資料　4－1

DDBJ Read Archive（DRA）とは

DDBJ
DNA Data Bank of Japan

大量高速シークエンサー由来の配列データベース

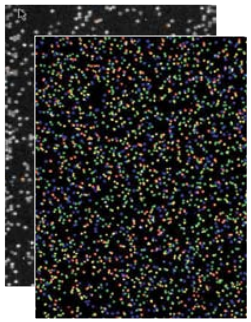
454 (Roche)

Solexa (Illumina)

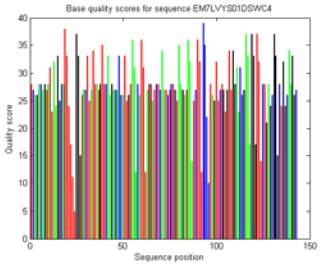
SOLiD (ABI)

454、Illumina、SOLiD などからのデータを登録

登録対象



画像データ



計測データ

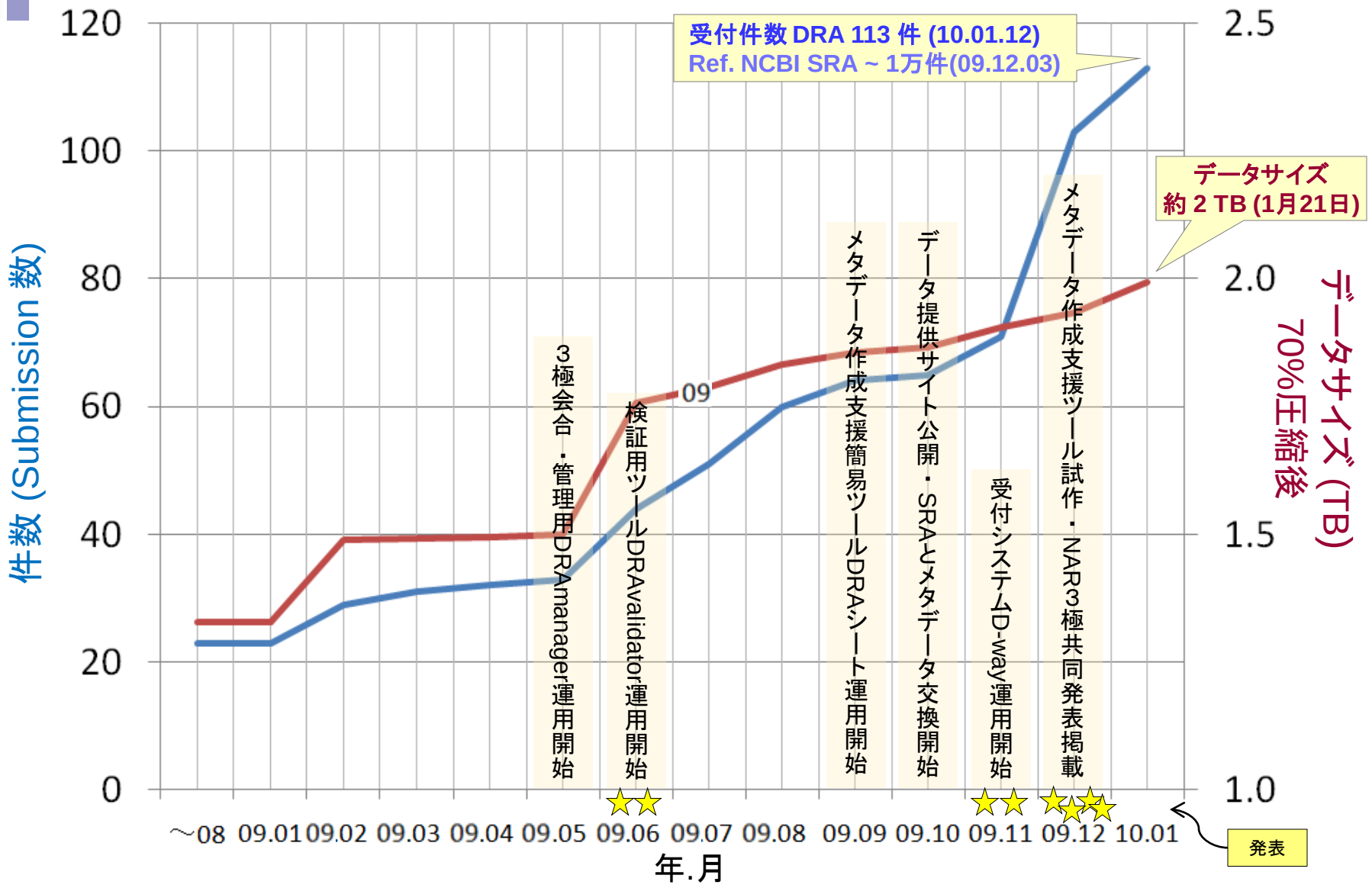
シグナル強度
ノイズ
ベースコール
Quality value 等

```
@SRR001654.1 9460:7:1:830:763 length=36
GTCAATATTAATCATACCAATATACTCAAAAAATAA
+SRR001654.1 9460:7:1:830:763 length=36
I+-&*4)%+5'#%/)&$%$#%"#&%'%"$%#%%!""
@SRR001654.2 9460:7:1:402:781 length=36
GGTCTAAAAAGCAAAATTCAGTCTTCAAAATAATTC
+SRR001654.2 9460:7:1:402:781 length=36
II+(%$+%'&+*-0+/*("%&+"*&"(*$""#%%&$
@SRR001654.3 9460:7:1:433:775 length=36
GTGCTTTTTTTTTTCCAGGAAGTTGTCTCCTCTATC
+SRR001654.3 9460:7:1:433:775 length=36
II3DI>IIIIIIIB7.,&%&'&)."+%,$"&$&"%#
```

fastq データ

塩基配列
Quality Value

DRA開発と受付の年間動向



登録機関数と公開件数

DDBJ
DNA Data Bank of Japan

登録機関数　13機関
　　国内　11機関
　　国外　　2機関(中国、タイ)

公開件数　　18件
　　国内　17件
　　国外　　1件（中国）

（注）ここまでの件数は、メタデータの中でいうところのSubmissionが単位。
メタデータの中でいうところのrunを単位とすると、登録件数が510件、
公開件数が91件となる。

2010年度成果の目標及び業務の方法

DDBJ
DNA Data Bank of Japan

2009年度はDDBJ Read Archive(DRA)の運用を開始した。すなわち、第2世代シーケンサ由来のデータを受付し、受付けたデータにIDを発行し、さらに、利用者へ提供する枠組みを実現した。また、NCBIのSRAならびにEBIのERAとの連携も具体化した。

2010年度は、このDRAの受付システム、公開システムならびに管理システムの高度化を進める。また、SRAならびにERAと協調して、研究社会へのデータ登録への理解を広げるべく努め、さらに、第3世代シーケンサからのデータやそれを利用した新たな研究分野の成果への対応についても取り組む。

これによって、技術進歩の歩みが止まらないシーケンサから生成され続ける塩基配列リードに対応し、また、高速に入手できる大量のデータを活用して進化し続ける研究動向に対応可能なDRA基盤を開発しつつ安定に運用する。

① 受付システムの高度化と運用

DRAの受付システムは、受付け窓口に相当するD-way、データの正当性を検証するDRAvalidator、メタデータ作成を補助するGUIならびにショートリードを解析するパイプラインからの出力を受付けるインターフェースといったサブシステムで構成される。2009年度に構築したこれらのサブシステムにおいて、検証機能を向上させ、登録者に対する利用者インターフェースを直感的に分かりやすいものにし、さらに、受付査定業務の効率化を図る。また、NCBIで開発が進められているSRA toolkitとの整合性も確保し、同時に、fastqデータより大規模になるが豊富な情報を含むSRAファイルにも対応する。

② 公開システムの高度化と運用

2009年度のデータ提供は限られた項目で検索しヒットしたデータセットをダウンロード可能にしたものであったが、2010年度には、メタデータに含まれる多様な項目による検索を可能として、その検索結果にヒットしたデータセットをダウンロード可能とする。特に、SRAファイルの提供に取り組む。

③ 管理システムの高度化と運用

2009年度はログファイルなどから手作業で集計していた各種統計情報の自動生成、データのバージョン管理、DRA側でのデータ検証、データ更新の自動反映、公開日管理の機能をシステム化して、得られる統計情報をシステム改善に活かしていく。

④ 国際連携

国際塩基配列データベースの国際実務者会議の機会などを生かして、SRAならびにERAと意識と情報を共有することによって、日米欧3極で整合性のあるリードアーカイブ構築を目指す。また、塩基配列データ登録の事例に倣って、シーケンサからの出力データをいずれかのアーカイブに登録するルール採用を、3極共同で学術雑誌などに働きかける。

① 受付システムの高度化と運用

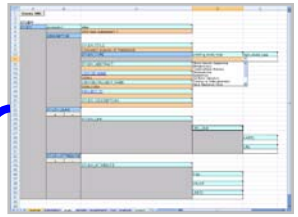
DDBJ
DNA Data Bank of Japan

開発済

今年度からの開発継続

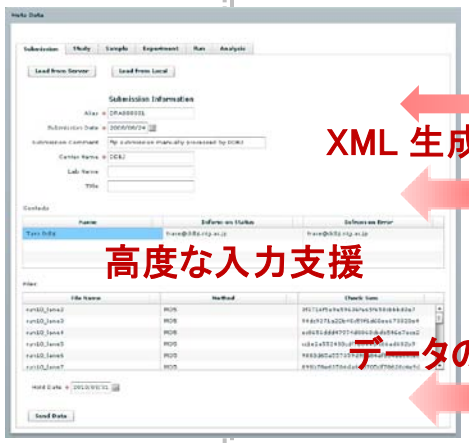
2011年度開発

“DRA シート”



XML ファイル

“ウェブ GUI ツール”



高度な入力支援

結果

XML 生成/読み込み

チェック

“DRAvalidator”

“登録受付システム D-way”

データの再利用

メタデータ

D-way サイトからアップロード

登録者

ランデータ

サーバにファイル転送

受付サーバ

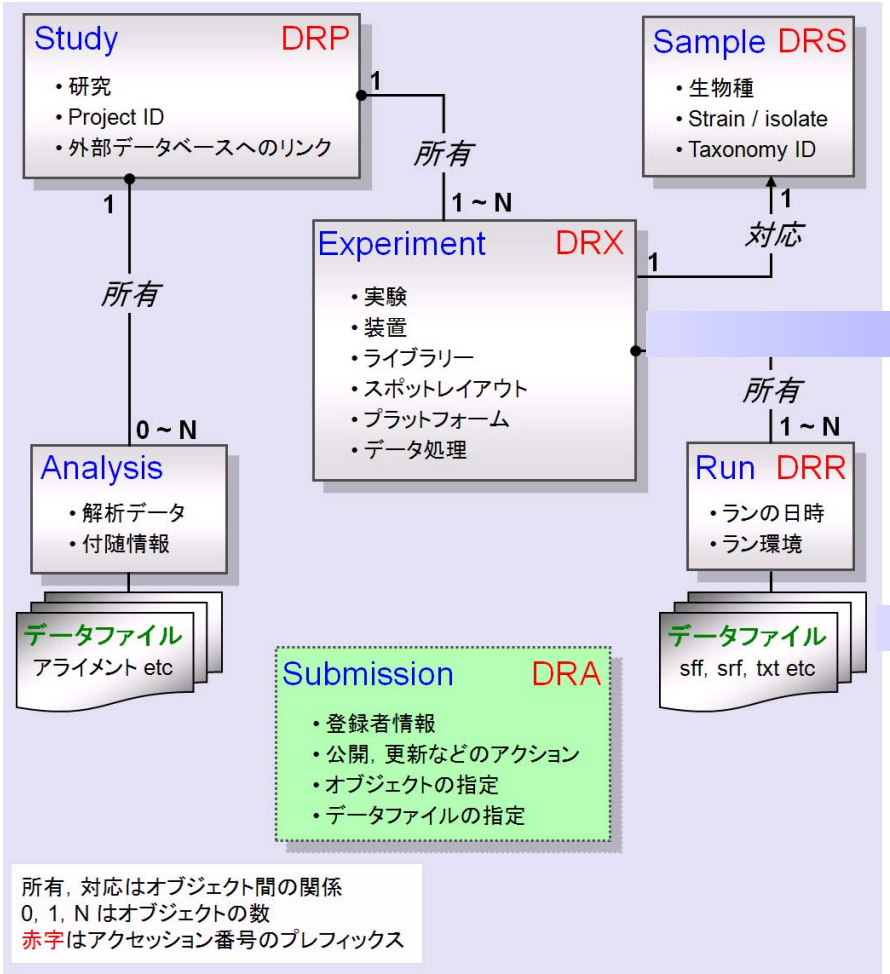
“SRA Toolkit”
ファイル形式変換

解析パイプライン

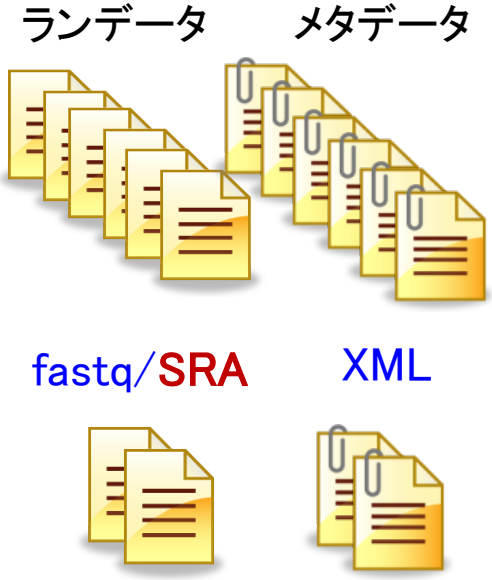
② 公開システムの高度化と運用

DDBJ
DNA Data Bank of Japan

メタデータの検索



ダウンロード提供
(FTP/Aspera)



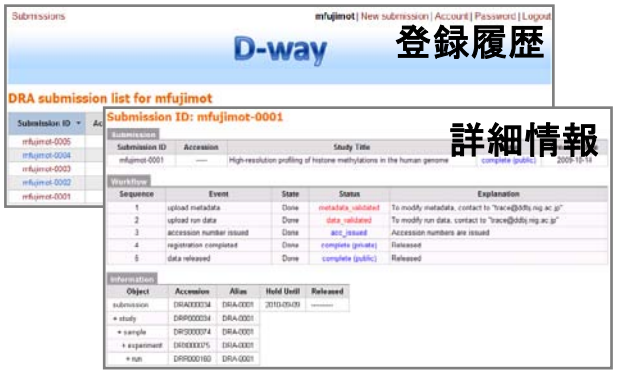
Native data → SRA in binary → fastq

③ 管理システムの高度化と運用

DDBJ
DNA Data Bank of Japan

“登録受付システム D-way”

“管理システム DRM”



統計情報

登録件数
研究の種類
データの種類 etc

アノテータ

“DRAvalidator”

バージョン管理

.3
.2
.1

反映

履歴記録

ログイン/データ更新

反映

DB

反映/公開

登録者

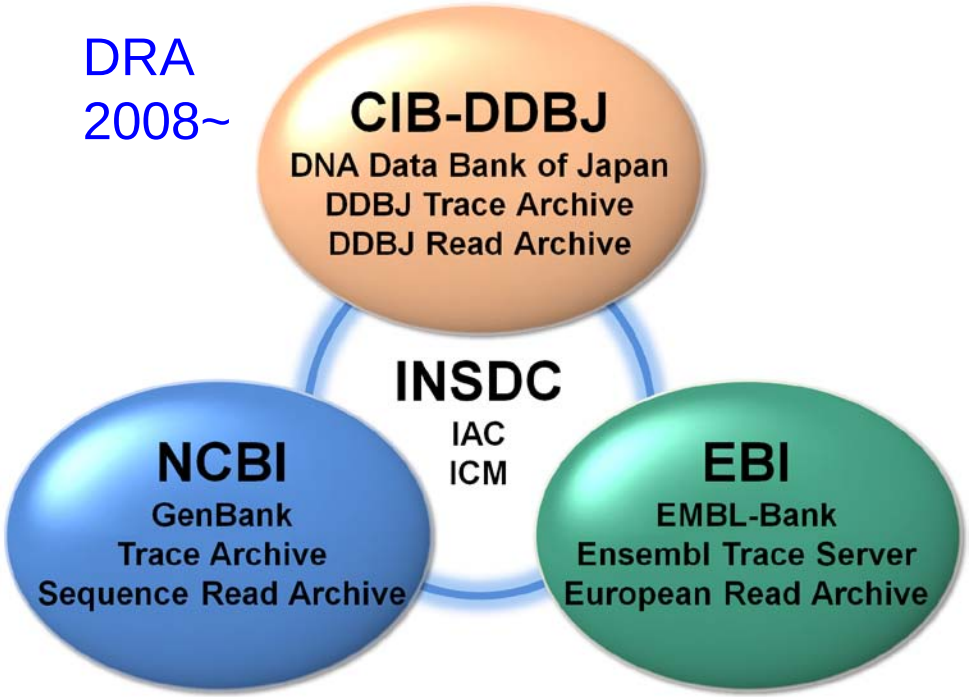
利用者

④ 国際連携

DRA
2008~

SRA
2007~

ERA
2008~



第1回 DRA/ERA/SRA 国際会議 2009.5.14-15 at NCBI

DRA/ERA/SRA 3極共同で国際アーカイブを構築することで合意

Shumway M, Cochrane G, Sugawara H (2010)
"Archiving next generation sequencing data." *Nucleic Acids Res* 38 (Database issue) :D870-871

第2回 2010.5.20-21 at EBI 予定

2010年度のデータ量は？

DDBJ
DNA Data Bank of Japan

国内におけるシークエンサーの導入と稼働の状況を把握しつつ、ハードウエアとソフトウエアのスケーラビリティーを継続的に検証

おおよその国内設置台数（一部見込みを含む）

Illumina	～ 60台
Illumina新型機	～ 3台
SOLiD	～ 20台
454	～ 20台
Heliscope	～ 4台

DDBJにおけるDRAの位置付け

DDBJ
DNA Data Bank of Japan