



**Phil Ewels**  
phil@seqera.io

# Adventures in Open Science

NextGenBUG 2022



# Open Science



**Open Science**

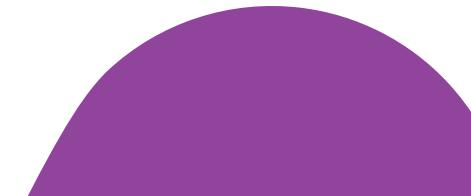
**Open Source**

**Open Data**

**Open Contributions**

**Open Community**

**Open Core**



# Open Source



# Open Source

Labrador  Search  Create New Project Log In / Register

## Filters

TEXT FILTER

PROJECT NAME [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#) [0-9](#)

SPECIES [Mus musculus](#) [Homo sapiens](#) [Harpegnathos saltator](#) [Mouse](#) [Saccharomyces cerevisiae](#) [Camponotus floridanus](#) [Polistes canadensis](#) [Arabidopsis thaliana](#) [Apis mellifera](#)

## Labrador Dataset Browser

A database of datasets processed by the BI Bioinformatics group.

You can use labrador to find and download processed data or request new datasets. Projects are annotated with how they were processed.

Key:  Processing Complete  Currently Processing  Not Started  Directory not found

Name	Datasets	Species	Cell Types	Data Types
Abad_2013	14	Mus musculus	ES cells, iPS cells	RNA-Seq
Adachi_2013	6	Mus musculus	ES cells	ChIP-Seq
Aiba_2009	88	Mus musculus	embryonic carcinoma cell, ES cell, TS cell, Neural stem cell, Placenta, SNL-STO cell line, MEF, Embryonic germ cells, iPS-	RNA ChIP Microarray

Labrador Dataset Browser



Video Tutorial

# Open Source

2010 First release

2012 Moved to GitHub

2015 Stopped active development

Labrador  Search Create New Project Log In / Register

**Filters**

TEXT FILTER

PROJECT NAME

A B C D E F G H I J K L M N  
O P Q R S T U V W X Y Z 0-9

SPECIES

Mus musculus  
Homo sapiens  
Harpegnathos saltator  
Mouse  
Saccharomyces cerevisiae  
Camponotus floridanus  
Polistes canadensis  
Arabidopsis thaliana  
Apis mellifera  
Bombyx mori  
Drosophila melanogaster  
Nasonia vitripennis

DATA TYPES

ChIP-Seq  
RNA-Seq  
DNA-Seq

**Labrador Dataset Browser** A database of datasets processed by the BI Bioinformatics group. You can use labrador to find and download processed data or request new datasets. Projects are annotated with how they were processed.

You can use the table below to browse the projects and datasets. You can filter the visible data using the tools on the left. If you're looking for something really specific, try the search bar at the top of the page.

Key:  Processing Complete  Currently Processing  Not Started  Directory not found

Name	Datasets	Species	Cell Types	Data Types
Abad_2013	14	Mus musculus	ES cells, iPS cells	RNA-Seq
Adachi_2013	6	Mus musculus	ES cells	ChIP-Seq
Aiba_2009	88	Mus musculus	embryonic carcinomal cell, ES cell, TS cell, Neural stem cell, Placenta, SNL-STO cell line, 3T3 cell line, MEF, Embryonic germ cells, iPS-MEF	RNA ChIP Microarray
Ang_2011	2	Mus musculus	ES cells, ES Cell	ChIP-Seq
Aravin_2008	9	Mus musculus	10 dpp Dnmt3L heterozygotes testes, 10 dpp Dnmt3L KO mice testes, 16.5 dpc embryos testes, 4-6 week ovaries, 2 day testes, 10 dpp embryos testes	Small RNA-Seq
Asp_2011	8	Mus musculus	C2C12 myoblasts, C2C12 myotubes	ChIP-Seq
Auerbach_2009	1	Homo sapiens	HeLa	ChIP-Seq
Boulker_2010	14	Mus musculus	ES Cells	ChIP-Seq

Labrador Dataset Browser Video Tutorial

# Open Source

2010 First release

2012 Moved to GitHub

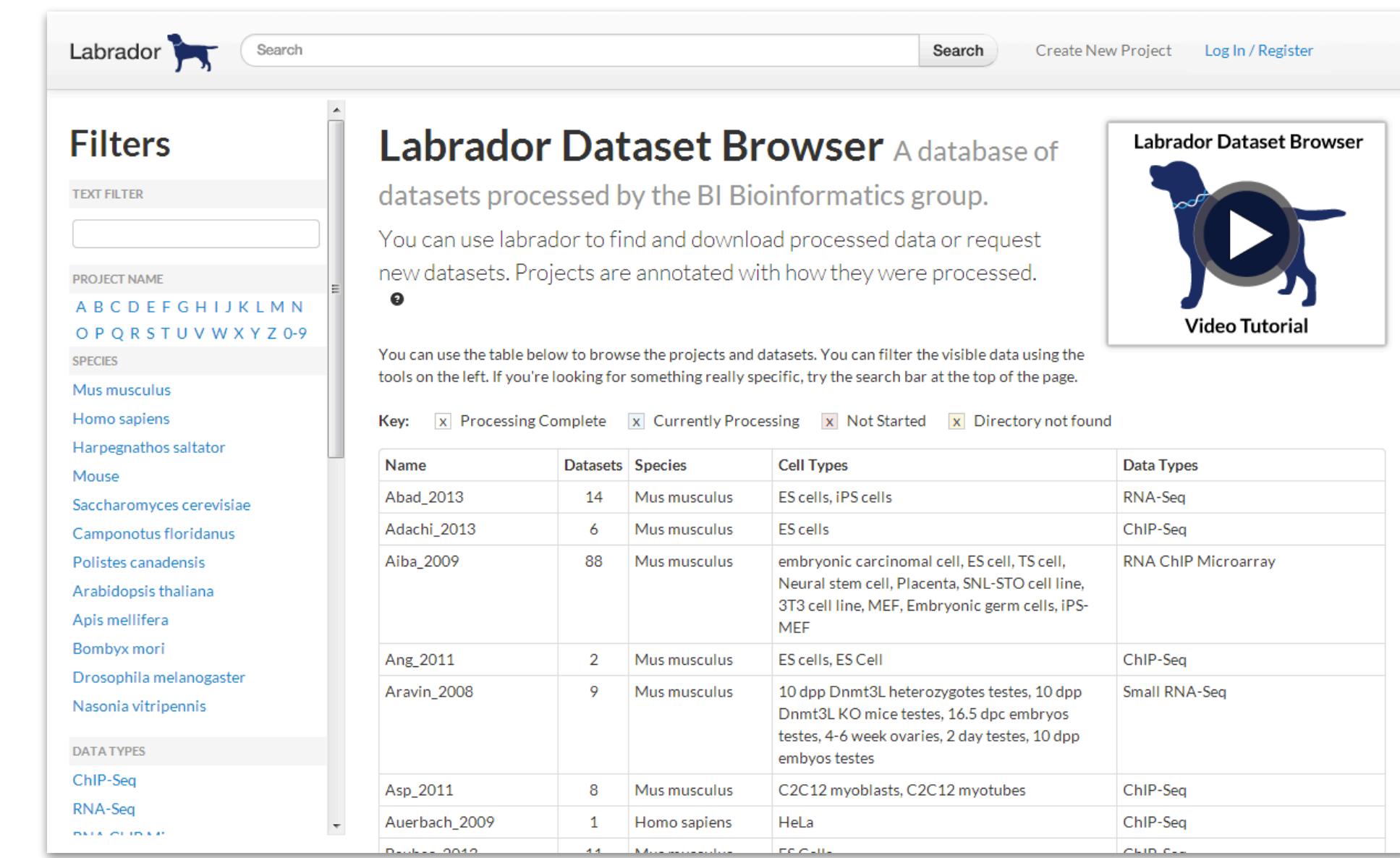
2015 Stopped active development

2017 Contributions from Russell Hamilton

2018 Fixes from Simon Andrews

2021 Updates from Steven Wingett

2022 Security patches from Altos Labs



The screenshot shows the Labrador Dataset Browser interface. At the top, there is a navigation bar with a dog icon, a search bar, and links for 'Create New Project' and 'Log In / Register'. Below the navigation bar, there is a section titled 'Filters' with dropdown menus for 'TEXT FILTER', 'PROJECT NAME' (with letters A-Z and 0-9), 'SPECIES' (listing Mus musculus, Homo sapiens, Harpegnathos saltator, Mouse, Saccharomyces cerevisiae, Camponotus floridanus, Polistes canadensis, Arabidopsis thaliana, Apis mellifera, Bombyx mori, Drosophila melanogaster, and Nasonia vitripennis), and 'DATA TYPES' (listing ChIP-Seq, RNA-Seq, and RNA-Seq). To the right of the filters, there is a main content area titled 'Labrador Dataset Browser' which describes it as a database of datasets processed by the BI Bioinformatics group. It says you can use the browser to find and download processed data or request new datasets. Projects are annotated with how they were processed. Below this text is a table listing various projects with their details:

Name	Datasets	Species	Cell Types	Data Types
Abad_2013	14	Mus musculus	ES cells, iPS cells	RNA-Seq
Adachi_2013	6	Mus musculus	ES cells	ChIP-Seq
Aiba_2009	88	Mus musculus	embryonic carcinomal cell, ES cell, TS cell, Neural stem cell, Placenta, SNL-STO cell line, 3T3 cell line, MEF, Embryonic germ cells, iPS-MEF	RNA ChIP Microarray
Ang_2011	2	Mus musculus	ES cells, ES Cell	ChIP-Seq
Aravin_2008	9	Mus musculus	10 dpp Dnmt3L heterozygotes testes, 10 dpp Dnmt3L KO mice testes, 16.5 dpc embryos testes, 4-6 week ovaries, 2 day testes, 10 dpp embryos testes	Small RNA-Seq
Asp_2011	8	Mus musculus	C2C12 myoblasts, C2C12 myotubes	ChIP-Seq
Auerbach_2009	1	Homo sapiens	HeLa	ChIP-Seq
Boulker_2010	14	Mus musculus	ES Cells	ChIP-Seq

On the right side of the content area, there is a 'Video Tutorial' button featuring a dog icon and a play button.

# **Open Data**



# Open Data

The screenshot shows a web browser window with the URL [sra-explorer.info](https://sra-explorer.info) in the address bar. The page title is "SRA-Explorer". A header bar includes a "SRA-Explorer" logo, a "0 saved datasets" button, and a search icon. The main content area features a large "SRA Explorer" heading and a subtext: "This tool aims to make datasets within the Sequence Read Archive more accessible." Below this are search input fields: "Search for:" containing "GSE30567[All Fields]", "Max Results" set to 100, and "Start At Record" set to 0. A note at the bottom suggests datasets like GSE30567, SRP043510, PRJEB8073, ERP009109, or human liver miRNA.

SRA-Explorer

0 saved datasets

# SRA Explorer

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

**Search for:** GSE30567[All Fields] \* Q

**Max Results** 100 ^ **Start At Record** 0 ^

Need inspiration? Try [GSE30567](#), [SRP043510](#), [PRJEB8073](#), [ERP009109](#) or [human liver miRNA](#).

Select relevant datasets and click *add to collection*. When you're finished, view all saved datasets with the button in the top right of the page, where you can copy the SRA URLs.

# Open Data

# AWS iGenomes

Common reference genomes hosted on AWS S3

This resource hosts commonly used bioinformatics reference genomes with the help of a grant from [AWS Programs for Research and Education](#).

In order to get the references, you need to sync the files from S3 to your EC2 environment. This web page contains two tools to help you with that - a [command builder](#) and a [command-line script](#).

For more information about this resource, please see the GitHub readme at <https://github.com/ewels/AWS-iGenomes>.

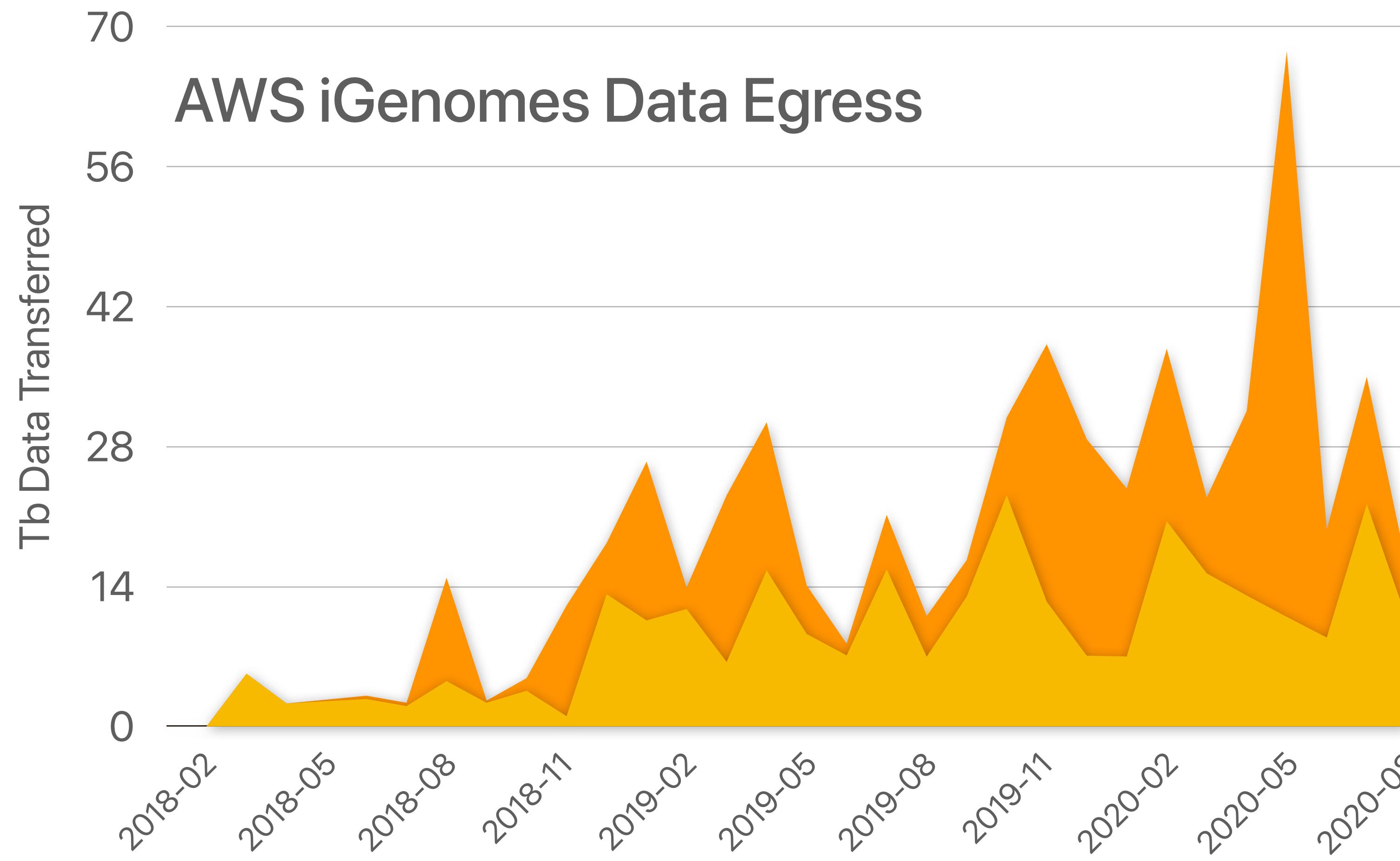
---

## Sync command builder

Use the dropdown boxes below to build an AWS S3 sync command for your reference.

Genome	Source	Build	Type
--------	--------	-------	------

# Open Data



AWS iGenomes

Common reference genomes hosted on AWS S3

This resource hosts commonly used bioinformatics reference genomes with the help of a grant from [AWS Programs for Research and Education](#).

In order to get the references, you need to sync the files from S3 to your EC2 environment. This web page contains two tools to help you with that - a [command builder](#) and a [command-line script](#).

For more information about this resource, please see the GitHub readme at <https://github.com/ewels/AWS-iGenomes>.

**Sync command builder**

Use the dropdown boxes below to build an AWS S3 sync command for your reference.

Genome	Source	Build	Type
Homo sapiens	UCSC	hg38	BWA

Region Currently just eu-west-1

Local directory Available variables: {genome}, {source}, {build}, {type\_path}

eu-west-1 ./references/{genome}/{source}/{build}/{type\_path}

Bucket URL

Copy s3://ngi-igenomes/igenomes/Homo\_sapiens/UCSC/hg38/Sequence/BWAIndex/

Sync command

Copy aws s3 --no-sign-request --region eu-west-1 sync s3://ngi-igenomes/igenomes/Homo\_sapiens/UCSC/hg38/Sequence/BWAIndex/ ./references/Homo\_sapiens/UCSC/hg38/Sequence/BWAIndex/

File list for selected download

Total storage: ~5.5TB

Human STAR index: ~30GB

Tb within AWS  
Tb to internet  
(stacked plot)

# **Open Contributions**



# Open Contributions



# Open Contributions

Packaged (easy to install)

Simple user interface

Usage documentation

Sensible defaults



# Open Contributions

Packaged (easy to install)

Simple user interface

Usage documentation

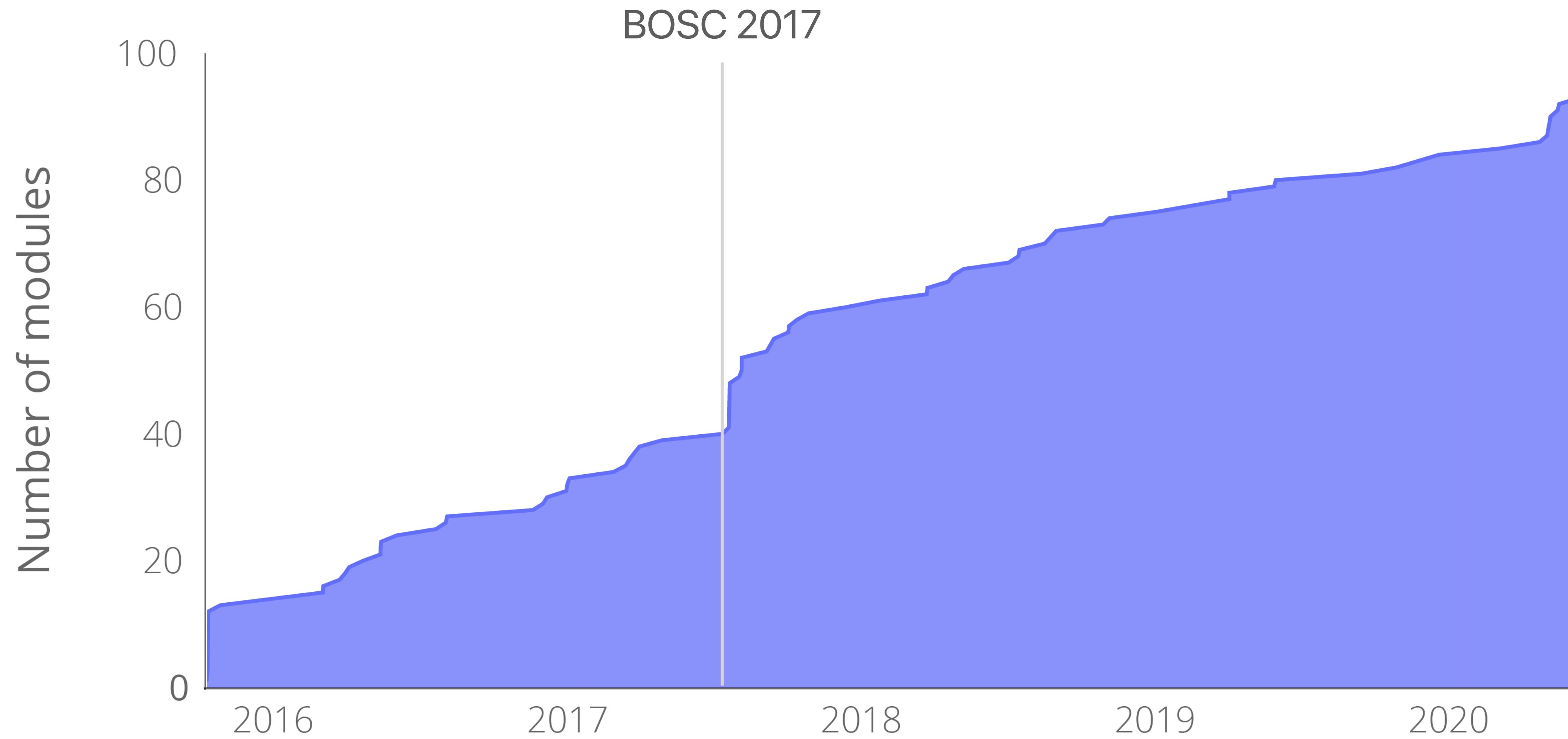
Sensible defaults

Extensibility

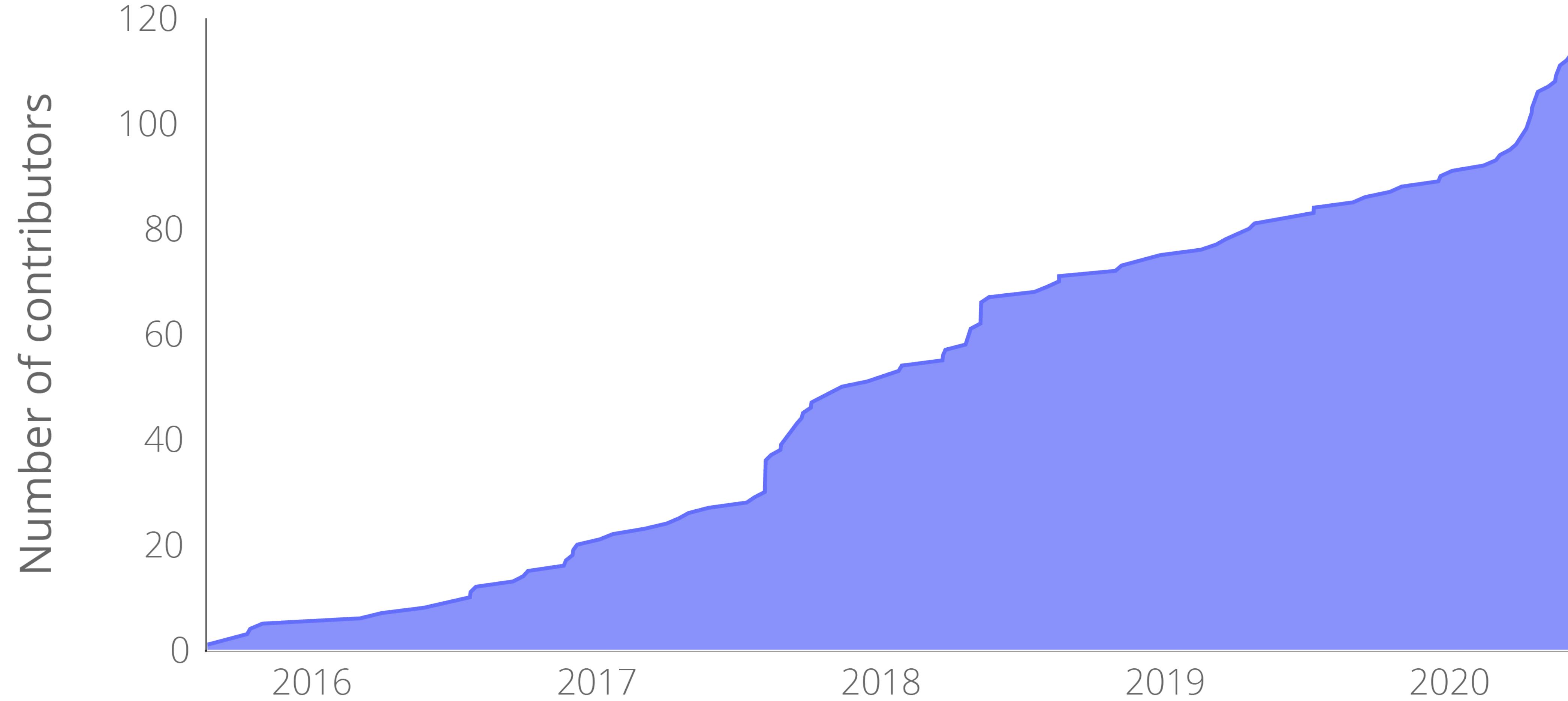
Developer documentation



# Open Contributions



# Open Contributions



# Open Contributions



0.45  per day

Open issue

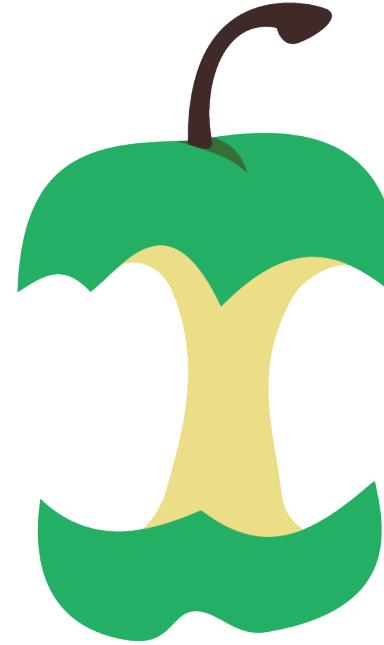
0.23  per day

Open pull request

# **Open Community**



# Open Community

**nf-core** 

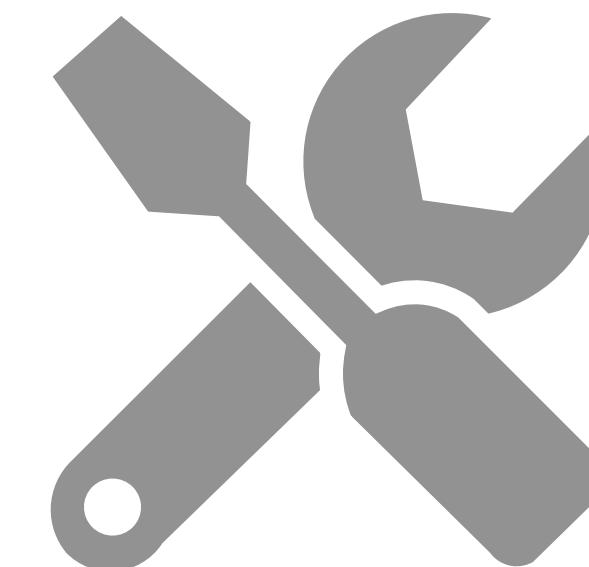
# nf-core



71

PIPELINES

<https://nf-co.re>

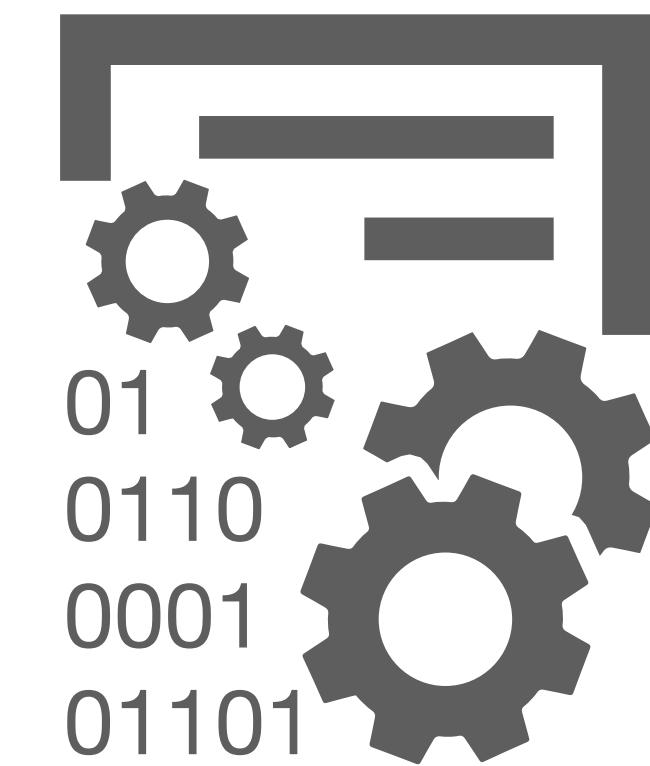


Running pipelines

Writing pipelines

Testing / automation

<https://nf-co.re>

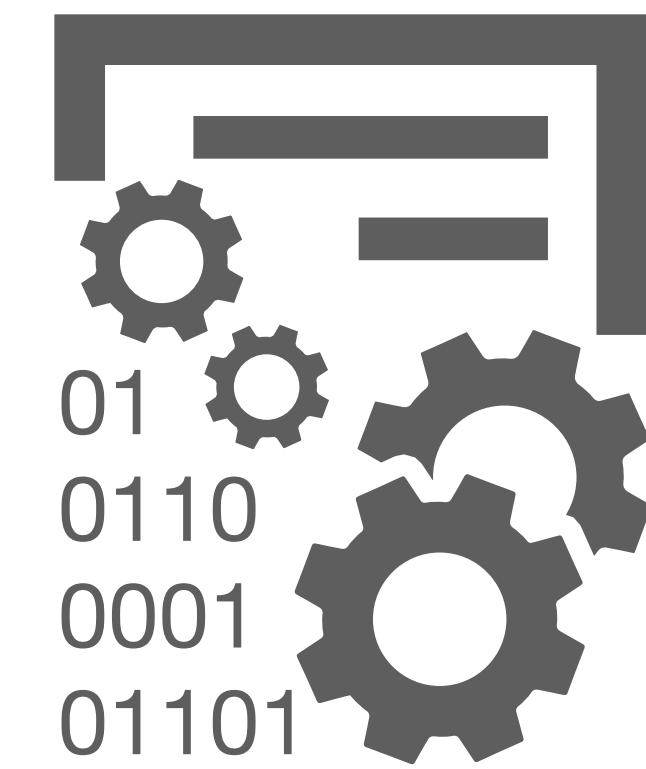


692

MODULES

<https://nf-co.re>

# nf-core



692

MODULES

21

SUB-  
WORKFLOWS

<https://nf-co.re>

# Open Community

Removal of institutional branding

Defined scope and guidelines

Generous access rights

Focus on developer tooling



# Open Community

Removal of institutional branding

Defined scope and guidelines

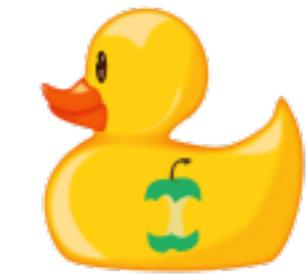
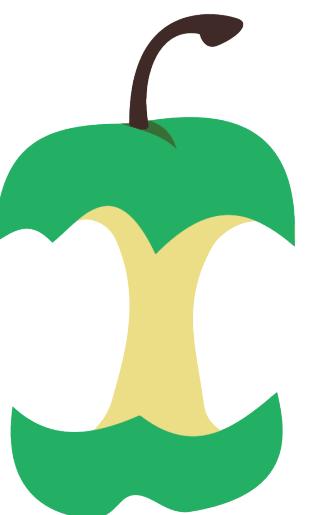
Generous access rights

Focus on developer tooling

Automated testing

Enforced code review

**nf-core**



# Open Community



> 4000

Members on Slack



# Open Community



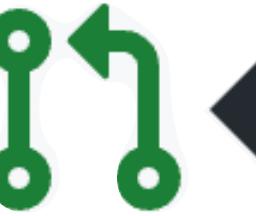
320



Open issue

*nf-core hackathon  
(3 days)*

298



Open pull request

# Open Core



# Open Core

Nextflow and nf-core are supported by:

Chan Zuckerberg Initiative



# Open Core



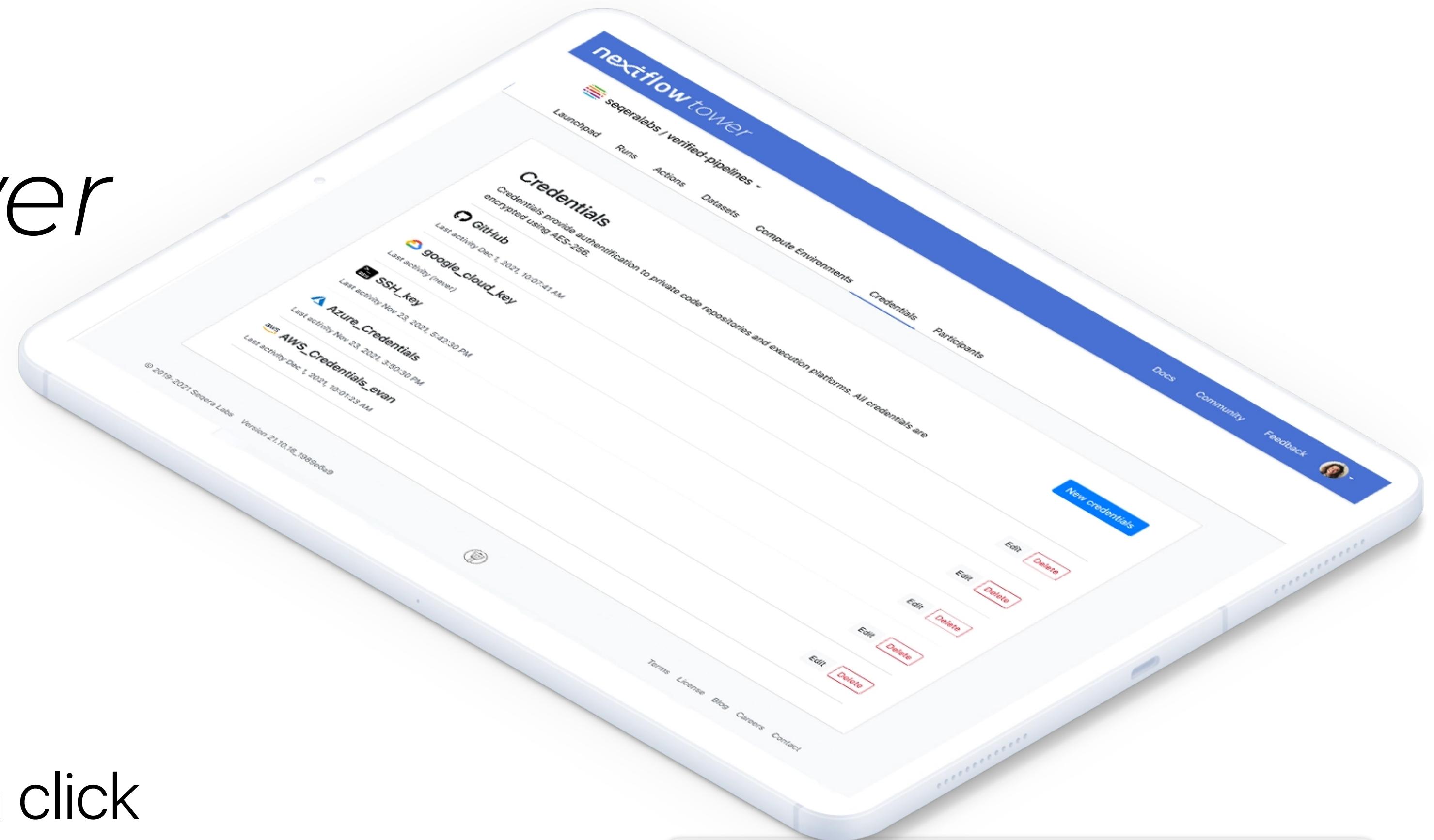
## nextflow tower

Intuitive launchpad interface

Launch, manage, and monitor

Share runs and work in teams

Create cloud infrastructure with a click

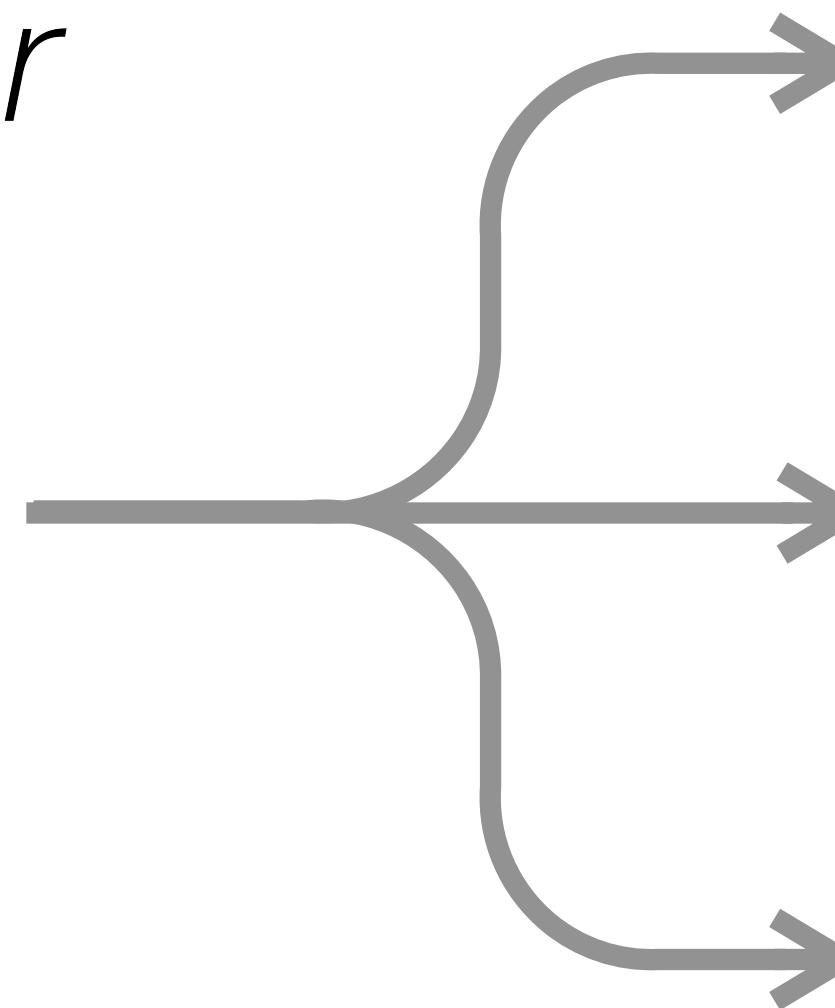
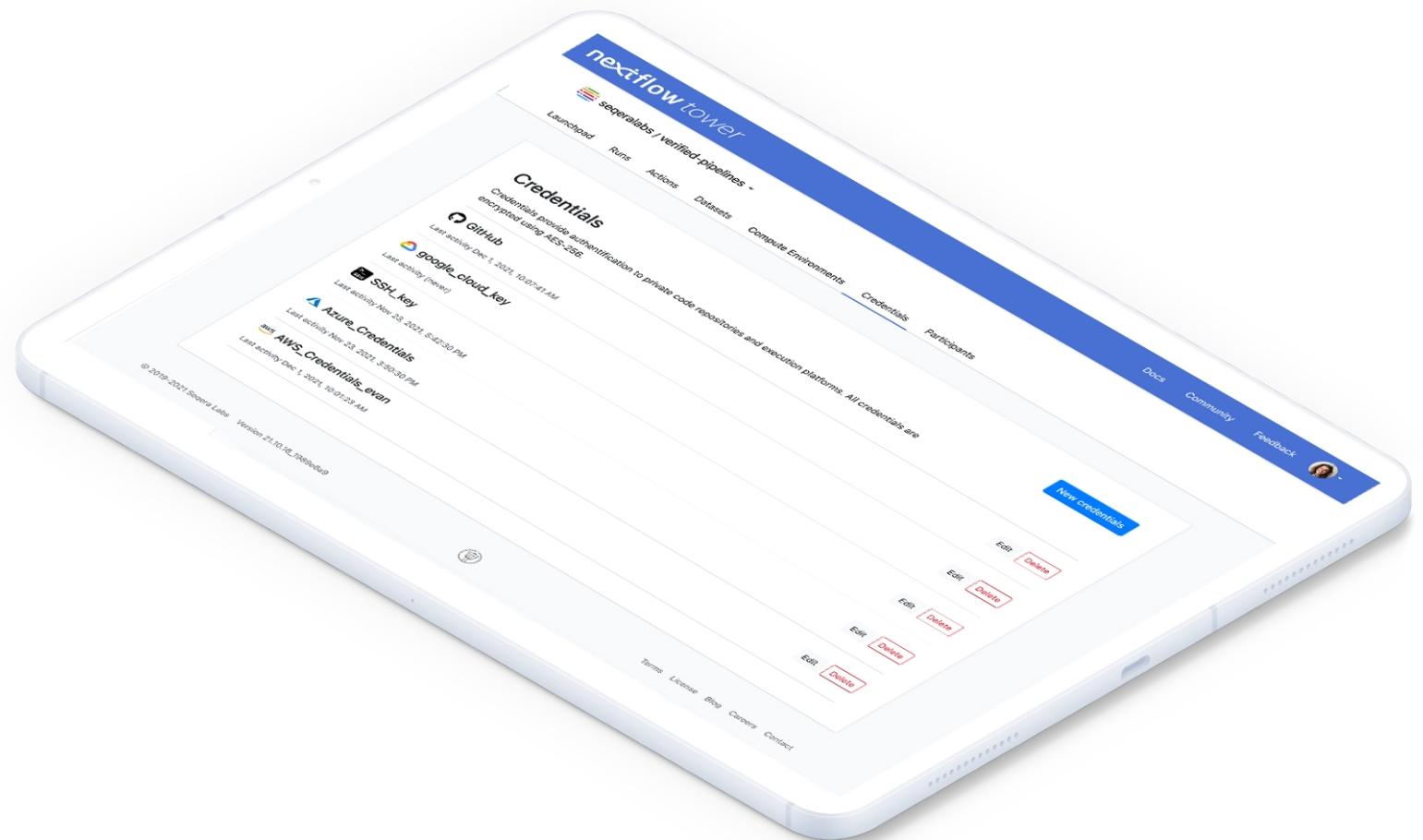


<https://tower.nf>

# Open Core



**nextflow tower**



**Community:** Open source

**Cloud:** Free & paid tiers

**Enterprise:** Commercial

<https://tower.nf>

# Open Core



Open training materials



Open source code



Open communities

# Phil Ewels

<https://phil.ewels.co.uk>

[phil@seqera.io](mailto:phil@seqera.io)

 tallphil

 ewels



**seqeralabs**

<https://seqera.io>

**Chan Zuckerberg  
Initiative**



## nextflow SUMMIT 2022

<https://summit.nextflow.io>



### Mentorship applications

Close November 1st 2022

### Nextflow / nf-core training

6-10 March 2023

### nf-core hackathon

20-24 March 2023

<https://nf-co.re/join>