Astrocyte: A Scientific Workflow Platform on the BioHPC Cluster

—

Peng Lian, Ph.D.

BioHPC, Lyda Hill Department of Bioinformatics
UT Southwestern Medical Center

Agenda

What is Astrocyte?

How does Astrocyte work?

The Astrocyte web interface

The Astrocyte command-line interface (CLI)

Use case: ChIP-seq analysis workflow

Demonstration: Astrocyte Alphafold workflow

What is Astrocyte?

A platform for scientific workflows

Features:

- Standardized workflows

- Good reproducibility

- Easy-to-use web interface

- Online documents

- Result visualization

How does Astrocyte work?

The Astrocyte web interface
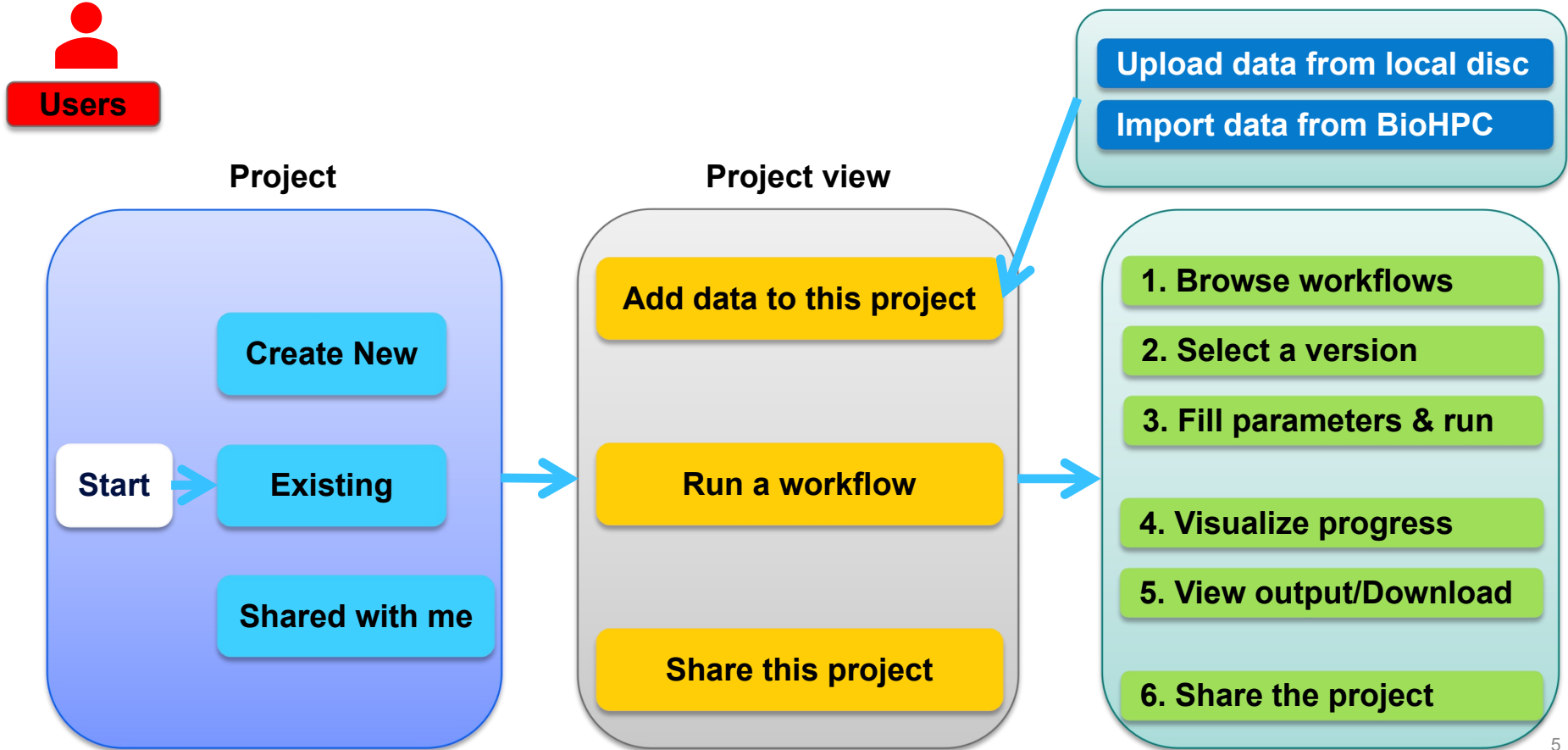

Users

Project

Start → Create New

Existing

Shared with me

Project view

Add data to this project

Run a workflow

Share this project

Upload data from local disc

Import data from BioHPC

1. Browse workflows

2. Select a version

3. Fill parameters & run

4. Visualize progress

5. View output/Download

6. Share the project

The Astrocyte command-line interface (CLI)

Version control of the workflows

Developers

workflow repositories

BioHPC Git

GitLab

Astrocyte

BICF

CRI

Radiology

Green Center

Tags for publishing

publish_2.1.3
test_1.0.2

PURPOSE:
For publishing/testing

MAJOR:
changes that break the
parameter structure of
previous versions

MINOR:
new functionality

PATCH:
small bug fixes

Astrocyte CLI: the YAML file

Developers

```
# ----------------------------------------------------------------
# BASIC INFORMATION
# ----------------------------------------------------------------

# A unique identifier for the workflow package, text/underscores only
name: 'example_wordcount'
# Who wrote this?
author: 'David Trudgian, Peng Lian'
# A contact email address for questions
email: 'biohpc-help@utsouthwestern.edu'
# A more informative title for the workflow package
title: 'Example Wordcount Workflow'
# A summary of the workflow package in plain text
description: |
  This is a minimal test workflow package that counts the occurences of words
  in a test file. It can be used as a template to develop workflows, and as to
  test the astrocyte platform. Please note that start from publish version
  0.0.8 and test version 0.0.2, this workflow will only support Astrocyte 0.4.0
  and above.

#### New Features in Astrocyte 0.4.0 and above ####
citation: |    You, 12 days ago * Add citation label
  Please cite individual programs and versions of pipeline
  used, and the overall pipeline doi: 12.3456/zenodo.9876543.
  Please cite in publications:
  Pipeline was developed by BioHPC.

# The minimum Astrocyte version that requires to run this workflow. For old pipelines, which do not have this label
# a default value of 0.3.1 will be assigned automatically. A request of minimum version less than 0.4.0 will be ignored.
minimum_astrocyte_version: '0.4.0'
# The Nextflow version that requires to run this workflow.  For old pipelines, which do not have this label
# a default value of 0.31.0 will be assigned automatically. Please make sure the requested nextflow version is available
# in the module list.
nextflow_version: '20.01.0'
# (Optional) The Nextflow config file to use for this workflow. If provided, the file should exist in workflow_configs
nextflow_config: 'biohpc.config'
# The container to use for this workflow, none/singularity. If omitted, the default value 'none' will be used.
container: 'singularity'
# The version of singularity to use. This is required if container == 'singularity'
singularity_version: '3.5.3'

# ----------------------------------------------------------------
# DOCUMENTATION
# ----------------------------------------------------------------

# A list of documentation file in .md format that should be viewable from the
# web interface. These files are in the 'docs' subdirectory. The first file
# listed will be used as a documentation index and is index.md by convention
# To supply a title for documentation use a pair of [ 'filename', 'title' ]

documentation_files:
  - [ 'index.md', 'Wordcount Help' ]

# ----------------------------------------------------------------
# NEXTFLOW WORKFLOW CONFIGURATION
# ----------------------------------------------------------------

# Remember - The workflow file is always named 'workflow/main.f'
#            The workflow must publish all final output into $baseDir

# A list of cluster environment modules that this workflow requires to run.
# Specify versioned module names to ensure reproducability.
workflow_modules:
  - Test
```

```
workflow_parameters:
  - id: story
    type: files
    required: true
    description: |
      A text file containing a story
    regex: ^.*(txt)$
    min: 1

  - id: test_int
    type: integer
    required: true
    default: 123
    min: 1
    max: 1000
    description: |
      This is an example integer field, it is not used by the workflow

  - id: test_real
    type: real
    required: true
    default: 123.456
    min: 1
    max: 1000
    description: |
      This is an example real field, it is not used by the workflow

  - id: test_string
    type: string
    required: true
    default: "Example"
    description: |
      This is an example string field, it is not used by the workflow

  - id: test_select
    type: select
    required: true
    default: "Dog"
    description: |
      This is an example select field, it is not used by the workflow
    choices:
      - [ 'Cat', 'Cat, Kitten' ]
      - [ 'Dog', 'Dog, Puppy' ]
      - [ 'Horse', 'Horse, Foal' ]

  - id: test_multiselect
    type: multiselect
    required: true
    default: "Banana"
    description: |
      This is an example multiselect field, it is not used by the workflow
    choices:
      - [ 'Apple', 'Green Apple' ]
      - [ 'Banana', 'Yellow Banana' ]
      - [ 'Strawberry', 'Red Strawberry' ]

# ----------------------------------------------------------------
# SHINY APP CONFIGURATION
# ----------------------------------------------------------------

# Remember - The vizapp is always 'vizapp/server.R' 'vizapp/ui.R'
#            The workflow must publish all final output into $baseDir

# Name of the R module that the vizapp will run against
vizapp_r_module: 'R/3.4.1-gccmkl'

# List of any CRAN packages, not provided by the modules, that must be made
# available to the vizapp
vizapp_cran_packages:
  - shiny
  - shinyFiles
```

Use case I: ChIP-seq analysis workflow
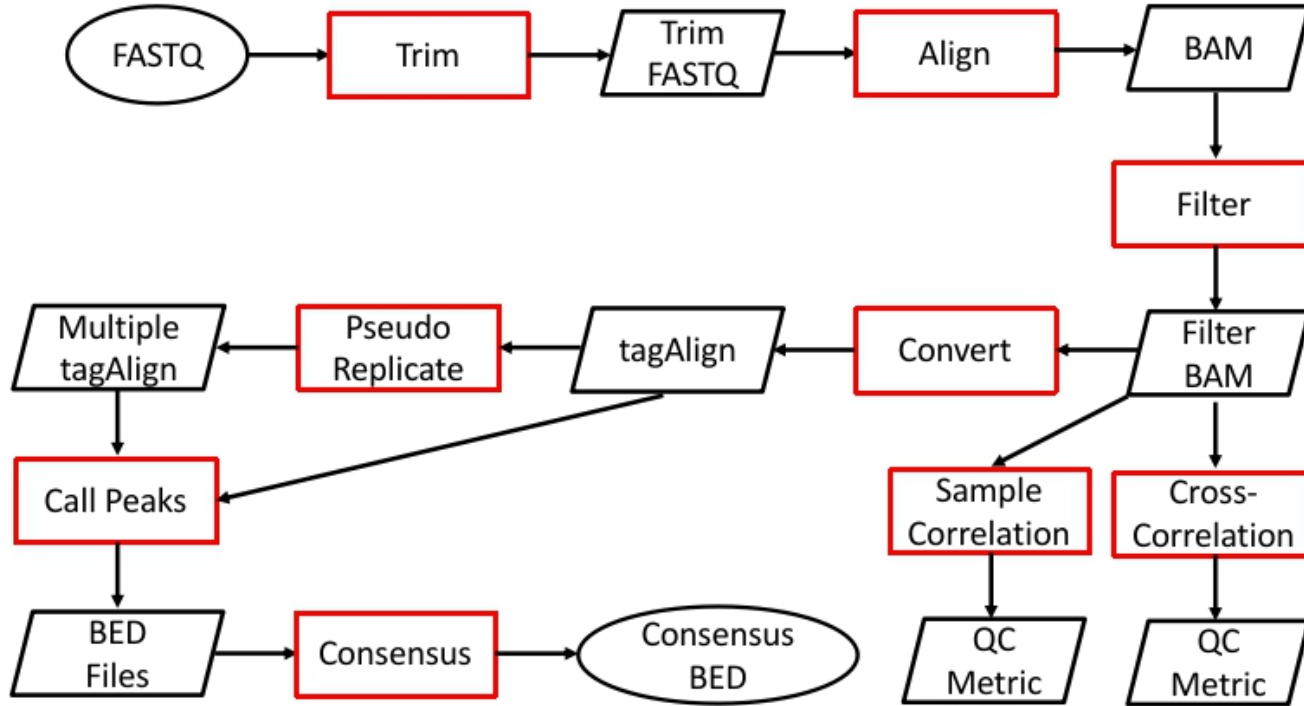
Astrocyte web: create a project

Users

My Projects

In Astrocyte **projects** are used to organize your work. You upload **input data** into a project, and can then run **workflows** against this input data. Try to separate your work into natural projects, so that you can easily share them with other users if required.

✚ Start a New Project

| Project Name | Create New Project |

👤 Existing Projects

ID	Name	Created	Workflows Run	Input Files	Size	Actions
1956	biohpc_chipseq	Feb. 22, 2022, 12:41 p.m.	2	1	775.7 MB	🗑 Delete
1905	alphafold	Oct. 3, 2021, 10:45 p.m.	3	3	394.8 MB	🗑 Delete
1885	chip_seq	Aug. 3, 2021, 11:06 p.m.	1	10	389.4 GB	🗑 Delete
1882	rshiny	July 14, 2021, 11:19 a.m.	3	4	484.0 MB	🗑 Delete
1716	wordcount	June 2, 2020, 2:26 p.m.	12	6	347.0 MB	🗑 Delete

↩ Projects Shared with Me

ID	Name	Created	Workflows Run	Input Files	Size	Actions
PRJ1909	Liver	Oct. 18, 2021, 2:49 p.m. by Xun WANG	8	9	11.6 GB	
PRJ1779	ERX_41_SUM159_WT_and_KO_4h	Sept. 23, 2020, 10:08 a.m. by Xihui Liu	23	9	1.8 TB	

10

Astrocyte web: add data to the project

Users

Project 1956 - biohpc_chipseq

Owner: Peng Lian (s190450)

Created: Feb. 22, 2022, 12:41 p.m. by Peng Lian (s190450)

📄 Input data in this project

To run a workflow against input data you need to upload it into this project. Click the button below to add new files from your web browser or the BioHPC cluster. You can also download or delete existing files from the project in the list below.

⊕ Add Data To This Project

Filename	Size	Actions
G1E_ER4_CTCT_chr9.fastq	29.2 MB	⊕ Download 🗑 Delete

☰ Workflows run in this project

Astrocyte provides many workflow created by different groups at UTSW for you to run against your data. To begin, make sure you have added input data into your project and then click the 'Run a workflow' button to choose a workflow to run.

⊕ Run a workflow in this project

ID	Name	Date	Workflow	Status	Size	Actions
3804	test_vizapp	Feb. 22, 2022, 1:16 p.m.	Astrocyte Example ChIPSeq Workflow Version test_0.0.3	Completed successfully	366.4 MB	🗑 Delete
3808	test_vizapp	Feb. 22, 2022, 1:42 p.m.	Astrocyte Example ChIPSeq Workflow Version publish_0.0.8	Completed successfully	380.2 MB	🗑 Delete

↻ Sharing

You can share access to this project with another user. Anyone you share the project with will be able to upload/remove data, run and remove workflows, but cannot delete the project itself. You can only share with a user that has logged into Astrocyte at least once.

--------- ∨   Share With User

11

Astrocyte web: choose a workflow to run

Available Workflows

Sort by: Popularity | Name (a-z) | Name (z-a)

UTSouthwestern Medical Center | BICF	**BICF RNASeq Analysis Workflow** This is a workflow package for the BioHPC/BICF RNASeq workflow system. It implements differential expression analysis, gene set enrichment analysis, gene fusion analysis and variant identification using RNASeq data.	**Current Version:** rnaseq_bicf - 2.0.0 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | BICF	**BICF ChIP-seq Analysis Workflow** This is a workflow package for the BioHPC/BICF ChIP-seq workflow system. It implements ChIP-seq analysis workflow and visualization application.	**Current Version:** chipseq_analysis_bicf - 1.1.3 **Author:** Holly Ruess, Spencer D. Barnes, Jeremy A. Mathews, Beibei Chen and Venkat Malladi **Contact:** bicf@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | WBMF	**aMAP Pipeline execution Workflow** This workflow automates various steps involved in Image registration and segmentation process. The input is stitched tif image files from tissuecyte microscopes and probability maps from ilastik training. The workflow performs the task of converting tif to nifty files and then registering them to the reference atlas. The output nii files are converted back to tif stack.	**Current Version:** aMAP_Workflow - 0.0.8 **Author:** Apoorva Ajay **Contact:** Apoorva.Ajay@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | BioHPC	**Example Wordcount Workflow** This is a minimal test workflow package that counts the occurrences of words in a test file. It can be used as a template to develop workflows, and as to test the astrocyte platform. Please note that start from publish version 0.0.8 and test version 0.0.2, this workflow will only support Astrocyte 0.4.1 and above.	**Current Version:** example_wordcount - 0.0.9 **Author:** David Trudgian, Peng Lian, Devin OKelly **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | BICF	**BICF SRA Download Pipeline** This is a workflow package for the BioHPC/BICF SRA download workflow system. It implements the SRA Download workflow.	**Current Version:** sra_pipeline_bicf - 1.1.0 **Author:** Jon Gesell, Jeremy Mathews **Contact:** bicf@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | BioHPC	**Astrocyte Example ChIPSeq Workflow** This is an example workflow package for the BioHPC astrocyte workflow system. It implements a simple ChIPSeq analysis workflow using BWA and MACS, plus a simple R Shiny visualization application.	**Current Version:** astrocyte_example - 0.0.7 **Author:** David Trudgian **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
CHILDREN'S MEDICAL CENTER RESEARCH INSTITUTE AT UT SOUTHWESTERN	**Normalized data profiling and visualization** This workflow uses R and Cluster to visualize the profile of a tabular dataset such as RNA-seq data, microarray data, proteomic data, and so on. We assume that the data have been normalized. If not, the workflow can be used to visualize some features of the original data but functions such as clustering and heatmap might be much less meaningful.	**Current Version:** Normalized_data_profiling - 0.0.64 **Author:** Liang Shi **Contact:** liang.shi@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | BICF	**BICF Germline Variant Analysis Workflow** This is a workflow package for the BioHPC/BICF Germline Variant workflow system. It implements a simple germline variant analysis workflow using TrimGalore, BWA, Speedseq, GATK, Samtools and Platypus. SNPs and Indels are integrated using BAYSIC; then annotated using SNPEFF and SnpSift.	**Current Version:** germline_bicf - 1.0.0 **Author:** Brandi Cantarel **Contact:** biohpc-help@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | BICF	**BICF ATAC-seq Analysis Workflow** This is a workflow package for the BICF ATAC-seq workflow system. It implements ATAC-seq analysis workflow and visualization application.	**Current Version:** atacseq_analysis_bicf - 2.0.4 **Author:** Holly Ruess, Spencer Barnes, Venkat Malladi **Contact:** bicf@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions
UTSouthwestern Medical Center | BICF	**BICF CellRanger count Workflow** This is a workflow package for the BICF/Strand Lab CellRanger count workflow system. It implements 10x CellRanger count analysis workflow application.	**Current Version:** cellranger_count - 2.2.0 **Author:** Gervaise H. Henry, Jeremy Mathews, Jon Gesell, and Venkat Malladi **Contact:** bicf@utsouthwestern.edu	▶ Run Workflow / 🗎 Documentation / ⊙ All Versions

Workflows per page: 10 25 50 100

« 1 2 3 »

Astrocyte web: select a version of the workflow

Astrocyte web: fill out parameter forms

Astrocyte web: submit and wait for the result of a run

Astrocyte web: launch the vizapp for online analysis

Astrocyte web: visualization of the result

Demonstration: Astrocyte Alphafold workflow

Summary

As a platform for scientific workflows, Astrocyte bridges the knowledge gap between scientists from different fields.

The web interface and the online documentations enable the user who has little HPC knowledge to run scientific workflows easily.

The Astrocyte CLI provides a toolkit for the workflow developers to compose and test their workflows.

There are already many scientific workflows from different institutes of UT Southwestern Medical Center are running on the Astrocyte platform.

Questions

All questions are welcome!