University of Warsaw
MIM
Faculty
of Mathematics, Informatics
and Mechanics

Uniwersytet Warszawski

Architecture of large projects in bioinformatics (ADP)

Lecture 10

Łukasz P. Kozłowski

Warsaw, 2025

lukaskoz@mimuw.edu.pl

Outline

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. Generic model Organism database (GMOD) project - assumptions, history and usage

6. Genome browsers, problem description and state of the solutions

7. Version control systems (CVS, SVN, git), and online collaboration ad distribution platforms (github, sourceforge).

8. Software testing, automated testing frameworks.

9. Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.

10. Literate programming idea and sweave, markdown, software documentation.

11. Interactive scripting platforms, Rstudio, Jupyter.

Outline

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. **Generic model Organism database (GMOD) project - assumptions, history and usage**

6. **Genome browsers, problem description and state of the solutions**

7. Version control systems (CVS, SVN, git), and online collaboration ad distribution platforms (github, sourceforge)

8. Software testing, automated testing frameworks.

9. **Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.**

10. **Literate programming idea and sweave, markdown, software documentation.**

11. Interactive scripting platforms, Rstudio, Jupyter.

Outline

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. Generic model Organism database (GMOD) project - assumptions, history and usage

6. Genome browsers, problem description and state of the solutions

7. Version control systems (CVS, SVN, git), and online collaboration ad distribution platforms (github, sourceforge)

8. Software testing, automated testing frameworks.

9. Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.

10. Literate programming idea and sweave, markdown, software documentation.

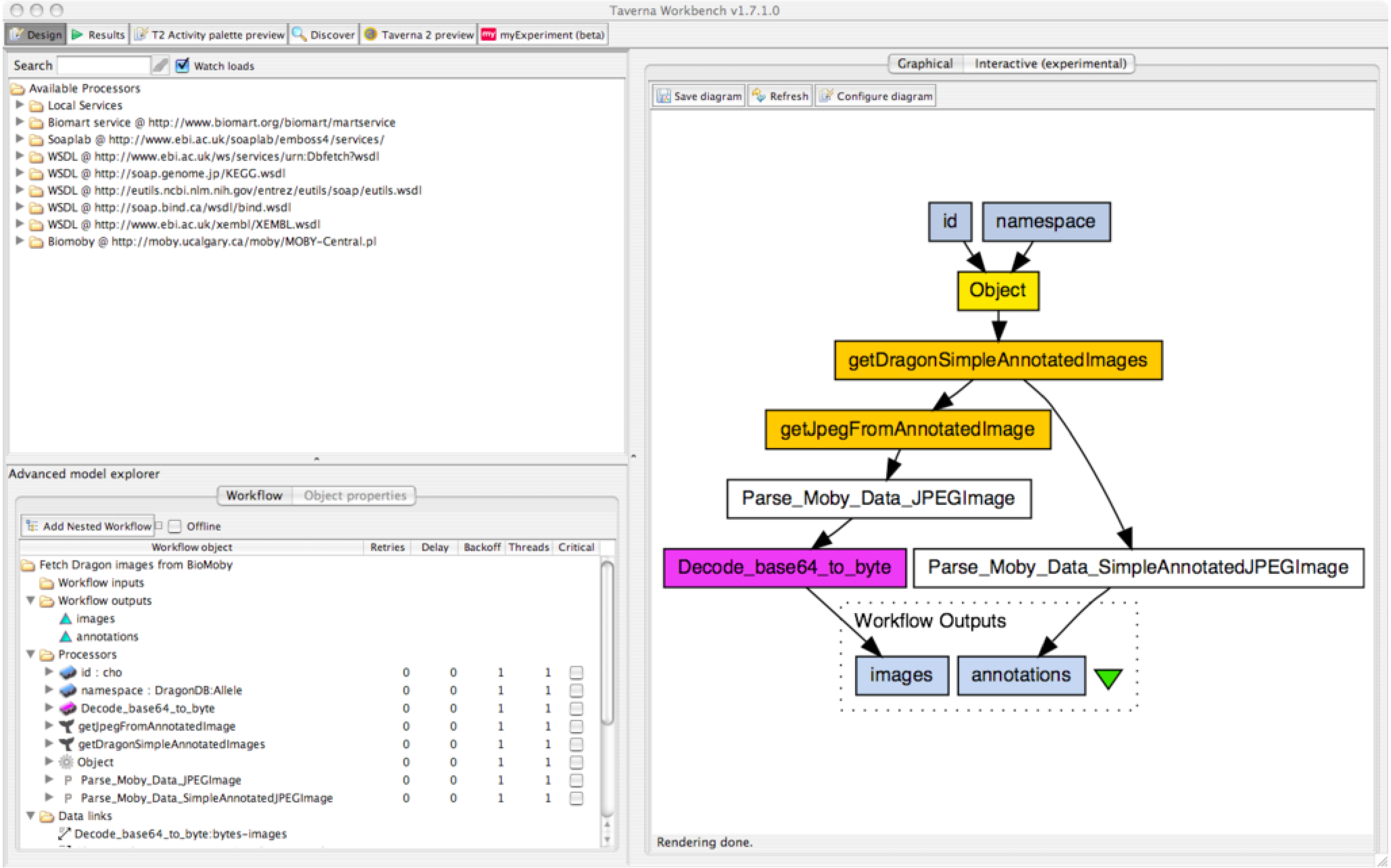
11. Interactive scripting platforms, Rstudio, Jupyter.

Scientific workflow systems

Scientific workflow systems

Taverna

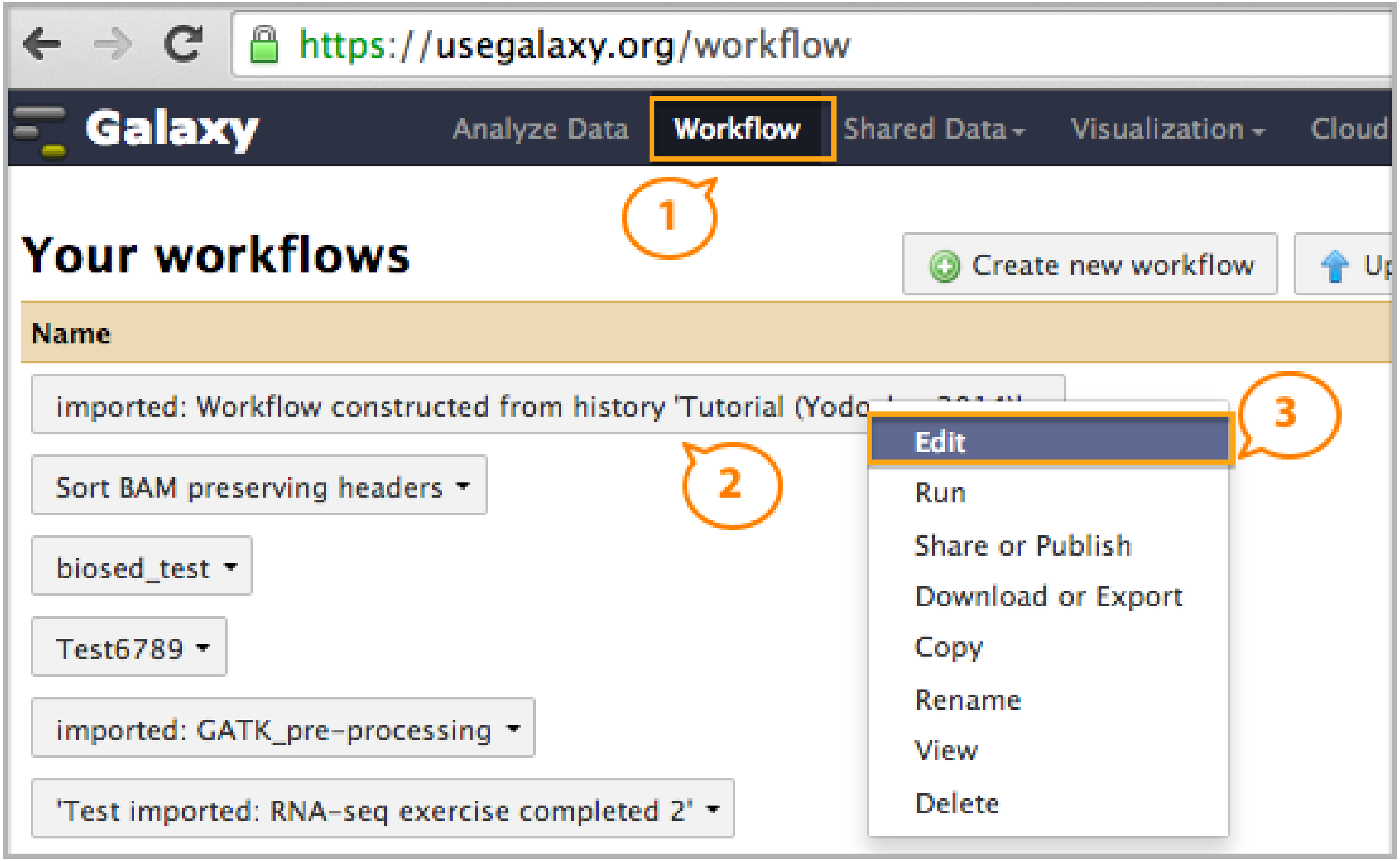
ADP



- Available Processors
  - Local Services
  - Biomart service @ http://www.biomart.org/biomart/martservice
  - Soaplab @ http://www.ebi.ac.uk/soaplab/emboss4/services/
  - WSDL @ http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl
  - WSDL @ http://soap.genome.jp/KEGG.wsdl
  - WSDL @ http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl
  - WSDL @ http://soap.bind.ca/wsdl/bind.wsdl
  - WSDL @ http://www.ebi.ac.uk/xembl/XEMBL.wsdl
  - Biomoby @ http://moby.ucalgary.ca/moby/MOBY-Central.pl

Workflow object	Retries	Delay	Backoff	Threads	Critical
Fetch Dragon images from BioMoby					
Workflow inputs					
Workflow outputs					
images					
annotations					
Processors					
id : cho	0	0	1	1	□
namespace : DragonDB:Allele	0	0	1	1	□
Decode_base64_to_byte	0	0	1	1	□
getJpegFromAnnotatedImage	0	0	1	1	□
getDragonSimpleAnnotatedImages	0	0	1	1	□
Object	0	0	1	1	□
Parse_Moby_Data_JPEGImage	0	0	1	1	□
Parse_Moby_Data_SimpleAnnotatedJPEGImage	0	0	1	1	□
Data links					
Decode_base64_to_byte:bytes-images					

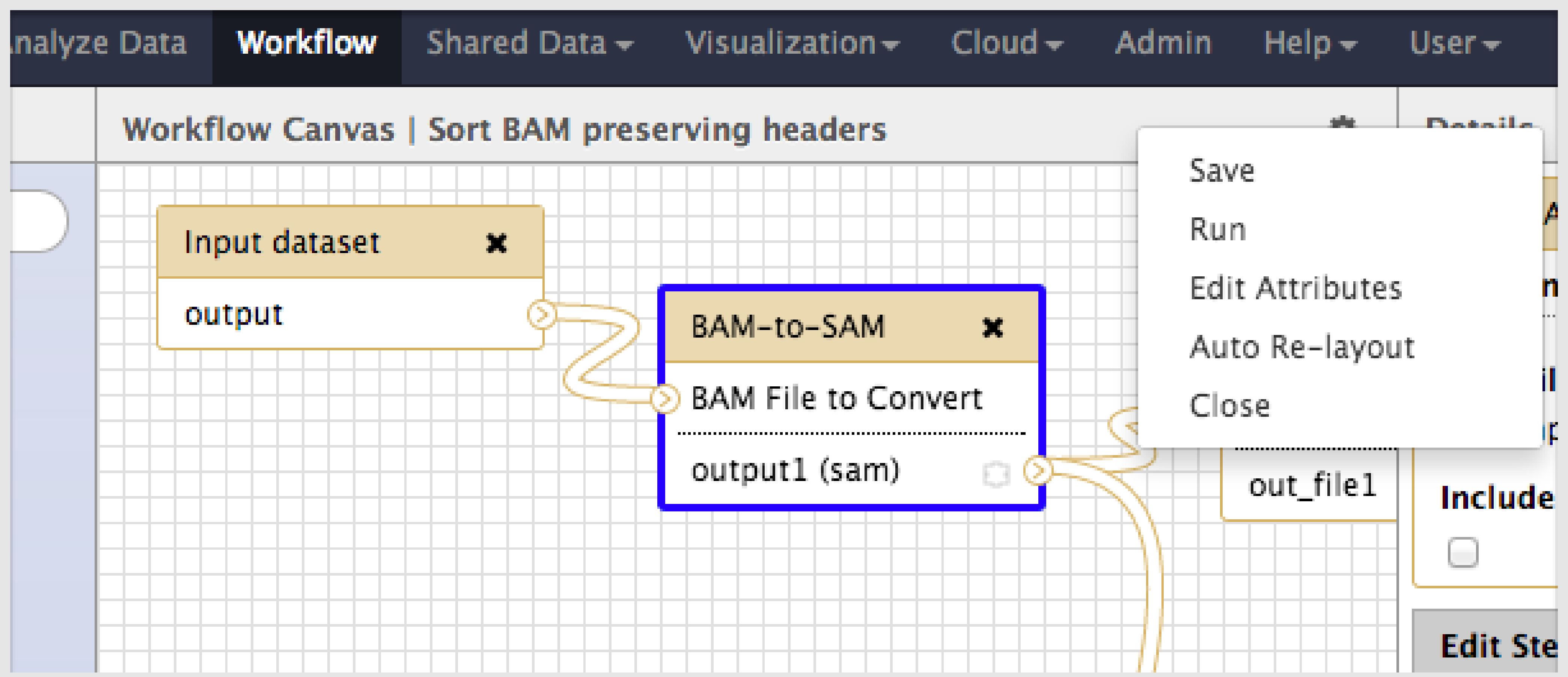
Rendering done.

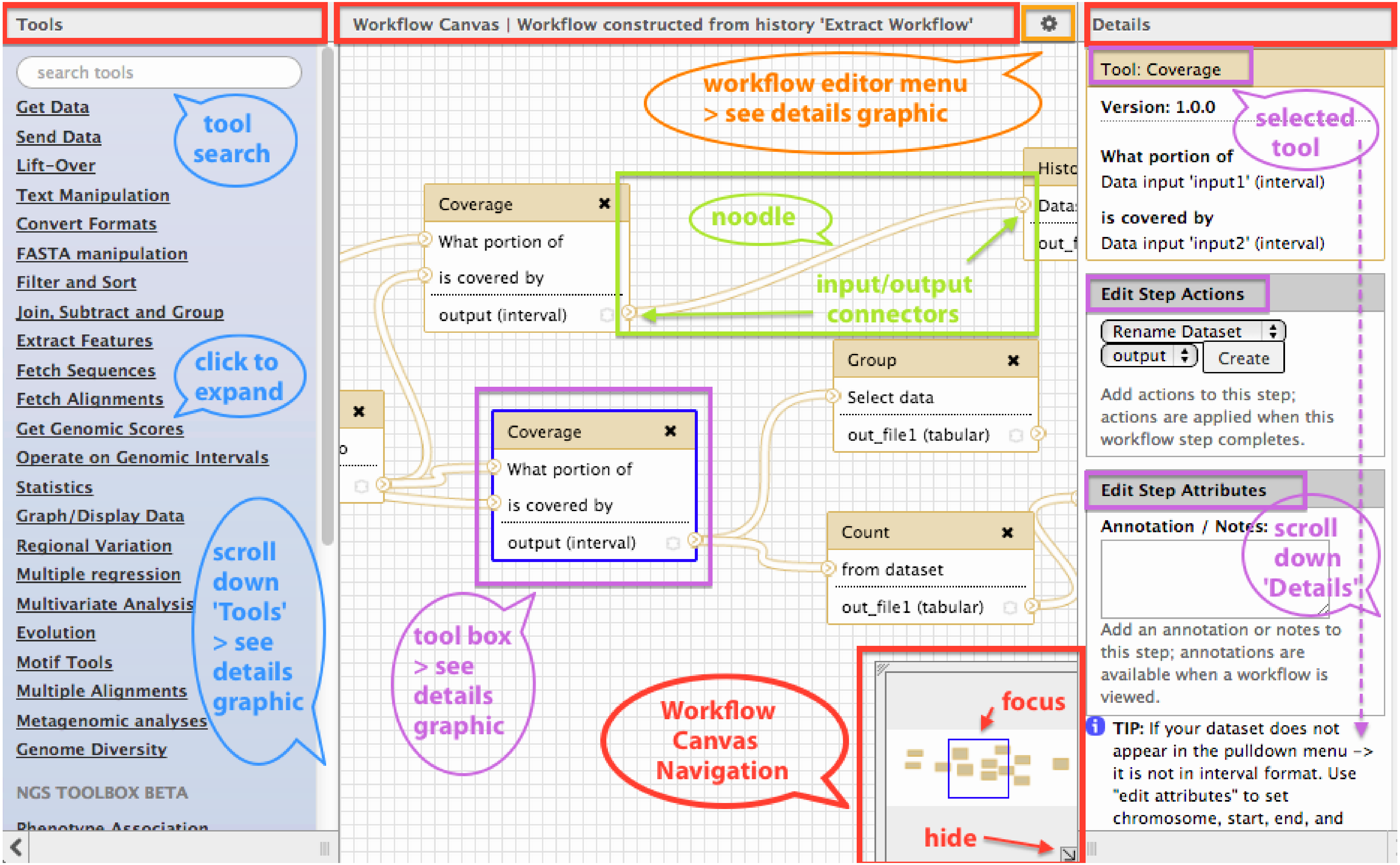


https://usegalaxy.org/workflow

Your workflows

Name
imported: Workflow constructed from history 'Tutorial (Yodo
Sort BAM preserving headers
biosed_test
Test6789
imported: GATK_pre-processing
'Test imported: RNA-seq exercise completed 2'


Analyze Data
Workflow
Shared Data
Visualization
Cloud
Admin
Help
User
Workflow Canvas | Sort BAM preserving headers
Input dataset
output
BAM-to-SAM
BAM File to Convert
output1 (sam)
out_file1
Save
Run
Edit Attributes
Auto Re-layout
Close
Include
Edit Ste



Version: 1.0.0

What portion of
Data input 'input1' (interval)

is covered by
Data input 'input2' (interval)

Add actions to this step; actions are applied when this workflow step completes.

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

TIP: If your dataset does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and

tool search

click to expand

scroll down 'Tools' > see details graphic

workflow editor menu > see details graphic

noodle

input/output connectors

tool box > see details graphic

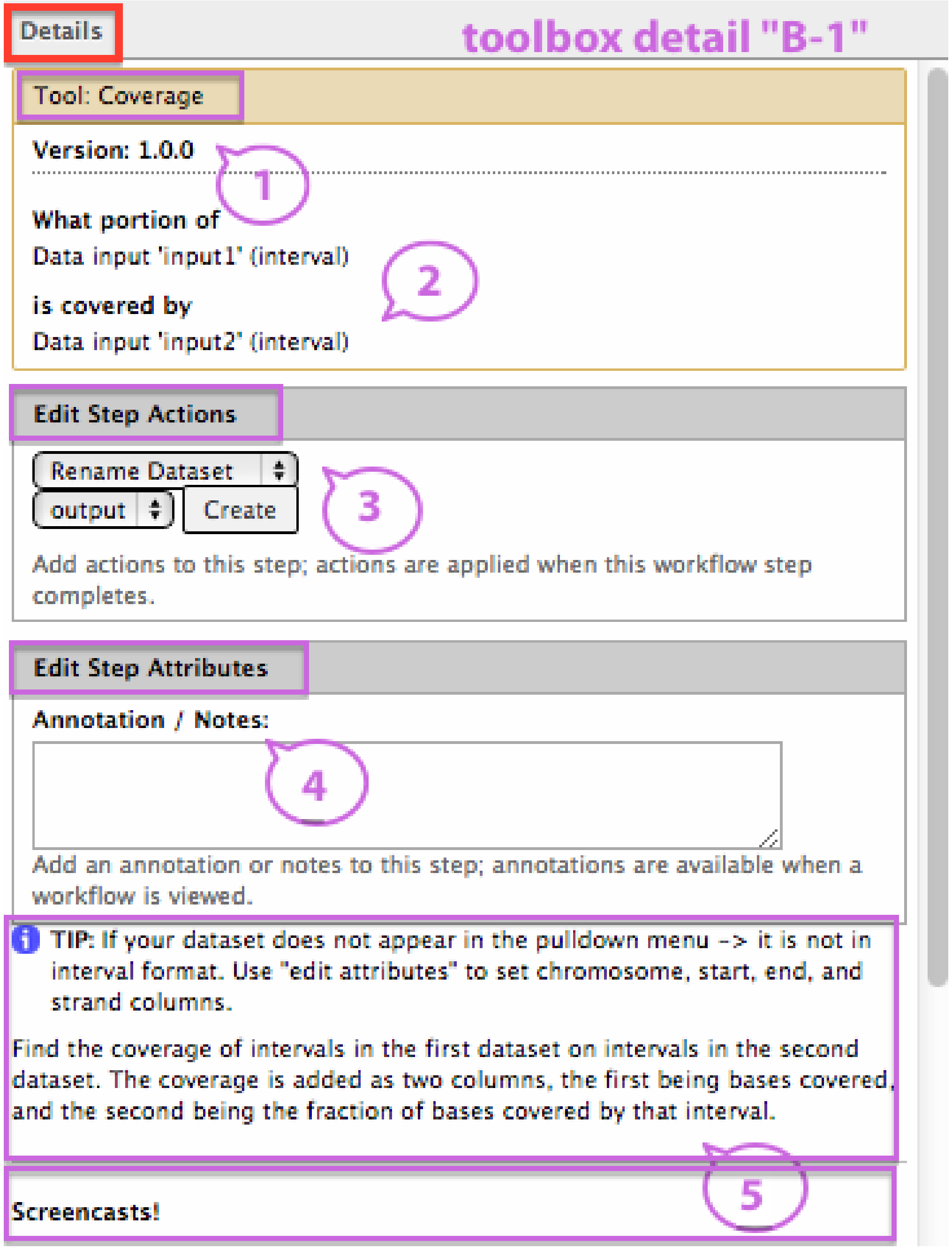
Workflow Canvas Navigation

focus

hide

selected tool

scroll down 'Details'



Details

Tool: Coverage

Version: 1.0.0

What portion of

Data input 'input1' (interval)

is covered by

Data input 'input2' (interval)

Edit Step Actions

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

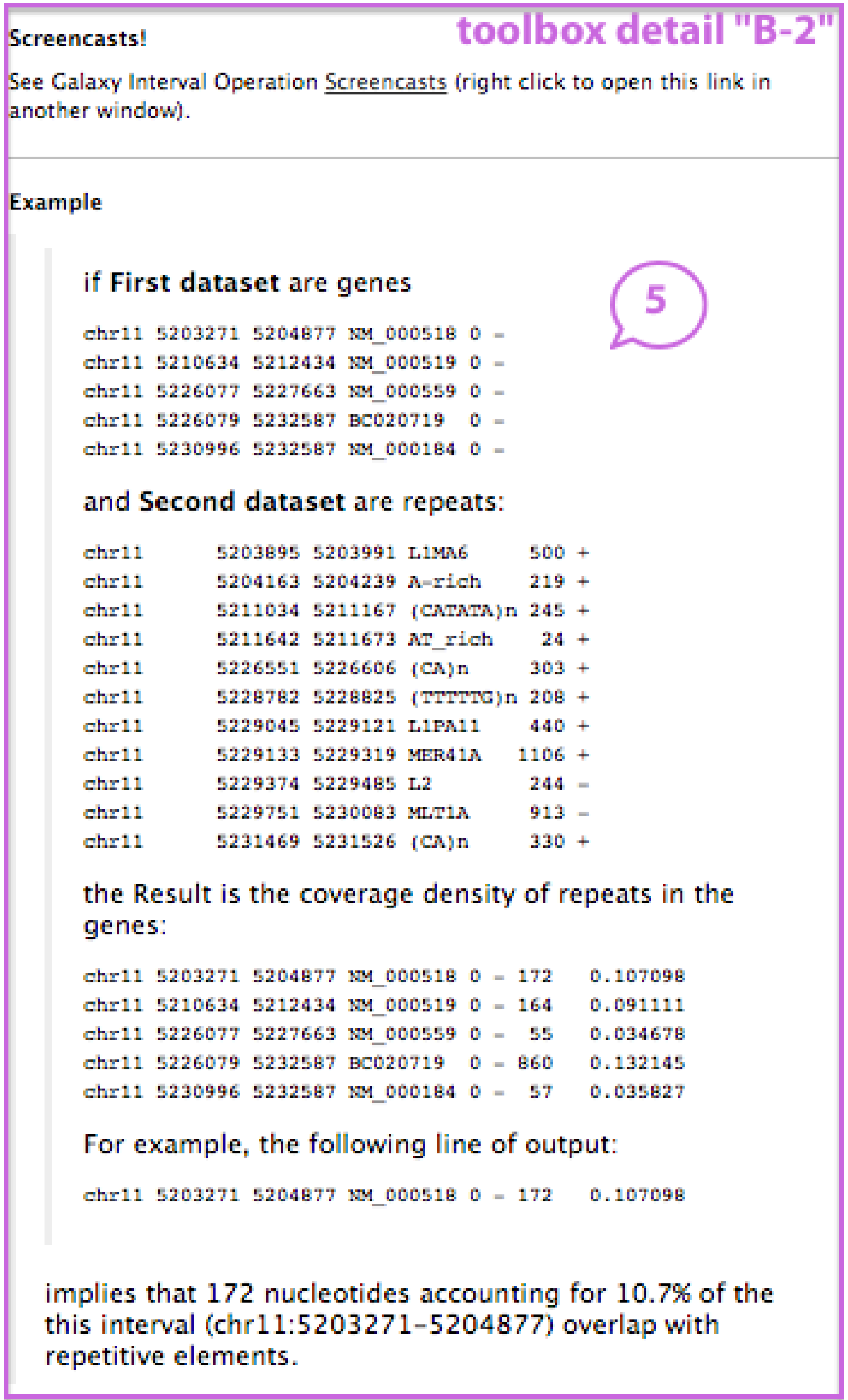
Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

TIP: If your dataset does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Find the coverage of intervals in the first dataset on intervals in the second dataset. The coverage is added as two columns, the first being bases covered, and the second being the fraction of bases covered by that interval.

Screencasts!



Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Example

if **First dataset** are genes

```
chr11 5203271 5204877 NM_000518 0 -
chr11 5210634 5212434 NM_000519 0 -
chr11 5226077 5227663 NM_000559 0 -
chr11 5226079 5232587 BC020719  0 -
chr11 5230996 5232587 NM_000184 0 -
```

and **Second dataset** are repeats:

```
chr11       5203895 5203991 L1MA6      500 +
chr11       5204163 5204239 A-rich     219 +
chr11       5211034 5211167 (CATATA)n 245 +
chr11       5211642 5211673 AT_rich     24 +
chr11       5226551 5226606 (CA)n      303 +
chr11       5228782 5228825 (TTTTTG)n 208 +
chr11       5229045 5229121 L1PA11     440 +
chr11       5229133 5229319 MER41A    1106 +
chr11       5229374 5229485 L2         244 -
chr11       5229751 5230083 MLT1A      913 -
chr11       5231469 5231526 (CA)n      330 +
```

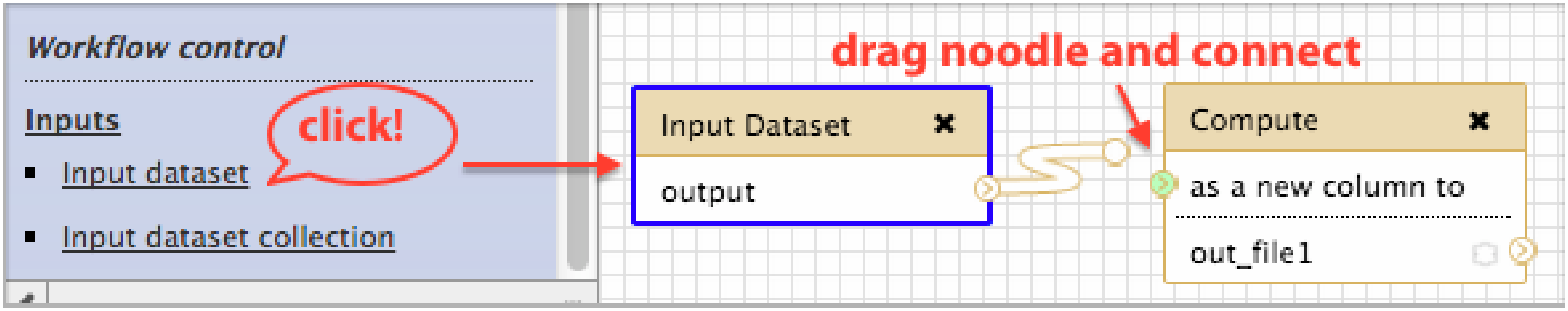
the Result is the coverage density of repeats in the genes:

```
chr11 5203271 5204877 NM_000518 0 - 172   0.107098
chr11 5210634 5212434 NM_000519 0 - 164   0.091111
chr11 5226077 5227663 NM_000559 0 -  55   0.034678
chr11 5226079 5232587 BC020719  0 - 860   0.132145
chr11 5230996 5232587 NM_000184 0 -  57   0.035827
```

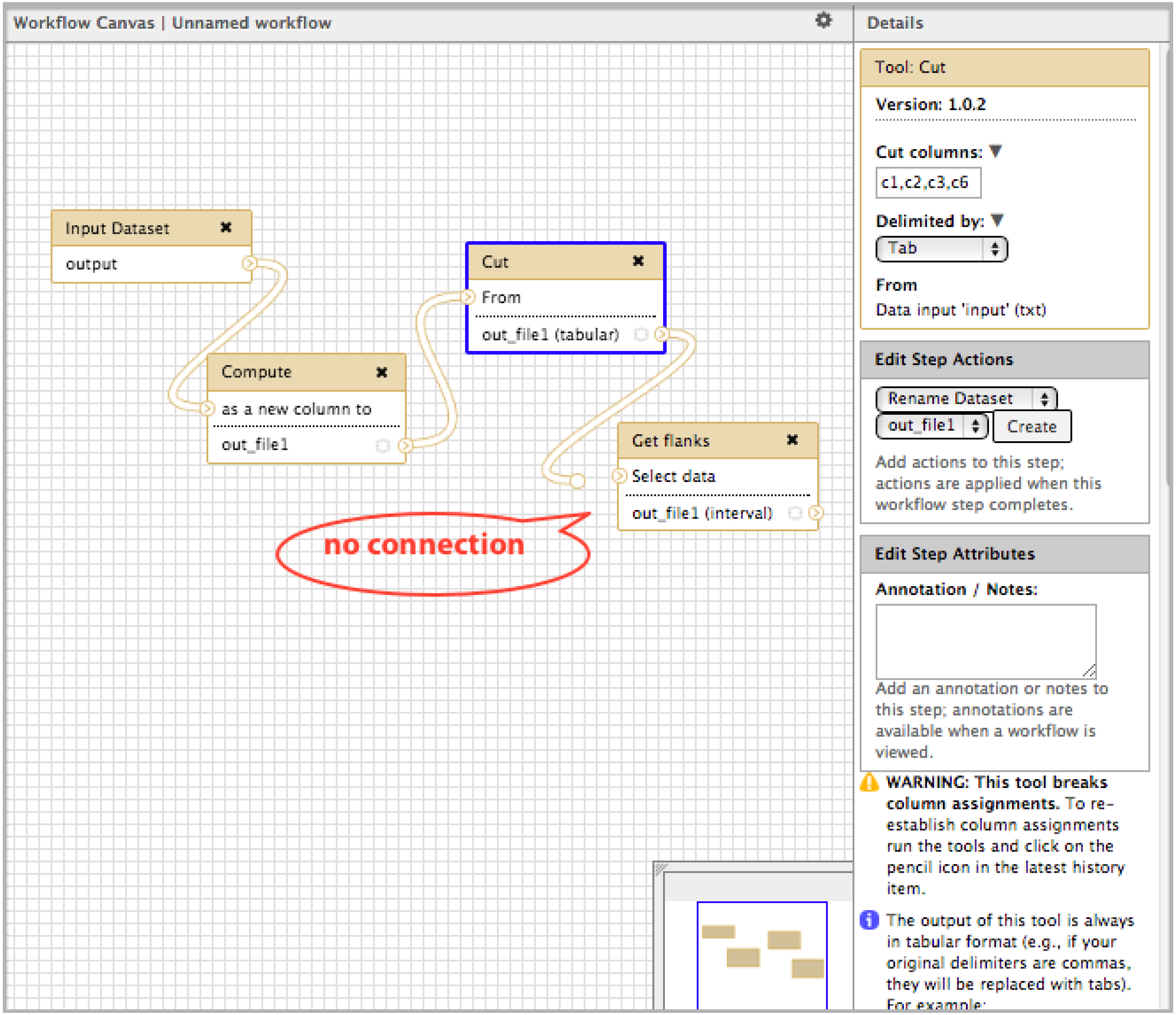
For example, the following line of output:

```
chr11 5203271 5204877 NM_000518 0 - 172   0.107098
```

implies that 172 nucleotides accounting for 10.7% of the this interval (chr11:5203271-5204877) overlap with repetitive elements.


Workflow control
Inputs
Input dataset
Input dataset collection
click!
drag noodle and connect
Input Dataset
output
Compute
as a new column to
out_file1

Galaxy ADP



Tool: Cut

Version: 1.0.2

Cut columns: c1,c2,c3,c6

Delimited by: Tab

From
Data input 'input' (txt)

Edit Step Actions

Rename Dataset | out_file1 | Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

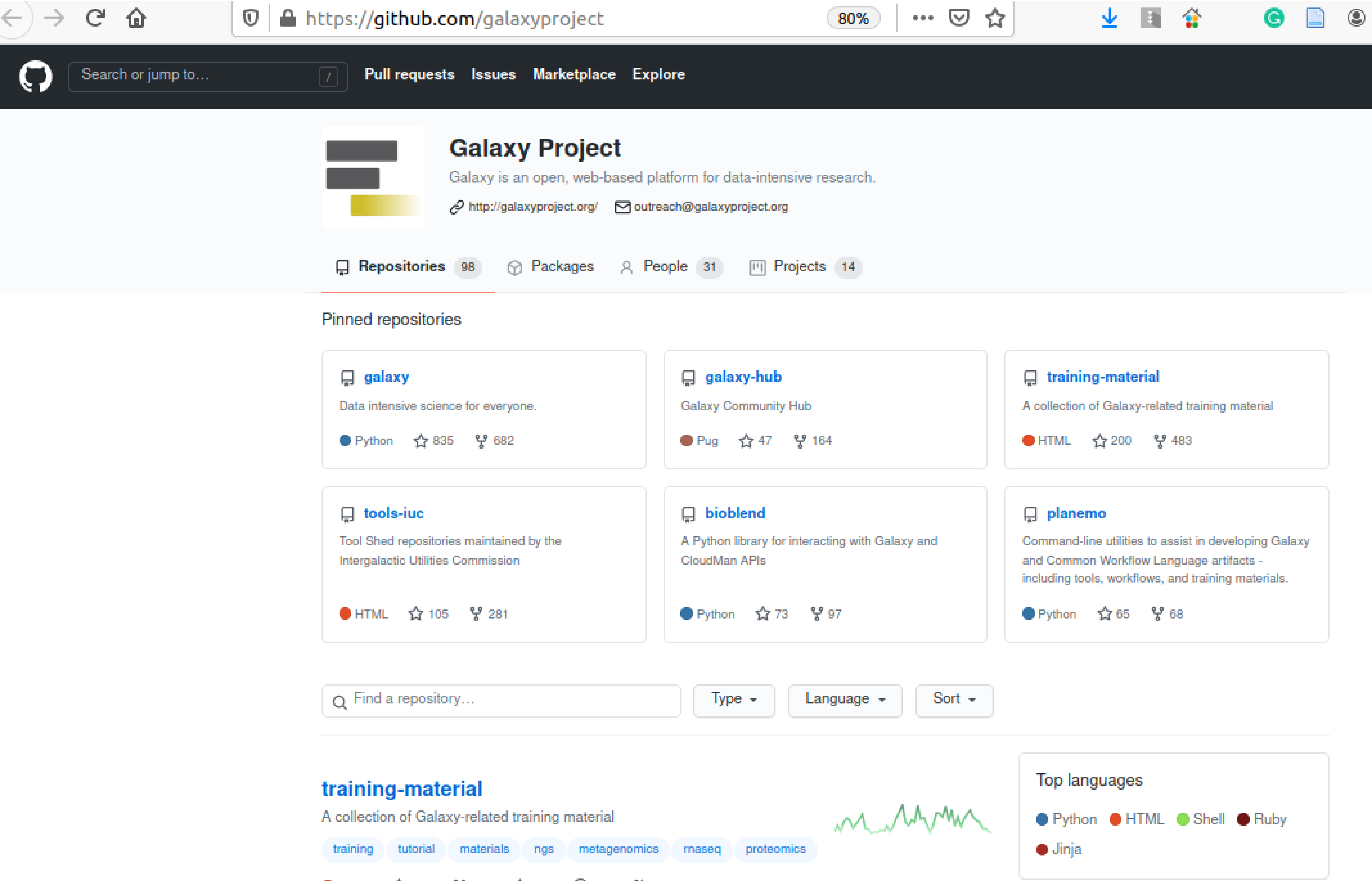
Add an annotation or notes to this step; annotations are available when a workflow is viewed.

⚠ **WARNING: This tool breaks column assignments.** To re-establish column assignments run the tools and click on the pencil icon in the latest history item.

ℹ The output of this tool is always in tabular format (e.g., if your original delimiters are commas, they will be replaced with tabs). For example:

Galaxy

ADP



Galaxy Project

Galaxy is an open, web-based platform for data-intensive research.

http://galaxyproject.org/ outreach@galaxyproject.org

Pinned repositories

galaxy
Data intensive science for everyone.
● Python ☆ 835 682

galaxy-hub
Galaxy Community Hub
● Pug ☆ 47 164

training-material
A collection of Galaxy-related training material
● HTML ☆ 200 483

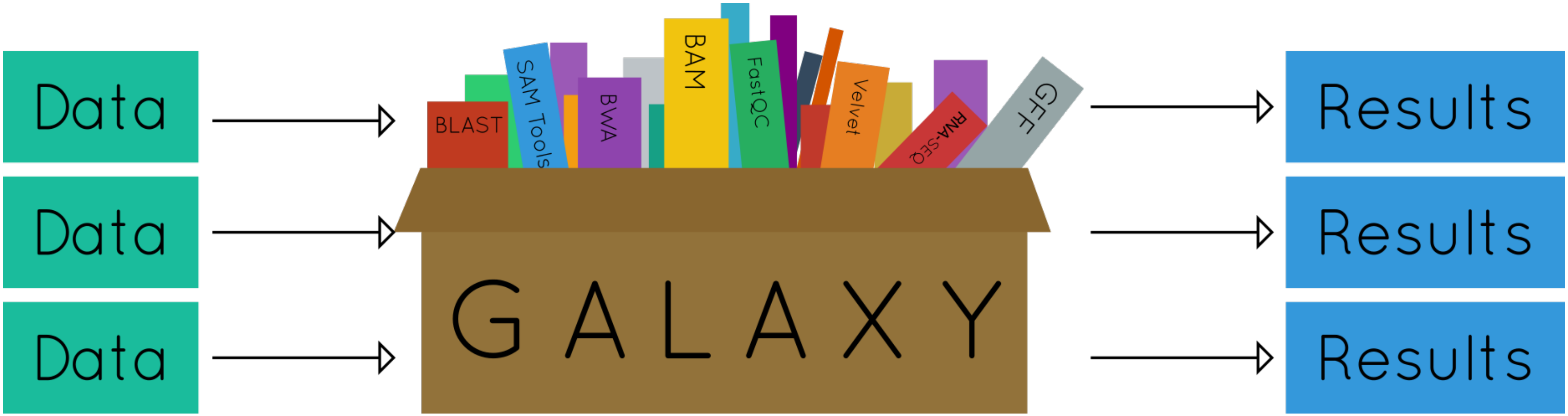
tools-iuc
Tool Shed repositories maintained by the Intergalactic Utilities Commission
● HTML ☆ 105 281

bioblend
A Python library for interacting with Galaxy and CloudMan APIs
● Python ☆ 73 97

planemo
Command-line utilities to assist in developing Galaxy and Common Workflow Language artifacts - including tools, workflows, and training materials.
● Python ☆ 65 68

training-material
A collection of Galaxy-related training material
training tutorial materials ngs metagenomics rnaseq proteomics

Top languages
● Python ● HTML ● Shell ● Ruby ● Jinja



Data

Data

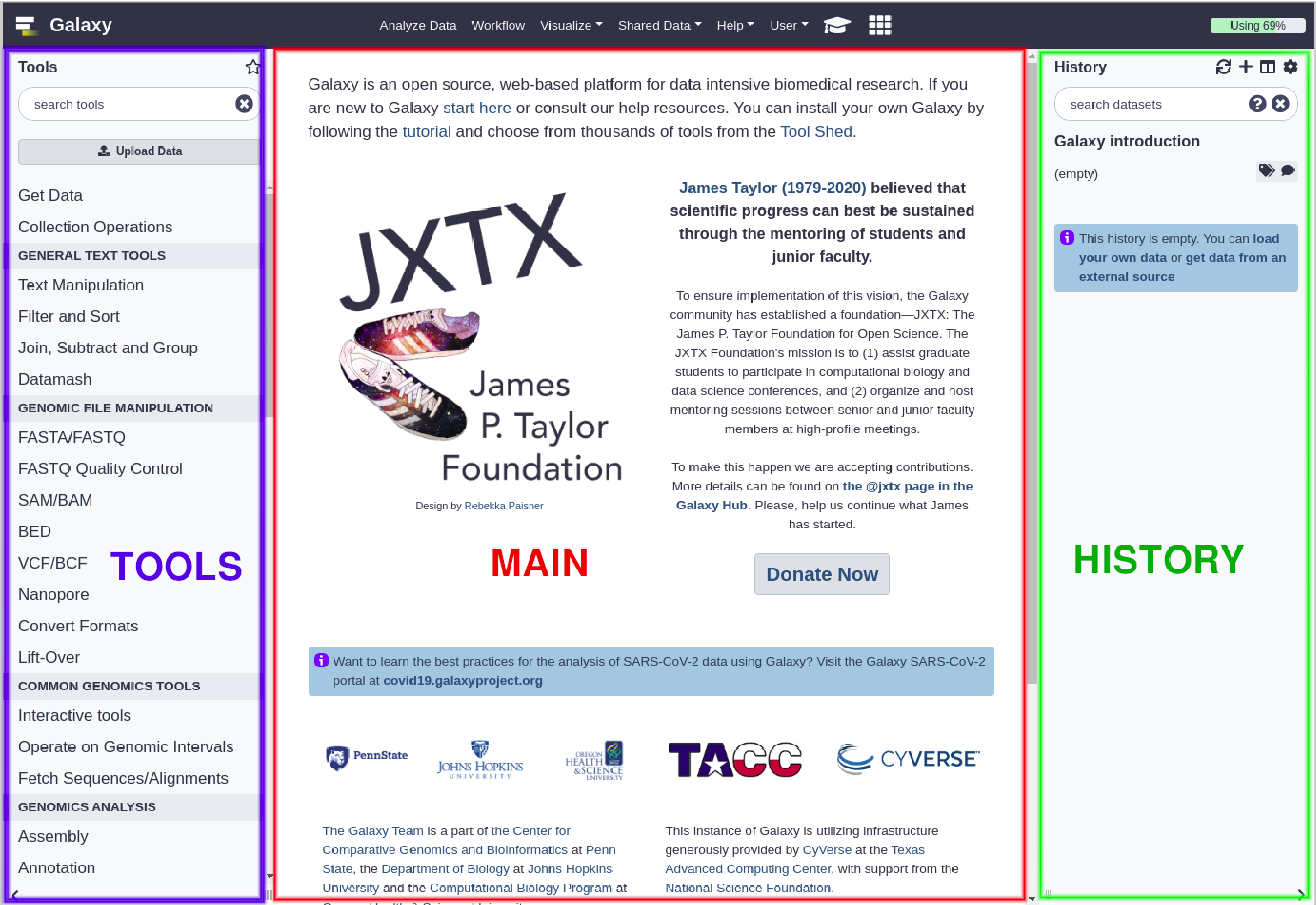
Data

Results

Results

Results

Galaxy



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Design by Rebekka Paisner

James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.

To ensure implementation of this vision, the Galaxy community has established a foundation—JXTX: The James P. Taylor Foundation for Open Science. The JXTX Foundation's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

To make this happen we are accepting contributions. More details can be found on **the @jxtx page in the Galaxy Hub**. Please, help us continue what James has started.

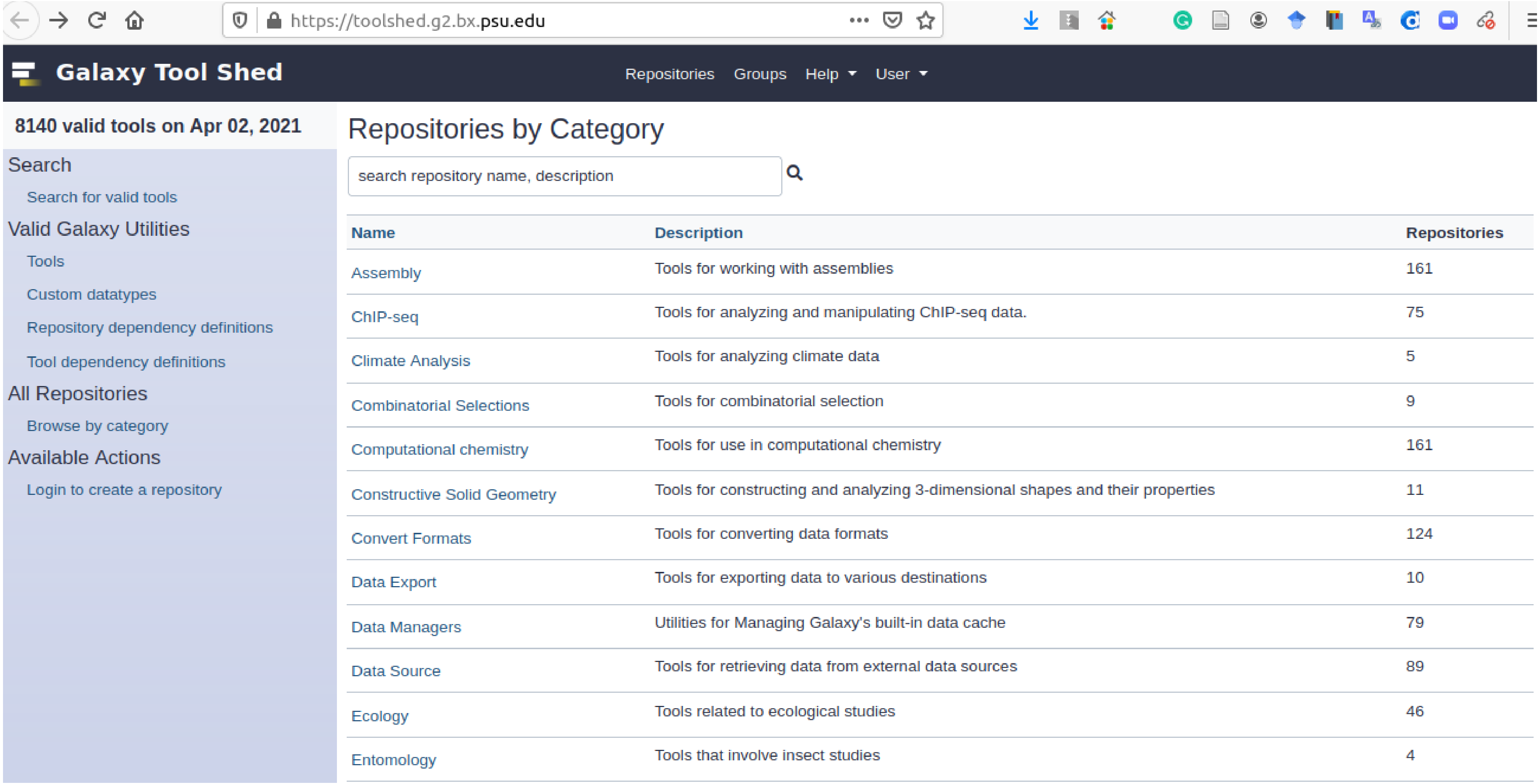
Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at **covid19.galaxyproject.org**

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology at Johns Hopkins University and the Computational Biology Program at

This instance of Galaxy is utilizing infrastructure generously provided by CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

This history is empty. You can **load your own data** or **get data from an external source**

Galaxy



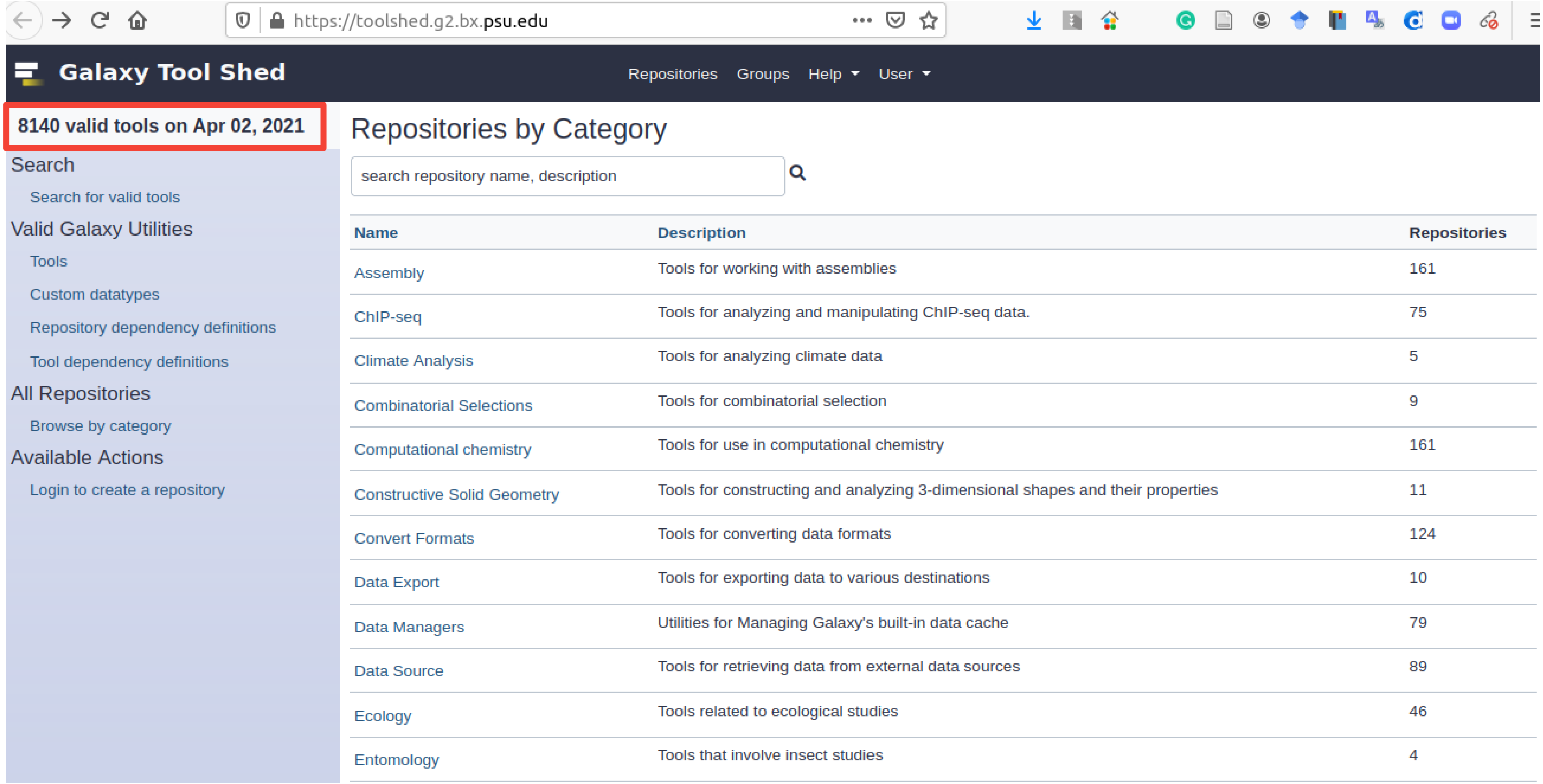
8140 valid tools on Apr 02, 2021

Repositories by Category

Name	Description	Repositories
Assembly	Tools for working with assemblies	161
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	75
Climate Analysis	Tools for analyzing climate data	5
Combinatorial Selections	Tools for combinatorial selection	9
Computational chemistry	Tools for use in computational chemistry	161
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	11
Convert Formats	Tools for converting data formats	124
Data Export	Tools for exporting data to various destinations	10
Data Managers	Utilities for Managing Galaxy's built-in data cache	79
Data Source	Tools for retrieving data from external data sources	89
Ecology	Tools related to ecological studies	46
Entomology	Tools that involve insect studies	4

Galaxy

ADP



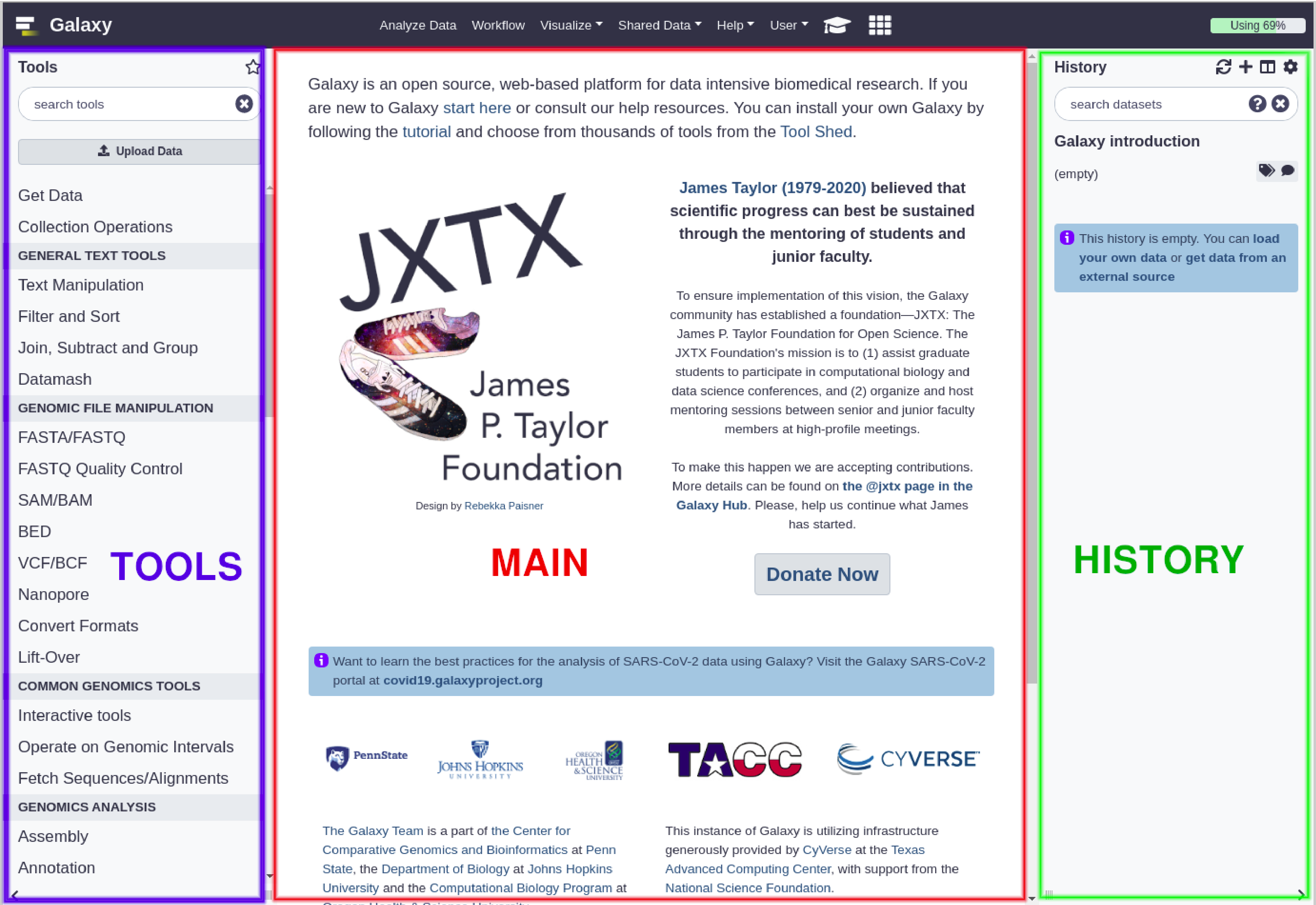
8140 valid tools on Apr 02, 2021

Repositories by Category

Name	Description	Repositories
Assembly	Tools for working with assemblies	161
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	75
Climate Analysis	Tools for analyzing climate data	5
Combinatorial Selections	Tools for combinatorial selection	9
Computational chemistry	Tools for use in computational chemistry	161
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	11
Convert Formats	Tools for converting data formats	124
Data Export	Tools for exporting data to various destinations	10
Data Managers	Utilities for Managing Galaxy's built-in data cache	79
Data Source	Tools for retrieving data from external data sources	89
Ecology	Tools related to ecological studies	46
Entomology	Tools that involve insect studies	4

Galaxy

ADP



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Design by Rebekka Paisner

James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.

To ensure implementation of this vision, the Galaxy community has established a foundation—JXTX: The James P. Taylor Foundation for Open Science. The JXTX Foundation's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

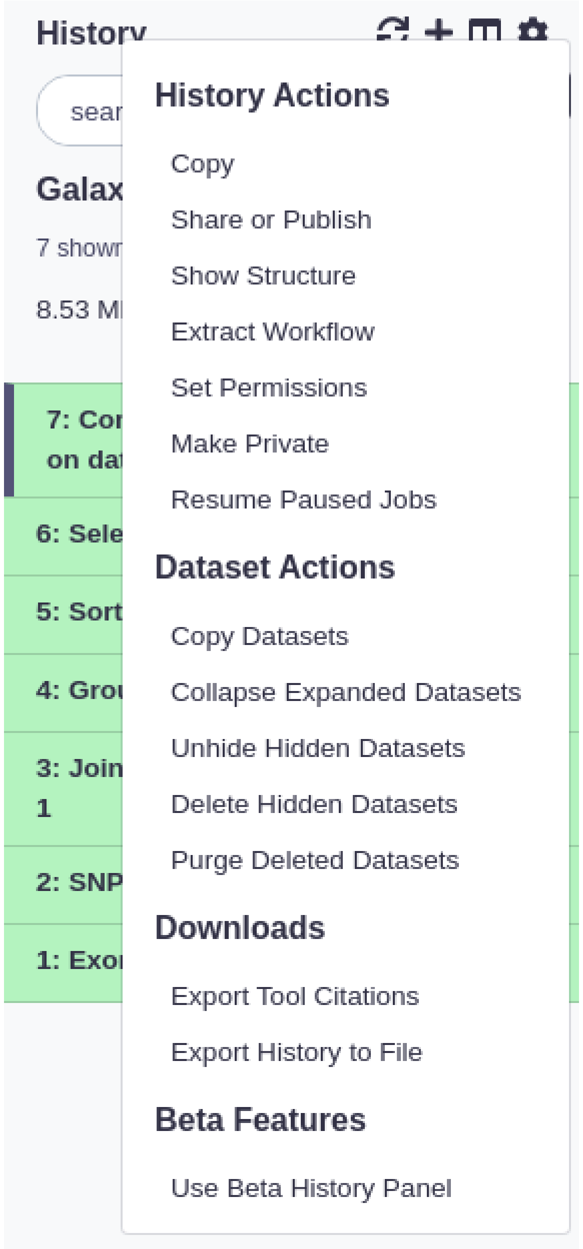
To make this happen we are accepting contributions. More details can be found on **the @jxtx page in the Galaxy Hub**. Please, help us continue what James has started.

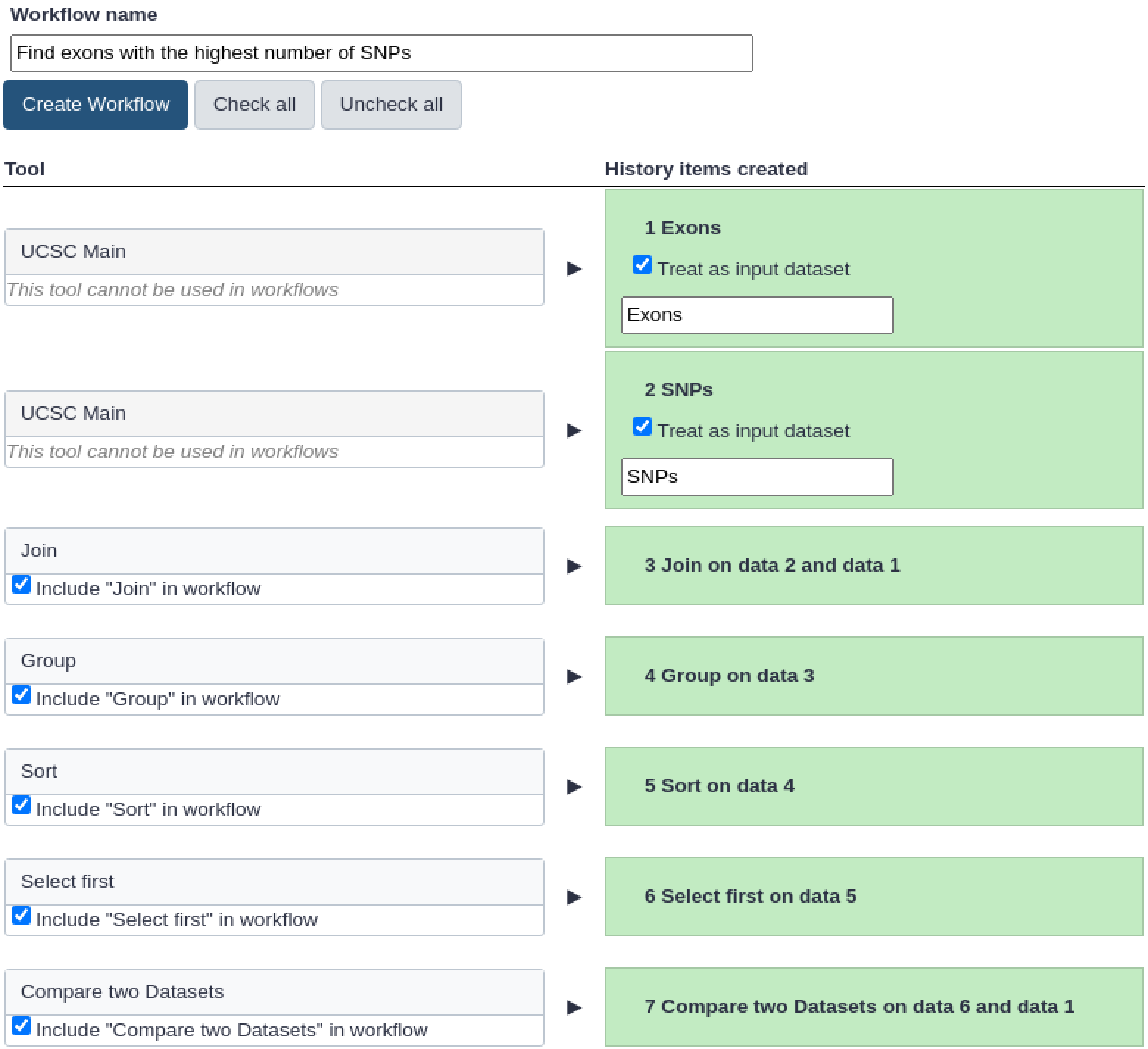
Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at **covid19.galaxyproject.org**

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology at Johns Hopkins University and the Computational Biology Program at

This instance of Galaxy is utilizing infrastructure generously provided by CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

This history is empty. You can **load your own data** or **get data from an external source**

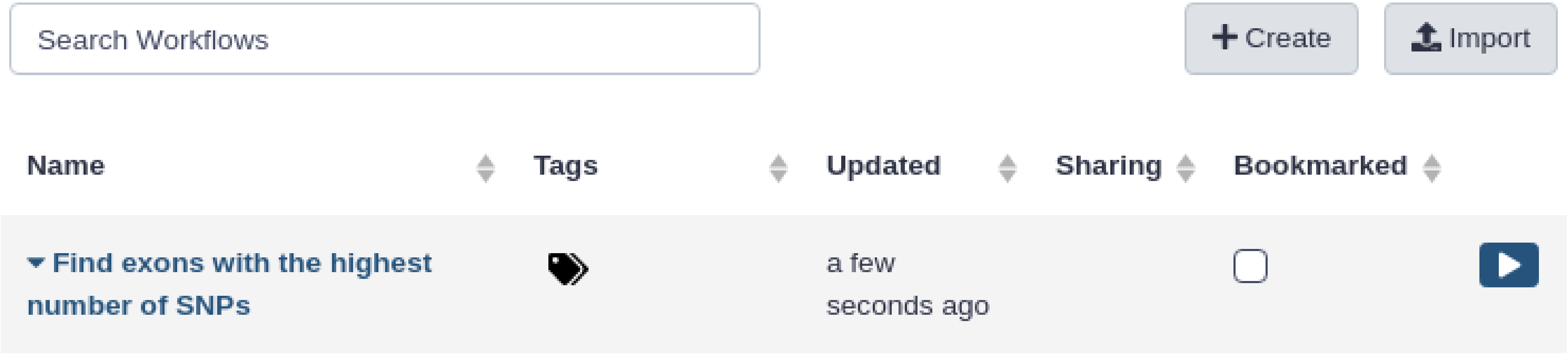

History
sear
Galax
7 shown
8.53 MB
7: Cor
on dat
6: Sele
5: Sort
4: Grou
3: Join
1
2: SNP
1: Exon
History Actions
Copy
Share or Publish
Show Structure
Extract Workflow
Set Permissions
Make Private
Resume Paused Jobs
Dataset Actions
Copy Datasets
Collapse Expanded Datasets
Unhide Hidden Datasets
Delete Hidden Datasets
Purge Deleted Datasets
Downloads
Export Tool Citations
Export History to File
Beta Features
Use Beta History Panel



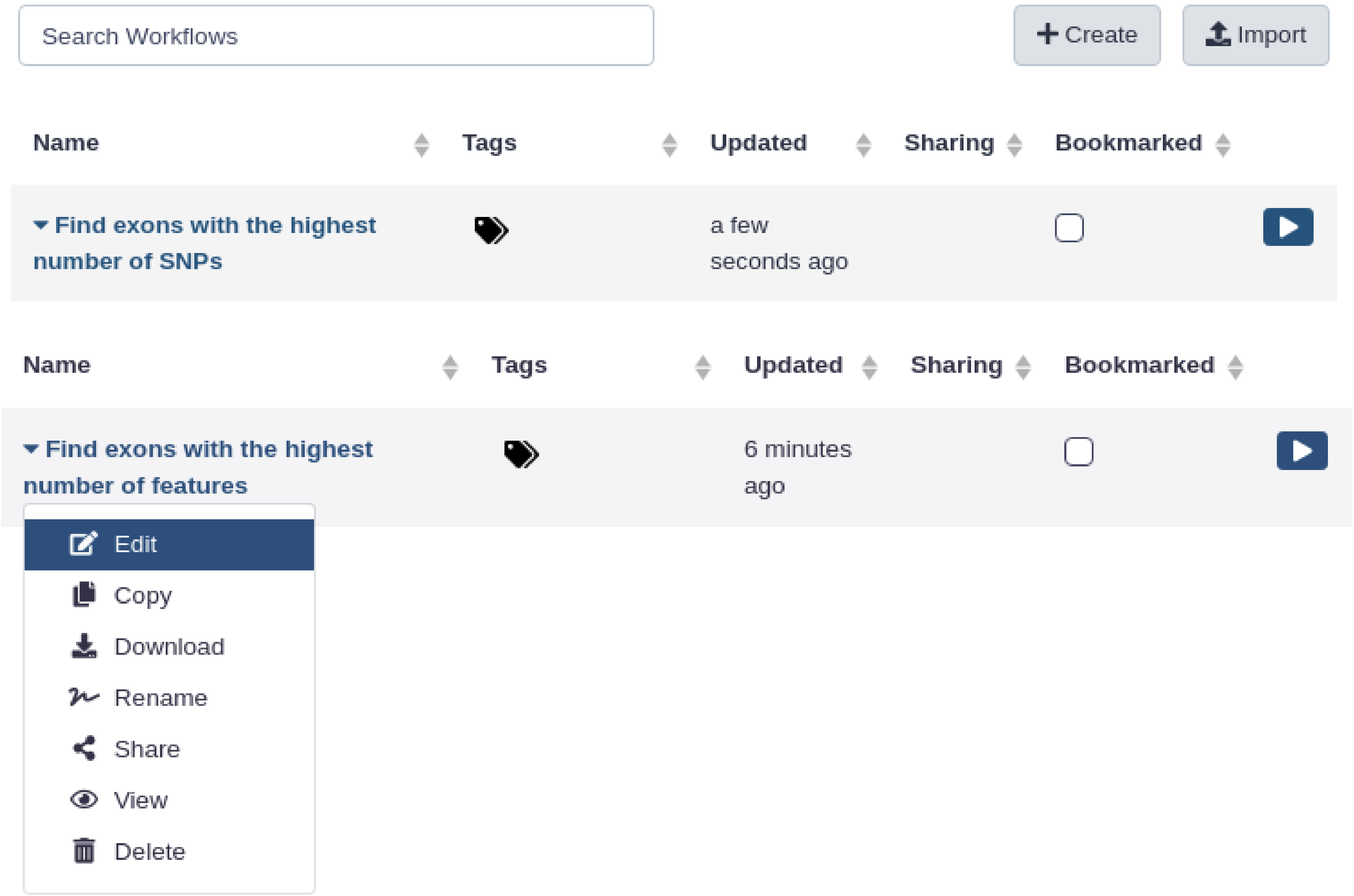
Workflow name

Find exons with the highest number of SNPs

Tool		History items created
UCSC Main *This tool cannot be used in workflows*	▶	**1 Exons** ☑ Treat as input dataset Exons
UCSC Main *This tool cannot be used in workflows*	▶	**2 SNPs** ☑ Treat as input dataset SNPs
Join ☑ Include "Join" in workflow	▶	**3 Join on data 2 and data 1**
Group ☑ Include "Group" in workflow	▶	**4 Group on data 3**
Sort ☑ Include "Sort" in workflow	▶	**5 Sort on data 4**
Select first ☑ Include "Select first" in workflow	▶	**6 Select first on data 5**
Compare two Datasets ☑ Include "Compare two Datasets" in workflow	▶	**7 Compare two Datasets on data 6 and data 1**



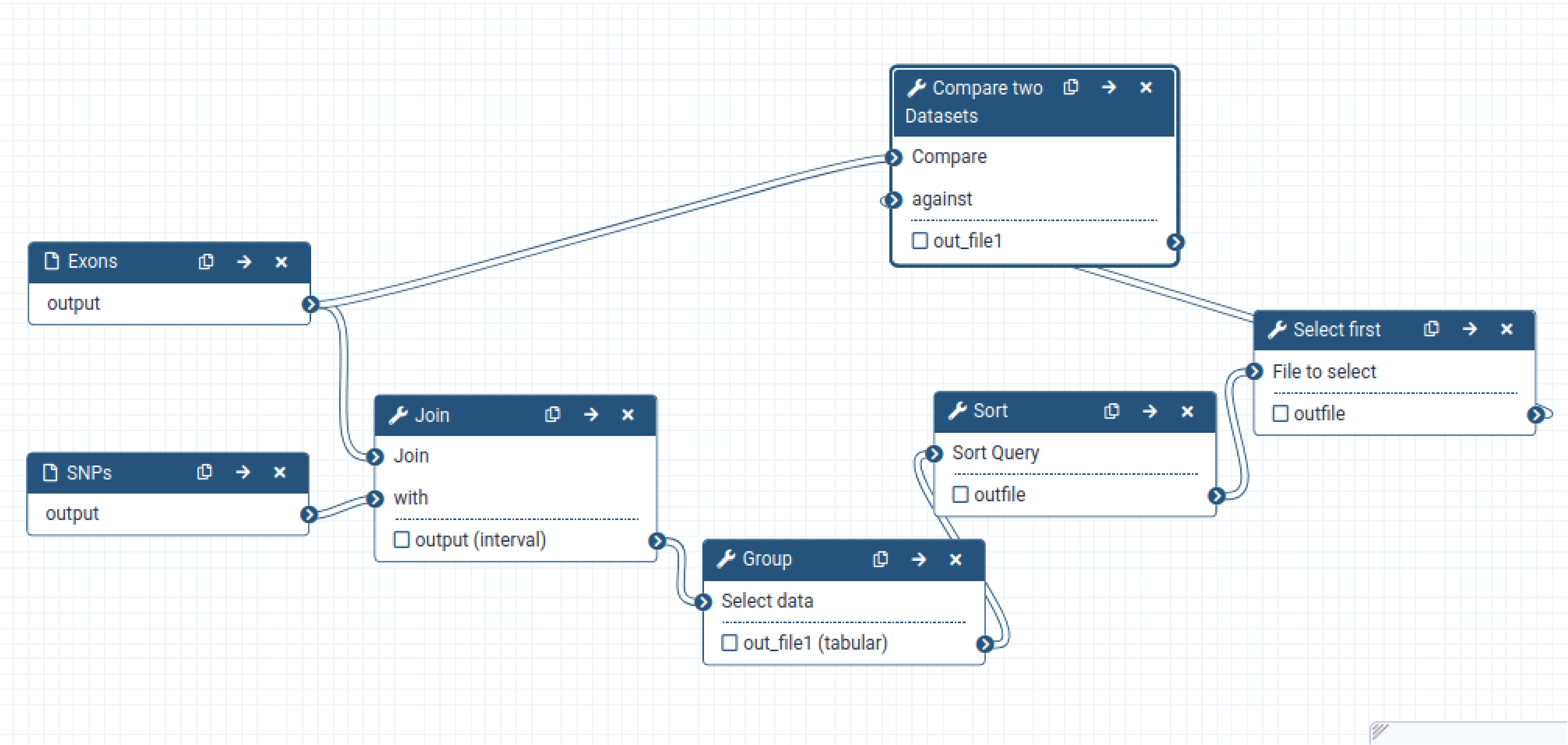
Name	Tags	Updated	Sharing	Bookmarked	
▾ Find exons with the highest number of SNPs		a few seconds ago		☐	▶

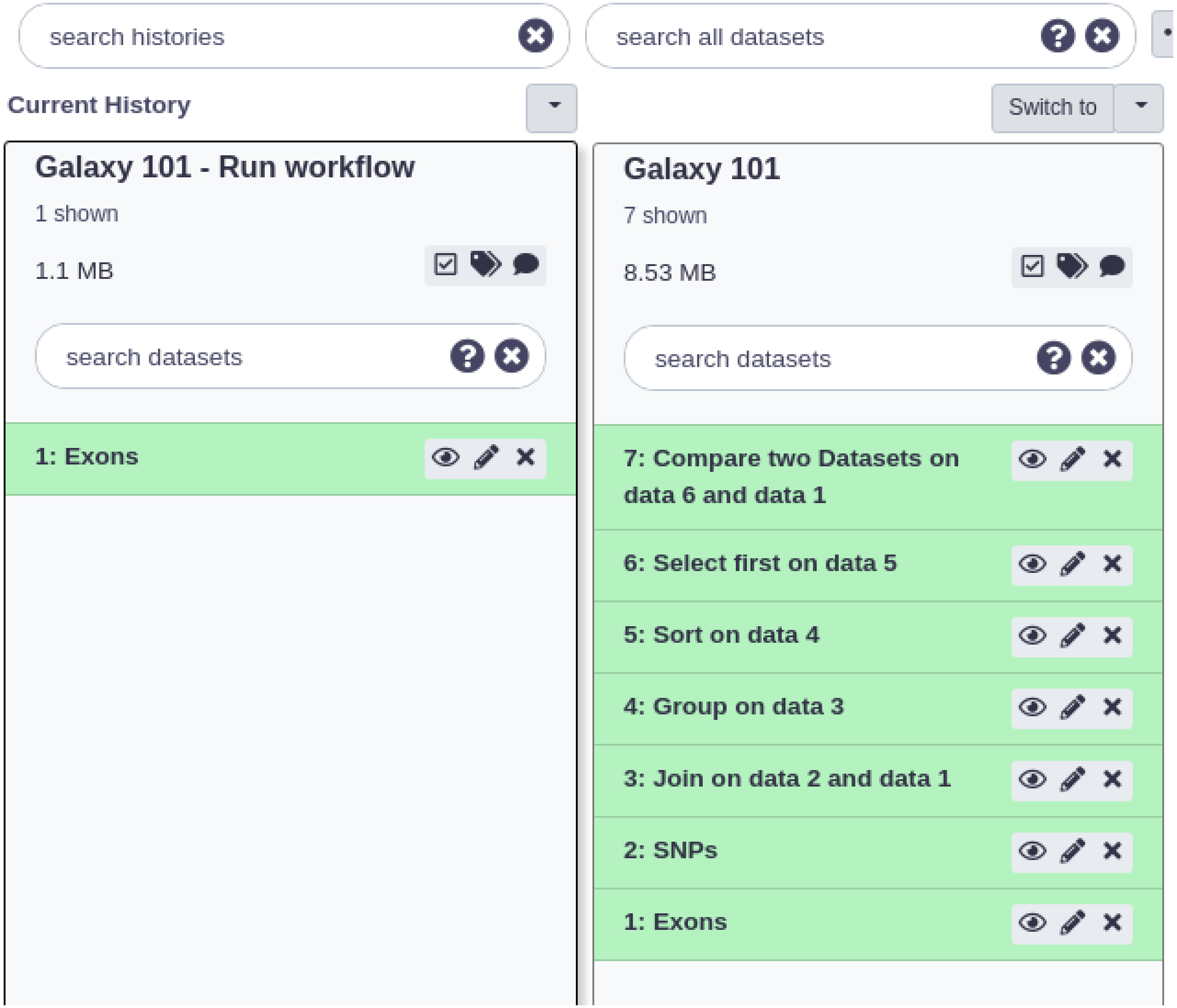


Name	Tags	Updated	Sharing	Bookmarked	
Find exons with the highest number of SNPs		a few seconds ago		□	▶

Name	Tags	Updated	Sharing	Bookmarked	
Find exons with the highest number of features		6 minutes ago		□	▶

Galaxy


Compare two Datasets
Compare
against
□ out_file1
Exons
output
Select first
File to select
□ outfile
Join
Join
with
□ output (interval)
Sort
Sort Query
□ outfile
SNPs
output
Group
Select data
□ out_file1 (tabular)



Galaxy 101 - Run workflow

1 shown

1.1 MB

1: Exons

Galaxy 101

7 shown

8.53 MB

7: Compare two Datasets on data 6 and data 1

6: Select first on data 5

5: Sort on data 4

4: Group on data 3

3: Join on data 2 and data 1

2: SNPs

1: Exons

Share your work

To share a history, click on the galaxy- **gear icon** in the history panel and select **Share or Publish**. On this page you can do 3 things:

Make History Accessible via Link. This generates a link that you can give out to others. Anybody with this link will be able to view your history.

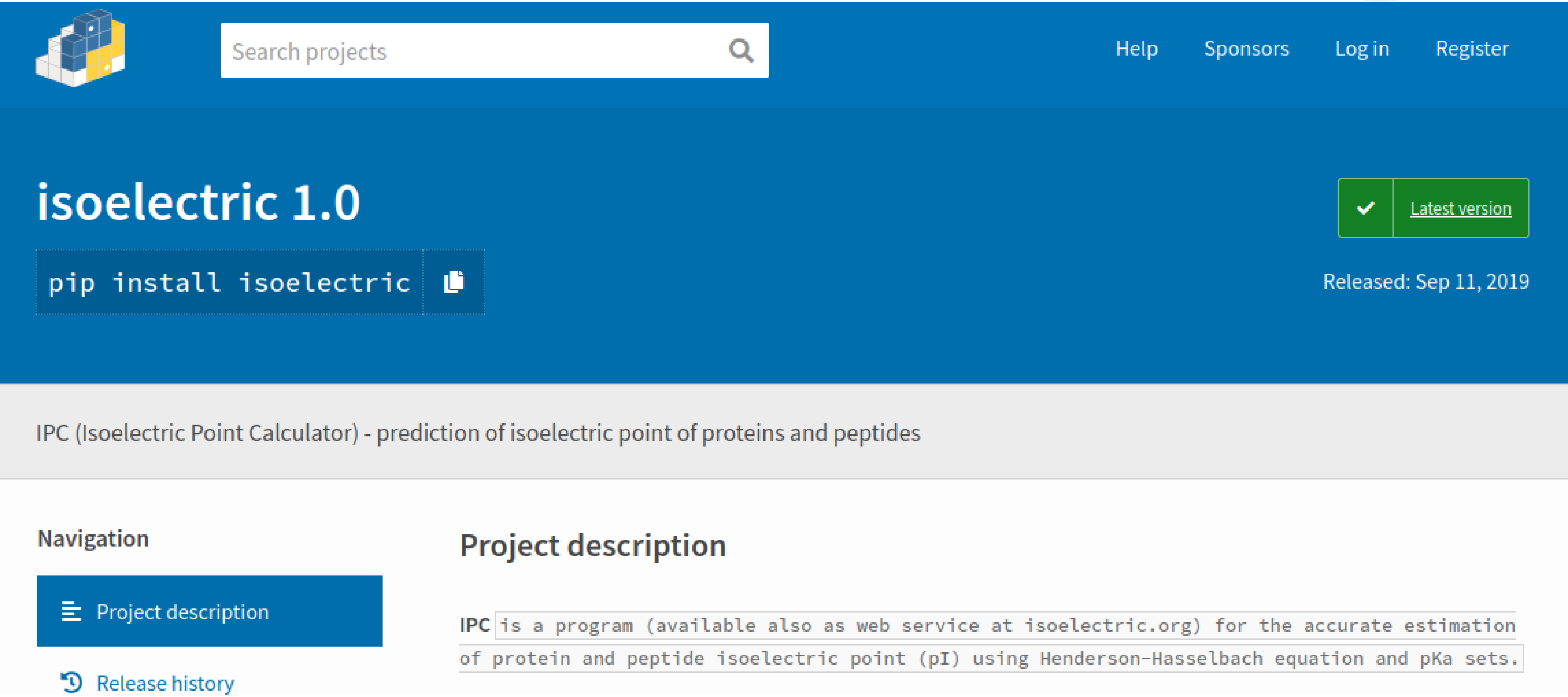
Make History Accessible and Publish. This will not only create a link, but will also publish your history. This means your history will be listed under Shared Data → Histories in the top menu.

Share with a user. This will share the history only with specific users on the Galaxy instance.

Share your work

Share your work

Python PyPi



isoelectric 1.0

```
pip install isoelectric
```

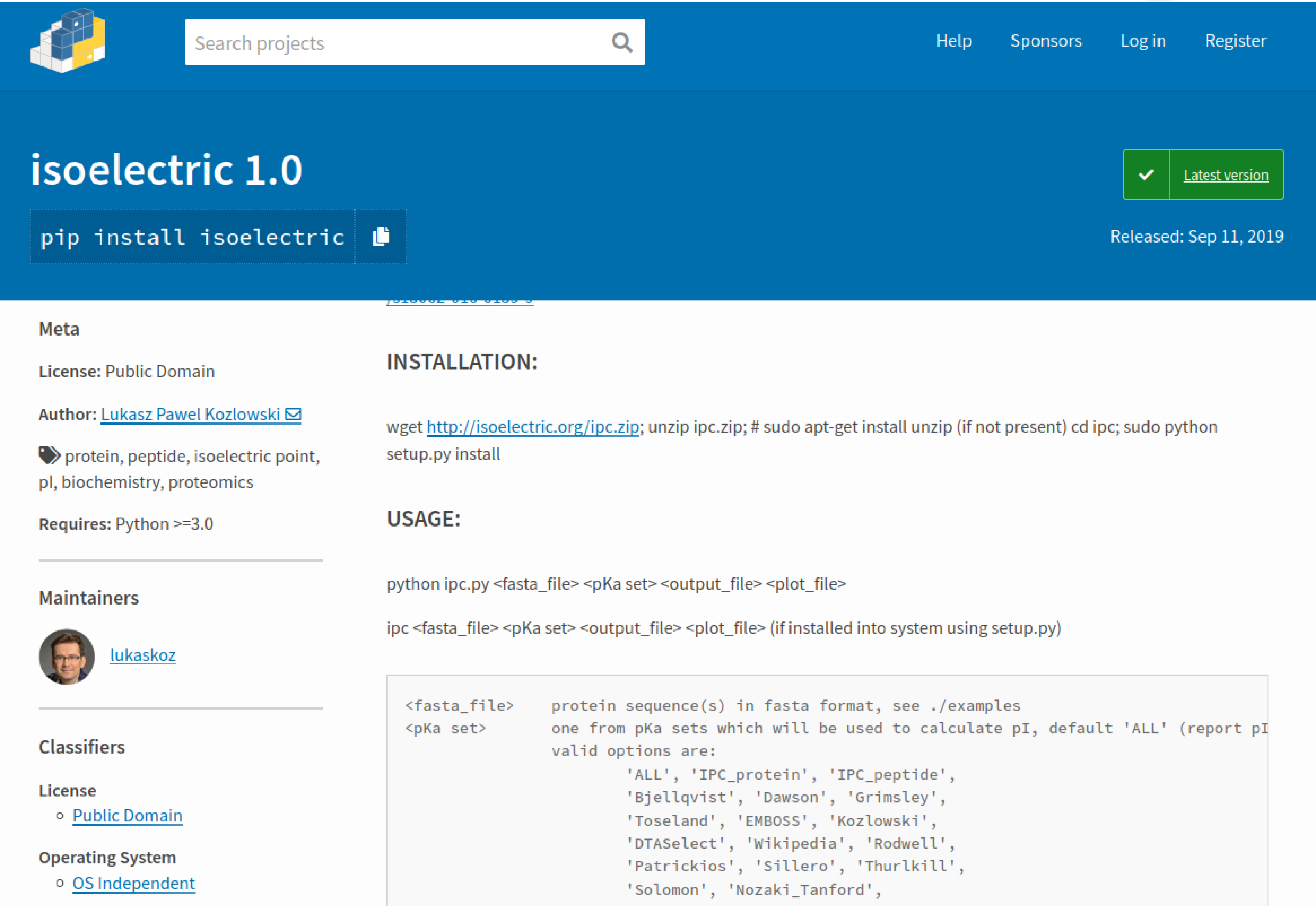
Released: Sep 11, 2019

IPC (Isoelectric Point Calculator) - prediction of isoelectric point of proteins and peptides

Project description

IPC `is a program (available also as web service at isoelectric.org) for the accurate estimation of protein and peptide isoelectric point (pI) using Henderson-Hasselbach equation and pKa sets.`

Python PyPi



isoelectric 1.0

```
pip install isoelectric
```

Released: Sep 11, 2019

/s13062-016-0159-9

Meta

License: Public Domain

Author: Lukasz Pawel Kozlowski

protein, peptide, isoelectric point, pI, biochemistry, proteomics

Requires: Python >=3.0

Maintainers

lukaskoz

Classifiers

License

- Public Domain

Operating System

- OS Independent

INSTALLATION:

wget http://isoelectric.org/ipc.zip; unzip ipc.zip; # sudo apt-get install unzip (if not present) cd ipc; sudo python setup.py install

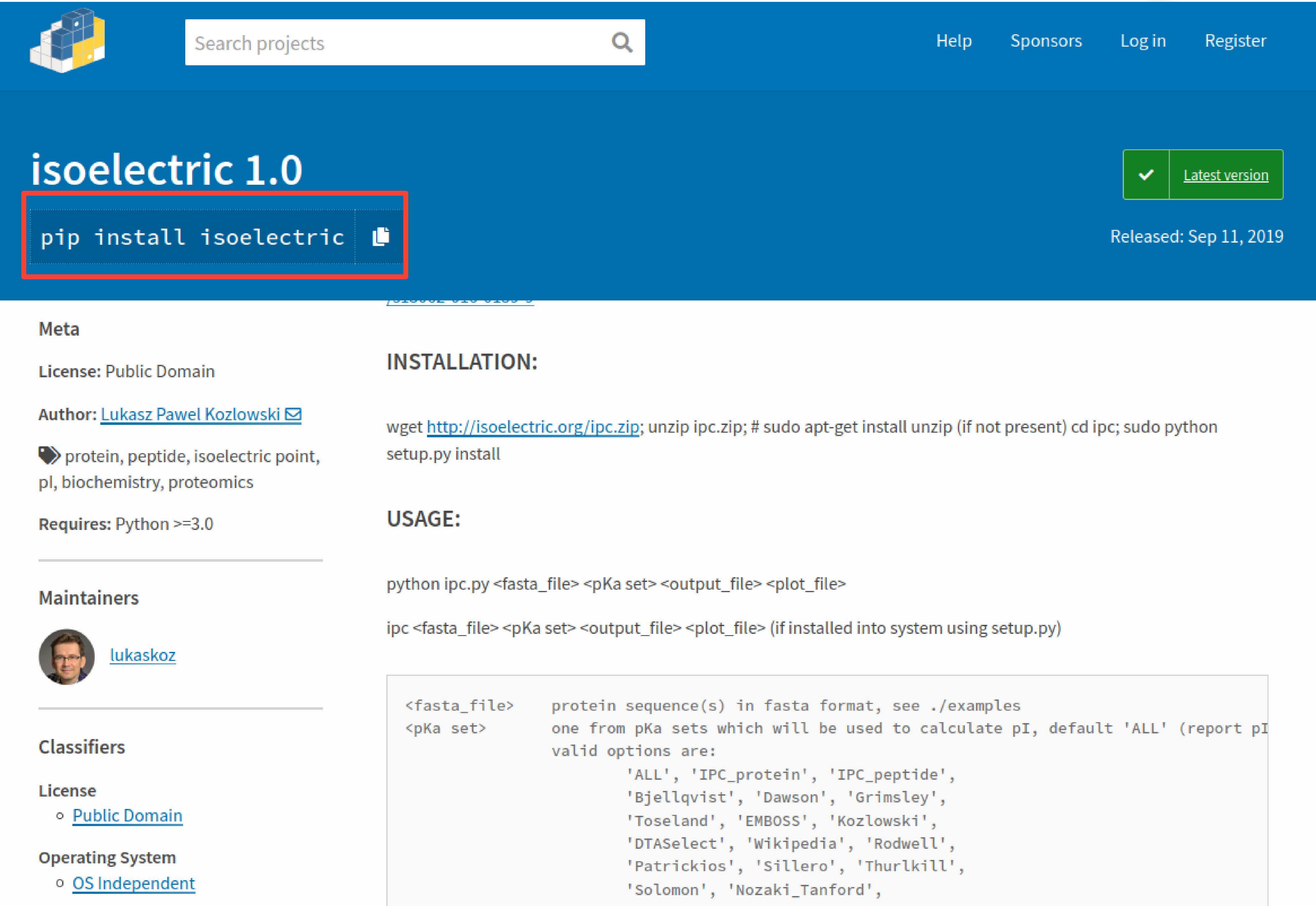
USAGE:

python ipc.py <fasta_file> <pKa set> <output_file> <plot_file>

ipc <fasta_file> <pKa set> <output_file> <plot_file> (if installed into system using setup.py)

```
<fasta_file>    protein sequence(s) in fasta format, see ./examples
<pKa set>       one from pKa sets which will be used to calculate pI, default 'ALL' (report pI
                valid options are:
                        'ALL', 'IPC_protein', 'IPC_peptide',
                        'Bjellqvist', 'Dawson', 'Grimsley',
                        'Toseland', 'EMBOSS', 'Kozlowski',
                        'DTASelect', 'Wikipedia', 'Rodwell',
                        'Patrickios', 'Sillero', 'Thurlkill',
                        'Solomon', 'Nozaki_Tanford',
```

Python PyPi



isoelectric 1.0

```
pip install isoelectric
```

Released: Sep 11, 2019

Meta

License: Public Domain

Author: Lukasz Pawel Kozlowski

protein, peptide, isoelectric point, pI, biochemistry, proteomics

Requires: Python >=3.0

Maintainers

lukaskoz

Classifiers

License

- Public Domain

Operating System

- OS Independent

INSTALLATION:

wget http://isoelectric.org/ipc.zip; unzip ipc.zip; # sudo apt-get install unzip (if not present) cd ipc; sudo python setup.py install

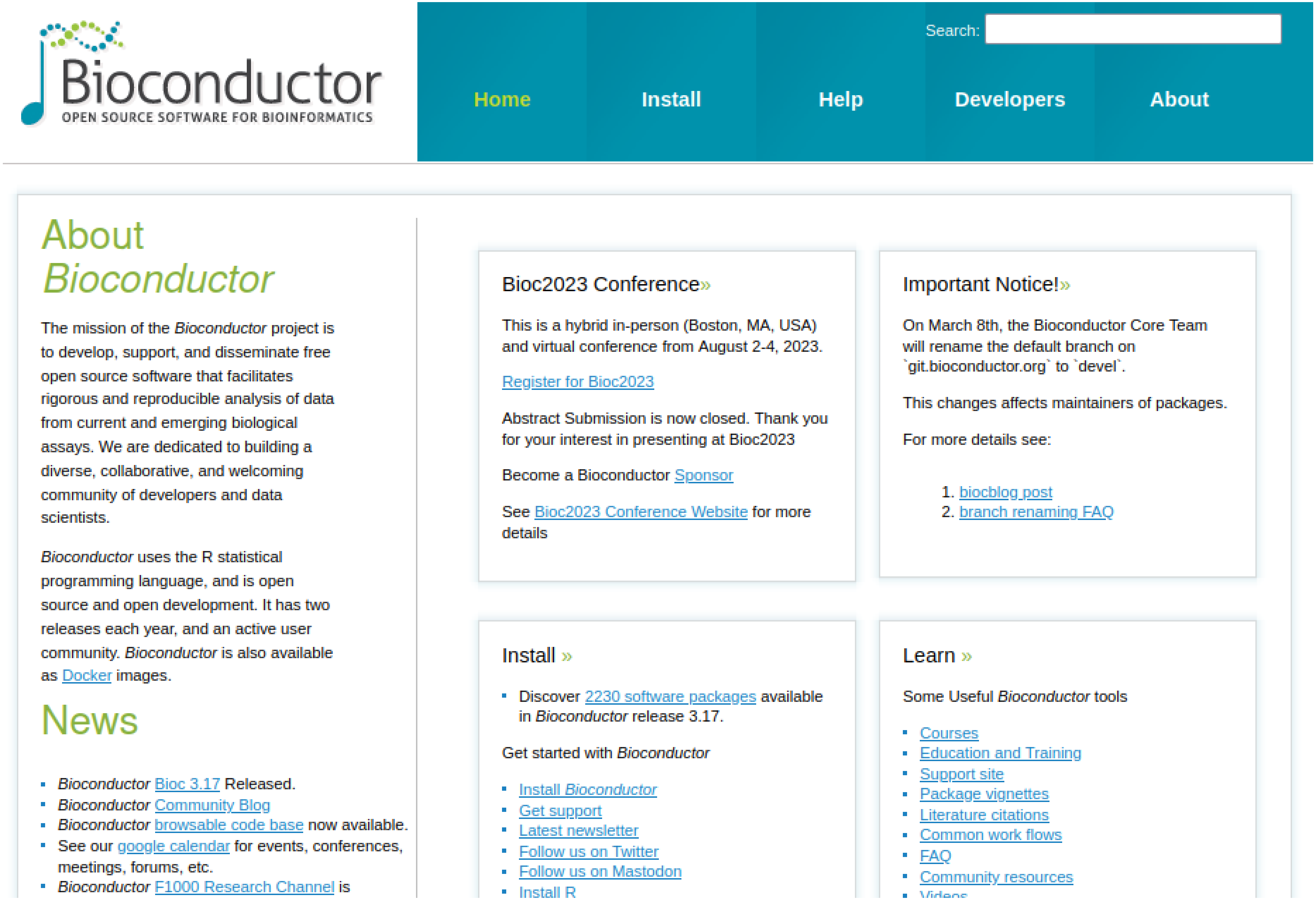
USAGE:

python ipc.py <fasta_file> <pKa set> <output_file> <plot_file>

ipc <fasta_file> <pKa set> <output_file> <plot_file> (if installed into system using setup.py)

```
<fasta_file>    protein sequence(s) in fasta format, see ./examples
<pKa set>       one from pKa sets which will be used to calculate pI, default 'ALL' (report pI
                valid options are:
                        'ALL', 'IPC_protein', 'IPC_peptide',
                        'Bjellqvist', 'Dawson', 'Grimsley',
                        'Toseland', 'EMBOSS', 'Kozlowski',
                        'DTASelect', 'Wikipedia', 'Rodwell',
                        'Patrickios', 'Sillero', 'Thurlkill',
                        'Solomon', 'Nozaki_Tanford',
```

R Bioconductor



About *Bioconductor*

The mission of the *Bioconductor* project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. We are dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. *Bioconductor* is also available as Docker images.

News

- *Bioconductor* Bioc 3.17 Released.
- *Bioconductor* Community Blog
- *Bioconductor* browsable code base now available.
- See our google calendar for events, conferences, meetings, forums, etc.
- *Bioconductor* F1000 Research Channel is

Bioc2023 Conference»

This is a hybrid in-person (Boston, MA, USA) and virtual conference from August 2-4, 2023.

Register for Bioc2023

Abstract Submission is now closed. Thank you for your interest in presenting at Bioc2023

Become a Bioconductor Sponsor

See Bioc2023 Conference Website for more details

Important Notice!»

On March 8th, the Bioconductor Core Team will rename the default branch on `git.bioconductor.org` to `devel`.

This changes affects maintainers of packages.

For more details see:

1. biocblog post
2. branch renaming FAQ

Install »

- Discover 2230 software packages available in *Bioconductor* release 3.17.

Get started with *Bioconductor*

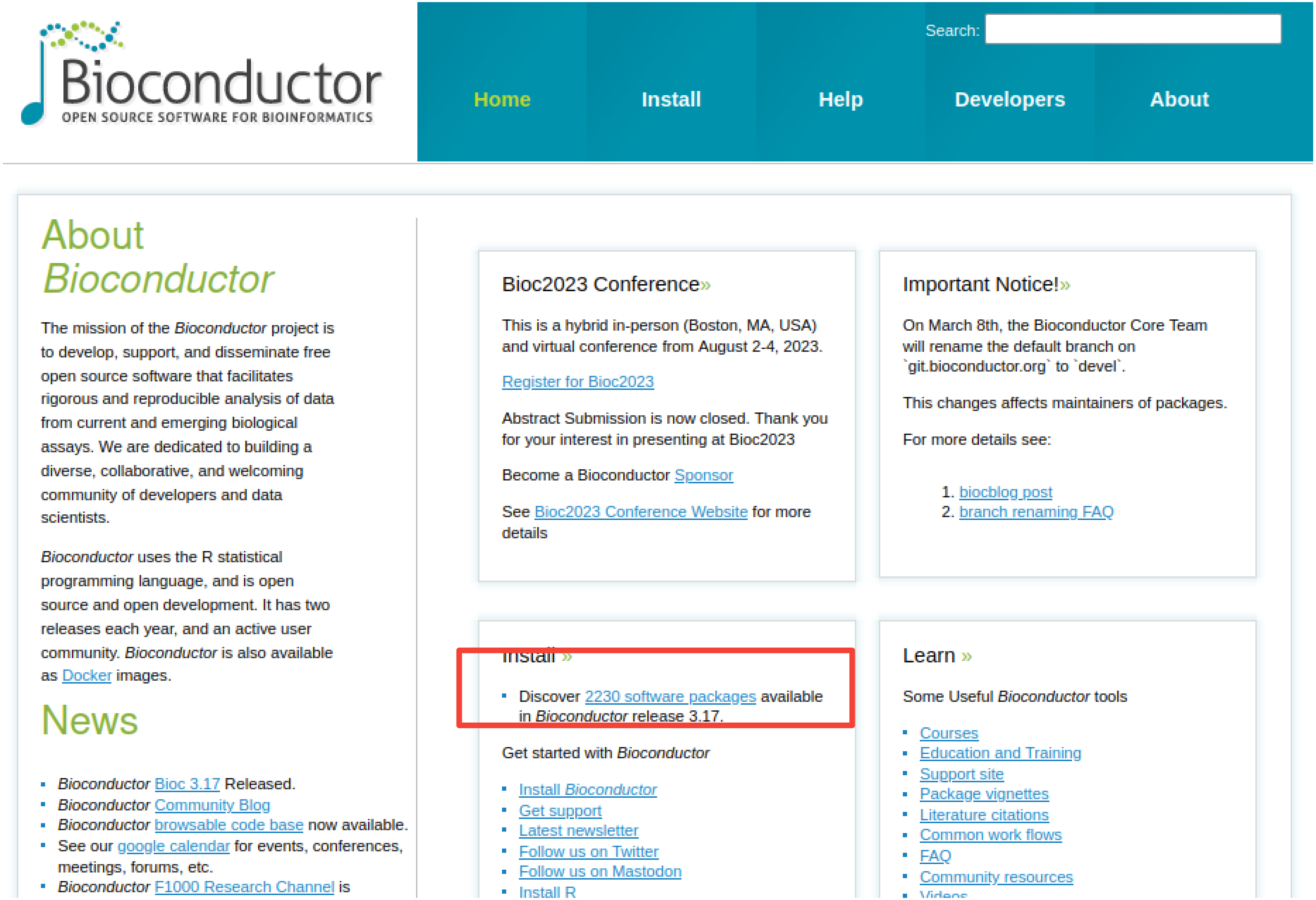
- Install *Bioconductor*
- Get support
- Latest newsletter
- Follow us on Twitter
- Follow us on Mastodon
- Install R

Learn »

Some Useful *Bioconductor* tools

- Courses
- Education and Training
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

R Bioconductor



About *Bioconductor*

The mission of the *Bioconductor* project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. We are dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. *Bioconductor* is also available as Docker images.

News

- *Bioconductor* Bioc 3.17 Released.
- *Bioconductor* Community Blog
- *Bioconductor* browsable code base now available.
- See our google calendar for events, conferences, meetings, forums, etc.
- *Bioconductor* F1000 Research Channel is

Bioc2023 Conference»

This is a hybrid in-person (Boston, MA, USA) and virtual conference from August 2-4, 2023.

Register for Bioc2023

Abstract Submission is now closed. Thank you for your interest in presenting at Bioc2023

Become a Bioconductor Sponsor

See Bioc2023 Conference Website for more details

Important Notice!»

On March 8th, the Bioconductor Core Team will rename the default branch on \`git.bioconductor.org\` to \`devel\`.

This changes affects maintainers of packages.

For more details see:

1. biocblog post
2. branch renaming FAQ

Install »

- Discover 2230 software packages available in *Bioconductor* release 3.17.

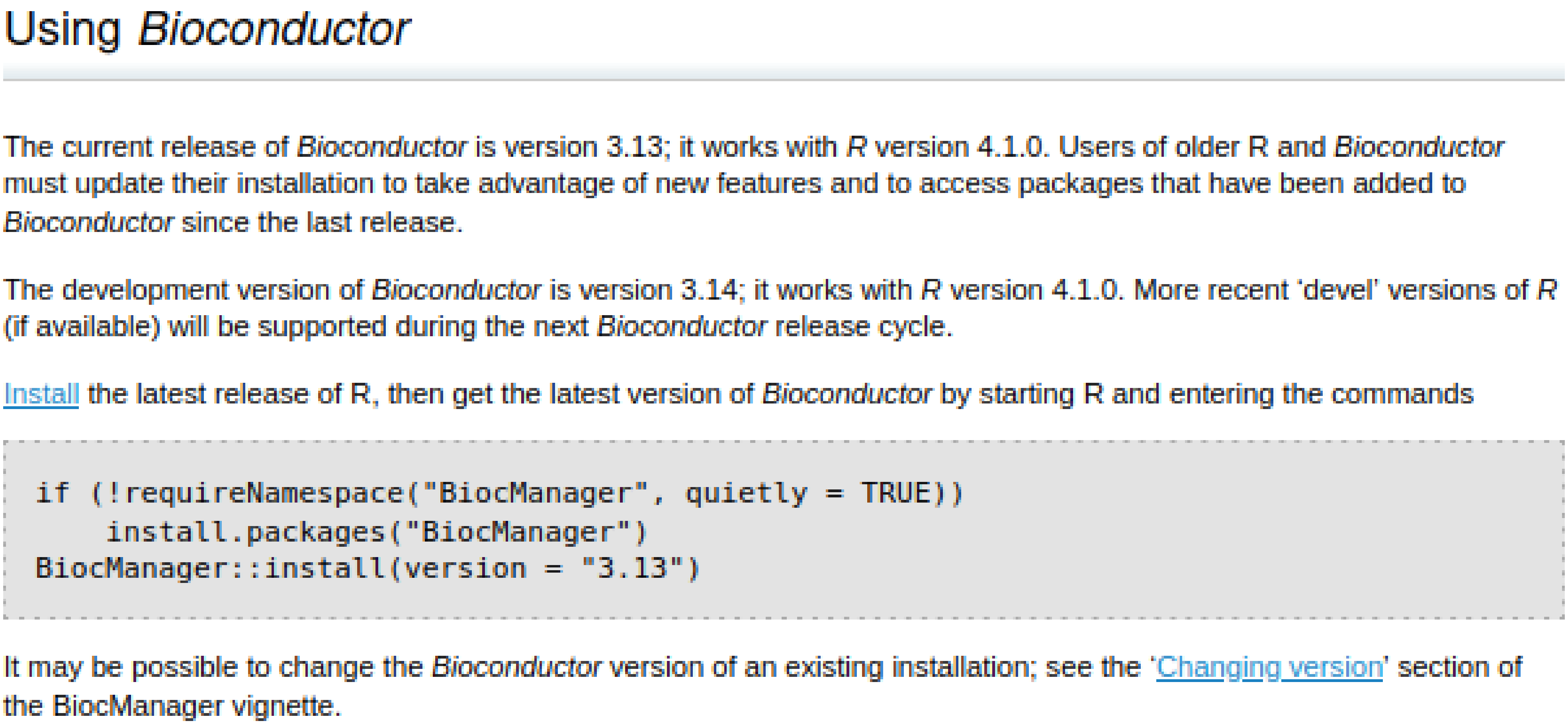
Get started with *Bioconductor*

- Install *Bioconductor*
- Get support
- Latest newsletter
- Follow us on Twitter
- Follow us on Mastodon
- Install R

Learn »

Some Useful *Bioconductor* tools

- Courses
- Education and Training
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos



Using *Bioconductor*

The current release of *Bioconductor* is version 3.13; it works with *R* version 4.1.0. Users of older R and *Bioconductor* must update their installation to take advantage of new features and to access packages that have been added to *Bioconductor* since the last release.

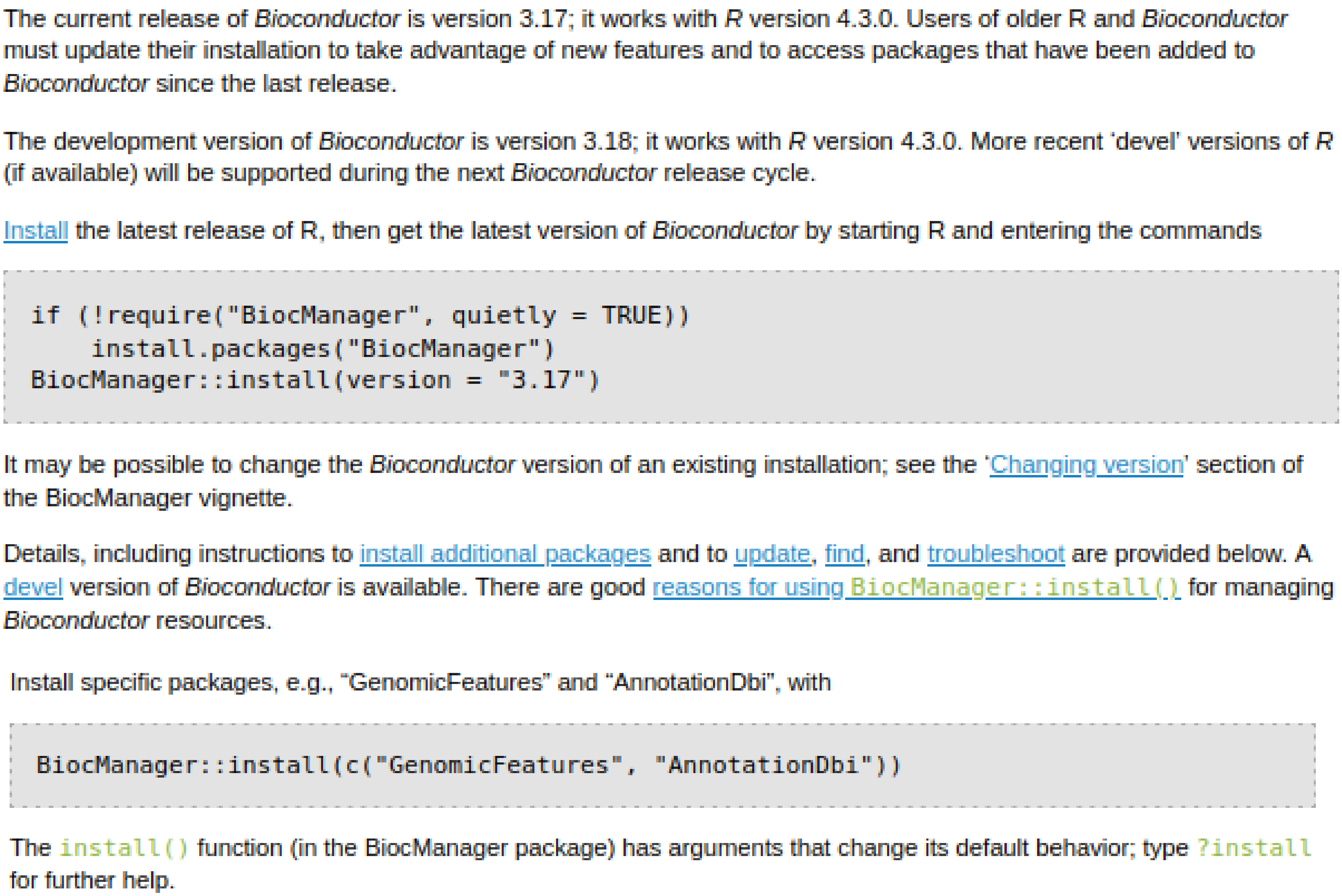
The development version of *Bioconductor* is version 3.14; it works with *R* version 4.1.0. More recent 'devel' versions of *R* (if available) will be supported during the next *Bioconductor* release cycle.

Install the latest release of R, then get the latest version of *Bioconductor* by starting R and entering the commands

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.13")
```

It may be possible to change the *Bioconductor* version of an existing installation; see the 'Changing version' section of the BiocManager vignette.

R Bioconductor



The current release of *Bioconductor* is version 3.17; it works with *R* version 4.3.0. Users of older R and *Bioconductor* must update their installation to take advantage of new features and to access packages that have been added to *Bioconductor* since the last release.

The development version of *Bioconductor* is version 3.18; it works with *R* version 4.3.0. More recent 'devel' versions of *R* (if available) will be supported during the next *Bioconductor* release cycle.

Install the latest release of R, then get the latest version of *Bioconductor* by starting R and entering the commands

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.17")
```

It may be possible to change the *Bioconductor* version of an existing installation; see the 'Changing version' section of the BiocManager vignette.

Details, including instructions to install additional packages and to update, find, and troubleshoot are provided below. A devel version of *Bioconductor* is available. There are good reasons for using `BiocManager::install()` for managing *Bioconductor* resources.

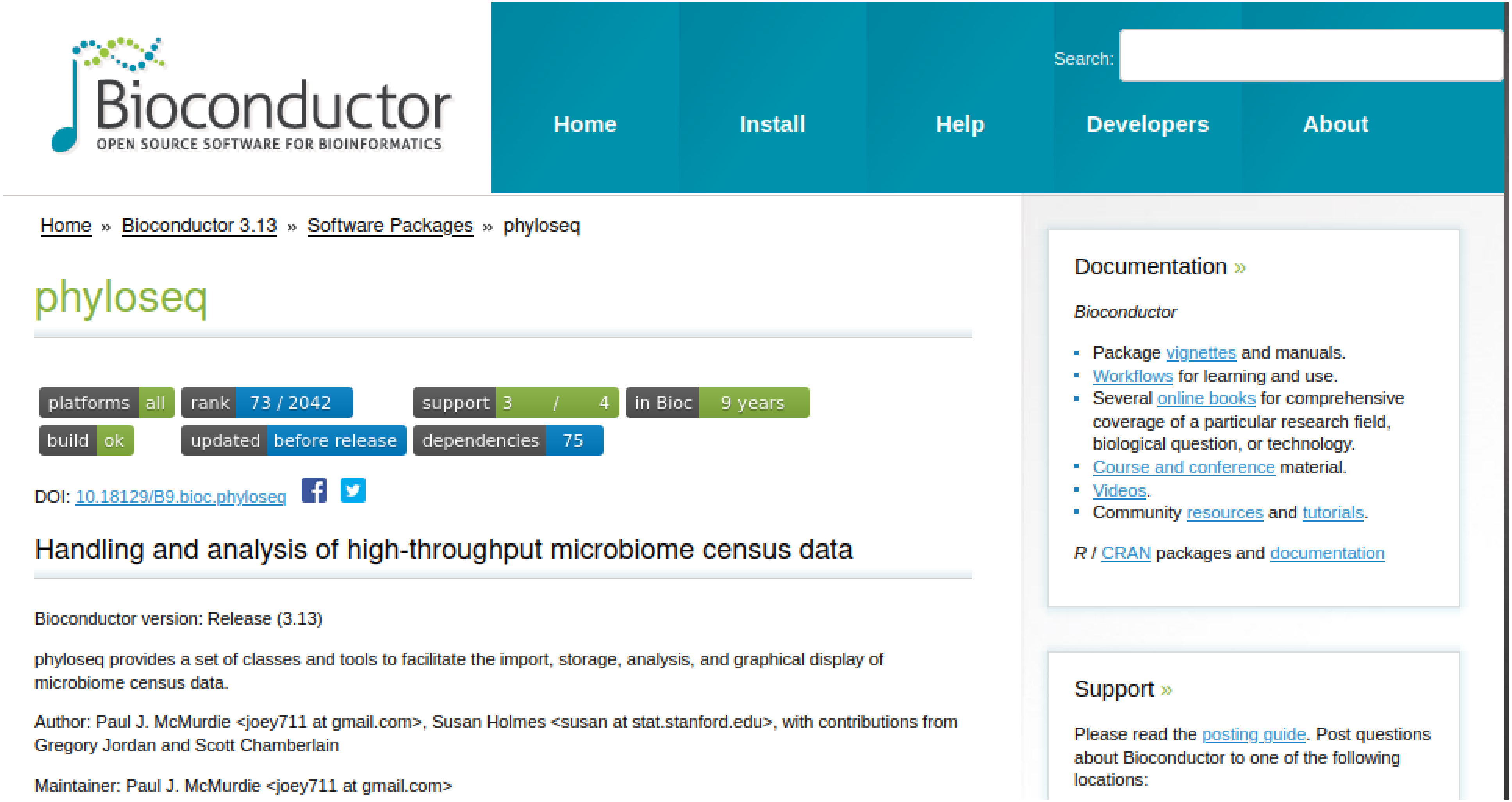
Install specific packages, e.g., "GenomicFeatures" and "AnnotationDbi", with

```
BiocManager::install(c("GenomicFeatures", "AnnotationDbi"))
```

The `install()` function (in the BiocManager package) has arguments that change its default behavior; type `?install` for further help.

https://www.bioconductor.org/install/

R Bioconductor



Home » Bioconductor 3.13 » Software Packages » phyloseq

phyloseq

DOI: 10.18129/B9.bioc.phyloseq

Handling and analysis of high-throughput microbiome census data

Bioconductor version: Release (3.13)

phyloseq provides a set of classes and tools to facilitate the import, storage, analysis, and graphical display of microbiome census data.

Author: Paul J. McMurdie <joey711 at gmail.com>, Susan Holmes <susan at stat.stanford.edu>, with contributions from Gregory Jordan and Scott Chamberlain

Maintainer: Paul J. McMurdie <joey711 at gmail.com>

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Several online books for comprehensive coverage of a particular research field, biological question, or technology.
- Course and conference material.
- Videos.
- Community resources and tutorials.

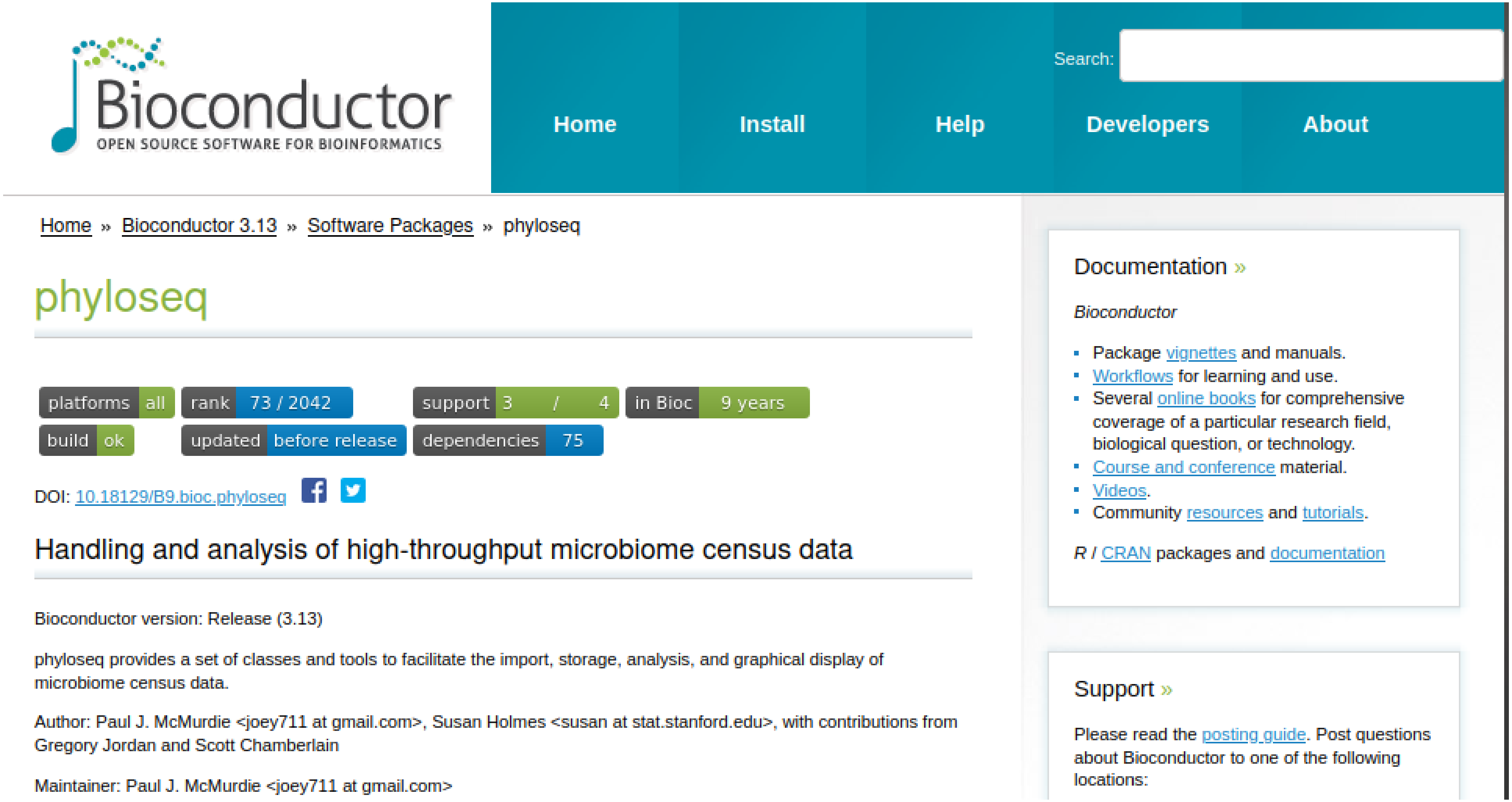
R / CRAN packages and documentation

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

R Bioconductor

ADP



Home » Bioconductor 3.13 » Software Packages » phyloseq

phyloseq

platforms all | rank 73 / 2042 | support 3 / 4 | in Bioc 9 years
build ok | updated before release | dependencies 75

DOI: 10.18129/B9.bioc.phyloseq

Handling and analysis of high-throughput microbiome census data

Bioconductor version: Release (3.13)

phyloseq provides a set of classes and tools to facilitate the import, storage, analysis, and graphical display of microbiome census data.

Author: Paul J. McMurdie <joey711 at gmail.com>, Susan Holmes <susan at stat.stanford.edu>, with contributions from Gregory Jordan and Scott Chamberlain

Maintainer: Paul J. McMurdie <joey711 at gmail.com>

Documentation »

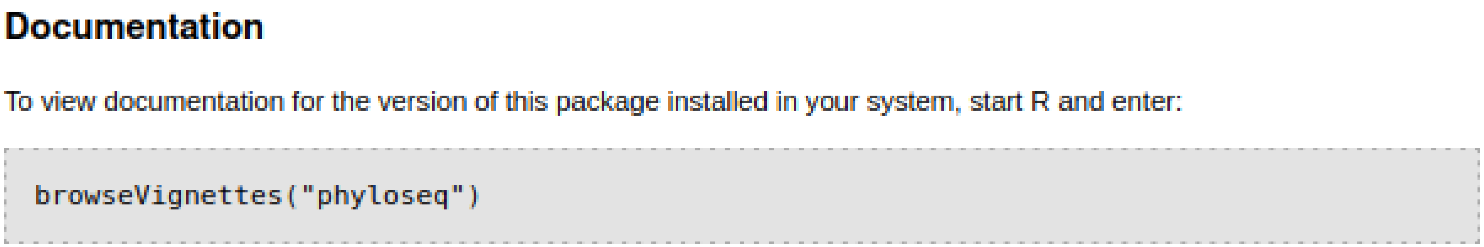
Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Several online books for comprehensive coverage of a particular research field, biological question, or technology.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

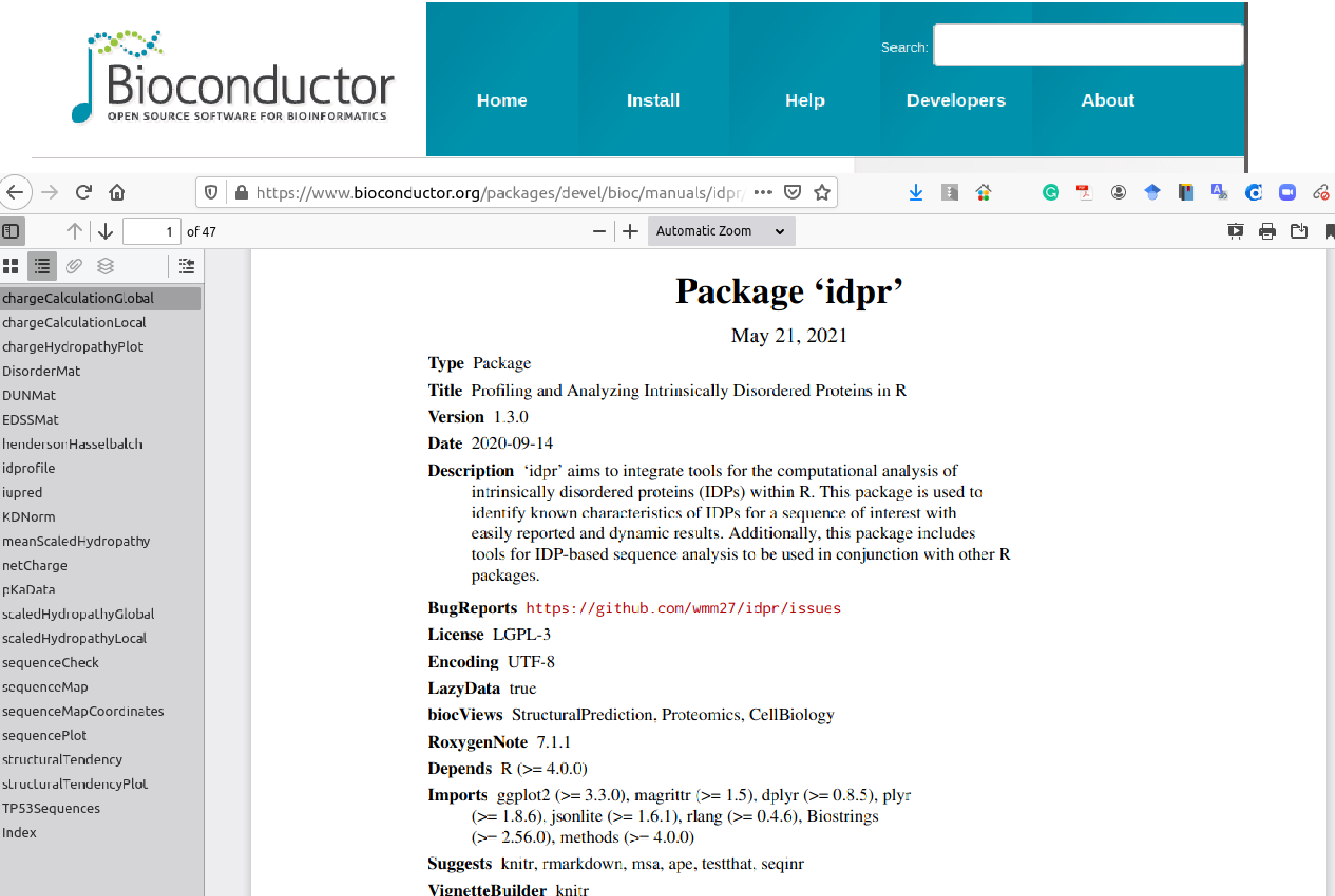


Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("phyloseq")
```

R Bioconductor



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Search:
Home
Install
Help
Developers
About
https://www.bioconductor.org/packages/devel/bioc/manuals/idpr/
1
of 47
Automatic Zoom

Package ‘idpr’

May 21, 2021

Type Package

Title Profiling and Analyzing Intrinsically Disordered Proteins in R

Version 1.3.0

Date 2020-09-14

Description ‘idpr’ aims to integrate tools for the computational analysis of intrinsically disordered proteins (IDPs) within R. This package is used to identify known characteristics of IDPs for a sequence of interest with easily reported and dynamic results. Additionally, this package includes tools for IDP-based sequence analysis to be used in conjunction with other R packages.

BugReports https://github.com/wmm27/idpr/issues

License LGPL-3

Encoding UTF-8

LazyData true

biocViews StructuralPrediction, Proteomics, CellBiology

RoxygenNote 7.1.1

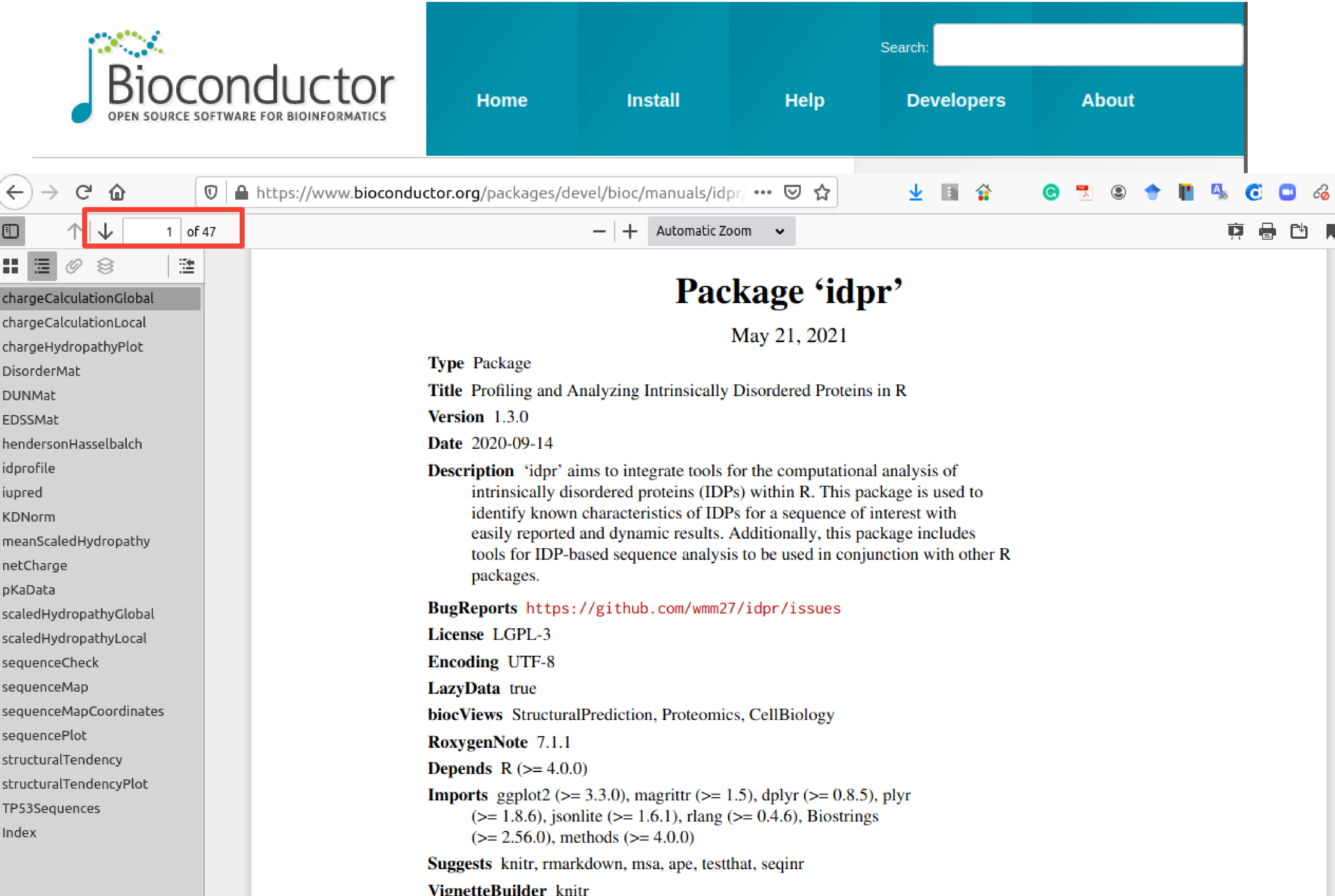
Depends R (>= 4.0.0)

Imports ggplot2 (>= 3.3.0), magrittr (>= 1.5), dplyr (>= 0.8.5), plyr (>= 1.8.6), jsonlite (>= 1.6.1), rlang (>= 0.4.6), Biostrings (>= 2.56.0), methods (>= 4.0.0)

Suggests knitr, rmarkdown, msa, ape, testthat, seqinr

VignetteBuilder knitr

R Bioconductor



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Search:
Home
Install
Help
Developers
About
https://www.bioconductor.org/packages/devel/bioc/manuals/idpr/
1
of 47
Automatic Zoom

- chargeCalculationGlobal
- chargeCalculationLocal
- chargeHydropathyPlot
- DisorderMat
- DUNMat
- EDSSMat
- hendersonHasselbalch
- idprofile
- iupred
- KDNorm
- meanScaledHydropathy
- netCharge
- pKaData
- scaledHydropathyGlobal
- scaledHydropathyLocal
- sequenceCheck
- sequenceMap
- sequenceMapCoordinates
- sequencePlot
- structuralTendency
- structuralTendencyPlot
- TP53Sequences
- Index

Package ‘idpr’

May 21, 2021

Type Package

Title Profiling and Analyzing Intrinsically Disordered Proteins in R

Version 1.3.0

Date 2020-09-14

Description ‘idpr’ aims to integrate tools for the computational analysis of intrinsically disordered proteins (IDPs) within R. This package is used to identify known characteristics of IDPs for a sequence of interest with easily reported and dynamic results. Additionally, this package includes tools for IDP-based sequence analysis to be used in conjunction with other R packages.

BugReports https://github.com/wmm27/idpr/issues

License LGPL-3

Encoding UTF-8

LazyData true

biocViews StructuralPrediction, Proteomics, CellBiology

RoxygenNote 7.1.1

Depends R (>= 4.0.0)

Imports ggplot2 (>= 3.3.0), magrittr (>= 1.5), dplyr (>= 0.8.5), plyr (>= 1.8.6), jsonlite (>= 1.6.1), rlang (>= 0.4.6), Biostrings (>= 2.56.0), methods (>= 4.0.0)

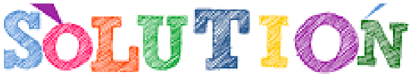
Suggests knitr, rmarkdown, msa, ape, testthat, seqinr

VignetteBuilder knitr

Scientific project timeline

General workflow in science

Problem to solve

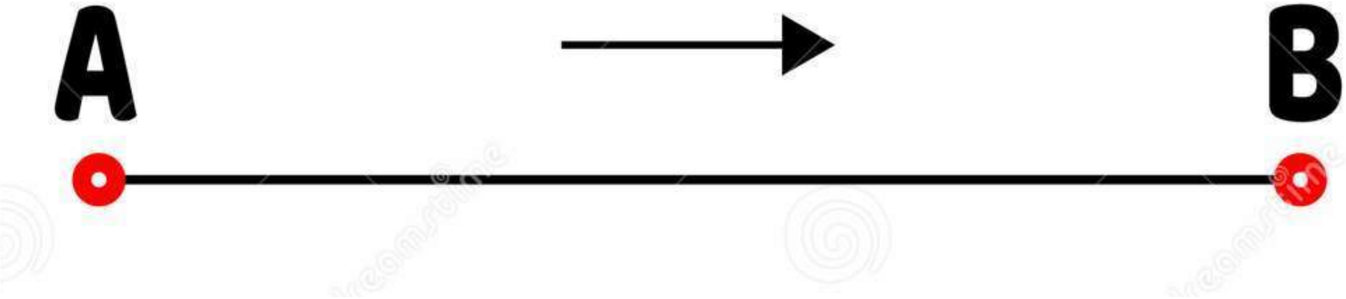


General workflow in science

Problem to solve

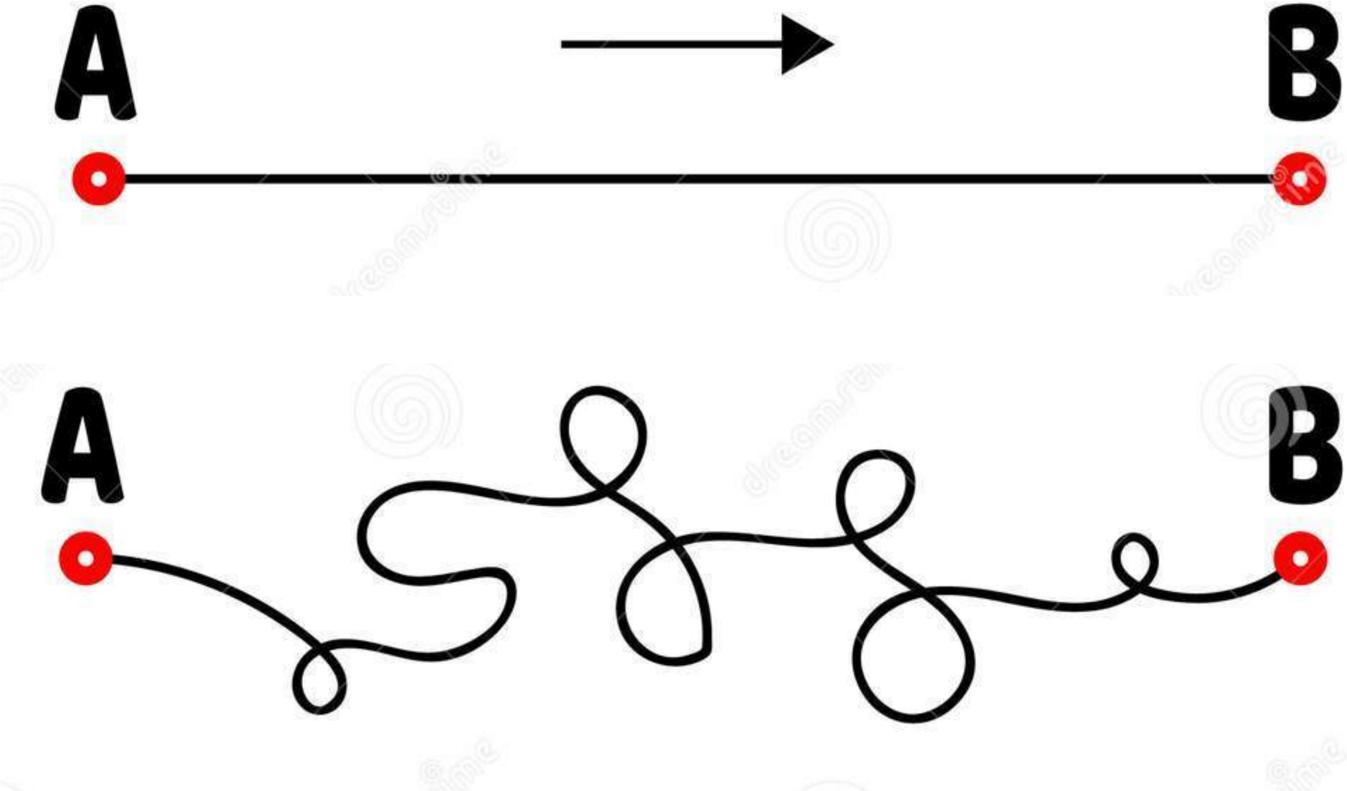
General workflow in science

Problem to solve



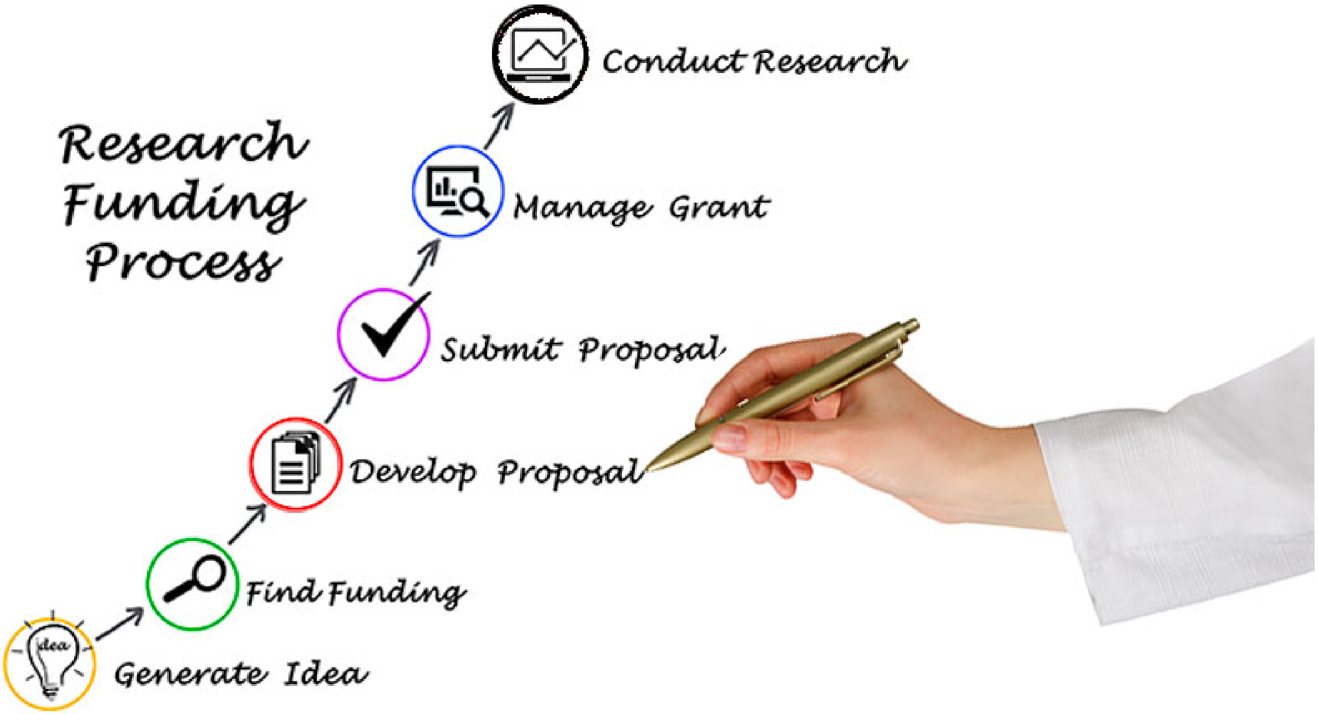
General workflow in science

Problem to solve



General workflow in science

Problem to solve



General workflow in science

Problem to solve

National
Science
Centre
Poland

The National Centre
for Research and Development

General workflow in science

Problem to solve

Pfizer

AstraZeneca

National Science Centre Poland

The National Centre for Research and Development

SEVENTH FRAMEWORK PROGRAMME

wellcome trust

BILL & MELINDA GATES foundation

HORIZON 2020

NIH National Institutes of Health

NSF National Science Foundation

MARIE CURIE ACTIONS

General workflow in science

Problem to solve

Grant assessment takes (at least) 6 months

(the bigger project, more parties involved, the longer)

Problem to solve

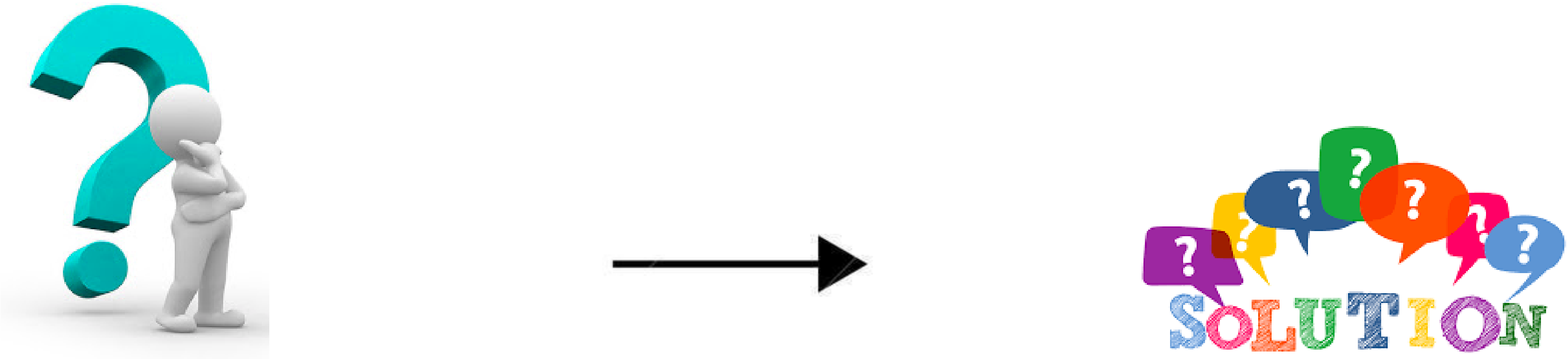
Grant assessment takes (at least) 6 months

(the bigger project, more parties involved, the longer)

Success rate is ~20%

(even if you have very good track record)

General workflow in science



Problem to solve

The research
(1-2 years, if you are lucky)

General workflow in science

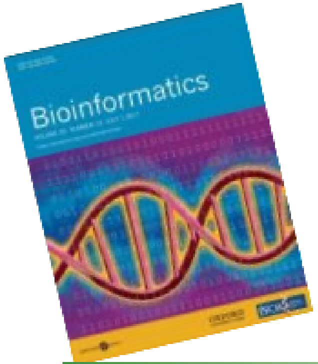
Problem to solve

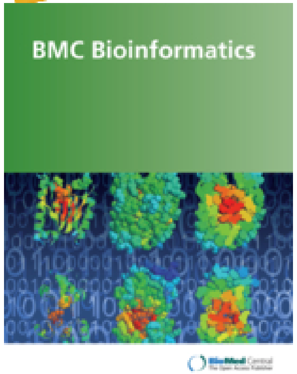
The research
(1-2 years, if you are lucky)

General workflow in science

Problem to solve

Presenting results



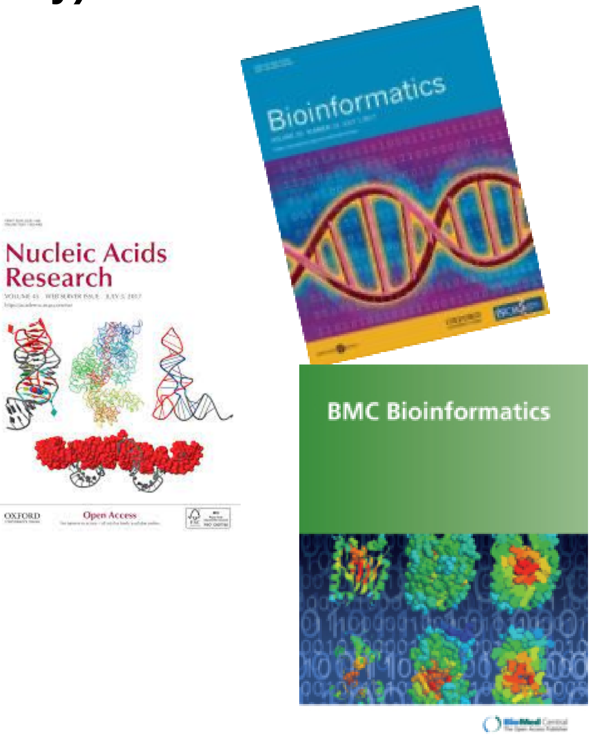


General workflow in science

Problem to solve

Presenting results

(6-12 months, if you are lucky)



General workflow in science

arXiv.org

PeerJ

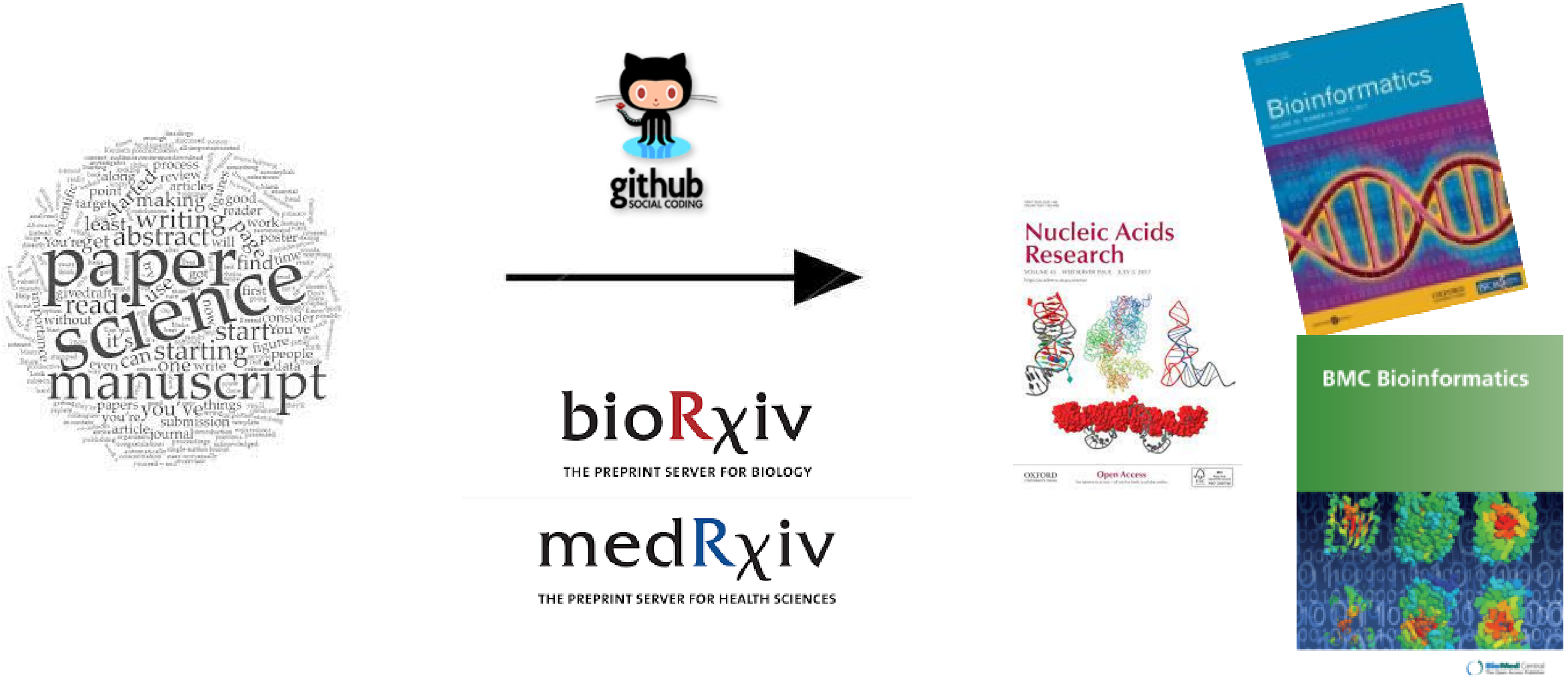
medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

SSRN
tomorrow's research today

ChemRxiv™

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

SOC ARXIV



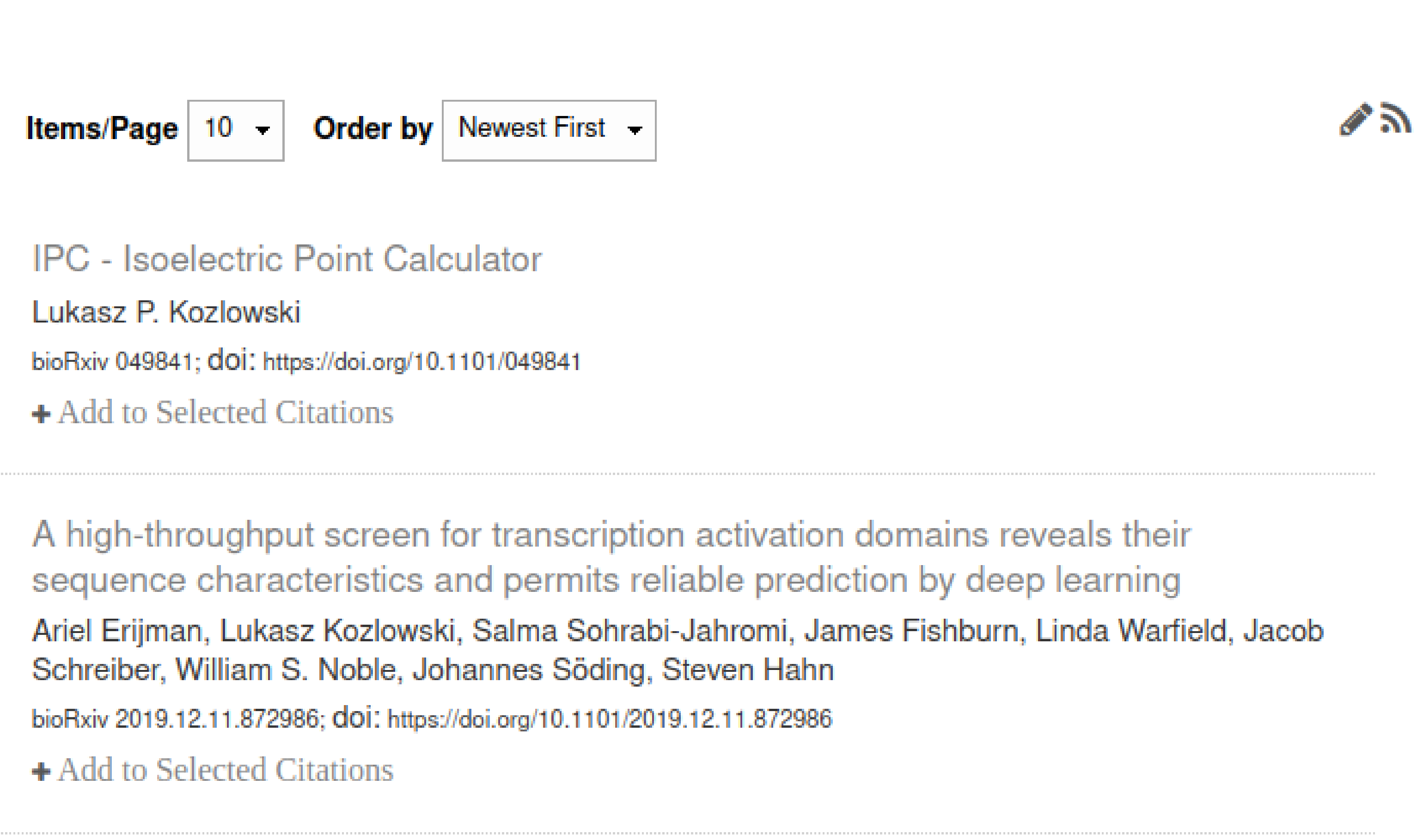
General workflow in science

CSH Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

8 Results for term "Lukasz Kozlowski"



Items/Page 10 ▾ Order by Newest First ▾

IPC - Isoelectric Point Calculator

Lukasz P. Kozlowski

bioRxiv 049841; doi: https://doi.org/10.1101/049841

+ Add to Selected Citations

A high-throughput screen for transcription activation domains reveals their sequence characteristics and permits reliable prediction by deep learning

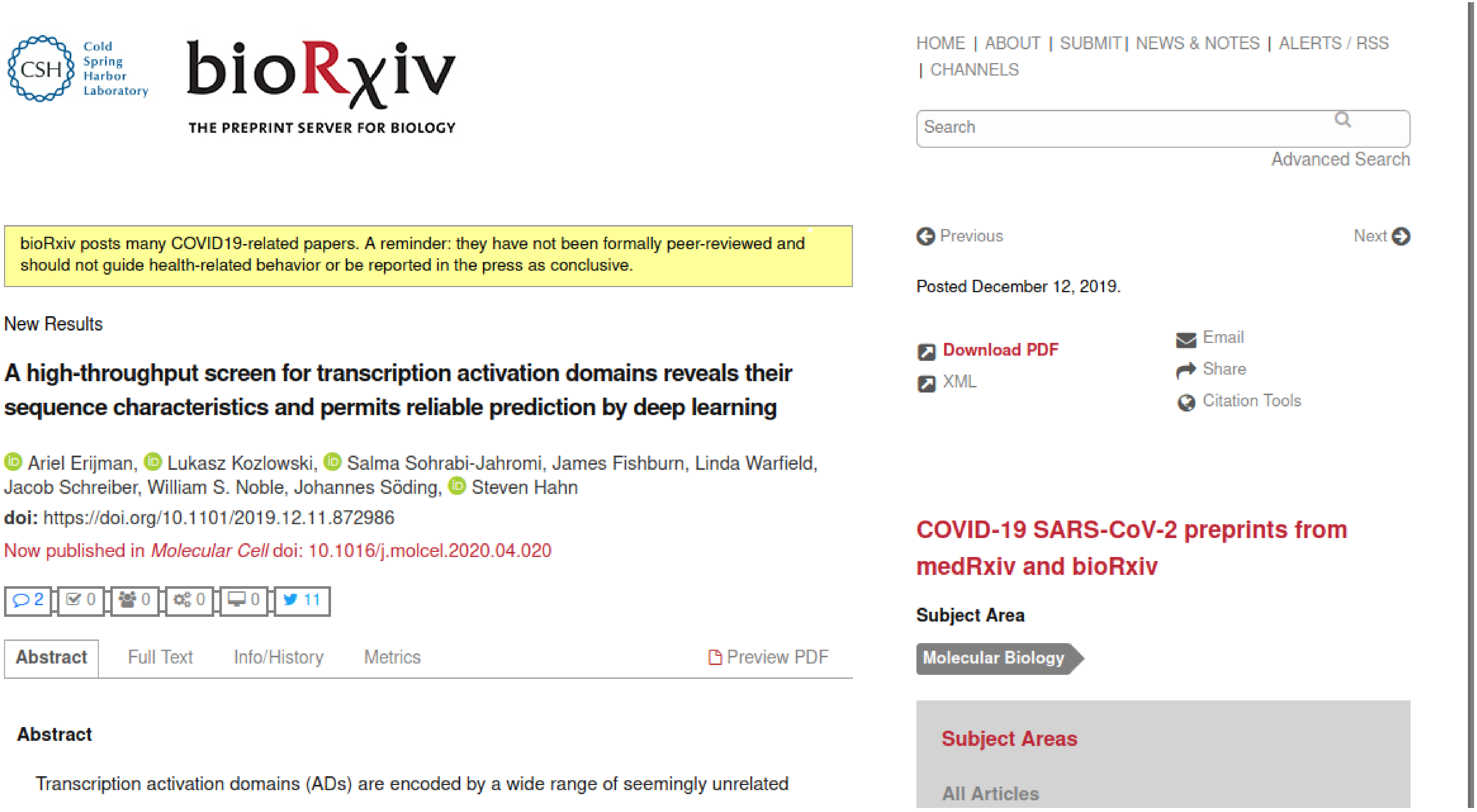
Ariel Erijman, Lukasz Kozlowski, Salma Sohrabi-Jahromi, James Fishburn, Linda Warfield, Jacob Schreiber, William S. Noble, Johannes Söding, Steven Hahn

bioRxiv 2019.12.11.872986; doi: https://doi.org/10.1101/2019.12.11.872986

+ Add to Selected Citations

General workflow in science

ADP



Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS | CHANNELS
Search
Advanced Search

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

A high-throughput screen for transcription activation domains reveals their sequence characteristics and permits reliable prediction by deep learning

Ariel Erijman, Lukasz Kozlowski, Salma Sohrabi-Jahromi, James Fishburn, Linda Warfield, Jacob Schreiber, William S. Noble, Johannes Söding, Steven Hahn

doi: https://doi.org/10.1101/2019.12.11.872986

Now published in *Molecular Cell* doi: 10.1016/j.molcel.2020.04.020

Posted December 12, 2019.

COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

Subject Area

Molecular Biology

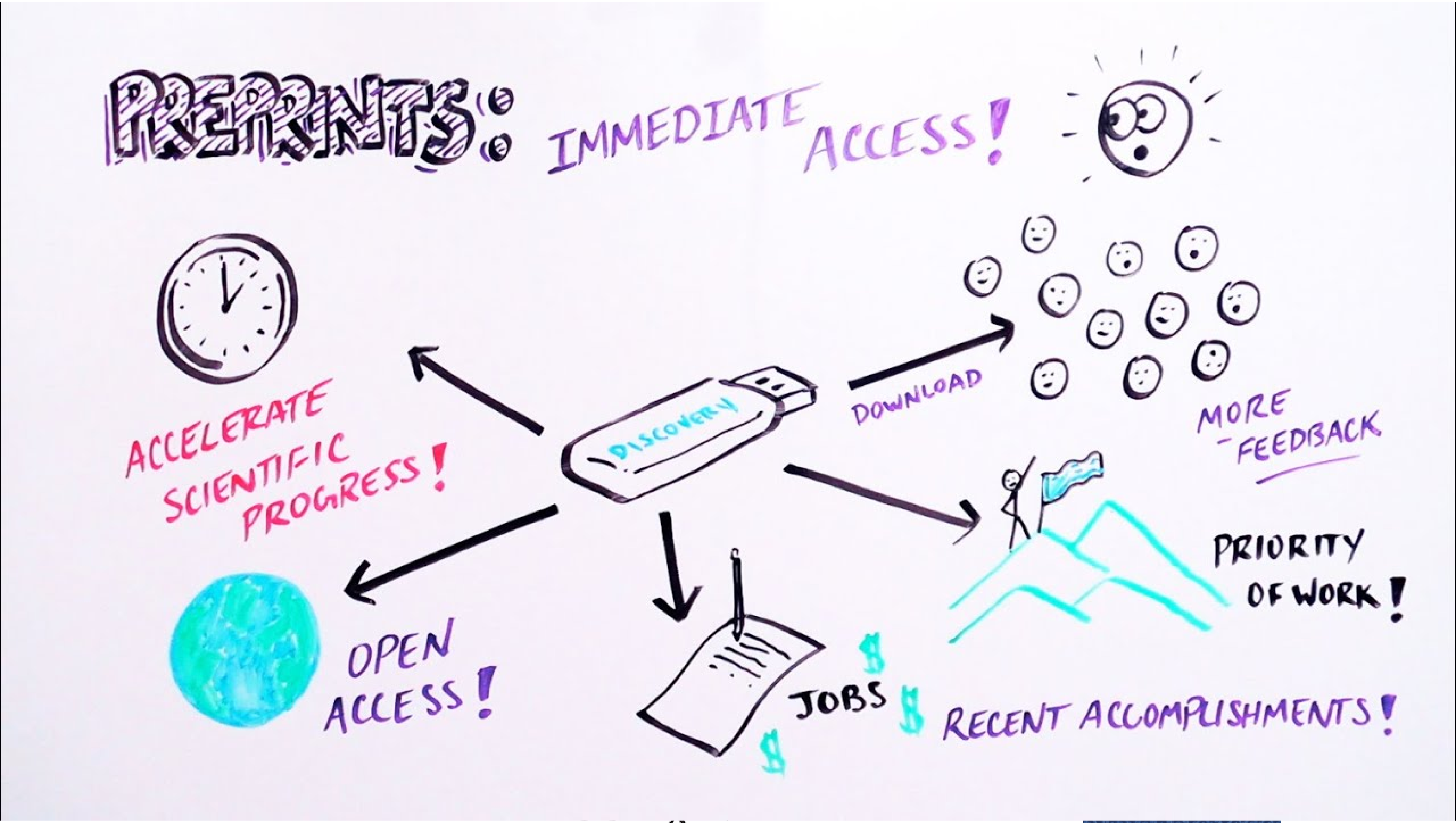
Subject Areas

All Articles

Abstract

Transcription activation domains (ADs) are encoded by a wide range of seemingly unrelated

General workflow in science



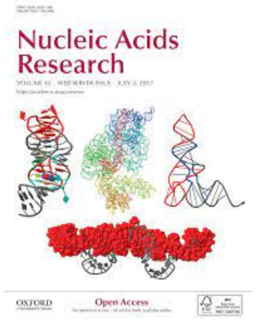
THE PREPRINT SERVER FOR HEALTH SCIENCES

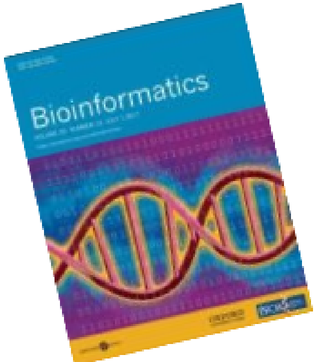
General workflow in science

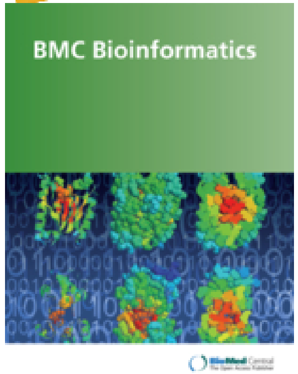
github SOCIAL CODING

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

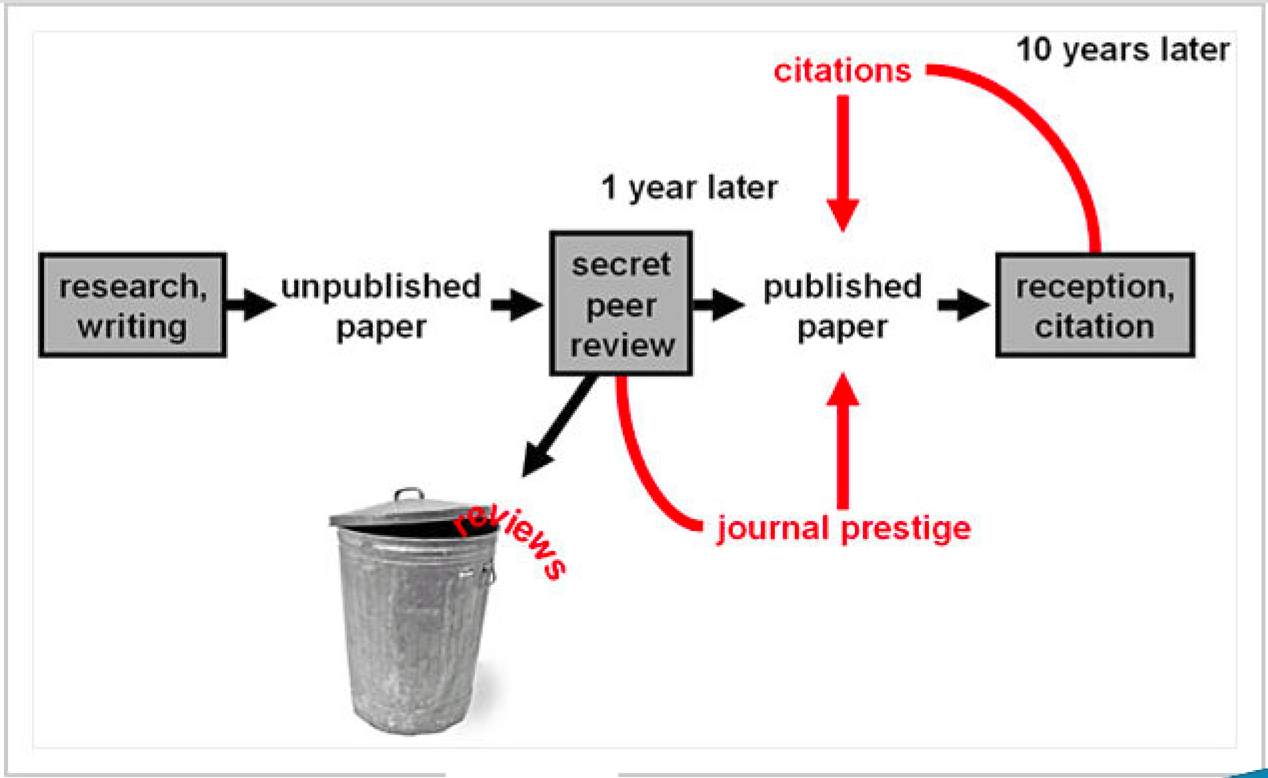
medRχiv
THE PREPRINT SERVER FOR HEALTH SCIENCES







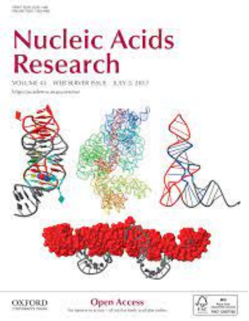
General workflow in science

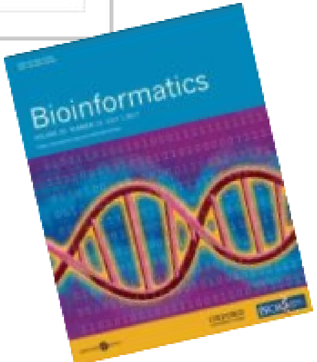


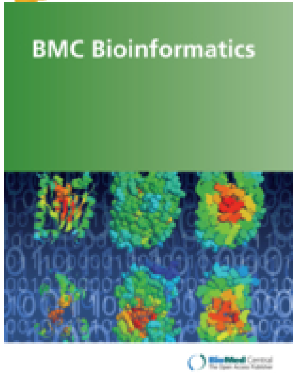
github
SOCIAL CODING

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

medRχiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

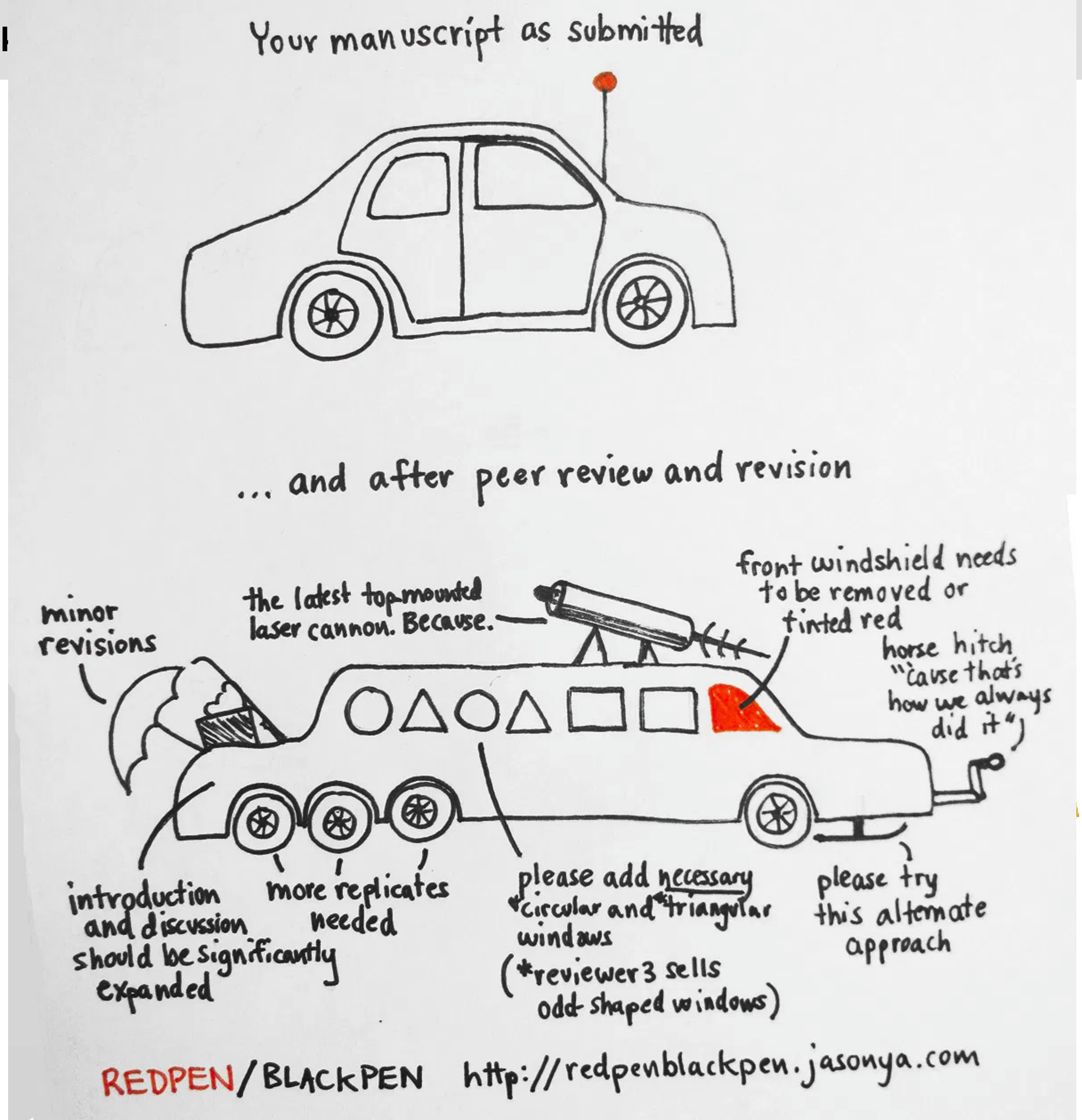






General wor

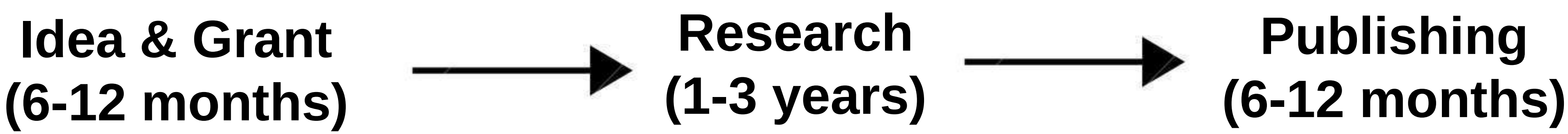
ADP


Your manuscript as submitted
... and after peer review and revision
minor revisions
the latest top-mounted laser cannon. Because.
front windshield needs to be removed or tinted red
horse hitch ("cause that's how we always did it")
introduction and discussion should be significantly expanded
more replicates needed
please add necessary *circular and *triangular windows
(*reviewer 3 sells odd-shaped windows)
please try this alternate approach
REDPEN/BLACKPEN http://redpenblackpen.jasonya.com

General workflow in science

Problem to solve

→



Idea & Grant
(6-12 months)

→

Research
(1-3 years)

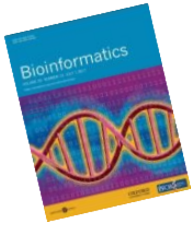
→

Publishing
(6-12 months)

General workflow in science

Problem to solve

GitHub



Idea & Grant
(6-12 months)

Research
(1-3 years)

Publishing
(6-12 months)

bioRχiv

General workflow in science

Problem to solve

→

Idea & Grant
(6-12 months)

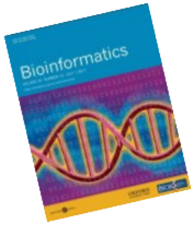
→

GitHub

Research
(1-3 years)

bioRxiv

→

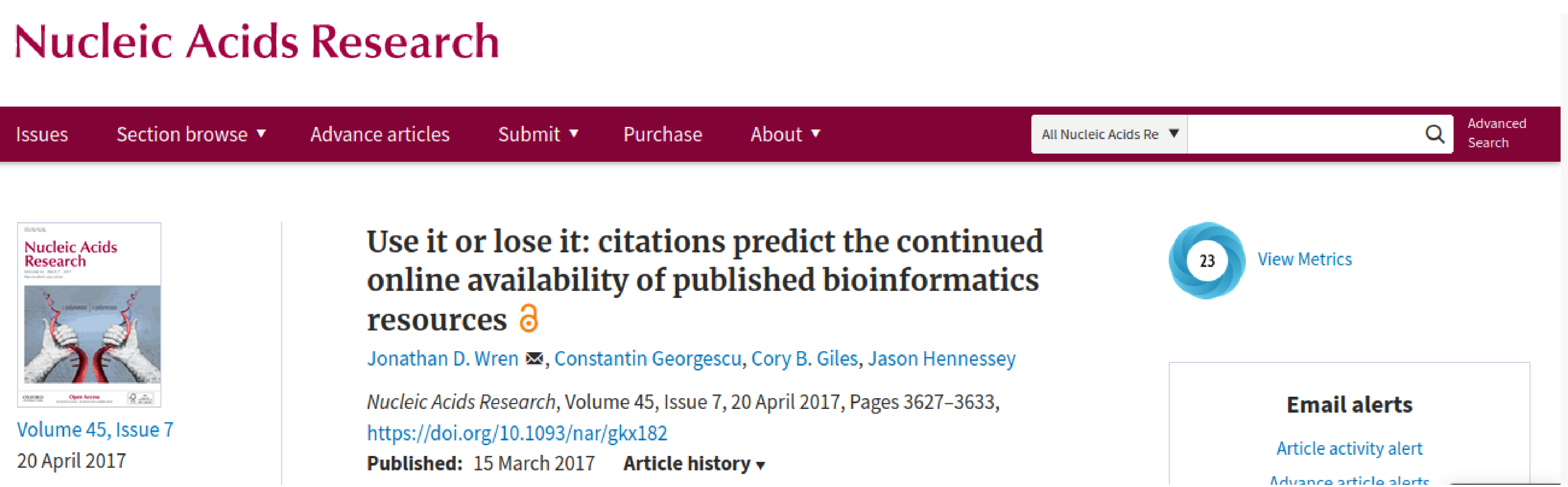


Publishing
(6-12 months)

→

Post Production
(at least 3-5 years)

General workflow in science



Nucleic Acids Research

Volume 45, Issue 7
20 April 2017

Use it or lose it: citations predict the continued online availability of published bioinformatics resources

Jonathan D. Wren, Constantin Georgescu, Cory B. Giles, Jason Hennessey

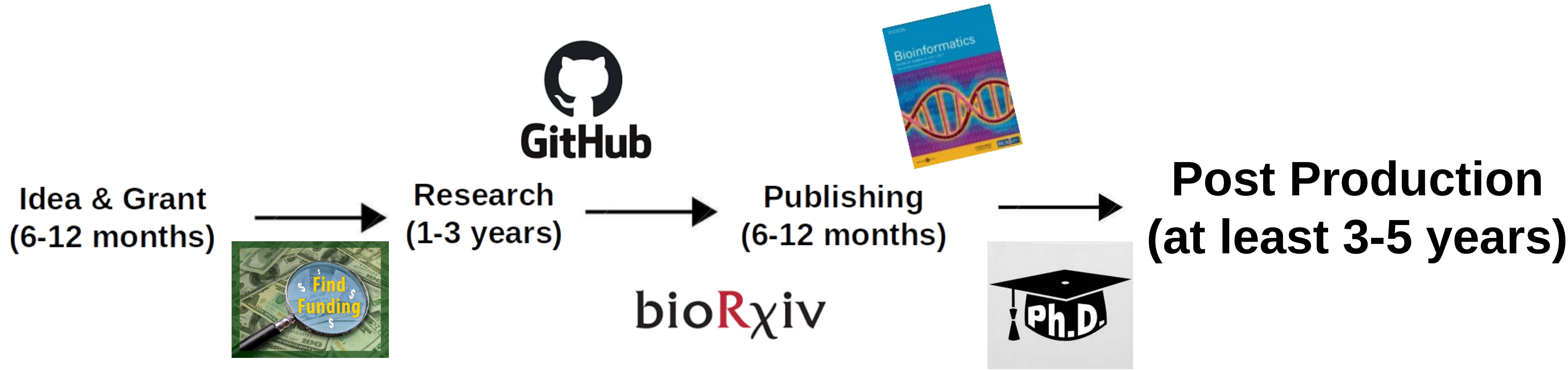
Nucleic Acids Research, Volume 45, Issue 7, 20 April 2017, Pages 3627–3633, https://doi.org/10.1093/nar/gkx182

Published: 15 March 2017 **Article history** ▾

Email alerts

Article activity alert

Advance article alerts



Idea & Grant
(6-12 months)

→

Research
(1-3 years)

→

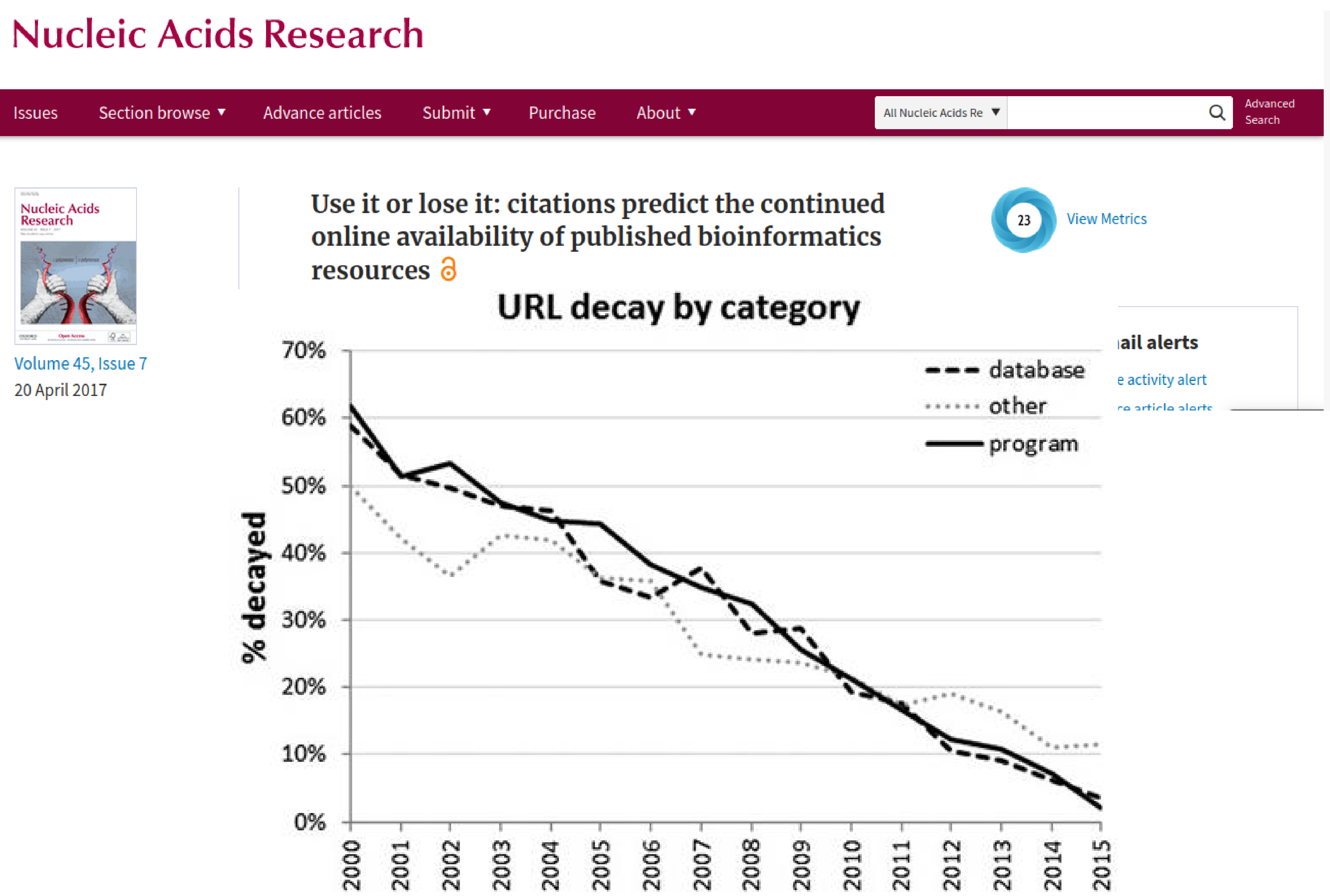
Publishing
(6-12 months)

→

Post Production
(at least 3-5 years)

General workflow in science

ADP



Nucleic Acids Research

Issues | Section browse ▾ | Advance articles | Submit ▾ | Purchase | About ▾ | All Nucleic Acids Re ▾ | Advanced Search

Use it or lose it: citations predict the continued online availability of published bioinformatics resources

Volume 45, Issue 7

20 April 2017

23 View Metrics

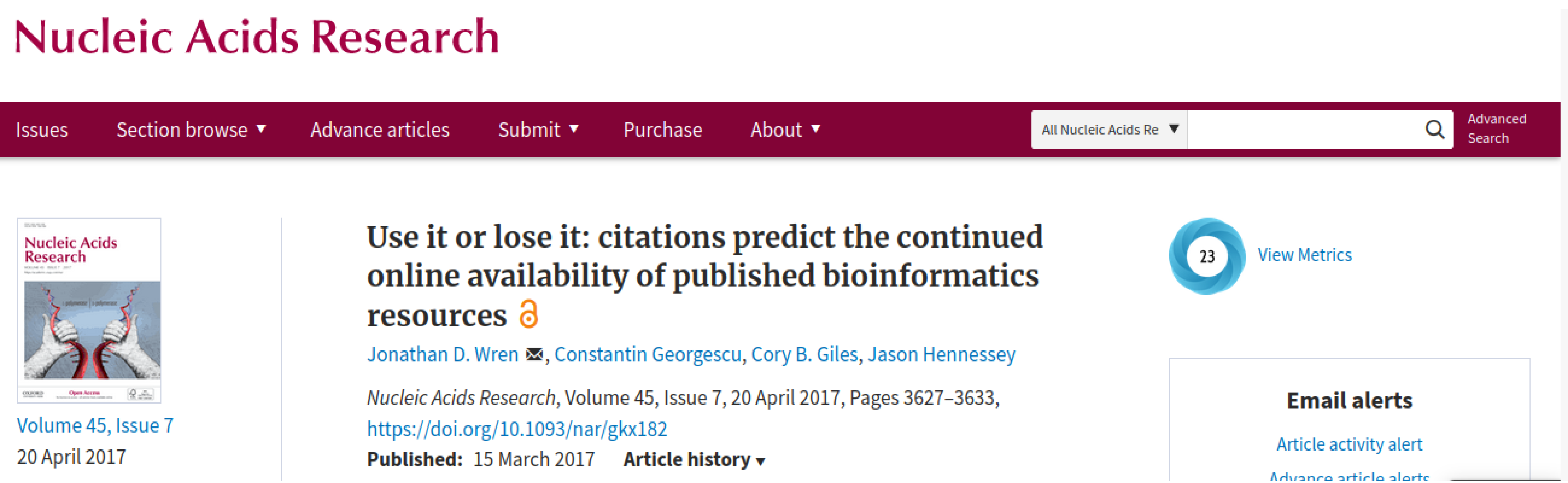
ail alerts

e activity alert

ce article alerts

General workflow in science

ADP



Volume 45, Issue 7

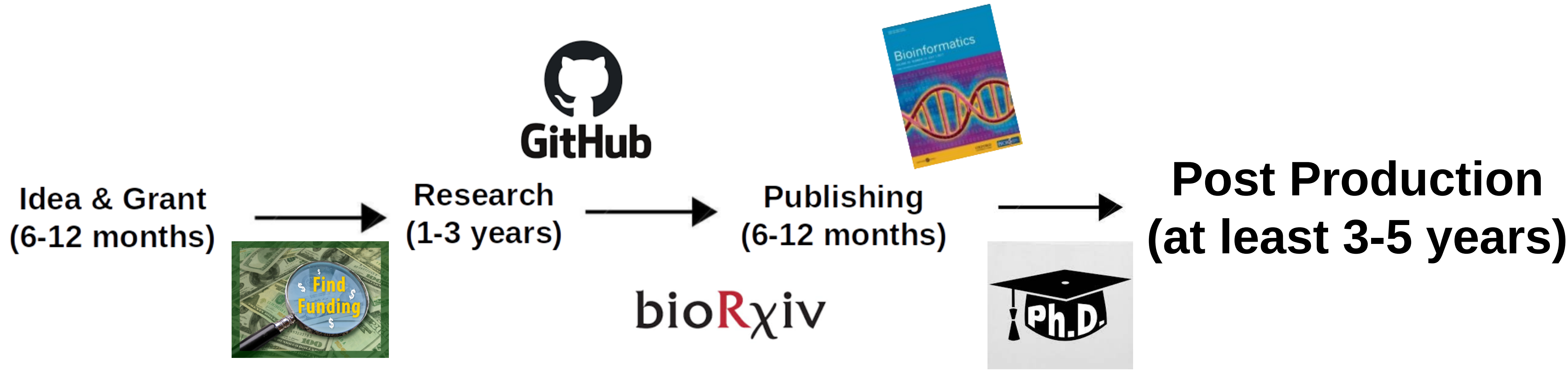
20 April 2017

Use it or lose it: citations predict the continued online availability of published bioinformatics resources

Jonathan D. Wren, Constantin Georgescu, Cory B. Giles, Jason Hennessey

Nucleic Acids Research, Volume 45, Issue 7, 20 April 2017, Pages 3627–3633, https://doi.org/10.1093/nar/gkx182

Published: 15 March 2017 **Article history** ▾



Idea & Grant (6-12 months) → Research (1-3 years) → Publishing (6-12 months) → Post Production (at least 3-5 years)

Projects

Projects presentation will be done during last week (11.06.24)

Formal requirements:

- the project has the github repository

- the project has the documentation

- license, data, scripts, test/examples are provided

Additionally, the project will be described by the leader of the Group (or few members) in short presentation (up to 20 min, plus 5-10 min for discussions).

The format: pdf/ppt and/or if possible interactive presentation of the project in the browser

Projects presentation will be done during last week (11.06.24)

Presentation should include:

- the introduction of the project (problem to be solved)

- the objectives (what you wanted to achieve)

- the results (what you achieved)

- the conclusions & possible directions for the future
(what the software can do and what can be improved if more people would join)

Projects presentation will be done during last week (11.06.24)

Presentation should include:

- the introduction of the project (problem to be solved)

- the objectives (what you wanted to achieve)

- the results (what you achieved)

- the conclusions & possible directions for the future (what the software can do and what can be improved if more people would join)

The project and presentations should be finished up to 09.06.
(please upload the presentation to the github of your repository up to 09.06)

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

lukaskoz@mimuw.edu.pl