Japanese Journal of Lactic Acid Bacteria

解　説

次世代シーケンサーデータの解析手法 第10回 DDBJへの塩基配列の登録（実践編）

谷澤 靖洋[1]、藤澤 貴智[1]、真島 淳[2]、李 慶範[2]、遠野 雅徳[3*]、坂本 光央[4,5]、大熊 盛也[4]、中村 保一[1,2*]、清水 謙多郎[6]、門田 幸二[6*]

[1] 国立遺伝学研究所 生命情報研究センター
[2] 国立遺伝学研究所 DDBJ センター
[3] 農業・食品産業技術総合研究機構 畜産研究部門
[4] 理化学研究所 バイオリソースセンター 微生物材料開発室
[5] 日本医療研究開発機構 PRIME
[6] 東京大学 大学院農学生命科学研究科

多くの学術雑誌は、その論文中で決定された塩基配列を国際塩基配列データベース（International Nucleotide Sequence Database Collaboration; INSDC）に登録することを掲載の条件としている。DNA Data Bank of Japan（DDBJ）は、INSDC を構成する日米欧三極の１つである。本稿では、DDBJ への登録用アカウントの取得から、*Lactobacillus acidipiscis* JCM 10692^T 株のゲノム配列登録作業について解説する。ウェブサイト（Rで）塩基配列解析（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html）中に本連載をまとめた項目（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_book_JSLAB）が存在する。ウェブ資料（以下、W）や関連ウェブサイトなどを効率的に活用してほしい。

Key words：D-way, INSDC, DDBJ, MSS, JCM

はじめに

本稿で用いる *L. acidipiscis* は、魚を塩漬けした発酵食品プラーラー（Pla-ra）から分離され、2000年に新種として提唱された乳酸菌である[1)]。今回は、この基準株 JCM 10692^T のドラフトゲノムの登録を行う。これは、Illumina MiSeq から出力された長さ 300 bp のペアエンドリードを、バクテリア用 *de novo* アセンブリプログラム Platanus_B（ver. 1.1.0）[2)] を実行して得られたものであり、全部で327件のコンティグ配列からなる。本稿では、DDBJ Fast Annotation and Submission Tool（DFAST）[3)] によるゲノムアノテーションを行った結果を用いて DDBJ[4)] への登録作業を行う。尚、生のリードデータの DDBJ Sequence Read Archive（DRA）[5)] への登録は任意であるため、今回は行わない。したがって、DDBJ への登録に必要な情報は、**図1** に示す DFAST 実行結果が基本となる［W1-1］。

DDBJ への塩基配列の登録に必要なアカウントは、DDBJ の登録ポータルである D-way（DDBJ Submission Portal D-way; https://trace.ddbj.nig.ac.jp/D-way/）から取得できる［W2-1］。「Register for a new account」のリンク先から、新規アカウントの取得およびパスワードの設定を行うことで、D-way にログインすることができる。DRA にデータを登録する場合は、取得したアカウントに center name と公開鍵を予め登録しておく（紐づけしておく）必要があるので注意されたい［W4-2］。DRA への登録の詳細については、DRA のトップページから辿れるハ

*To whom correspondence should be addressed.
Phone ：+81-3-5841-2395
Fax ：+81-3-5841-1136
E-mail：yn@nig.ac.jp、tohno@affrc.go.jp
kadota@bi.a.u-tokyo.ac.jp

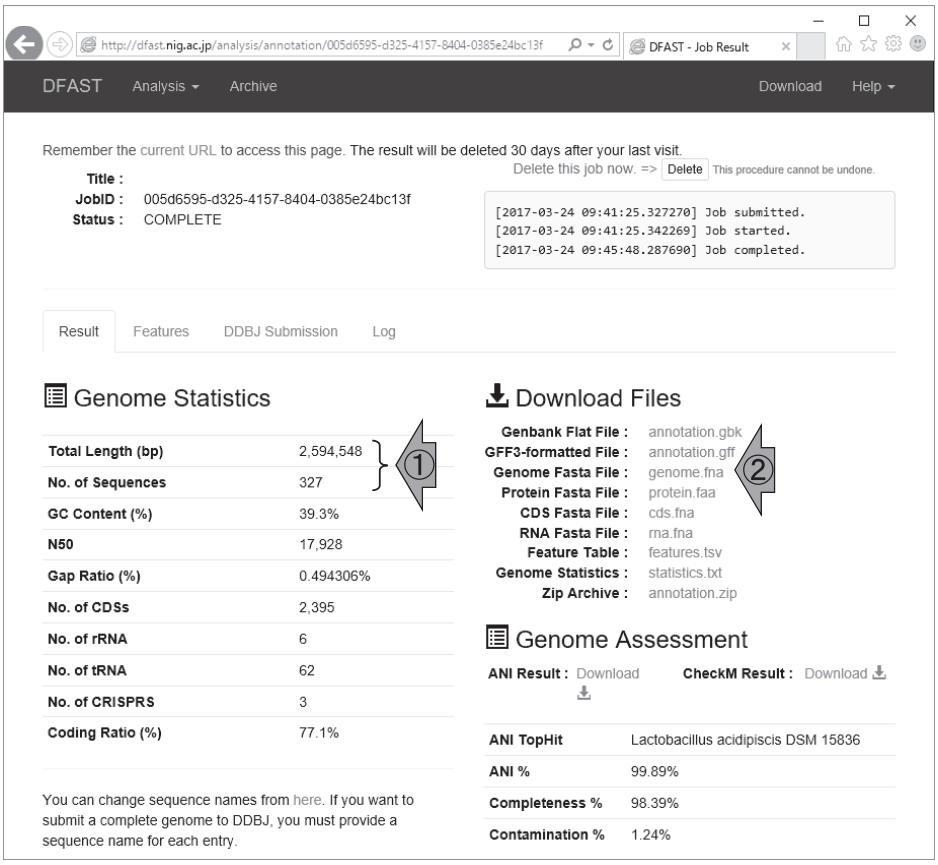

図 1. DFAST 実行結果画面
Platanus_B（ver. 1.1.0）による *de novo* アセンブリ実行結果から、300 bp 以下の配列がフィルタリングされたものが DFAST の入力として与えられている。①その配列の総塩基数は 2,594,548 bp、配列数は 327 個である。入力配列情報は、② genome.fna の中身と同じである。

ンドブックを参考にしてほしい［W4-3］。

DDBJ への登録の概要

DDBJ センターでは、目的に合わせて複数のデータベース（以下、DB）が運用されている。単独の遺伝子配列や本稿で扱うドラフトゲノムは、DDBJ センターが提供する**塩基配列 DB としての（狭義の）DDBJ** への登録対象となる。これは NCBI が提供している塩基配列 DB である GenBank [6] に相当するものであり、DDBJ センターで運用されている他の DB と区別する意味で **traditional DDBJ** とも呼ばれる。他の DB の例としては、次世代シーケンサーの生データを扱う **DRA** や、主として従来のキャピラリ式シーケンサーの生データを扱う DDBJ Trace Archive（**DTA**）、そして後述する **BioProject** や **BioSample** などが挙げられる。

DDBJ への塩基配列登録手段は 2 つある。1 つは、DDBJ Nucleotide Sequence Submission System（**NSSS**; http://www.ddbj.nig.ac.jp/sub/websub-j.html）と呼ばれる対話型の web 版塩基配列登録システムである［W4-4］。記載する内容がそれほど多くない場合に有効なツールであり、単独の遺伝子配列など小規模な登録に適している。そしてもう 1 つは、ゲノム全体の配列など大規模な登録に適した Mass Submission System（**MSS**; http://www.ddbj.nig.ac.jp/sub/mss_flow-j.html）である［W4-5］。DFAST で作成可能な DDBJ 登録用ファイルは MSS 用のものであり、本稿でも MSS を利用した DDBJ への登録手順について解説する。論文を投稿する直前に登録手続きを行おうとしても、予想外に時間がかかってしまうこともある。このため、全体像を理解し早めに登録を終えておくことが大切である。登録の手順は、大まかに以下の通りである：

1. BioProject の登録
2. BioSample の登録
3. MSS への登録ファイルの送付

BioProject（http://trace.ddbj.nig.ac.jp/bioproject/index.html）は、研究プロジェクトとそのプロジェクトに由来するデータをまとめるための DB である［W4-6］。近年、大規模な研究プロジェクトが増え、大量のデータが生み出されている。それに伴い、データの種類に応じて複数の関連 DB に分散して登録されることも増えてきた。BioProject

のメリットは、これらのデータを包括的に扱えるという点である。例えば比較ゲノムのプロジェクトであれば、プロジェクト内に複数のゲノムデータが紐付けられることになる。RNA-seq のプロジェクトであれば、例えば転写産物の配列と発現量データが紐付けられることになる。

BioSample（http://trace.ddbj.nig.ac.jp/biosample/index.html）は、データが由来するサンプルについての情報を収集した DB である［W4-7］。BioSample に登録されたデータには、生物種名だけでなく、その生物試料が採取された環境、疾患や個人の医学的情報といった表現型情報を記述するための**サンプル属性**が含まれる。現在、この属性名は Genomics Standard Consortium（GSC）が規定した Minimum Information about any（x）Sequence（MIxS）に準拠した詳細な記述ができるようになっている[7]［W4-8］。以下では、最もシンプルな登録形態として、1 つの BioProject の下に 1 つの BioSample データとそれに由来するゲノム塩基配列が存在する場合の登録手順を紹介する。

BioProject の登録

BioProject の登録は、D-way を通して行う（W5-1;**図 2**）。ログイン後、BioProject の New submission ボタンを押し、必要事項をフォームに入力しながら進めていく。主な登録内容は、登録者名［W5-3］、プロジェクト名とその説明、文献情報（通常は投稿する予定の論文の情報）などである。必要であればグラント情報を記入することもできるので、事前にグラントの課題番号などを手元に用意しておくとよい。

PROJECT TYPE タブでは、プロジェクトの種類を指定する［W5-10］。今回我々はゲノム配列データを登録するので、Project data type では "Genome Sequencing" にチェックを入れる［W5-11］。それ以外の項目についてもプルダウンメニューやチェックボックスから該当する選択肢を選んでいけばよい。第 9 回[8] で述べた Locus tag prefix の取得申請もできるので、希望のもの（ここでは **Lacidipiscis** とした）を入力する［W6-1］。TARGET タブでは、対象生物に関する情報を入力する。単独のゲノムであれば、その生物種名を記入すればよい。比較ゲノムのプロジェクトであれば、属名など共通する階層までの分類名を記入するケースもあるだろう。生物種名情報は最低限必要であるが、オプションとしてグラム染色性・運動性・細胞形態なども記入することができる［W6-2］。原著論文があれば、PubMed ID や DOI 情報を入力する。全体をチェックし、問題がなければ Submit ボタンを押して終了となる［W6-6］。このときは、約 2 日後に BioProject のアクセッション番号 **PRJDB5682** が発行された［W6-7］。

BioSample の登録

引き続き BioSample の登録を行う。登録作業中に BioProject ID を指定する欄があるので、先に取得した PRJDB5682 を入力する。これにより BioProject と BioSample が紐づけられる。なお、BioProject ID については登録途中で発行される PSUB で始まる Submission ID（今回の場合は "PSUB007096"［W5-2］）を仮 ID として入力することができるため、登録手続き完了を待つことなく BioSample の登録を並行して進めることができる。登録手続きは、BioProject と同様に D-way にログイン後、BioSample の New submission ボタンを押して開始する［W7-1］。SUBMITTER と GENERAL INFO については、BioProject と同じ要領で進めていけばよい

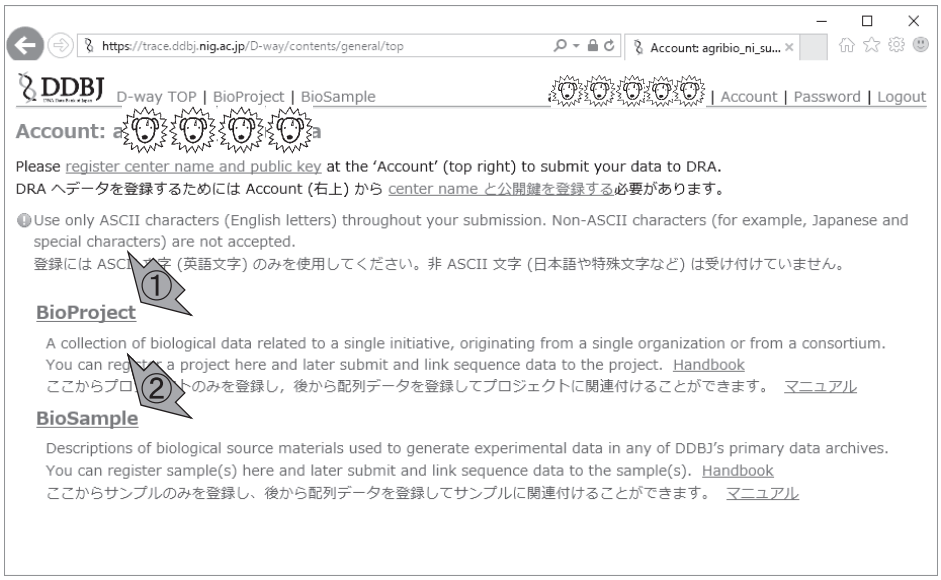

図 2. D-way ログイン後の画面
① BioProject と② BioSample の登録はここから行う。

［W7-3］。SAMPLE TYPE では、前述の MIxS [7] に準拠した Core Package と Environmental Package を候補の中から選択する［W7-4］。今回のように単離した菌株のゲノムデータの場合、Core Package については「Genome, metagenome or marker sequences」、そして「Cultured Bacterial/Archaeal Genomic Sequences」を選択すればよい。Environmental Package については、「miscellaneous or artificial」を選択した。

ATTRIBUTE では、サンプル属性情報の登録を行う［W8-1］。サンプル属性の種類は、前の画面で指定した package によって異なる。入力項目数も非常に多いため、画面での入力ではなくひな形となるファイル（MIGS.ba.miscellaneous.txt）をダウンロードして記入後にアップロードする形態となっている。ファイルはタブ区切りテキスト形式なので、著者らは Excel で読み込んで作業を行う［W8-2］。属性ごとの定義と書式に関する説明などを参考にしながら［W8-3］、必須項目を中心に埋めていく［W8-5］。入力すべき内容が不明な場合には、"not applicable" や "missing" などを使い分けて入力すればよい［W8-7］。非必須項目においても、株名（strain）や菌株保存機関における登録番号などは記入しておいたほうがいいだろう［W8-9］。菌株登録番号は、culture_collection という項目に入力することが推奨されている。この属性はテンプレートには含まれていないが、自分で項目を追加して記入することができる。使用可能なサンプル属性の一覧は、ヘルプ画面で "List all sample attributes" を押せばよい［W8-3］。

一通り入力を終えたタブ区切りテキストファイル（MIGS.ba.miscellaneous_after.txt［W8-10］）をアップロードし、全体をチェックして問題がなければ、Submit ボタンを押して終了となる［W8-13］。このときは、約 2 日後に BioSample のアクセッション番号 **SAMD00078754** が発行された［W9-1］。メール本文をよく見ると、DDBJ アノテーターによって属性の計 3 か所で修正がなされていることがわかる［W9-2］。今回の submission では、ひな形には存在していなかった isolation_source という項目に "fermented fish（Pla-ra）" が、そして type_strain という項目に "yes" が追加された。また、host という項目中に記載していた "fermented fish" が削除された。isolation_source はサンプルの分離源を表すものであり、微生物サンプルの登録においては使用頻度の高い項目といえる。次回以降同様の登録を行う場合は、これらを予め追記しておくとよいだろう。

MSS への登録ファイルの送付

MSS を利用した塩基配列およびアノテーション情報の登録を行うべく、「MSS 申し込みフォーム」上で必要事項を入力していく［W10-1］。ゲノム配列を登録する場合は、データ種別の欄で complete genome または draft genome を選択することになる［W10-3］。Complete genome の場合は、系統分類に基づく登録区分に割り振られる。バクテリアの complete genome の場合は、BCT DIVISION への登録となる［W10-7］。今回のデータが該当する draft genome の場合は、Whole Genome Shotgun（WGS）エントリとして取り扱われる［W10-8］。WGS は、整理が不十分な段階の大量の DNA 断片を対象とした登録区分である。アクセッション番号も他の登録区分とは異なる書式で与えられ、アノテーション情報を含まない塩基配列のみの登録も可能である。

ギャップを含んだドラフトゲノムの場合は、2 通りの登録方法がある。1 つめは、ギャップを除いて分割した配列を WGS エントリとして登録し、エントリ間の連結情報を AGP ファイルによって登録する方法（CON エントリという）である［W10-9］。そしてもう 1 つは、ギャップを含んだ配列のまま WGS エントリとして登録し、assembly_gap feature を用いてギャップ情報を記載する最近可能となった方法である。DFAST を用いたドラフトゲノムの登録は、後者の手順に沿って行う。一通りの入力を終えて送信すると［W10-6］、ほどなくして DDBJ から手順の詳細に関するメールが届く［W11-1］。メール本文中には「MSS 関連資料、および、ツールダウンロード」および「MSS 用データファイル作成」手順に関する URL を参照しながら登録作業を行うように書かれているが、これらの作業は DFAST 上で行うことができる［W11-2］。

第 9 回でも解説したように、DFAST 実行結果画面の "DDBJ Submission" タブをクリックし、画面の指示に従って入力項目を埋めていけば登録用ファイルを作成できる。前回との違いは、Locus Tag Prefix、BioProject、そして BioSample の項目に、仮ではなく本物の情報を入力した点である［W11-4］。最後の Format Check の結果として、翻訳されたアミノ酸配列の 2 か所で不明アミノ酸残基（X）が含まれている旨の警告が出た［W11-11］。これは不明塩基 N を含む領域に遺伝子が予測された場合に起こりうる。DFAST は、X が連続して 2 つ以上並んだ場合には登録には不適切として結果から削除している。このため、X が含まれていたとしても最大で 1 アミノ酸残基である。X の数が多いと DDBJ のアノテーターから修正を要求されることがあるが、この程度であれば通常は許容範囲である。

チェックが済んだ塩基配列およびアノテーションファイルは、ダウンロード後に添付ファイルとして DDBJ にメールで送付するのが一般的な手順である。今回はファイルサイズが大きいこともあり、DFAST の結果ページの URL を DDBJ にメールで伝える方法で行った［W11-13］。

DDBJ とのやりとり、そして公開

メール送信の約 2 日後、DDBJ より登録内容をチェックした結果が返されてきた。今回の DDBJ からの問い合わ

せ内容は、下記の 3 点であった［W12-2］：

1. アセンブリに用いたソフトウェア Platanus のバージョンは "B 1.1.0" でよいか？
2. rRNA において 5S rRNA のみ予測されているが 16S や 23S は予測できなかったのか？
3. 今回のデータは登録完了後、即日公開でよいか？

我々は下記のように回答した［W12-3］：

1. アセンブリに使用されたソフトは Platanus_B であり、version は 1.1.0。
2. 予測ソフト Barrnap を用いたところ、検出されたものは 5S のみでした。
3. 登録完了後、即日公開でお願いします。BioProject および BioSample ID についても同じタイミングで公開してください。

尚、1. については単純に我々の誤記であり、3. についてはデータ公開予定日（Hold Date）を指定していなかったために確認された事項である。Hold Date を指定すると、指定日になるか公開申請手続きを行うまでは非公開となる。投稿した論文が受理されるまでは非公開扱いにするという指定方法が一般的であろう。

DDBJ からは下記のような返信が寄せられ、この段階でやり取りは終了した［W12-4］：

1.「Platanus_B v. 1.1.0」のように記載する。
2. rRNA 領域の予測について承知した。
3. 登録データと BioProject/BioSample が連携されているので連動公開される。

今回登録したのは、327 コンティグからなる塩基配列である。アクセッション番号は配列ごとに与えられ、DDBJ からの定型メールに添付されたファイル（SAMD00078754_JCM10692.acclist.txt［W12-5］）の中身が、配列ごとのアクセッション番号リストとなっている。今回は、4 月 21 日（金）の夜に DDBJ にメールし［W11-13］、4 月 25 日（火）の夕方に一連のやりとりが終了した［W12-5］。そして 4 月 27 日（木）の未明に BioSample ID（SAMD00078754［W13-2］）および BioProject ID（PRJDB5682［W13-4］）が公開された旨のメールが届いた。**図 3** は、DDBJ のデータベース検索ツール getentry［W13-5］を用いて、最初のコンティグのアクセッション

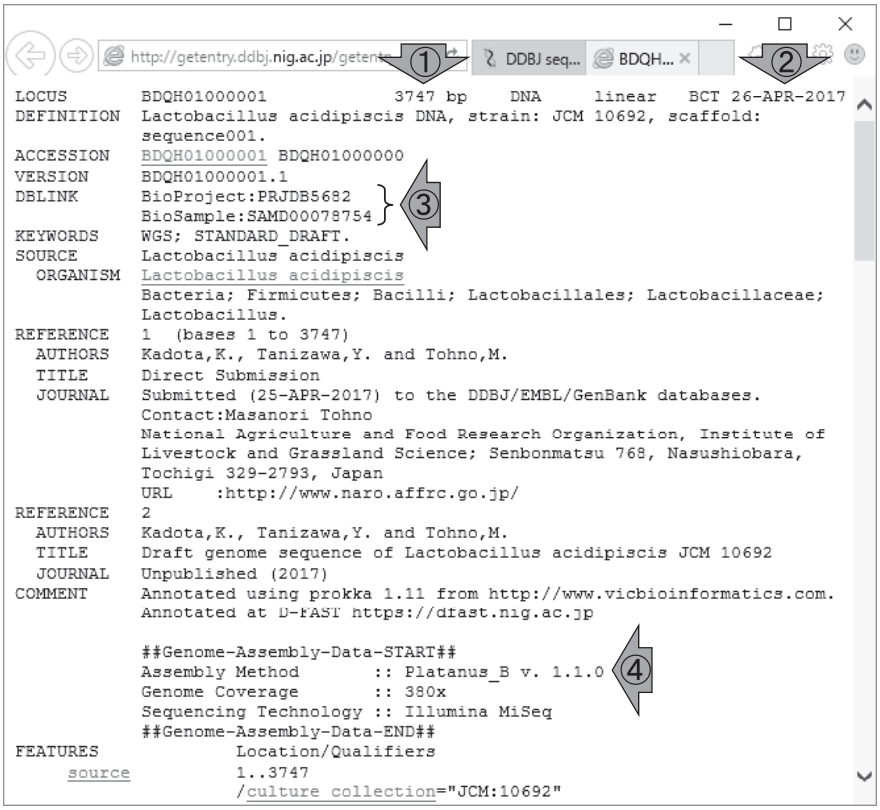

図 3. 公開されたアクセッション番号 BDQH01000001
①このコンティグの配列長は 3,747 bp。②公開日時は 2017 年 4 月 26 日。③ BioProject および BioSample ID もみられる。④ DDBJ とのやりとりで修正された内容も正しく反映されていることがわかる。

番号（BDQH01000001）で検索した結果である［W13-6］。2017年4月26日付の公開になっており、DDBJとのやりとりで修正された内容になっていることもわかる。

本稿での登録作業はこれで終わりとなるが、論文投稿にともなって配列の登録を行った場合には、論文が公開された後に論文タイトル・雑誌名・巻および頁番号などの文献情報の更新依頼を行う。これはDDBJホームページ上の申し込みフォームから行うことができる。

おわりに

本稿では、*L. acidipiscis* JCM 10692^{T} の *de novo* アセンブリおよびDFAST実行結果を用いてDDBJに登録する一連の手順を解説した。DDBJは登録依頼に対し5営業日以内に何らかの応答を返すことを目安としているが、登録の集中する時期には対応が追いつかないこともあり得る点に注意してほしい。今回の登録に要した日数は、1つの目安である。

第9回でも述べた通り、塩基配列のINSDCへの登録は学術雑誌への論文掲載に付随する義務である[9]。この考えは、GenBankの前身であるLos Alamos Sequence Databaseの設立メンバーのWalter Goadらの提案に基づくものである。彼らはまた、登録された塩基配列は公知のものであり、誰でもその情報を入手して利用可能にすべきだと主張した。この考えは、科学的知見は人類共通の財産であるというオープンサイエンスの基礎となる理念へと継承されてきた。正しい形式でのファイルの準備や細かな情報の記述は、登録者側からすれば手間である。しかし、正確で詳細な記述は研究結果の再現性の担保やデータ再利用の促進へとつながり、ひいては登録者自身の研究意義を高めることにもなるであろう。

謝　辞

本連載の一部は、情報・システム研究機構国立遺伝学研究所（遺伝研）との共同研究（2012-2088, 2013-2070）の成果によるものです。また、JSPS科研費JP25712032, JP15K06919の助成を受けたものです。

参　考　文　献

1) Tanasupawat S, Shida O, Okada S, Komagata K. (2000) *Lactobacillus acidipiscis* sp. nov. and *Weissella thailandensis* sp. nov., isolated from fermented fish in Thailand. Int J Syst Evol Microbiol **50**: 1479-1485.
2) Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, et al. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res **24**: 1384-1395.
3) Tanizawa Y, Fujisawa T, Kaminuma E, Nakamura Y., Arita M. (2016) DFAST and DAGA: Web-based integrated genome annotation tools and resources. Biosci Microbiota Food Health **35**: 173-184.
4) Mashima J, Kodama Y, Fujisawa T, Katayama T, Okuda Y, et al. (2017) DNA Data Bank of Japan. Nucleic Acids Res **45**: D25-D31.
5) Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration. (2012) The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res **40**: D54-56.
6) Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ. (2017) GenBank. Nucleic Acids Res **45**: D37-D42.
7) Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, et al. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol **29**: 415-420.
8) 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（2017）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．日本乳酸菌学会誌 **28**：3-11.
9) Strasser BJ. (2008) Genetics. GenBank--Natural history in the 21st Century? Science **322**: 537-538.

Methods for analyzing next-generation sequencing data X. Registration to DDBJ through Mass Submission System

Yasuhiro Tanizawa[1], Takatomo Fujisawa[1], Jun Mashima[2], Kyungbum Lee[2], Masanori Tohno[3], Mitsuo Sakamoto[4,5], Moriya Ohkuma[4], Yasukazu Nakamura[1,2], Kentaro Shimizu[6], and Koji Kadota[6]

[1]Center for Information Biology, National Institute of Genetics.
[2]DDBJ Center, National Institute of Genetics.
[3]Institute of Livestock and Grassland Science, National Agriculture and Food Research Organization.
[4]Microbe Division/Japan Collection of Microorganisms, RIKEN BioResource Center.
[5]PRIME, Japan Agency for Medical Research and Development (AMED).
[6]Graduate School of Agricultural and Life Sciences, The University of Tokyo.

Abstract

The International Nucleotide Sequence Database Collaboration (INSDC) has maintained a primary sequence database that collects experimentally-determined nucleotide sequence data directly from researchers. Now data deposition to the INSDC is mandatory for research publication at most of the scientific journals. However, the procedure to deposit data to the INSDC is a big burden, especially for those who are not familiar with computer skills. Recently, we have developed a genome annotation pipeline DFAST, which also assists data deposition to the INSDC. In this article, we show the instruction to deposit annotated genome sequence data using the DFAST web service. Supplementary materials are available at our web site, http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_book_JSLAB.