文部科学省委託研究開発事業「統合データベースプロジェクト」

次世代シーケンサデータを活用するための目次サイトの構築

Functional indexing and curation of next-generation sequencing data

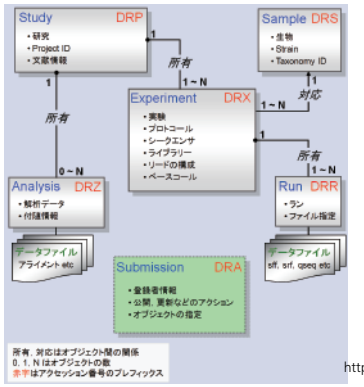
○仲里 猛留*、坊農 秀雅、高木 利久
Takeru Nakazato, Hidemasa Bono, Toshihisa Takagi
*：nakazato@dbcls.rois.ac.jp
情報・システム研究機構 ライフサイエンス統合データベースセンター
Database Center for Life Science (DBCLS), Research Organization of Information and Systems (ROIS)

Introduction

塩基配列 → DDBJ/EMBL/GenBank
遺伝子発現 → GEO ArrayExpress
文献 → PubMed
次世代シーケンサ → Short Read Archive (SRA)

これまで、塩基配列が GenBank に、マイクロアレイが GEO にと、さまざまな実験データが公共データベースに登録されてきた。同様に次世代シーケンサによるデータも sequence read archive (SRA) として NCBI、EBI、DDBJ に登録、公開するしくみが構築されている。

SRA でのデータ構造

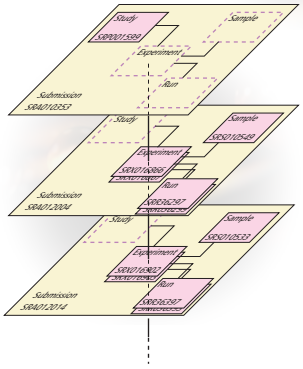


登録された次世代シーケンサによるデータにはリード（検出された塩基配列）だけでなく、実験情報（プロジェクト名、サンプルの生物種名、測定機器名）の情報がメタデータとして登録されている。これらの情報は、６種類の XML ファイル (submission, study, experiment, run, sample, analysis) に記載され、提供されている。しかしながら、１つの登録（Submission）にこれら６種類のメタデータが付与されているわけではない。（すでにあるプロジェクトに対して追加でデータを登録する場合があるため）

http://trace.ddbj.nig.ac.jp/dra/documentation.shtml より引用

Submission	Study	Experiment	Run	Sample	Analysis	
✓			✓			15535
✓				✓		3443
✓	✓	✓	✓	✓		3349
✓		✓				1426
✓		✓	✓			708
✓						273
✓	✓					262
✓		✓	✓	✓		254
✓	✓	✓		✓		102
✓	✓			✓		48
✓	✓	✓	✓			34
✓	✓	✓	✓	✓	✓	3
✓	✓	✓				3
✓					✓	3
✓		✓		✓		1

※ 2010-12-2 現在　　total 25444 (submissions)



Methods

```
<?xml version="1.0" encoding="UTF-8"?>
<EXPERIMENT_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <EXPERIMENT alias="4NG_TG-P3_044sA-FLX" accession="SRX003641">
    <TITLE/>
    <STUDY_REF refname="1000Genomes Project Pilot 3" accession="SRP000033"/>
    <DESIGN>
      <DESIGN_DESCRIPTION>454 Sequencing of Human Single-Direction Library from Nimblegen
Capture</DESIGN_DESCRIPTION>
      <SAMPLE_DESCRIPTOR refname="NA19179" accession="SRS000790"/>
      <LIBRARY_DESCRIPTOR>
        ...
      </LIBRARY_DESCRIPTOR>
      <SPOT_DESCRIPTOR>
        ...
    </DESIGN>
    <PLATFORM>
      <LS454>
        <INSTRUMENT_MODEL>GS FLX</INSTRUMENT_MODEL>
        <FLOW_SEQUENCE>TACG</FLOW_SEQUENCE>
        <FLOW_COUNT>100</FLOW_COUNT>
      </LS454>
      ...
    </PLATFORM>
    ...
```

SRA008310.experiment.xml

メタデータの XML ファイルを解析し、対応するリンクを抽出することで各種のメタデータを対応づけた。また、XML 中からプロジェクト名やプラットフォームなどの必要な情報を抽出した。

Results and Discussions

統計値

※ 2010-12-2 現在

Study Types

Whole Genome Sequencing	2149
Transcriptome Analysis	613
Metagenomics	404
Epigenetics	262
Resequencing	200
Other	192
Population Genomics	65
RNASeq	60
Gene Regulation Study	28
Cancer Genomics	11
Synthetic Genomics	3
Forensic or Paleo-genomics	2
Total (studies)	3989

Platforms

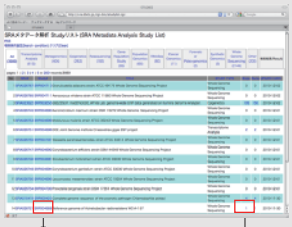
Illumina Genome Analyzer II	1053
454 GS FLX	1024
454 Titanium	805
Illumina Genome Analyzer	437
Solexa 1G Genome Analyzer	222
GS FLX	205
GS 20	135
unspecified	111
454 GS 20	48
AB SOLiD System 2.0	31
AB SOLiD System 3.0	27
AB SOLiD System	26
Illumina HiSeq 2000	8
Helicos HeliScope	8
454 GS	5
Illumina Genome Analyzer IIx	3
UNKNOWN	2
Total (studies)	4150

Species of samples (top 10)

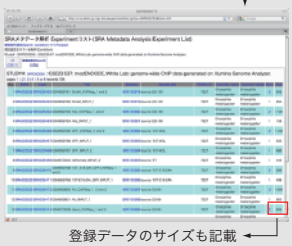
Homo sapiens	335
Mus musculus	175
metagenome sequence	169
Drosophila melanogaster	127
marine metagenome	81
Arabidopsis thaliana	44
Caenorhabditis elegans	40
Saccharomyces cerevisiae	38
synthetic construct	37
Panicum virgatum	21
...	
Total (studies)	4590

ウェブサイト

SRAs: survey of read archives
http://sra.dbcls.jp/



NCBI SRA の当該エントリへ
対応する Exp./Run のリストへ



登録データのサイズも記載

登録された次世代シーケンサによるデータについて目次サイトの構築を行った。Study (プロジェクト) ごとや対応する experiment のデータを見ることができる。また、実験目的 (Whole genome sequencing など) やプラットフォーム (Illumina Genome Analyzer II など) から対応する実験を表示させることもできる。本サービスは http://sra.dbcls.jp/ より自由に利用可能である（CC-BY ライセンス）。

現在、次世代シーケンサによる登録データを用いた論文から、当該データの閲覧も行えるよう開発を行っている。
（＋お茶大・理・情報　寺田、山田）

Conclusions

- 次世代シーケンサのデータは、SRA にアーカイブされており、リード（塩基配列）だけでなく、実験条件もメタデータとして付与されている。
- 追加の実験によるデータは既存の登録の追加でなく新規の登録とされるため、メタデータ内のリンクをたどることにより、対応する Submission を結びつけた。また必要な情報も抽出した。
- 次世代シーケンサ登録データの目次を構築した。http://sra.dbcls.jp/

BMB2010（第 33 回日本分子生物学会年会、第 83 回日本生化学会大会合同大会）
会場:神戸ポートアイランド　会期:2010年12月7日(火)～10日(金)

大学共同利用機関法人　情報・システム研究機構
ライフサイエンス統合データベースセンター
〒113-0032　東京都文京区弥生2-11-16　東京大学工学部12号館　TEL 03-5841-6754　FAX 03-5841-8090