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

ZIFO RnD SOLUTIONS

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**

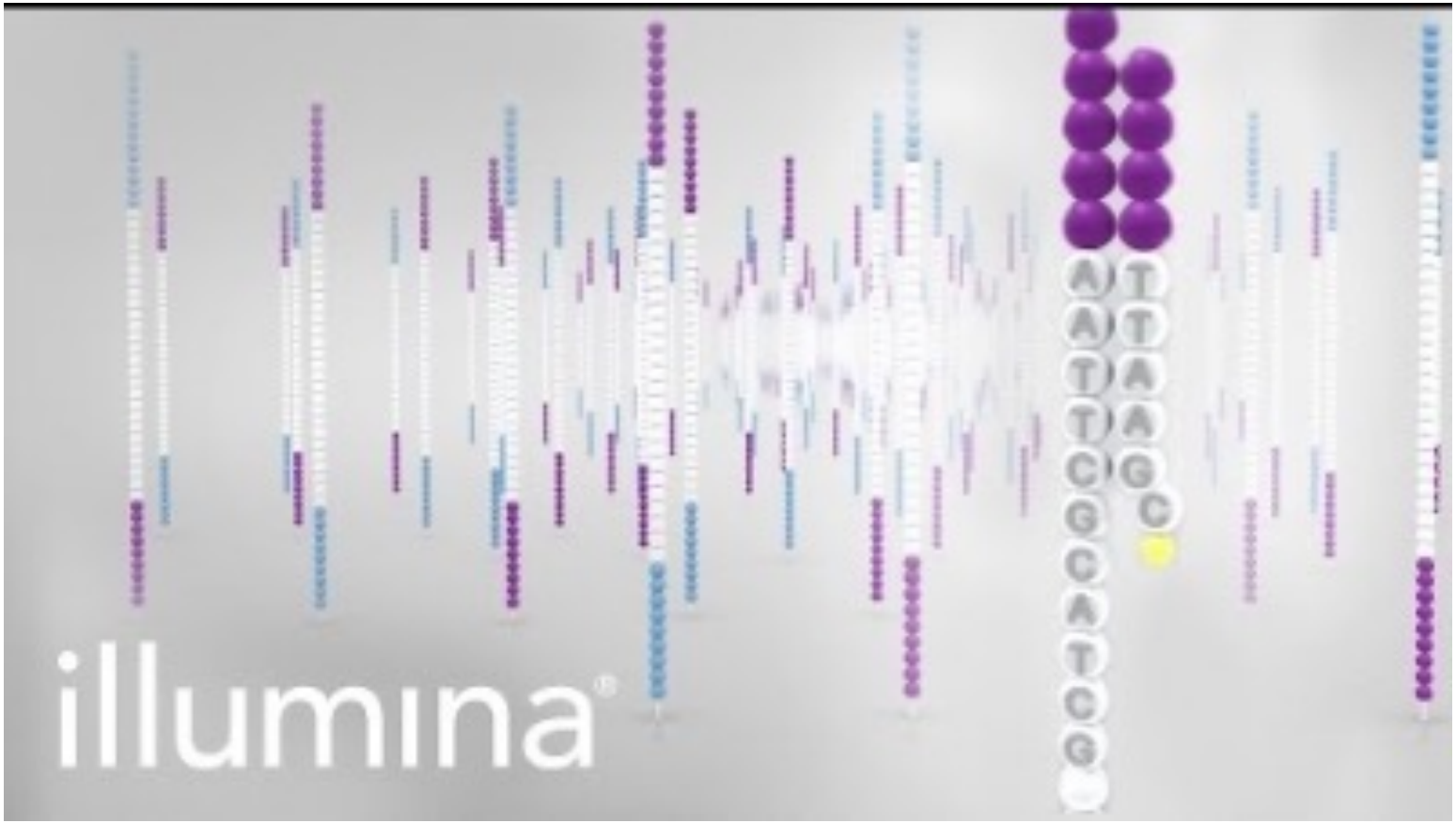
ILLUMINA

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)



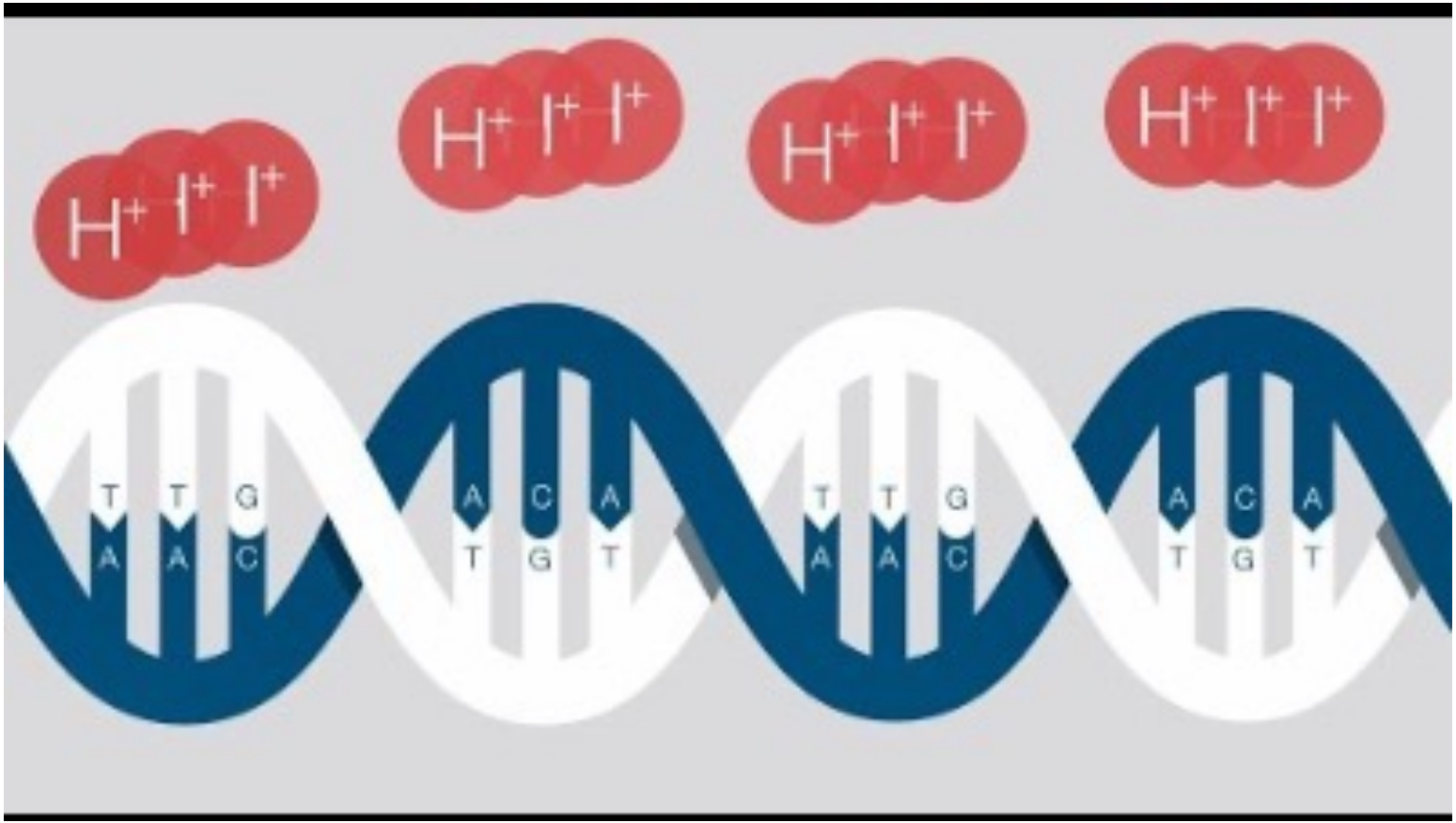
THERMOFISHER

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)



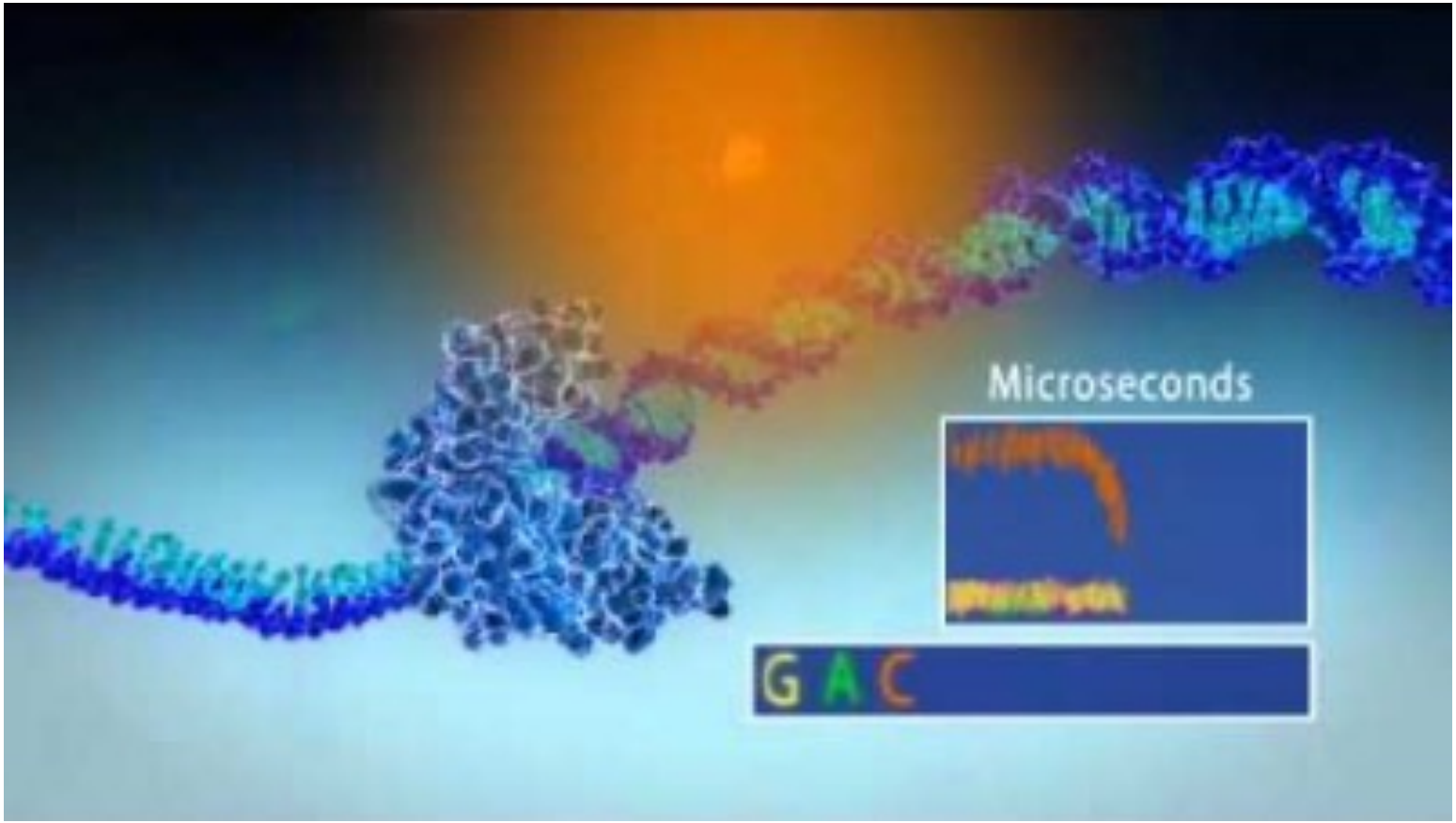
PACIFIC BIOSCIENCES

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.

NANOPORE

WHAT IS THE BEST ?

- **Design your experiment based on the scientific question.**
- **Chose 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 unform shorter pieces.
- Repairing ends.

Addition of Adapters

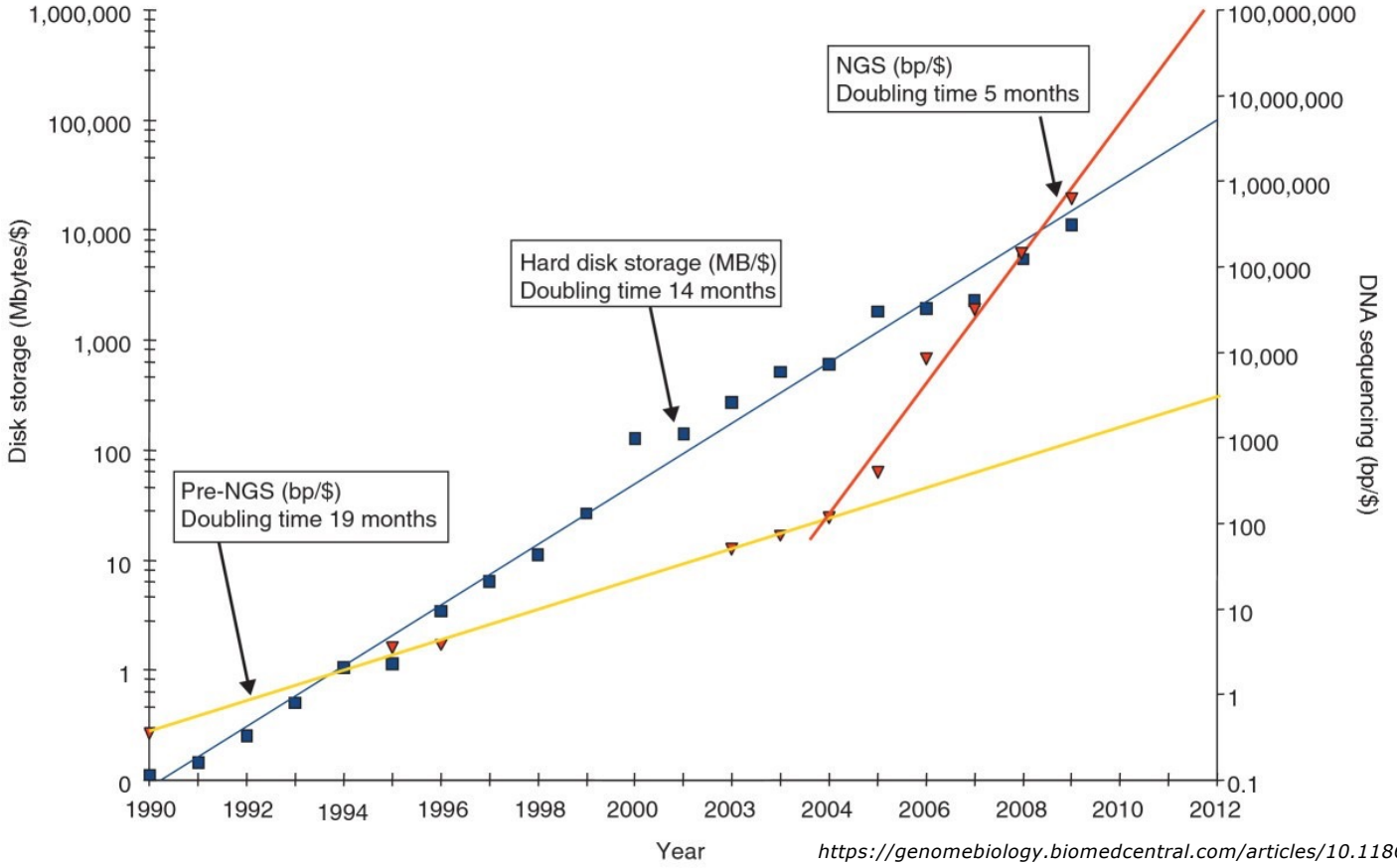
- Linking adapters to the DNA fragments.
- Adapters serves multiple purposes: attaching to flow cell, index to demultiplex samples during analysis.

PCR Amplification

- Optional step

Sequence Database

SEQUENCING STORAGE / COST



https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-5-207

RAW DATA ARCHIVING

Biologist's Perspective

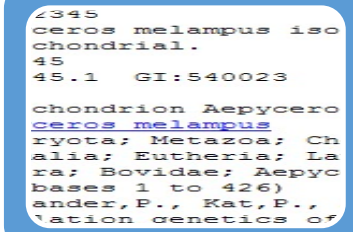
Bioinformatician Perspective

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

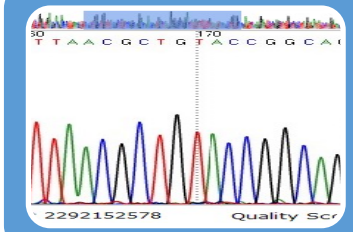


GenBank

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

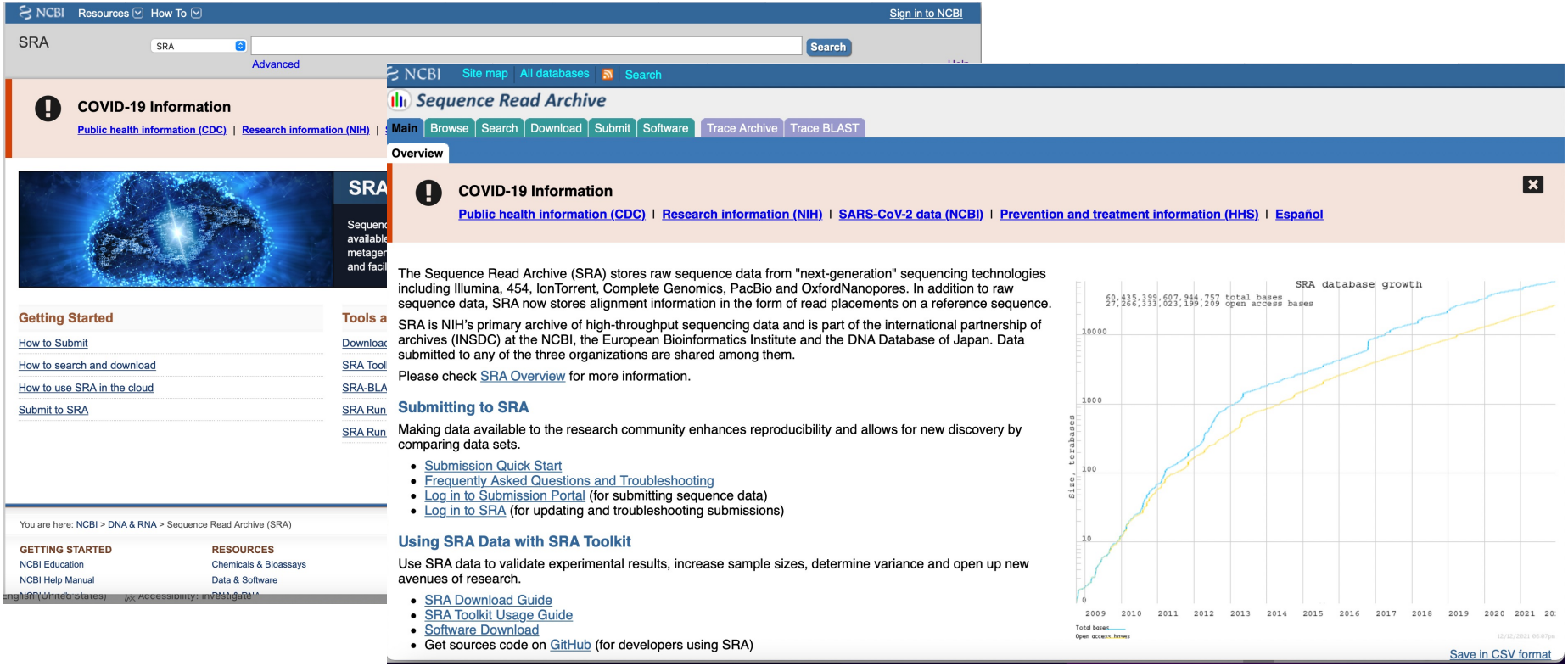
Inputs: SFF, SRF, FASTQ, SOLiD native, Illumina native, etc…

Bulk Output: SRA format

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.**

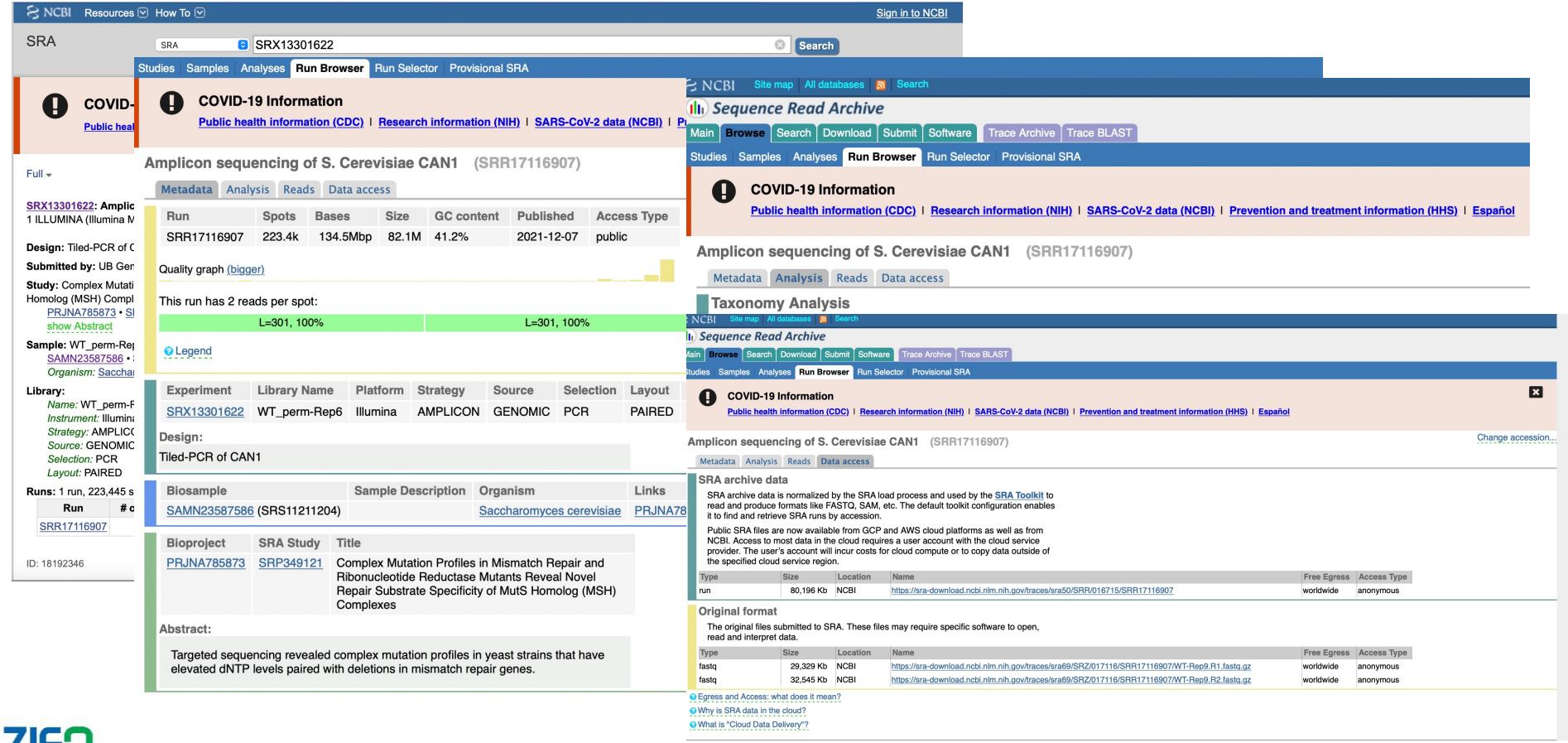
NCBI-SEQUENCE READ ARCHIVE (SRA)



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

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

SRA: WALK THROUGH



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 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, MacOSX, 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.)

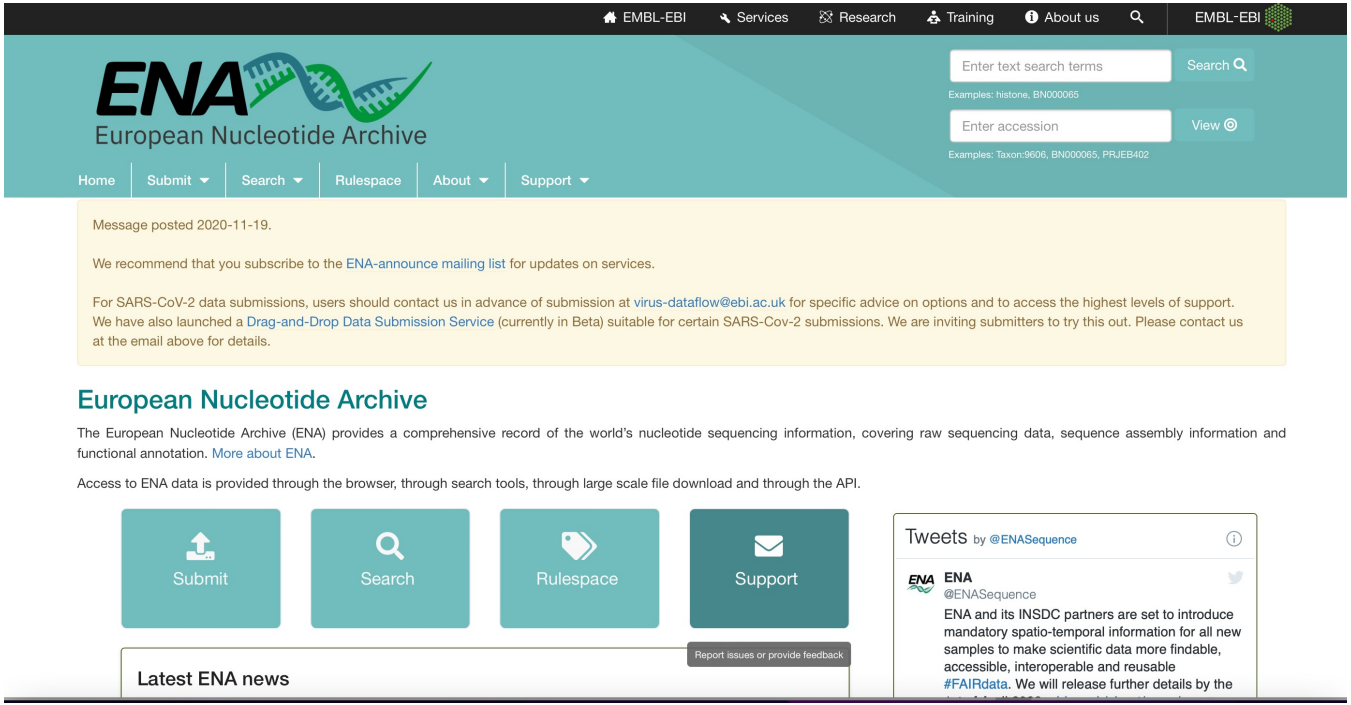
vdb-dump: Output the native VDB format of SRA data.

vdb-encrypt: Encrypt non-SRA dbGaP data ("phenotype data")

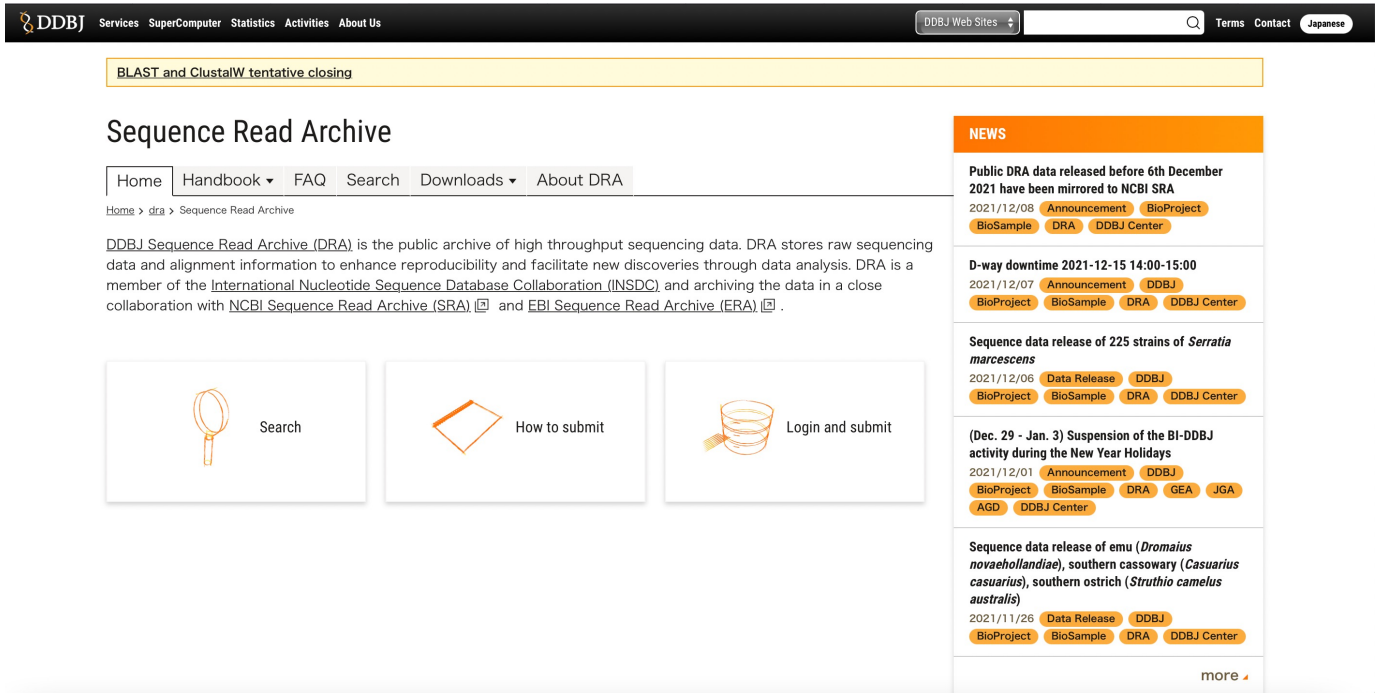
vdb-validate: Validate the integrity of downloaded SRA data

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

European Nucleotide Archive (ENA)



DDBJ SEQUENCE READ ARCHIVE (DRA)




That's all Folks!


ZIFO
RnD SOLUTIONS