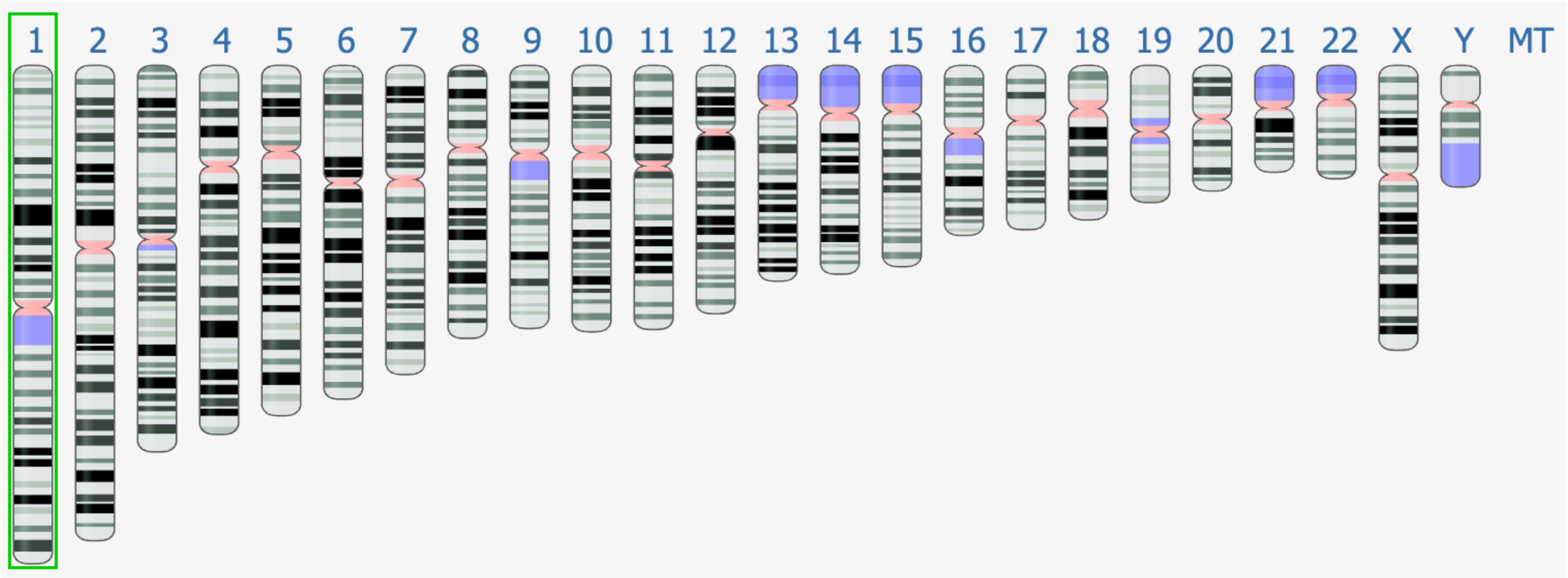


Genomics104

BIOL647
Digital Biology

Rodolfo Aramayo

Genomics104

NCBI Databases and Quality Control

A Brief Introduction To NCBI: Small Reads Archive (SRA)

SRA Read Archive Overview

Sequence Read Archive

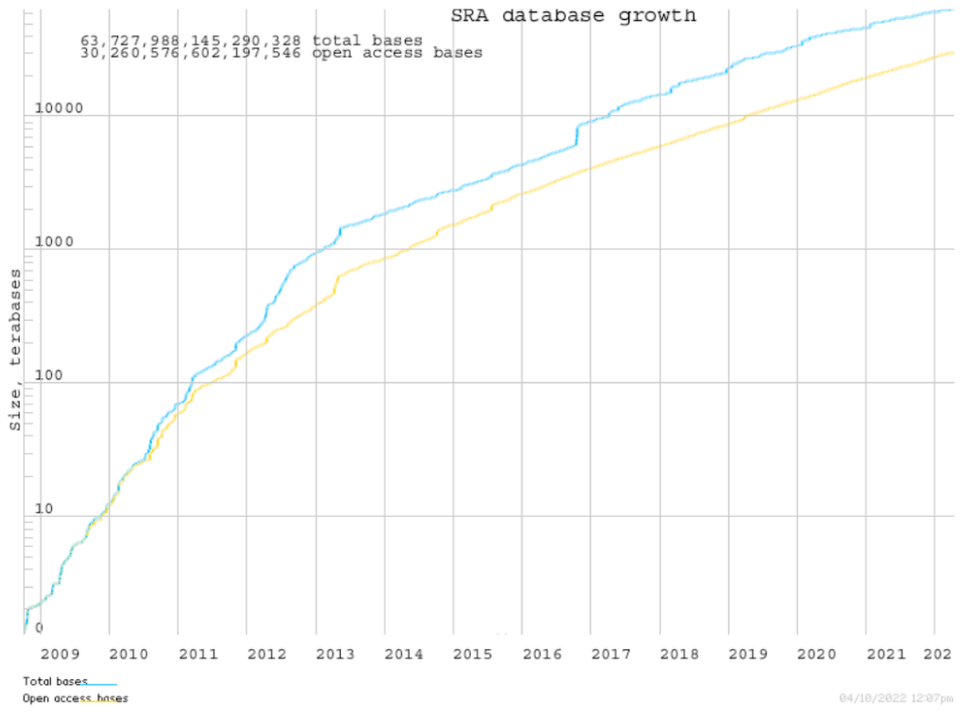
Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace BLAST

Overview

The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, 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)

Genomics104

NCBI Databases and Quality Control

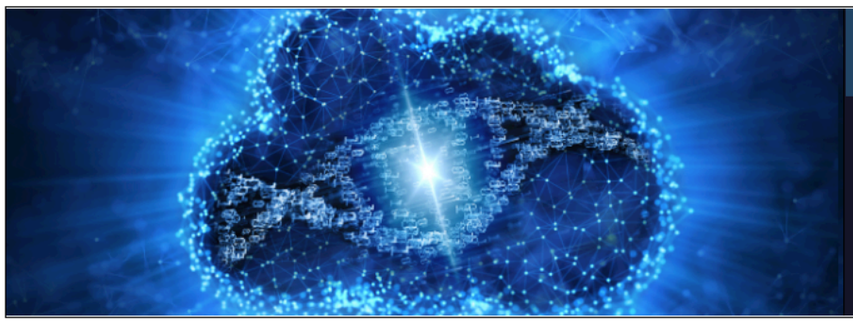
A Brief Introduction To NCBI: Small Reads Archive (SRA)

SRA Read Archive

SRA

SRA | Search

Advanced | Help



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- dbGaP Home
- BioProject
- BioSample

Genomics104

NCBI Databases and Quality Control

A Brief Introduction To NCBI: Small Reads Archive (SRA)

NCBI SRA Toolkit Download

NCBI SRA Toolkit

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

SRA Toolkit

Compiled binaries/install scripts of February 10, 2022, version 3.0.0:

- 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:

- HISAT2 version 2.2.1-ngs.3.0.0 - 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

- NCBI VDB Software Development Kit – February 10, 2022, version 3.0.0 release
- NCBI SRA Toolkit – February 10, 2022, version 3.0.0 release
- NCBI NGS Toolkit – February 10, 2022, version 3.0.0 release

File checksums

You may validate downloaded files with md5 checksums computed using **md5sum -b**

Genomics104

NCBI Databases and Quality Control

A Brief Introduction To NCBI: Small Reads Archive (SRA)

NCBI SRA Toolkit Documentation

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

Genomics104

NCBI Databases and Quality Control

A Brief Introduction To NCBI: Small Reads Archive (SRA)

SRA Working Examples

Tetrahymena

Database	Access		all
	public	controlled	
BioSample	794		794
BioProject	87		87
dbGaP			
GEO Datasets	530		530

Neurospora crassa

Database	Access		all
	public	controlled	
BioSample	5,420		5,420
BioProject	2,770		2,770
dbGaP			
GEO Datasets	1,472		1,472

Saccharomyces cerevisiae

Database	Access		all
	public	controlled	
BioSample	128,764		128,764
BioProject	3,143		3,143
dbGaP		1	1
GEO Datasets	57,839		57,839

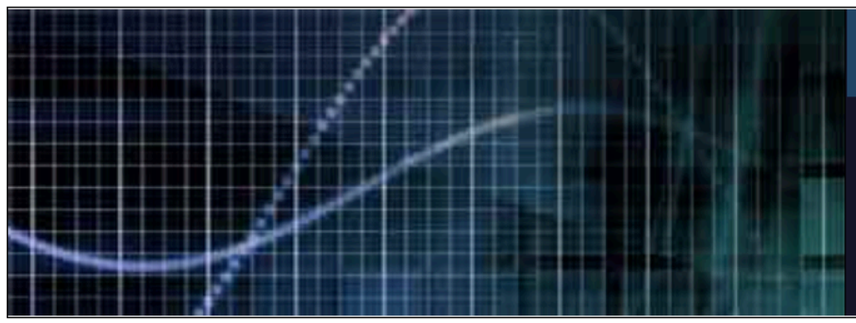
Homo sapiens

Database	Access		all
	public	controlled	
BioSample	7,275,933	727,750	8,003,683
BioProject	40,765	969	41,734
dbGaP		8	8
GEO Datasets	1,019,877		1,019,877

Genomics104

NCBI Databases and Quality Control

Introduction To BioProjects



BioProject

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

Using BioProject

- Frequently Asked Questions
- BioProject Help
- BioProject Overview
- Submission

Browse BioProject

- By Project attributes UPDATED
- Download (FTP)

Large Initiatives

- 1000 Genomes
- ENCODE
- HMP

NCBI Resources

- BioSample
- dbGaP
- Genome

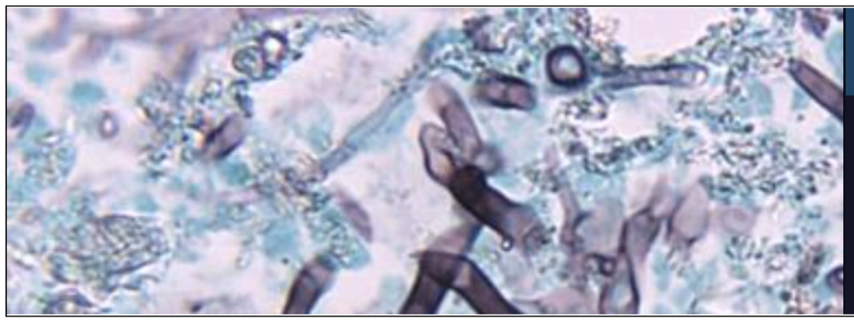
External Resources

- Genome projects at DOE
- Genome News Network
- GOLD - Genome On Line Database

Genomics104

NCBI Databases and Quality Control

Introduction To BioSamples



BioSample

The BioSample database contains descriptions of biological source materials used in experimental assays.

Using BioSample	Sources	Authenticated Cell Line
BioSample Overview	GenBank	Background, Search and Submit
BioSample Documentation	SRA	Browse Human Cell Line STR Profiles
Submission FAQ	Coriell	Browse Known Misidentified Cell Lines
Search Help	ATCC	
Submit	ICLAC	

Example Searches	
bacteria of genus Shigella for which SRA data is available	shigella[organism] AND biosample sra[filter]
MIGS/MIMS/MIMARKS.water-complaint samples released in first quarter of 2013	package migs/mims/mimarks water[Properties] AND 2013/1:2013/3[Publication date]
mouse samples for which strain and age information is available	(strain[Attribute Name] AND age[Attribute Name]) AND Mus musculus[organism]
fibroblast cell samples	cell type fibroblast[Attribute]

Genomics104

NCBI Databases and Quality Control

Introduction To SRA Explorer

SRA Explorer

This tool aims to make datasets within the Sequence Read Archive more accessible.

Search for: Enter accession number

Max Results 100 **Start At Record** 0

Need inspiration? Try `GSE30567`, `SRP043510`, `PRJEB8073`, `ERP009109` or `human liver miRNA`.

SRA-Explorer was written by Phil Ewels. Source code is available under a GNU GPLv3 licence at https://github.com/ewels/sra-explorer.

Here a lot? It might be worth taking a look at some alternative tools..

Genomics104

NCBI Databases and Quality Control

A Database Example

Pervasive, coordinated protein level changes driven by transcript isoform switching during meiosis (baker's yeast)

Accession: PRJNA428526 ID: 428526

To better understand the gene regulatory mechanisms that program developmental processes, we carried out simultaneous, genome-wide measurements of mRNA, translation and protein through meiotic differentiation in budding yeast. **More...**

See Genome Information for Saccharomyces cerevisiae

Accession	PRJNA428526; GEO: GSE108778
Scope	Multiisolate
Organism	**Saccharomyces cerevisiae** [Taxonomy ID: 4932] Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces; Saccharomyces cerevisiae
Publications	Cheng Z *et al.*, "Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis.", *Cell*, 2018 Feb 22;172(5):910-923.e16
Submission	Registration date: 4-Jan-2018 **UC Berkeley**
Relevance	Model Organism

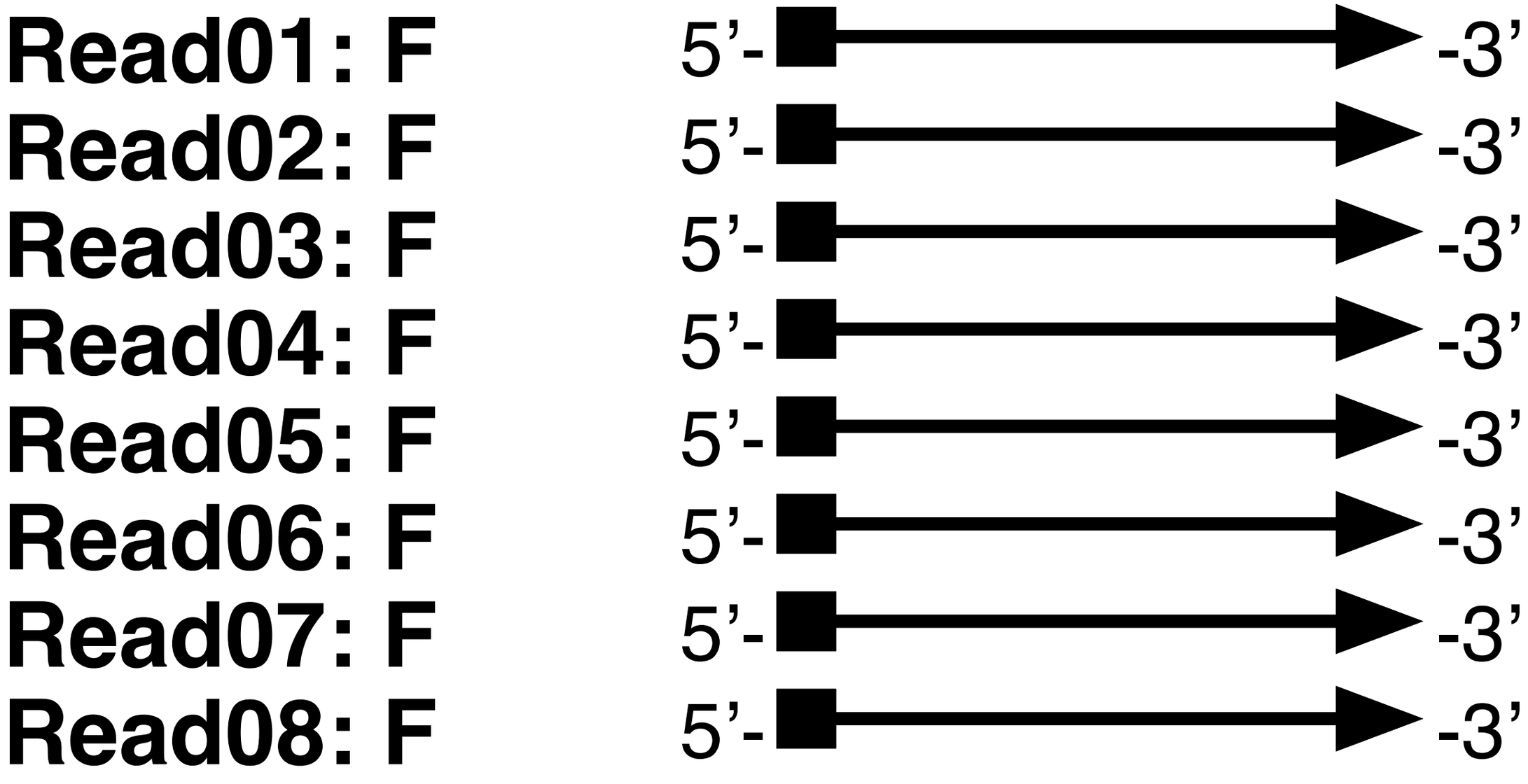
NAVIGATE ACROSS

4217 additional projects are related by organism.

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Single or Unpaired Reads:

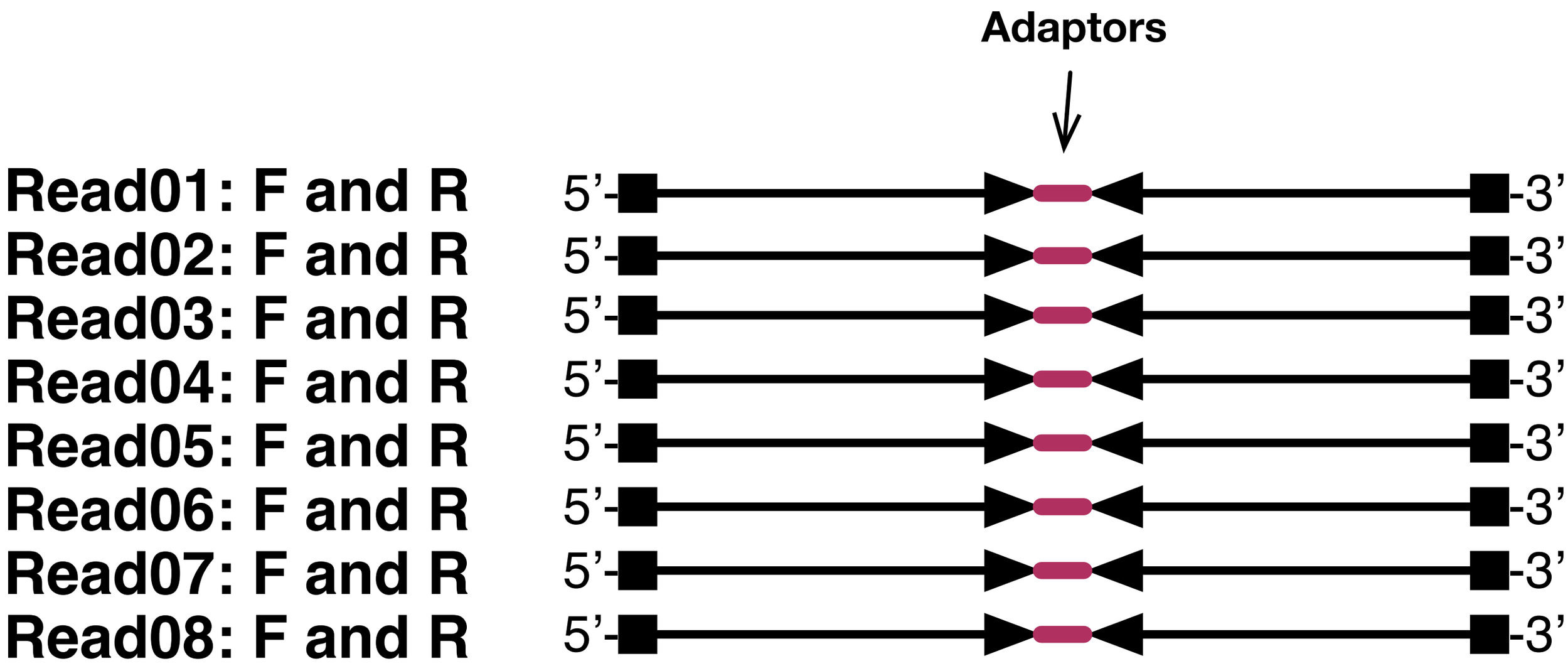


Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Paired Reads:

Unsplit Files:

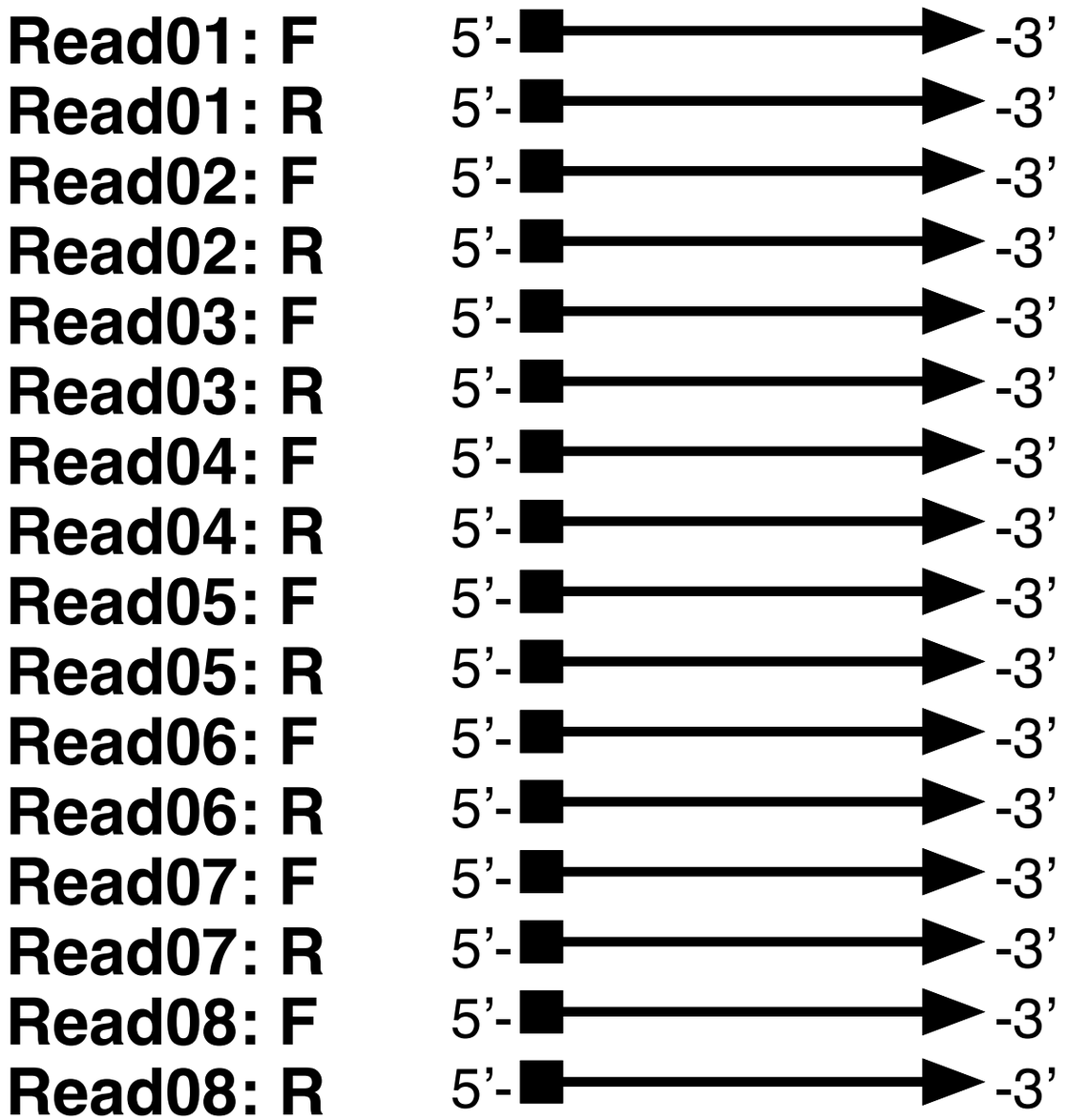


Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Paired Reads:

Single Interleaved Files:



Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Paired Reads:

Split Files:

File01:

Read01: F 5’- ➡ -3’
Read02: F 5’- ➡ -3’
Read03: F 5’- ➡ -3’
Read04: F 5’- ➡ -3’
Read05: F 5’- ➡ -3’
Read06: F 5’- ➡ -3’
Read07: F 5’- ➡ -3’
Read08: F 5’- ➡ -3’

File02:

Read01: R 5’- ➡ -3’
Read02: R 5’- ➡ -3’
Read03: R 5’- ➡ -3’
Read04: R 5’- ➡ -3’
Read05: R 5’- ➡ -3’
Read06: R 5’- ➡ -3’
Read07: R 5’- ➡ -3’
Read08: R 5’- ➡ -3’

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Paired Reads:

Split Files After Quality Control:

Read01: F	5’- ■ ⟶ -3’	**Ok**
Read01: R	5’- ■ ⟶ -3’	**Ok**
Read02: F	5’- ■ ⟶ -3’	**Ok**
Read02: R	5’- ■ ⟶ -3’	**Fails**
Read03: F	5’- ■ ⟶ -3’	**Fails**
Read03: R	5’- ■ ⟶ -3’	**Ok**
Read04: F	5’- ■ ⟶ -3’	**Fails**
Read04: R	5’- ■ ⟶ -3’	**Fails**
Read05: F	5’- ■ ⟶ -3’	**Ok**
Read05: R	5’- ■ ⟶ -3’	**Ok**
Read06: F	5’- ■ ⟶ -3’	**Ok**
Read06: R	5’- ■ ⟶ -3’	**Ok**
Read07: F	5’- ■ ⟶ -3’	**Fails**
Read07: R	5’- ■ ⟶ -3’	**Ok**
Read08: F	5’- ■ ⟶ -3’	**Ok**
Read08: R	5’- ■ ⟶ -3’	**Ok**

Which Results in:

File01 Paired:

Read01: F	5’- ■ ⟶ -3’
Read05: F	5’- ■ ⟶ -3’
Read06: F	5’- ■ ⟶ -3’
Read08: F	5’- ■ ⟶ -3’

File02 Paired:

Read01: R	5’- ■ ⟶ -3’
Read05: R	5’- ■ ⟶ -3’
Read06: R	5’- ■ ⟶ -3’
Read08: R	5’- ■ ⟶ -3’

File03 Singletons:

Read02: F	5’- ■ ⟶ -3’
Read03: R	5’- ■ ⟶ -3’
Read07: R	5’- ■ ⟶ -3’

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

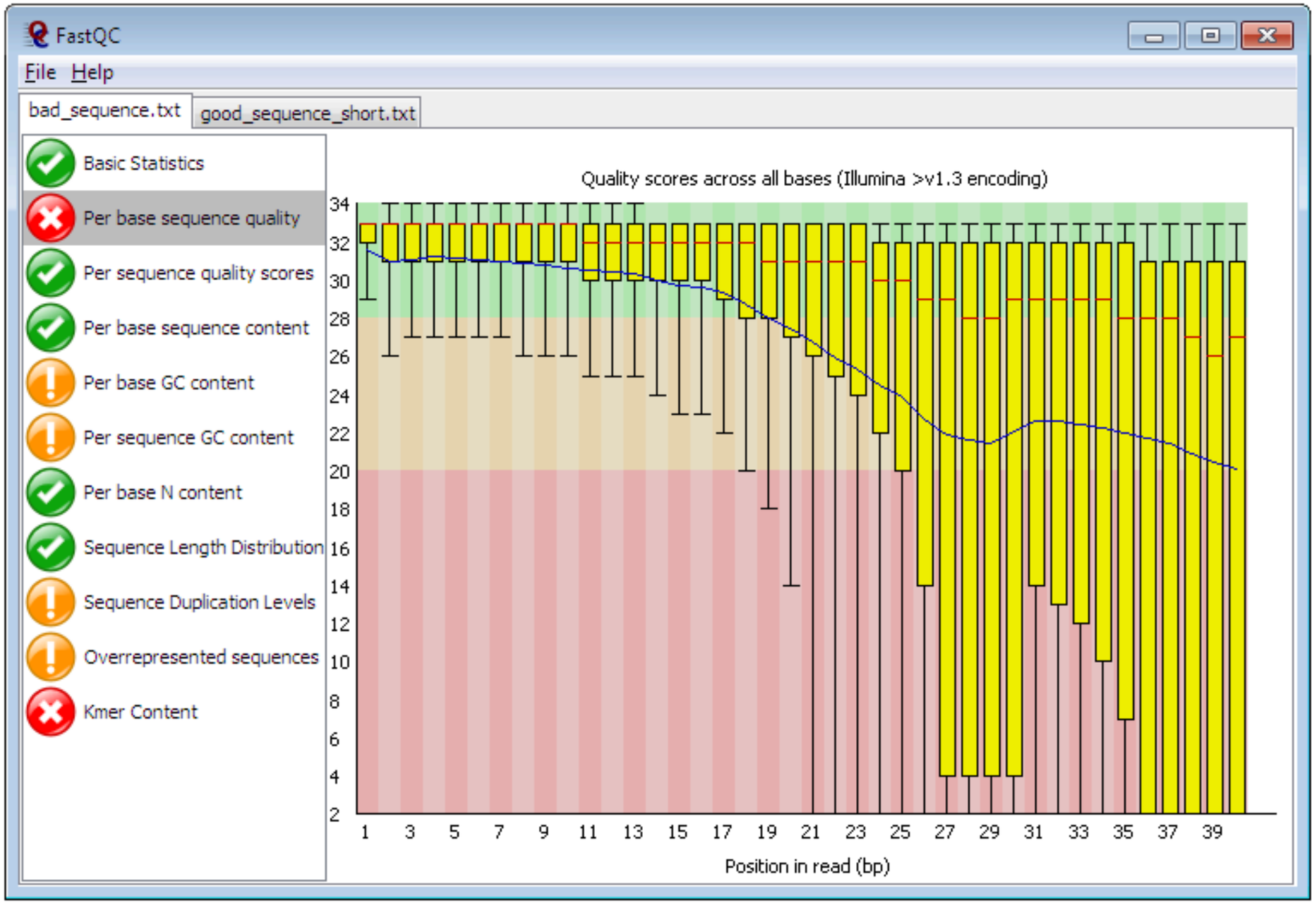
- *Download the SRA file*
- *Process Downloaded files with SRA Toolkit using fastq-dump either:*
  - *Preserving reads in one file or*
  - *Splitting reads into two files*
- *Upload the files to Galaxy and/or to your working directory*
- *Get Reads Statistics*
  - *Count the number of reads*
  - *Make sure length is right*

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

- ***Run FastQC***



Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

- *Compute Quality Statistics*
- *Determine the Reads Size Distribution*
- *Split the reads, if necessary*
  - *FASTQ splitter on joined paired end reads*
  - *By barcodes: Barcode Splitter*
- *Trim Non-Native Bases, if needed*
- *Remove Ns bases*
    - *By Removing reads carrying Ns*
    - *By trimming reads and then determining read size*
- *Remove artifacts: Remove sequencing artifacts*

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

- ***Remove adaptors***
  - ***Using Cutadapt***

Cutadapt

Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.

Cleaning your data in this way is often required: Reads from small-RNA sequencing contain the 3' sequencing adapter because the read is longer than the molecule that is sequenced. Amplicon reads start with a primer sequence. Poly-A tails are useful for pulling out RNA from your sample, but often you don't want them to be in your reads.

Cutadapt helps with these trimming tasks by finding the adapter or primer sequences in an error-tolerant way. It can also modify and filter single-end and paired-end reads in various ways. Adapter sequences can contain IUPAC wildcard characters. Cutadapt can also demultiplex your reads.

Cutadapt is available under the terms of the MIT license.

Cutadapt development was started at TU Dortmund University in the group of Prof. Dr. Sven Rahmann. It is currently being developed within NBIS (National Bioinformatics Infrastructure Sweden).

If you use Cutadapt, please cite DOI:10.14806/ej.17.1.200 .

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

- ***Remove adaptors***
  - ***Using Trimmomatic***

Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Downloading Trimmomatic

starting on version 0.40 we also offer a github page (as well as older versions)

Version 0.39: binary, source and manual

Version 0.36: binary and source

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

- ***Filter reads by Quality Score***
- ***Prepare reads for assembly:***
  - ***FASTQ interlacer (Galaxy) or PEAR on paired end reads***
  - ***Run KmerGenie k-mer histograms analysis***

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

- ***Filter reads by Quality Score***
- ***Prepare reads for assembly:***
  - ***FASTQ interlacer (Galaxy) or PEAR on paired end reads***
  - ***Run KmerGenie k-mer histograms analysis***

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Logical Steps

- ***Run MultiQC***

MultiQC

Aggregate results from bioinformatics analyses across many samples into a single report

MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.

English Español

Introduction to MultiQC *(1:19)*

Installing MultiQC *(4:33)*

Running MultiQC *(5:21)*

Using MultiQC Reports *(6:06)*

Citations 2.1k

GitHub

Python Package Index

Documentation

114 supported tools

Publication / Citation

Get help on Gitter

Quick Install

```
pip install multiqc    # Install
multiqc .              # Run
```

pip conda manual

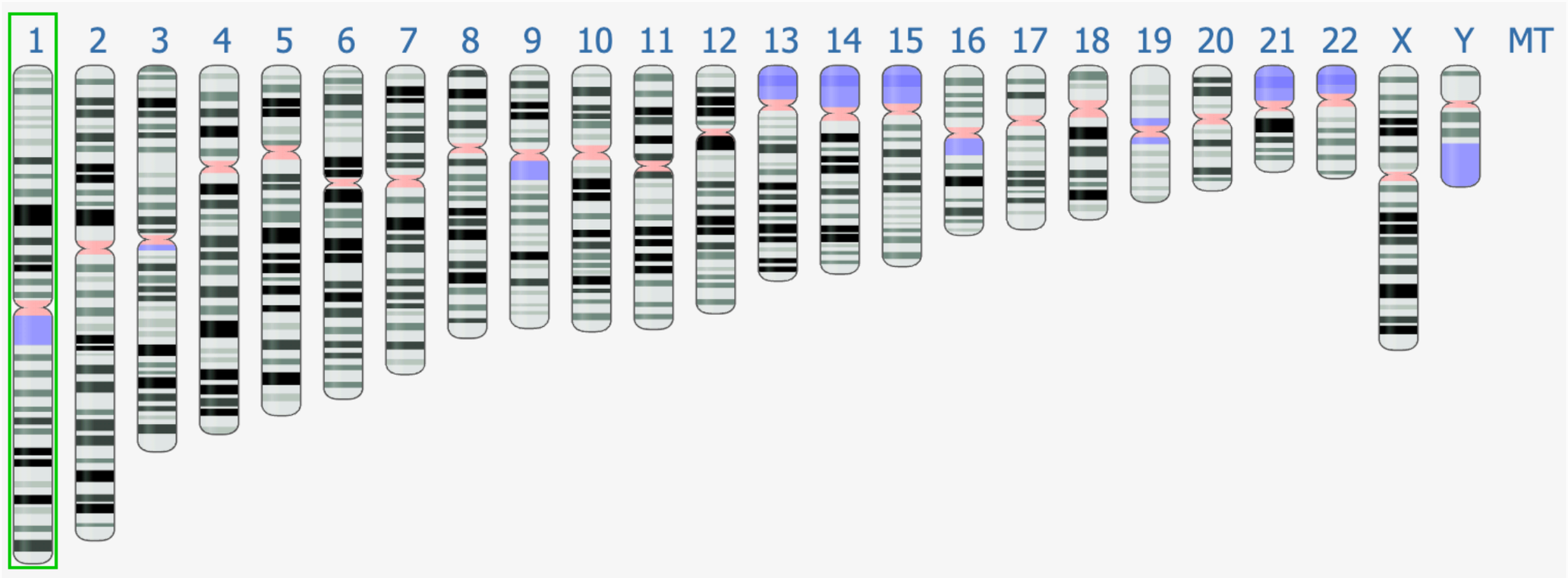
Need a little more help? See the full installation instructions.

Genomics104

Small Reads Quality Control (QC) Files Fundamentals

Small Reads Quality Control (QC) Practical Examples

Dataset	SRA Accession Number
1	SRX040240
2	SRX691982
3	SRX099287
4	SRX099497
5	SRX152734



Genomics104

BIOL647
Digital Biology

Rodolfo Aramayo