資料８－１

委託業務の題目

DDBJ Trace Archive / DDBJ Sequence Read Archive データベース構築事業

受託機関名

国立遺伝学研究所

１．委託事業の９月末時点の判断基準になる目標

・ 登録データから SRA Toolkit により SRA ファイルを作成し、データリストとともに DRA から公開する（9 月末時点）

２．９月末時点の達成状況

・ SRA Toolkit により登録データを SRA ファイルに変換するシステムを整備した。公開日になると、SRA ファイルから汎用されている fastq ファイルを作成し、fastq とメタデータを DRA から公開している。同時に SRA ファイルとメタデータはミラーリングのため NCBI Sequence Read Archive に送付している。これにより、データの受付、作成、管理の一連の流れで NCBI に依存していた部分がなくなり、全ての処理を DRA 自前でできるようになった。なお、NCBI に確認したところ「データリストは不要」とのことであったので、データリストは作成していない。
・ メタデータ作成支援ツール MetaDefine に登録データのもととなるテンプレートを提供する機能を実装した（現在、Whole Genome Shotgun、Transcriptome と Barcode)。
・ MetaDefine で、同一のランで由来サンプルをタグ配列で区別している「Barcode データ」のメタデータを作成可能にした。
・ DDBJ Sequence Read Archive の登録受付件数が 336 件に達した。
・ DDBJ Sequence Read Archive と DDBJ Omics Archive についての論文を公表した。
・ Kodama Y *et al* (2010) Biological Databases at DNA Data Bank of Japan in the Era of Next-Generation Sequencing Technologies. *Adv Exp Med Biol* 680:125-135
・ Sequence Read Archive への DDBJ/EBI/NCBI の各極の取り組みを Nucleic Acids Research 誌に投稿した。
・ NEDO「完全長 cDNA 構造解析プロジェクト」由来の波形データ計 2,445,650 件を DDBJ Trace Archive へ登録した。本データは DDBJ Trace Archive と NCBI Trace Archive から公開されている。

３．上記達成状況を踏まえたプロジェクト終了までの目標

・ メタデータの検索系、及び、それと連動した fastq ファイルのダウンロード提供システムを完成させる（3 月末時点）

４．成果の概要

DRA に登録された各種シークエンサ由来のランデータから SRA Toolkit により SRA ファイルを作成するシステムを整備した。

【従来】※青字が NCBI SRA に依存していた部分

登録データ　->　NCBI SRA にアップロード　->　SRA ファイル作成

->　SRA ファイルと fastq ファイル公開 ->　SRA から fastq 取得、DRA で公開

【9 月から】

登録データ　->　SRA ファイル作成　->　DRA で fastq ファイルを作成、公開

>　SRA ファイルを NCBI にアップロード　->　NCBI から公開

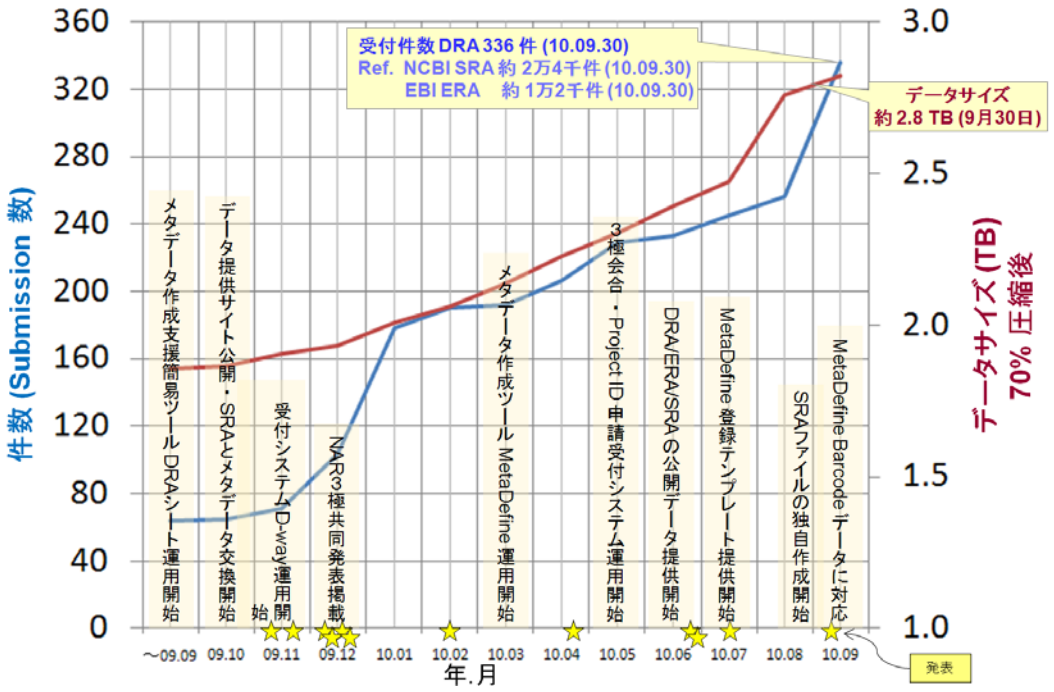

表 1　累積登録件数、圧縮後データサイズと DRA の主な活動

DRA/DTA ウェブサイト

	2010年1月	2月	3月	4月	5月	6月	7月	8月	9月
アクセス件数	1684	1453	1836	2703	2642	2475	2196	2972	2753
バイト数 (MB)	146	133	193	284	345	335	315	364	380

DRA/DTA FTP サイト

	2010年1月	2月	3月	4月	5月	6月	7月	8月	9月
アクセス件数	468	682	333	594	440	1530	65399	10998	564
バイト数 (GB)	31.77	1.47	27.19	66.18	35.88	314.72	5155.48	3292.2	29.31

表 2　アクセス件数と転送データサイズ

7 月と 8 月に機械的なデータダウンロードアクセスがあり、件数が急増した