NGS - quality control, alignment, visualisation

Quality control + database retrieval

Why Quality control?

1. How is the base quality?
2. What is the read length?
3. Are there adapters/barcodes in my sequences?
4. Are there overrepresented sequences?

Dedicated software

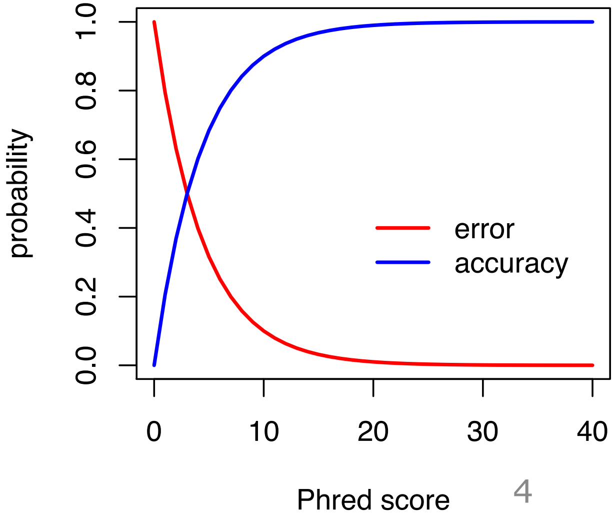
- Manufacturers’ software
- Illumina: fastQC
- ONT: pycoQC
- ONT + PacBio: NanoPlot
- ..

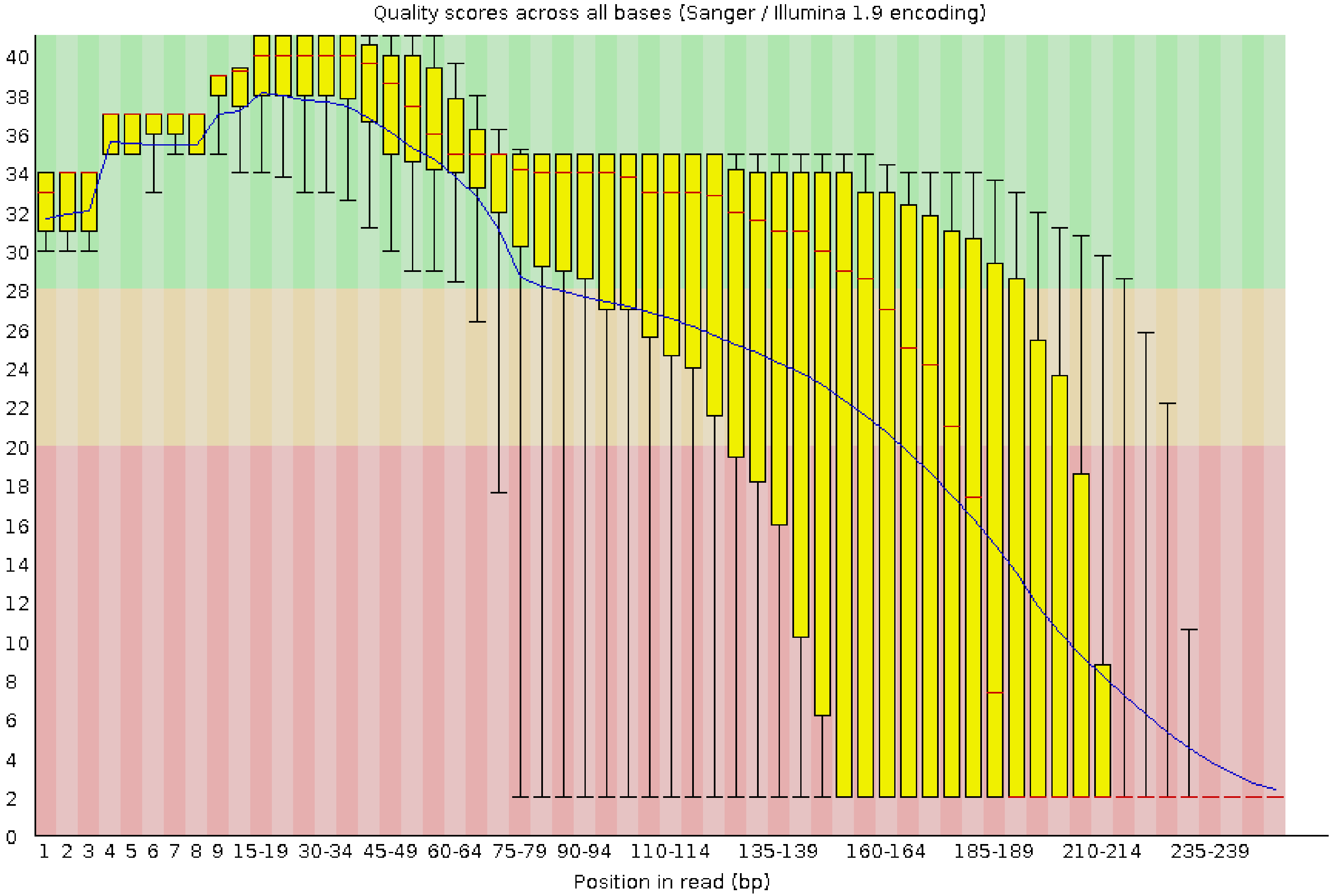
fastq

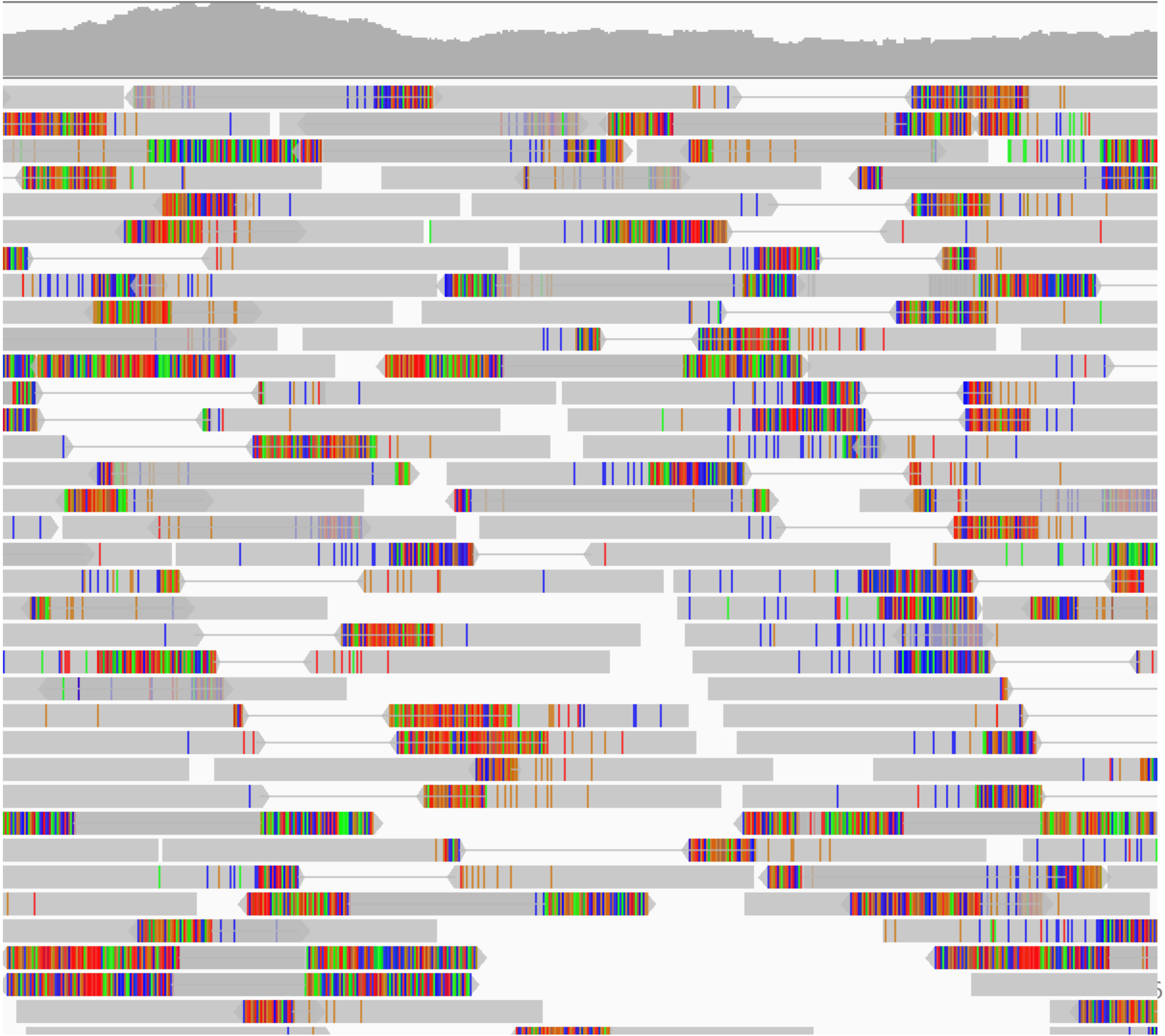
fasta + basequality (fasta + q = fastq)

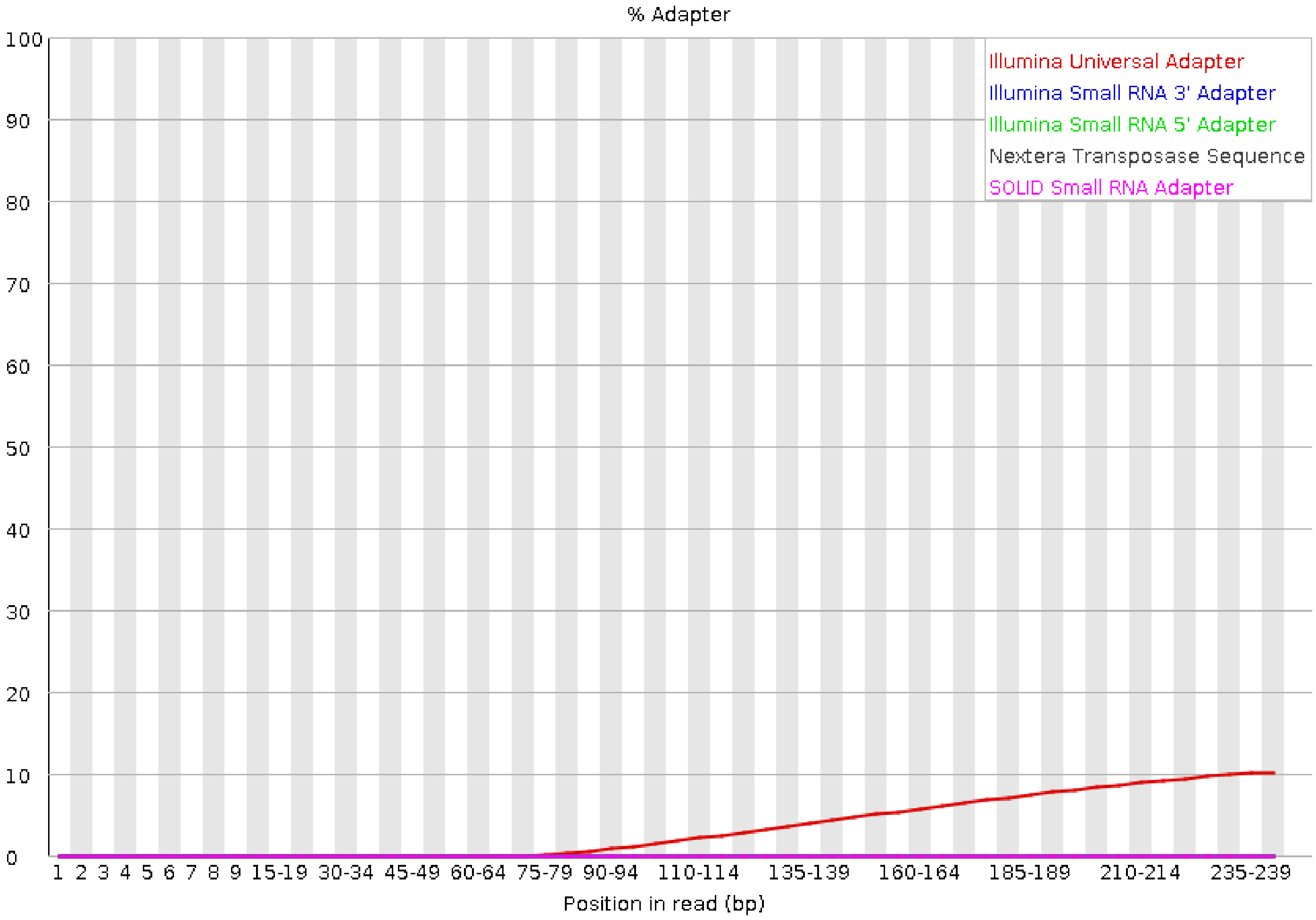
$$BASEQ = -10log_{10}\Pr\{base\ is\ wrong\}$$

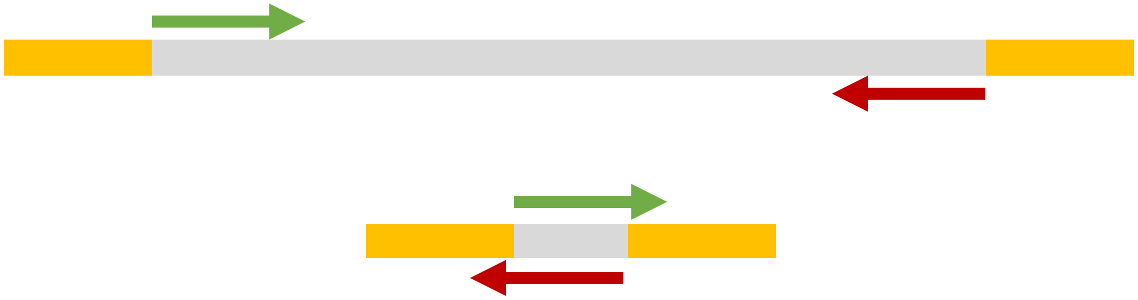
$$-10log_{10}(0.01) = 20$$
$$-10log_{10}(0.1) = 10$$
$$-10log_{10}(0.5) = 3$$




Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)

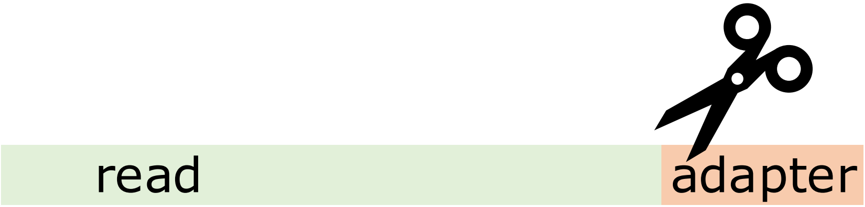



% Adapter
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
15-19
30-34
45-49
60-64
75-79
90-94
110-114
135-139
160-164
185-189
210-214
235-239
Position in read (bp)



Trimming

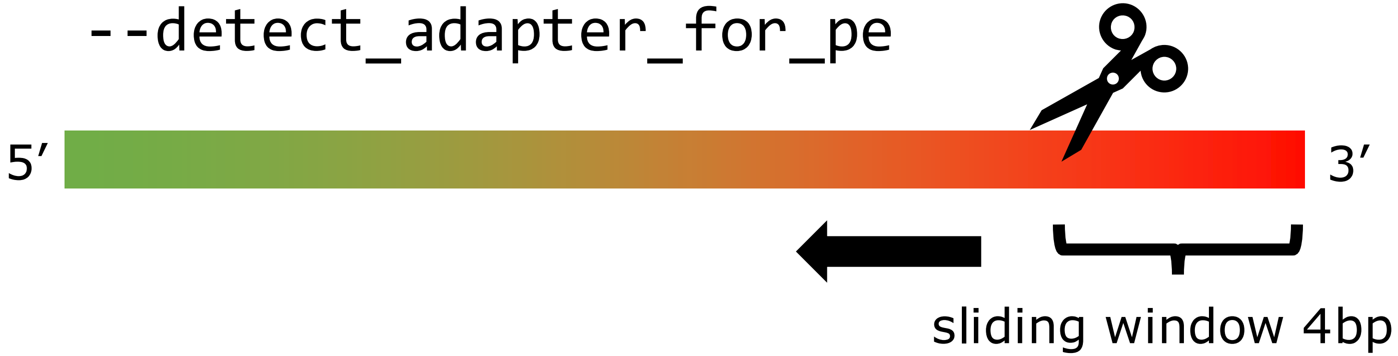
- Find and remove:
  - Regions or reads with low base quality
  - Adapter sequences
  - poly G sequences (e.g. with NovaSeq 6000)
- Software: `fastp` (or `cutadapt`, `trimmomatic`, `trim_galore`, `bbduk ..`)
- Articles on frequently occurring issues: sequencing.qcfail.com



Quality trimming with fastp

- Default:
  - Remove reads with >40% bases <Q15
  - Trim poly N (and poly G)
  - Autodetect adapters in R1, for both:

    `--detect_adapter_for_pe`

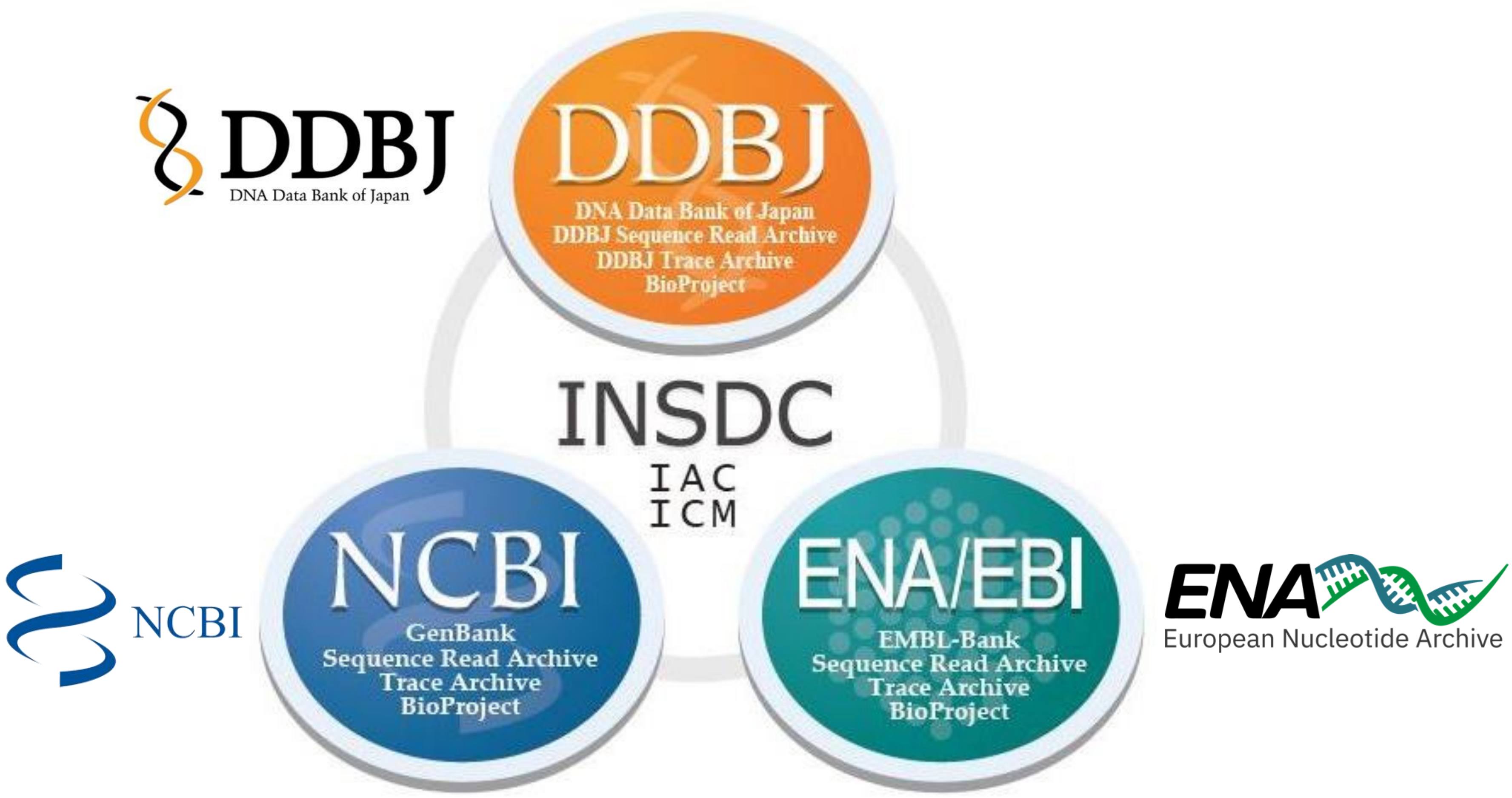


- ‘Classical’ trimming: sliding window
  - options `--cut_front` and `--cut_tail`

Nucleic acid sequence databases

- There are three major sites for finding information about nucleic acids (DNA and/or RNA sequences) on the Web, and all of them contain basically the same information
- The methods and databases that you will want to use will depend mainly on how much data you want and in what form.

3 synchronized databases



DDBJ
DNA Data Bank of Japan

DDBJ
DNA Data Bank of Japan
DDBJ Sequence Read Archive
DDBJ Trace Archive
BioProject

INSDC
IAC
ICM

NCBI

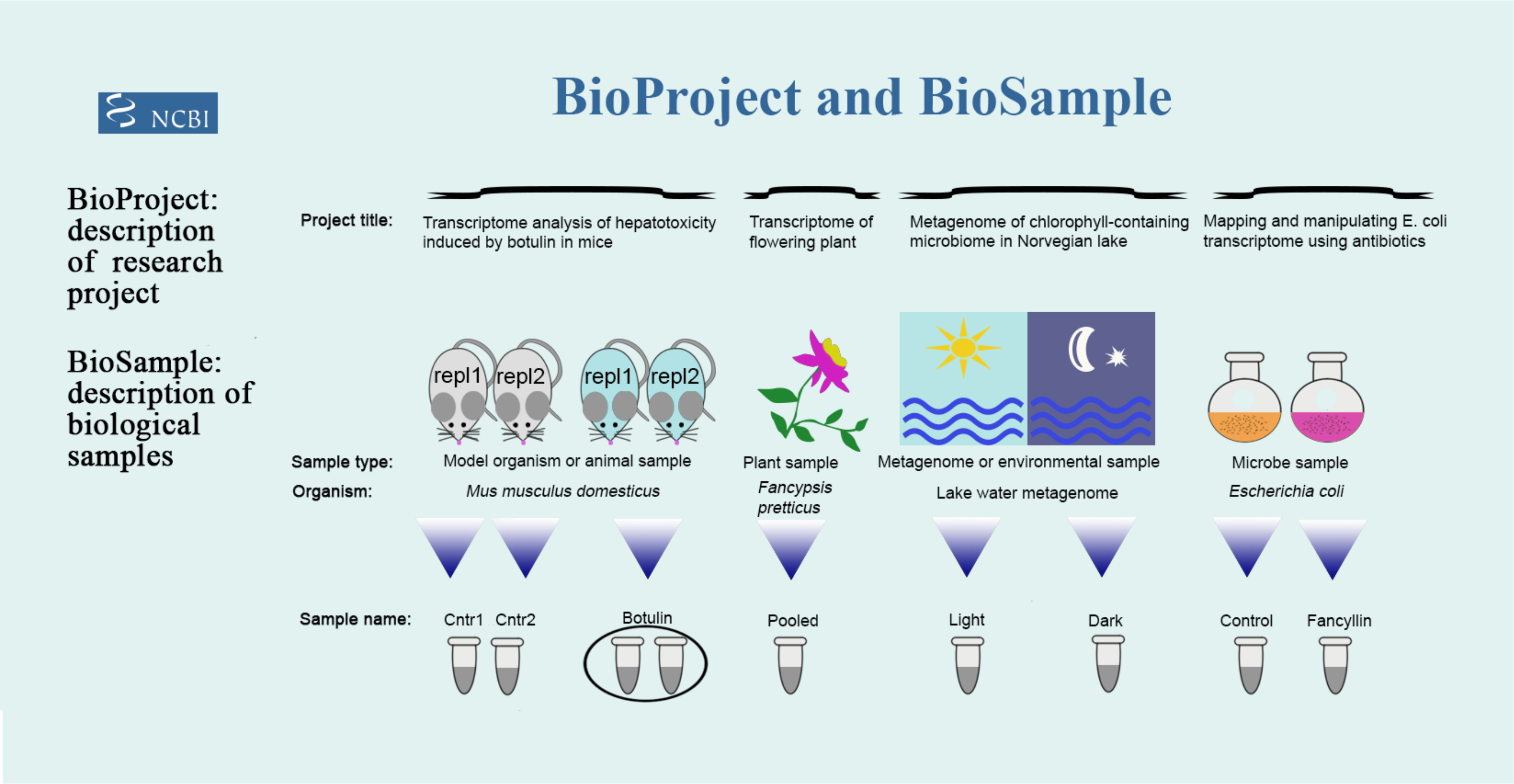
NCBI
GenBank
Sequence Read Archive
Trace Archive
BioProject

ENA/EBI
EMBL-Bank
Sequence Read Archive
Trace Archive
BioProject

ENA
European Nucleotide Archive

INSDC: **I**nternational **N**ucleotide **S**equence **D**atabase **C**ollaboration

BioProject and BioSample



BioProject and BioSample

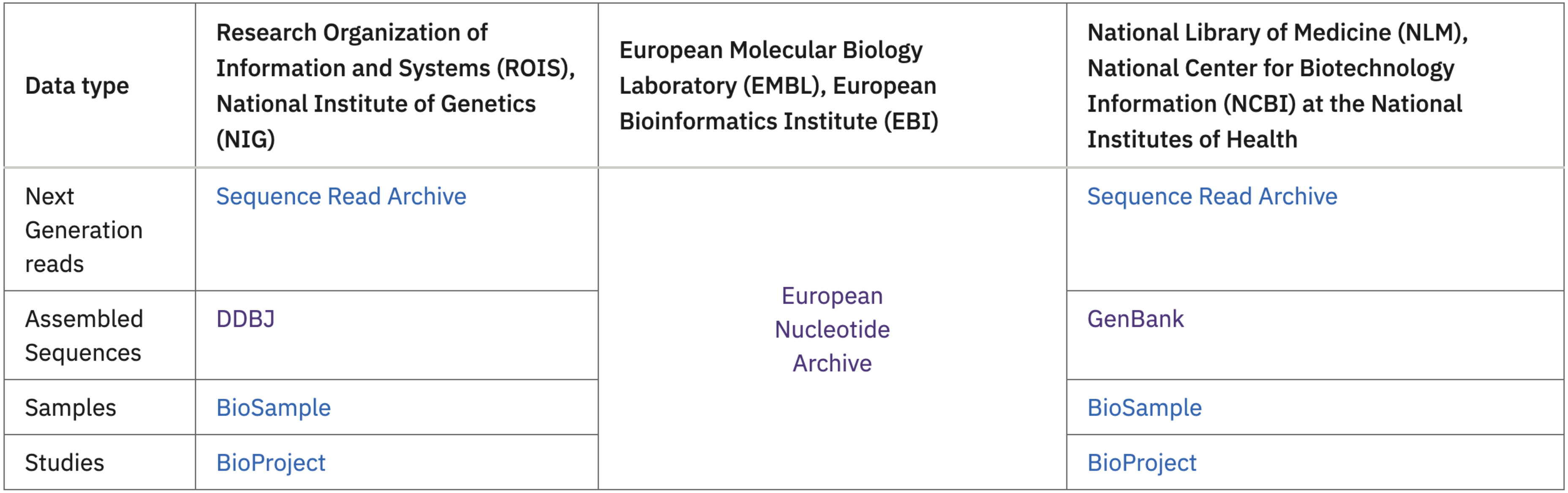
NCBI

BioProject: description of research project

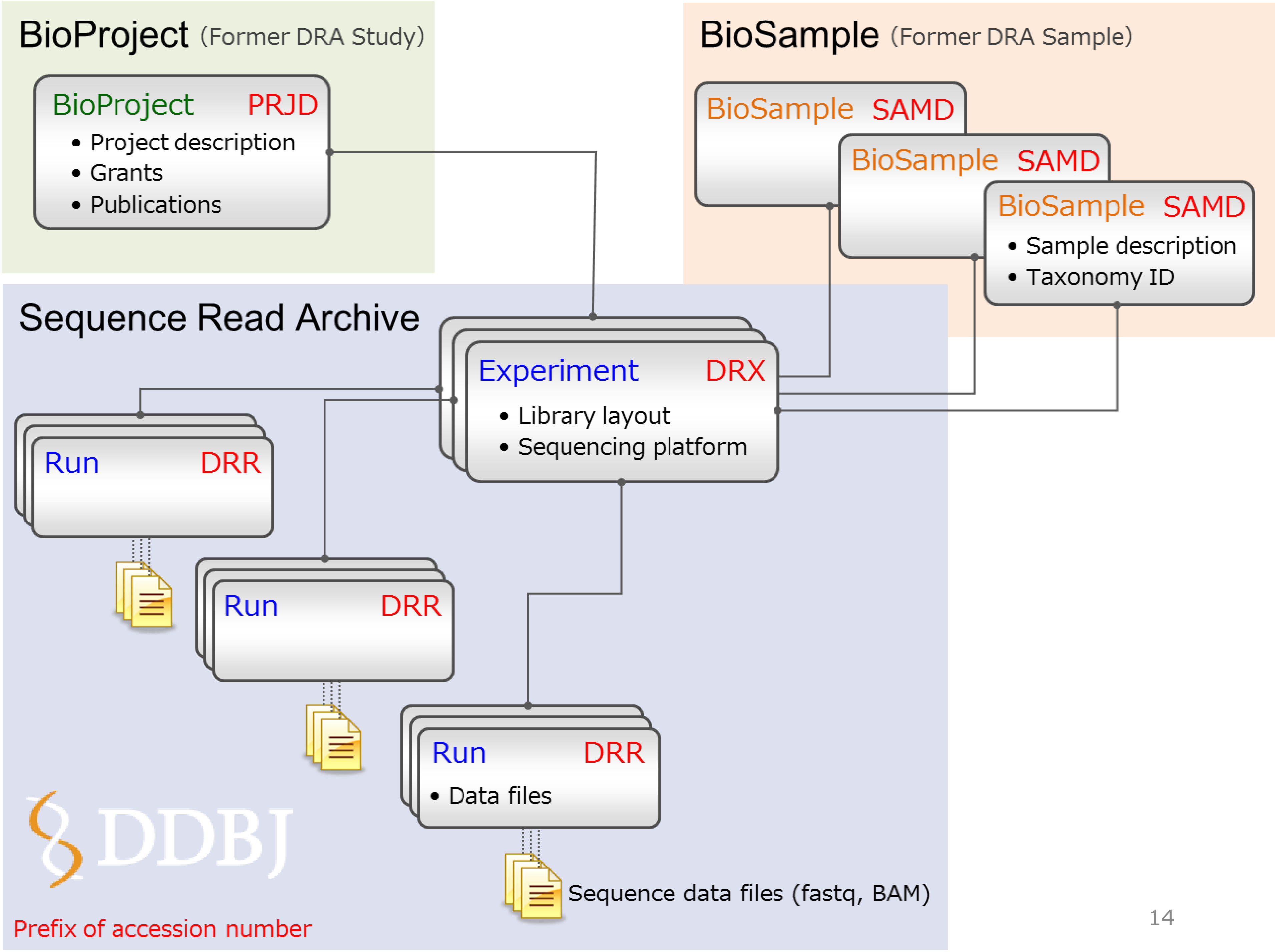
BioSample: description of biological samples

Project title:
Transcriptome analysis of hepatotoxicity induced by botulin in mice
Transcriptome of flowering plant
Metagenome of chlorophyll-containing microbiome in Norvegian lake
Mapping and manipulating E. coli transcriptome using antibiotics
repl1
repl2
repl1
repl2
Sample type:
Model organism or animal sample
Plant sample
Metagenome or environmental sample
Microbe sample
Organism:
Mus musculus domesticus
Fancypsis pretticus
Lake water metagenome
Escherichia coli
Sample name:
Cntr1
Cntr2
Botulin
Pooled
Light
Dark
Control
Fancyllin

The entries in the EMBL, GenBank and DDBJ databases are synchronized on a daily basis, and the accession numbers are managed in a consistent manner between these three centers.



Data type	Research Organization of Information and Systems (ROIS), National Institute of Genetics (NIG)	European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EBI)	National Library of Medicine (NLM), National Center for Biotechnology Information (NCBI) at the National Institutes of Health
Next Generation reads	Sequence Read Archive	European Nucleotide Archive	Sequence Read Archive
Assembled Sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject



BioProject (Former DRA Study)

BioProject PRJD
- Project description
- Grants
- Publications

BioSample (Former DRA Sample)

BioSample SAMD

BioSample SAMD

BioSample SAMD
- Sample description
- Taxonomy ID

Sequence Read Archive

Experiment DRX
- Library layout
- Sequencing platform

Run DRR

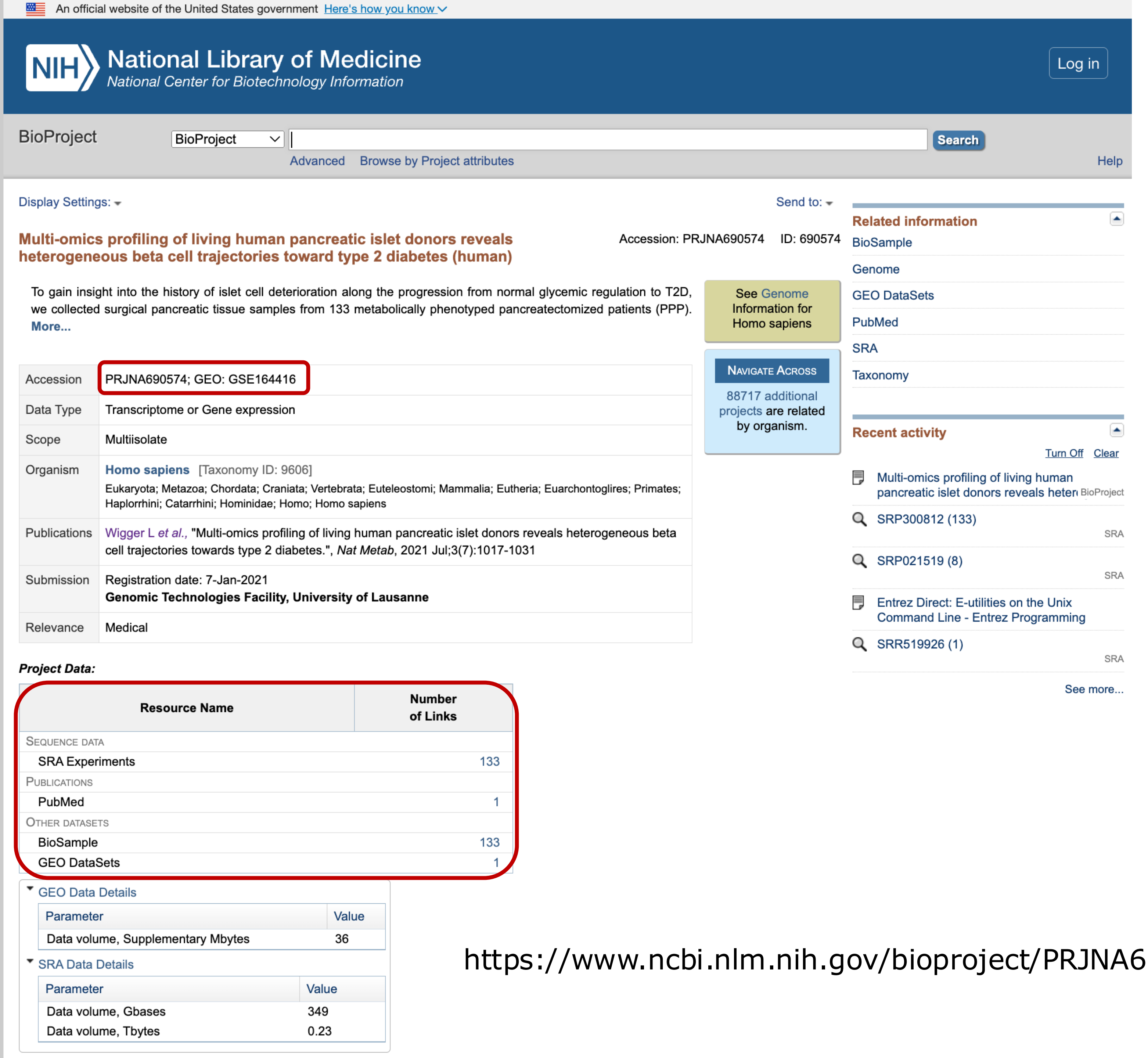
Run DRR

Run DRR
- Data files

Sequence data files (fastq, BAM)

DDBJ

Prefix of accession number



An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BioProject

BioProject | Search

Advanced Browse by Project attributes Help

Display Settings: Send to:

Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories toward type 2 diabetes (human)

Accession: PRJNA690574 ID: 690574

To gain insight into the history of islet cell deterioration along the progression from normal glycemic regulation to T2D, we collected surgical pancreatic tissue samples from 133 metabolically phenotyped pancreatectomized patients (PPP). **More...**

See Genome Information for Homo sapiens

NAVIGATE ACROSS
88717 additional projects are related by organism.

Accession	PRJNA690574; GEO: GSE164416
Data Type	Transcriptome or Gene expression
Scope	Multiisolate
Organism	**Homo sapiens** [Taxonomy ID: 9606] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens
Publications	Wigger L *et al.*, "Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories towards type 2 diabetes.", *Nat Metab*, 2021 Jul;3(7):1017-1031
Submission	Registration date: 7-Jan-2021 **Genomic Technologies Facility, University of Lausanne**
Relevance	Medical

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	133
PUBLICATIONS	
PubMed	1
OTHER DATASETS	
BioSample	133
GEO DataSets	1

GEO Data Details

Parameter	Value
Data volume, Supplementary Mbytes	36

SRA Data Details

Parameter	Value
Data volume, Gbases	349
Data volume, Tbytes	0.23

Related information

- BioSample
- Genome
- GEO DataSets
- PubMed
- SRA
- Taxonomy

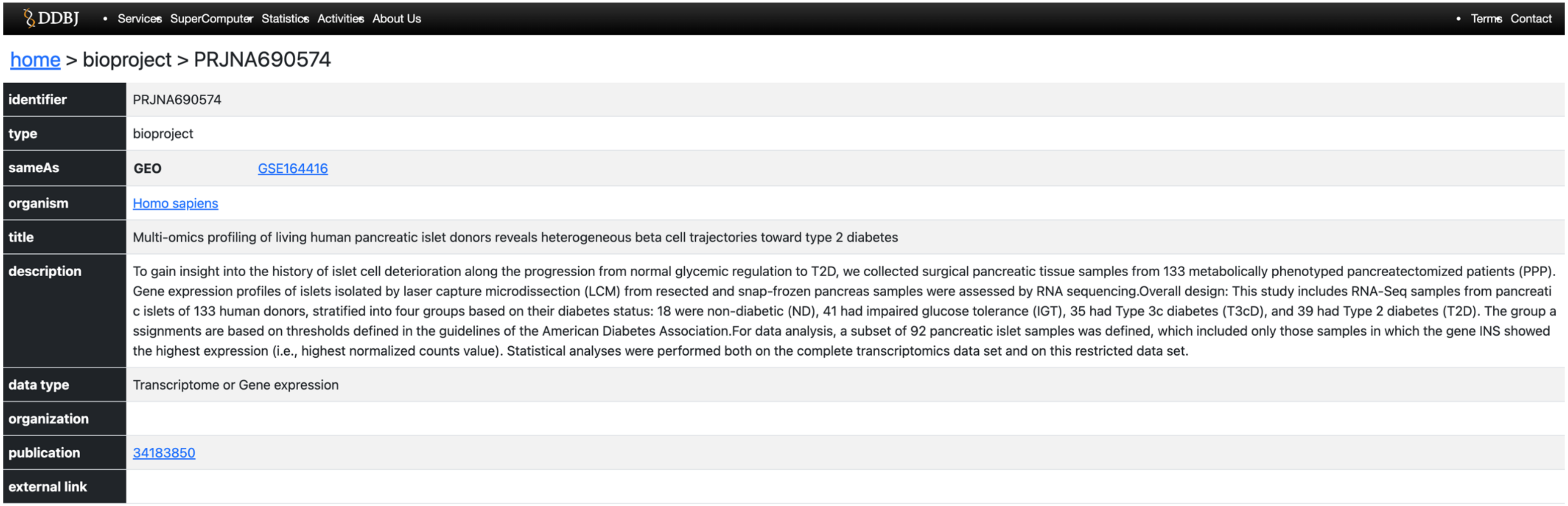
Recent activity

Turn Off Clear

- Multi-omics profiling of living human pancreatic islet donors reveals heter BioProject
- SRP300812 (133) SRA
- SRP021519 (8) SRA
- Entrez Direct: E-utilities on the Unix Command Line - Entrez Programming
- SRR519926 (1) SRA

See more...

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA690574



home > bioproject > PRJNA690574

identifier	PRJNA690574	
type	bioproject	
sameAs	**GEO**	GSE164416
organism	Homo sapiens	
title	Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories toward type 2 diabetes	
description	To gain insight into the history of islet cell deterioration along the progression from normal glycemic regulation to T2D, we collected surgical pancreatic tissue samples from 133 metabolically phenotyped pancreatectomized patients (PPP). Gene expression profiles of islets isolated by laser capture microdissection (LCM) from resected and snap-frozen pancreas samples were assessed by RNA sequencing.Overall design: This study includes RNA-Seq samples from pancreatic islets of 133 human donors, stratified into four groups based on their diabetes status: 18 were non-diabetic (ND), 41 had impaired glucose tolerance (IGT), 35 had Type 3c diabetes (T3cD), and 39 had Type 2 diabetes (T2D). The group assignments are based on thresholds defined in the guidelines of the American Diabetes Association.For data analysis, a subset of 92 pancreatic islet samples was defined, which included only those samples in which the gene INS showed the highest expression (i.e., highest normalized counts value). Statistical analyses were performed both on the complete transcriptomics data set and on this restricted data set.	
data type	Transcriptome or Gene expression	
organization		
publication	34183850	
external link		

For sensitive human data

EUROPEAN
GENOME-PHENOME
ARCHIVE

See also:
https://ega-archive.org/about/projects-and-funders/federated-ega/

“For each dataset that requires access control, there is a corresponding Data Access Committee (DAC) who determines access permissions. Data access is not the responsibility of the EGA. “

Command line tools

- Retrieve raw data: SRA-tools
  - `prefetch`
  - `fastq-dump`
- Retrieve sequences: Entrez Direct
  - `esearch`
  - `efetch`