Sequence Read Archive

Validation, Archival, and Distribution of Raw Sequencing Data

By Eugene Yaschenko – Head of Molecular Software Section, NCBI
11 years of software development

In 30 minutes

- Introduction
- Mass production revolution in Sequencing
- Trace Archive
- Next Generation of Massively Parallel Sequencing
- Sequence Read Archive
- SRA Toolkit
The Sequence Read Archive (SRA) was created and engineered at the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Department of Health and Human Services.

The SRA is part of a cluster of sequencing data repositories called the "Trace Archives“, and is located under the "Primary Data Archives" at NCBI, which includes GenBank.

The SRA is part of the International Nucleotide Sequence Database Collaboration (INSDC). The data model, data transfer protocols, and accession space are shared with the INSDC collaborators: European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ).
NCBI was created by Congress in 1988 to develop information systems, such as GenBank, to support the biomedical research community. NCBI was also mandated to conduct basic and applied research and, as part of the NIH Intramural Program, NCBI scientists work in areas of gene and genome analysis, computational structural biology and mathematical methods for sequence analysis.

As a part of National Library of Medicine, NCBI also creates archives of Medical and Biological literature, consumer health information sites.
NCBI is most widely known for
Bioinformatics resources at NCBI
Primary Data Archives are submitter-driven.

Data Archive is different from file archive. It stores data not original files.

Data Archive is not necessarily lossless. Some controlled loss of information or precision should be allowed.

Internal storage format should be sufficiently uniform to enable validation, searching, sub-setting, etc...

Extra effort is needed to support conversion from input formats as well as produce output formats. Large variety of formats significantly stresses archive’s resources.

Additional benefit of conversion is that all data are validated before archival.
Primary Data Archives

GenBank
- Internal storage: ASN.1
- Inputs: GenBank, EMBL, FASTA, GFF, ASN.1, etc...
- Bulk Outputs: GenBank, FASTA, ASN.1, BlastDB

TraceArchive
- Internal storage: Relational Database
- Inputs: SFF, ZTR, SCF, ABI
- Bulk Outputs: FASTA, Quality

SRA
- Internal storage: SRA format
- Inputs: SFF, SRF, FASTQ, SOLiD native, Illumina native, etc...
- Bulk Output: SRA format
Biologist’s perception of Raw Data Archive

Bioinformatician’s perception of Raw Data Archive
But similar perspective
Mass production in Sequencing

1998 - Applied Biosystems and Hitachi develops revolutionary ABI PRISM 3700 DNA Analyzer based on Sanger sequencing.

Fully automated sequencing, lab technicians were needed to prepare and load biological samples.

A few years later updated 3730 DNA Analyzer significantly improved automation and data quality.

The era of mass production of sequences begins. Since there were very little manual supervision of produced data – we labeled it as raw.
Trace Archive – first steps

Created by request of Mouse Sequencing Consortium in 2000

• Requested to contain 55 million records
• Input file format 200Kbyte per record
• Total input data size 11Tb (Total Usenet was 1.5 Tb)

Data Modeling

• Relational Approach – model all data and metadata as a rectangular table
• Normalize low-cardinality metadata
• Put bulky columns in separate tables
• Use data-warehousing databases for indexing

Data Compression

• Put a lot of effort into compressing signal data
• Managed to reduce footprint of a single record to < 20Kb
### Trace Archive Layout

**Design**

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

**RDBMS**

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

**DataWarehouse**

<table>
<thead>
<tr>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>
Trace Archive User Experience

User

Web Server

Data Storage

Index

query

list of ids

data

list of ids

update
Trace Archive – growth

Number of records

Prompted design corrections
- Vertical split of tables with large data
- Moving split tables into separate databases
- Finally, replacing non-active databases with a file-based storage

The design requirements were exceeded in a year.

Trace Archive went on beyond mouse.

Design requirement
Next Generation Sequencing

2005 – Trace Archive was approached by “454” company about archiving new type of data.

NCBI and 454 develop a strategy of storing this data in Trace Archive.

We were surprised to learn that technology is capable of producing 400,000 reads in 7 hours, but we did not suspect that this is just a beginning of new wave in the Sequencing Revolution.
Next Generation Sequencing

2006 – We start to hear reports about new technology from *Solexa* capable of producing tens of millions sequences in just a few days.

We come to realization that Trace Archive design is obsolete and will not be able to handle new wave.

We started to look for new approaches in designing Sequence Read Archive.
Trace Archive limits

Global integer identifier (ti) for all sequences

Data and metadata are mixed as columns in a simple rectangular table

Number of traces exceeded 2 billions

No convenient identifier for common project, sample, and methodology

Every search results in a large list of integer ids to be retrieved one-by-one
Trace Archive most important lessons

Model Data and Metadata separately

Metadata
- Model metadata not only for completeness but by ability to easily group the data
- Use of commercial relational databases is acceptable
- Heavy indexing is the requirement
- Assign accessions (simple identifiers) to metadata objects

Data
- Model data in uniform and compact format
- Try to make this format usable outside of the archive
- Use regular file system instead of relational databases

Software Toolkit
- Support internal format through software toolkit
Trace Archive – dangerously high growth

Number of records

- 454 submission started to increase

Design requirement

- Millions of records

- Time period: Jun-00 to Apr-07
Trace Archive - crisis averted

Number of records

454 submission diverted to SRA
THE SEQUENCE EXPLOSION

At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence - the three main databases for "finished" sequences, the National Center for Biotechnology Information (NCBI), the DNA DataBank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. This number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace Archive and Sequence Read Archive (SRA).

DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration

The main repositories of "finished" sequence span a wide range of organisms, representing the many priorities of scientific workflows.

HOW MANY HUMAN GENOMES?

The graphic shows all published, fully sequenced human genomes since 2000, including nine from the first quarter of 2010. Some are resequencing efforts on the same person and the list does not include unpublished completed genomes.
SRA design challenges

Multiple sequencing platforms
- Roche 454
- Illumina Genome Analyzer
- Applied Biosystems SOLiD System
- Helicos Heliscope
- Complete Genomics
- Pacific Biosciences SMRT
- Ion Torrent

Multiple input formats
- 454 Standard Flowgram Format (SFF)
- Sequence Read Format (SRF)
- ASCII record-based formats (fasta, fastq, etc.)
- ASCII tabular formats (qseq, etc.)
- HDF5

Multiple sequencing application
- Whole Genome Sequencing
- Whole Exome Sequencing
- Medical resequencing
- Metagenome sequencing
- Functional Genome
Storage size per base pair is important

Cancer paired whole genome sequence - samples (thousands)

Assumptions
500 gigabases of sequence per sample pair (tumor + normal) • curation cost = $40.00 per sample pair
dbGaP registration fee = $1000.00 per study • Participant consent for dbGaP distribution
• storage cost (disk + tape) = $1500.00 per terabyte

NCBI 2011 SRA cost models

$ cost (w) = 40 (u samples) + \frac{750}{8} (v \text{ bits per base})(u \text{ samples}) + 1000
SRA Model

Metadata is lifted to the level of physical organization of the platform

Metadata is modeled with XML schema

Metadata is stored in XML-aware RDBMS

Data is order-independent array of internal elements of the platform

Data is divided into parallel series, compressed and stored in binary files.
SRA User Experience

User

query

file ids

Web Server

Indexed Metadata

File Servers – http/ftp/fasp/etc..
### Comparative Table Layout

#### Trace Archive

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Billions of rows

#### Sequence Read Archive

<table>
<thead>
<tr>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

Billions of rows

Millions of rows
SRA Data Processing Pipeline

Sequencing Machine generates data in machine format

Data is converted into common storage format

Data is stored and indexed for retrieval

Data is converted into user preferred format

Data is analyzed

Toolkit

Past
Present
Future
SRA Data – common denominator

Raw Data

Bytes/Base

Thousands

40 million clusters per flow cell

20 microns
SRA Data – common denominator

<table>
<thead>
<tr>
<th>Raw Data</th>
<th>Signal Data</th>
<th>Bytes/Base</th>
<th>Thousands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 - 50</td>
</tr>
</tbody>
</table>
SRA Data – common denominator

<table>
<thead>
<tr>
<th>Raw Data</th>
<th>Signal Data</th>
<th>Basecalls and Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thousands</td>
<td>4 - 50</td>
<td>0.6 - 0.9</td>
</tr>
</tbody>
</table>

One channel quality score

```
tcagGGGGGAGCTTAATTTGAAACTAGAAAAATTTTTGAAACAAAATAATCATATTGTGTA
GCTGATGAAAAACTAGAAAAGATTTTTGAGTTggaaccgaaagggttgaattc
cccccttttcggggcattccacGCTATCCGTAAGGTCATCCCTGCTCTGGGATACAGCTAG
CTCCCAATTTCCATAAACAAACTCCTTGTAAGTAAACCTCCTTTTGAACAGGGGTACTGAG
CGGGCTGGCAAGGCC
```

Bytes/Base
### SRA Data – common denominator

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Bytes/Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Data (not stored)</td>
<td>Thousands</td>
</tr>
<tr>
<td>Signal Data</td>
<td>4 - 50</td>
</tr>
<tr>
<td>Basecalls and Quality</td>
<td>0.6 - 0.9</td>
</tr>
<tr>
<td>original SRA</td>
<td>1 - 50</td>
</tr>
</tbody>
</table>

![Image of DNA sequence](image)
SRA Data – common denominator

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Description</th>
<th>Bytes/Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Data (not stored)</td>
<td>Thousands</td>
<td></td>
</tr>
<tr>
<td>Signal Data (not stored*)</td>
<td>4 - 50</td>
<td></td>
</tr>
<tr>
<td>Basecalls and Quality</td>
<td>0.6 - 0.9</td>
<td></td>
</tr>
<tr>
<td>current SRA</td>
<td>0.6 - 0.9</td>
<td></td>
</tr>
</tbody>
</table>

* All signal data was removed from the archive, except for the 454 platform
Divide data into series, store by column
Eliminate repeated cells
Apply type-aware compression
Tune for retrieval speed vs. size tradeoff
Fully indexed long-term online/nearline archive storage
Evolution-aware backward compatibility
Fagernes Airport, Leirin (ENFG/VDB) | Victor David Brenner (designer of a U.S. one-cent coin) | Paul Vanden Boeynants, Belgian prime minister | Frank Vandenbroucke, cyclist | Vrijzinnig Democratische Bond, a Dutch political party | Video Data Bank | Virtual Database | Visual Database | Voluntary Denied Boarding | Verband der Bahnindustrie in Deutschland e.V., the German railway industry association | /var/db/pkg, the installed package database in FreeBSD and Gentoo Linux
Columns group like-data for better compression

Good for accessing a reduced set of columns

Schema-driven

Virtualization layer – data from store or created from available information

Highly distributed - takes advantage of host file-system

Flexible - allows removal of individual columns in archive
Write Side of SRA Toolkit

Input Data Format

Format Parser

Input Columns
- A
- B
- C
- D

Writing Schema
- XF1

Physical Columns
- X
- B
- Y
- Z

Encoding Schema
- EF1
- EF2

Serialization
- X
- B
- Y
- Z

Directory Structure

SRA File Format

vdb

kdb

SRA000001
- col
- CLIP_QUALITY
- QUALITY
- READ
- SIGNAL

SRA Toolkit Input Data Format

Writing Schema

Physical Columns

Encoding Schema

Serialization

Directory Structure

SRA File Format
Read Side of SRA Toolkit

SRA Data Directory or SRA File Format

Deserialization

Decoding Schema

Physical Columns

Read Schema

Output Columns

Format Generator

API to Read Columns

Output File Format
Schema Examples

- simple assignment
  
  ```
  ascii S = B;
  ```

- assignment using plugin function
  
  ```
  U32 A= <U32> clip <15, 10000> (B);
  ```

- conditional assignment
  
  ```
  ascii S = S1 | S2;
  ```

- Defining encoding rule using plugin functions
  
  ```
  physical <type T> T izip_encoding #1.0 {
  decode { return (T) iunzip (@); }
  encode { return izip (@); }
  }
  ```

- Using encoding rules
  
  ```
  extern column <U32> izip_encoding A;
  extern column <I64> izip_encoding C;
  ```
Schema: using existing functions

AATCGGTAAACCG

<ascii,U8>map<'ACGT', [ 0, 1, 2, 3 ]> (@)

{0,0,3,1,2,2,3,0,0,1,1,2}

pack<8,2> (@)

00001101,10101100,00010110
Privacy Issues

SRA Content by organism

- **Homo sapiens**: 61%
- **human metagenome**: 6%
- **Mus musculus**: 5%
- **other metagenome**: 4%
- **Plasmodium falciparum**: 2%
- **Drosophila melanogaster**: 2%
- **Danio rerio**: 1%
- **Mustela putorius furo**: 1%
- **Latimeria chalumnae**: 0%
- **Otolemur garnettii**: others

**Amount of human sequencing will only increase**

In a series of studies, NIH is planning to sequence tens of thousands people.

Genetic data is highly personal.

Most of human sequencing data will require careful protection from unauthorized distribution.
Encrypted SRA format

Single file SRA Format

32Kb  32Kb  32Kb  32Kb  32Kb  32Kb  15Kb

256 bit chained encryption

• Original file information
• Length, checksums
• Extra parameters used for randomizing encryption

• Encrypted block checksum and length
• Needed for checking transfers without decrypting the file
Compression by Reference

Currently, massively parallel sequencing created enough fragments to cover subject’s genome several times (4x-100x)

This is done to improve signal/noise ratio

In many studies the very next step after sequencing is mapping to the Reference Genome

After the mapping, large fraction of fragments perfectly match the Reference
Basic components of aligned sequences

References

Alignments

Sequences

cSRA
Redundancy in aligned data

Many sequences repeat

Only need to store differences
Unaligned sequences can be separated from aligned
All matching bases are dropped. Only the mismatched bases are stored. Qualities are retained.

Unaligned sequences are treated as traditional SRA data.
The original sequence may be restored by applying a function to the reference and the stored differences. There is no loss of information.
• References to GenBank and RefSeq accessions are typically stored as Remote
• References to de-novo assemblies or modified sequences will be stored as Local
• Local Reference may be useful for storing Transcriptome and Metagenome assemblies, which will make those projects completely contained within SRA
cSRA

VDB-2 table structure for “Archive” version

Needs:
- Access to Remote Ref.

Stores:
- Reference
  - Local Sequences, Statistics

Produces:
- Restored Sequences
- Restored Sequences

Sequences From Ref.
Primary Alignment For mapped reads

Primary Alignment
Secondary Alignment

Primary Alignment

Sequences

Placements, Mismatches, Indels

Sequence

Qualities, unplaced reads, groups

All Basecalls And Qualities

Alternate Placements, Indels

Restored Sequences

Restored Sequences
New Data Dimension

Sequence Read Archive

- Metadata
- Data

Millions of rows

SRA Multisample View

<table>
<thead>
<tr>
<th>Data1</th>
<th>Data2</th>
<th>Data3</th>
<th>Data4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Query Range

Reference Position
Credits

**SRA Toolkit Development**  sra-tools@ncbi.nlm.nih.gov
- Kenneth Durbrow
- Anton Golikov
- William Killian
- Andrei Klymenko
- Wolfgang Raetz
- Kurt Rodarmer
- Eugene Yaschenko  yaschenk@ncbi.nlm.nih.gov

**SRA Pipeline Development**
- Michael Kimelman
- Yuri Ostapchuk

**Project manager:**
- Reza Safarnejad

**SRA Curators**
- Zinaida Belaya
- Christopher O'Sullivan
- Robert Sanders
- Martin Shumway
- Yuriy Skripchenko
- Adam Stine

**SRA Website**
- Sergiy Ponomarev