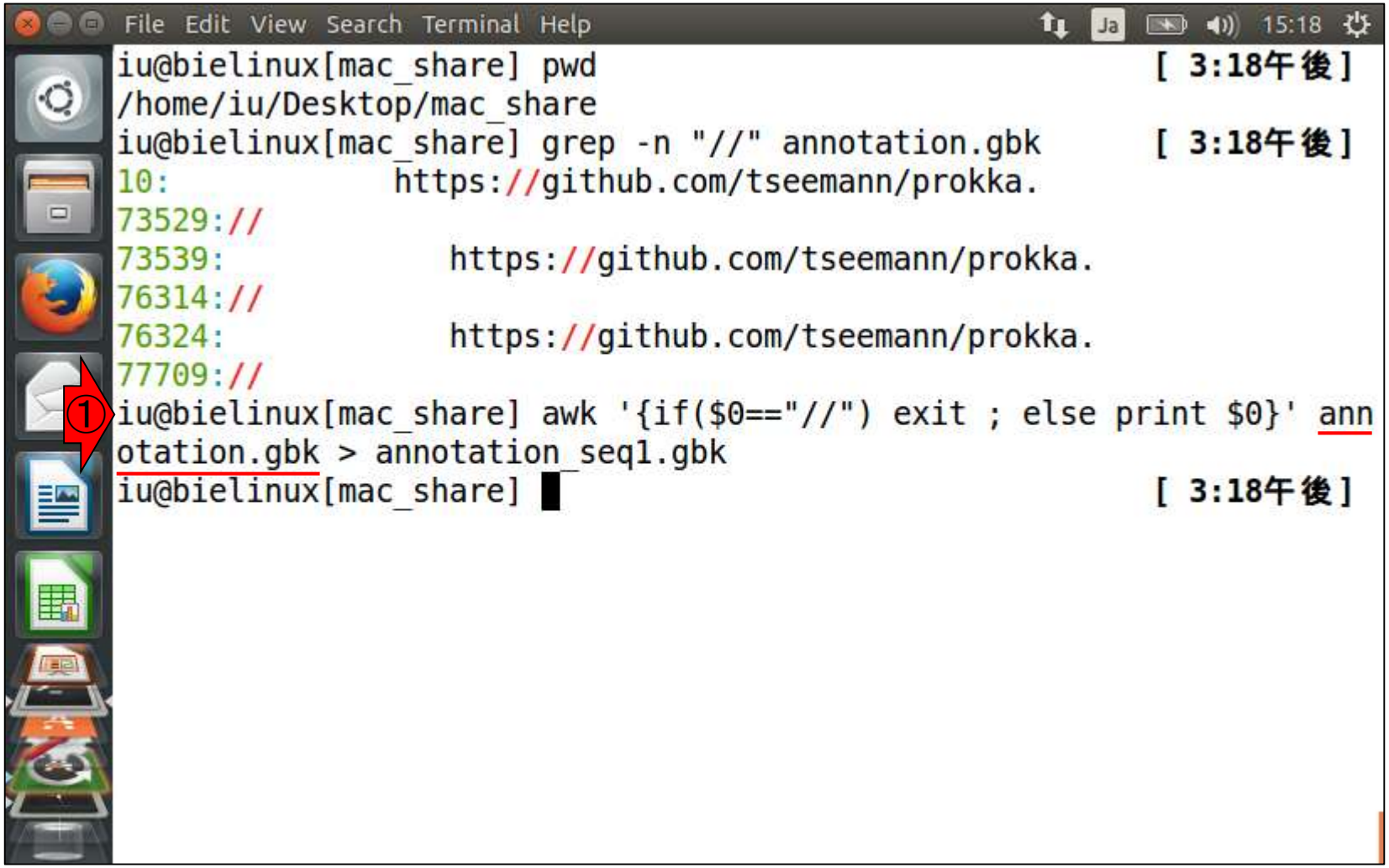
W14-1：awk

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



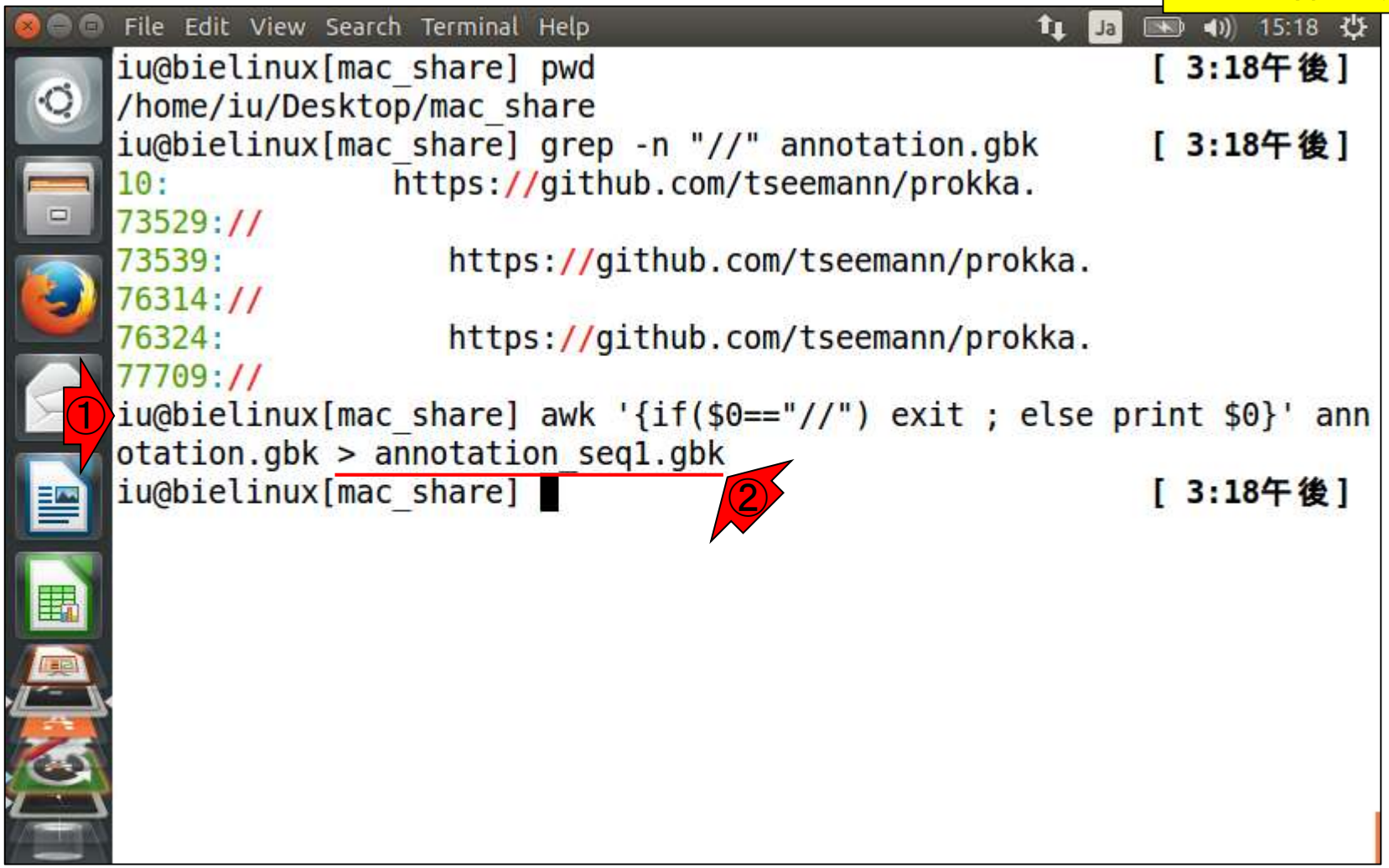
W14-1：awk

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



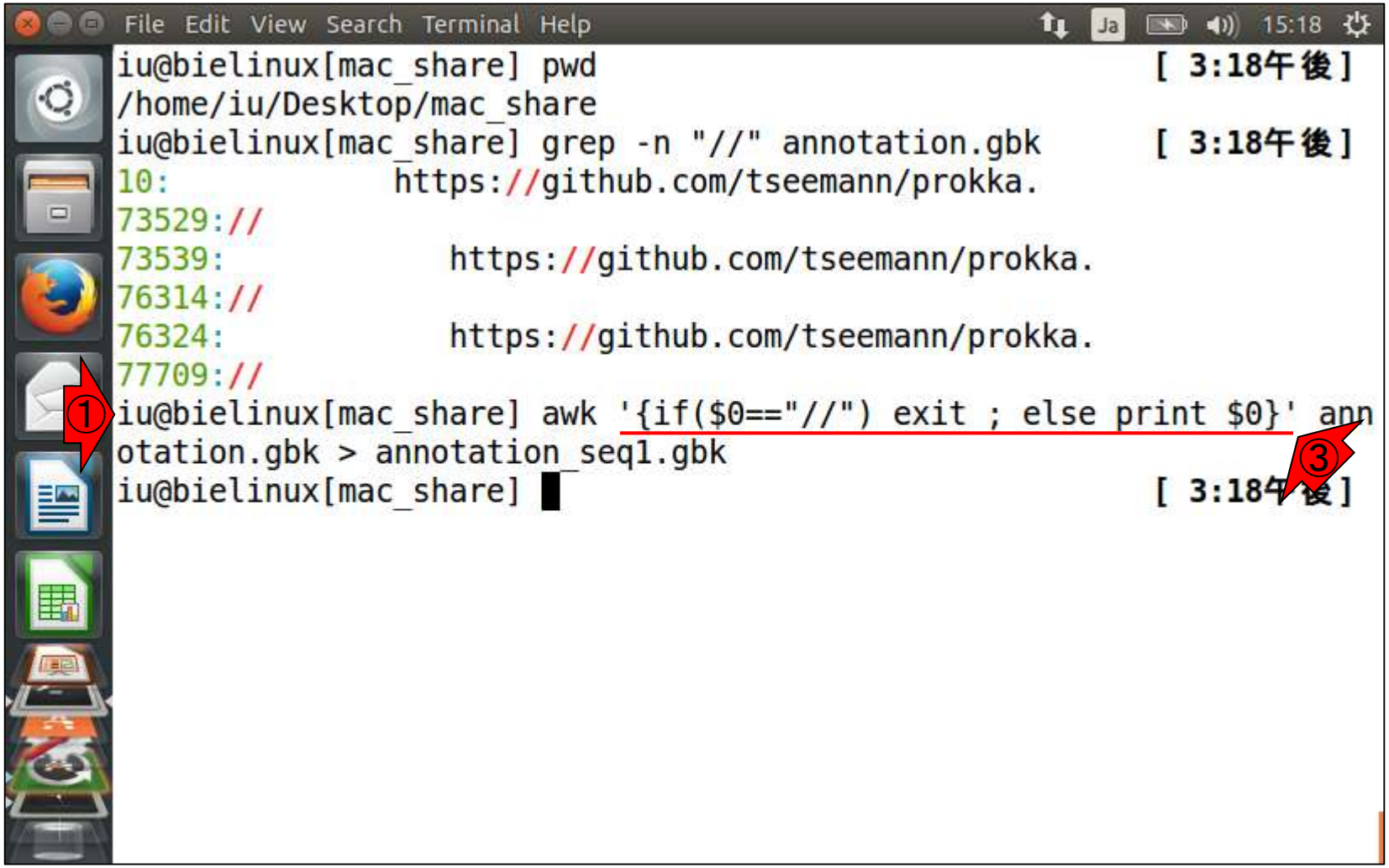
W14-1：awk

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



③がやってもらいたいことを書く部分

W14-1：awk

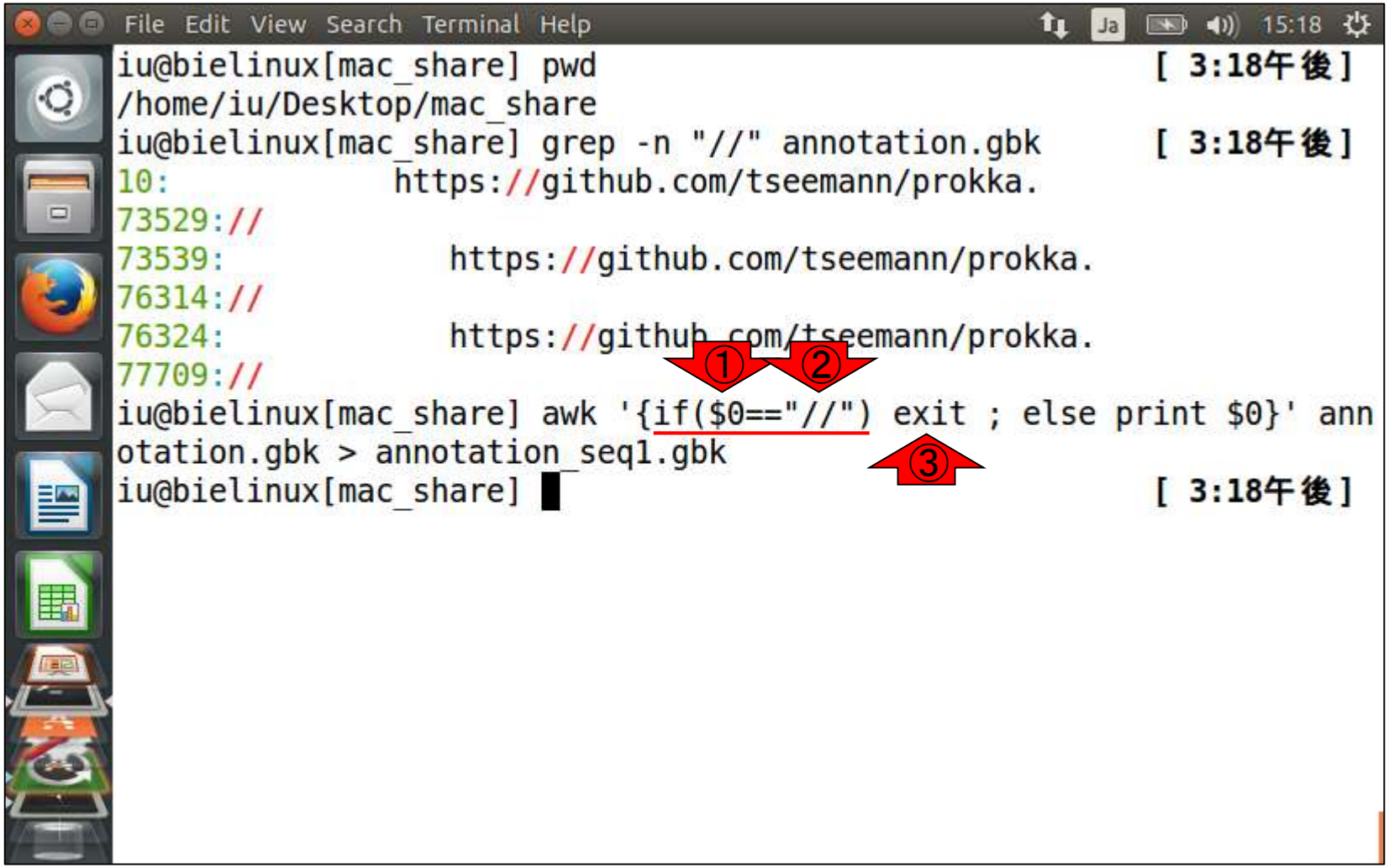


W14-2：解説

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

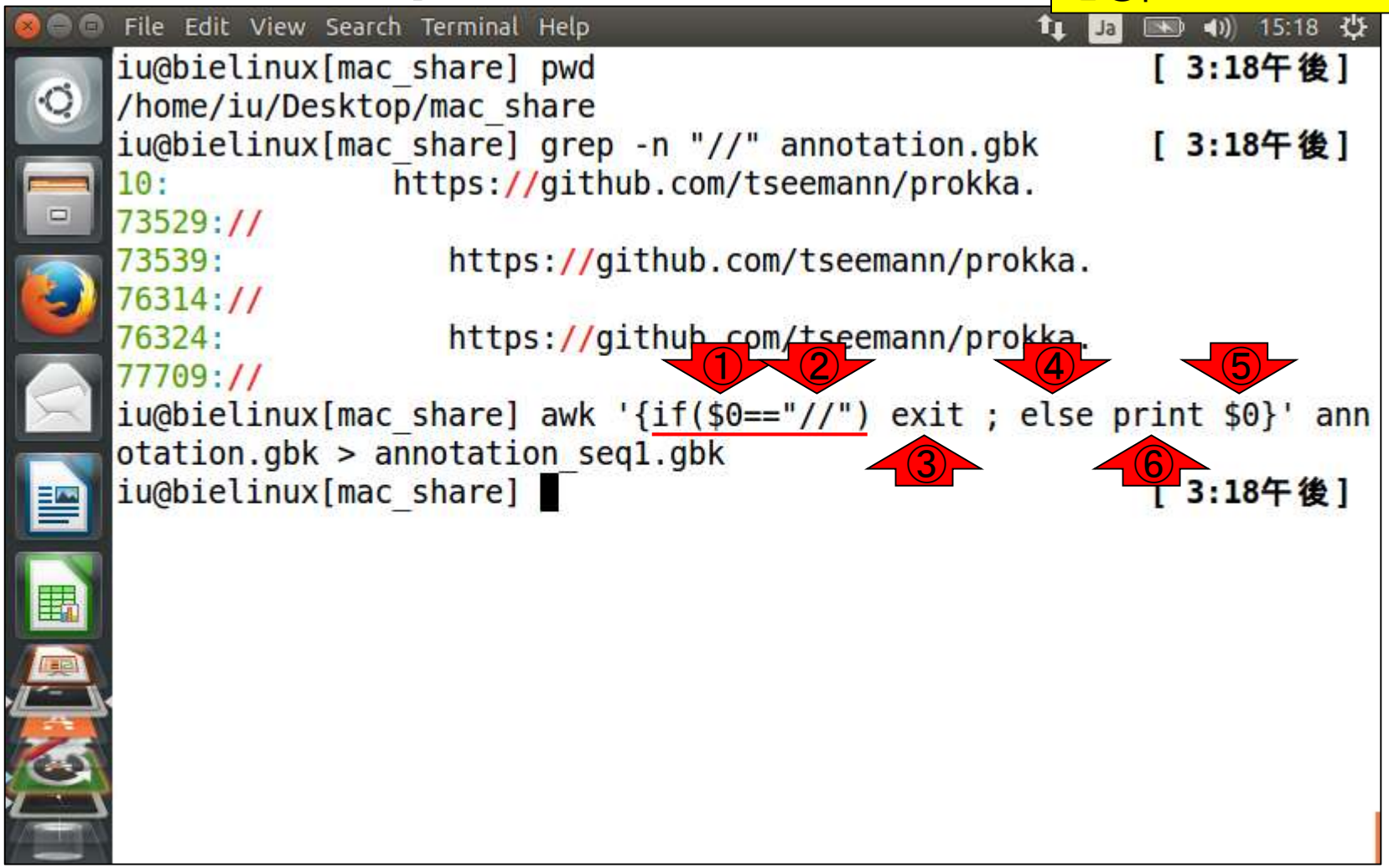
W14-2：解説

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



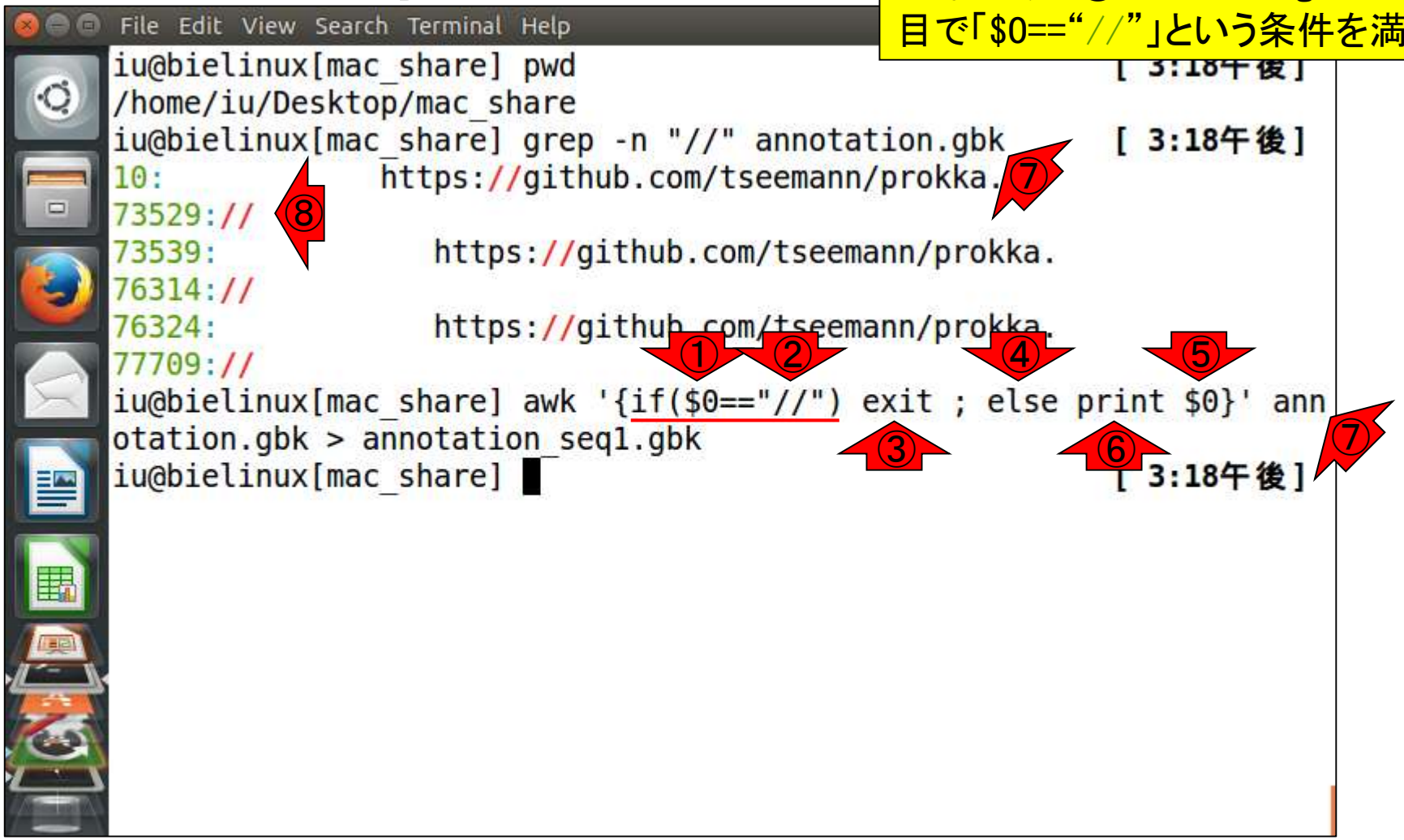
W14-2：解説

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



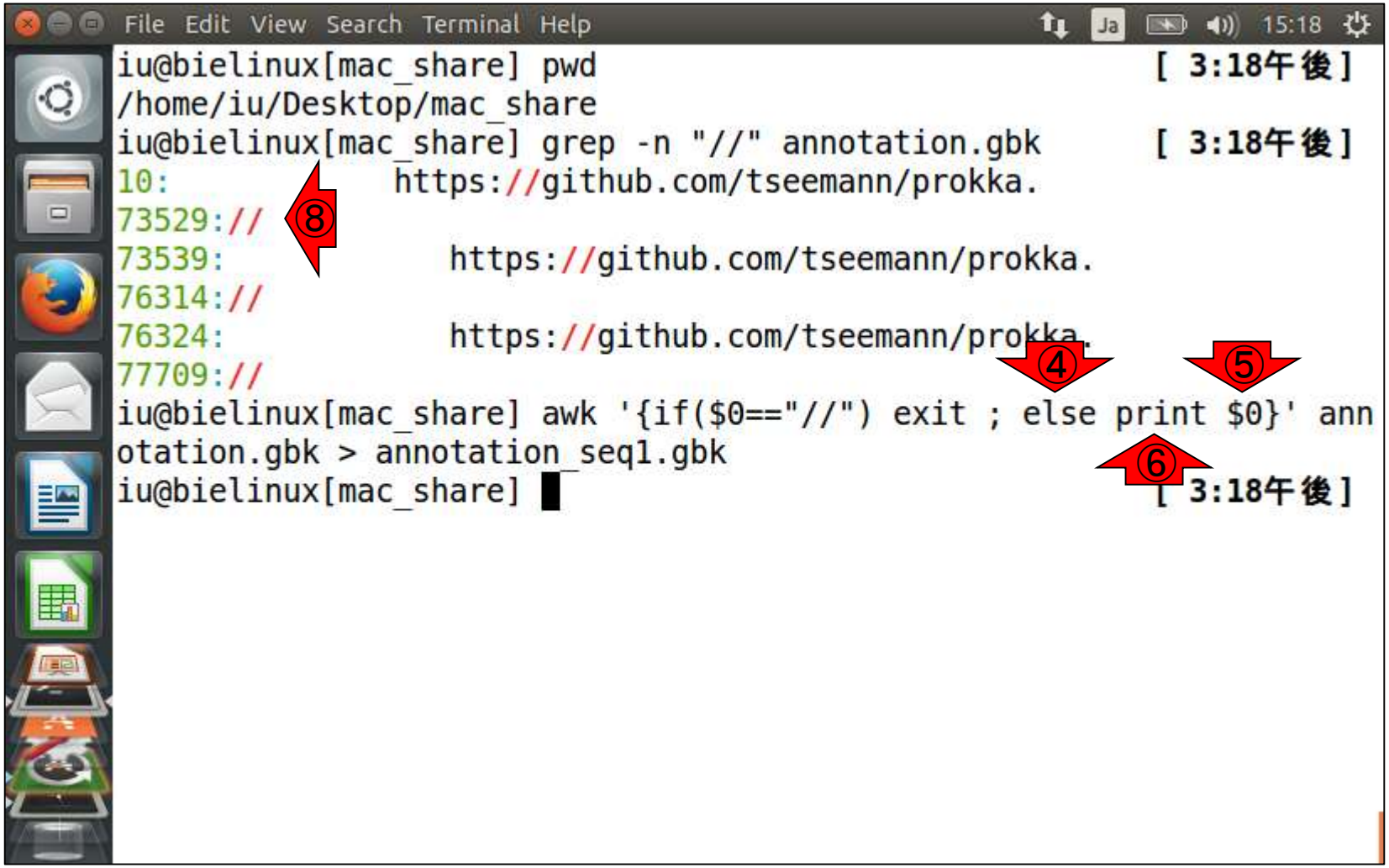
W14-2：解説

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…



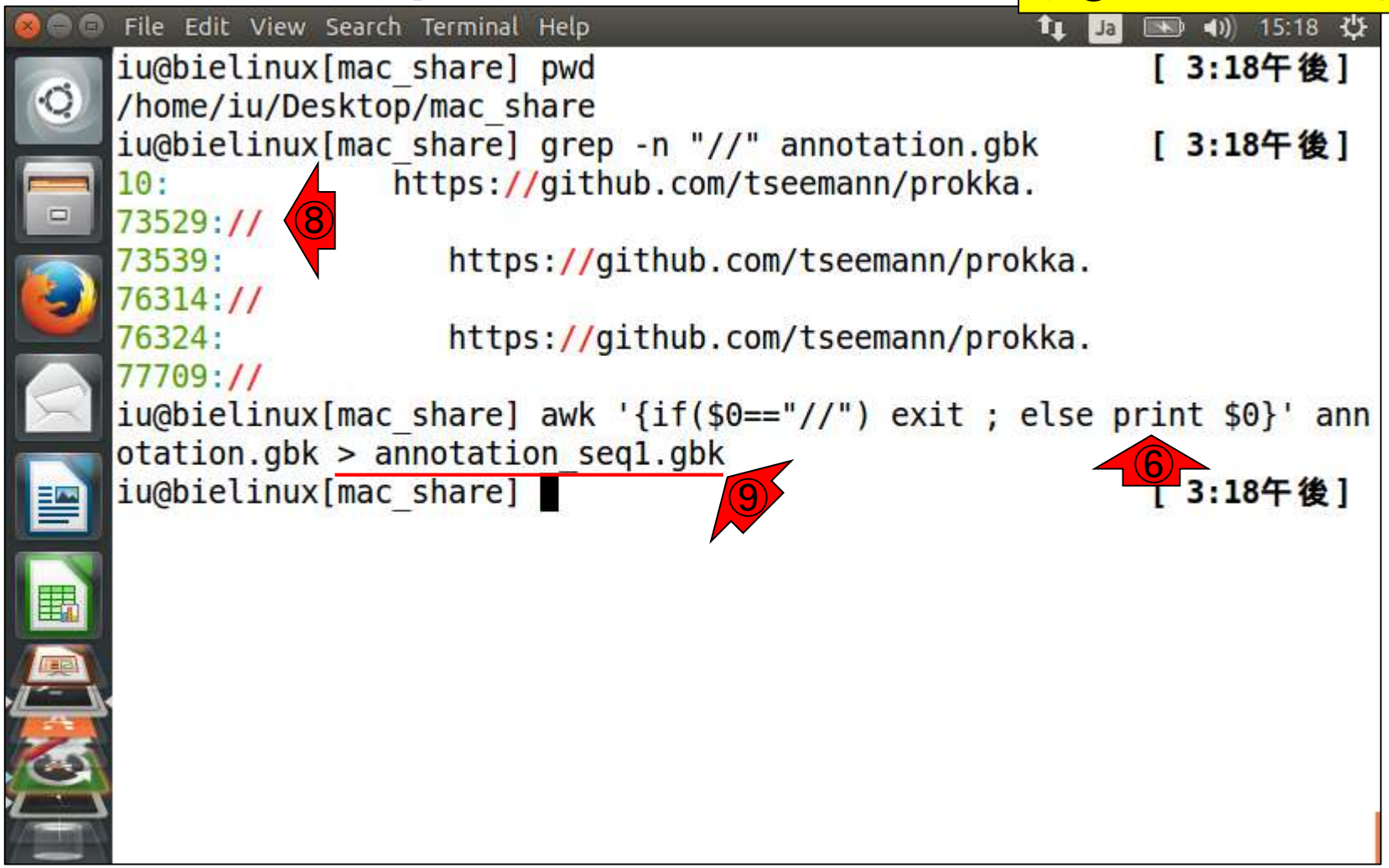
W14-2：解説

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



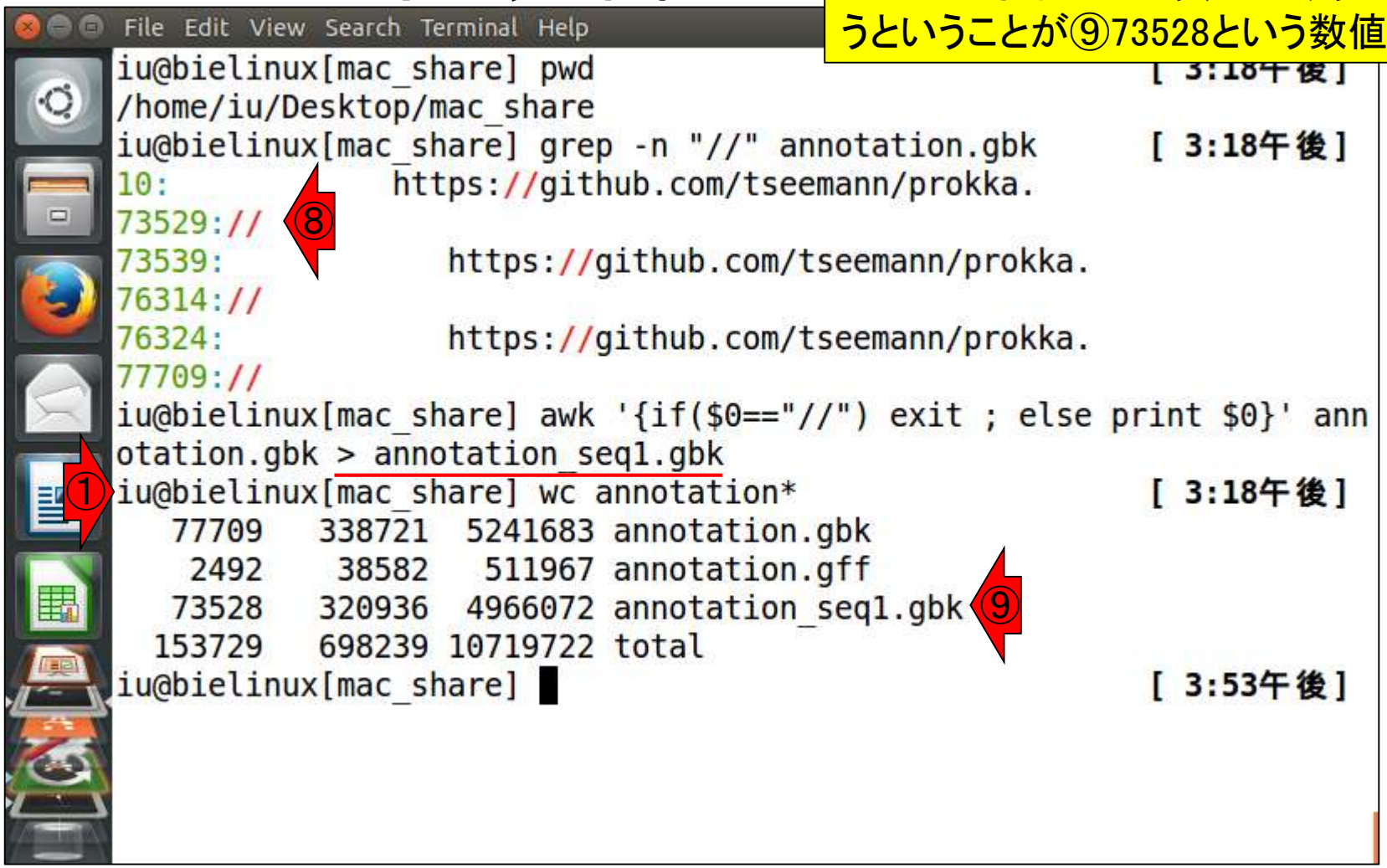
W14-2：解説

⑥print先である、⑨annotation_seq1.gbkの中身は、「$0==“//”」という条件を満たす⑧73529行目の手前の行までの部分



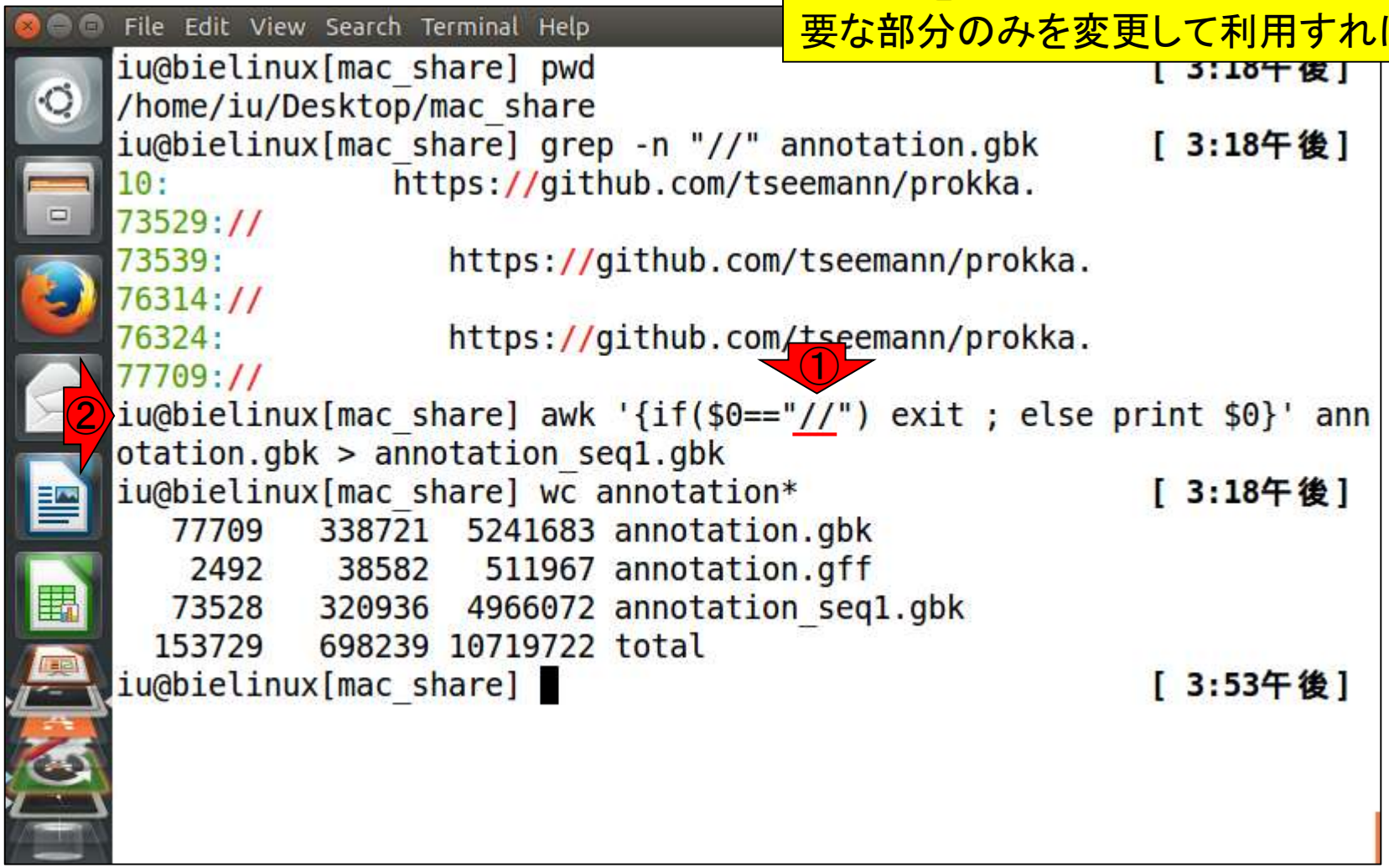
W14-3：行数確認

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



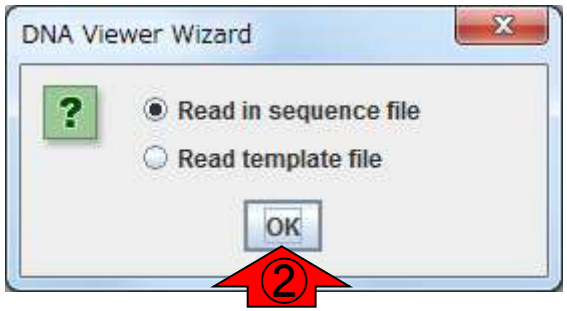
W14-4：awk

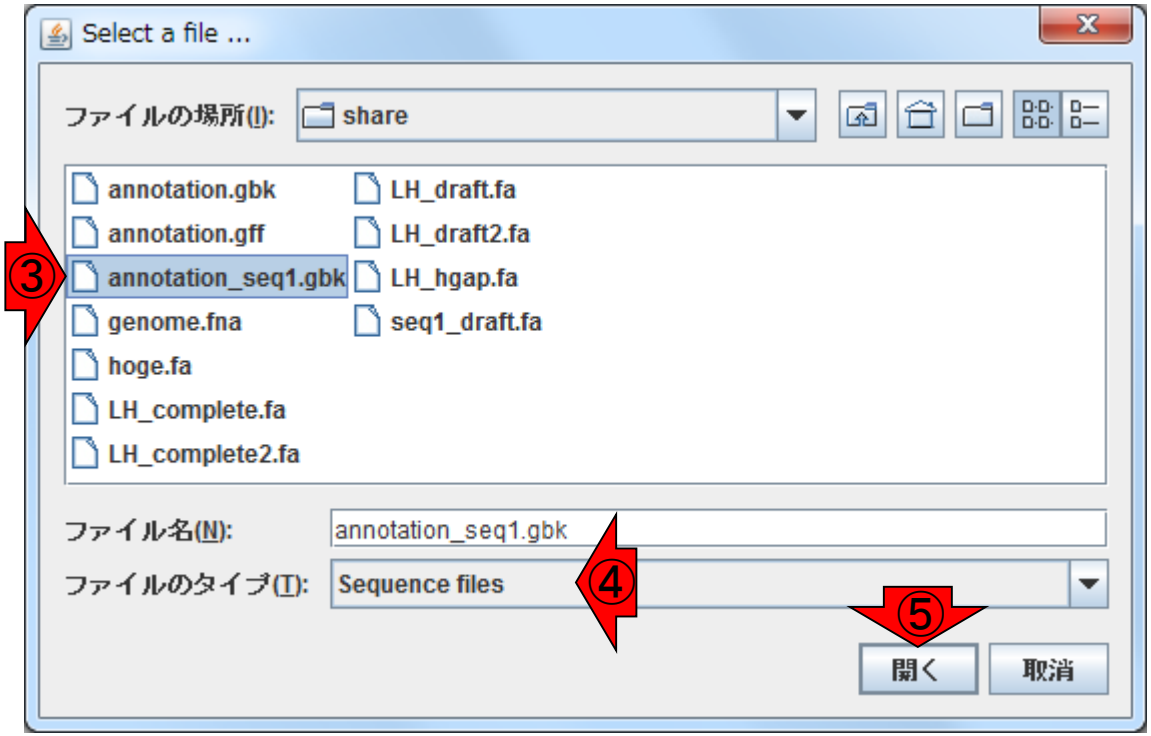
一見すごくややこしいが、①で指定した文字列と合致する行（の手前）までを出力したい場合に、②全体のスクリプトをテンプレートとして利用できる。コピペして必要な部分のみを変更して利用すればよい、ということ



W15-1：DNAPlotter

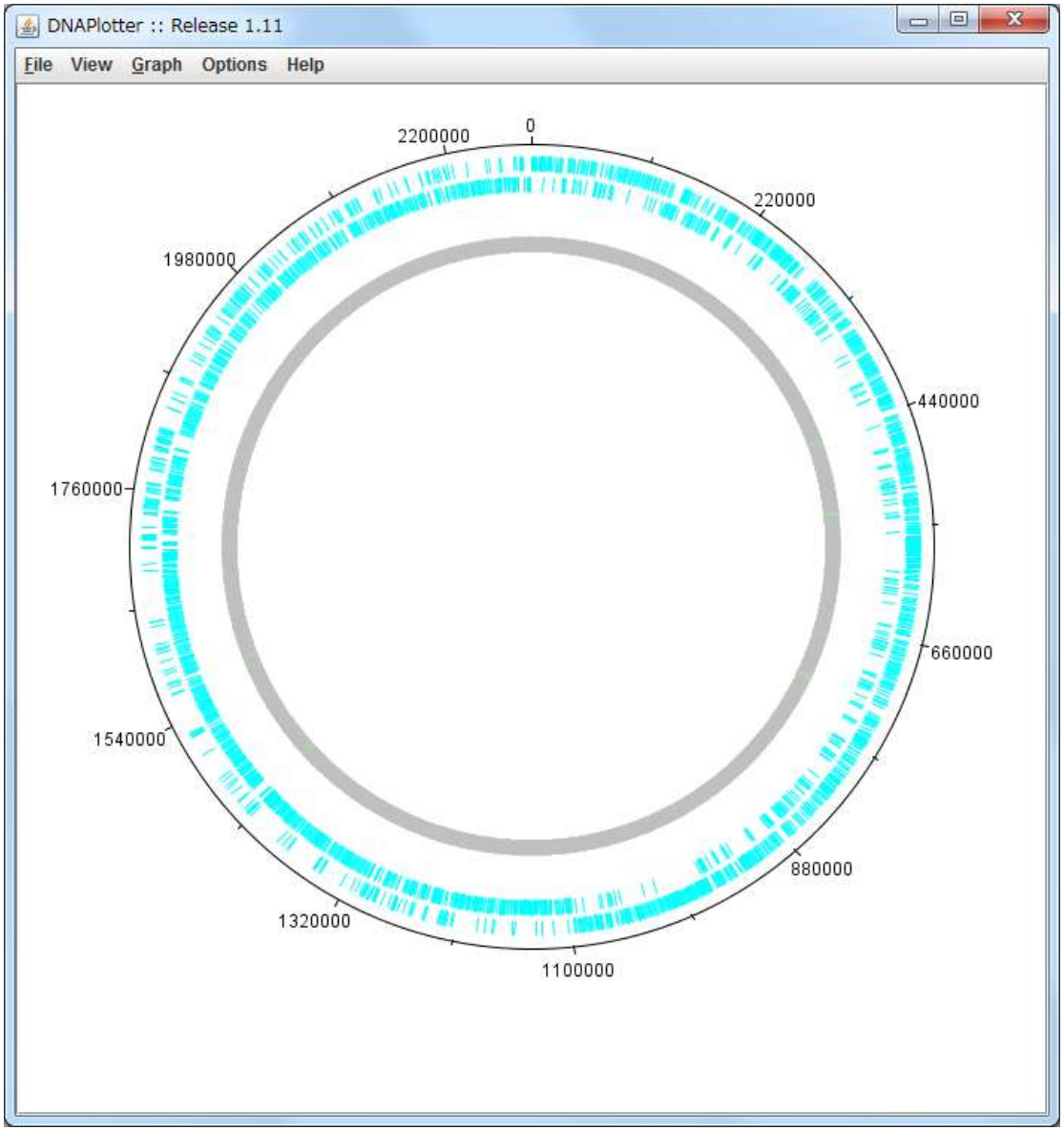
①②DNAPlotterを起動して、作成した③annotation_seq1.gbkを、④ファイルのタイプはSequence filesのままでよいので、⑤開く





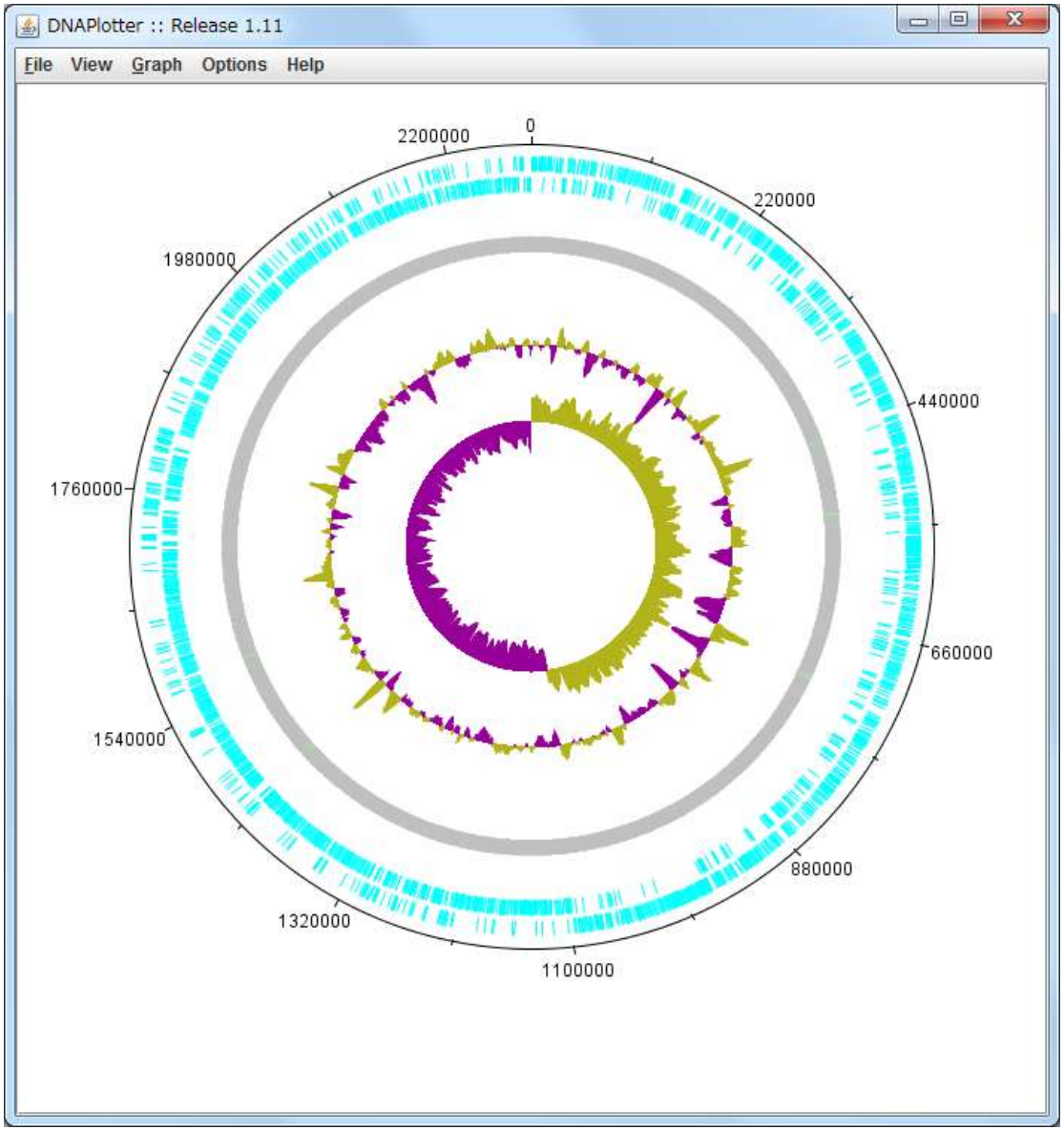
W15-1：DNAPlotter

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



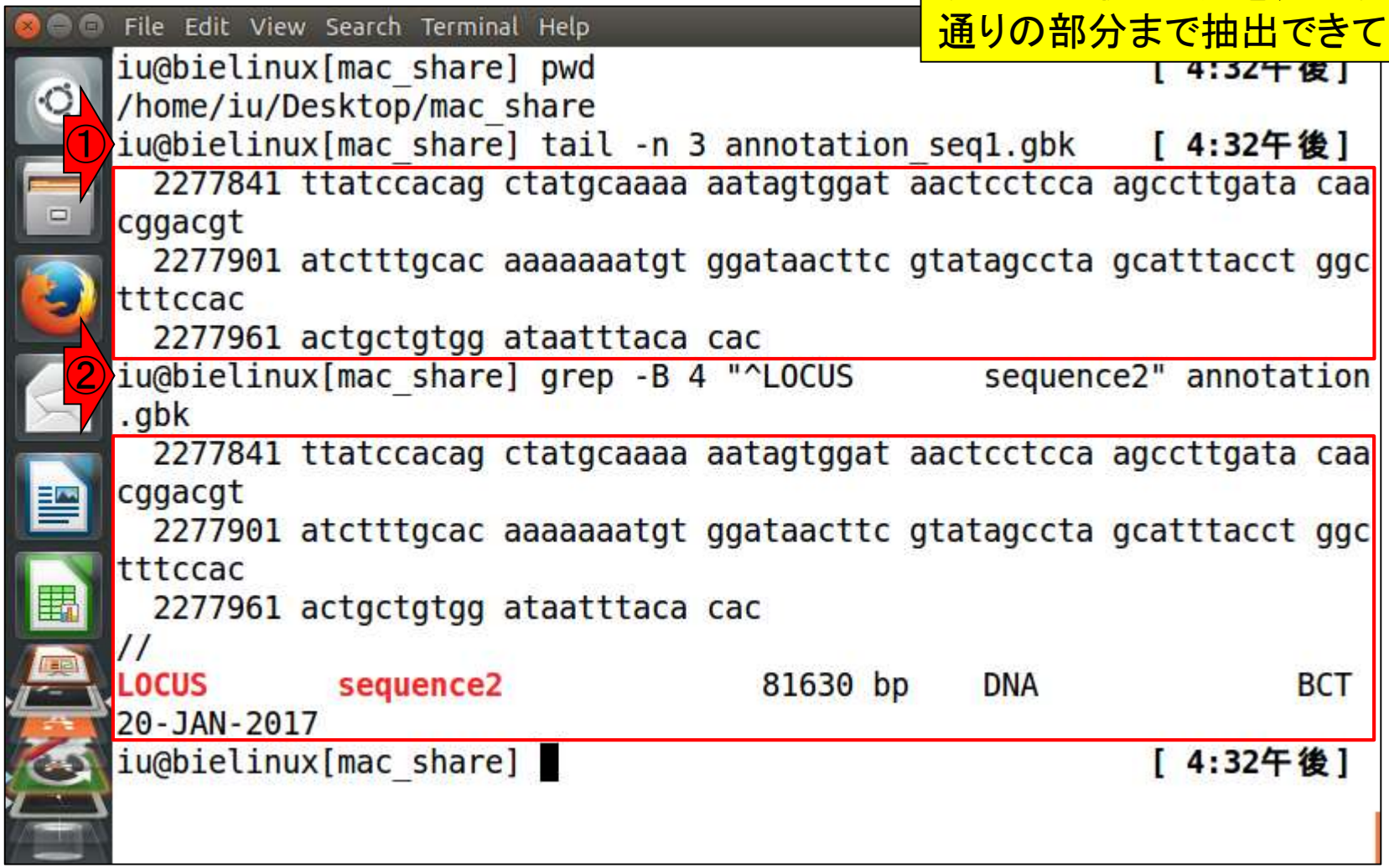
W15-1：DNAPlotter

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



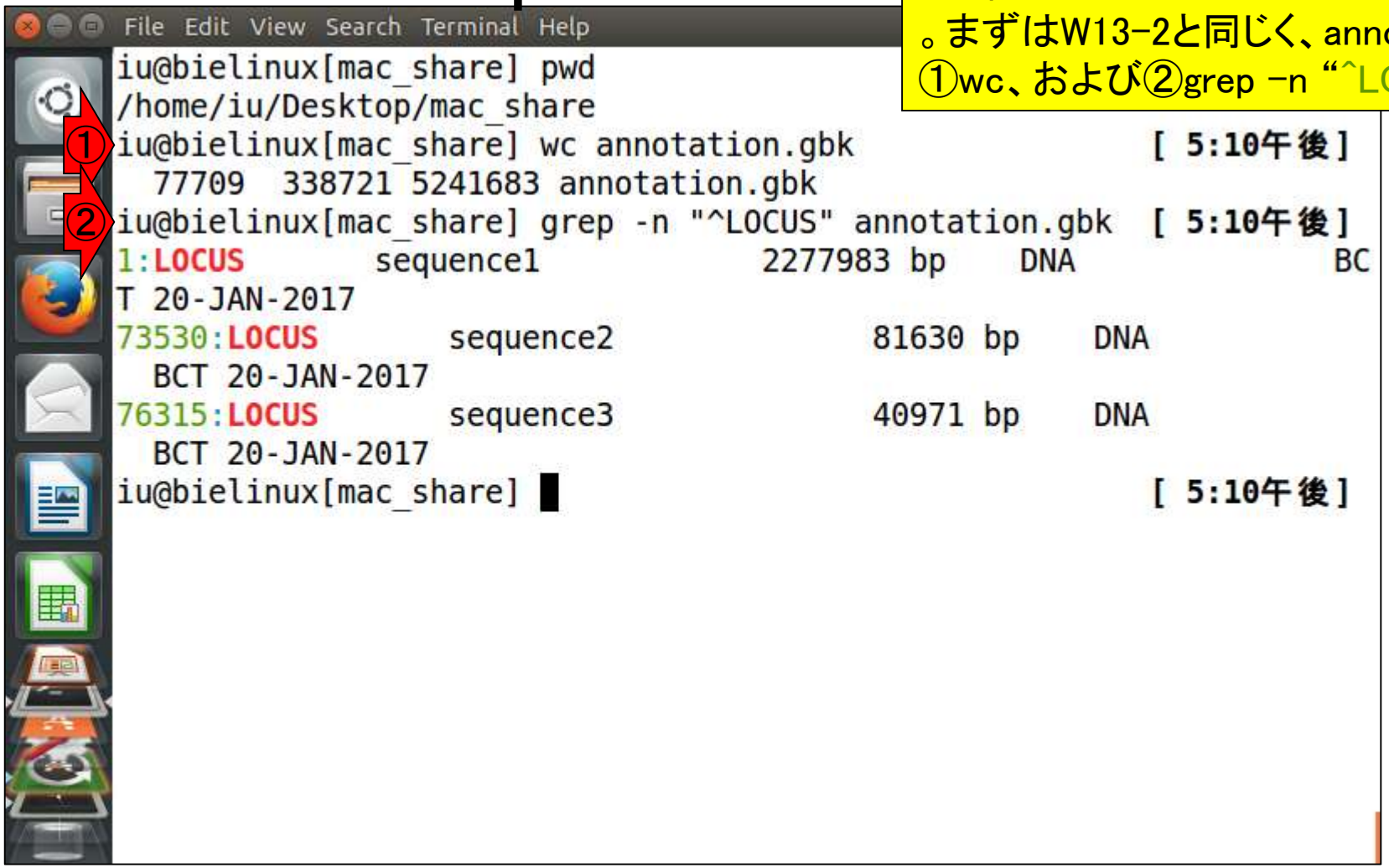
W15-2：tail

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



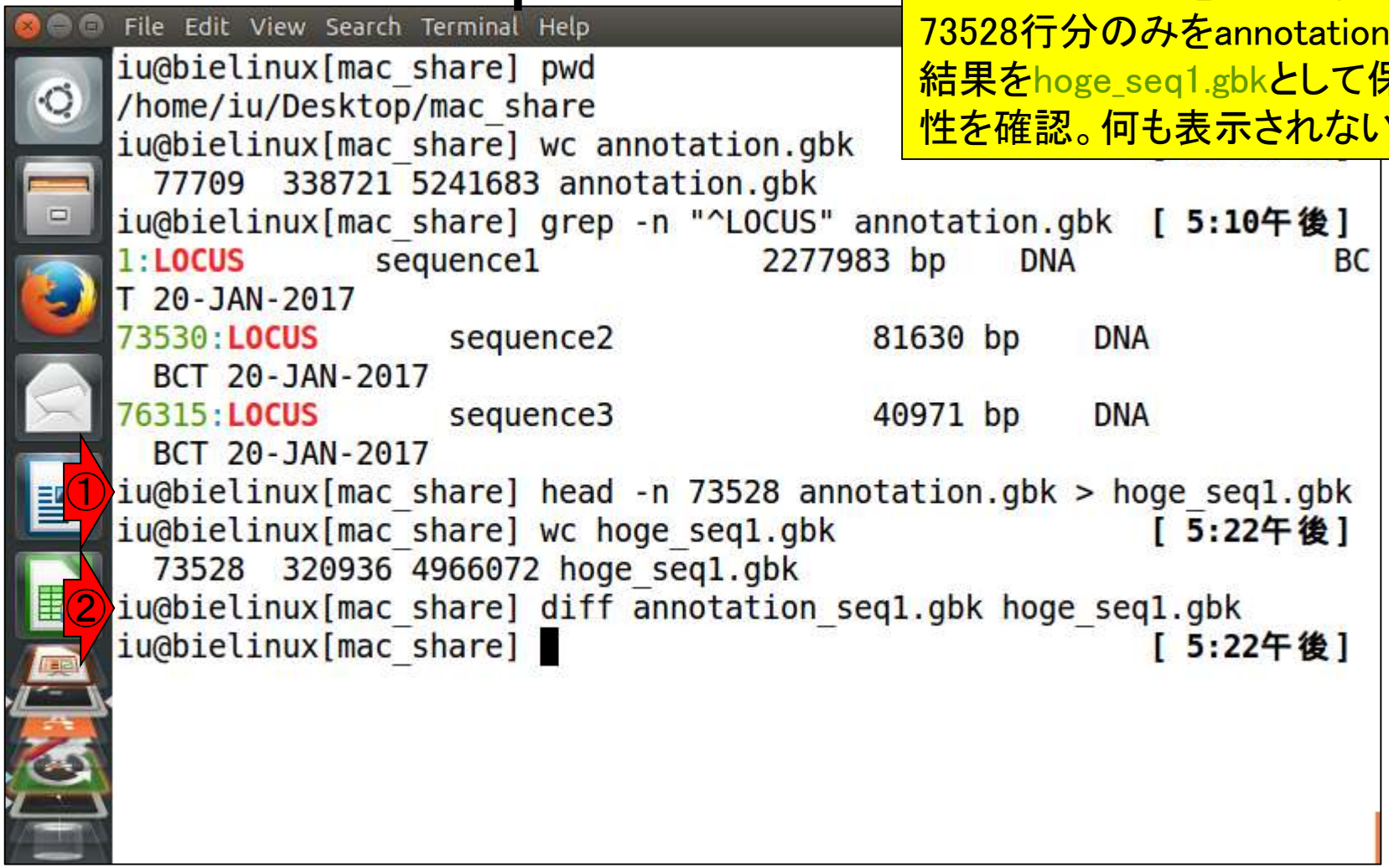
W16-1：sequence1

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



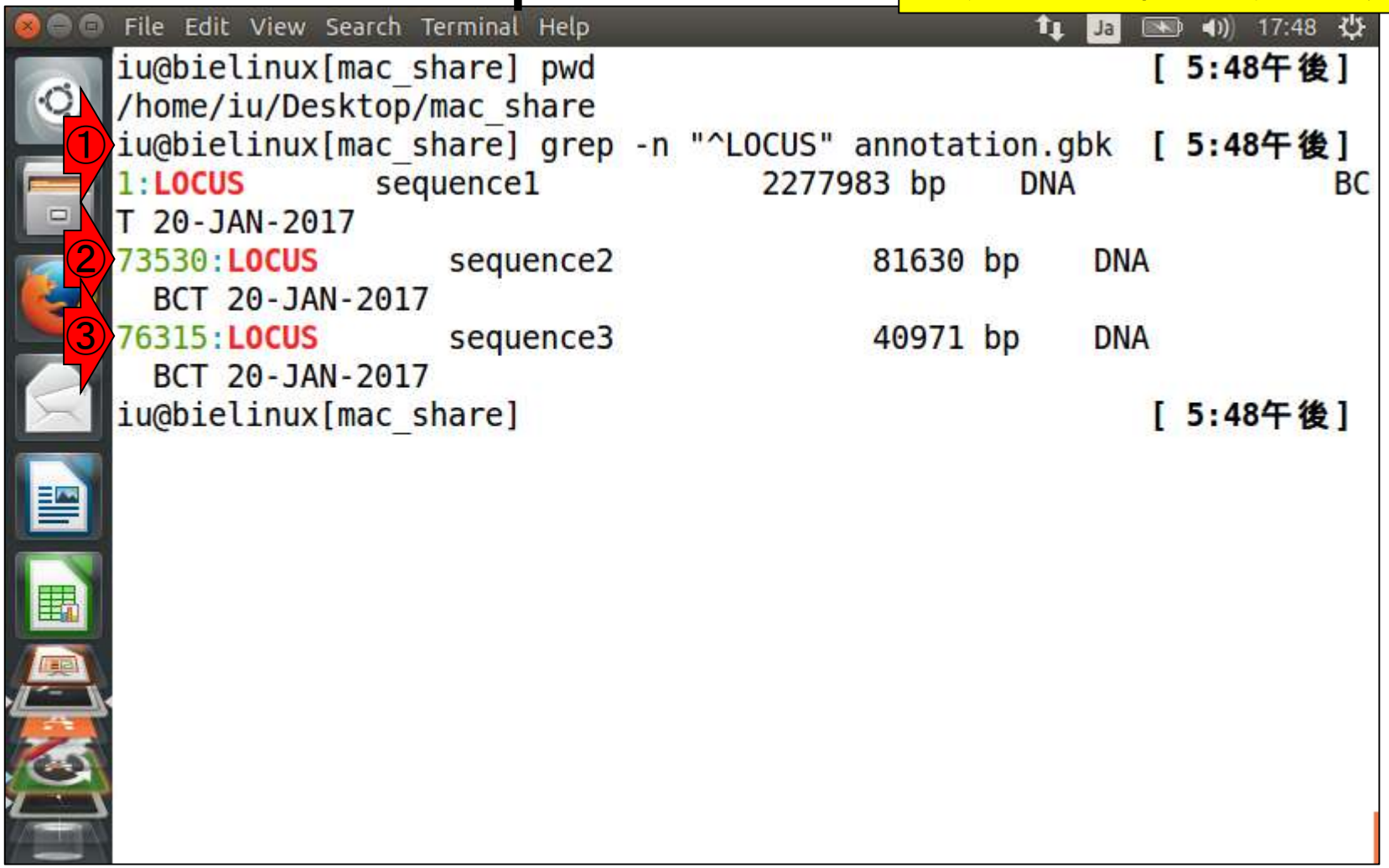
W16-1：sequence1

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。何も表示されないので同一である



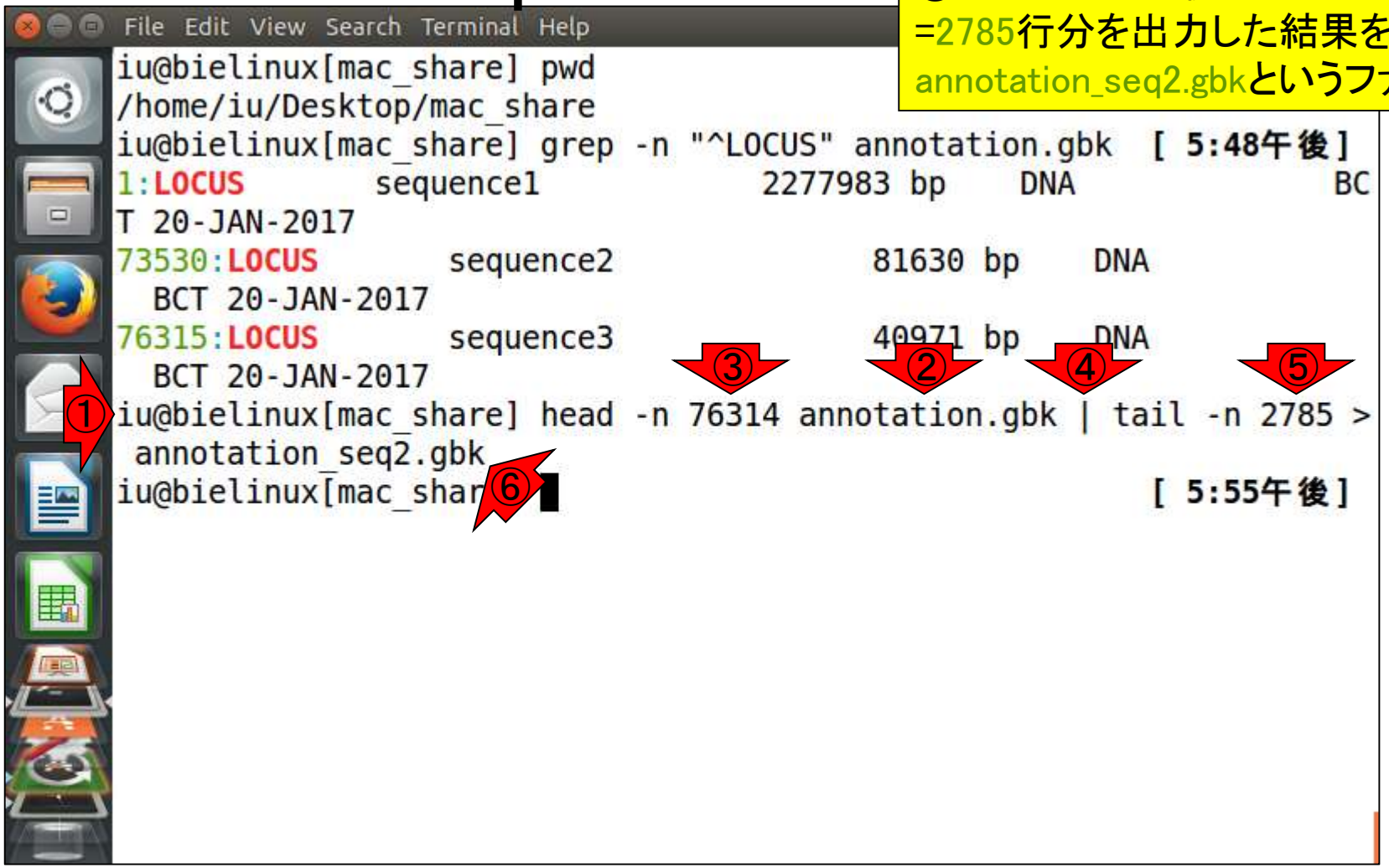
W16-2：sequence2

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



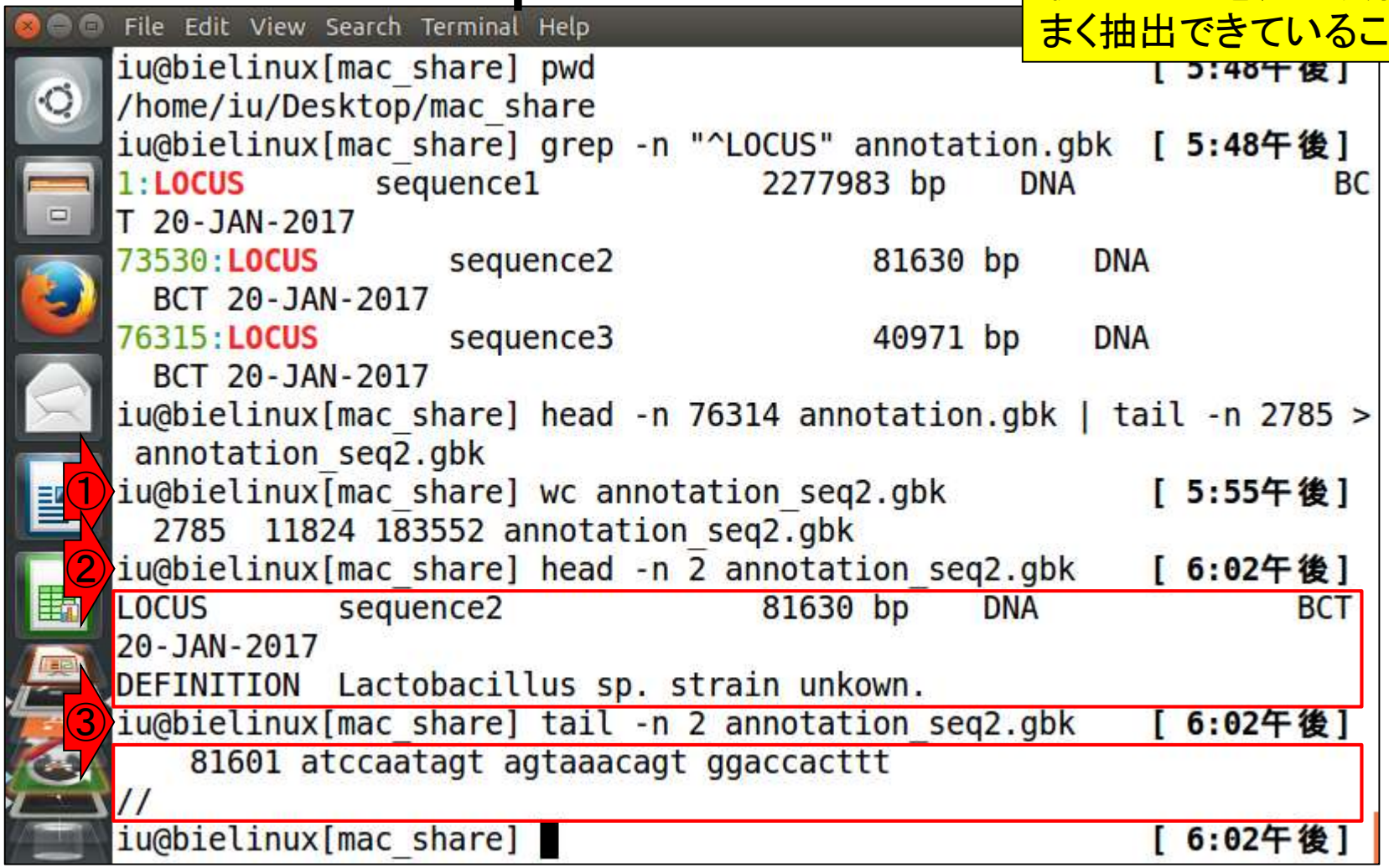
W16-2：sequence2

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



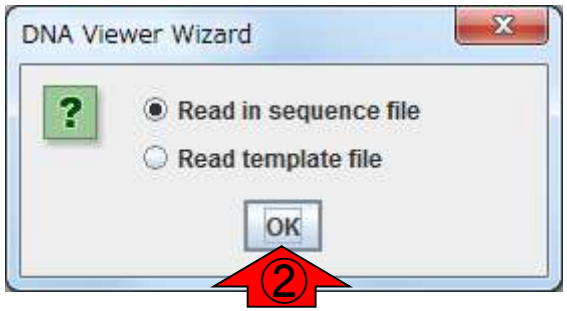
W16-2：sequence2

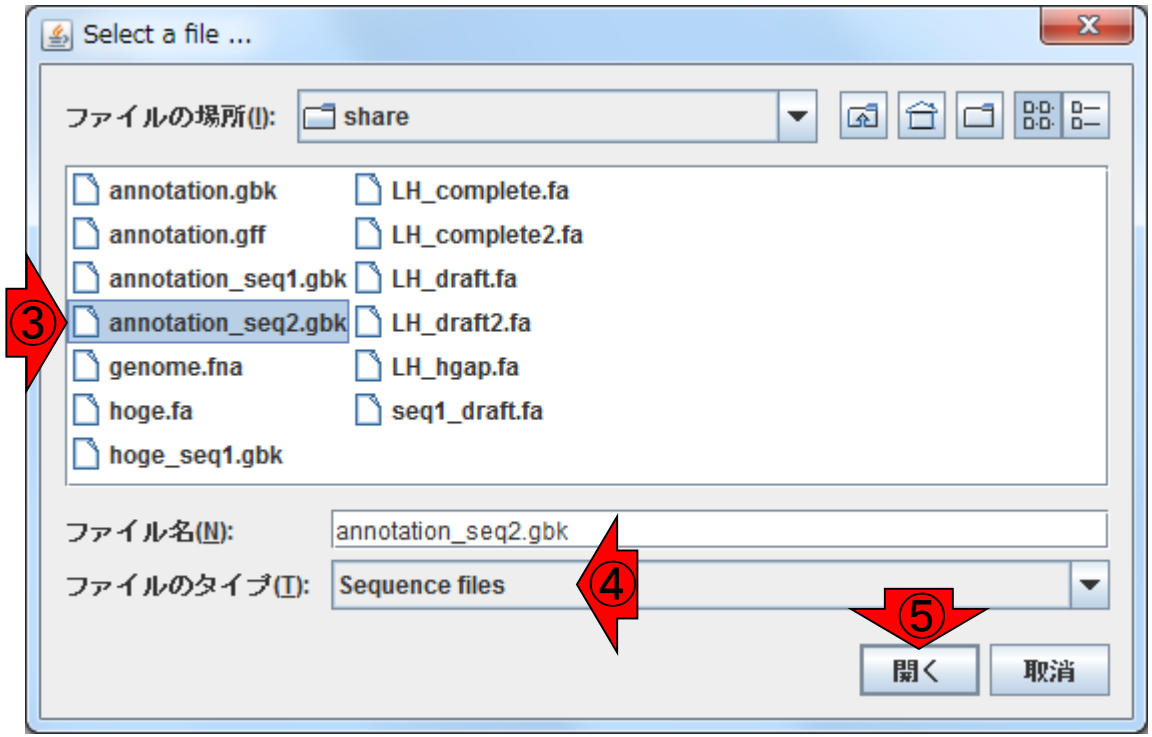
このあたりはただの確認です。得られたannotation_seq2.gbkの②最初の2行と③最後の2行を表示。赤枠部分を眺めてうまく抽出できていることを確認しています



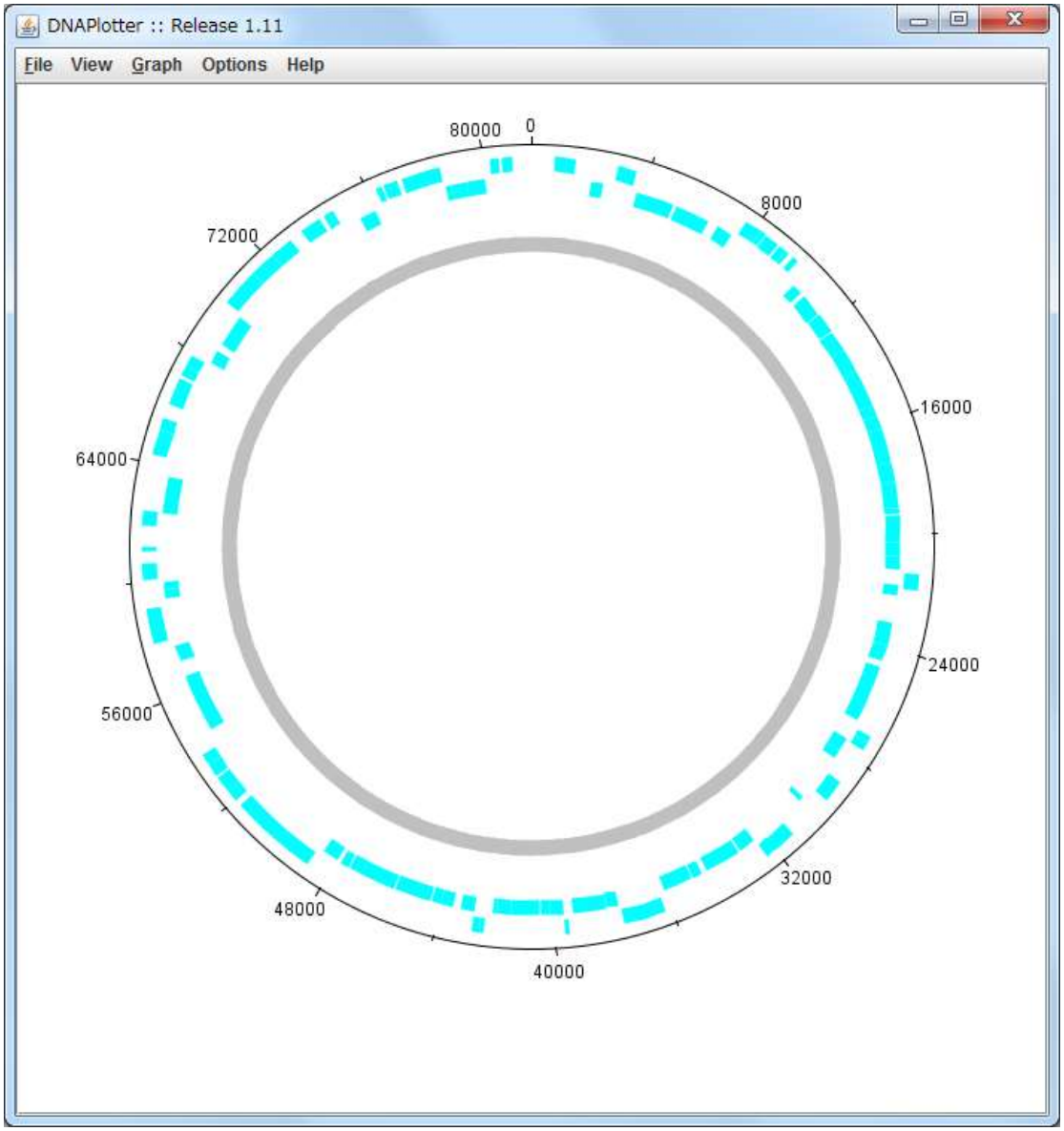
W16-3：DNAPlotter

①②DNAPlotterを起動して、作成した③annotation_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



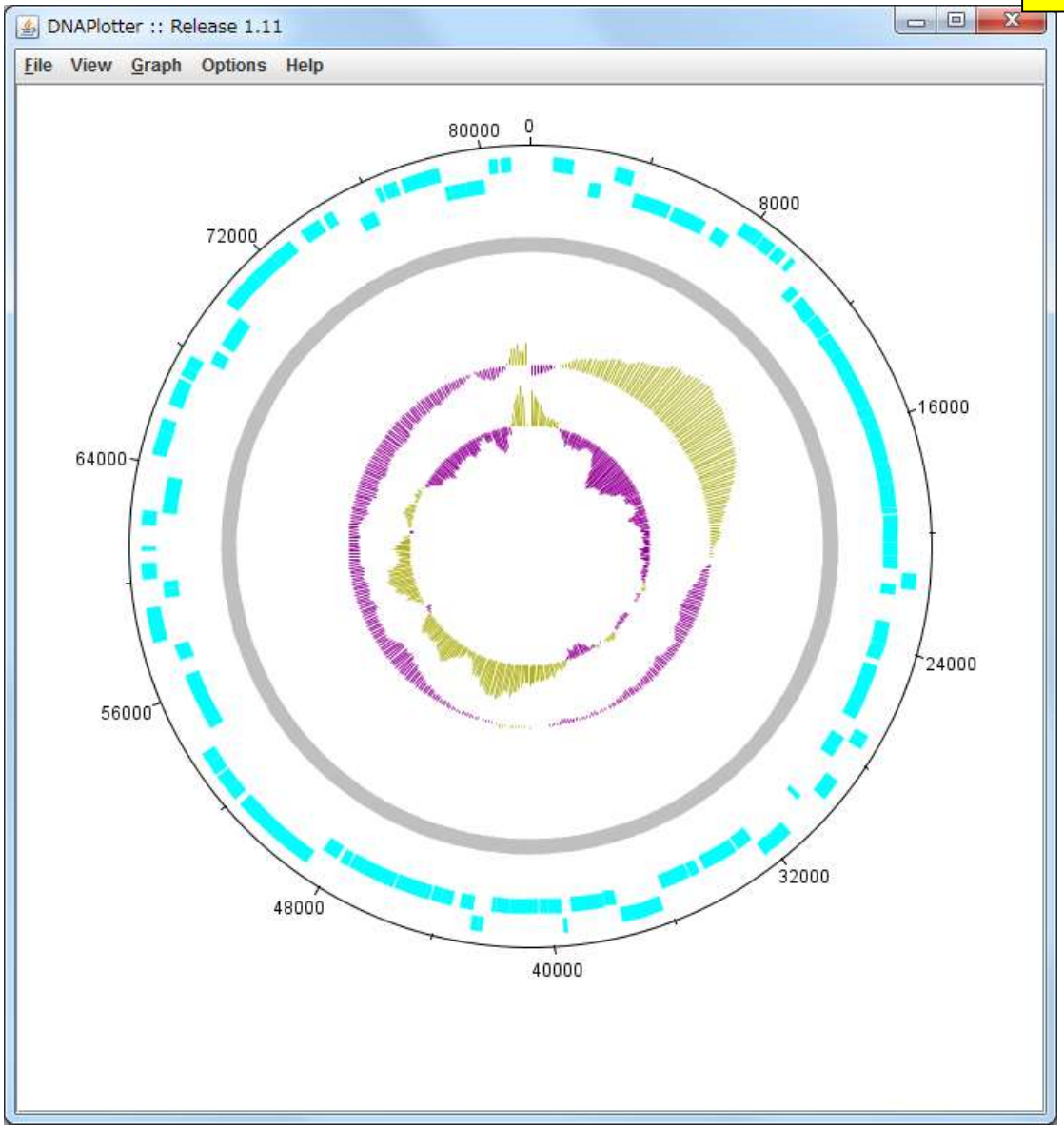


W16-3：DNAPlotter



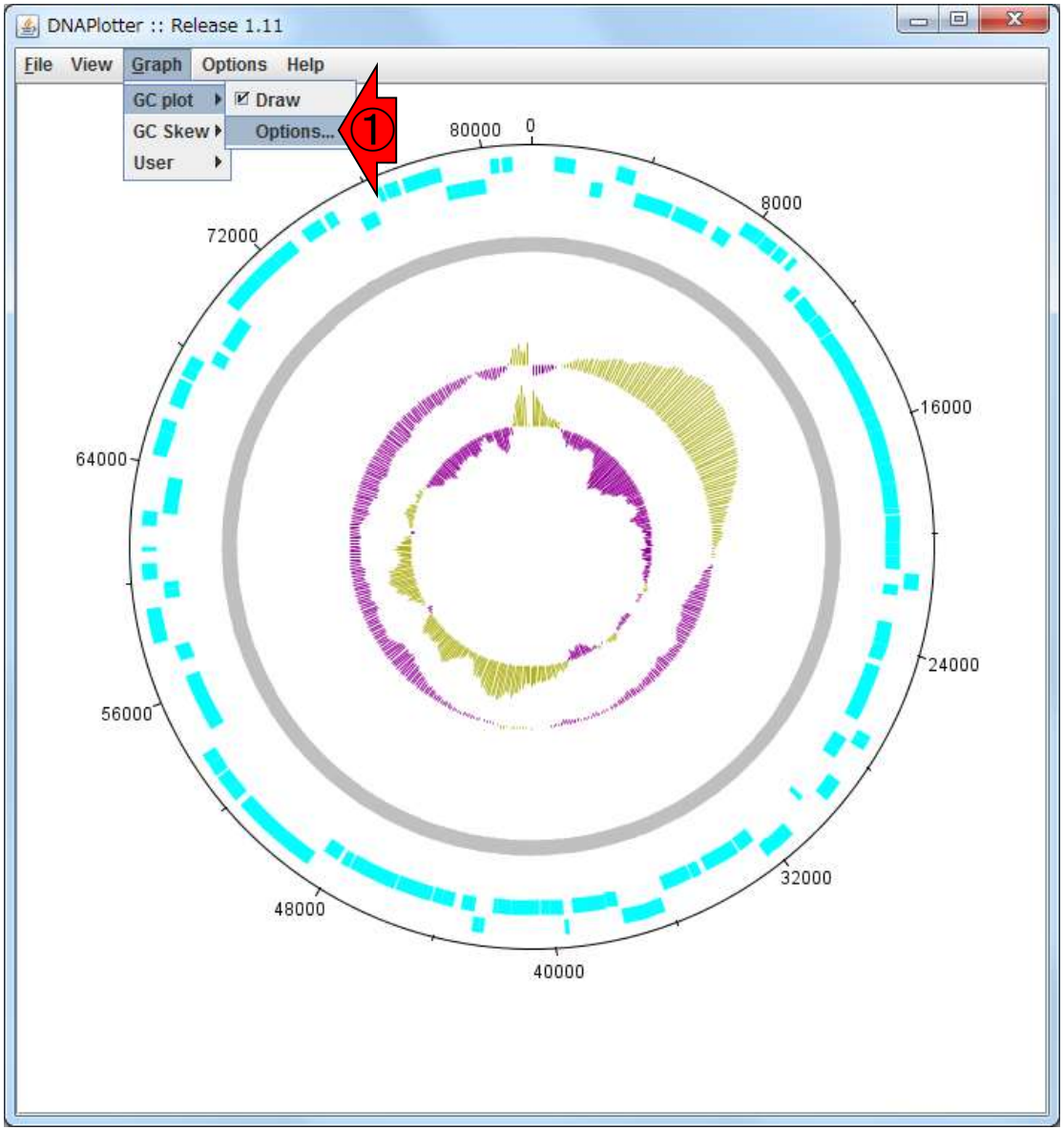
W16-3：DNAPlotter

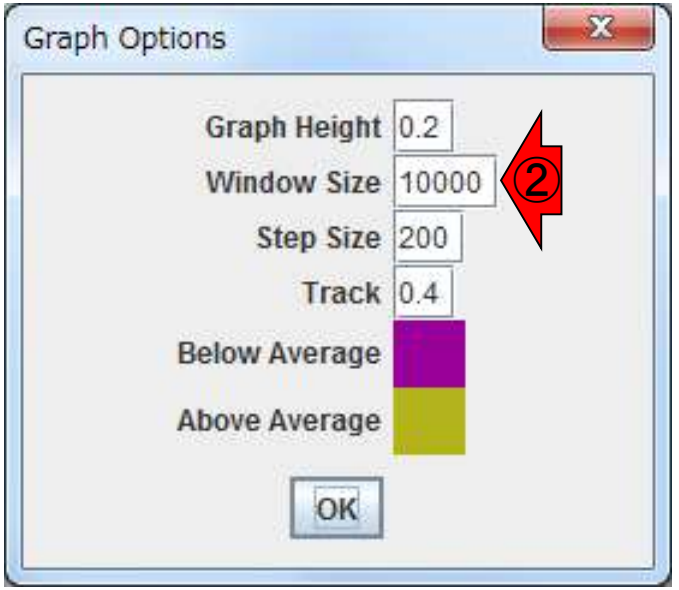
GC plotとGC skewを表示させた結果。ちゃんと表示できたことに最初は感動したが…、あまりに平滑化しすぎてるんじゃないかという疑念を抱く



W16-3：DNAPlotter

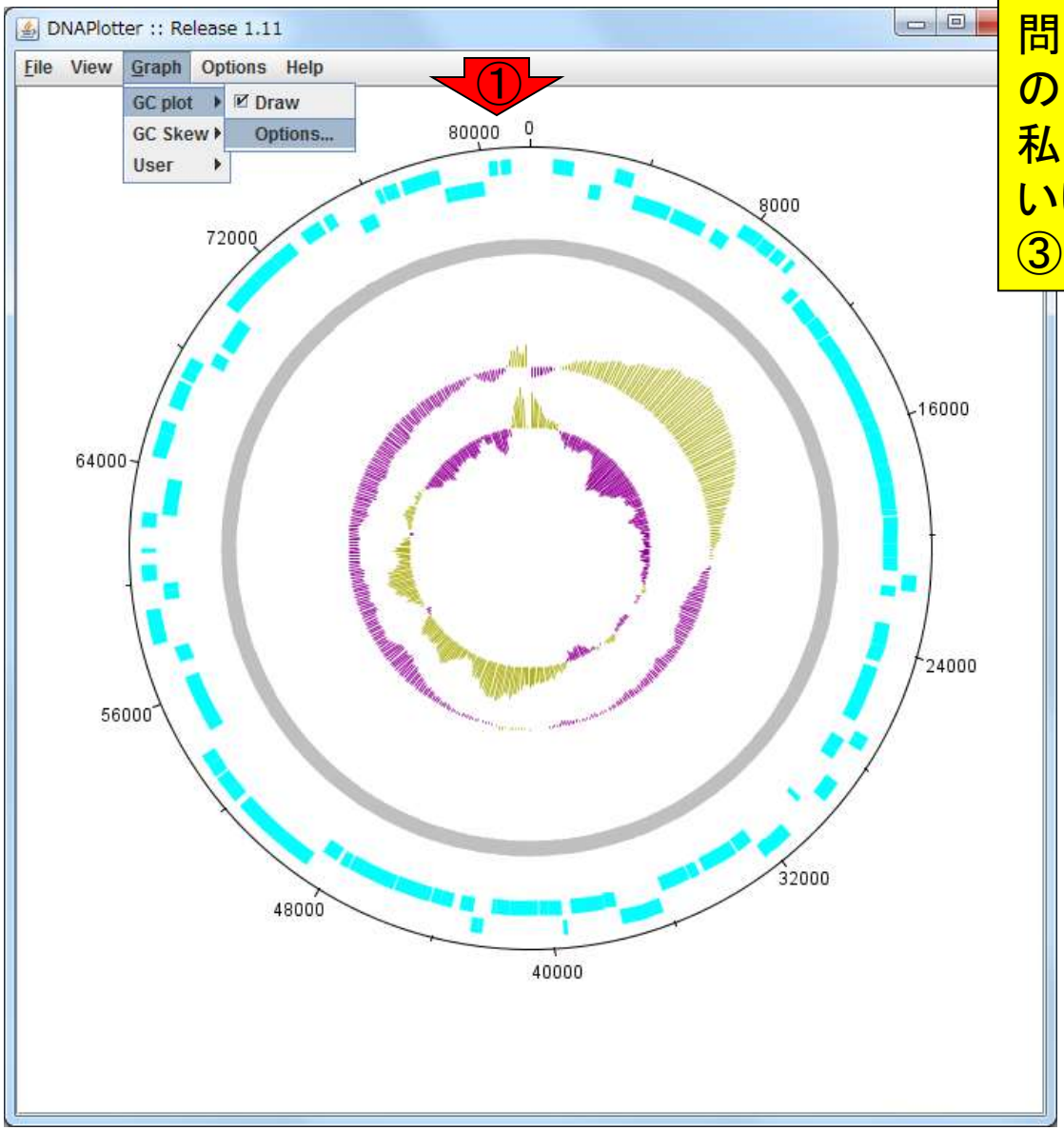
①Graph – GC plot – Options…。②
Window Sizeは、Window（つまり幅）を
10000塩基とって、その範囲の平均のGC
含量を計算せよという指定だった(W12-15)

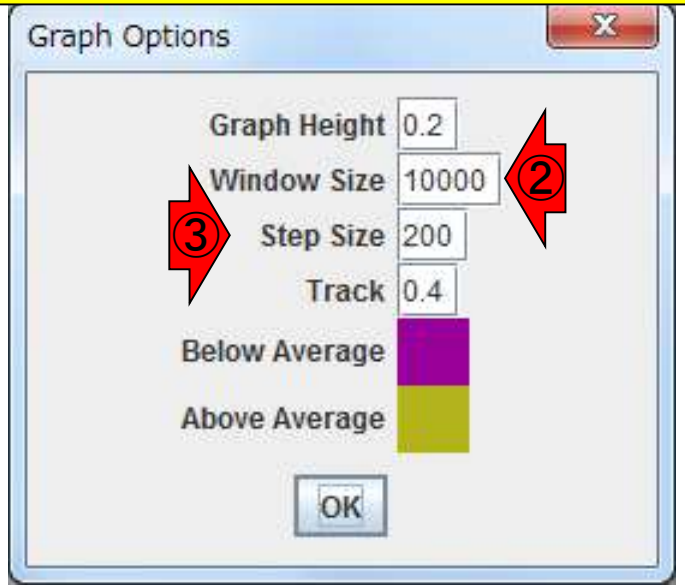




W16-3：DNAPlotter

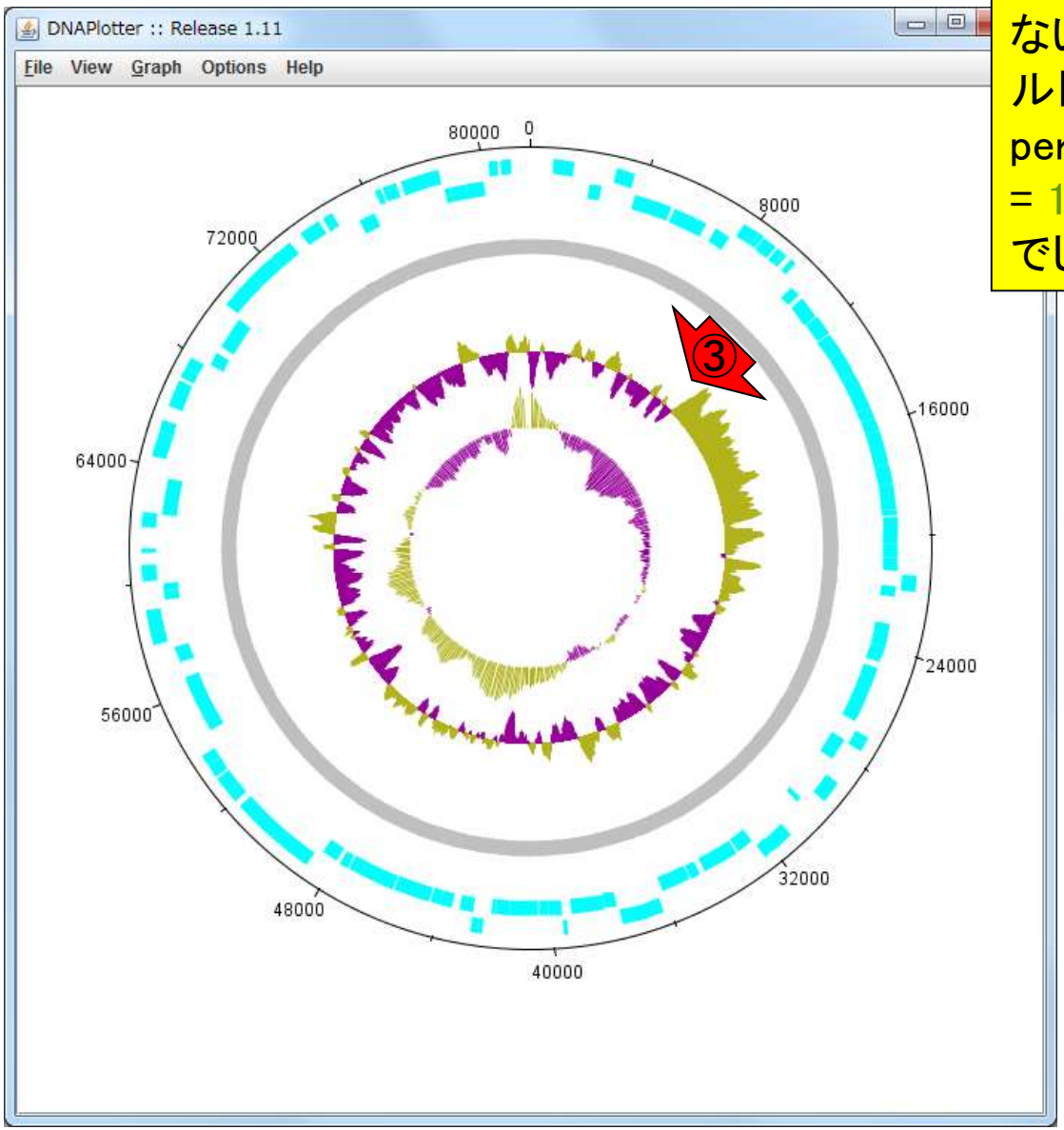
①しかしsequence2は配列長が81,630 bpの小さい環状プラスミド。2,277,983 bpの染色体ゲノムのときであれば②Window Size = 10000でも特に問題なかったであろうが、配列長8万ちょっとのものに対しては明らかに大きすぎる値（というのが私の常識的な感覚）。1000とか500とかのほうがいいんじゃないだろうか…。またそれに伴って、③Step Sizeも20とか10くらいですかねえ…

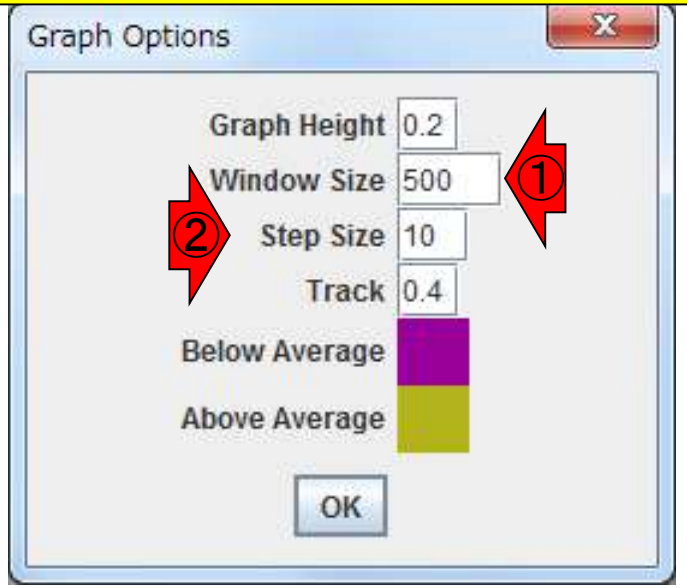




W16-3：DNAPlotter

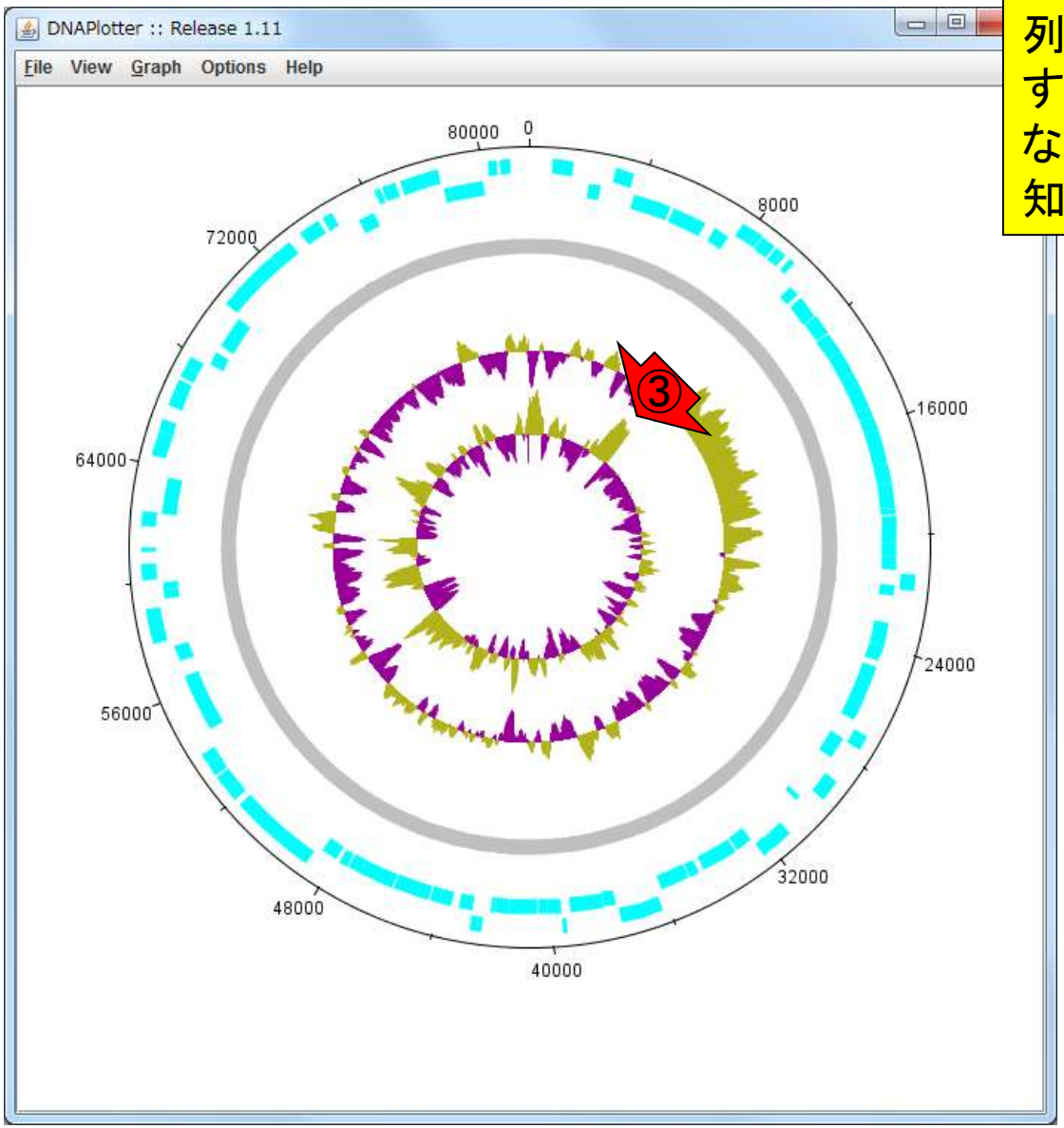
①Window Size = 500、②Step Size = 10で再描画した結果。特に③のあたりがはっきりとGC含量分布の見栄えとして異なっているのがわかるのではないでしょうか。投稿論文中で示すときは、デフォルトの数値から変えたときは「DNAPlotter was performed with Window Size = 500 and Step Size = 10 for GC plot」みたいな感じで書いておくべきでしょう（英語はテキトーです）。

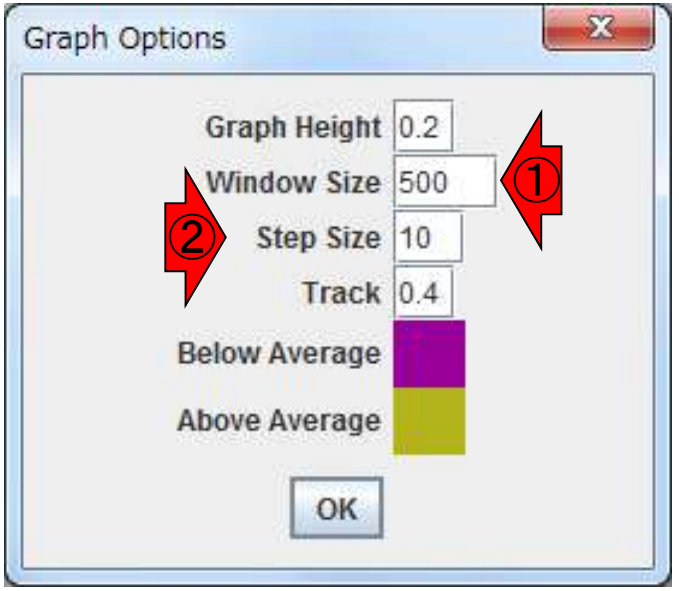




W16-3：DNAPlotter

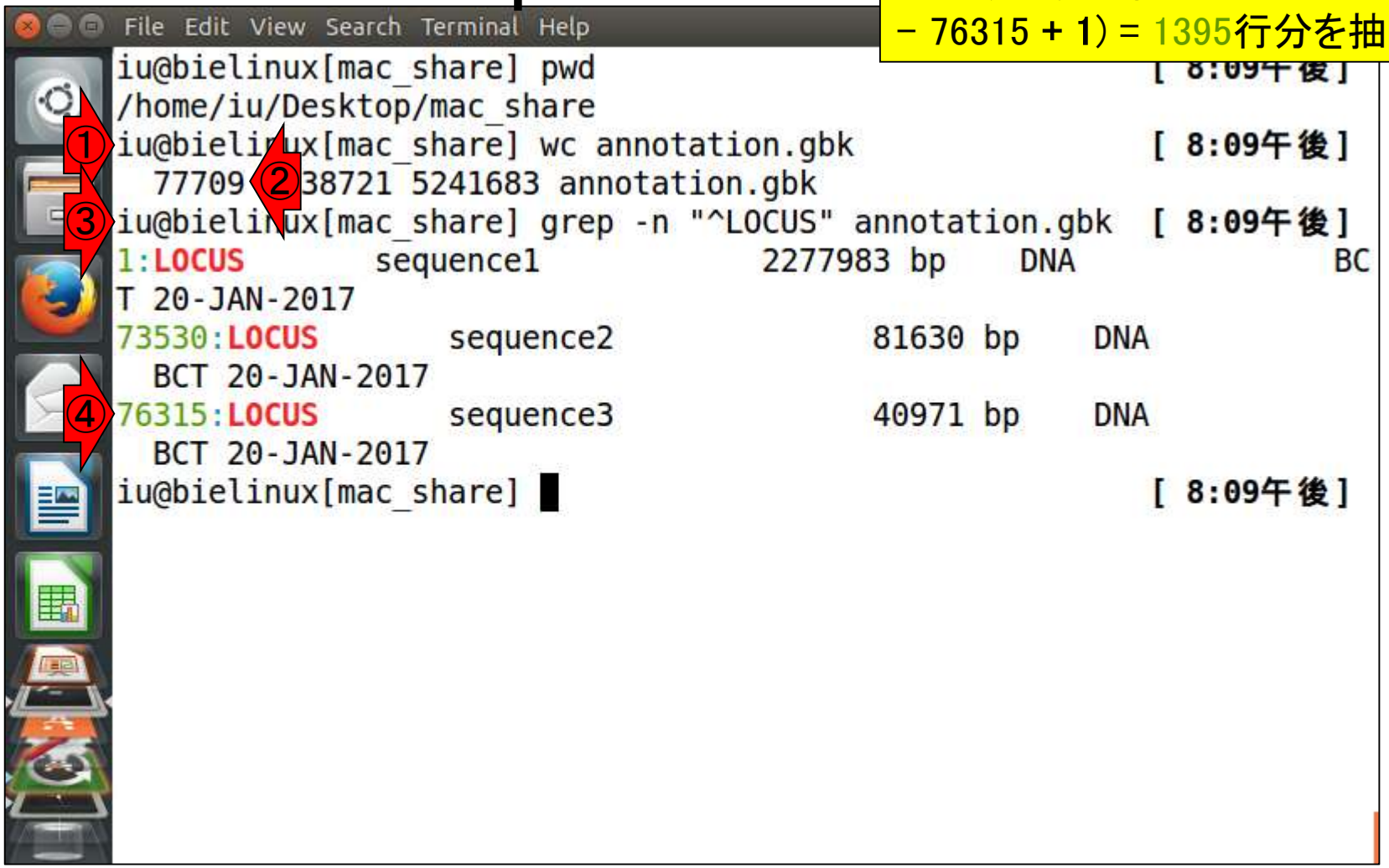
GC skewについても、①Windows Size = 500、②Step Size = 10で再描画。例えば③のあたりの見栄えが変わっているのがわかります。全体の配列長とWindow Size（およびそれに関連して変更するStep Size）の関係についてのガイドライン的なものはどこかにあるのかもしれませんが、私は知りません





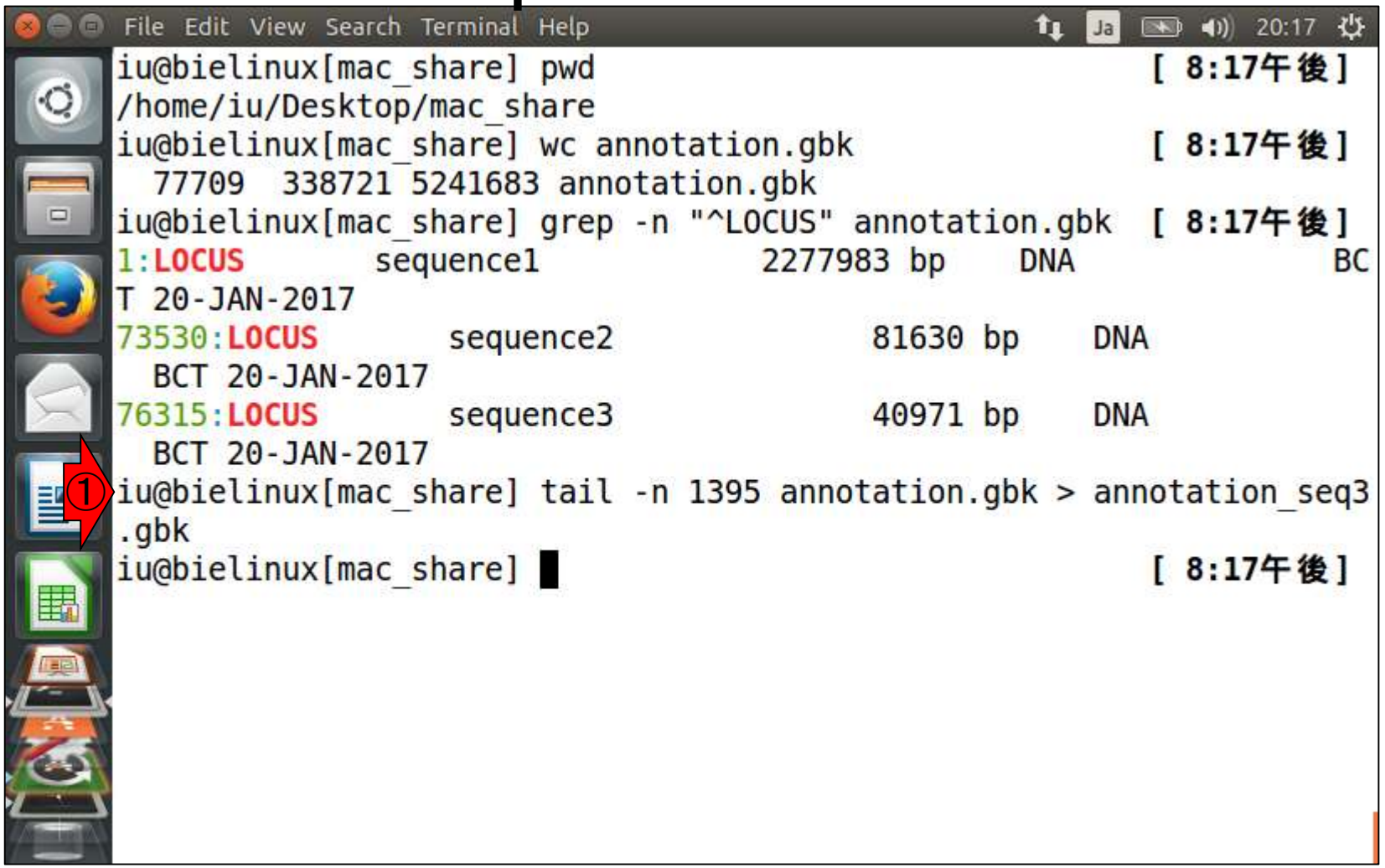
W16-4：sequence3

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



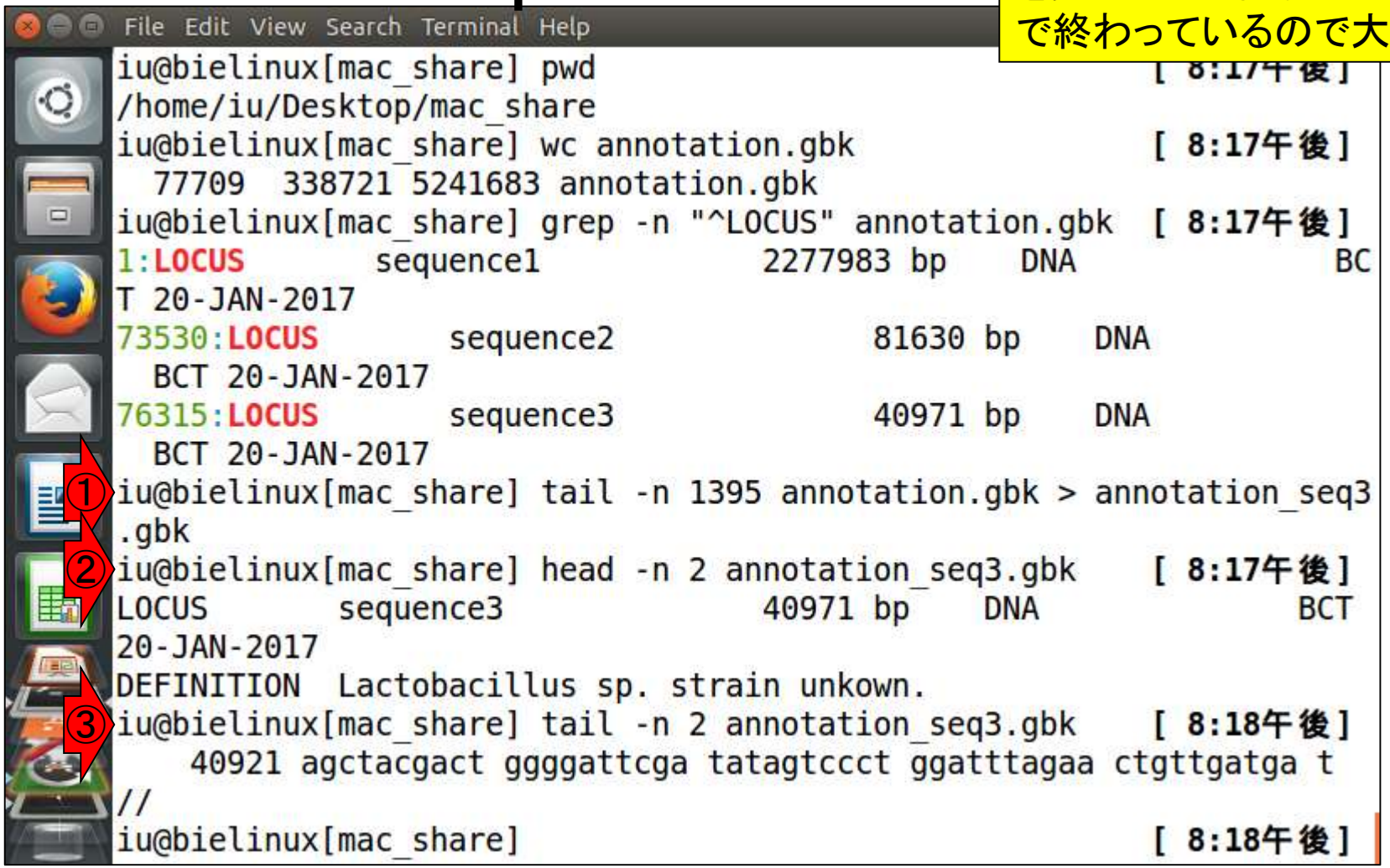
①こんな感じです

W16-4：sequence3



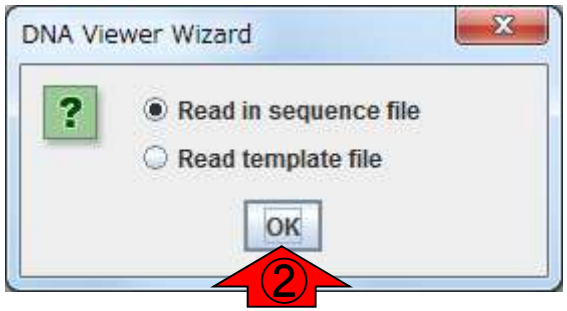
W16-4：sequence3

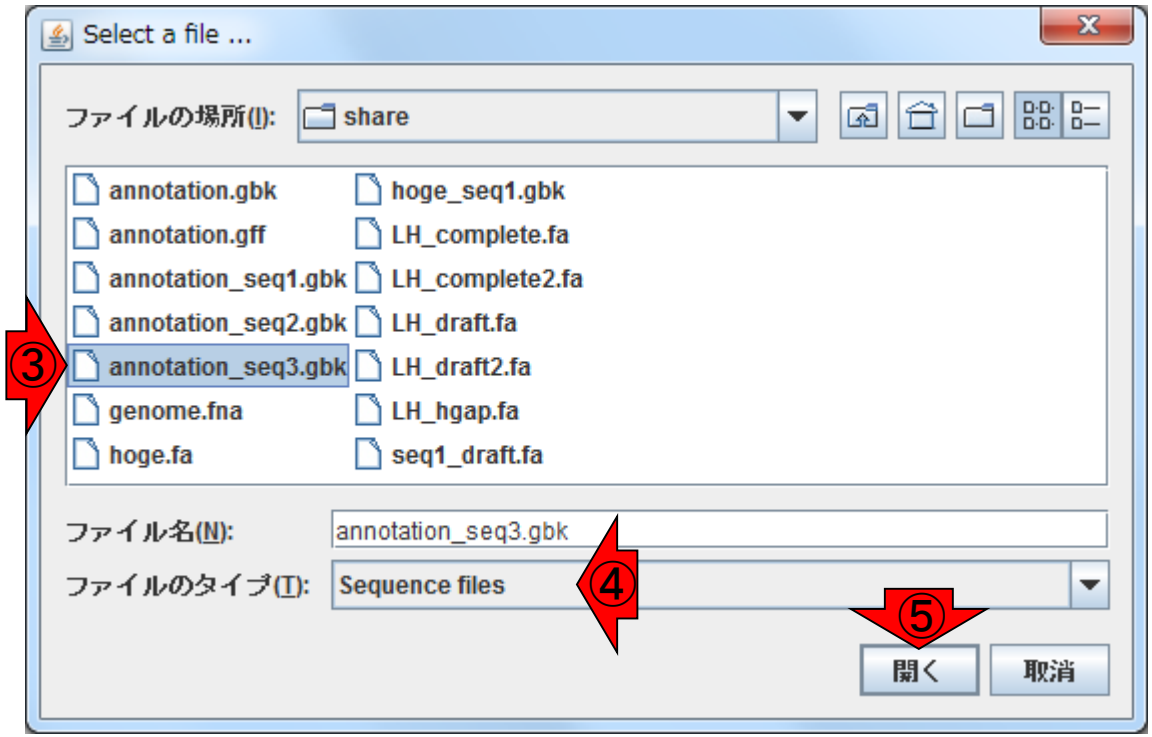
一応得られたannotation_seq3.gbkに対して、②headと③tailで最初と最後の2行分を実行して確認。LOCUSから始まって、//で終わっているので大丈夫そうだと判断



W16-5：DNAPlotter

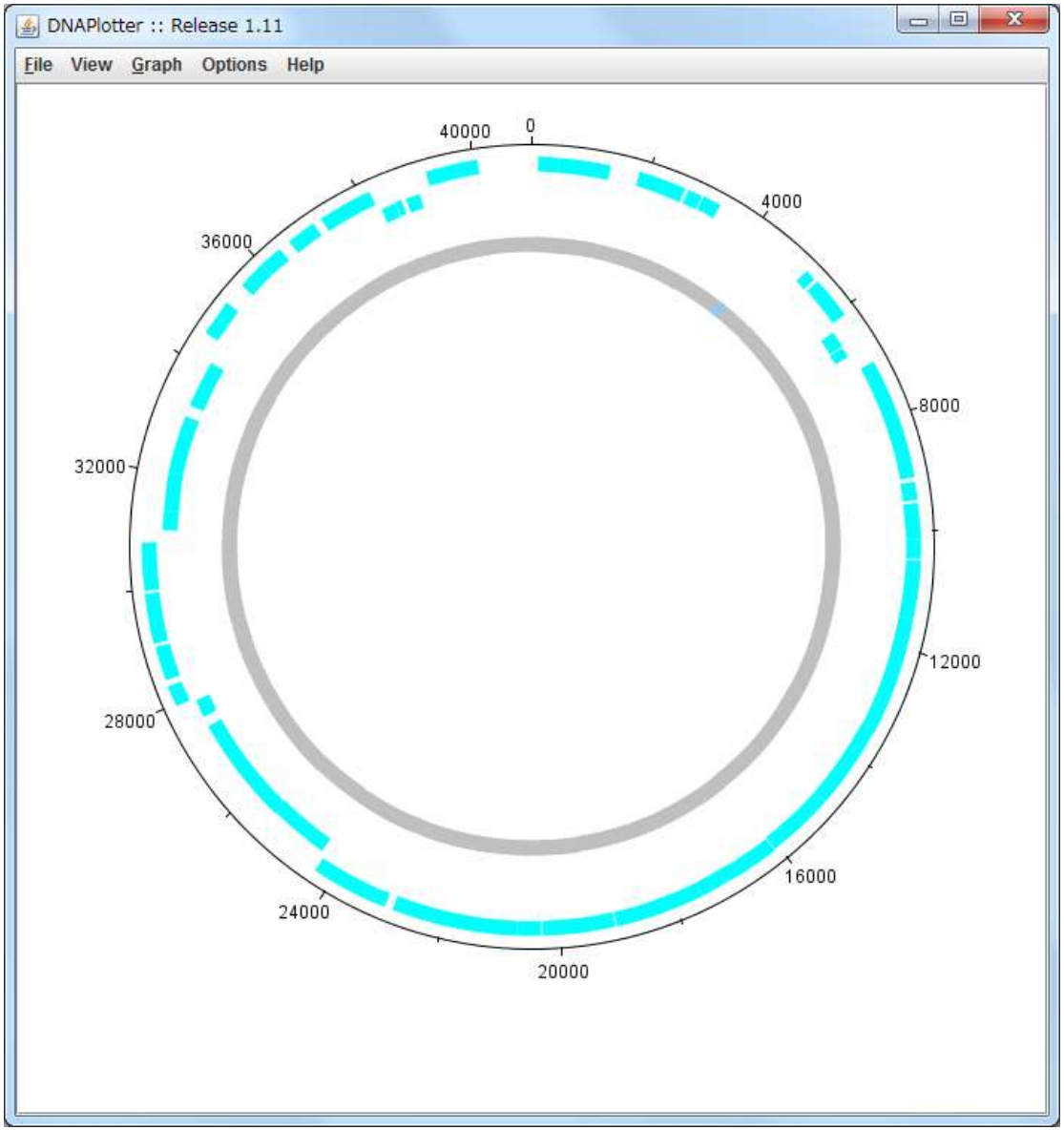
①②DNAPlotterを起動して、作成した③annotation_seq3.gbkを、④ファイルのタイプはSequence filesのままでよいので、⑤開く



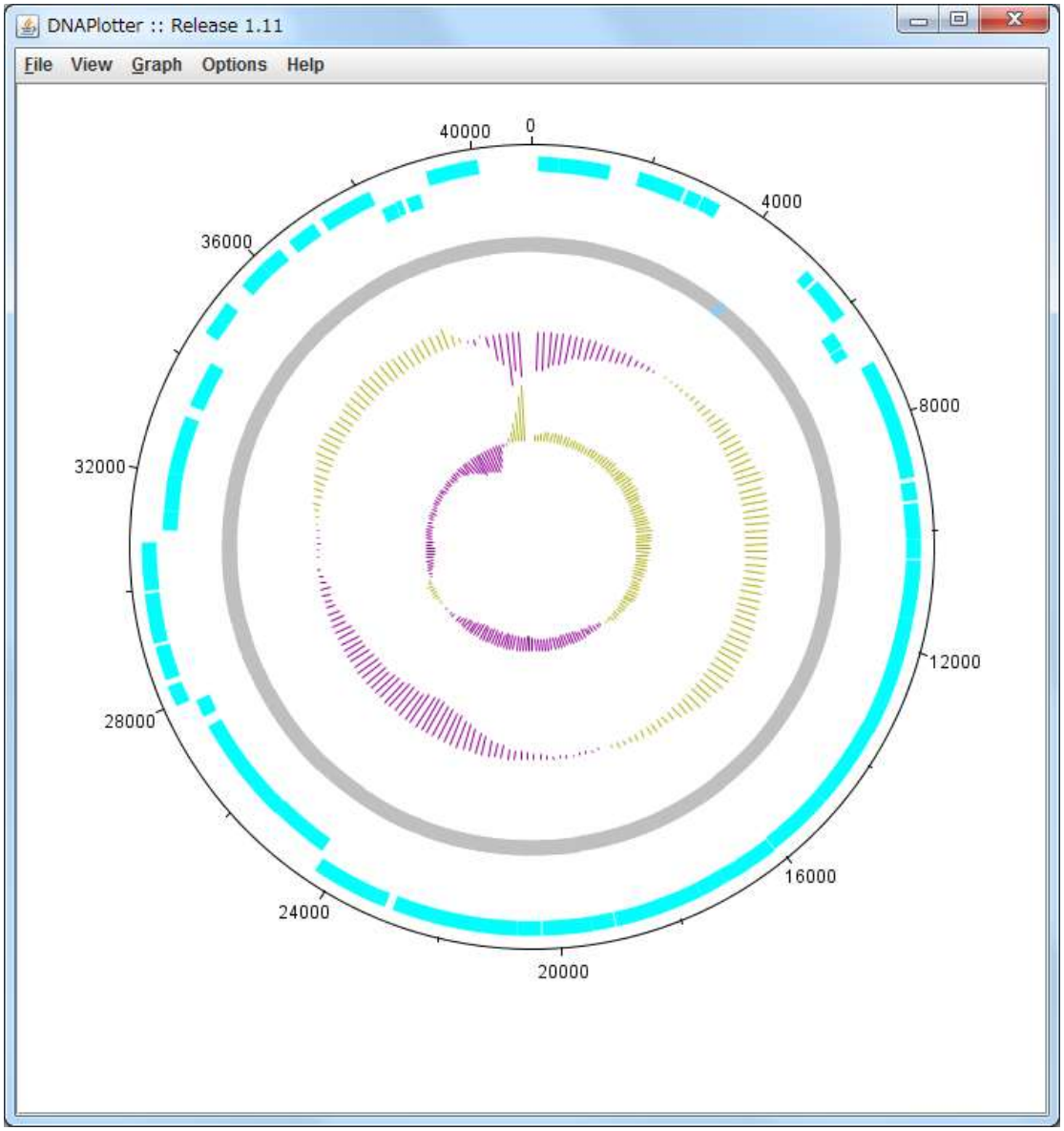


W16-5：DNAPlotter

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

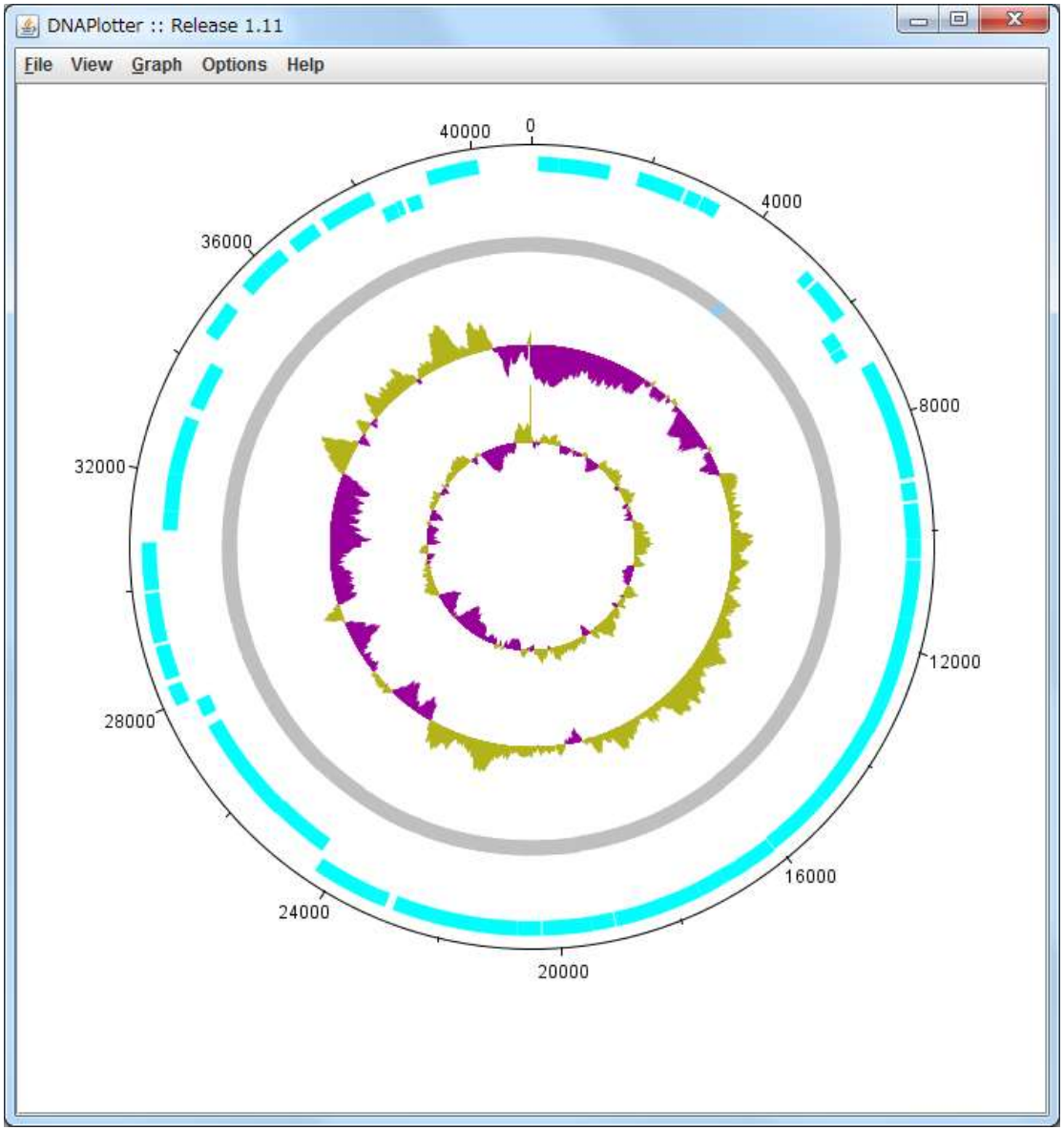


W16-5：DNAPlotter



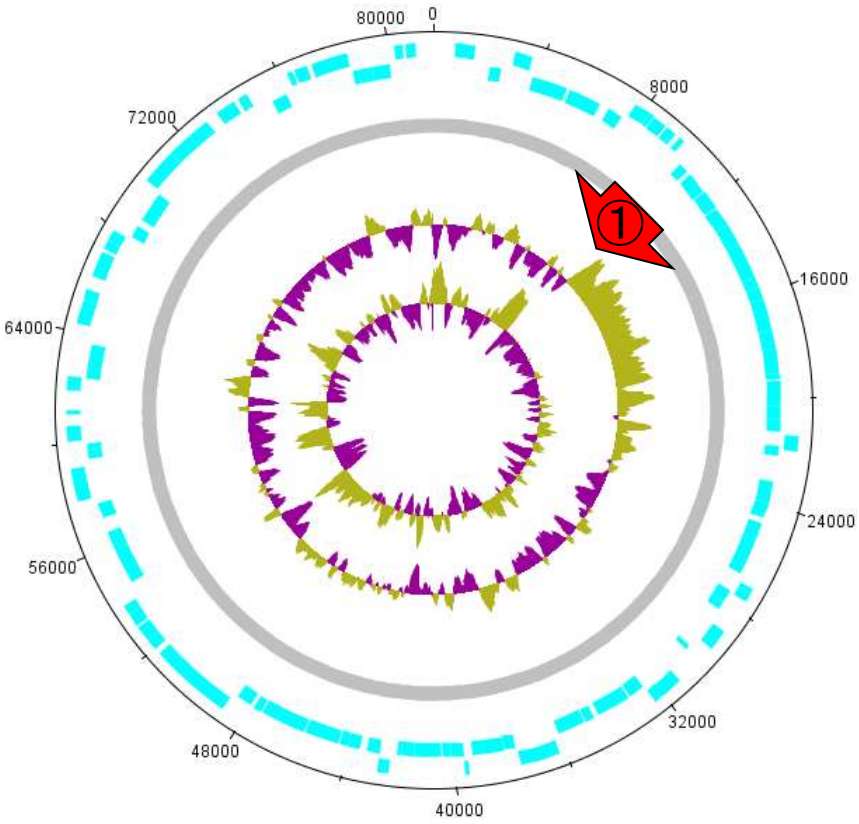
GC plotとGC skewを表示させた結果。デフォルトのWindow Size = 10000およびStep Size = 200のままなので、sequence2のときと同じくWindow Size = 500、②Step Size = 10にする

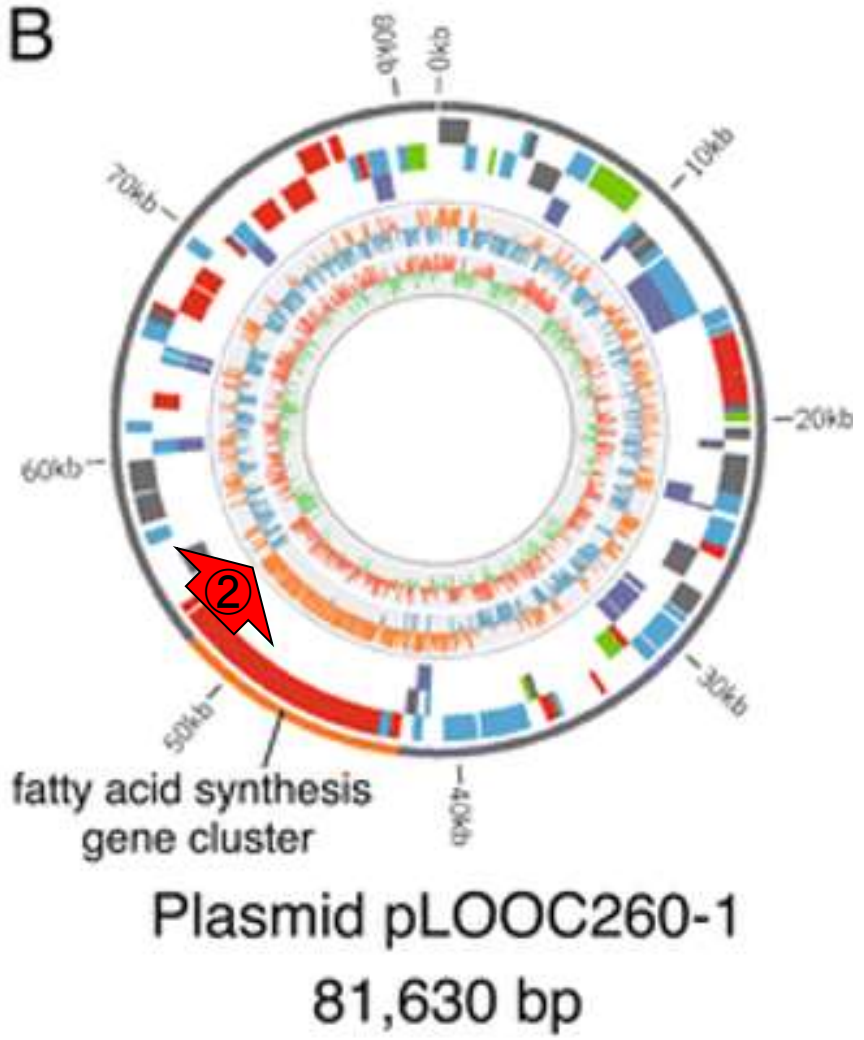
W16-5：DNAPlotter



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ね。配列長は同じ

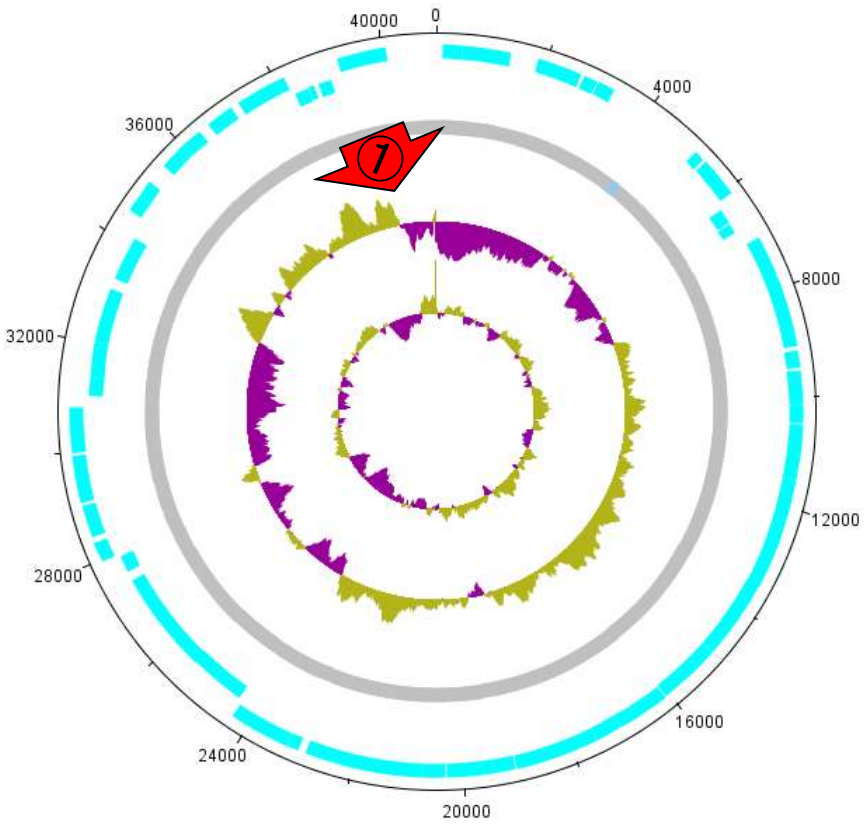


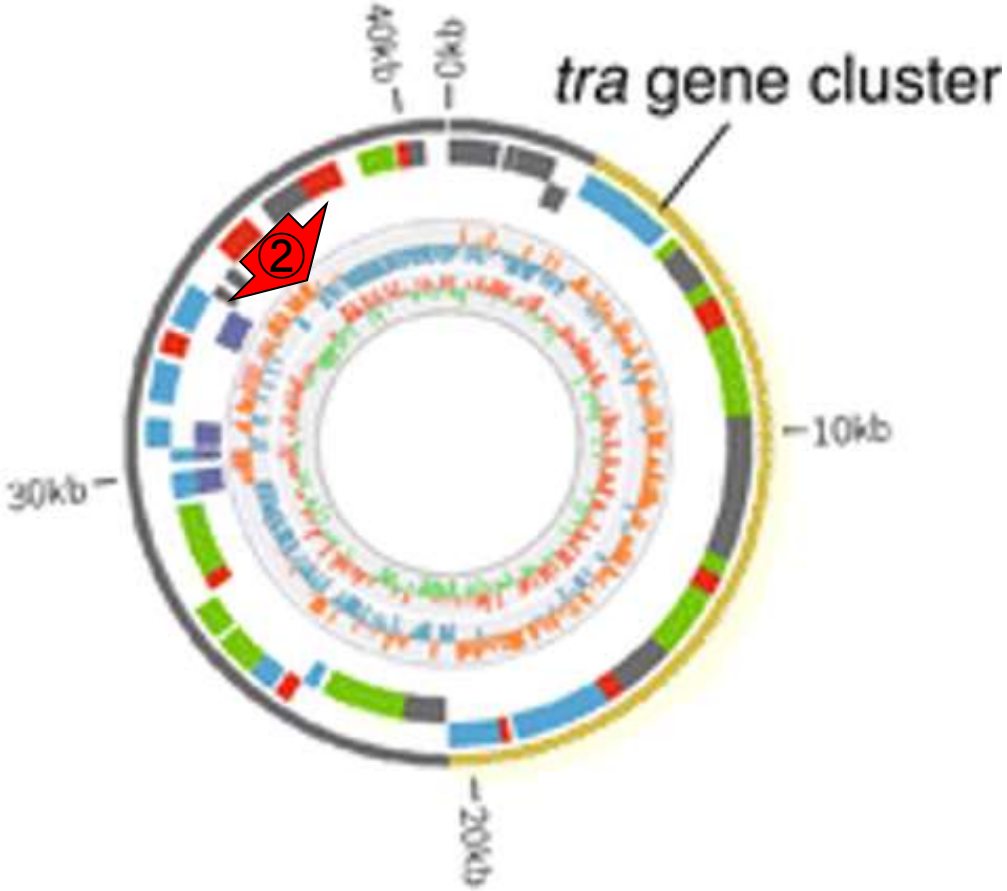


Tanizawa et al., *BMC Genomics*, **16**: 240, 2015

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





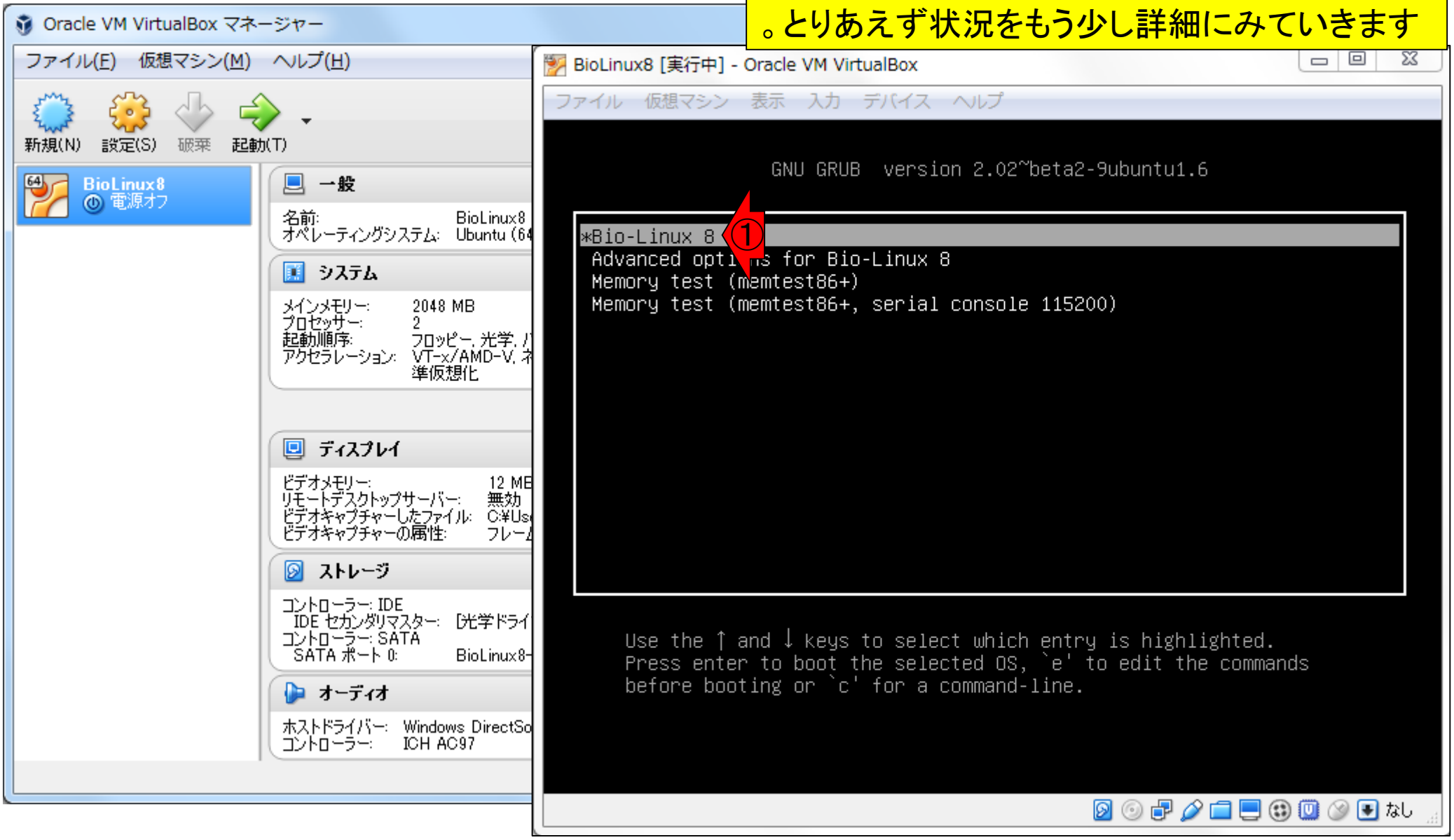
Plasmid pLOOC260-2

40,971 bp

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

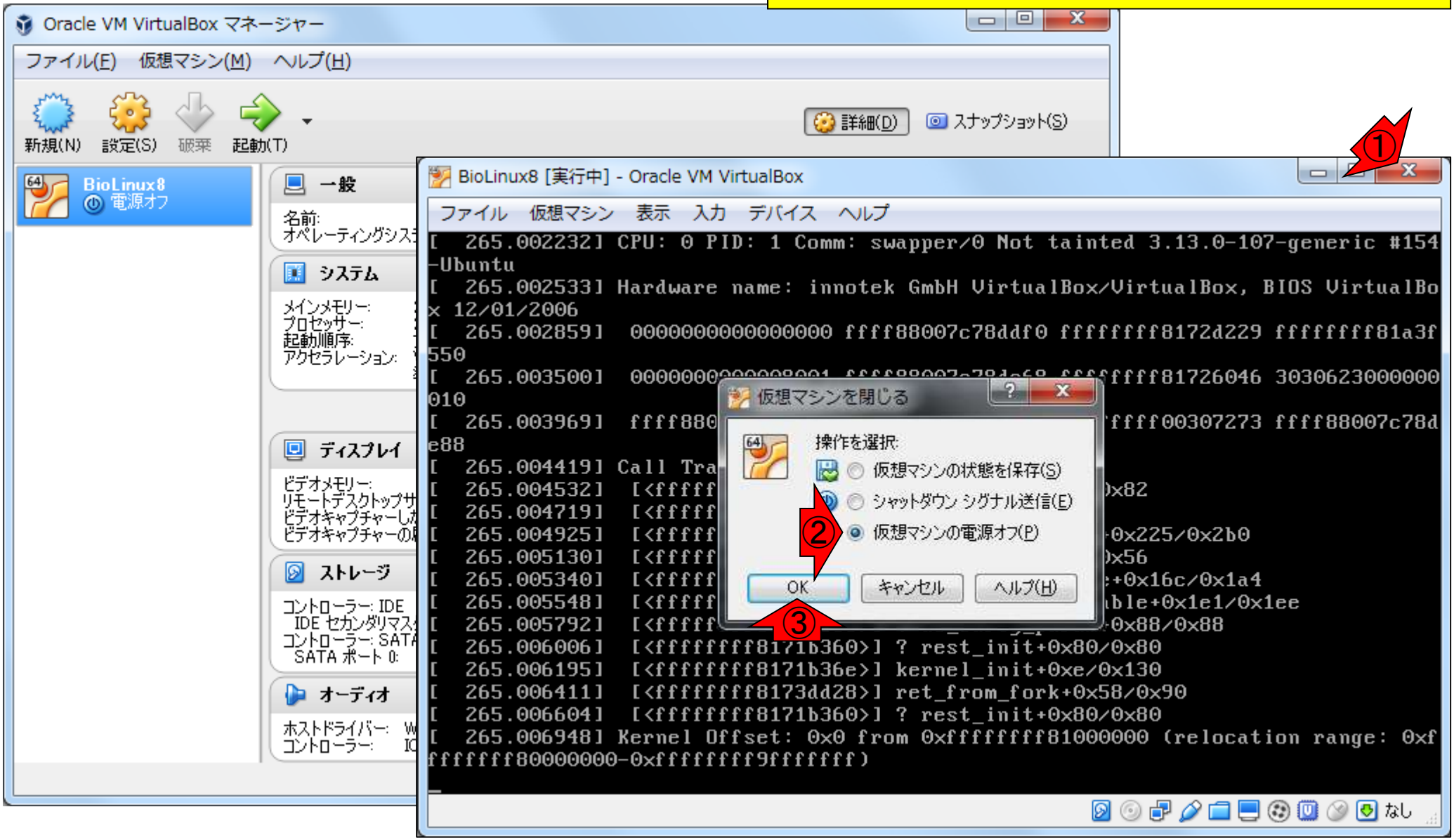
W18-1：調子が…

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



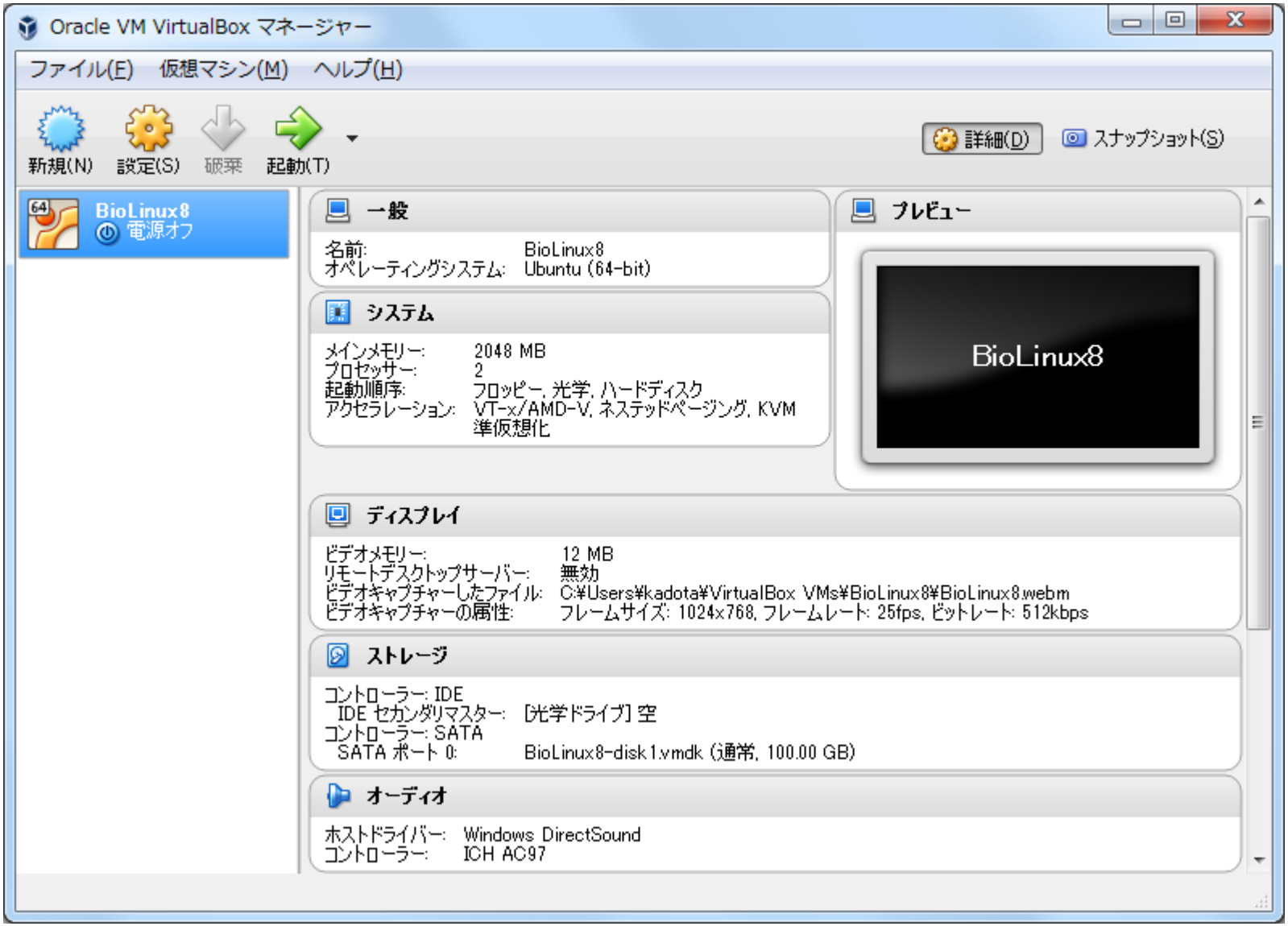
W18-1：調子が…

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



この状態から…

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われますが、第5回のリクエストしか来ませんw。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版に従ってコピーしてRと必要なパッケージを インストール済みであるという前提で記述 しています。初心者の い。本ウェブページを

What's new?

- 「書籍 | 日本乳酸菌 た。「検索文字列の ですm(_ _)m(2017/

- 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 | について (last modified 2016/12/13)
- 書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション (last modified 2016/12/22)
- 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 2015/11/26)
- 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 2015/12/07)
- 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 2016/06/03)
- 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 2016/07/05)
- 書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ modified 2016/06/17)
- 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 2017/01/26) NEW
- 書籍 | 日本乳酸菌学会誌 | 第8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、DDBJへの登録 (last modified 2017/01/31)
- イントロ | 一般 | ランダム

①

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)。

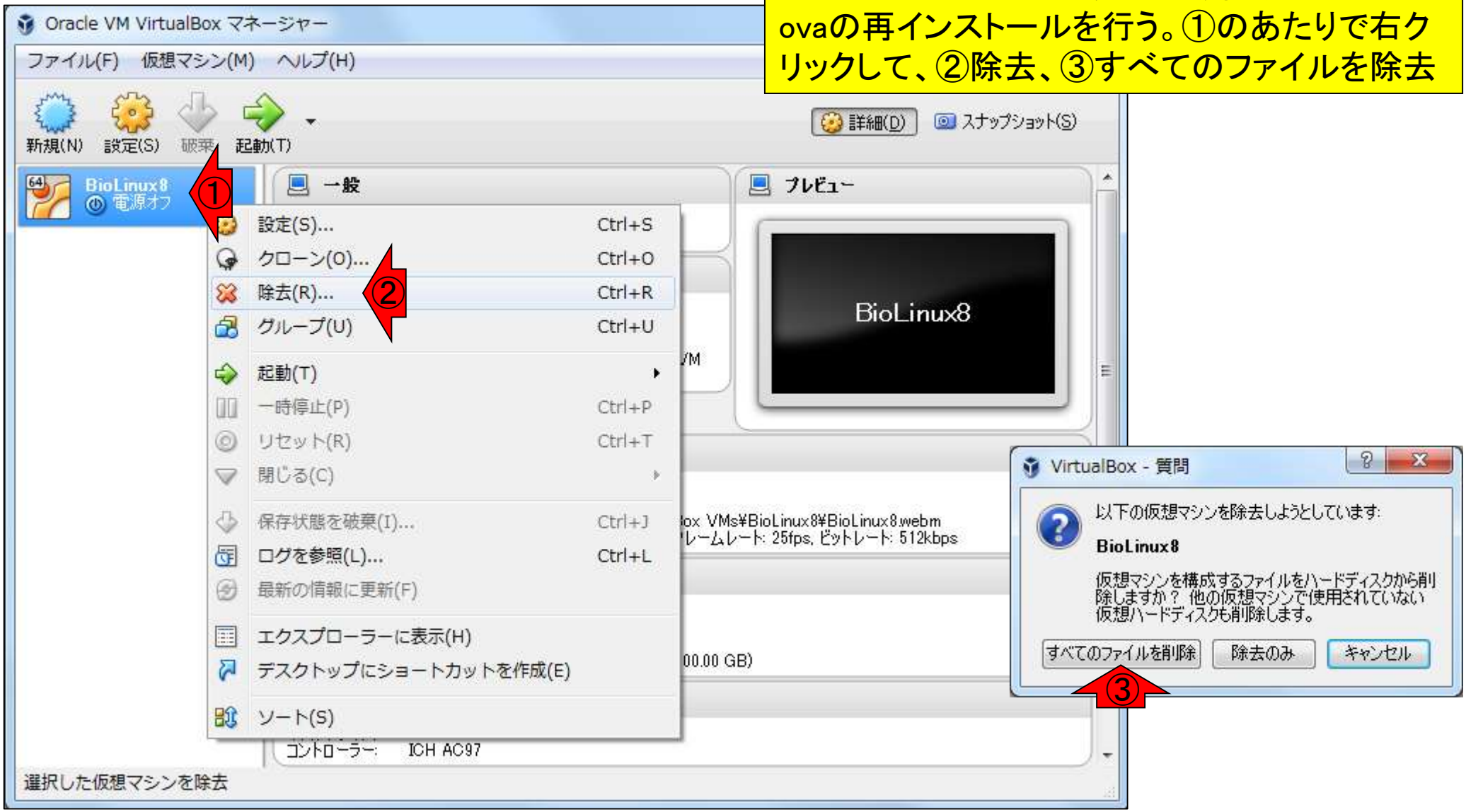
- 原稿PDF
- ウェブ資料PDF
  - Windows用(2016.06.17版; 約20MB)
  - Macintosh用

- (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)

②

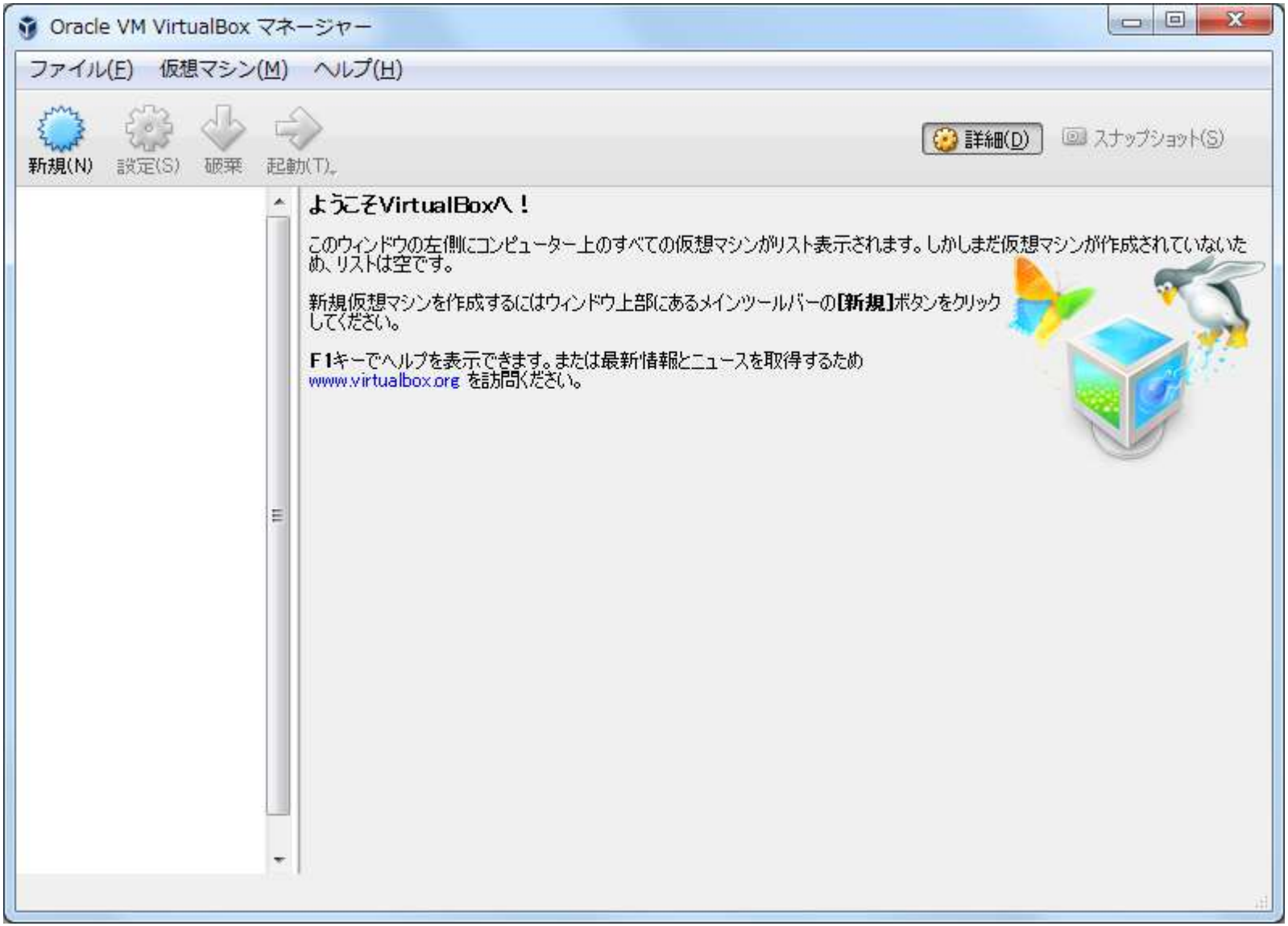
W18-3：ova除去

「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



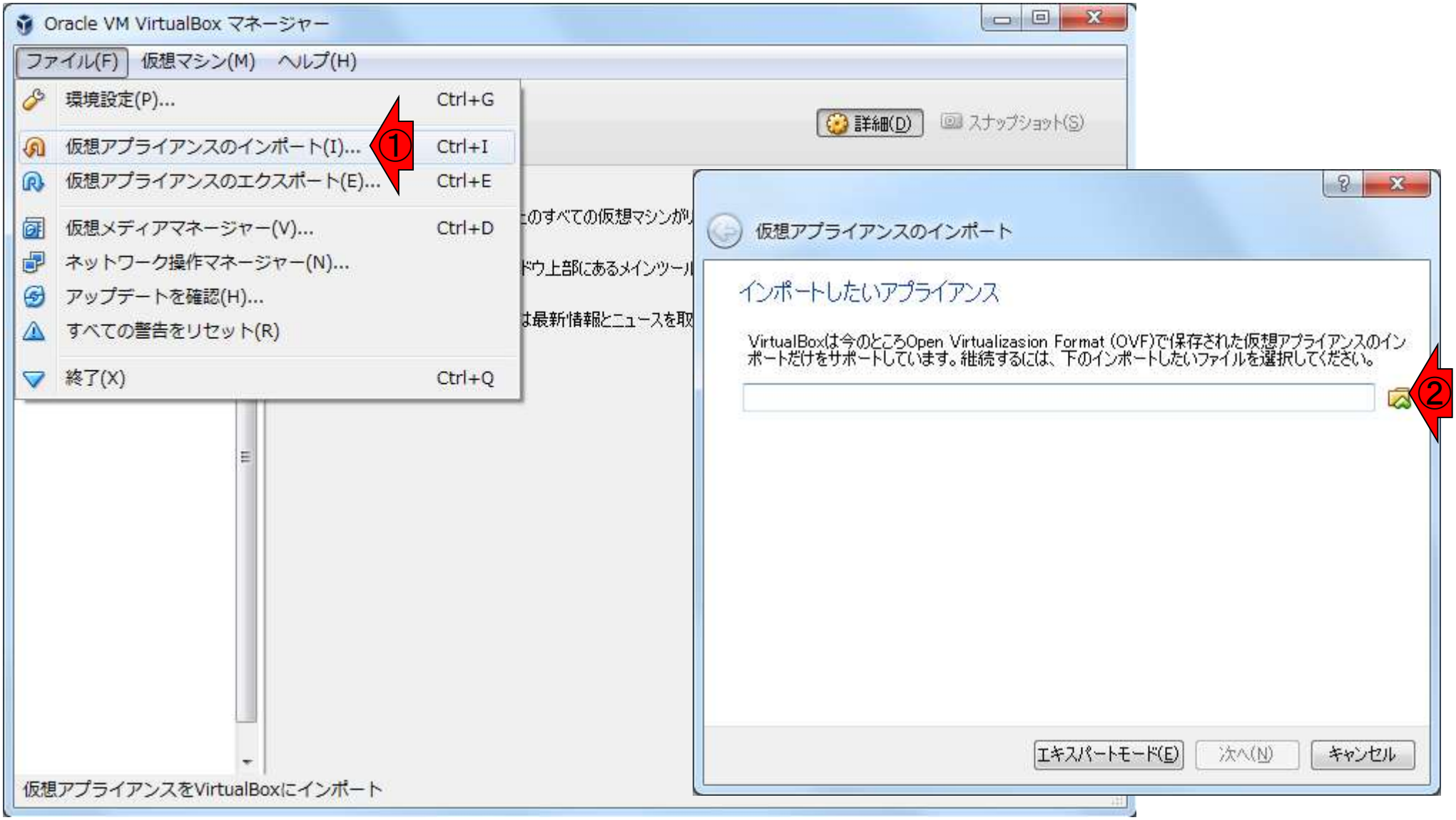
除去後はこんな感じになります

W18-3：ova除去

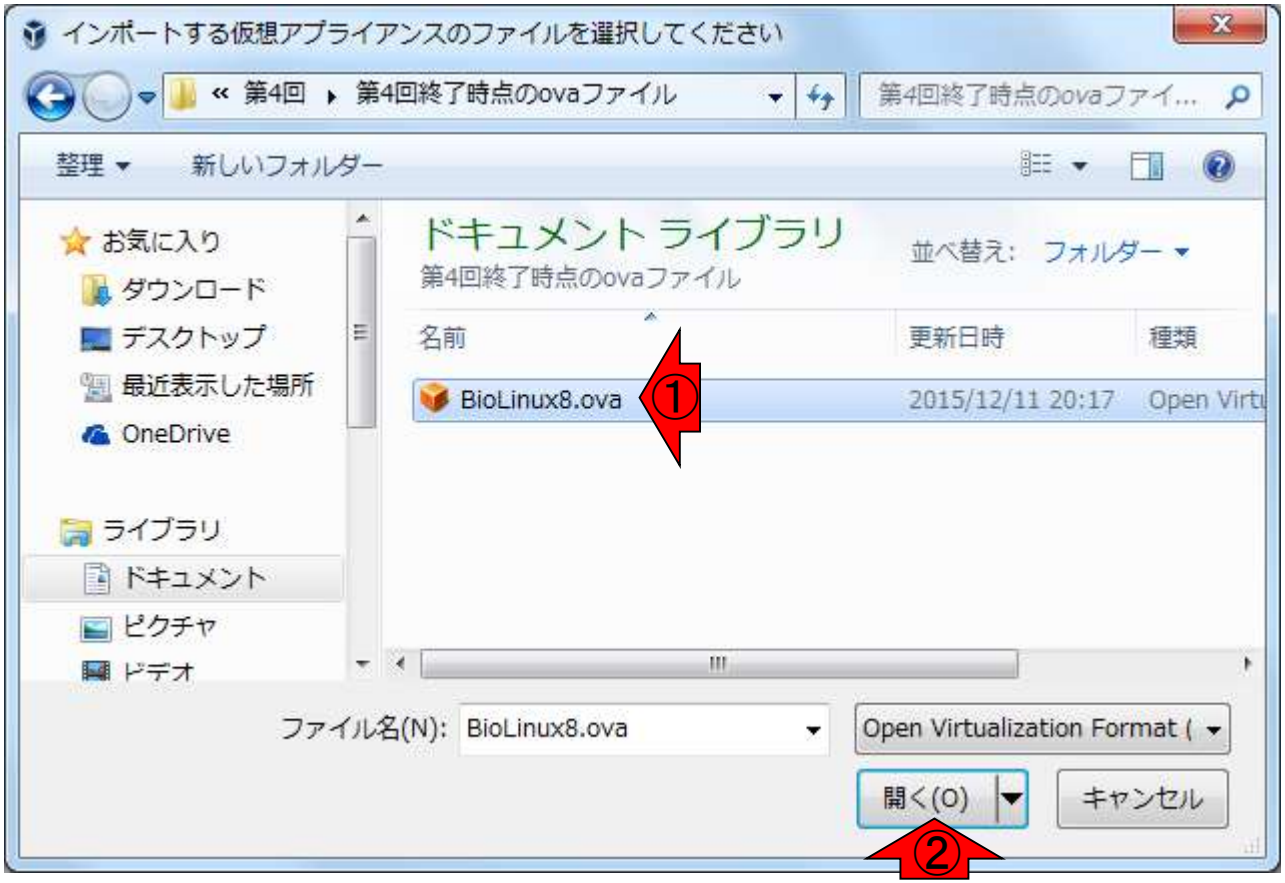


W18-4：ovaをインポート

①仮想アプライアンスのインポート、②のところを押す



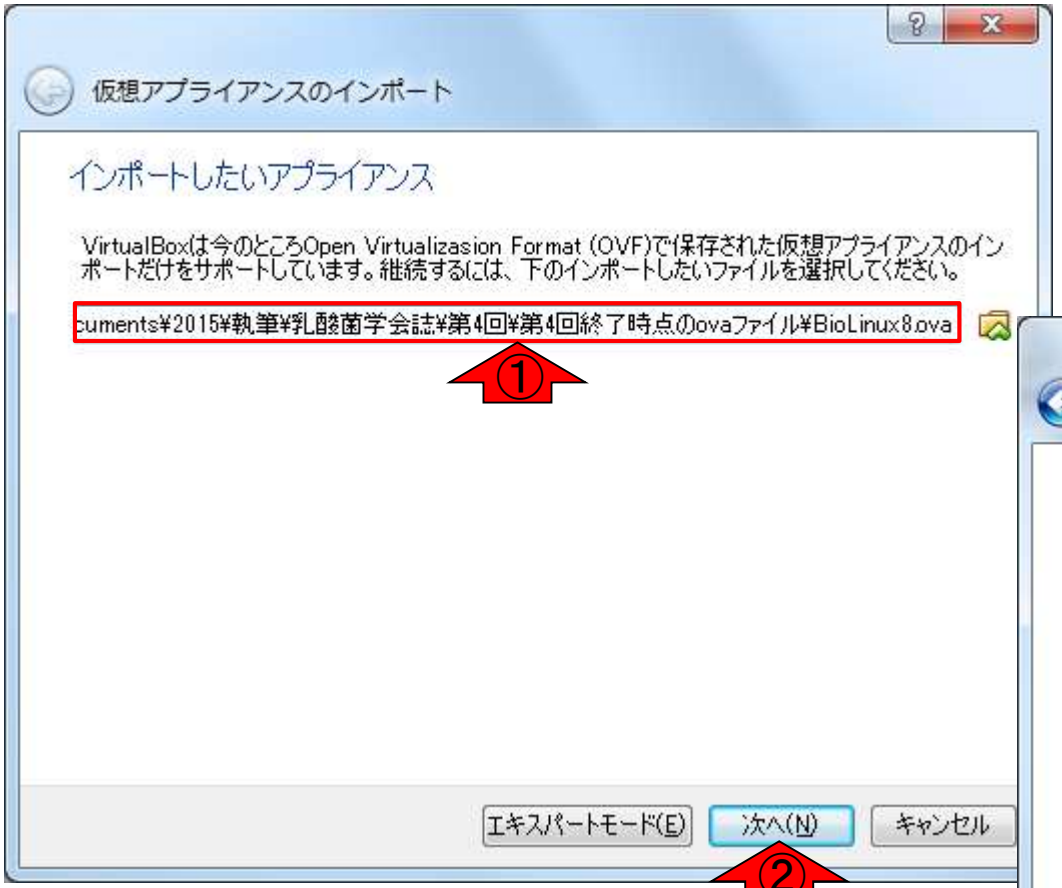
W18-5：BioLinux8.ovaを選択



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

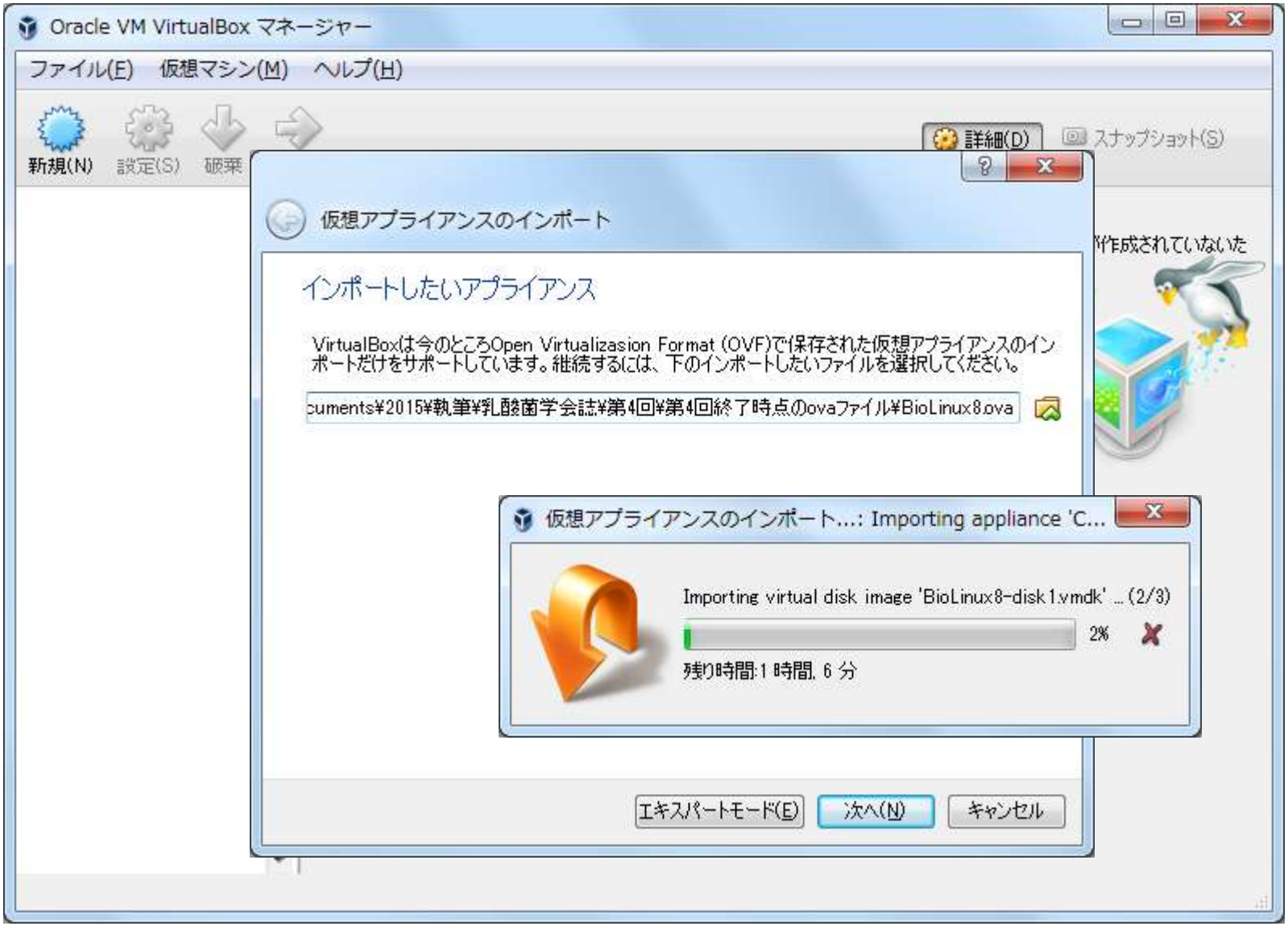
W18-6：ovaをインポート

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



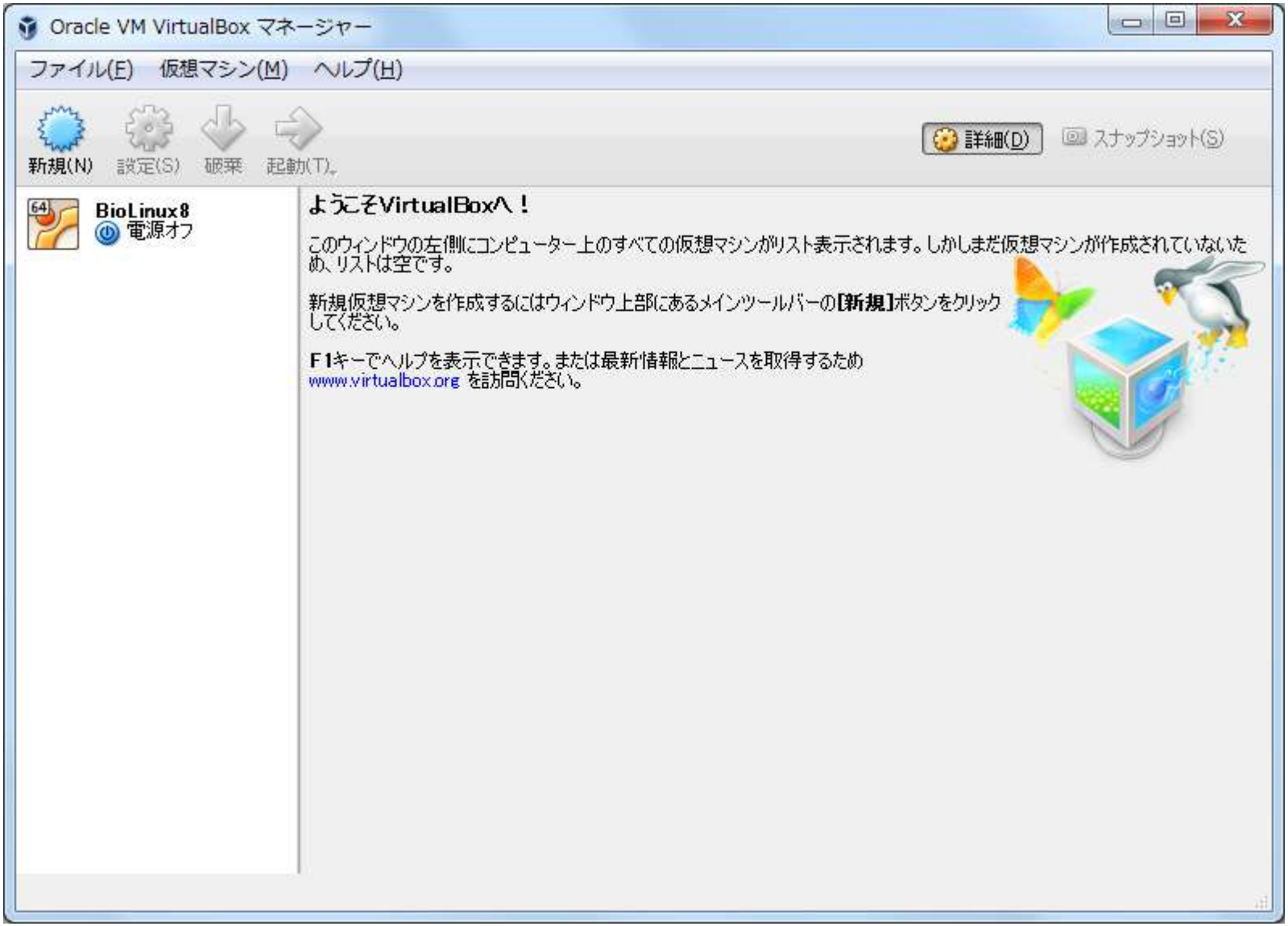
だいたい5-10分程度で終わります

W18-7：インポート中



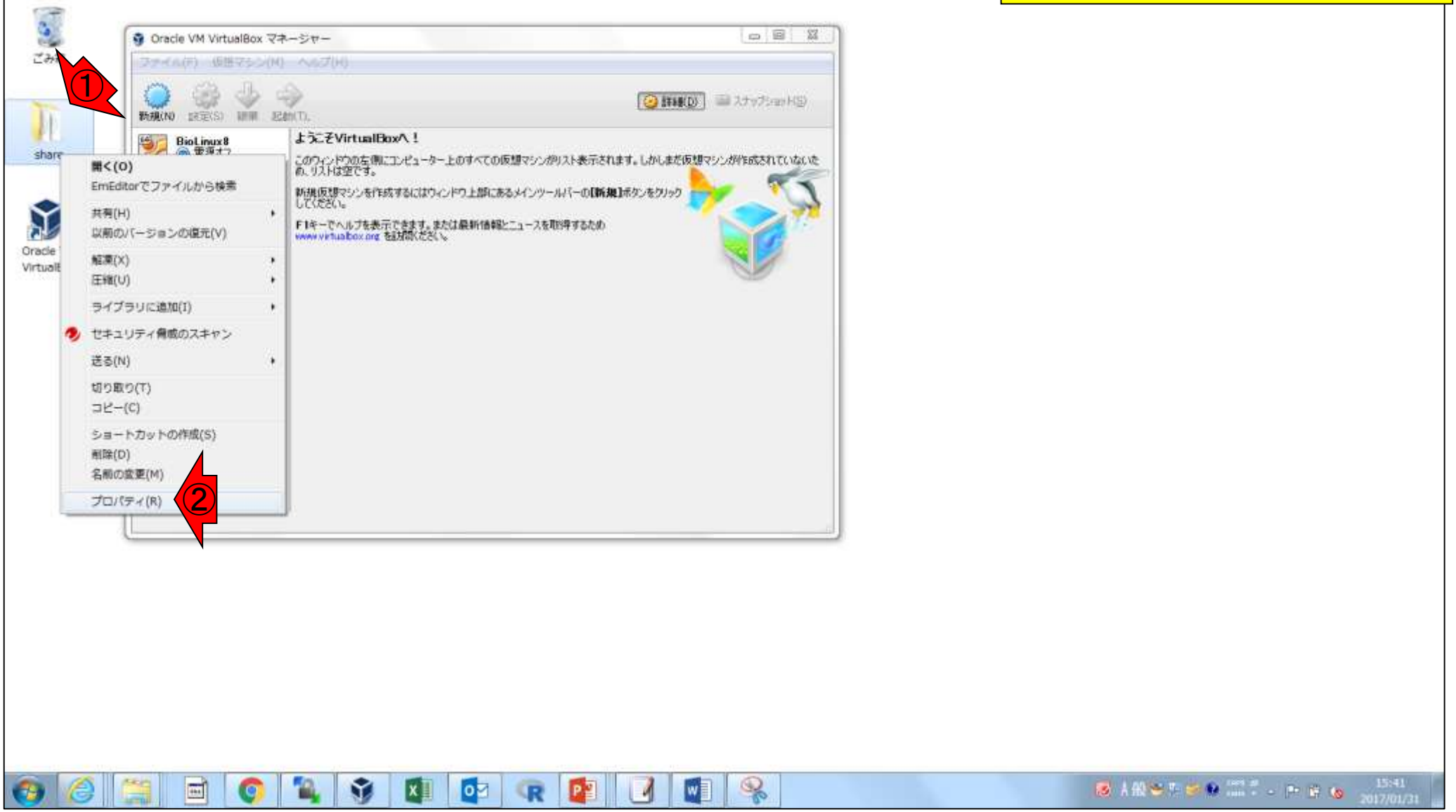
こんな感じになっているはずです

W18-8：インポート完了



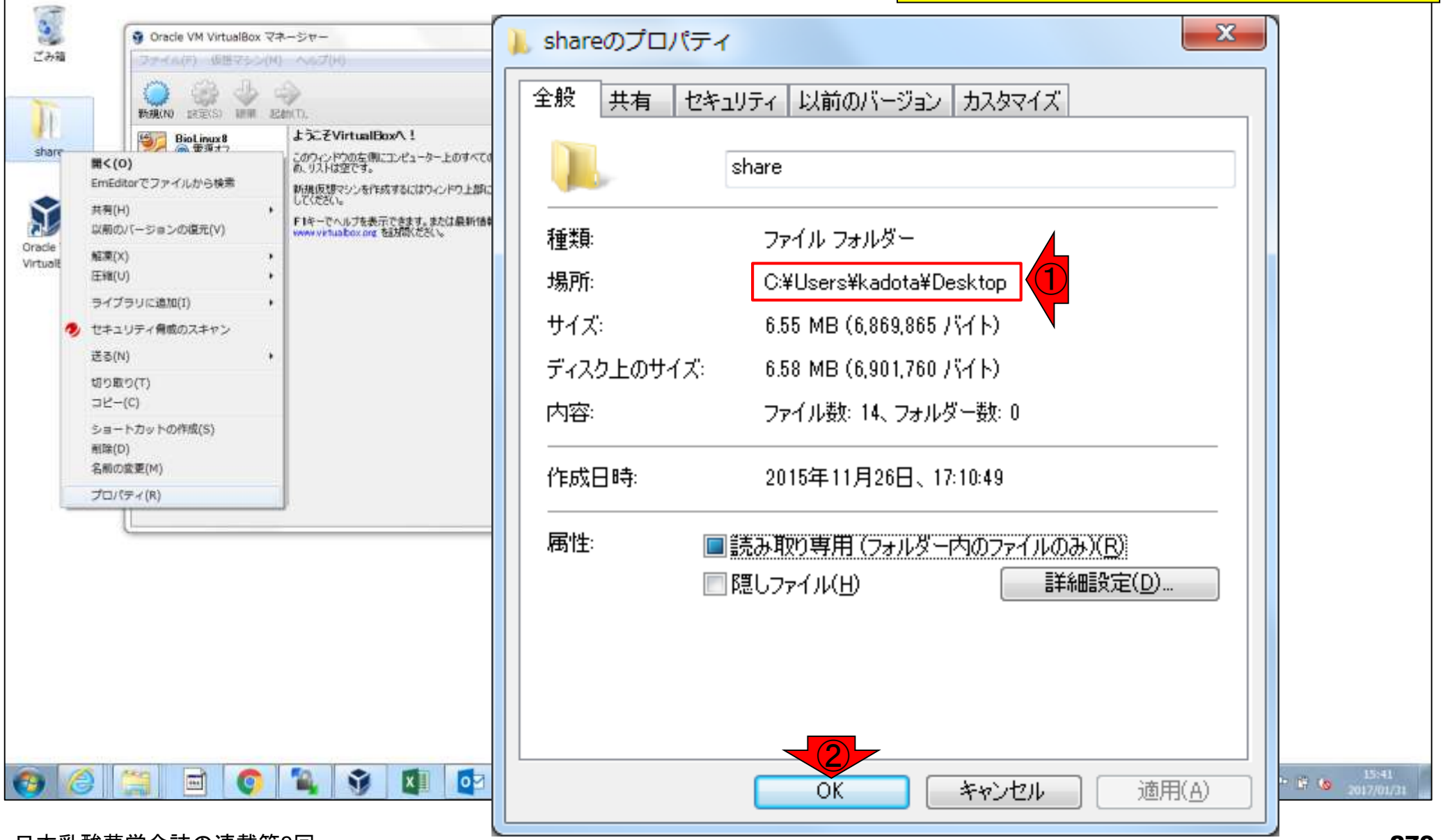
W18-9：shareフォルダ確認

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



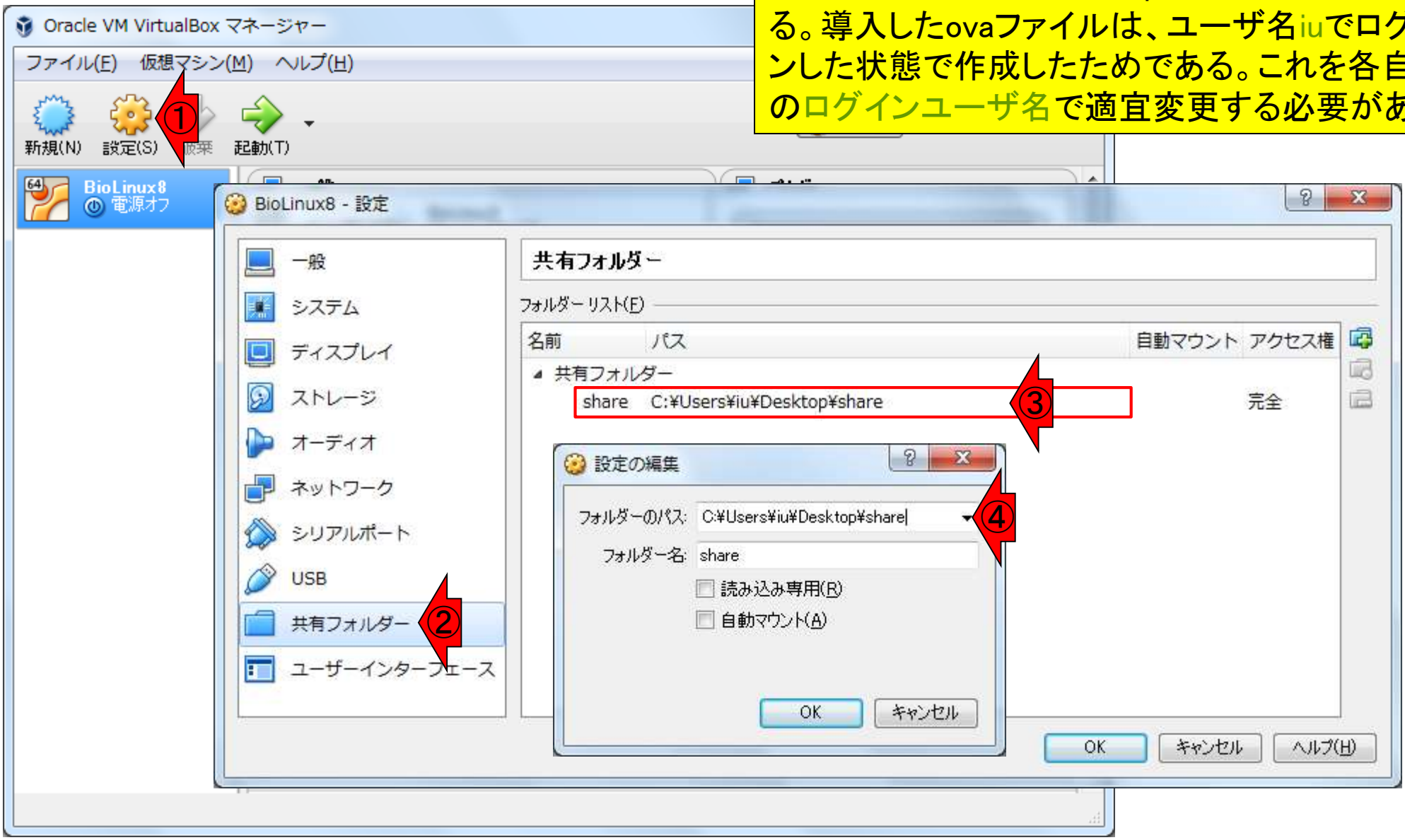
W18-9：shareフォルダ確認

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



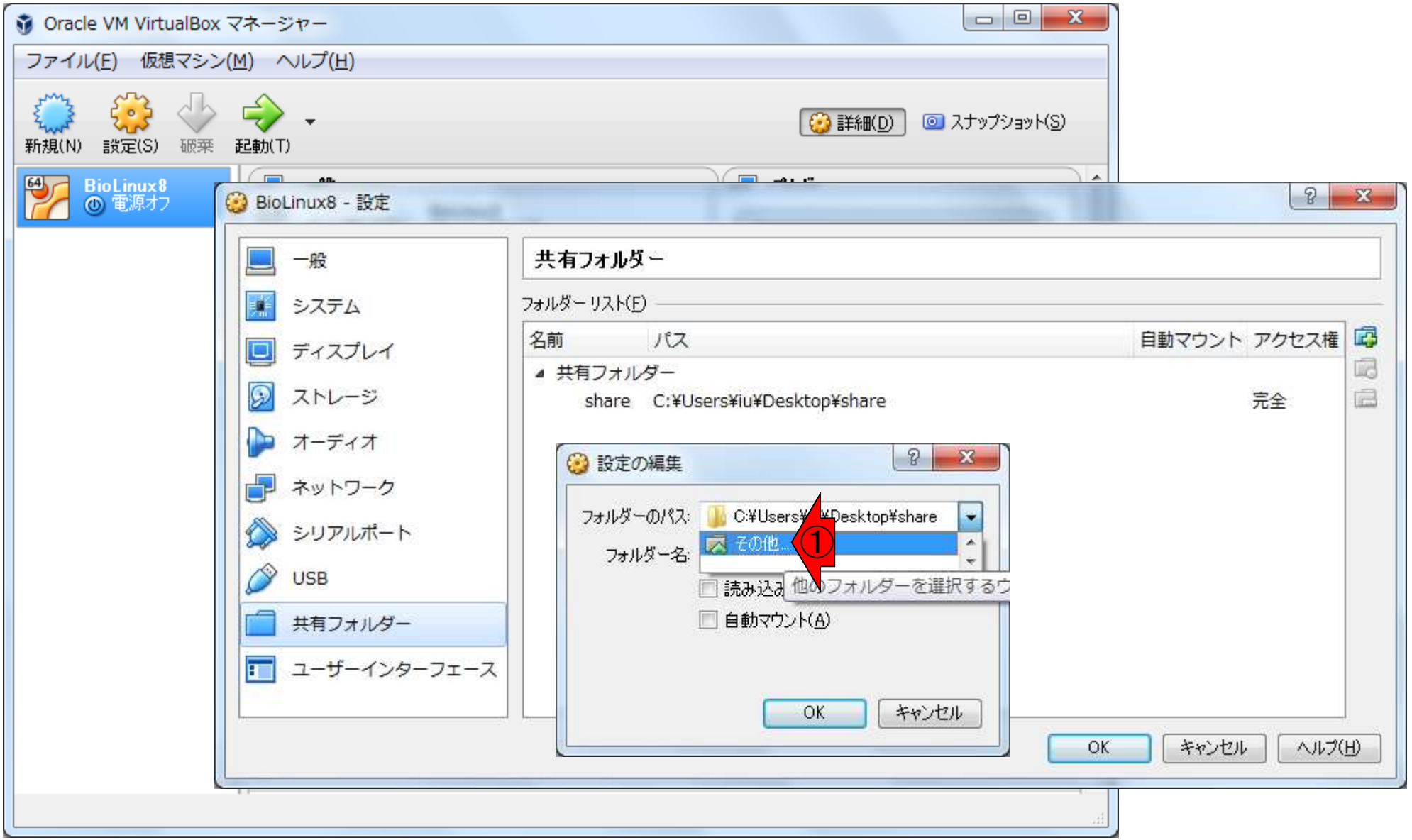
W18-10：パス設定

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



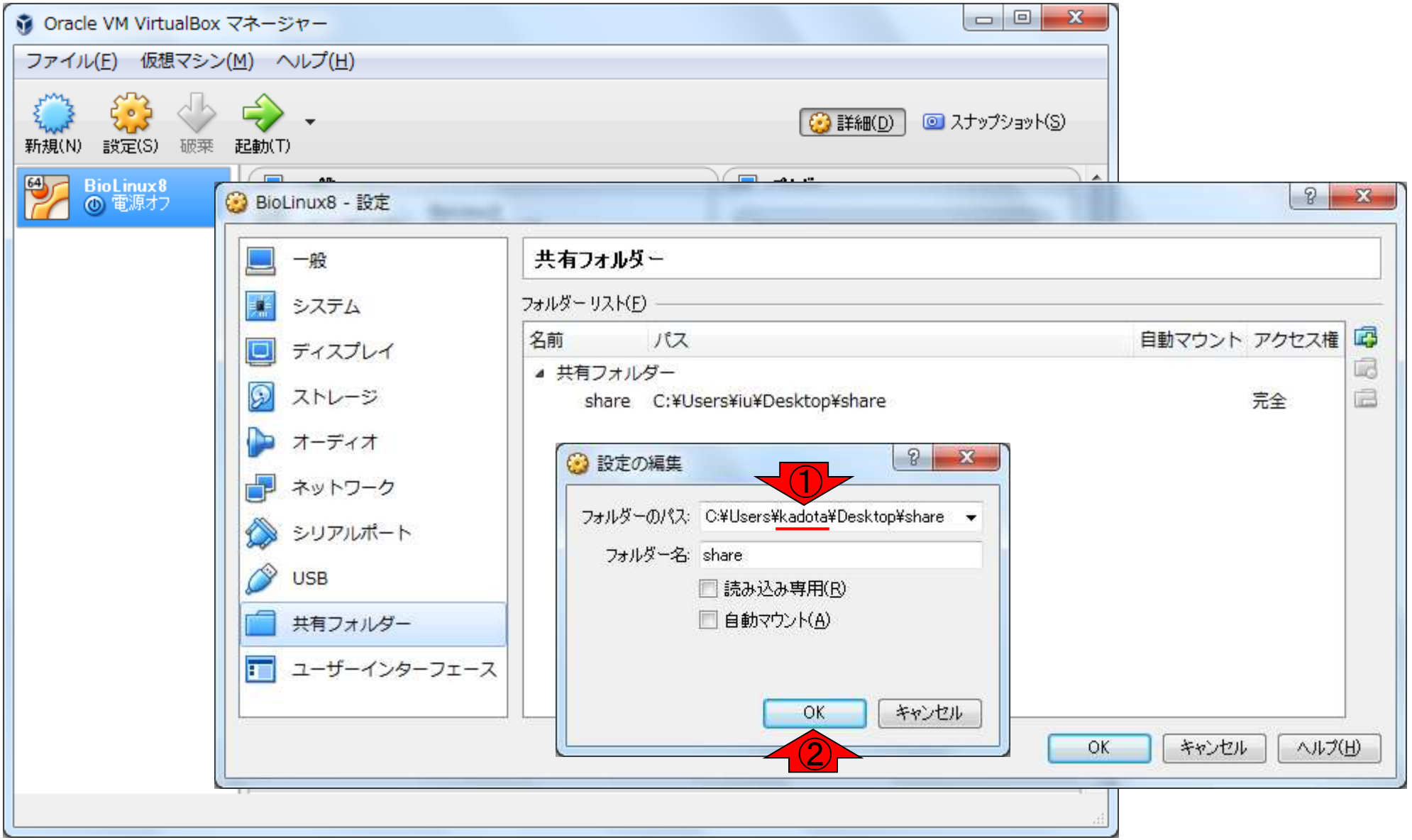
W18-10：パス設定

①「その他」を選んで、本来の「デスクトップ – share」のパスを指定



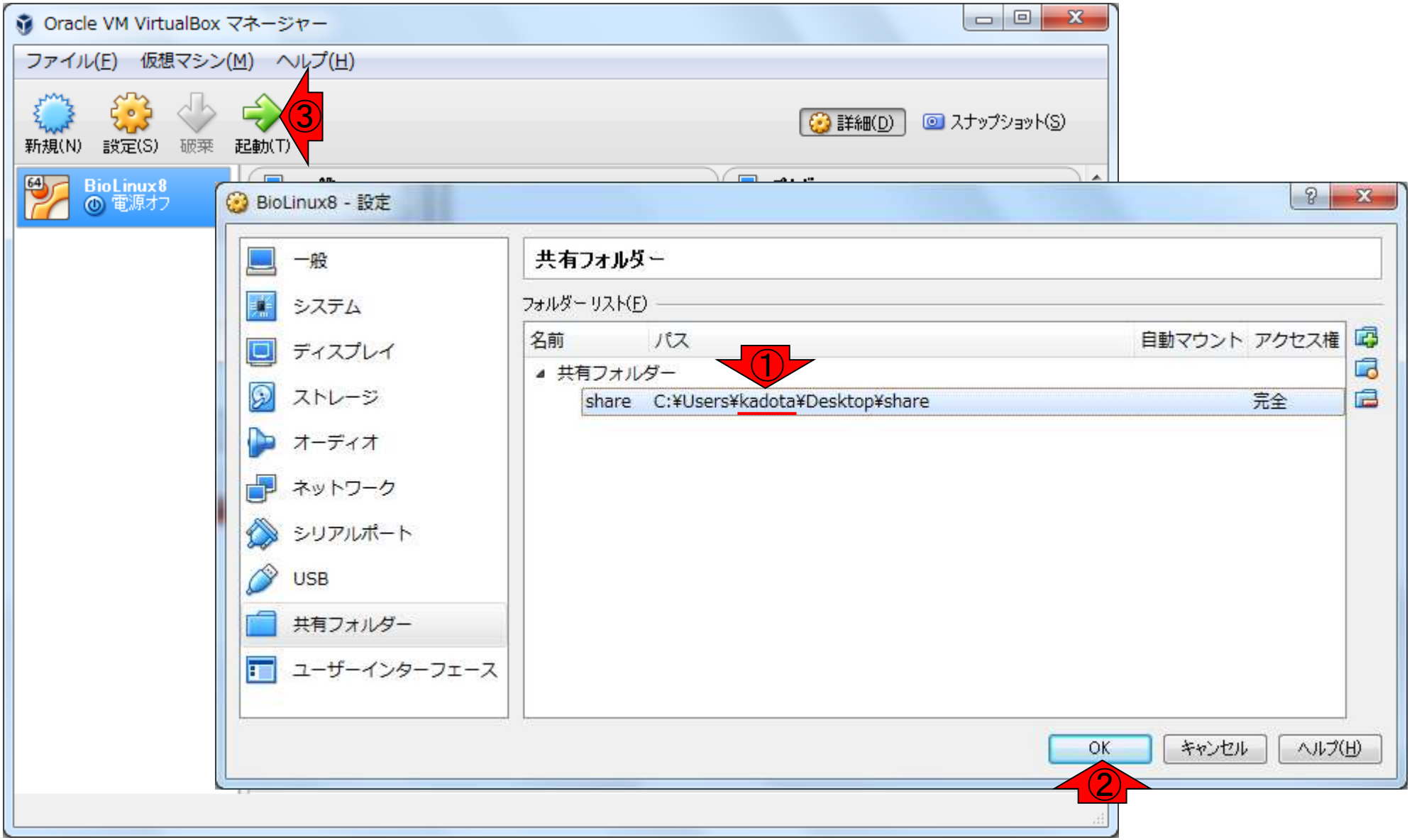
W18-10：パス設定

ユーザ名kadotaのPC環境では①のようになる。②OK



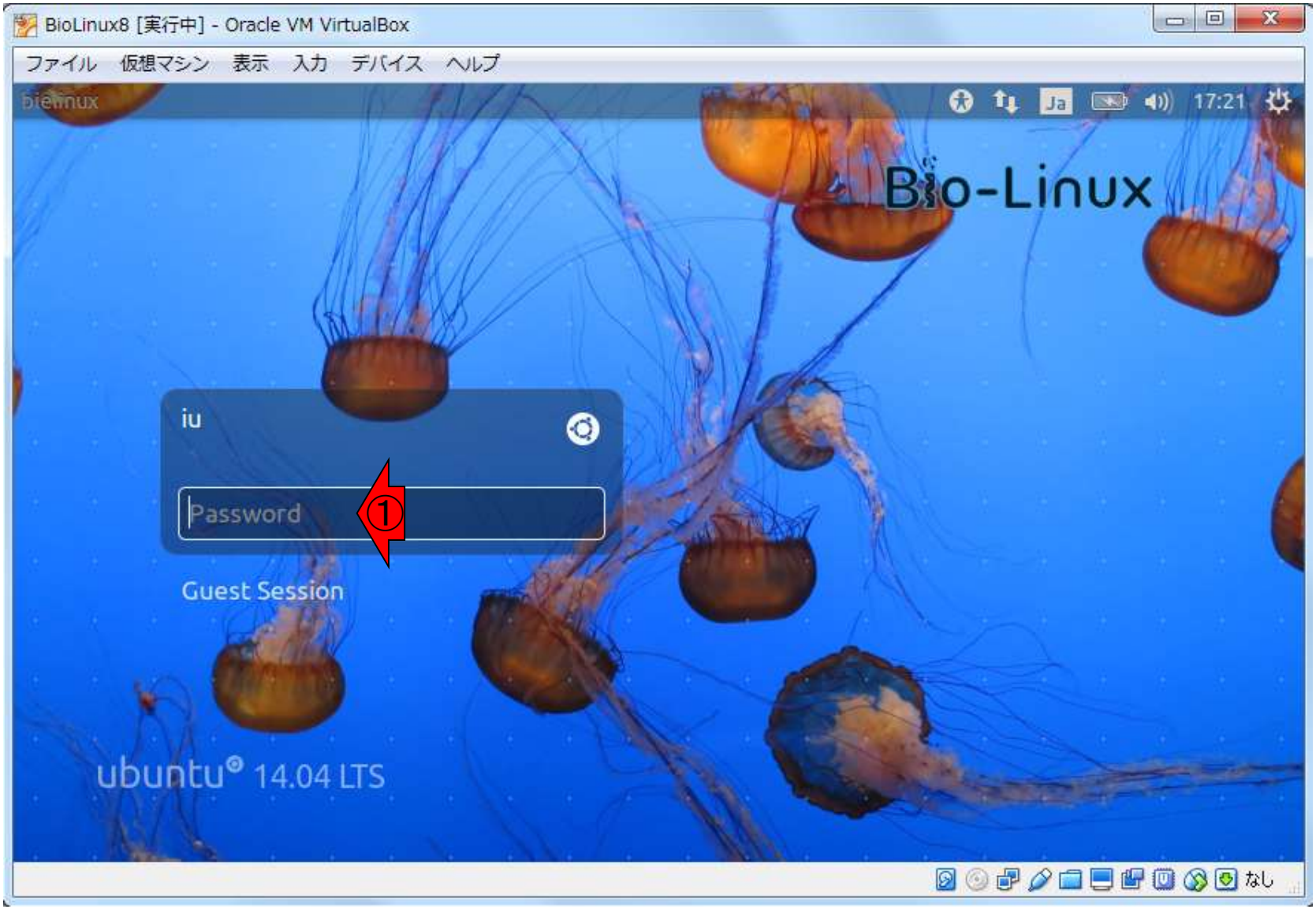
W18-10：パス設定

①の部分が変更されていることがわかります。②OK、③起動



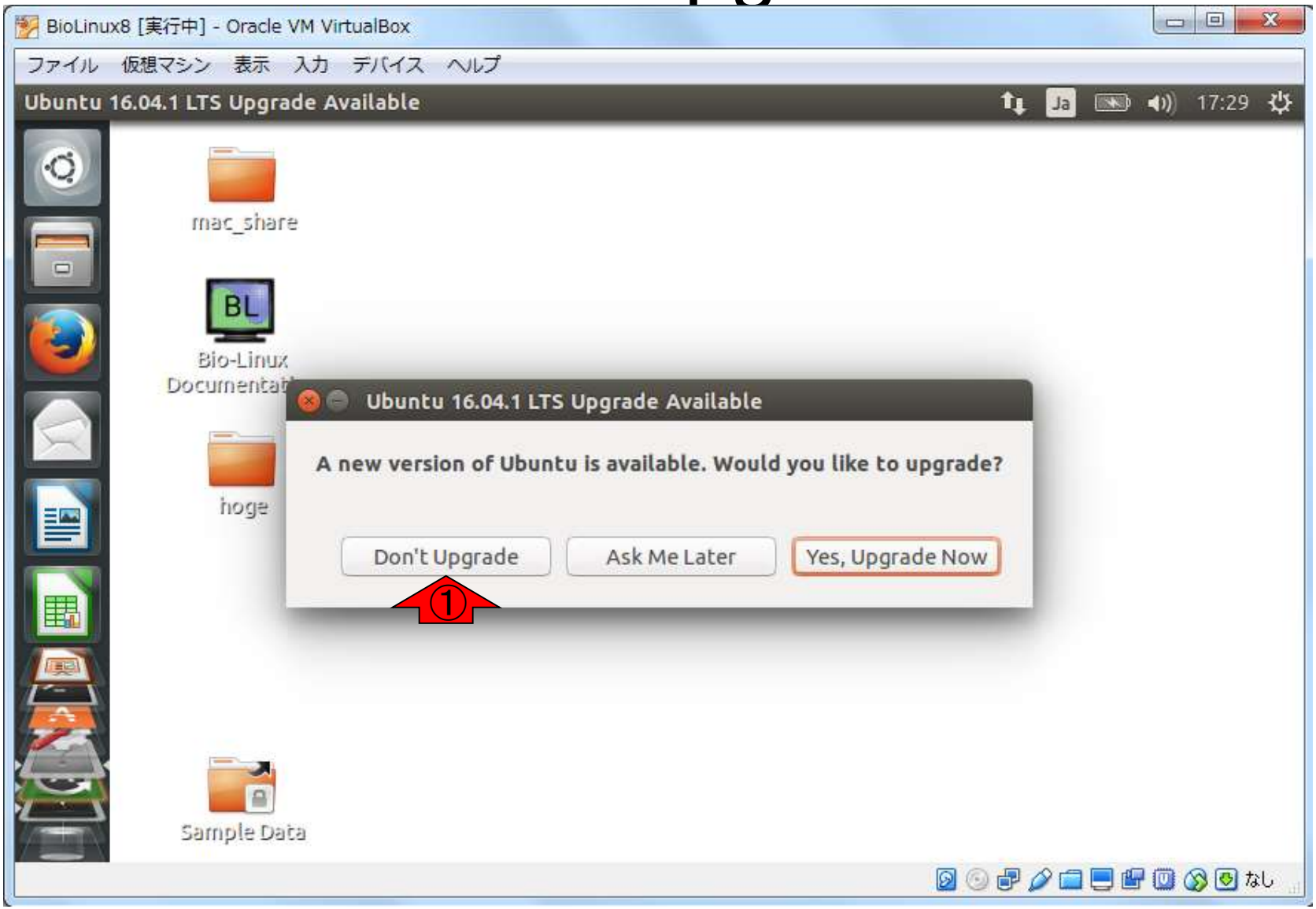
W18-11：起動

無事復活。①パスワード(pass1409)を入力して起動



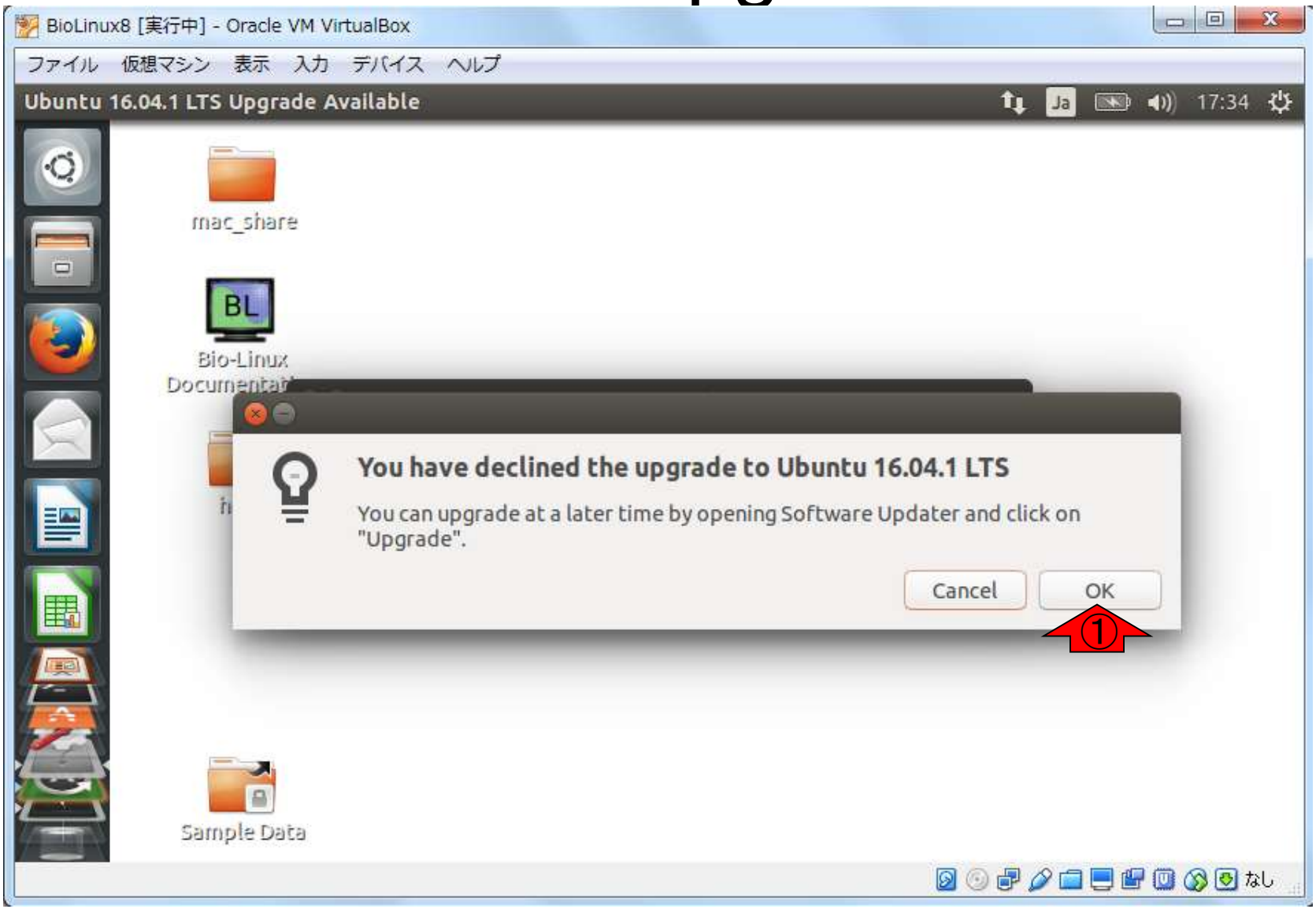
W18-11：Don’t Upgrade

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



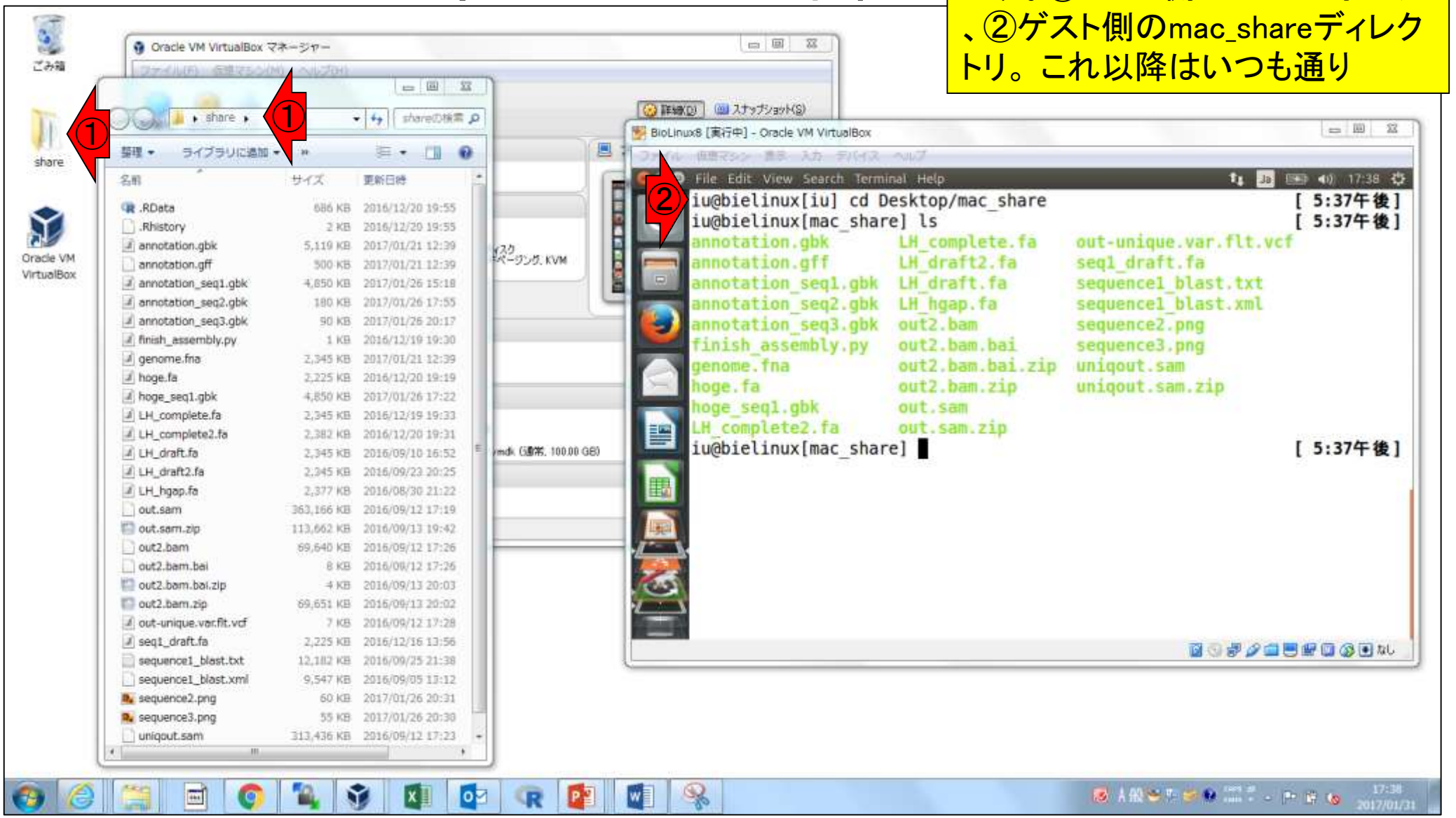
①OK

W18-11：Don’t Upgrade



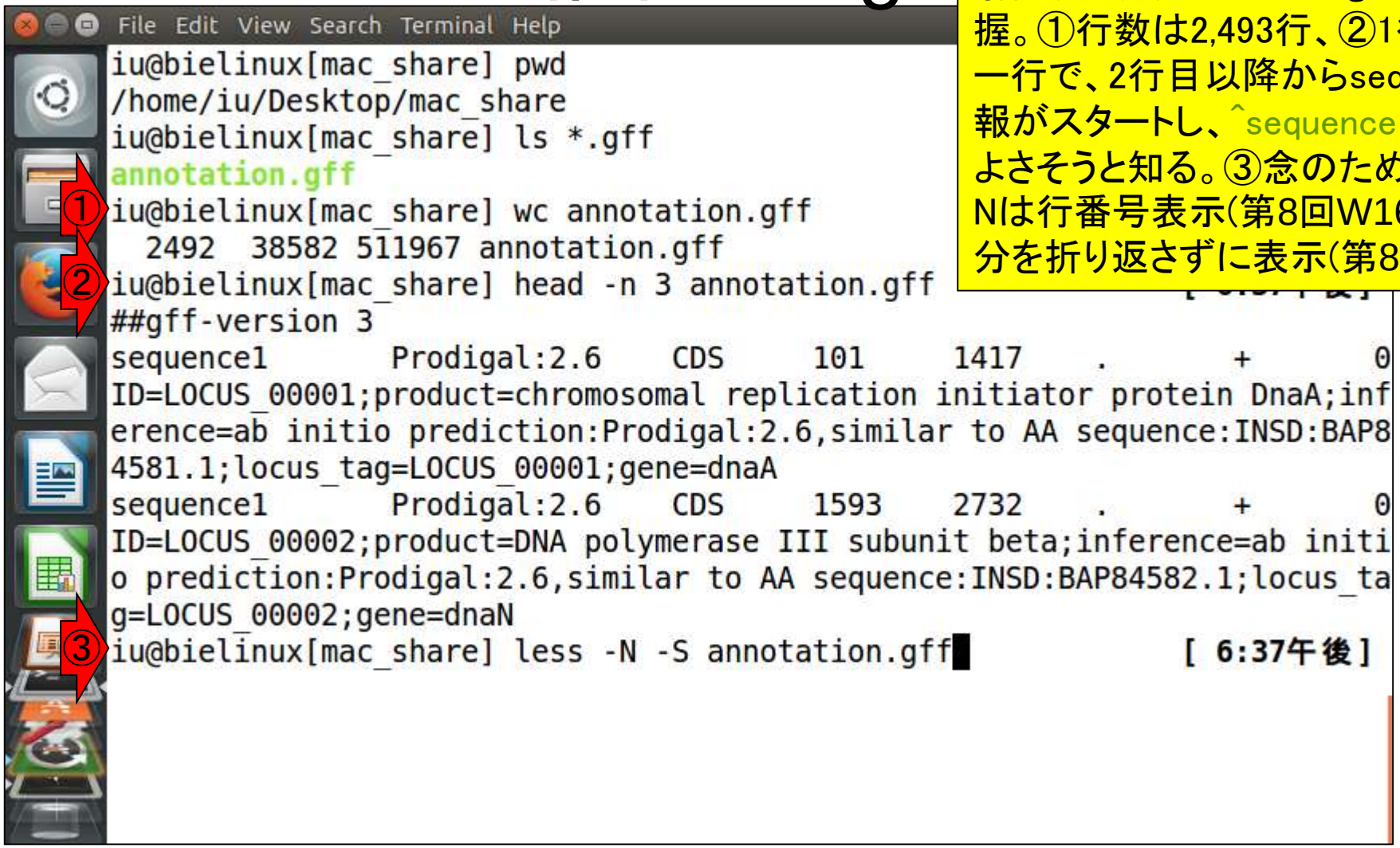
W18-12：共有フォルダ確認

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



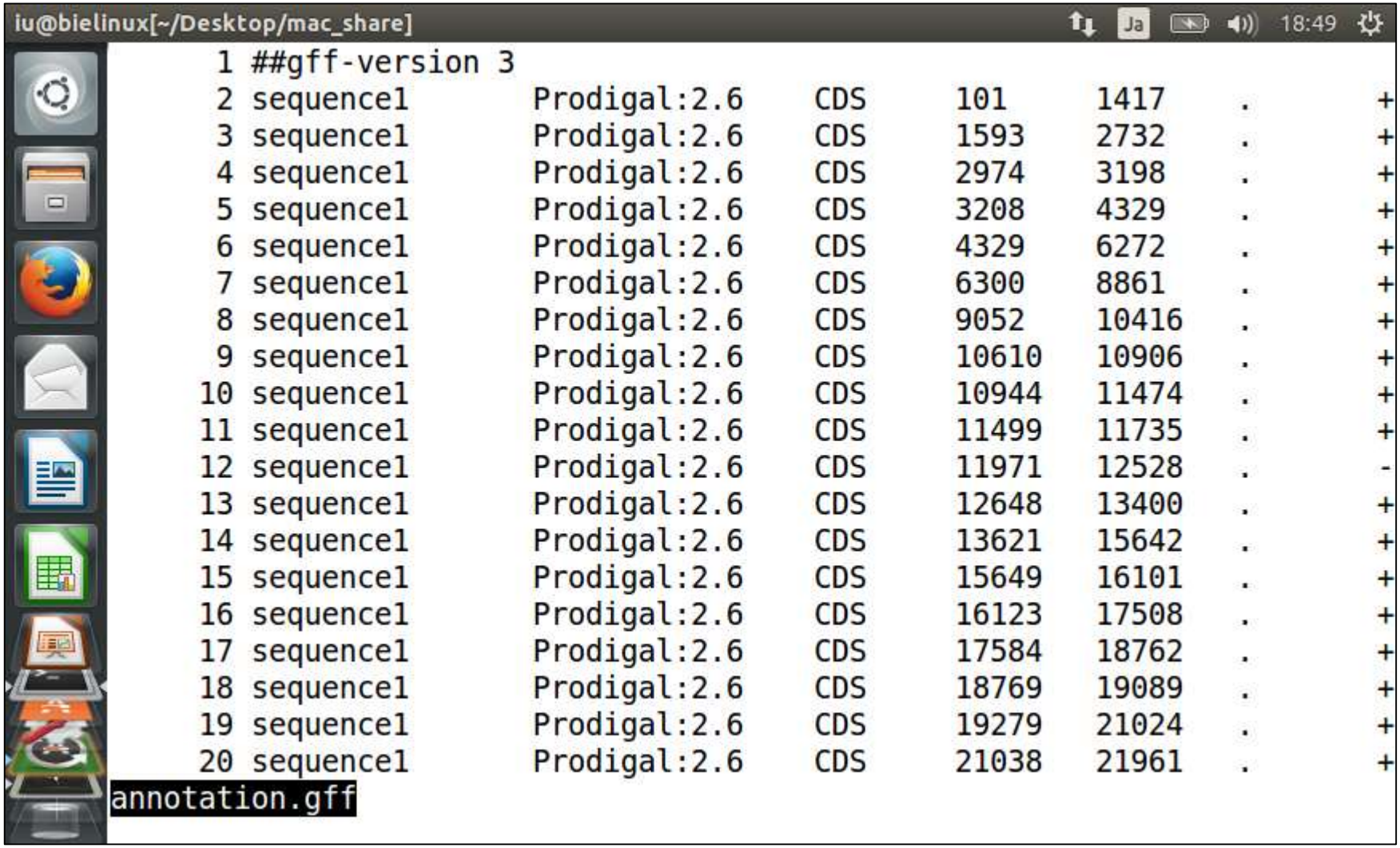
W19-1：配列情報つきgff

W11-4で「”配列情報つきの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行目がヘッダー行で、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)



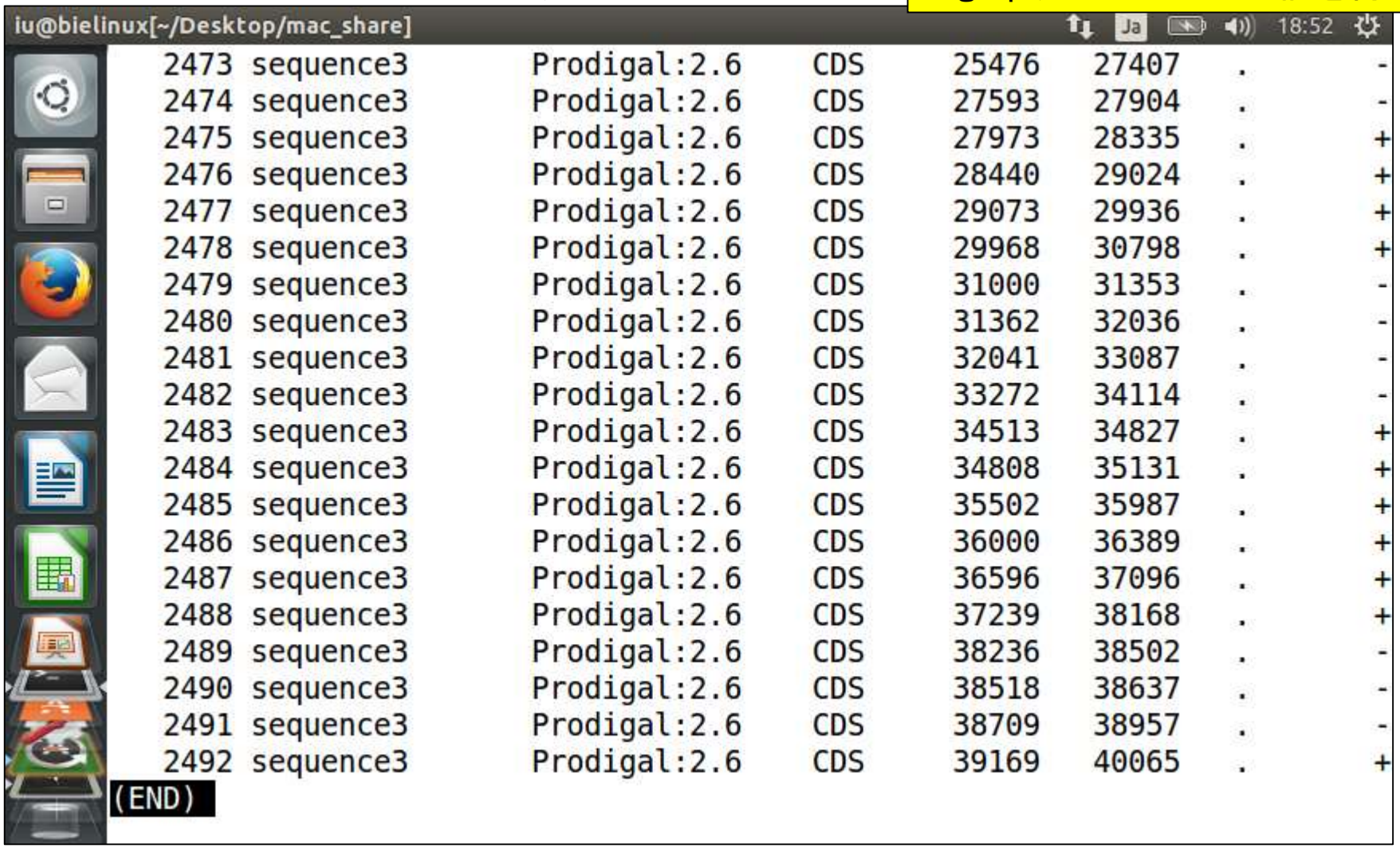
「less -N -S annotation.gff」実行結果。

W19-2：less復習



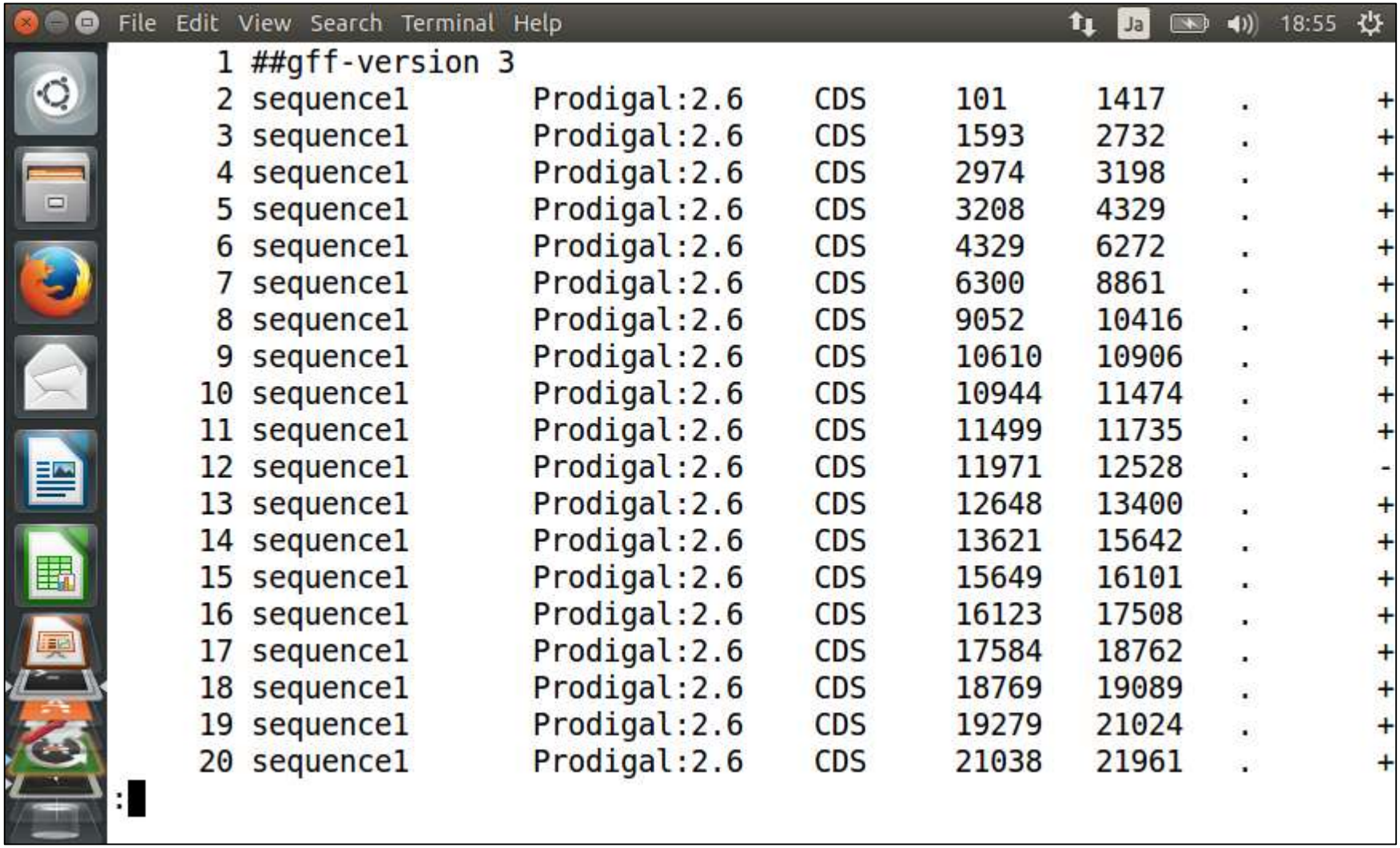
W19-2：less復習

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



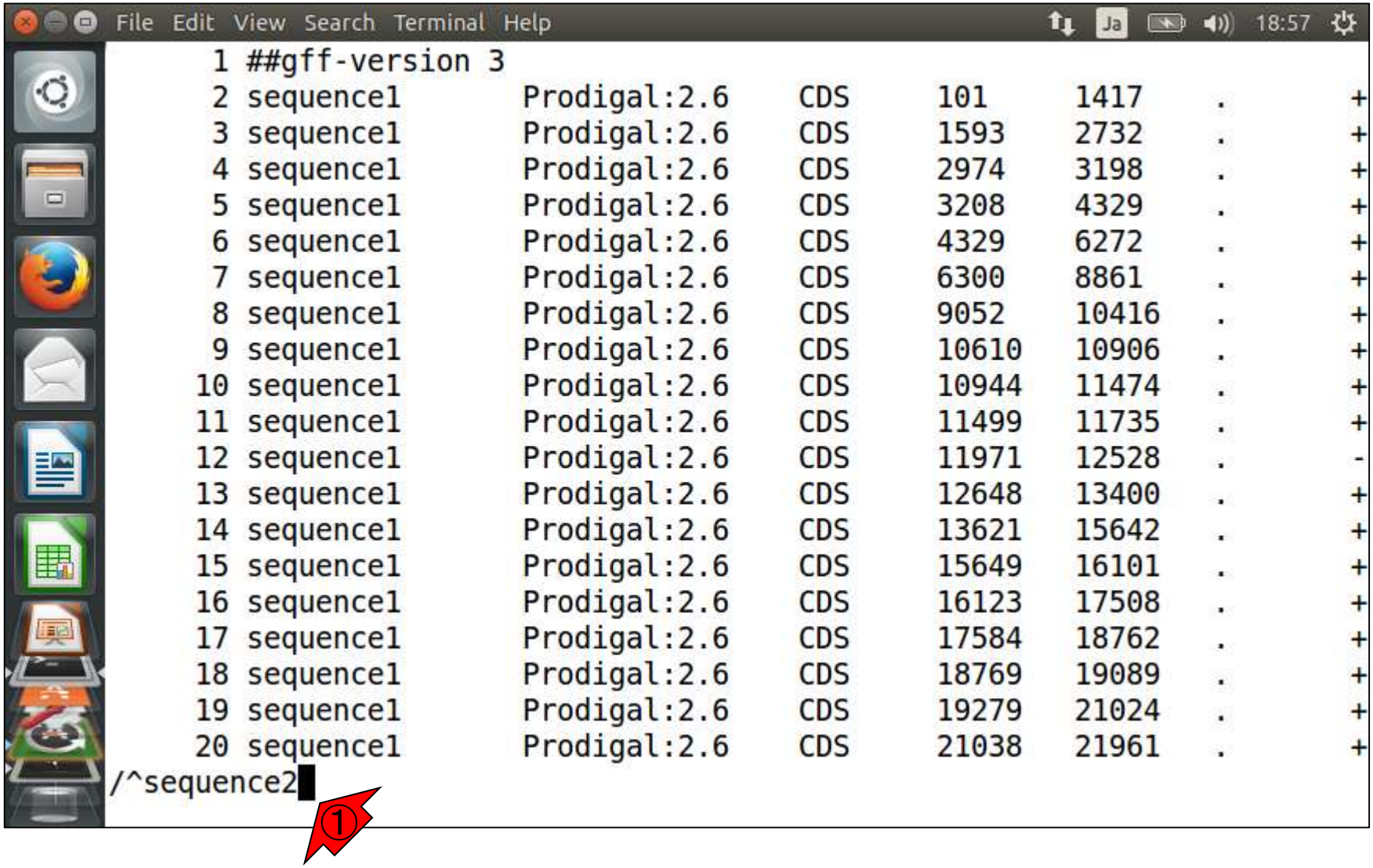
小文字のgを打ち込んでファイルの先頭に移動

W19-2：less復習



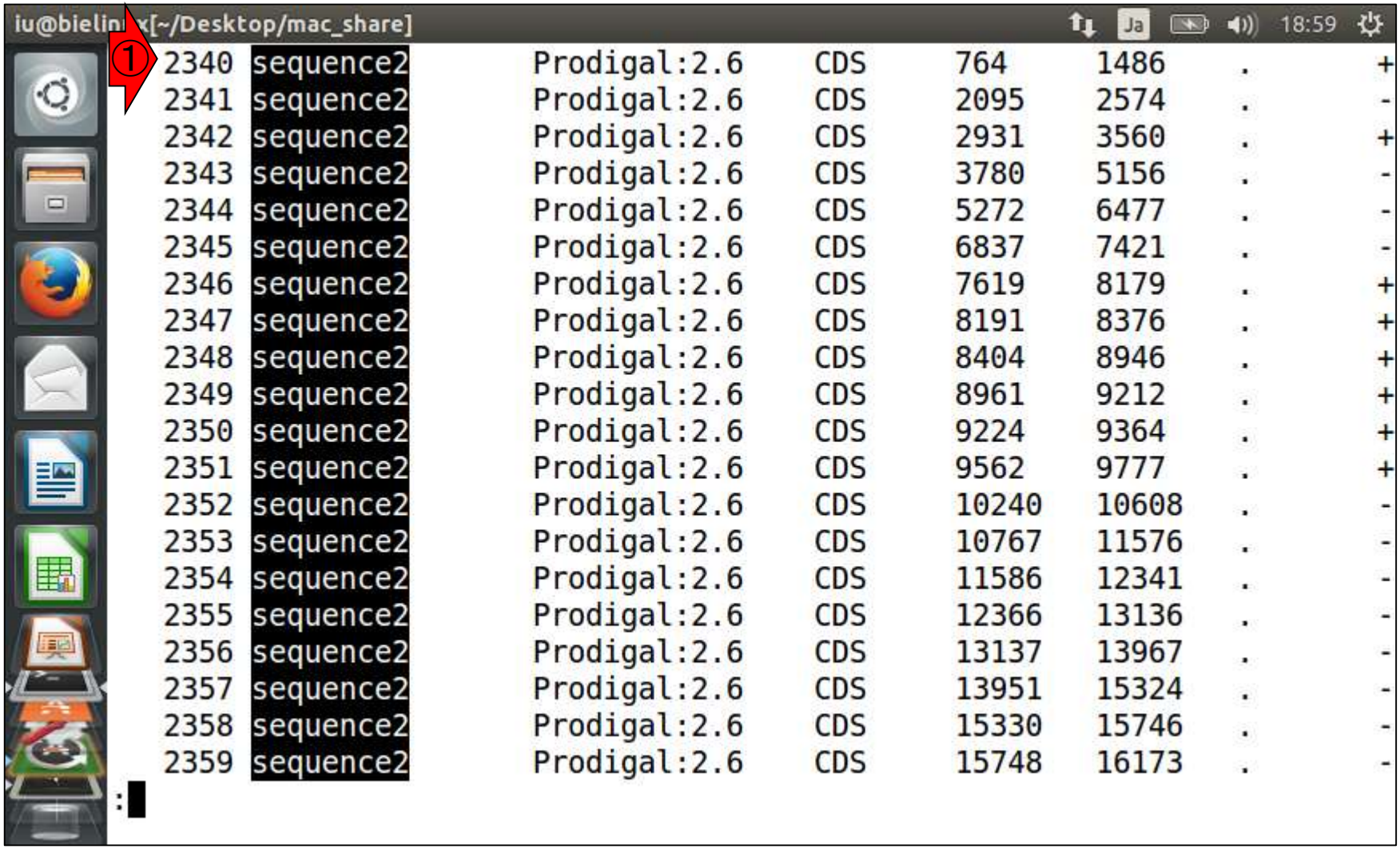
W19-2：less復習

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



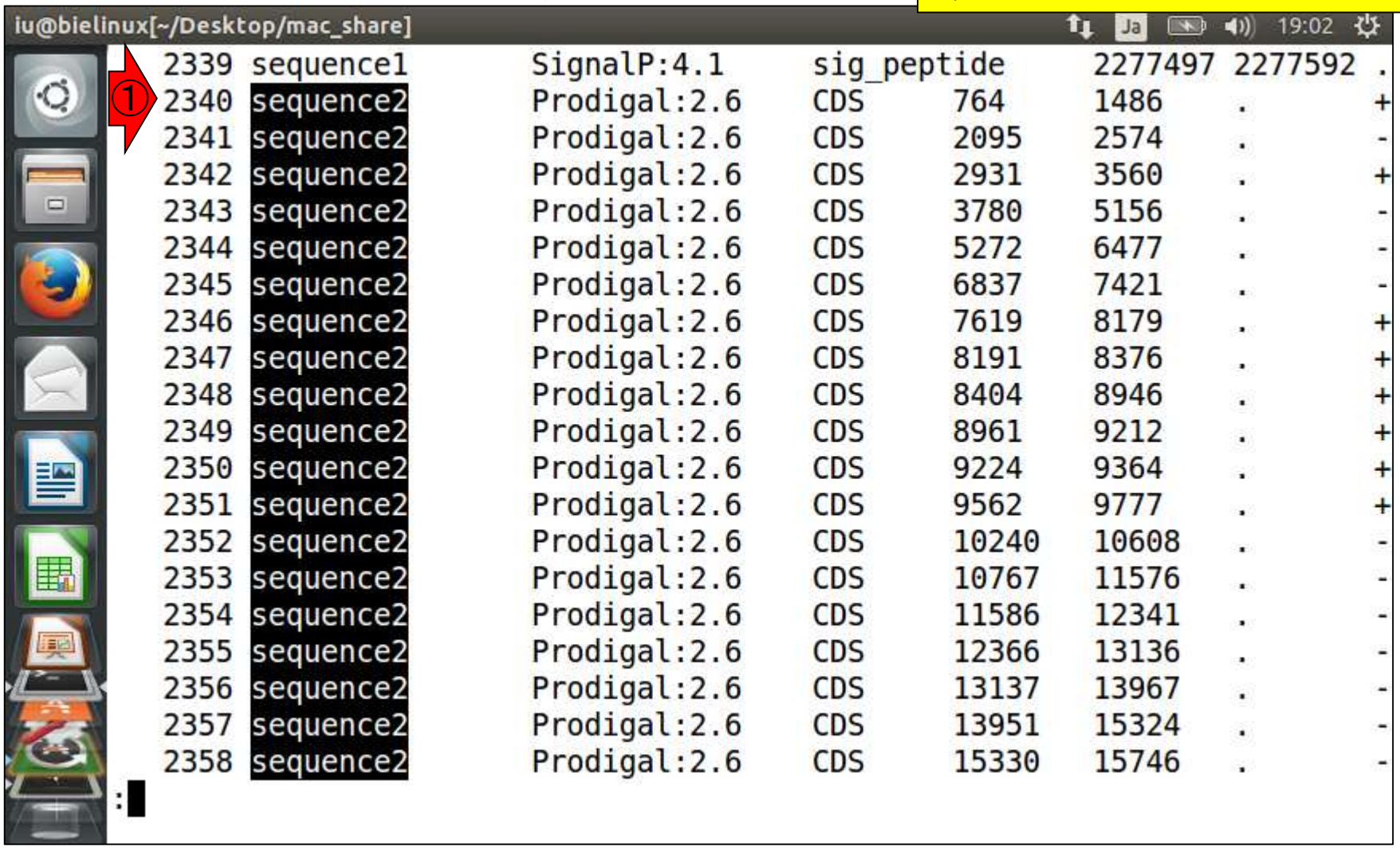
W19-2：less復習

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



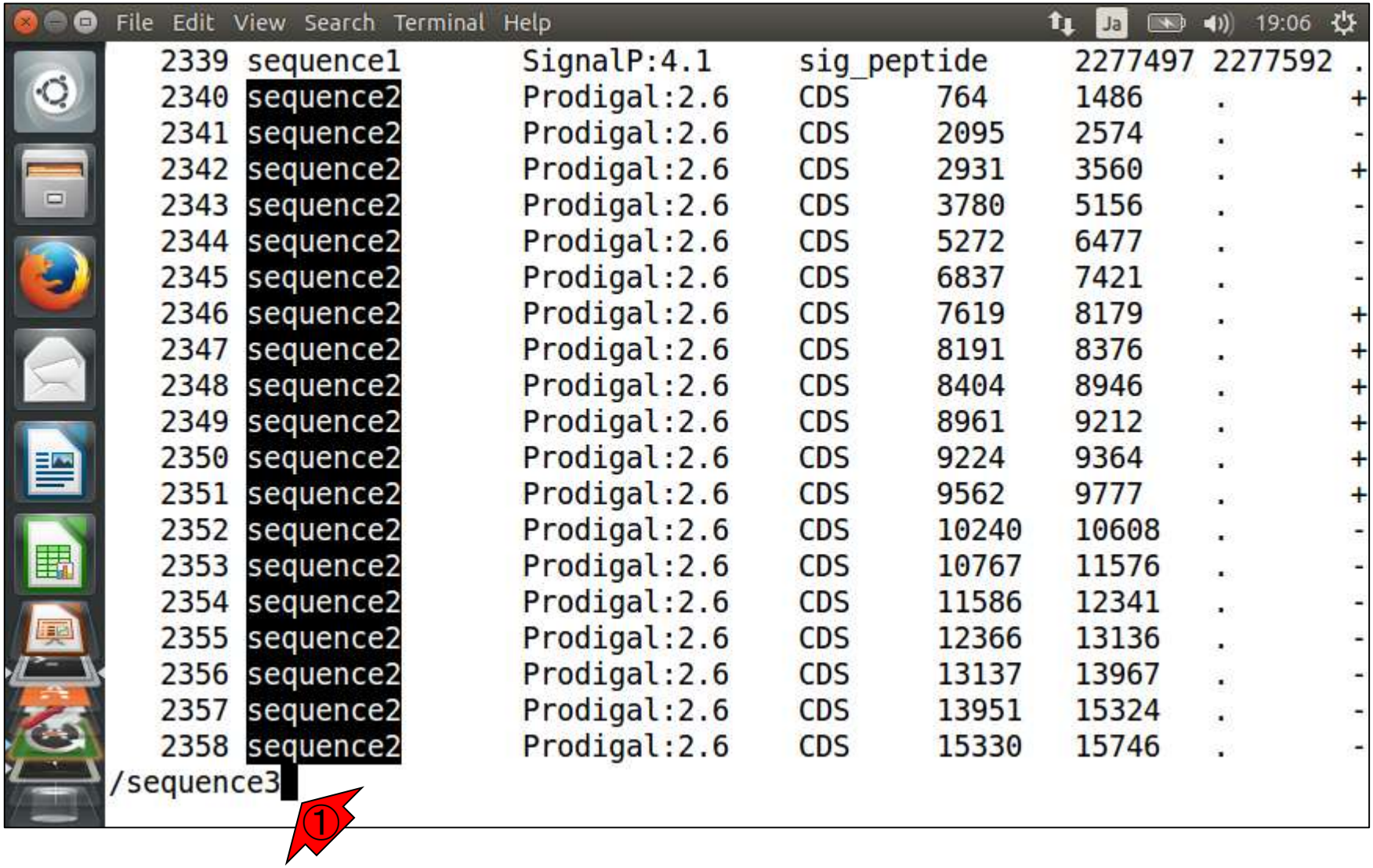
W19-2：less復習

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



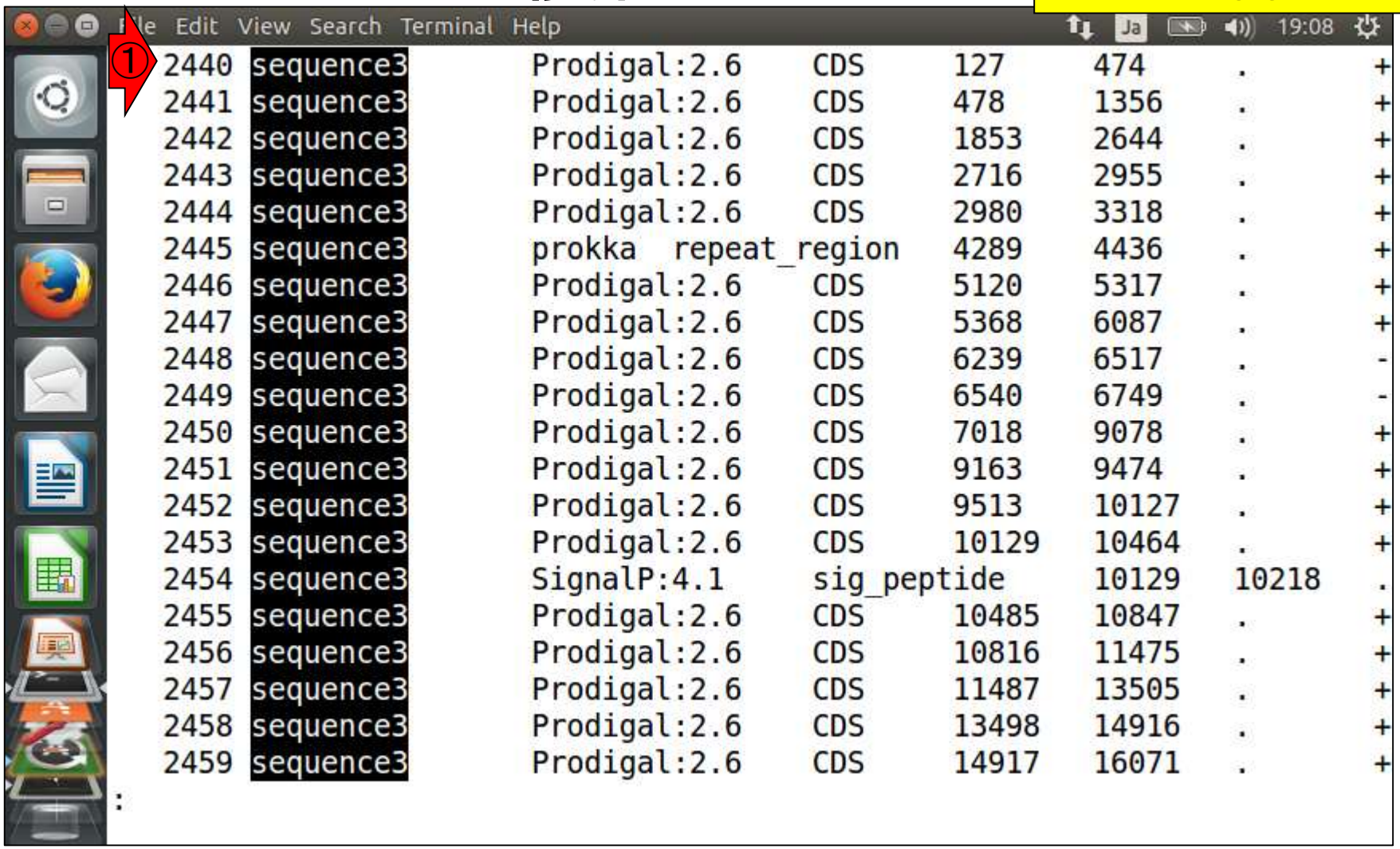
①「/sequence3」と打ってリターン

W19-2：less復習



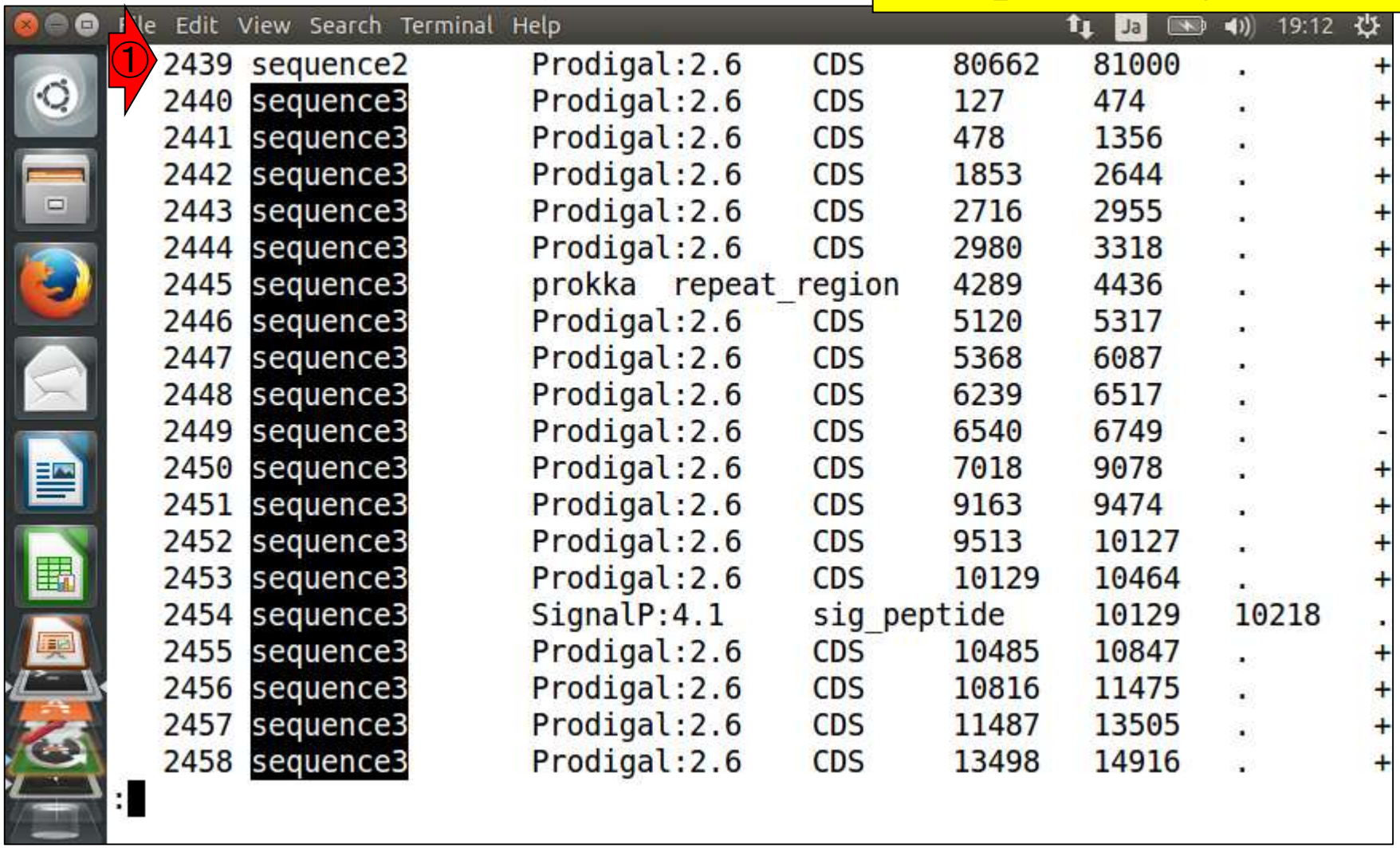
W19-2：less復習

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



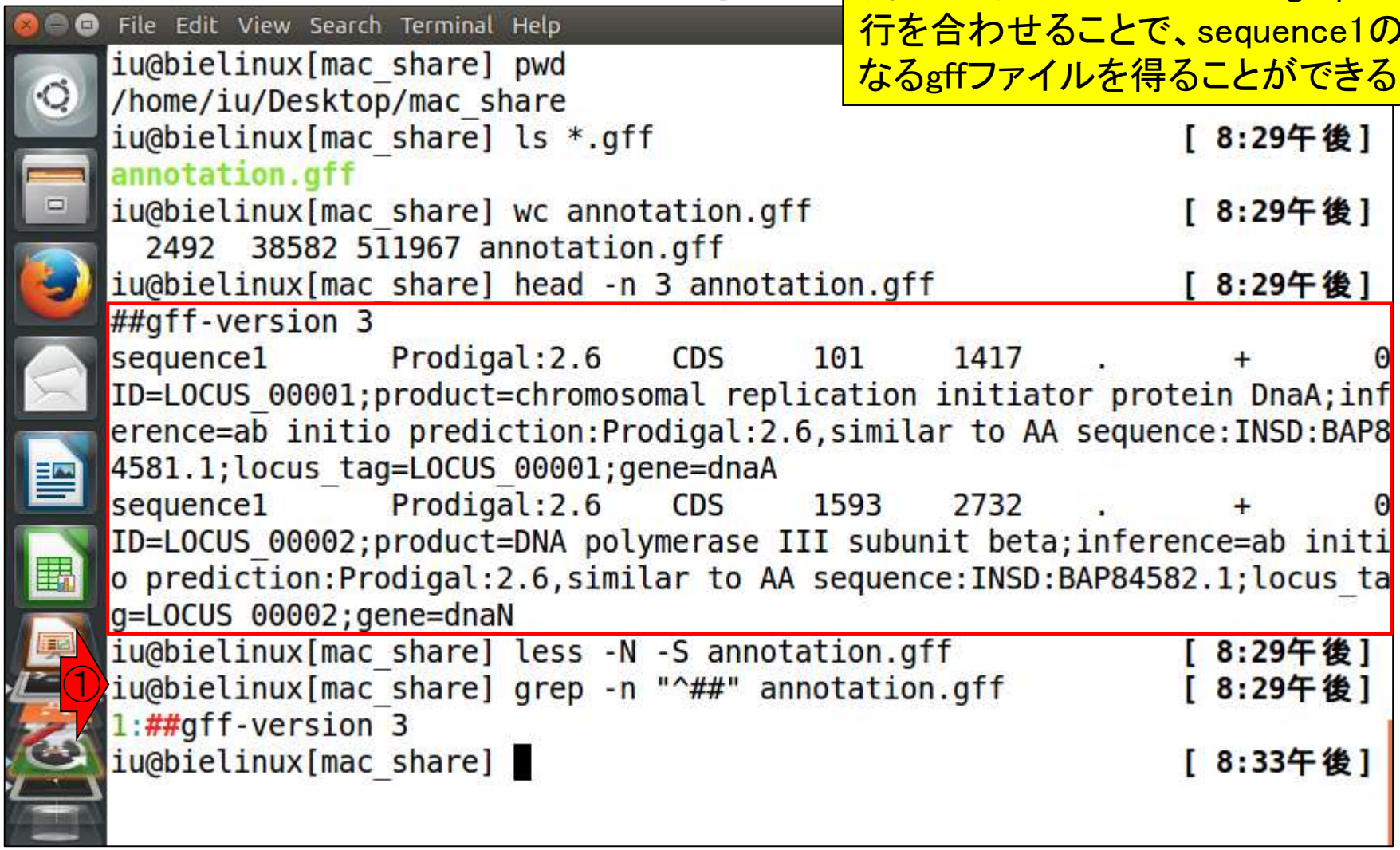
W19-2：less復習

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



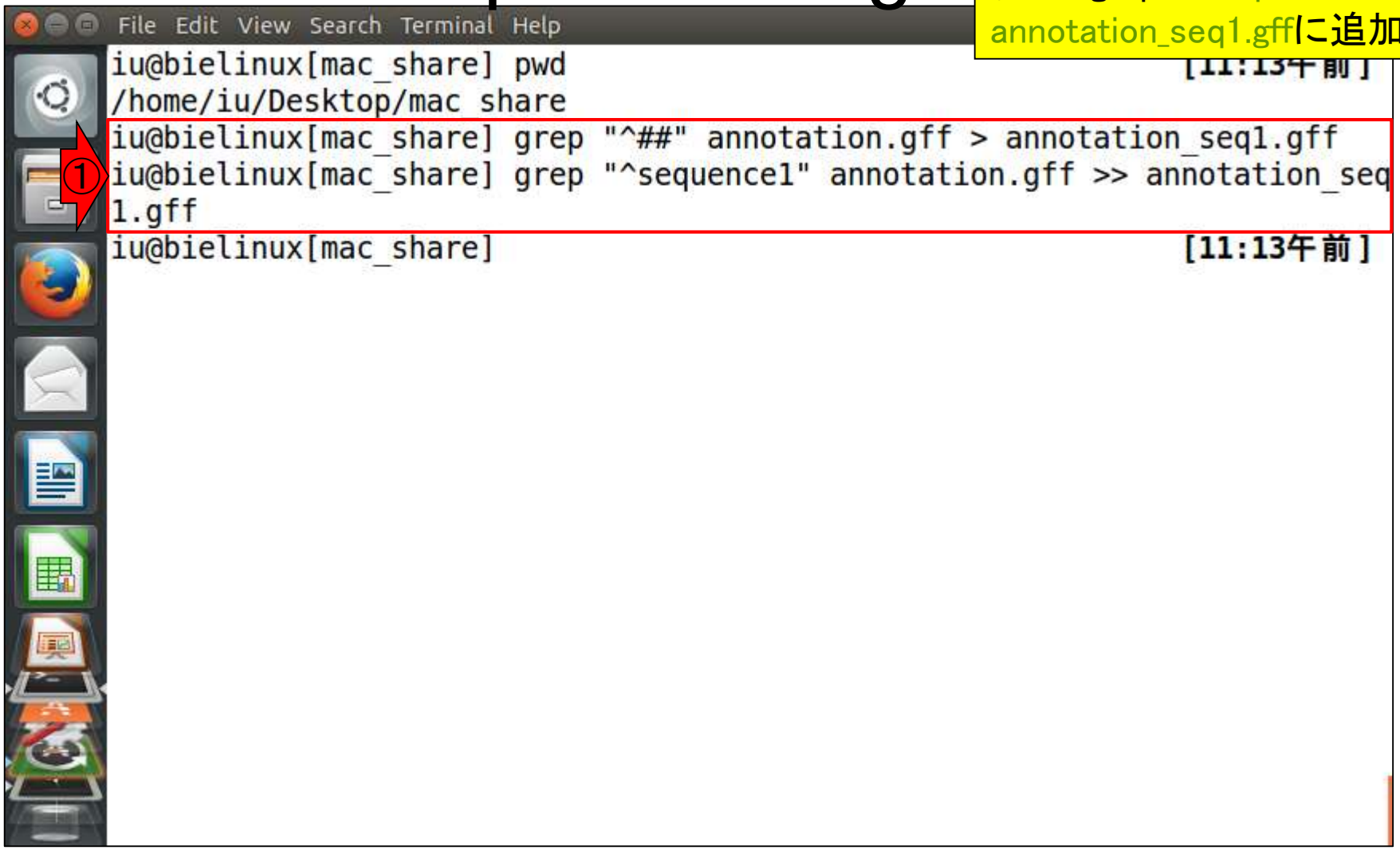
W19-3：ヘッダー部分

①grepの結果からもわかるように、”^##”でannotation.gff中のヘッダー行情報を取り出せる。従って、このヘッダー行と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…



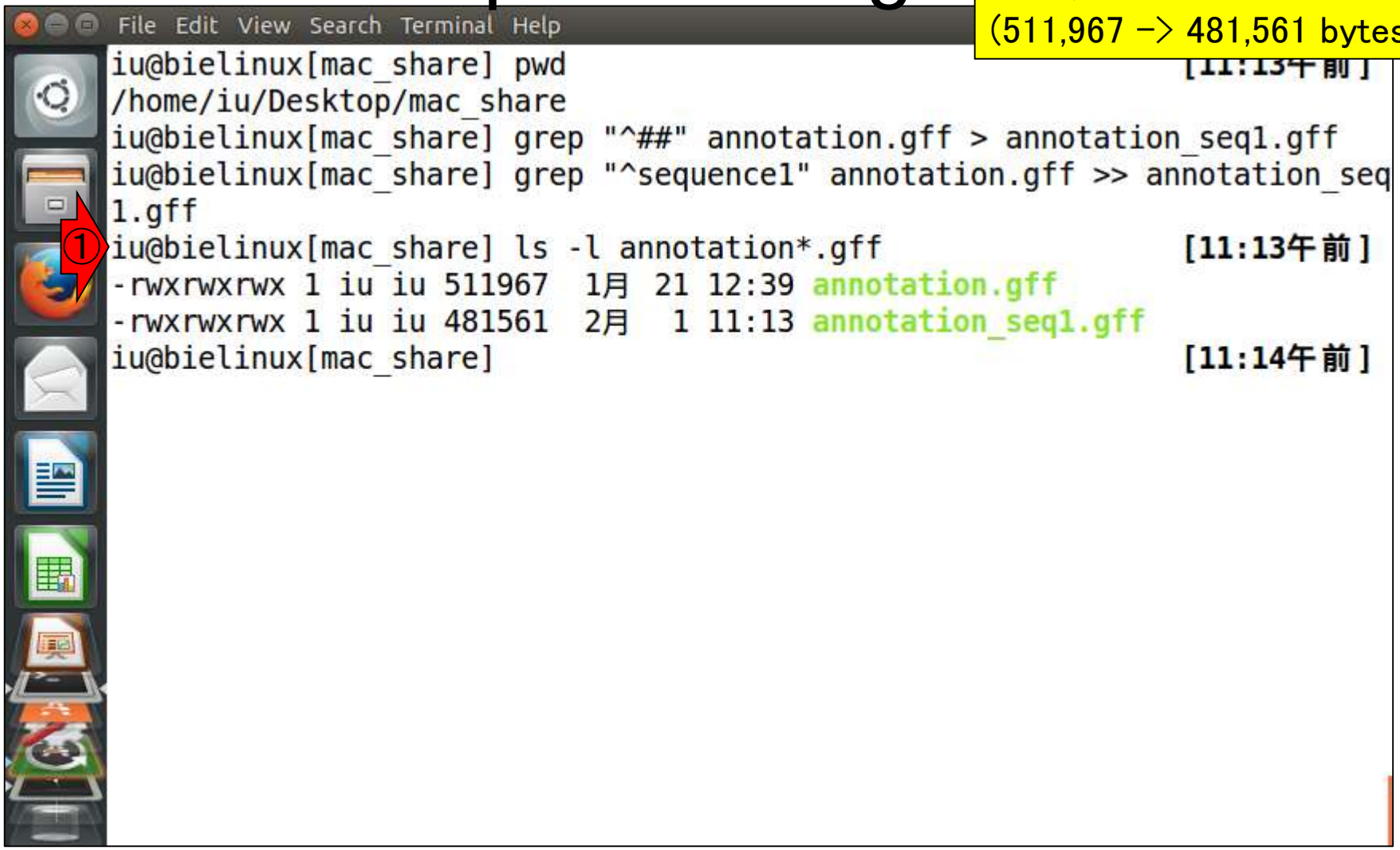
W19-4：sequence1のgff

①最初のgrep “^##”で、gffのヘッダー行を取り出してannotation_seq1.gffに保存。次にgrep “^sequence1”を満たす行をannotation_seq1.gffに追加保存



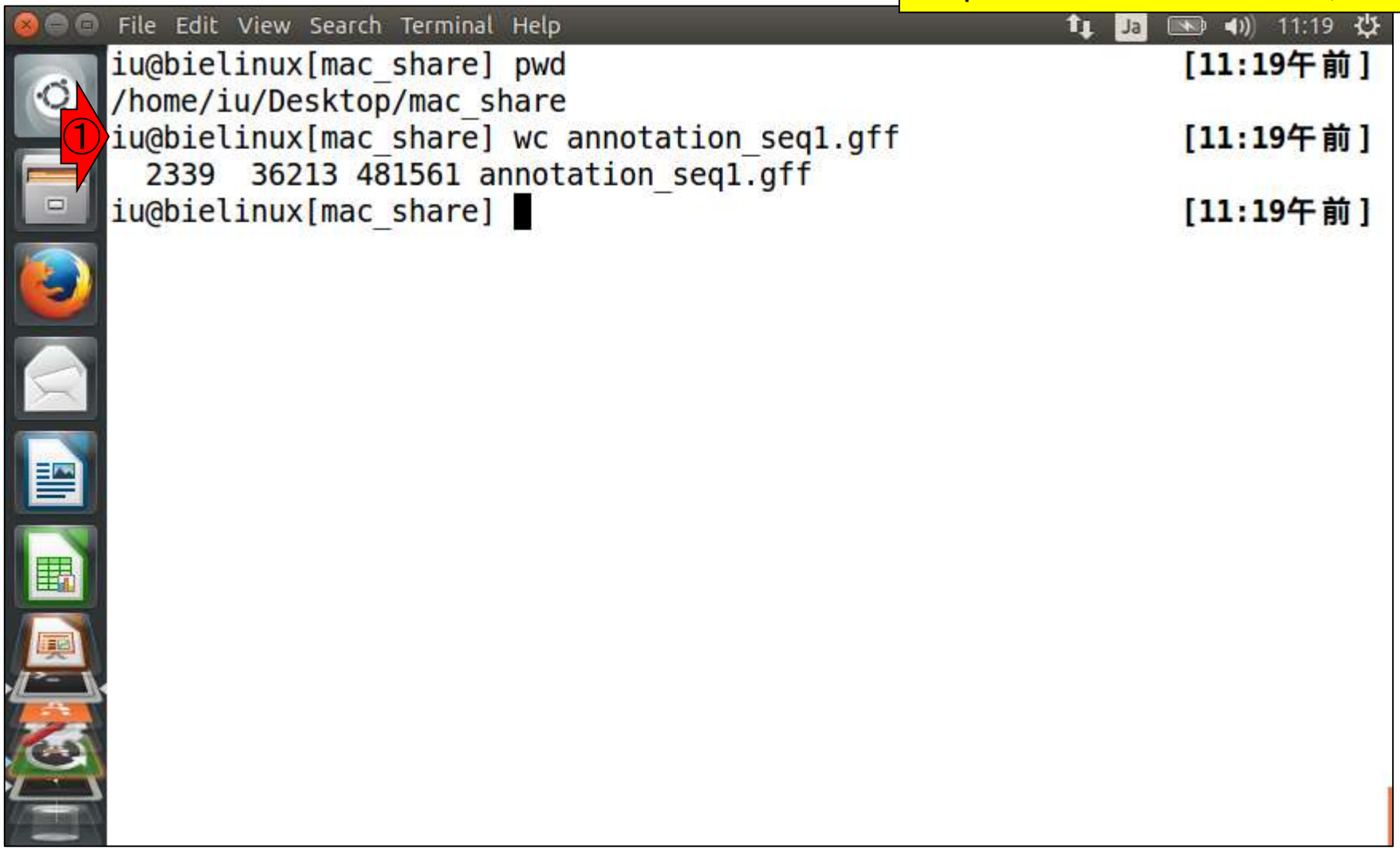
W19-4：sequence1のgff

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



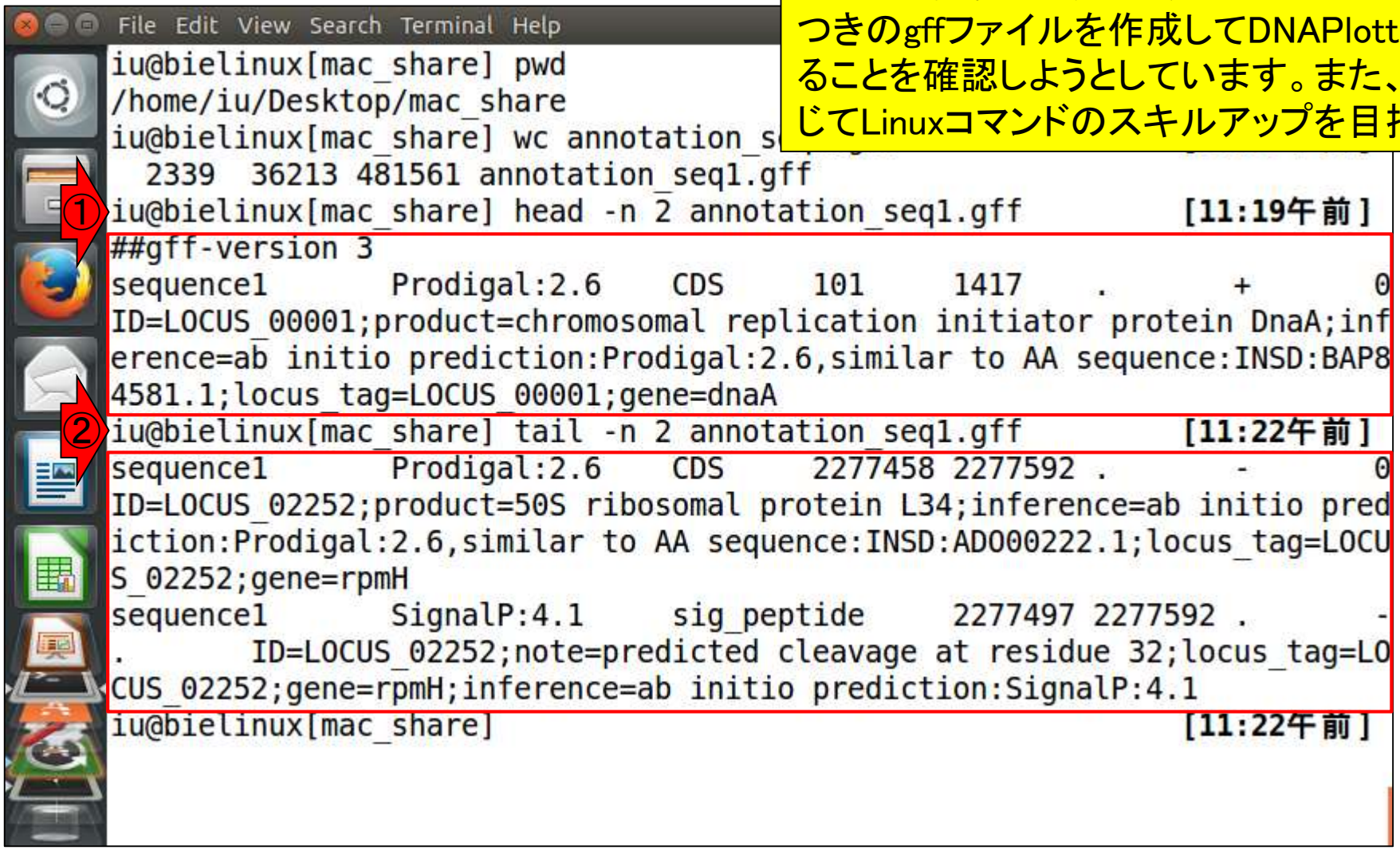
W19-5：中身を確認

①annotation_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



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、配列情報つきの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

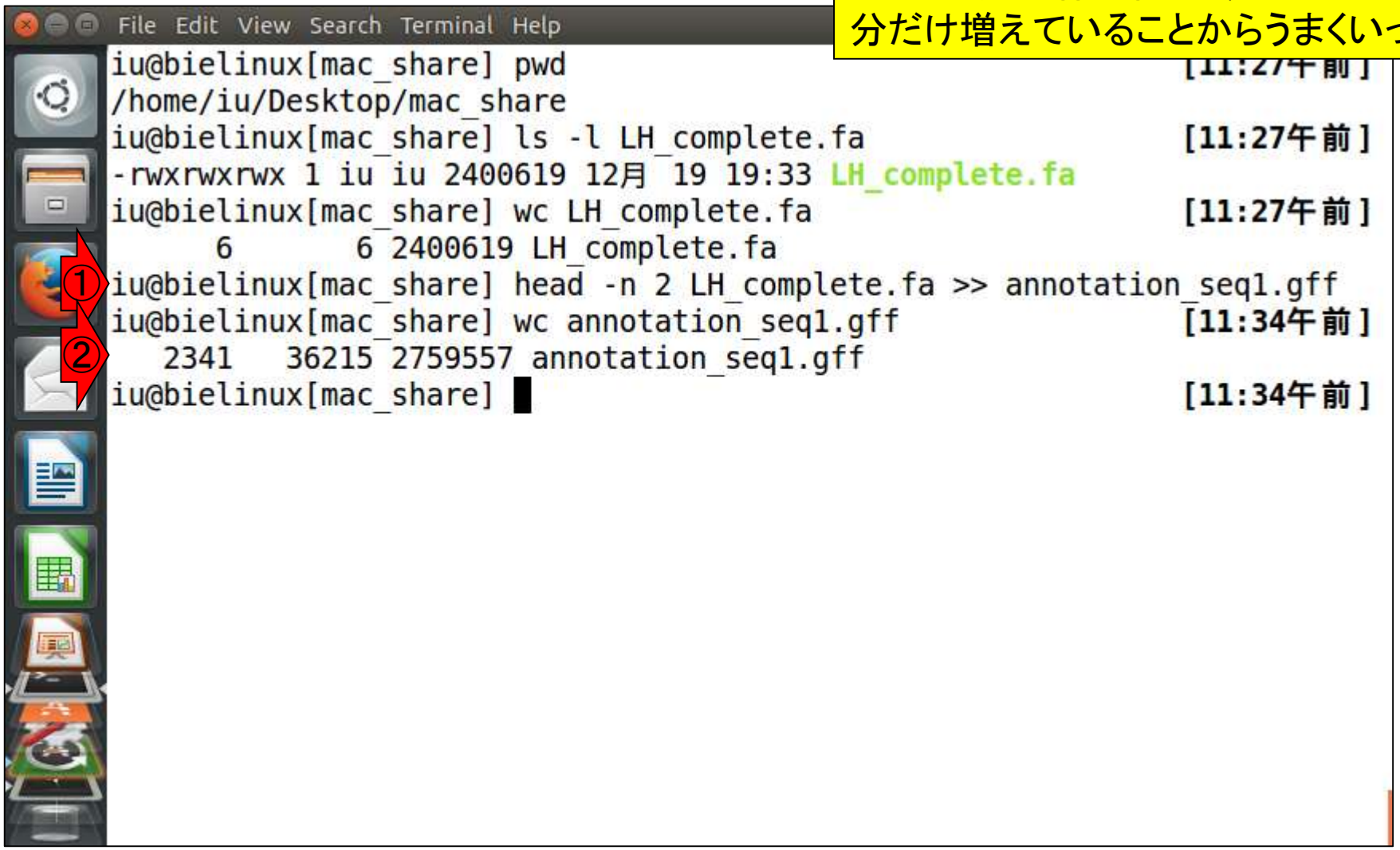


W19-6：配列を結合

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

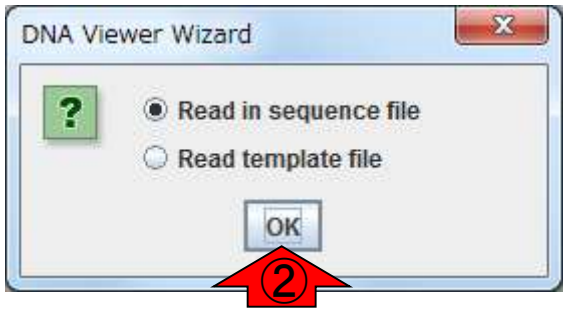
W19-6：配列を結合

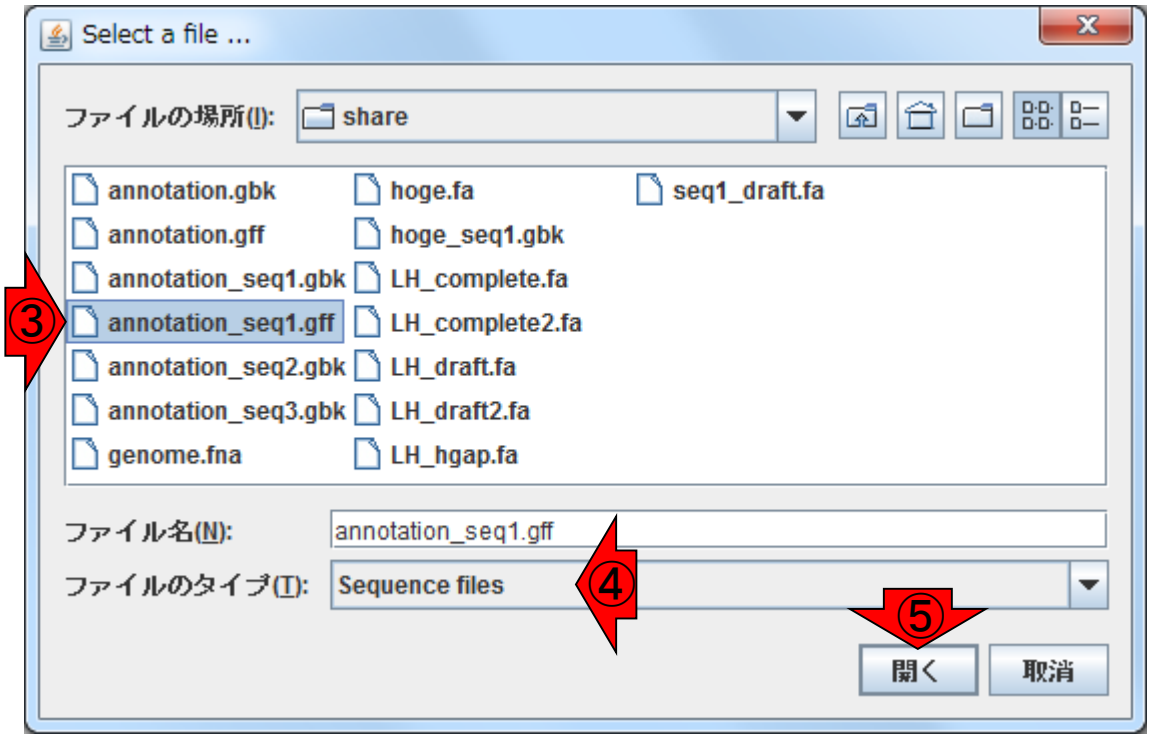
①LH_complete.faの最初の2行分を抽出してannotation_seq1.gffに追加保存。②その後のwcコマンドで追加保存前の2,339行から2,341行と2行分だけ増えていることからうまくいっていると判断



W19-7：DNAPlotter

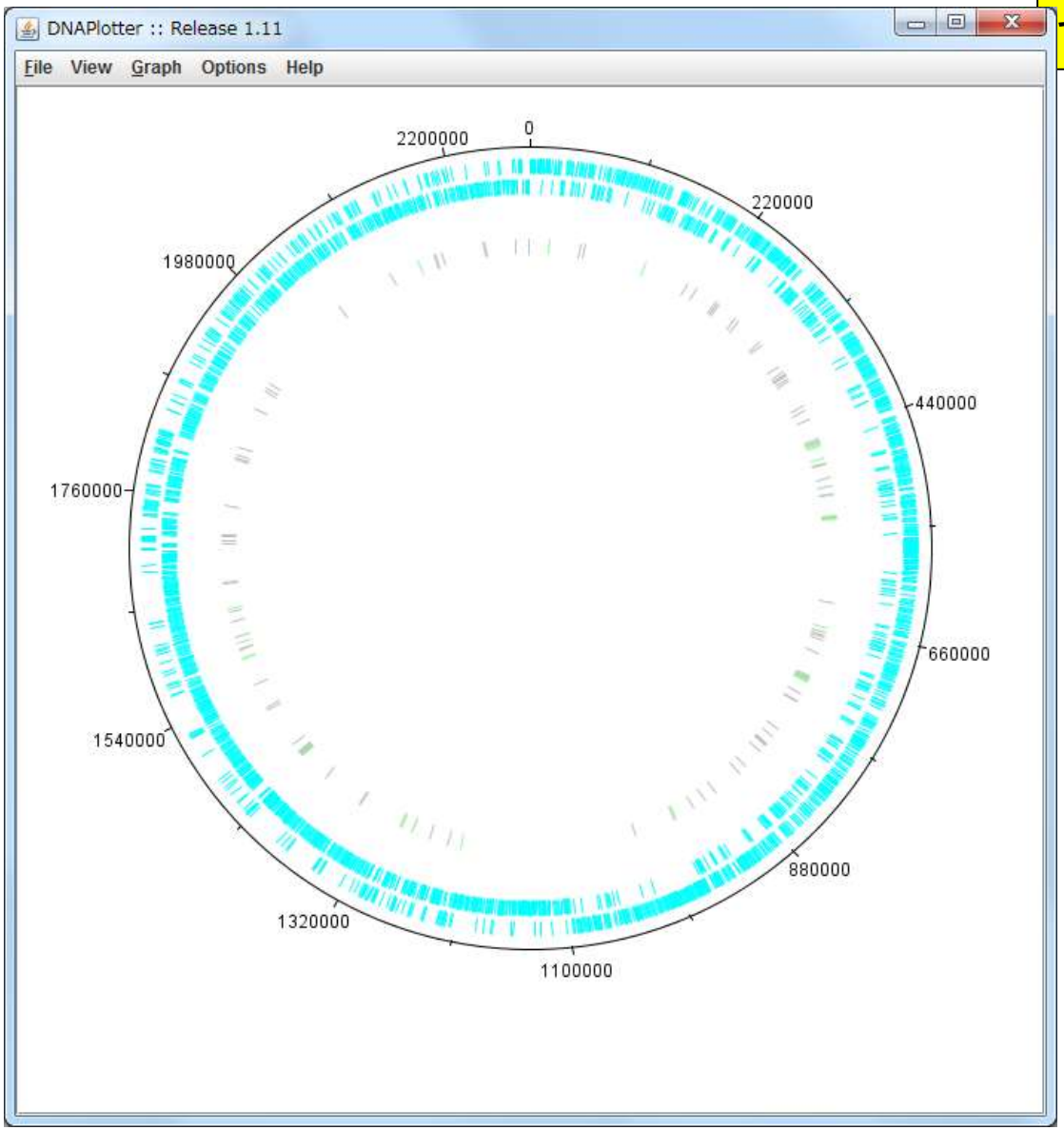
①②DNAPlotterを起動して、作成した③annotation_seq1.gffを、④ファイルのタイプはSequence filesのままでよいので、⑤開く

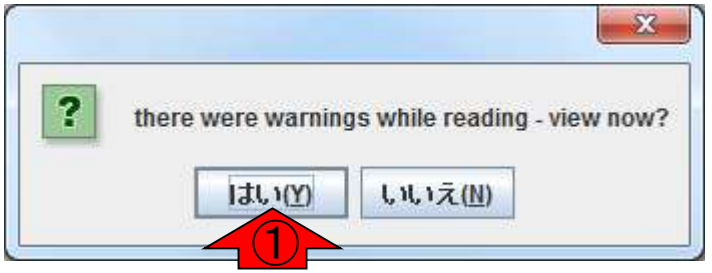




W19-7：DNAPlotter

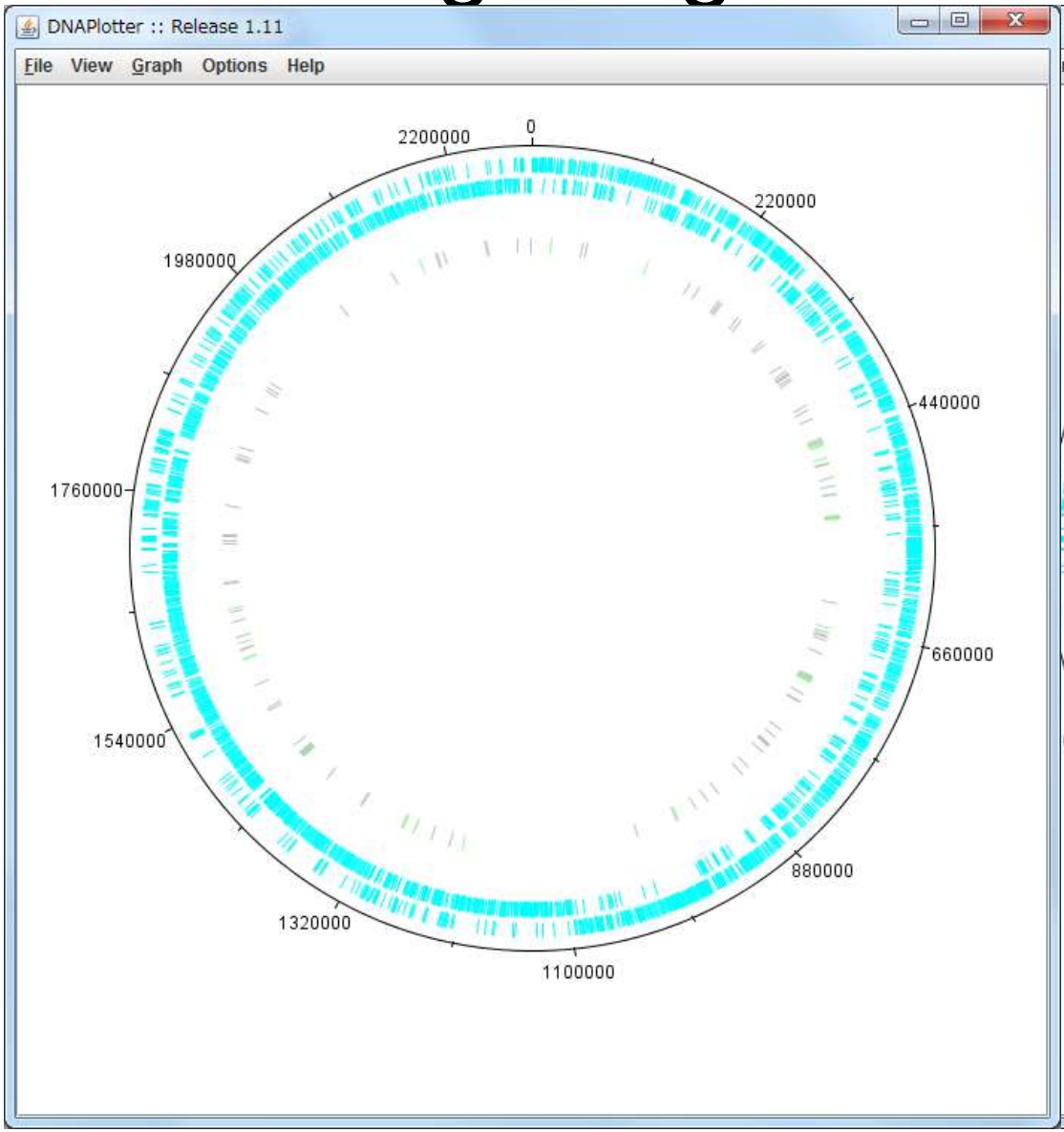
①なんか多少警告は出るが気にしない。gbkファイルを読み込んで得られたプロット(W12-3)と若干の違いはありますが、配列情報つきのgffで無事読み込めたことがわかります。

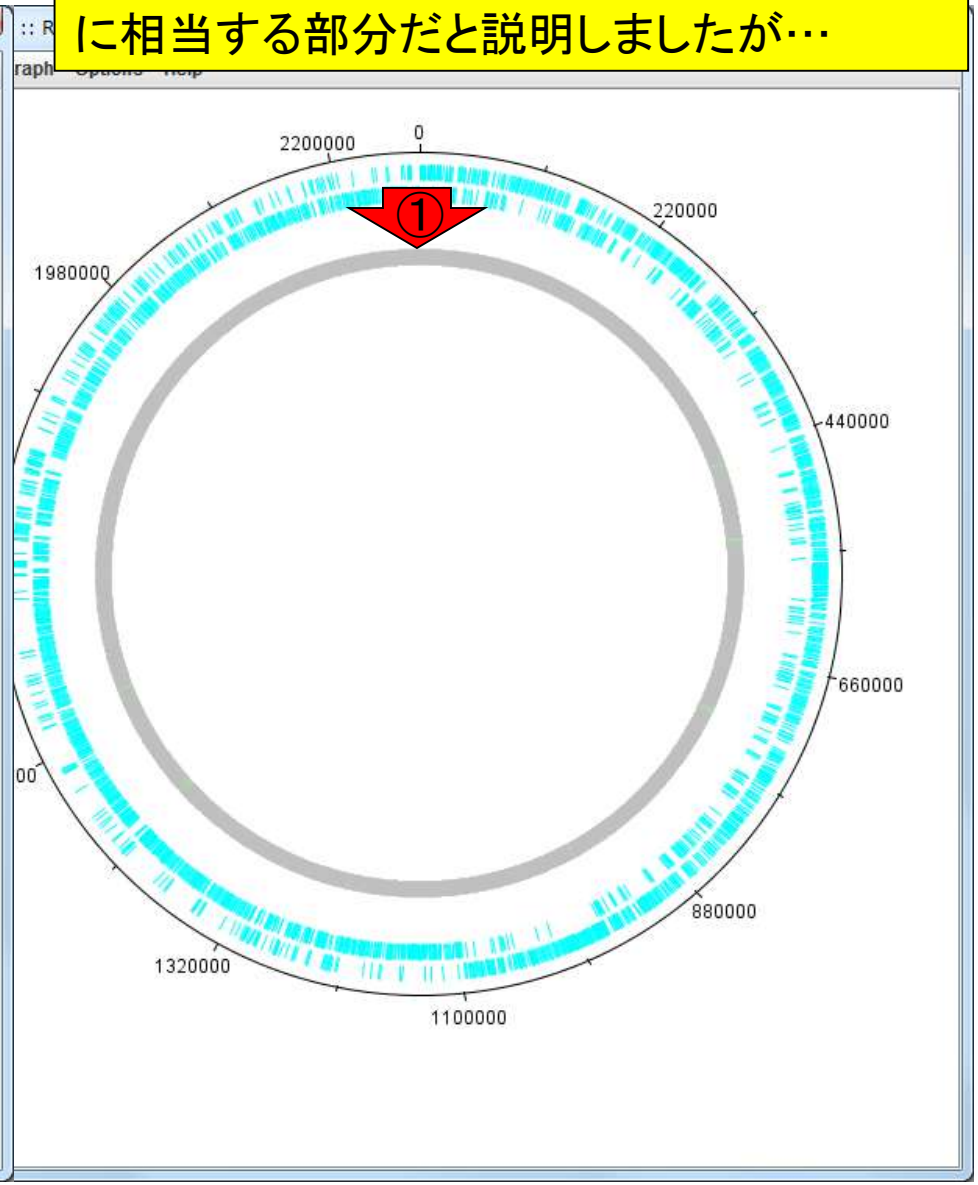




W19-8：gbkとgffの違い

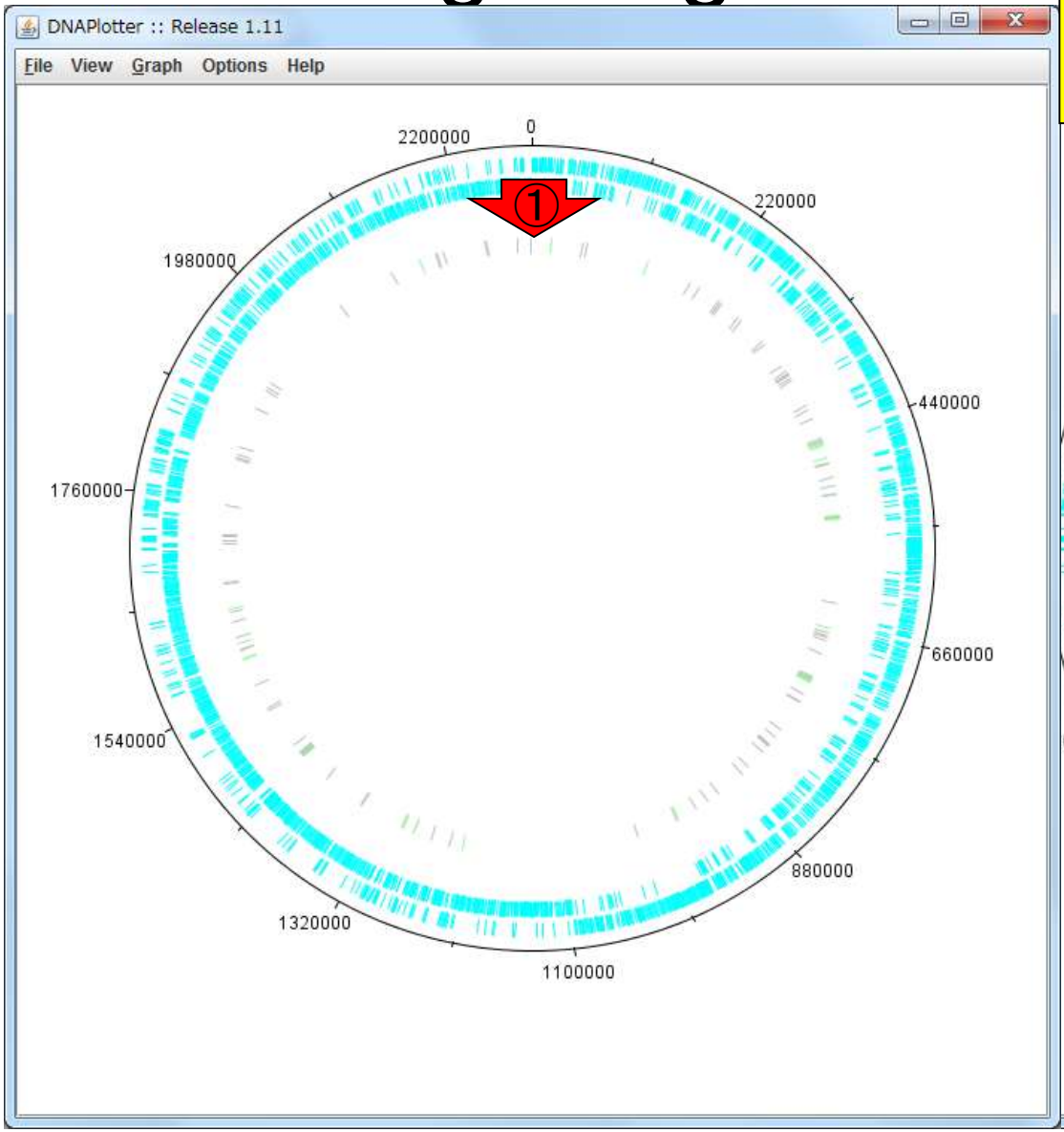
gbkファイルを読み込んで得られたプロット（右）と比較すると、①の部分に違いがあることがわかります。これは、W12-10ではtRNAに相当する部分だと説明しましたが…

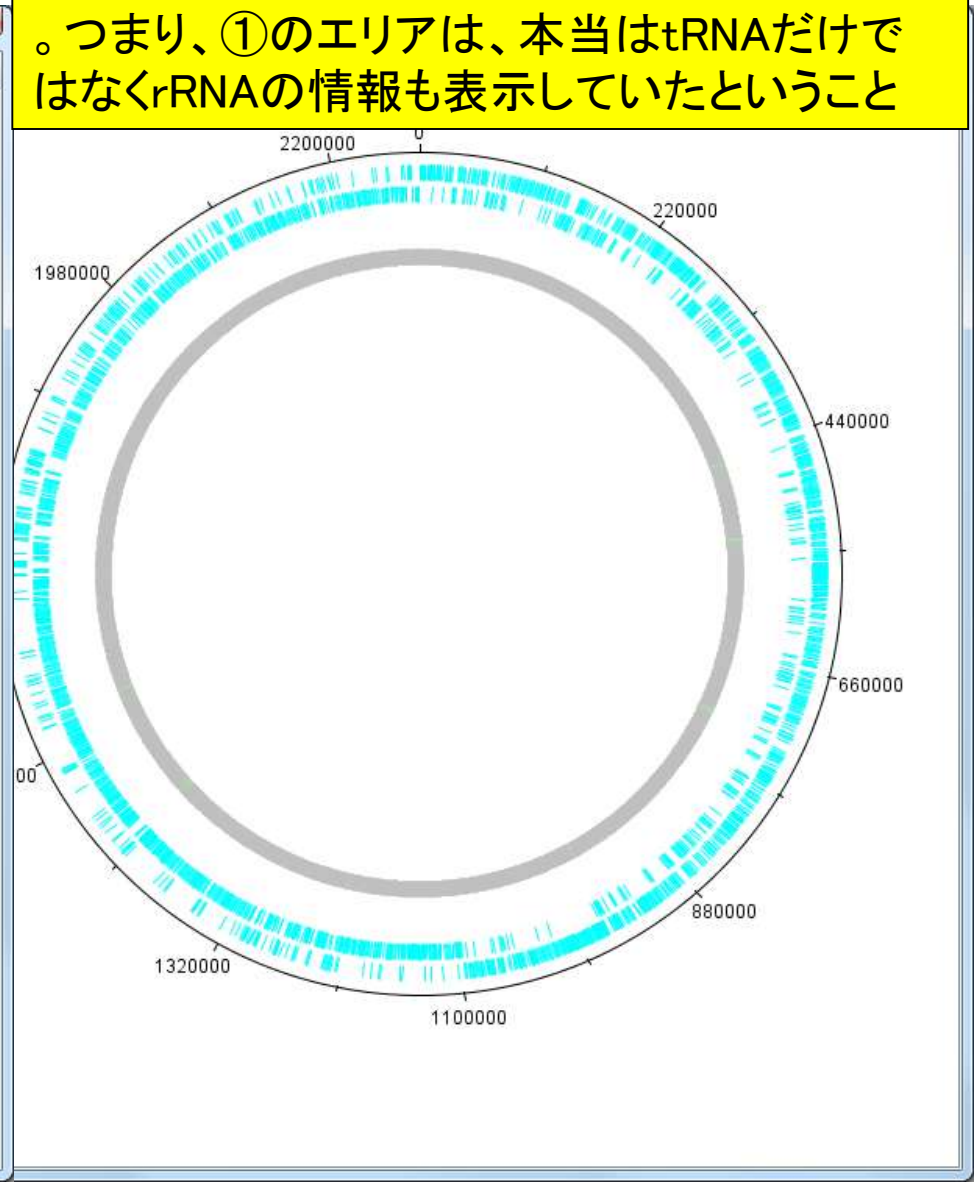




W19-8：gbkとgffの違い

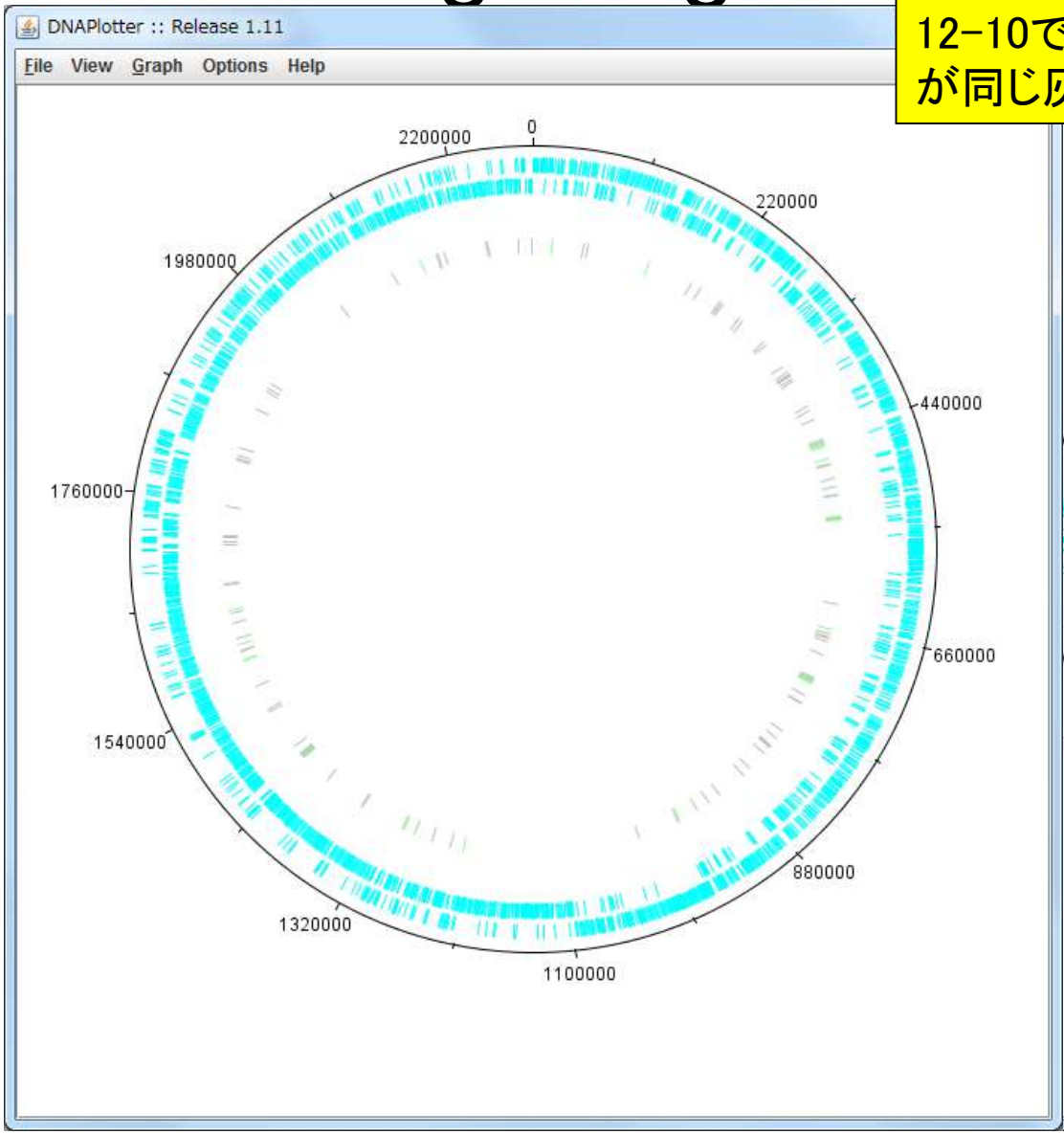
配列情報つきのgffファイルを読み込んで得られたプロット（左）を眺めると、①rRNAが灰色、そしてtRNAが緑色で表示されているようです。つまり、①のエリアは、本当はtRNAだけではなくrRNAの情報も表示していたということ

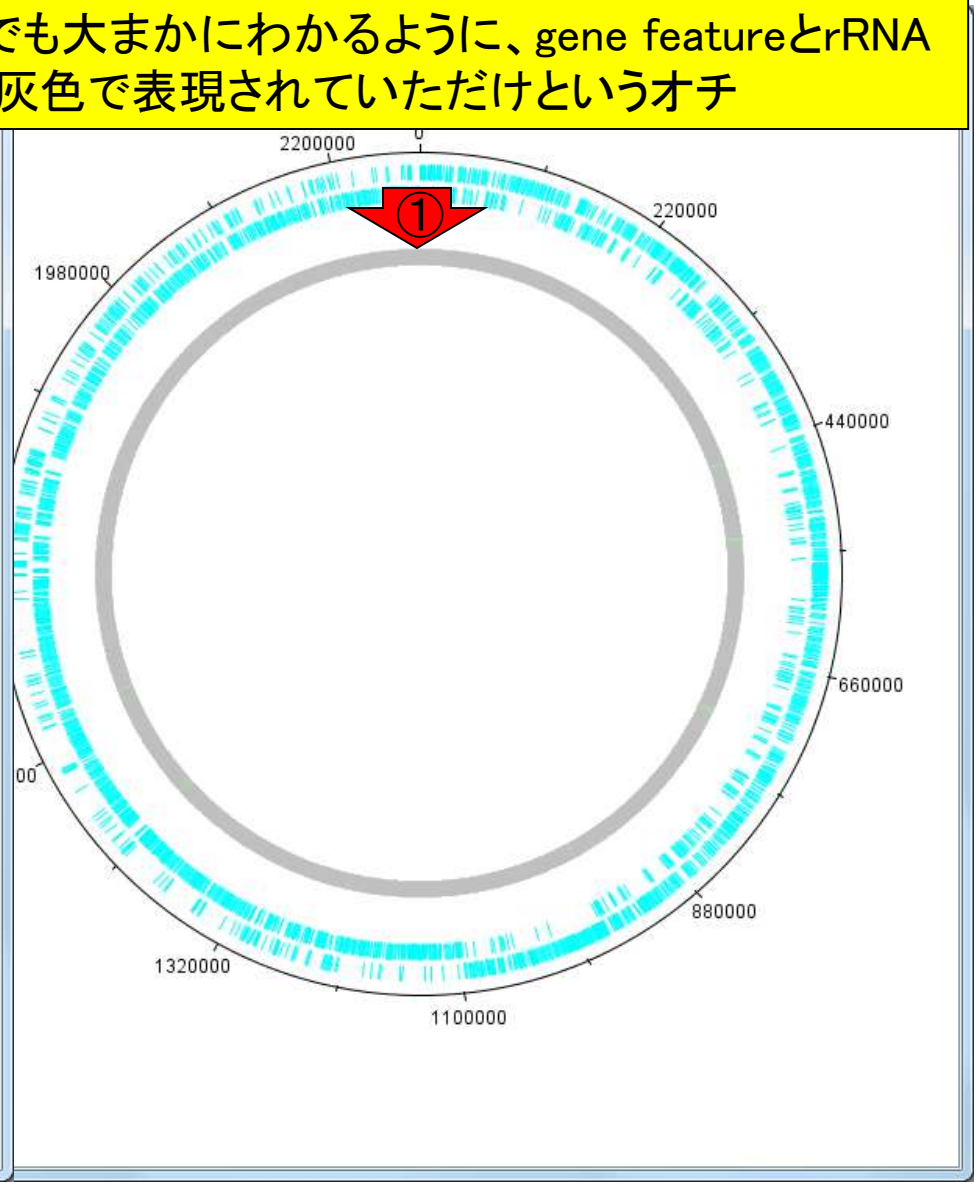




W19-8：gbkとgffの

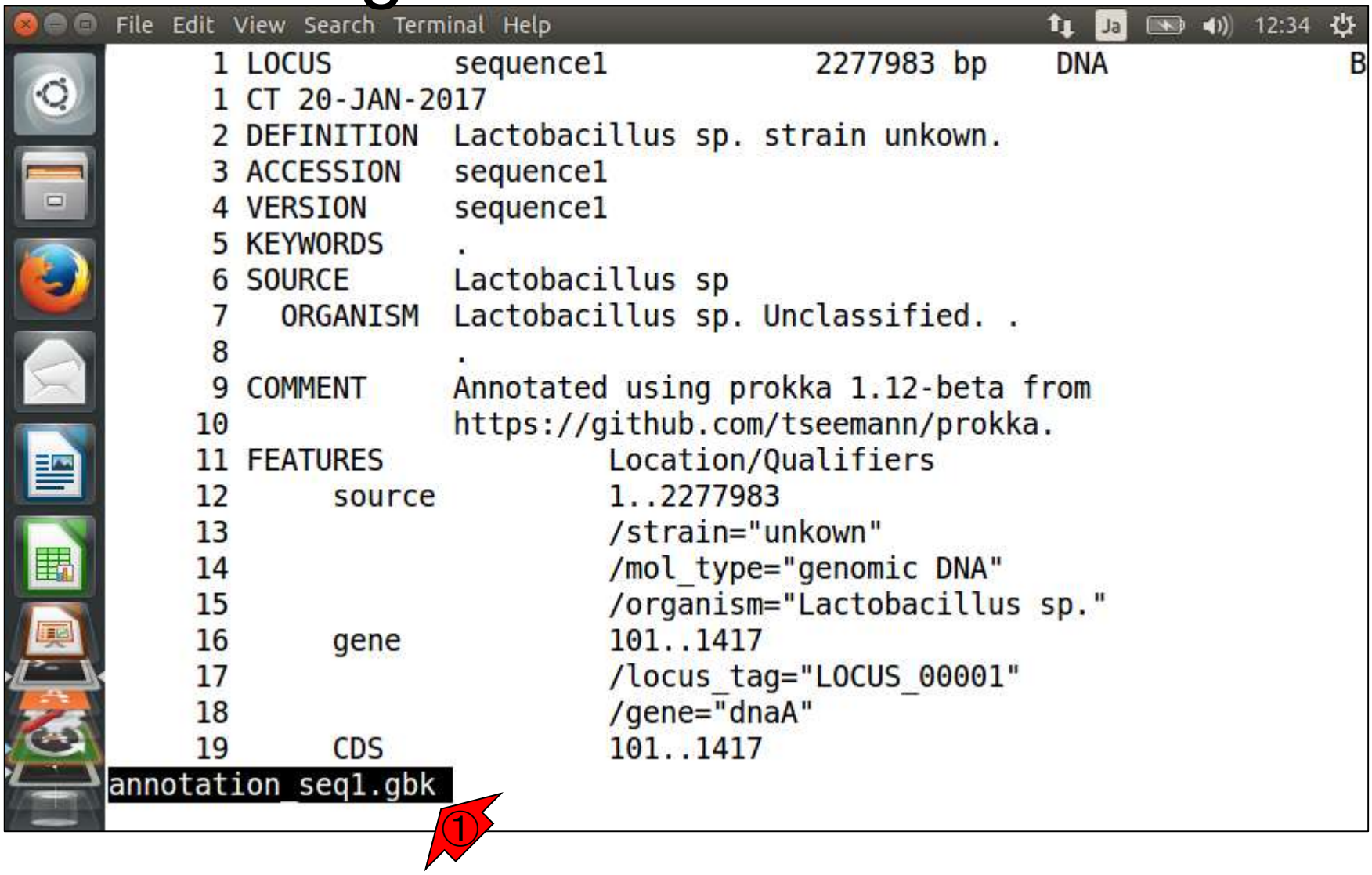
確かにW12-10で見られるようにDNAPlotterのウェブページ上でも、tRNAとrRNAが一体であるような印象を受けます。①が灰色ばかりになっていたのは、W12-9や12-10でも大まかにわかるように、gene featureとrRNAが同じ灰色で表現されていただけというオチ





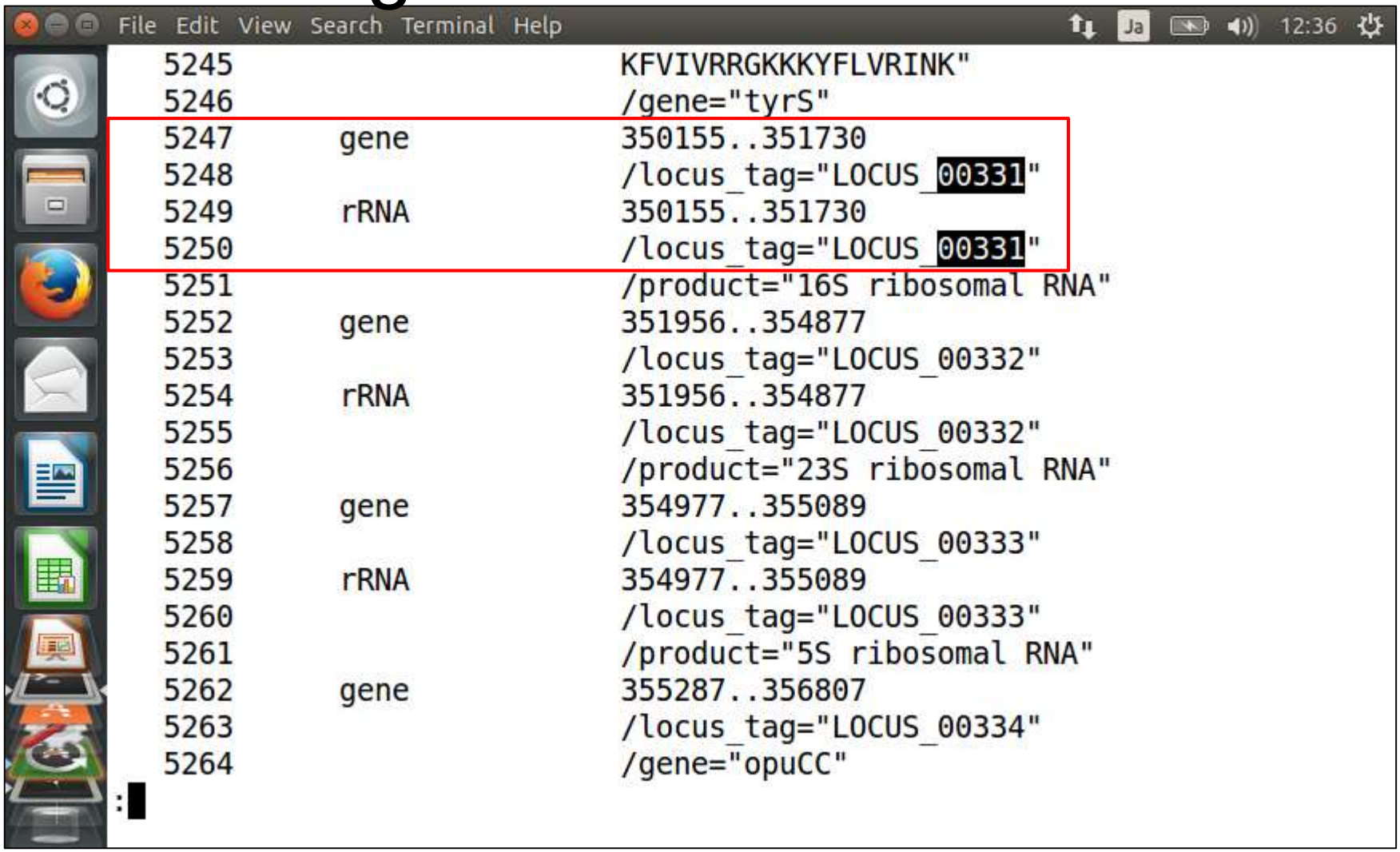
①annotation_seq1.gbkをless -Nで開いたところ

W19-9：gene feature



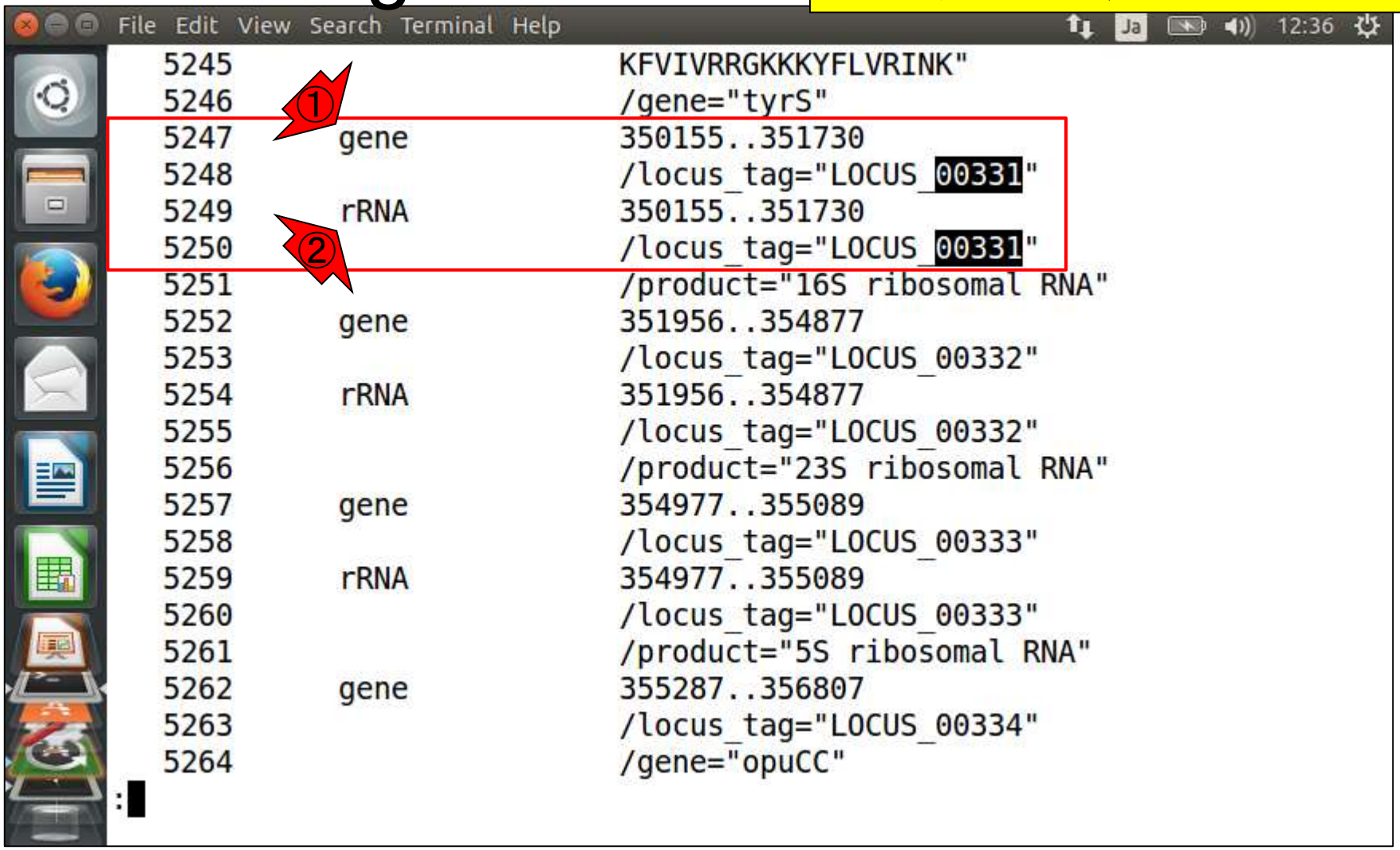
W19-9：gene feature

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

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 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

File Edit View Search Terminal Help

```
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
:
```