「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」に関する成果報告（H21 年度）

科学技術振興機構バイオインフォマティクス推進センター事業研究開発課題
「バイオ情報資源の高準化と共用化」平成２１年度研究開発実施報告書から抜粋

3.4 DDBJ Read Archive (DRA)

研究計画書では新世代シーケンサ由来のデータへの取り組みを、国立遺伝学研究所の分担課題の中の「総括、メソッド・オントロジーならびに Web サービスとワークフローに関する研究」に「新しい型のデータへの取り組み」として位置付けたが、平成２１年度の研究開発内容として大きな比重を占め、また、今後コミュニティーとの連携がますます重要になっていくと思われることから、独立の項を立てて報告する。

【研究開発目的】

DDBJ にとっても喫緊の課題になってきたトレース・アーカイブ（Trace archive、以下 TA）とショート・リード・アーカイブ（Short read archive、以下 SRA）の高準化と共用化にも取り組む。

【方法】

本課題において平成２０年度までは、主として塩基配列、タンパク構造、パスウエーなどを対象とする情報資源に取り組んできたが、新たに、トレース（trace archive (TA)）とショートリード（short read(SR)）を対象とする情報資源の高準化と共用化にも取り組みを始める。このために、NCBI ならびに EBI とデータ標準を調整しつつ、データの受付・査定・DRA(*)番号の発行・公開のためのデータシステムを構築する。

(*)注（2011 年１月現在）：研究計画書提出時には、新世代シーケンサ由来データのアーカイブを Short Read Archive と呼んでいたが、新世代シーケンサの進歩も踏まえて、Sequence Read Archive と呼ぶことになった。また、３極の Sequence Read Archive をそれぞれ、SRA（Sequence Read Archive)、ERA(European Sequence Read Archive)、DRA（DDBJ Sequence Read Archive)と呼んでいる。

【成果】

(1) DRA 登録受け付けの仕組みの確立

従来、DDBJ での Short Read 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った。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った。

- 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/
- 登録公開実績（2010/02/22 時点）：受付 187 件、公開 18 件

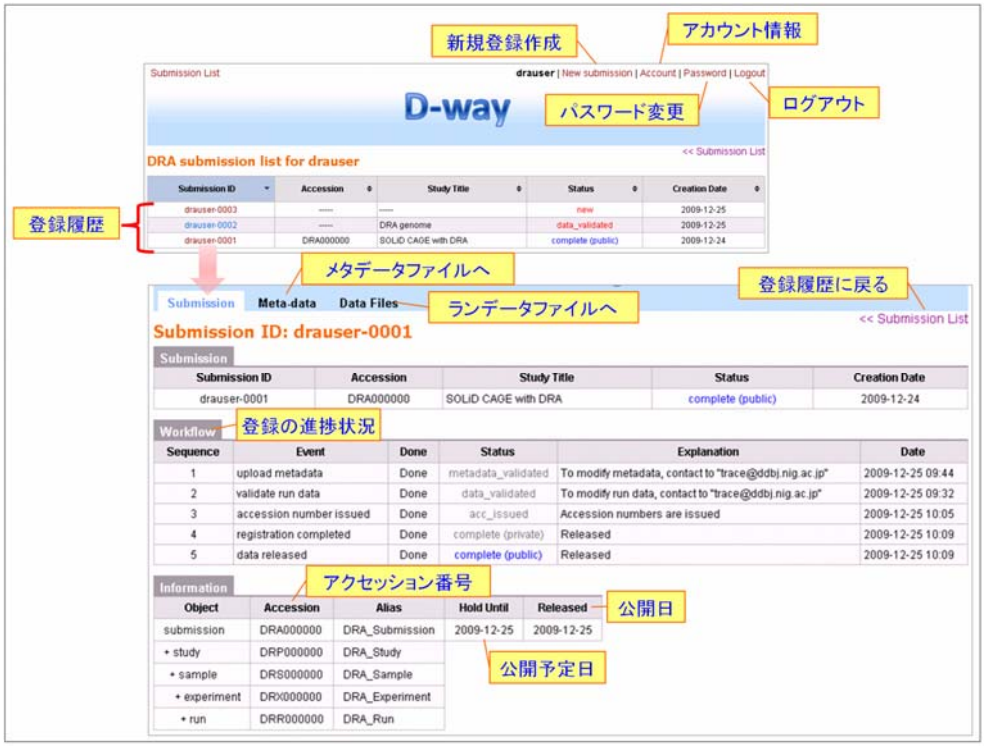


図 7-1 データ受付システム D-way 概要

図 7-1 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 7-2 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。

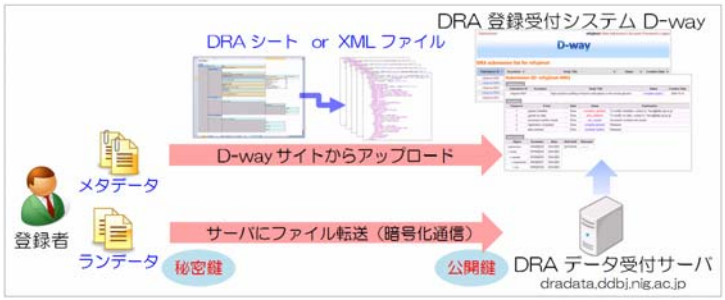
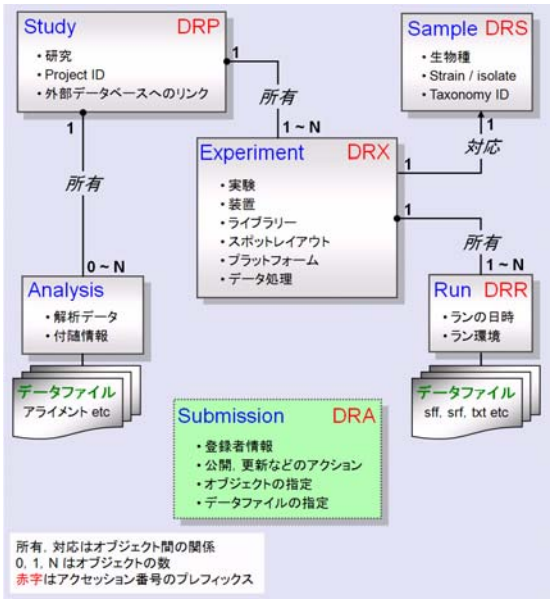
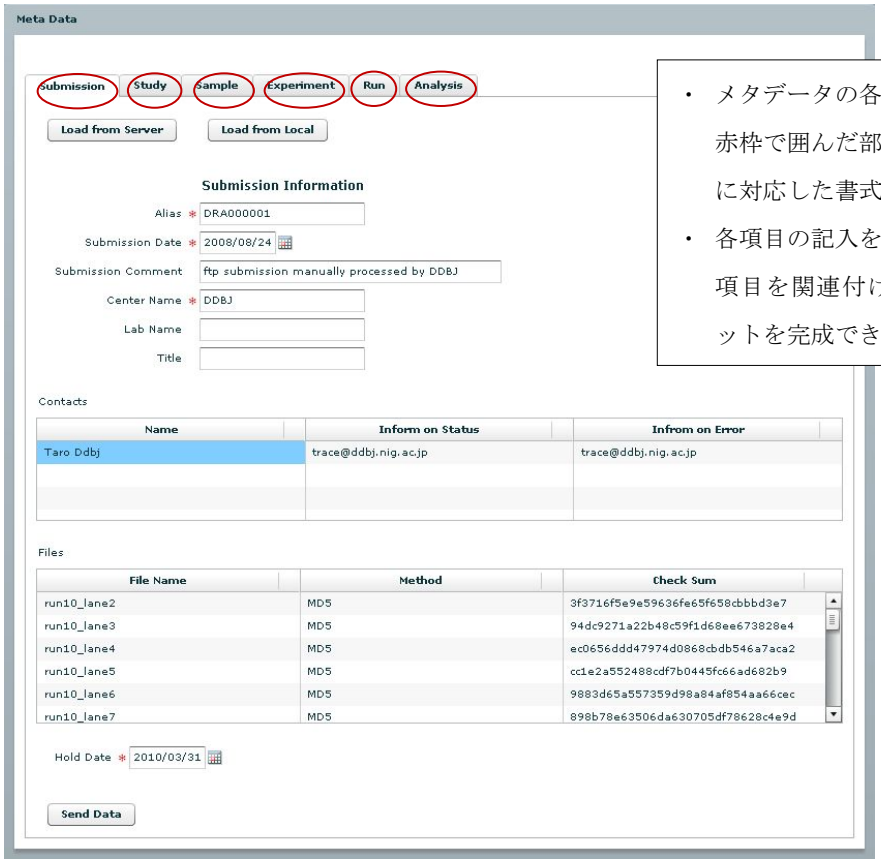


図 7-2 DRA へのメタデータとランデータのアップロード

- 2009 年 12 月に、図 7-3 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 7-4 にその画面例を紹介する。

図 7-3 メタデータの構造





- メタデータの各項目に対応するタブ（図の赤枠で囲んだ部分）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。

図 7-4 メタデータ作成支援ツールの画面

(2) DDBJ 独自の ID 発行開始に向けた環境整備

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たため、NCBI の ID（SRA 番号）が割り振られていた。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、DDBJ もそれにならい、受け付けた登録データは DDBJ 独自に ID を割り振ることを目指した。

- 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）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 電子版にて、3 極が共同して次世代シーケンサからのデータをアーカイブしていくことを発表した(Shumway, Cochrane and Sugawara (2010))。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(NCBI’s Sequence Read Archive: A Core Enabling Infrastructure (2009.12.23)http://www.bio-itworld.com/BioIT_Article.aspx?id=94069)。

- 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(Kodama et al (2010))。

(3) SRA データの利用者への提供開始

前項にもあるように、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 ならびに EBI との国際協調のもとで非常に大規模になることが明らかなアーカイブ全体を効率良く構築・運用するために、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た。このため、DDBJ がデータを NCBI に送付するという従来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。そこで、read 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目指した。

- 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 8 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。

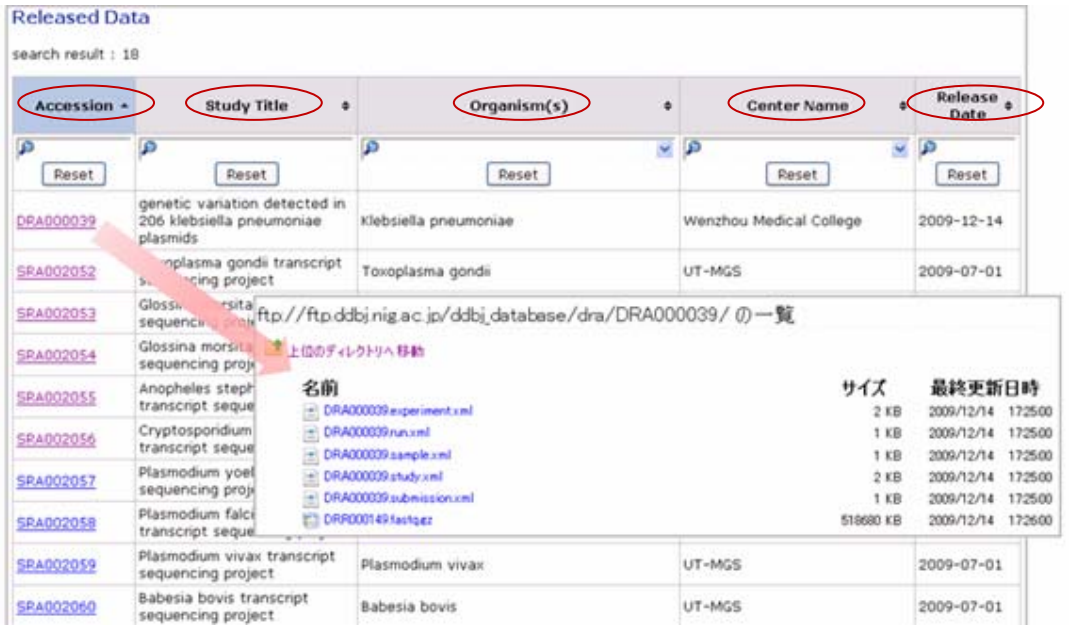


図 8 DRA データ公開サイト

(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 9 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。すなわち、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 やライフサイエンス統合データベースセンター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 MiGAP 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。

［用語解説］

DDBJ Read Archive（DRA）：次世代シーケンサ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

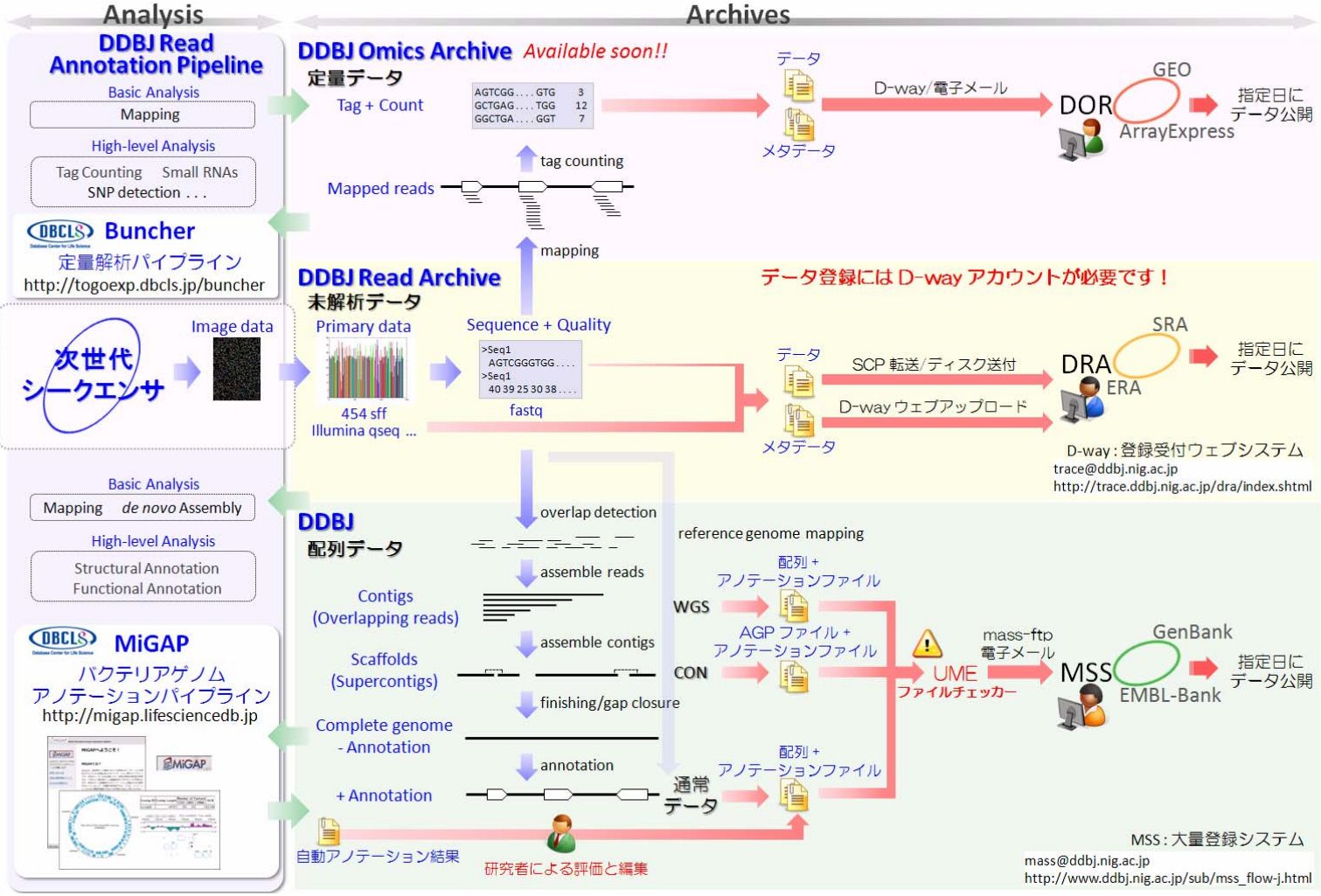


図 9 DDBJ の全体像のなかでの DRA の位置付け

(5) まとめ：DRA のデータ処理実績とシステム開発の経緯(図 10)

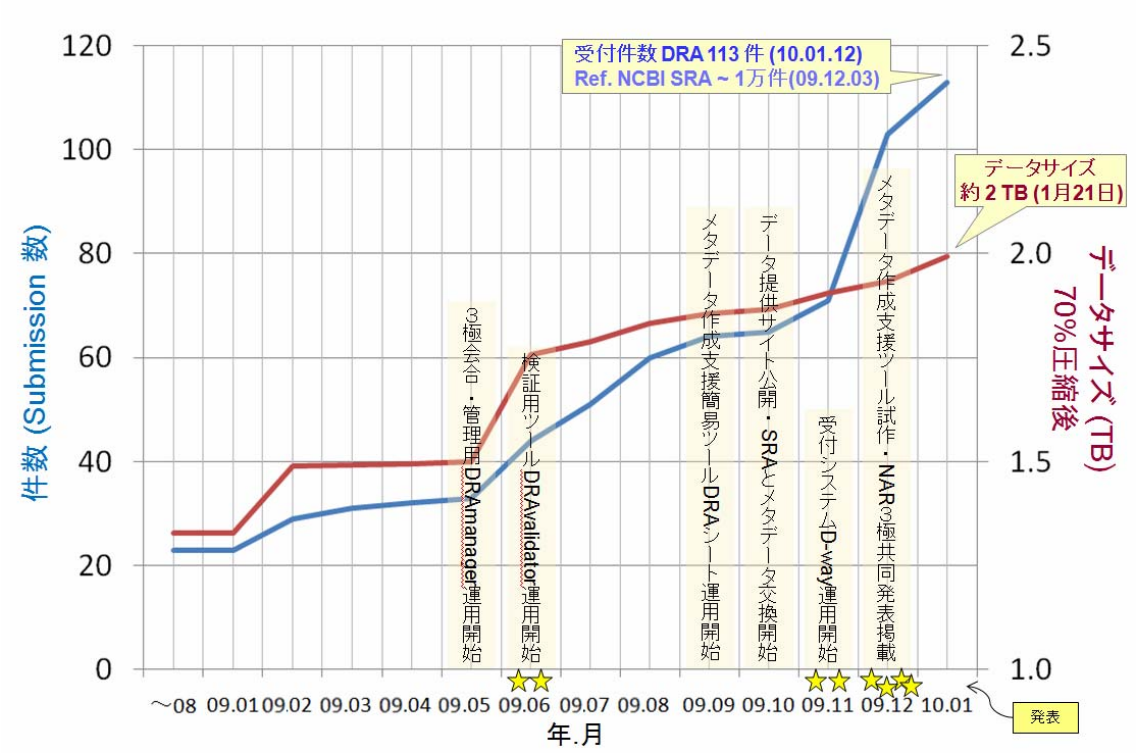


図 10　DRA 受付件数（Submission 数）とデータサイズならびにシステム構築・運用の進捗記録

4　研究開発実施体制(DRA 関係)

研究開発参加者名（所属、役職、研究開発項目）

菅原秀明	国立遺伝学研究所生命情報・DDBJ 研究センター・特任教授
大久保公策	国立遺伝学研究所生命情報・DDBJ 研究センター・センター長
五條堀　孝	国立遺伝学研究所生命情報・DDBJ 研究センター・副所長
猿橋　智	国立遺伝学研究所生命情報・DDBJ 研究センター・特任研究員
舘野義男	国立遺伝学研究所生命情報・DDBJ 研究センター・特任教授
中村保一	国立遺伝学研究所生命情報・DDBJ 研究センター・教授

5　研究開発成果の発表（DRA 関係）

【論文発表】

（国際誌２件（内査読あり２件））

Shumway M, Cochrane G, Sugawara H (2010)
Archiving next generation sequencing data
Nucleic Acids Research, Vol.38, Database issue, pp.D870-871, Dec. 2009.（査読あり）

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】

（海外１件（内査読あり１件)、国内１件（内査読あり１件））

Eli Kaminuma, Yuichi Kodama, Satoshi Saruhashi, Takeshi Konno, Takako Mochizuki, Hidemasa Bono, Hideaki Sugawara, Kousaku Okubo, Toshihisa Takagi, Yasukazu Nakamura, DDBJ Read Archive and DDBJ Read Annotation Pipeline:An archive database and an analytical tool for next-generation sequence data.
The 20th International Conference on Genome Informatics (GIW2009), P113, Yokohama, Japan, Poster Oral 3, Dec. 2009.（査読あり）

児玉悠一、猿橋智、五條堀孝、舘野義男、神沼英里、中村保一、菅原秀明
「DDBJ Read Archive」： 次世代シーケンサからの出力データのためのアーカイブ
第32回日本分子生物学会年会，2P-0052，横浜，2009年12月.（査読あり）

【研究会など】

児玉悠一、猿橋智、五條堀孝、舘野義男、中村保一、菅原秀明
次世代シーケンサの出力データのためのアーカイブについて
統合データベースプロジェクトシンポジウム2009，東京，2009年6月.

児玉悠一
DDBJ Read Archive
第23回システムバイオロジー研究会，東京，2009年6月.

猿橋智、児玉悠一、神沼英里、五條堀孝、舘野義男、中村保一、菅原秀明

DDBJ Read Archive のご紹介（新型シーケンサへの対応）
バイオインフォマティクス推進センター事業(BIRD)第 5 回研究開発成果報告会, 東京, 2009 年 11 月.

神沼英里、中村保一
DNA Data Bank of Japan (DDBJ) の新展開
情報・システム研究機構シンポジム「情報とシステム 2009」、2009 年 11 月.

児玉悠一
次世代シーケンサーデータの DRA(DDBJ Read Archive)への登録
理化学研究所横浜研究所第 1 回シーケンサ利用技術講習会, 横浜, 2009 年 12 月.

【プログラム・データベース等】

(1) DRA 一式
概要：DRA のデータ受付・査定・公開のためのシステム
URL：http://trace.ddbj.nig.ac.jp/dra/index.shtml
公開予定日：2009 年 7 月 30 日

(2) DRA 受付公開実績（2010 年 2 月 22 日現在）

機関	受付件数	公開件数
（国内）	185	17
東京大学	159	13
理化学研究所	9	4
国立感染症研究所	5	0
農業生物資源研究所	4	0
北里大学	2	0
慶応義塾大学	2	0
遺伝学研究所	1	0
沖縄科学技術振興センター	1	0
海洋研究開発機構	1	0
京都大学	1	0
（海外）	2	1
Wenzhou Medical College (中国)	1	1
BIOTEC (タイ)	1	0
合計	187	18

6 ワークショップ等

【DRA打合せ】

- 新世代シーケンサを使いこなしている主要な研究機関と個別に、円滑なデータ登録方法について打合せを行った。
- 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打合せを行い、新世代シーケンサ由来データのファイル形式やソフトウエアの情報を収集した。
- シーケンシング受託解析業者と打合せを行い、顧客に対して DRA への登録について協力を得られることになった。

以上