平成２１年度計画案成果「塩基配列アーカイブのデータベース構築と統合への貢献」

NCBI ならびに EBI と協調しながら、DDBJ としてリードアーカイブ DRA を立ち上げて、実データの受付から公開まで実績を重ねており、当初目的を十二分に達成している。具体的な成果を、当初計画の項目に準じて以下に報告する。

1. Short Read Archive 登録受け付けの仕組みの確立

【計画】

現在は、DDBJ での Short Read データの登録受け付けは試行的な段階にとどまっており、支援ツールやシステム類がまったく未整備で、査定作業はアノテータの手作業と目視で行われ、多くのデータ登録を頻繁に処理できる状態とは言い難い状況である。そこで、今後のサービス恒常化のために、登録データの査定作業の支援ツールを整備するなど、NCBI SRA と協調しながら登録受け付けや査定作業の仕組みの改善、より効率的な仕組みの確立を図る。

【成果】

- 2009 年 5 月に内部用データ管理システム DRA Manager（DRM）の運用を開始
- 2009 年 6 月にデータ検証ツール DRAvalidator の運用を開始
- 2009 年 9 月に Microsoft Excel ファイルの項目を埋めるだけで、簡単にメタデータを作成できる支援ツール DRA シートの運用を開始
- 2009 年 11 月から DRA データ受付システム D-way の本運用を開始。
  https://trace.ddbj.nig.ac.jp/D-way/
- 2010 年 1 月 12 日までの登録公開実績：［国内］受付 106 件、公開 17 件
  ［国外］受付 2 件、公開 1 件

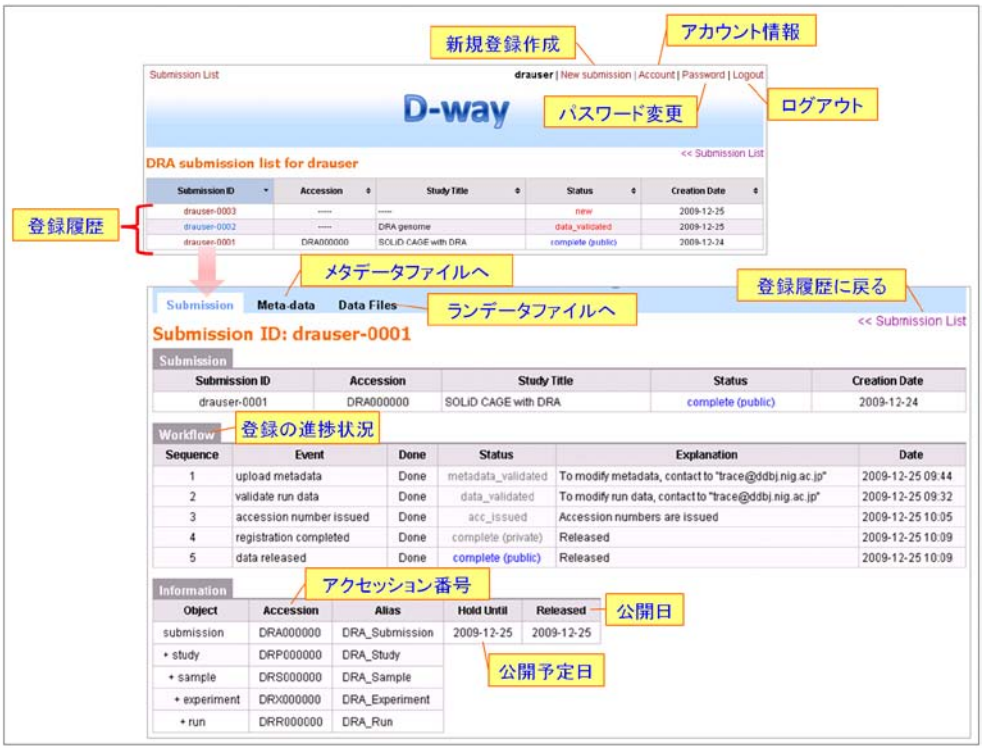

図 1-1 データ受付システム D-way 概要

図 1-1 に示すように、登録者はログイン後、新規登録作成、登録の進捗状況確認、登録履歴の参照やメタデータのアップロードなどができる。また、図 1-2 のように、DRA データ受付サーバに転送したランデータと、D-way からアップロードしたメタデータ間の整合性を D-way からチェックできる。

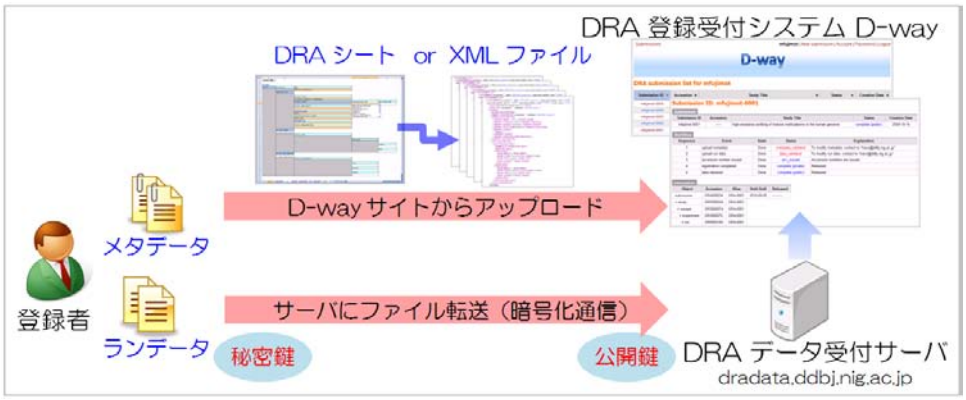

図 1-2 DRA へのメタデータとランデータのアップロード

- 2009 年 12 月に、図 1-3 のような複雑な構造を持つメタデータの作成支援のため、Flex 技術をベースにした、オンラインのメタデータ作成ツールを試作した。図 1-4 にその画面例を紹介する。

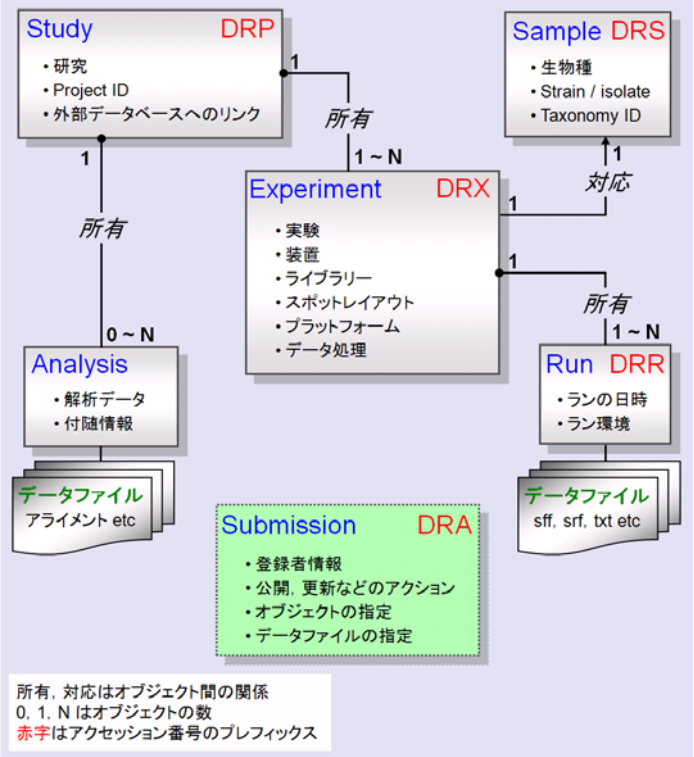

図 1-3 メタデータの構造

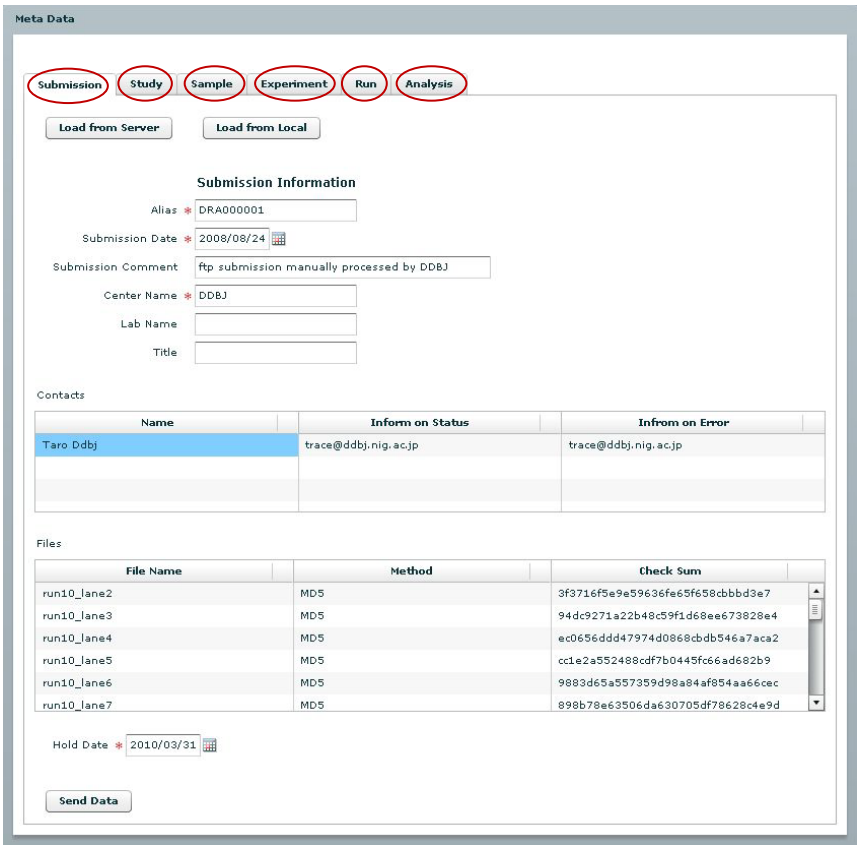

図 1-4 メタデータ作成支援ツールの画面

メタデータの各項目に対応するタブ（図の中で赤枠で囲んだ部分）を選択するとその項目に対応した書式が表示される。各項目の記入を済ませた後、GUI を介して項目を関連付けることでメタデータのセットを完成できる。

2. DDBJ 独自の ID 発行開始に向けた環境整備

【計画】

現在は、DDBJ で受け付けたデータについても、NCBI に登録処理を依頼しているため、NCBI の ID（SRA 番号）が割り振られている。一方、DDBJ に先行してサービス開始している EBI(European Bioinformatics Institute)の ERA(European Read Archive)は、自局で受け付けた登録データには ERA 独自の ID を割り振り、ERA から公開している。NCBI への過度の依存を避け、独立的に業務を行うためには、DDBJ もそれにならい、受け付けた登録データは DDBJ 独自に ID を割り振ることが望ましい。

よって、あくまでも NCBI / EBI との協調を前提にではあるが、DDBJ 独自の ID を発行、管理可能とすることを目標として、NCBI との交渉・調整や、必要なソフトウェアシステムの整備などを進める。

【成果】

- 2009 年 5 月に NCBI において NCBI の SRA ならびに EBI の ERA 担当チームと 2 日間にわたる打合せを行い 3 極協力体制確立の緒についた。
- 受付システムの整備が進んだこともあり、DDBJ は DDBJ Read Archive(DRA)として、SRA ならびに ERA と同等の立場で国際的に認可されたアクセション番号を発行していくことについて、NCBI ならびに EBI と合意した。

- 2009 年 10 月から NCBI Short Read Archive (SRA) とメタデータの交換を開始した。
- 2009 年 12 月に Nucleic Acids Research 電子版（文献１）にて、３極が共同して次世代シーケンサーからのデータをアーカイブしていくことを発表した。また、ニュースレターでも SRA、ERA ならびに DRA の３極共同事業が紹介された（文献２）。
- DRA と DDBJ 解析パイプライン、DRA と定量データのためのアーカイブ DDBJ Omics Archive (DOR)との連携について発表した（文献３）。

3. SRA データの利用者への提供開始

【計画】

現在は､DDBJ で受け付けたデータについても､NCBI への登録依頼を行っているため、データの本体は NCBI に送られ、従って NCBI から公開も行われる。しかし一方、NCBI と EBI の間の協調方針として、今後、データ量の増大を懸念し、データに関する情報と fastq 形式の塩基配列および品質情報に絞ってデータを交換し、サイズが巨大なシーケンサー出力の生データ自体は登録を処理した側だけに保持し公開する方向で議論が進んでいる。DDBJ が今後、アクセッション番号の独自発行など、SRA サービスをより自立的に開始するに当たり、データを NCBI に送付するという現在の方法は廃れる可能性があり、そうなった場合には、必然的に、DDBJ から独自にデータを公開することが必須となる。したがって、SRA データを公開する手段を整え、利用者に提供開始することを、3 つ目の目標とする。SRA データについては利用の形態もまだ未確立であることもあり、公開の形態についても検討し、優先順位をつけて開発を行う。

【成果】

- 2009 年 10 月に DRA 受付データの提供サイトを公開した：
  http://trace.ddbj.nig.ac.jp/registered/

  図 3 で赤枠で囲んだように、登録データの ID（Accession 番号）を検索または選択するか、あるいは、Study 名、生物種名、データを登録したセンター名または公開日で検索し、該当するデータをダウンロードすることができる。

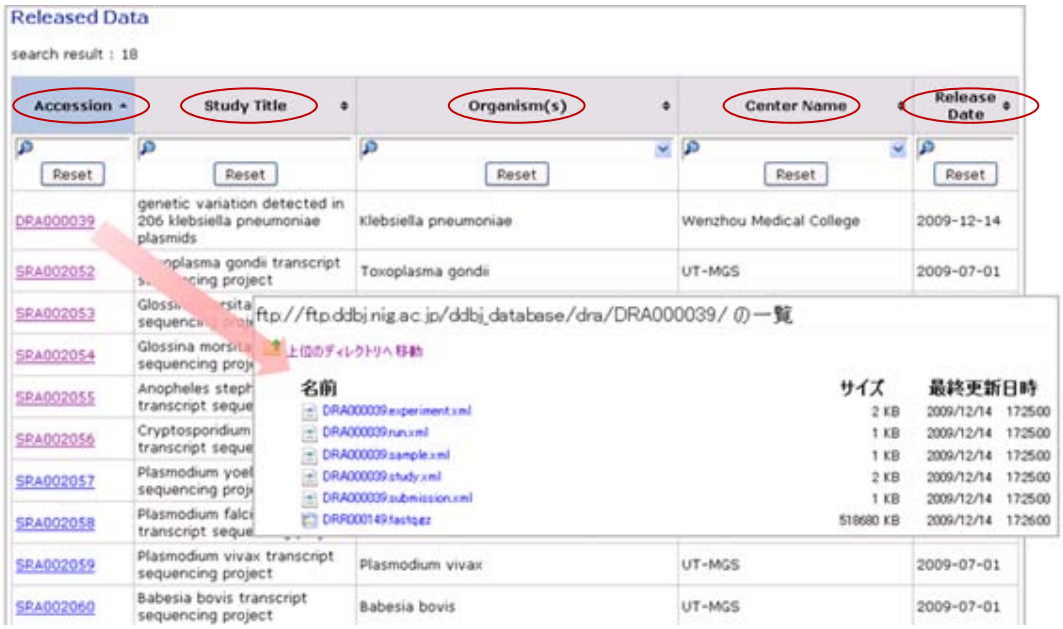

図３ DRA データ公開サイト

4.　DDBJ 事業における位置づけ

第２世代ならびに第３世代のシーケンサー由来のデータは、DRA に限らず、従来の塩基配列データベースや、遺伝子発現データベースにも波及していくことを認識し、図 4 に示す DDBJ におけるデータ登録提供事業の全体像の中に DRA を位置付け、研究開発を進めている。

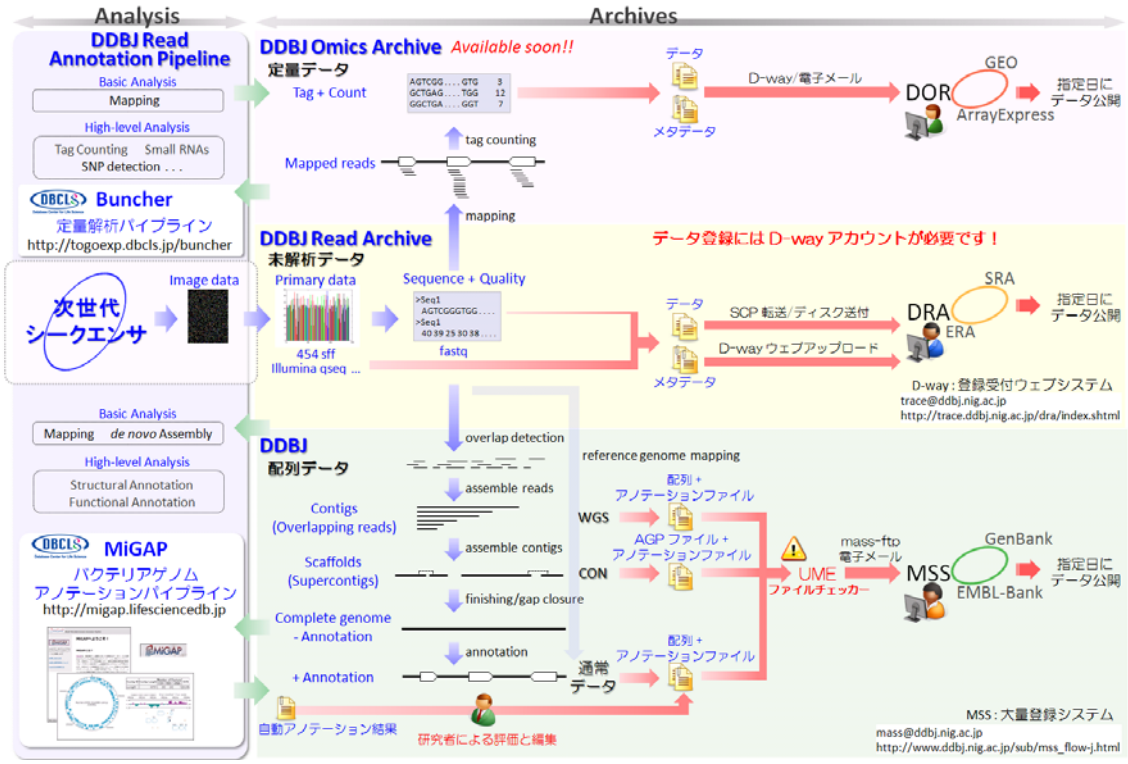

図４　DDBJ の全体像のなかでの DRA の位置付け

シーケンサからのリードが DRA に登録されるとともに、登録者の解析システムあるいは DDBJ や DBCLS が提供するリードアノテーションパイプライン、定量解析パイプラインまたはバクテリアゲノムアノテーションパイプラインで処理された結果が、DOR、MSS（DDBJ の大量塩基配列登録システム）を介して DDBJ に蓄積され DDBJ から公開される。

DDBJ Read Archive（DRA）：

次世代シーケンサーからの解析処理されていないデータを，メタデータとともに受付

DNA Data Bank of Japan（DDBJ）：

解析処理を経た配列データを受付

DDBJ Omics Archive（DOR）：

解析処理を経た定量データを受付

DDBJ Read Annotation Pipeline：

次世代シーケンサーからの fastq データを受付けて、解析結果を DDBJ，もしくは，DOR へ登録可能な形式で出力

参考文献

1. Shumway M, Cochrane G, Sugawara H (2010) Archiving next generation sequencing data. *Nucleic Acids Res* 38 (Database issue) : D870-871 Epub 2009.12.3
2. NCBI's Sequence Read Archive: A Core Enabling Infrastructure (2009.12.23) http://www.bio-itworld.com/BioIT_Article.aspx?id=94069
3. Kodama Y, Kaminuma E, Saruhashi S, Ikeo K, Sugawara H, Tateno Y, Nakamura Y (2010) Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technologies. *Advances in Computational Biology* Accepted

学会発表など

1. 統合データベースプロジェクトシンポジウム 2009（6 月、東京）
2. 第 23 回システムバイオロジー研究会（6 月、東京）
3. BIRD 研究成果発表会（11 月、品川）
4. 情報・システム研究機構シンポジウム（11 月、品川）
5. 第 32 回日本分子生物学会年会（12 月、横浜）
6. GIW2009（12 月、横浜）
7. 第 1 回シーケンサー利用技術講習会（12 月、理研横浜）
8. ロシュ社、イルミナ社やアプライドバイオシステムズ社の担当者と次世代シーケンサー由来データのファイル形式などについて情報交換。
9. シーケンシング受託解析業者に DRA を紹介し、データ登録における協力について意見交換。また、いくつかの大口データ登録者と円滑なデータ登録方法について打ち合わせ。