次世代シークエンサの出力データのためのアーカイブについて

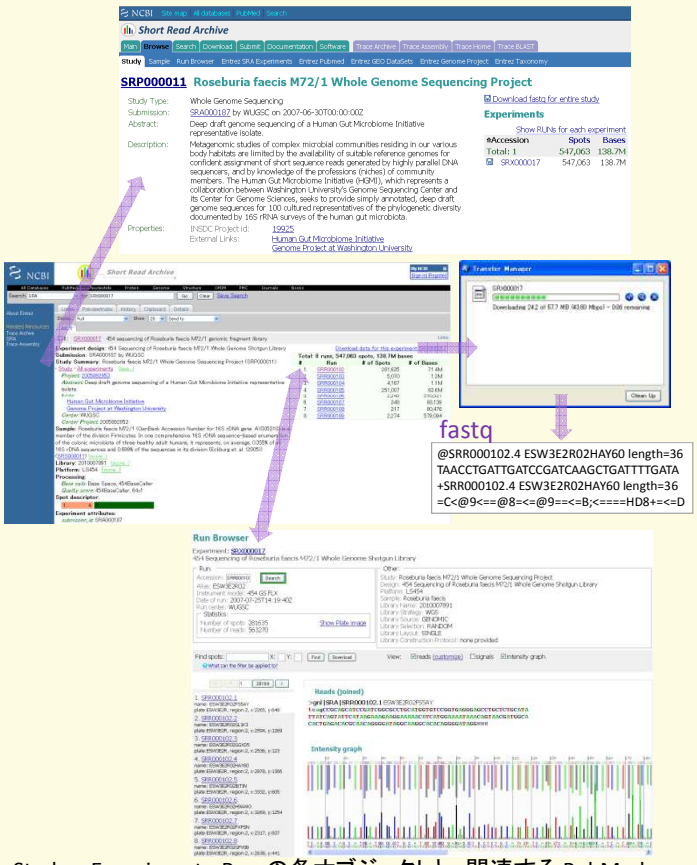
児玉悠一、猿橋智、五條堀孝、舘野義男、中村保一、菅原秀明 / 生命情報・DDBJ研究センター

次世代シークエンサは、様々な生物学分野に浸透し、従来を大幅に上回るペースでデータを産み出している この膨大なデータをアーカイブすべく、NCBI は2007年に Short Read Archive (SRA) を、EBI は2008年に European Nucleotide Archive – Reads (ERA) を稼働させた。 5月の時点で、SRA には総計6兆塩基対 (100 TB) 以上のデータが蓄積されている。 DDBJ も年内に DDBJ Read Archive (DRA) を立ち上げるべく2008年秋から活動を開始し、現在までに11件のデータ登録に取組んできた

DDBJ/EBI/NCBI は20年以上に渡り国際塩基配列データベース (INSDC) を共同構築してきた実績があり、次世代シークエンサからの出力データのためのアーカイブ DRA/ERA/SRA 構築も、今後、国際共同事業として推進していく

[DRA はライフサイエンス統合データベースプロジェクトならびに JST-BIRD 事業の支援を受けています]

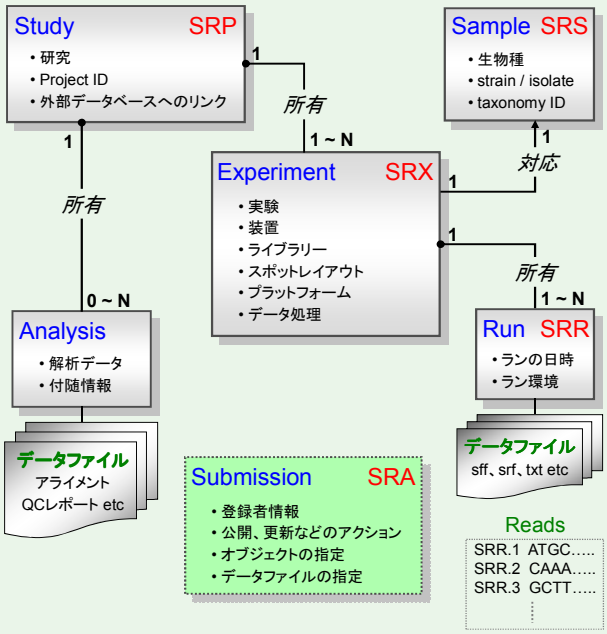
NCBI SRA 公開システム



Study – Experiment - Run の各オブジェクトと、関連する PubMed、GEO、Taxonomy データへの Entrez リンクをたどることができる
個々のリードの配列や Quality Value の閲覧、ダウンロードができる

データモデル

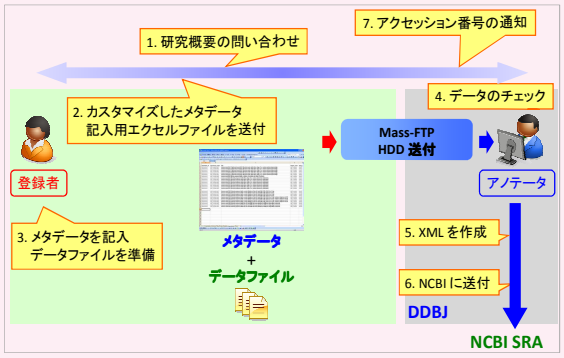
メタデータ + データ



所有、対応はオブジェクト間の関係
0、1、N はオブジェクトの数
赤字はアクセッション番号のプレフィックス

SRA では XML 形式の "メタデータ" と "データ" を分離、メタデータは6つのオブジェクトで構成：Submission、Study、Experiment、Sample、Run、Analysis 各オブジェクトに赤字で示したプレフィックスのアクセッション番号を発行し、個々の配列 (Reads) に "Run 番号 + 連番" を自動的に割り振る

DDBJ へのデータ登録



http://www.ddbj.nig.ac.jp/sub/trace_sra-j.html

登録受付実績

項番	機関名	生物種	機種	件数	容量 (圧縮前)	SRA発行	作業期間	状況
1	東京大学 新領域 菅野・鈴木研	微生物6種と媒介昆虫2種	Solexa	1億件	327 GB	2008/9/14	約2ヶ月	完了
2	理研 林崎研	ヒト	Solexa	20万件	38 MB (fastq)	2008/11/12	約2ヶ月半	完了
3	非公開	非公開	Solexa	200万件	23 GB	2008/12/12	約2ヶ月	完了
4	非公開	非公開	Solexa	4000万件	454 GB	2009/2/27	約2ヶ月	完了
5	非公開	非公開	454	不明	752 MB	2009/1/30	約1ヶ月半	完了
6	非公開	非公開	Solexa	不明	604 GB	2009/5/4	約1ヶ月半	完了
7	非公開	非公開	Solexa	不明	2.3 GB (fastq)	2009/5/7	約1ヶ月半	NCBI アップロード済
8	非公開	非公開	Solexa	不明	167 GB	未完了	未完了	対応中
9	非公開	非公開	Solexa	不明	3.8 GB (fastq)	2009/5/14	約1ヶ月	NCBI アップロード済
10	非公開	非公開	454	不明	3.9 GB	未完了	未完了	対応中
11	非公開	非公開	Solexa	不明	不明	未完了	未完了	対応中

現在、DDBJ は NCBI SRA への登録を仲介していますが、DDBJ にて 登録、査定、登録番号の発行を行うべく準備を進めています。データ登録には、加工処理していない出力データ、もしくは、fastq データが必要です 登録者はメタデータを記入したエクセルファイルとデータファイルを DDBJ に送付してください

登録のご希望は☞ trace@ddbj.nig.ac.jp