



This is a repository copy of *CGAT-core: a python framework for building scalable, reproducible computational biology workflows*.

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/145016/>

Version: Submitted Version

---

**Article:**

Cribbs, A.P., Luna-Valero, S., George, C. et al. (10 more authors) (Submitted: 2019)  
CGAT-core: a python framework for building scalable, reproducible computational biology workflows. F1000Research. ISSN 2046-1402 (Submitted)

10.12688/f1000research.18674.1

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:  
<https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.




[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>



## SOFTWARE TOOL ARTICLE

# CGAT-core: a python framework for building scalable, reproducible computational biology workflows [version 1; peer review: awaiting peer review]

Adam P. Cribbs <sup>1,2</sup>, Sebastian Luna-Valero<sup>2</sup>, Charlotte George<sup>2</sup>, Ian M. Sudbery<sup>3</sup>, Antonio J. Berlanga-Taylor<sup>4</sup>, Stephen N. Sansom<sup>5</sup>, Tom Smith<sup>6</sup>, Nicholas E. Illott<sup>5</sup>, Jethro Johnson<sup>7</sup>, Jakub Scaber<sup>2</sup>, Katherine Brown<sup>8</sup>, David Sims<sup>2</sup>, Andreas Heger<sup>2</sup>

<sup>1</sup>The Botnar Research Center, University of Oxford, Oxford, OX3 7LD, UK

<sup>2</sup>MRC WIMM Centre for Computational Biology, University of Oxford, Oxford, OX3 9DS, UK

<sup>3</sup>Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, S10 2TN, UK

<sup>4</sup>MRC-PHE Centre for Environment and Health, Department of Epidemiology & Biostatistics, Imperial College London, Oxford, W2 1PG, UK

<sup>5</sup>Kennedy Institute of Rheumatology, University of Oxford, Oxford, OX3 7FY, UK

<sup>6</sup>Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, CB2 1QR, UK

<sup>7</sup>The Jackson Laboratory for Genomic Medicine, Farmington, CT, 06032, USA

<sup>8</sup>Division of Virology, Department of Pathology, University of Cambridge, Cambridge, CB2 1QP, UK

---

**v1** First published: 04 Apr 2019, 8:377 (<https://doi.org/10.12688/f1000research.18674.1>)

Latest published: 04 Apr 2019, 8:377 (<https://doi.org/10.12688/f1000research.18674.1>)

---

## Abstract

In the genomics era computational biologists regularly need to process, analyse and integrate large and complex biomedical datasets. Analysis inevitably involves multiple dependent steps, resulting in complex pipelines or workflows, often with several branches. Large data volumes mean that processing needs to be quick and efficient and scientific rigour requires that analysis be consistent and fully reproducible. We have developed CGAT-core, a python package for the rapid construction of complex computational workflows. CGAT-core seamlessly handles parallelisation across high performance computing clusters, integration of Conda environments, full parameterisation, database integration and logging. To illustrate our workflow framework, we present a pipeline for the analysis of RNAseq data using pseudo-alignment.

## Keywords

workflow, pipeline, python, genomics

## Open Peer Review

**Referee Status:** *AWAITING PEER*

*REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the **Python Collection** collection.

**Corresponding authors:** Adam P. Cribbs ([adam.cribbs@imm.ox.ac.uk](mailto:adam.cribbs@imm.ox.ac.uk)), David Sims ([david.sims@imm.ox.ac.uk](mailto:david.sims@imm.ox.ac.uk)), Andreas Heger ([andreas.heger@gmail.com](mailto:andreas.heger@gmail.com))

**Author roles:** **Cribbs AP:** Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Luna-Valero S:** Software, Writing – Review & Editing; **George C:** Software, Writing – Review & Editing; **Sudbery IM:** Software, Writing – Review & Editing; **Berlanga-Taylor AJ:** Software, Writing – Review & Editing; **Sansom SN:** Software, Writing – Review & Editing; **Smith T:** Software, Writing – Review & Editing; **Ilott NE:** Software, Writing – Review & Editing; **Johnson J:** Software, Writing – Review & Editing; **Scaber J:** Software, Writing – Review & Editing; **Brown K:** Software, Writing – Review & Editing; **Sims D:** Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Heger A:** Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was funded by the Medical Research Council (UK) Computational Genomics Analysis and Training programme (G1000902).

**Copyright:** © 2019 Cribbs AP *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Cribbs AP, Luna-Valero S, George C *et al.* **CGAT-core: a python framework for building scalable, reproducible computational biology workflows [version 1; peer review: awaiting peer review]** F1000Research 2019, 8:377 (<https://doi.org/10.12688/f1000research.18674.1>)

**First published:** 04 Apr 2019, 8:377 (<https://doi.org/10.12688/f1000research.18674.1>)

## Introduction

Genomic technologies have given researchers the ability to produce large amounts of data at relatively low cost. Bioinformatic analyses typically involve passing data through a series of manipulations and transformations, called a pipeline or workflow. The need for tools to manage workflows is well established, with a wide range of options available from graphical user interfaces such as Galaxy<sup>1</sup> and Taverna<sup>2</sup>, aimed at non-programmers, to Snakemake, Nextflow, Toil, CGAT-Ruffus and others<sup>3-10</sup> developed with computational biologists in mind. These tools differ in their portability, scalability, parameter handling, extensibility, and ease of use. In a recent survey<sup>11</sup>, the tool rated highest for ease of pipeline development was CGAT-Ruffus<sup>12</sup>, a Python package that wraps pipeline steps in discrete Python functions, called 'tasks'. It uses Python decorators to track the dependencies between tasks, ensuring that dependent tasks are completed in the correct order and independent tasks can be run in parallel. If a pipeline is interrupted before completion, or new input files are added, only data sets that are missing or out-of-date are re-run. CGAT-Ruffus implements a wide range of decorators that allow complex operations on input files including: conversion of a single input file to a single output file; splitting of a file into multiple files (and vice versa) and conditional merging of multiple input files into a smaller number of outputs. More advanced options include combining combinations or permutations of input files and conditional execution based on input parameters. Use of decorators means that CGAT-Ruffus pipelines are native Python scripts, rather than the domain specific languages (DSLs) used in other workflow tools. A key advantage of this, in addition to python being an already widely understood language in computational biology, is that individual steps can use arbitrary python code, both in how they are linked together and in the actual processing task.

Here, we introduce Computational Genomics Analysis Toolkit (CGAT)-core<sup>13</sup>, an open-source python library that extends the functionality of CGAT-Ruffus by adding cluster interaction, parameterisation, logging, database interaction and Conda environment switching.

## Methods

CGAT-core<sup>13</sup> extends the functionality of CGAT-Ruffus by providing a common interface to control distributed resource management systems using Distributed Resource Management Application API (DRMAA). Currently, we support interaction with Sun Grid Engine, Slurm and PBS-pro/Torque. The execution engine enables tasks to be run locally or on a high-performance computing cluster and supports cluster distribution of both command line scripts (*cgatcore.run*) and python functions (*cgatcore.cluster*). System resources (number of cores to use, amount of RAM to allocate) can be set on a per-pipeline, per-task, or per task-instance basis, even allowing allocation to be based in variables, for example input file size.

## Operation

The parameter management component encourages the separation of workflow/tool configuration from implementation to build re-usable workflows. Algorithm parameters are collected in a

single human-readable yaml configuration file. Thus, parameters can be set specifically for each dataset, without the need to modify the code. For example, sequencing data can be aligned to a different reference genome, by simply changing the path to the genome index in the yaml file. Both pipeline-wide and job-local parameters are automatically substituted into command line statements at execution-time.

To assist with reproducibility, record keeping and error handling CGAT-core provides multi-level logging during workflow execution, recording full details of runtime parameters, environment configuration and tracking job submissions. Additionally, CGAT-core provides a simple, lightweight interface for interacting with relational databases such as SQLite (*cgatcore.database*), facilitating loading of analysis results at any step of the workflow, including combining output from parallel steps in single wide- or long-format tables.

CGAT-core can load a different Conda environment for each step of the analysis, enabling the use of tools with conflicting software requirements. Furthermore, providing Conda environment files alongside pipeline scripts ensures that analyses can be fully reproduced.

CGAT-core workflows are Python scripts, and as such are stand-alone command line utilities that do not require the installation of a dedicated service. In order to reproducibly execute our workflows, we provide utility functions for argument parsing, logging and record keeping within scripts (*cgatcore.experiment*). Workflows are started, inspected and configured through the command line interface. Therefore, workflows become just another tool and can be re-used within other workflows. Furthermore, workflows can leverage the full power of Python, making them completely extensible and flexible.

## Implementation

CGAT-core is implemented in Python 3 and installable via Conda and PyPI with minimal dependencies. We have successfully deployed and tested the code on OSX, Red Hat and Ubuntu. We have made CGAT-core and associated repositories open-source under the MIT licence, allowing full and free use for both commercial and non-commercial purposes. Our software is fully documented (<https://pypi.org>), version controlled and has extensive testing using continuous integration (<https://travis-ci.org/cgat-developers>.) We welcome community participation in code development and issue reporting through GitHub.

## Use case

To illustrate a simple use case of CGAT-core, we have built an example RNAseq analysis pipeline, which performs read counting using Kallisto<sup>14</sup> and differential expression using DESeq2<sup>15</sup>. This workflow and Conda environment are contained within our CGAT-showcase repository (<https://github.com/cgat-developers/cgat-showcase>). The workflow highlights how simple pipelines can be constructed using CGAT-core, demonstrating how the pipeline can be configured using a yaml file, how third-party tools can be executed efficiently across a cluster or on a local machine, and how data can be easily loaded into a

database. Furthermore, we and others have been extensively using CGAT-core to build pipelines for computational genomics (<https://github.com/cgat-developers/cgat-flow>).

## Discussion

CGAT-core<sup>13</sup> extends the popular Python workflow engine CGAT-Ruffus by adding desirable features from a variety of other workflow systems to form an extremely simple, flexible and scalable package. CGAT-core provides seamless high-performance computing cluster interaction and adds Conda environment integration for the first time. In addition, our framework focuses on simplifying the pipeline development and testing process by providing convenience functions for parameterisation, database interaction, logging and pipeline interaction.

The ease of pipeline development enables CGAT-core to bridge the gap between exploratory data analysis and building production workflows. A guiding principle is that it should be as easy (or easier) to complete a series of tasks using a simple pipeline compared to using an interactive prompt, especially once cluster submission is considered. CGAT-core enables the production of analysis pipelines that can easily be run in multiple environments to facilitate sharing of code as part of the publication process. Thus, CGAT-core encourages a best-practice reproducible research approach by making it the path of least resistance. For example, exploratory analysis in Jupyter

Notebooks can be converted to a Python script or used directly in the pipeline. Similarly, exploratory data analysis in R, or any other language, can easily be converted to a script that can be run by the pipeline. This lightweight wrapping of quickly prototyped analysis forms a lab book, enabling rapid reproduction of analyses and reuse of code for different data sets.

## Data availability

All data underlying the results are available as part of the article and no additional source data are required.

## Software availability

Source code available from: <https://github.com/cgat-developers/cgat-core>.

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.2598115><sup>13</sup>.

Licence: MIT License.

## Grant information

This work was funded by the Medical Research Council (UK) Computational Genomics Analysis and Training programme (G1000902).

## References

- Afgan E, Baker D, van den Beek M, *et al.*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.** *Nucleic Acids Res.* 2016; **44**(W1): W3–W10.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wolstencroft K, Haines R, Fellows D, *et al.*: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W557–61.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Okonechnikov K, Golosova O, Fursov M, *et al.*: **Unipro UGENE: a unified bioinformatics toolkit.** *Bioinformatics.* 2012; **28**(8): 1166–7.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Golosova O, Henderson R, Vaskin Y, *et al.*: **Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses.** *PeerJ.* 2014; **2**: e644.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nocq J, Celton M, Gendron P, *et al.*: **Harnessing virtual machines to simplify next-generation DNA sequencing analysis.** *Bioinformatics.* 2013; **29**(17): 2075–83.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gafni E, Luquette LJ, Lancaster AK, *et al.*: **COSMOS: Python library for massively parallel workflows.** *Bioinformatics.* 2014; **30**(20): 2956–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Vivian J, Rao AA, Nothhaft FA, *et al.*: **Toil enables reproducible, open source, big biomedical data analyses.** *Nat Biotechnol.* 2017; **35**(4): 314–316.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Fisch KM, Meißner T, Gioia L, *et al.*: **Omics Pipe: a community-based framework for reproducible multi-omics data analysis.** *Bioinformatics.* 2015; **31**(11): 1724–8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Köster J, Rahmann S: **Snakemake—a scalable bioinformatics workflow engine.** *Bioinformatics.* 2012; **28**(19): 2520–2.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Leipzig J: **A review of bioinformatic pipeline frameworks.** *Brief Bioinform.* 2017; **18**(3): 530–536.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goodstadt L: **Ruffus: a lightweight Python library for computational pipelines.** *Bioinformatics.* 2010; **26**(21): 2778–9.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Heger A, Cribbs A, Luna-Valero S, *et al.*: **cgat-developers/cgat-core: First public release of code (Version v0.5.10).** *Zenodo.* 2019.  
<http://www.doi.org/10.5281/zenodo.2598115>
- Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–7.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 550.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**