



“The important thing is not to stop questioning. Curiosity has its own reason for existing.” - Albert Einstein



Drug Discovery | Clinical Development | Manufacturing Labs | Genomics

R&D IT | Clinical Biometrics | Compliance Solutions | Scientific Process Management

TOPICS COVERED

- **Sequencing Technologies**
- **Exploring Sequencing Database (SRA, ERA, DRA)**



Sequencing Technologies

SEQUENCING: Definition

- In genetics and biochemistry, **sequencing** means to determine the primary structure (sometimes incorrectly called the primary sequence) of an unbranched biopolymer. (<https://en.wikipedia.org/wiki/Sequencing>)
 - » Types
 - » Principles
 - » Their “+” and “-”

TECHNOLOGIES

Company	Platform	Amplification	Sequencing Method
Illumina	MiSeq NextSeq NovaSeq	Bridge PCR	Synthesis (SBS)
ThermoFisher	Ion Torrent	emPCR	Synthesis (pH)
Pacific Biosciences	Sequel II	None	Synthesis
Oxford Nanopore Technologies	MinION GridION PromethION	None	Flow

DIFFERENCES BETWEEN PLATFORMS

- **Technology : Chemistry + Signal Detection**
- **Run Time : Varies from hours to days**
- **Product Range : Mb – Gb**
- **Read Length : < 100 bp to >20 Kbp**
- **Accuracy per base: 0.1 % to 15%**
- **Cost per base: Varies**

ILLUMINA

Instrument	Yield and Run Time*	Read Length	Error Rate#	Error Type
MiSeq	15 Gb, 4-55 hrs.	2 X 300 bp	0.47	Substitution
NextSeq	330 Gb, 11-48 hrs.	2 X 150 bp	0.59	Substitution
NovaSeq	6000 Gb, 13-44 hrs.	2 X 250 bp	0.11	Substitution

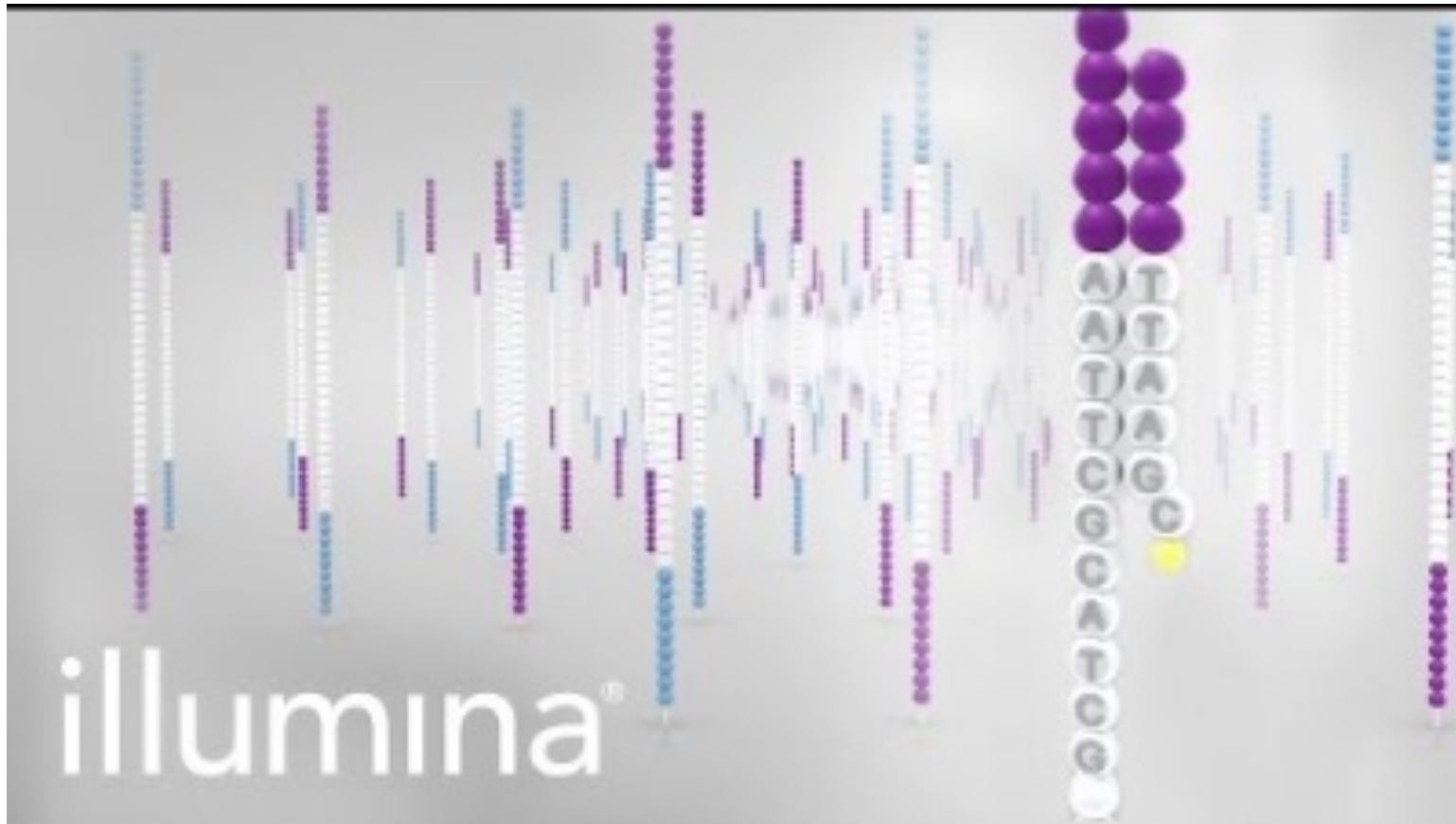
* <https://sapac.illumina.com/systems/sequencing-platforms.html>

#<https://academic.oup.com/nargab/article/3/1/lqab019/6193612>

■ Main Application

- » Whole Genome, Whole Exome, Targeted Panel Sequencing
- » Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling, miRNA)
- » Single-Cell Profiling (scRNA-Seq, snDNA-Seq, scDNA-Seq, oligo tagging assays)
- » Methylome and ChIPSeq
- » Metagenomic Profiling (16S RNA, shotgun metagenomics, metatranscriptomics)

ILLUMINA: Sequencing By Synthesis (SBS)



THERMOFISHER

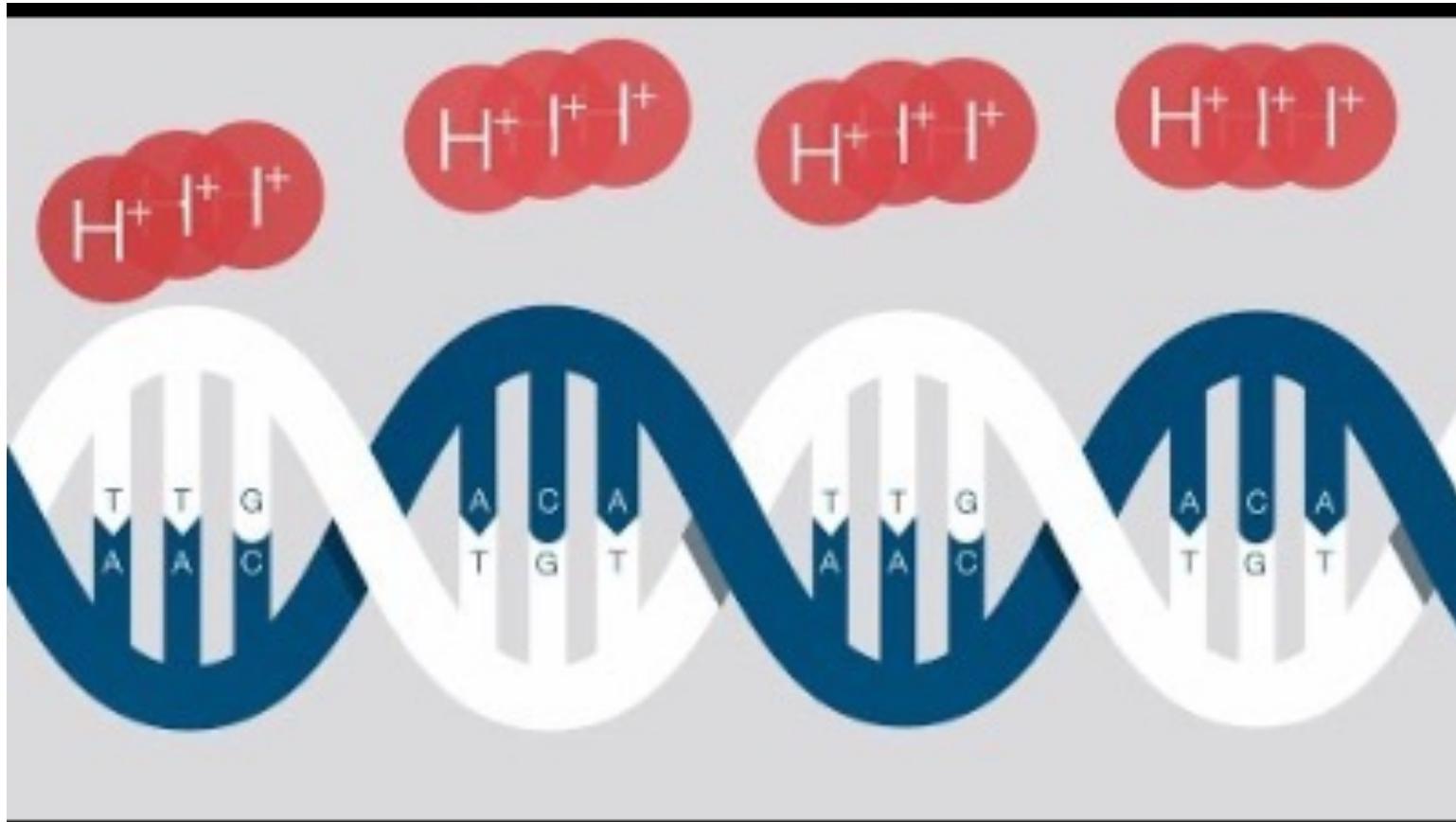
Instrument	Yield and Run Time	Read Length	Error Rate [#]	Error Type
Ion Torrent Genexus System	50 Gb, 3-22 hrs.	200-600 bp	1.5%	InDel

#<https://www.nature.com/articles/s41598-017-08139-y>

■ Main Application

- » Small Genomes, Whole Exome, Targeted Panel Sequencing
- » Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling, miRNA)
- » Metagenomic Profiling (16S RNA)

ION SEMICONDUCTOR SEQUENCING (pH)



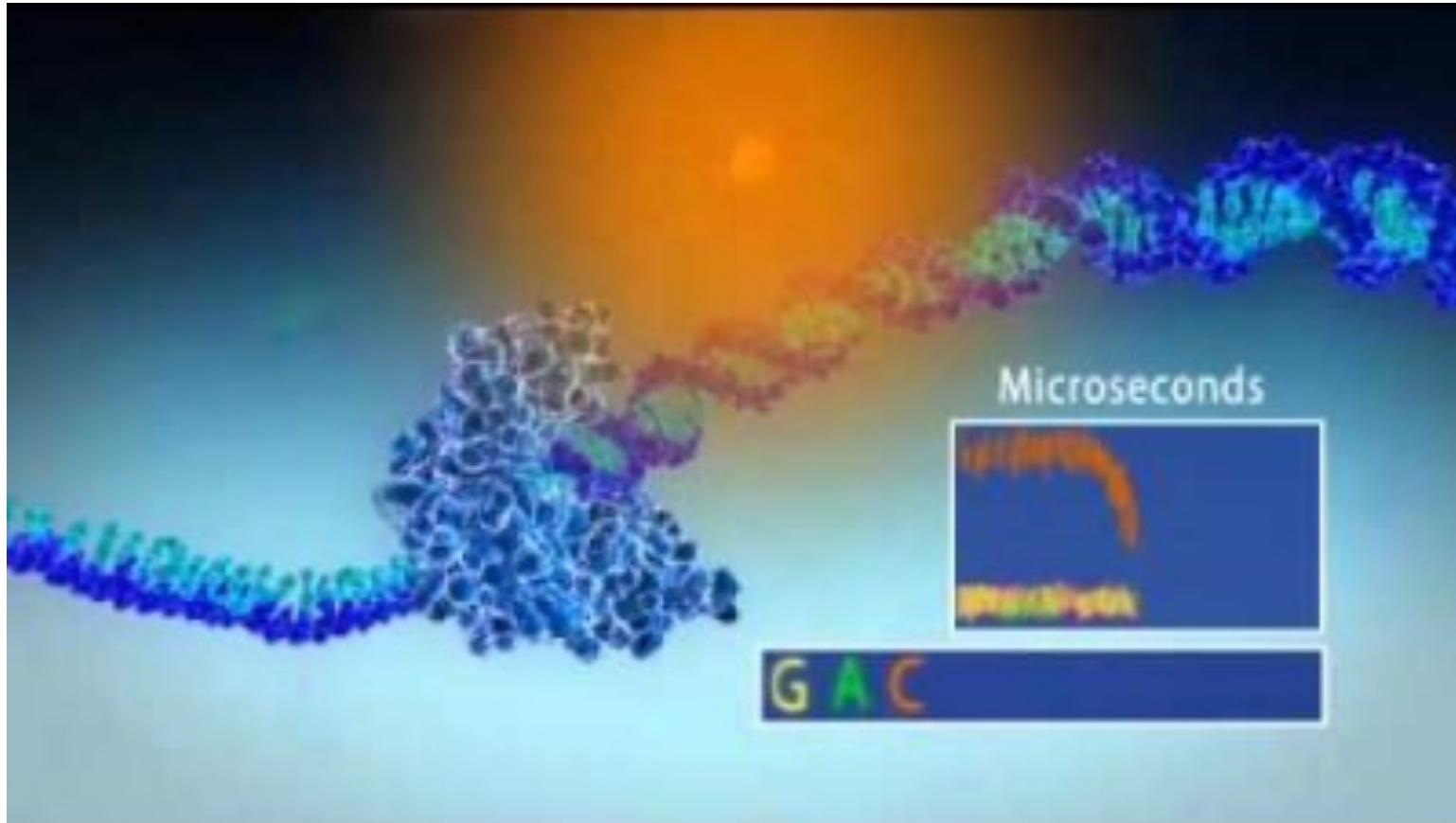
Instrument	Yield and Run Time	Read Length	Error Rate#	Error Type
Sequel II	250-450 Gb, ~30 hrs.	~ 15-20 Kbp	13-15%	Mismatches

#<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553090/>

■ Main Application

- » Large Genomes
- » Transcriptome Sequencing (total RNA-Seq, mRNA-Seq)
- » Metagenomic Profiling (Shotgun Sequencing)

SINGLE MOLECULE REAL TIME SEQUENCING (SMRT)



OXFORD NANOPORE TECHNOLOGIES

Instrument	Yield and Run Time	Read Length*	Error Rate [#]	Error Type
MinION	10-50 Gb, ~72 hrs.	~ 882 Kbp	~30%	Mismatches

*<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5889714/>

#<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5553090/>

■ Main Application

- » Large Genomes, Targeted Sequencing
- » Transcriptome Sequencing (total RNA-Seq, mRNA-Seq)
- » Metagenomic Profiling (Shotgun Sequencing)
- » Epigenetics

NANOPORE SEQUENCING

THE NANOPORE

The heart of our technology

- In nature, protein nanopores function as gateways between two systems.
- We have carefully engineered protein nanopores through mutating key residues in the barrel of the pore.



© Copyright 2019 Oxford Nanopore Technologies. All rights reserved.



WHAT IS THE BEST ?

- Design your experiment based on the scientific question.
- Choose the best suited application for your project.
- Find the most optimal sequencing technology.
- Answer all questions about our technologies and applications, as well as bioinformatics (down stream analysis).

LIBRARY PREPARATION FOR SEQUENCING

Fragmentation & End Repair

- Sequencers like Illumina cannot analyse long strands.
- Generating uniform shorter pieces.
- Repairing ends.

Addition of Adapters

- Linking adapters to the DNA fragments.
- Adapters serve multiple purposes: attaching to flow cell, index to demultiplex samples during analysis.

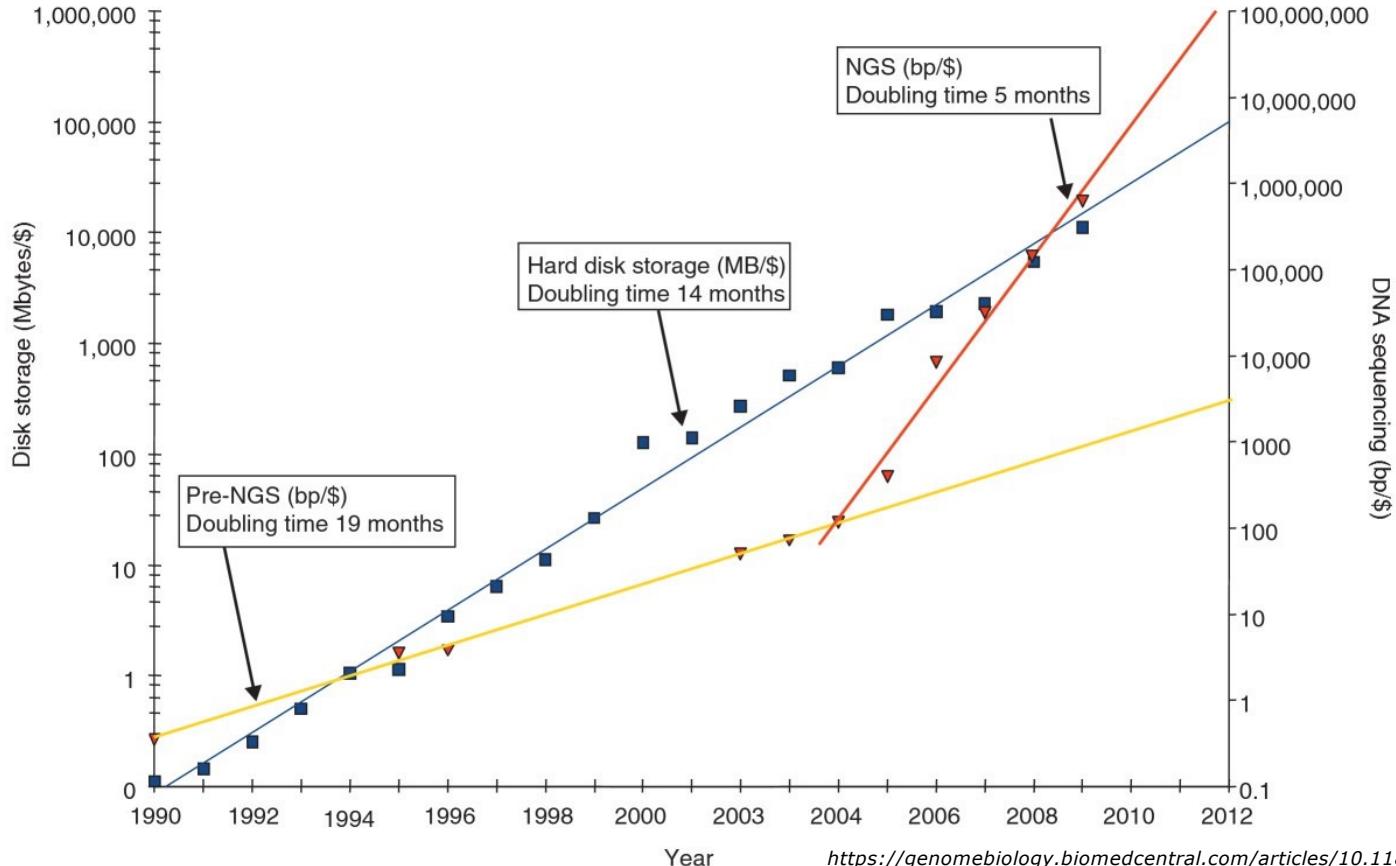
PCR Amplification

- Optional step



Sequence Database

SEQUENCING STORAGE / COST



RAW DATA ARCHIVING

Biologist's Perspective



Bioinformatician Perspective



SEQUENCING READ ARCHIVE (SRA)

- 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).

PRIMARY DATA ARCHIVE

RAW Data Archives

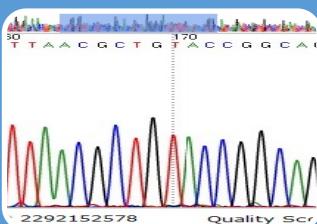
```
2345  
ceros melampus iso  
chondrial.  
45  
45.1 GI:540023  
  
chondrion Aepycero  
ceros melampus  
ryota; Metazoa; Ch  
alia; Eutheria; La  
ra; Bovidae; Aepyc  
bases 1 to 426)  
ander, P., Kat, P.,  
lation genetics of
```

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

```
MO  
i:1, x:684, y:3338  
  
RX  
i:1, x:80, y:843  
  
RI  
i:1, x:1654, y:3340  
  
MO  
i:1, x:1747, y:1248  
  
I86  
i:1, x:1906, y:1384  
  
Inte
```

SRA

Internal storage: SRA format

Inputs: SFF, SRF, FASTQ, SOLiD native, Illumina native, etc...

Bulk Output: SRA format

PRIMARY DATA ARCHIVE: KEY FEATURES

- 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.

NCBI-SEQUENCE READ ARCHIVE (SRA)

NCBI Resources How To Sign in to NCBI

SRA SRA Advanced Search

COVID-19 Information Public health information (CDC) | Research information (NIH)

Sequence Read Archive Main Browse Search Download Submit Software Trace Archive Trace BLAST Overview

COVID-19 Information Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, Ion Torrent, Complete Genomics, PacBio and OxfordNanopores. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence. SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them. Please check [SRA Overview](#) for more information.

Submitting to SRA
Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

- [Submission Quick Start](#)
- [Frequently Asked Questions and Troubleshooting](#)
- [Log in to Submission Portal](#) (for submitting sequence data)
- [Log in to SRA](#) (for updating and troubleshooting submissions)

Using SRA Data with SRA Toolkit
Use SRA data to validate experimental results, increase sample sizes, determine variance and open up new avenues of research.

- [SRA Download Guide](#)
- [SRA Toolkit Usage Guide](#)
- [Software Download](#)
- Get sources code on [GitHub](#) (for developers using SRA)

SRA database growth

Year	Total Bases (Terabases)	Open Access Bases (Terabases)
2009	~1	~1
2010	~10	~10
2011	~100	~100
2012	~1000	~1000
2013	~10000	~10000
2014	~100000	~100000
2015	~1000000	~1000000
2016	~10000000	~10000000
2017	~100000000	~100000000
2018	~1000000000	~1000000000
2019	~10000000000	~10000000000
2020	~100000000000	~100000000000

12/12/2022 06:07pm
Save in CSV format

<https://www.ncbi.nlm.nih.gov/sra>

<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement>

SRA: WALK THROUGH

NCBI Resources How To Sign in to NCBI

SRA SRX13301622 Search

Studies Samples Analyses Run Browser Run Selector Provisional SRA

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Amplicon sequencing of *S. Cerevisiae* CAN1 (SRR17116907)

Full ▾

SRX13301622: Amplic 1 ILLUMINA (Illumina V)

Design: Tiled-PCR of C

Submitted by: UB Ger

Study: Complex Mutation Homolog (MSH) Compl PRJNA785873 • Show Abstract

Sample: WT_perm-Rep SAMN23587586 : Organism: Saccharomyces cerevisiae

Library:
Name: WT_perm-F
Instrument: illumina;
Strategy: AMPLICON
Source: GENOMIC
Selection: PCR
Layout: PAIRED

Runs: 1 run, 223,445 s
Run # SRR17116907

ID: 18192346

SRX13301622 Metadata Analysis Reads Data access

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR17116907	223.4k	134.5Mbp	82.1M	41.2%	2021-12-07	public

Quality graph (bigger)

This run has 2 reads per spot:
L=301, 100% L=301, 100%

Legend

Experiment SRX13301622 **Library Name** WT_perm-Rep6 **Platform** Illumina **Strategy** AMPLICON **Source** GENOMIC **Selection** PCR **Layout** PAIRED

Design: Tiled-PCR of CAN1

Biosample SAMN23587586 (SRS11211204) **Sample Description** Sample Description **Organism** *Saccharomyces cerevisiae* **Links** PRJNA785873

Bioproject PRJNA785873 **SRA Study** SRP349121 **Title** Complex Mutation Profiles in Mismatch Repair and Ribonucleotide Reductase Mutants Reveal Novel Repair Substrate Specificity of MutS Homolog (MSH) Complexes

Abstract:
Targeted sequencing revealed complex mutation profiles in yeast strains that have elevated dNTP levels paired with deletions in mismatch repair genes.

NCB Site map All databases Search

Sequence Read Archive Main Browse Search Download Submit Software Trace Archive Trace BLAST

Studies Samples Analyses Run Browser Run Selector Provisional SRA

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Amplicon sequencing of *S. Cerevisiae* CAN1 (SRR17116907)

Metadata Analysis Reads Data access

Taxonomy Analysis

NCBI Site map All databases Search

Sequence Read Archive Main Browse Search Download Submit Software Trace Archive Trace BLAST

Studies Samples Analyses Run Browser Run Selector Provisional SRA

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Amplicon sequencing of *S. Cerevisiae* CAN1 (SRR17116907)

Metadata Analysis Reads Data access

SRA archive data
SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Size	Location	Name	Free Egress	Access Type
run	80,196 Kb	NCBI	https://sra-download.ncbi.nlm.nih.gov/traces/sra69/SRR017116/SRR17116907/WT-Rep9.R1.fastq.gz	worldwide	anonymous

Original format
The original files submitted to SRA. These files may require specific software to open, read and interpret data.

Type	Size	Location	Name	Free Egress	Access Type
fastq	29,329 Kb	NCBI	https://sra-download.ncbi.nlm.nih.gov/traces/sra69/SRR017116/SRR17116907/WT-Rep9.R1.fastq.gz	worldwide	anonymous
fastq	32,545 Kb	NCBI	https://sra-download.ncbi.nlm.nih.gov/traces/sra69/SRR017116/SRR17116907/WT-Rep9.R2.fastq.gz	worldwide	anonymous

Egress and Access: what does it mean? Why is SRA data in the cloud? What is "Cloud Data Delivery"?

SRA: DOWNLOAD (SRA Toolkit)

NCBI SRA Toolkit

Below are the latest releases of various tools and release checksum file.

SRA Toolkit

Compiled binaries/install scripts of October 25, 2021, version 2.11.3:

- [CentOS Linux 64 bit architecture](#) - non-sudo tar archive
- [Ubuntu Linux 64 bit architecture](#) - non-sudo tar archive
- [Cloud - apt-get install script](#) - for Debian and Ubuntu - requires sudo permissions
- [Cloud - yum install script](#) - for CentOS - requires sudo permissions
- [MacOS 64 bit architecture](#)
- [MS Windows 64 bit architecture](#)
- [Docker image repository](#)
- [md5 checksums](#)

Magic-BLAST

Magic-BLAST is a tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome.

- Magic-BLAST executables for LINUX, Mac OSX, and Windows as well as the source files are available on the [FTP site](#)
- Read more about Magic-BLAST on the [FTP site](#)

Third Party Software

Builds of Third Party Software Tools with SRA support:

- Genome Analysis Toolkit (GATK):
 - [version 3.6-6/ngs.2.11.2](#) - including direct support of SRA (NGS 2.11.2 release)
 - [version 4.2.0.0](#) - including NGS 2.11.2 release
- HISAT2 version 2.2.1-ngs.2.11.2 - graph-based alignment of next generation sequencing reads to a population of genomes with direct support of SRA, built for:
 - [CentOS Linux 64 bit architecture](#)
 - [MacOS 64 bit architecture](#)

Latest Source Code

- [NGS Software Development Kit](#) – October 7, 2021, version 2.11.2 release
- [NCBI VDB Software Development Kit](#) – October 7, 2021, version 2.11.2 release
- [NCBI SRA Toolkit](#) – October 25, 2021, version 2.11.3 release
- [NCBI NGS Toolkit](#) – October 7, 2021, version 2.11.2 release

File checksums

You may validate downloaded files with [md5 checksums](#) computed using `md5sum -b`

SRA Toolkit Documentation

[SRA Toolkit Installation and Configuration Guide](#)

[Protected Data Usage Guide](#)

Frequently Used Tools:

- [fastq-dump](#): Convert SRA data into fastq format
- [prefetch](#): Allows command-line downloading of SRA, dbGaP, and ADSP data
- [sam-dump](#): Convert SRA data to sam format
- [sra-pileup](#): Generate pileup statistics on aligned SRA data
- [vdb-config](#): Display and modify VDB configuration information
- [vdb-decrypt](#): Decrypt non-SRA dbGaP data ("phenotype data")

Additional Tools:

- [abi-dump](#): Convert SRA data into ABI format (csfasta / qual)
- [illumina-dump](#): Convert SRA data into Illumina native formats (qseq, etc.)
- [sff-dump](#): Convert SRA data to sff format
- [sra-stat](#): Generate statistics about SRA data (quality distribution, etc.)
- [vdb-dump](#): Output the native VDB format of SRA data.
- [vdb-encrypt](#): Encrypt non-SRA dbGaP data ("phenotype data")
- [vdb-validate](#): Validate the integrity of downloaded SRA data

<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>

European Nucleotide Archive (ENA)



ENAS

European Nucleotide Archive

Home | Submit | Search | Rulespace | About | Support

Message posted 2020-11-19.

We recommend that you subscribe to the [ENA-announce mailing list](#) for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a [Drag-and-Drop Data Submission Service](#) (currently in Beta) suitable for certain SARS-CoV-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

DDBJ SEQUENCE READ ARCHIVE (DRA)

 DDBJ Services SuperComputer Statistics Activities About Us

DDBJ Web Sites Terms Contact Japanese

BLAST and ClustalW tentative closing

Sequence Read Archive

Home Handbook ▾ FAQ Search Downloads ▾ About DRA

Home > dra > Sequence Read Archive

DDBJ Sequence Read Archive (DRA) is the public archive of high throughput sequencing data. DRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis. DRA is a member of the International Nucleotide Sequence Database Collaboration (INSDC) and archiving the data in a close collaboration with NCBI Sequence Read Archive (SRA)  and EBI Sequence Read Archive (ERA) .

 Search  How to submit  Login and submit

NEWS

Public DRA data released before 6th December 2021 have been mirrored to NCBI SRA
2021/12/08 Announcement BioProject BioSample DRA DDBJ Center

D-way downtime 2021-12-15 14:00-15:00
2021/12/07 Announcement DDBJ BioProject BioSample DRA DDBJ Center

Sequence data release of 225 strains of *Serratia marcescens*
2021/12/06 Data Release DDBJ BioProject BioSample DRA DDBJ Center

(Dec. 29 - Jan. 3) Suspension of the BI-DDBJ activity during the New Year Holidays
2021/12/01 Announcement DDBJ BioProject BioSample DRA GEA JGA AGD DDBJ Center

Sequence data release of emu (*Dromaius novaehollandiae*), southern cassowary (*Casuarius casuarius*), southern ostrich (*Struthio camelus australis*)
2021/11/26 Data Release DDBJ BioProject BioSample DRA DDBJ Center

more 



