scientific data

OPEN

ARTICLE

WorkflowHub: a registry for computational workflows

Ove Johan Ragnar Gustafsson[1], Sean R. Wilkinson[2], Finn Bacall[3], Stian Soiland-Reyes[3,4], Simone Leo[5], Luca Pireddu[5], Stuart Owen[3], Nick Juty[3], José M. Fernández[6,7], Tom Brown[8], Hervé Ménager[9,10], Björn Grüning[11], Salvador Capella-Gutierrez[6,7], Frederik Coppens[12] & Carole Goble[3] ✉

The rising popularity of computational workflows is driven by the need for repetitive and scalable data processing, sharing of processing know-how, and transparent methods. As both combined records of analysis and descriptions of processing steps, workflows should be reproducible, reusable, adaptable, and available. Workflow sharing presents opportunities to reduce unnecessary reinvention, promote reuse, increase access to best practice analyses for non-experts, and increase productivity. In reality, workflows are scattered and difficult to find, in part due to the diversity of available workflow engines and ecosystems, and because workflow sharing is not yet part of research practice. WorkflowHub provides a unified registry for all computational workflows that links to community repositories, and supports both the workflow lifecycle and making workflows findable, accessible, interoperable, and reusable (FAIR). By interoperating with diverse platforms, services, and external registries, WorkflowHub adds value by supporting workflow sharing, explicitly assigning credit, enhancing FAIRness, and promoting workflows as scholarly artefacts. The registry has a global reach, with hundreds of research organisations involved, and more than 800 workflows registered.

Introduction

In an era of Big Data and data-driven science, the need for repetitive, scalable, reproducible and quality-assured data processing and analysis methods has contributed to a surge in popularity for computational workflows[1]. The past two decades have seen a handful of workflow management systems (WMS) expand to hundreds[2] and workflows applied across a growing number of domains, including biosciences[3], astronomy[4] and the physical sciences[5].

Computational workflows are a special kind of software for handling multi-step, multi-code data pipelines, analysis, and simulations, and are ultimately intended to automate and standardise complex data-handling processes. They come in many forms, and generally will share the following features: a step-by-step set of analytical processes, described in a high-level language that can be executed by a dedicated WMS that manages data flow and code execution; a specific composition of modular software building blocks (e.g. code, program, or even other sub-workflows) that achieve the required step-by-step processes, and which can be remixed; and a tendency to be closely associated, even intertwined with the data on which they will operate[6]. Important scientific goals like repeatability, replicability, and reproducibility become more realistic when scientists specify their experiment's analysis processes as a computational workflow[7]. In practice, this would mean arriving at the same results when a workflow is executed again in the same computational environment (repeatability), or similar results when either an independent but similar computational environment (replicability), or completely different environment (reproducibility), are used[8]. For example, computational workflows have become central

[1]Australian BioCommons, University of Melbourne, Melbourne, Victoria, Australia. [2]Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA. [3]Department of Computer Science, The University of Manchester, Manchester, UK. [4]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands. [5]Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Pula, Cagliari, Italy. [6]Barcelona Supercomputing Center (BSC), Barcelona, Spain. [7]Spanish National Bioinformatics Institute (INB), Barcelona, Spain. [8]Leibniz Institute for Zoo- and Wildlife Research, Berlin, Germany. [9]Institut Pasteur, Université Paris Cité, Bioinformatics of Biostatistics Hub, Paris, 75015, France. [10]CNRS, UMS 3601, Institut Français de Bioinformatique, Evry, France. [11]Albert-Ludwigs-Universität Freiburg, Freiburg, Germany. [12]VIB Data Core, VIB Technologies, Ghent, Belgium. ✉e-mail: carole.goble@manchester.ac.uk

to major international science missions that require systematic, reproducible, and shared data analysis. Recent examples include the global response to the Coronavirus disease 2019 (COVID-19) pandemic and the analyses of severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2)[3], and the large-scale sequencing efforts currently in-flight for the Vertebrate Genomes Project (VGP)[9]. While these large consortia with defined collaborative research programs are key drivers for computational workflow creation and deployment, workflows are also being adopted across scientific disciplines as their computational requirements increase, and the emphasis on reproducibility and portability increases[10–14].

Increasingly, scientists are also being asked to share their data and associated research objects (i.e. workflows), in ways others can reuse (e.g. the Nelson Memo https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf, NASA SPD-41a https://smd-cms.nasa.gov/wp-content/uploads/2023/08/smd-information-policy-spd-41a.pdf). The idea is to accelerate scientific progress and spur innovation by enabling scientists to avoid reinventing each others' work, and to explicitly support confidence in published results by removing ambiguity surrounding the approach taken to create research outcomes. In addition, scientific activity often includes the exploration of analysis variance; modifying workflows to understand effects and changes on data products is simpler when those workflows are clearly described and available. To this end, Wilkinson *et al.* published guiding principles for scientific data management and stewardship, providing guidelines for making data and other research objects Findable, Accessible, Interoperable, and Reusable (FAIR) by others[15]. The FAIR principles have sparked an entire movement in the international community towards adopting FAIR practices, and further work has been undertaken to extend the principles to research software[16], AI models[17], and computational workflows[18]. A fundamental step towards supporting FAIR workflows[19] is to enable the sharing of workflows and their descriptions[20] and make them findable.

Researchers find software by searching: the web (i.e. search engines), public software project repositories (e.g. GitHub), the literature, mailing lists, discussion groups (e.g. StackOverflow), dependencies in the software itself, relevant registries (e.g. The Comprehensive R Archive Network, CRAN https://cran.r-project.org/, communities of practice like nf-core[21,22]), and even social media[23]. High-quality machine-processable metadata markup is needed to make workflows more findable and understandable in such a "search context": in other words, the descriptors for workflows must be themselves standardised, accessible, and discoverable. Current mechanisms for sharing workflows do not achieve this outcome for the entire workflow ecosystem. Sharing source code includes options such as version control platforms (e.g. GitHub https://github.com/or GitLab https://about.gitlab.com/), and WMS-specific curated git repositories (e.g. Intergalactic Workflow Commission, IWC https://github.com/galaxyproject/iwc[24], nf-core, Snakemake catalogue https://snakemake.github.io/snakemake-workflow-catalog/[25]). Creators can also publish their workflows, either in public generalist repositories (e.g. Zenodo https://zenodo.org/, DataVerse https://dataverse.org/), conventional journals (e.g. GigaScience https://academic.oup.com/gigascience) or software journals (e.g. the Journal of Open Source Software, JOSS https://joss.theoj.org/). Finally, a creator can register the workflow using either a platform-specific (e.g. Knime Community Hub https://hub.knime.com/, BinderHub https://binderhub.readthedocs.io/en/latest/, nf-core https://nf-co.re/) or platform-agnostic solution. The latter includes Dockstore, a registry that supports the sharing and running of containerised tools, workflows and notebooks across diverse cloud computing environments (https://dockstore.org/)[26].

Discovery of these many resources and platforms can be impeded by non-standardised descriptors that are not necessarily visible to search. Even once a computational workflow has been found, a divergent ecosystem does not lend itself to better integration of services, adoption of standards, or achieving FAIR outcomes for workflows. A registry that serves as a hub for these various mechanisms, and their specific benefits, would begin to address these challenges. It could support workflow developers to share and gain credit for their work, integrate with the platforms, services and infrastructures that developers and users rely on to both create and use workflows, and support making workflows more FAIR. Structurally, a registry should be flexible, extensible, and use internationally recognised standards that accommodate rich metadata. To capture and present the breadth of the global computational workflow ecosystem back to the research community, a registry should be agnostic to domains and WMS, and embrace community standards. Finally, it should provide mechanisms that can link workflows to other digital objects that provide context for a research project, including documents, standard operating procedures (SOPs) and publications.

Here, we present a public and inclusive registry dedicated specifically to the sharing of computational workflows: WorkflowHub (https://workflowhub.org)[14,27,28]. The registry is distinct from the WorkflowHub Framework[29], now called WfCommons (https://wfcommons.org/)[30], which is designed for analysing and simulating workflow executions. The WorkflowHub registry is designed to allow any scientist, regardless of expertise level, to contribute and share computational workflows. It indexes workflows from any scientific domain, in any format, in any workflow language, regardless of whether it uses a WMS, and supports users to increase the FAIRness of their workflows. Here, we describe in detail how WorkflowHub's structure, design, standards, community engagement, and continued evolution support: 1) collaboration, sharing and credit for workflow developers, projects, and consortia; 2) integration with added-value services, platforms, and capabilities that support the workflow life cycle (i.e. creation, version control, execution, maintenance, reuse and citation); and 3) wizards and inbuilt features that ease the process of sharing workflows alongside the constellation of associated digital artefacts that give a workflow its scientific context.

Results

A registry for computational workflows. The WorkflowHub is a registry for describing, sharing and publishing scientific computational workflows, irrespective of their type, development and maintenance location, or discipline. On the landing page for WorkflowHub, a new user is presented with a description of the platform and its purpose, the latest workflow additions, what content is discoverable, and how to join the WorkflowHub community. Underpinning WorkflowHub is the implementation of open tools and standards, which are further described herein. Collaborating Teams are supported by registry features that support workflow reuse, and include integration with native workflow repositories, assignment of credit, import and export, and the creation of curated Collections of workflows that are enriched by other digital objects (e.g. publications, SOPs). Figure 1 provides a conceptual view of the WorkflowHub's capabilities and its relationship to the workflow development and publishing ecosystem outlined above and discussed later in more depth.

WorkflowHub was launched in 2020, as part of the European Open Science Cloud (EOSC)-Life Workflow Collaboratory[14], to support the registration of workflows required for the response to the Coronavirus disease (COVID) pandemic. WorkflowHub now houses 67 COVID-related workflows (see https://workflowhub.eu/workflows?filter%5Bquery%5D=covid), including those that support the ongoing global analysis of intra-host variation as new samples become available[3,31].

Figure 2 highlights the discipline origins of the contributing groups, also known as Teams, on WorkflowHub based on a high level manual curation by the registry maintainers. Many of the contributions to the registry originate from the life science and health science domains. However, because WorkflowHub is designed to be agnostic to domain, discipline and workflow type, the registry is open to adoption by a wide spectrum of researchers and other stakeholders. For example, there are Teams from engineering, earth, and computer sciences on WorkflowHub (see Fig. 2). Examples of workflow contributions from non-life science domains include eFlows4HPC (https://workflowhub.eu/programmes/35#workflows) and Galaxy Climate (https://workflowhub.eu/projects/18#workflows). You can explore the complete workflow set by visiting the main workflows list (see https://workflowhub.eu/workflows) and by using the complete list of tags (https://workflowhub.eu/tags).

At the time of submission, 860 workflows were registered for an array of workflow types (https://workflowhub.eu/workflow_classes), and the registry had 958 users from 266 Organisations across 36 countries / locations. Figure 3 provides a breakdown of the workflow types registered with WorkflowHub.

A central ambition of WorkflowHub is to be of practical use in advancing the application of computational workflows in research science by *supporting the needs of the communities that it serves*. WorkflowHub meets community requirements and supports the workflow life cycle in three key ways. Firstly, the registry provides structures that directly support collaboration, sharing knowledge and distributing credit. Secondly, woven into this structure are multiple integrations with other elements of the global research ecosystem that support the workflow life cycle: creation, development, discovery, reuse, and citation of workflows (see Table 1 for a list of platforms, services, and standards). Finally, WorkflowHub provides a registration wizard that guides users in leveraging these structures and integrations. This approach is deliberate and will continue to evolve in lock step with the requirements of the community. In the following sections, we will first describe how the data model and metadata framework of WorkflowHub support the registry's core functions – namely registering, finding, and launching workflows and associated digital data objects. In turn we highlight how the registry design, including the use of wizards to guide best practice, allows it to act as an integrating hub across the workflows ecosystem and to support each stage of the workflow life cycle. Finally, we will describe how the WorkflowHub engages with the workflow community and highlight some key use cases for the registry.

A data model that reflects the real-world collaborations that create workflows.

Science is a collaborative enterprise, and infrastructure platforms should reflect this quality to be of practical use in accelerating science missions. Research programs and projects are also intertwined with diverse computational approaches and ways of sharing research outcomes, and these are subject to the same requirements for findability, credit and impact assessment[6]. As a result, WorkflowHub is structured to reflect real-world collaboration and assign complex credit well.

The data model of WorkflowHub provides access to three elements for every registered user: Organisations, Teams, and Spaces. A user can specify one or more Organisations (i.e. affiliations) as part of their user profile. They must also belong to at least one Team, which must also belong to an administrative Space. A user can create a new Team if the existing ones are not suitable or appropriate. A Team member can specify multiple Organisations for each Team they join, which allows someone to be affiliated with different Organisations for different Teams. Multiple creators and Teams can be specified for a single workflow, additional credit can be assigned to contributors, and a distinction is made between creators and submitters. In effect, users belong to Organisations that are specified in Teams, and Teams belong to Spaces. Credit is therefore able to cascade as required from a workflow to creators, contributors, submitters, the Teams and Spaces to which they belong, the consortia and Organisations that these represent, and even new workflows which are derived from the original (see Fig. 4). This nested structure is capable of addressing sharing and credit for a diverse set of workflow contributors, including, but not limited to, individuals (e.g. workflow developers), research groups, institutions (e.g. universities), communities of practice (e.g. nf-core), and major research consortia (e.g. Biodiversity Genomics Europe (BGE) https://biodiversitygenomics.eu/, https://doi.org/10.3030/101059492). For example, single users or research groups may only require a single Team to represent their workflow(s), and in this case they would add their Team to the default "Independent Teams" Space. However, a consortium representing multiple research groups or institutes may create a distinct Team for each one of its collaborating entities, and add the Teams to either the Independent Teams Space, or create a new Space to administer all these Teams. An individual, group, or consortium can thus establish a presence on WorkflowHub that reflects their real world structure, and which the registry uses to assign credit.

With respect to communities of practice, Spaces and Teams can be used to organise contributors, share workflows, and support the sharing of knowledge within that community. This is achieved by registering workflows and linking them to other research outputs, including events, presentations, documents, publications, data files and SOPs - these assets can either be hosted by WorkflowHub or added by reference. A workflow developer can even nominate relevant community channels where they can connect with users, and curated "Collections" of

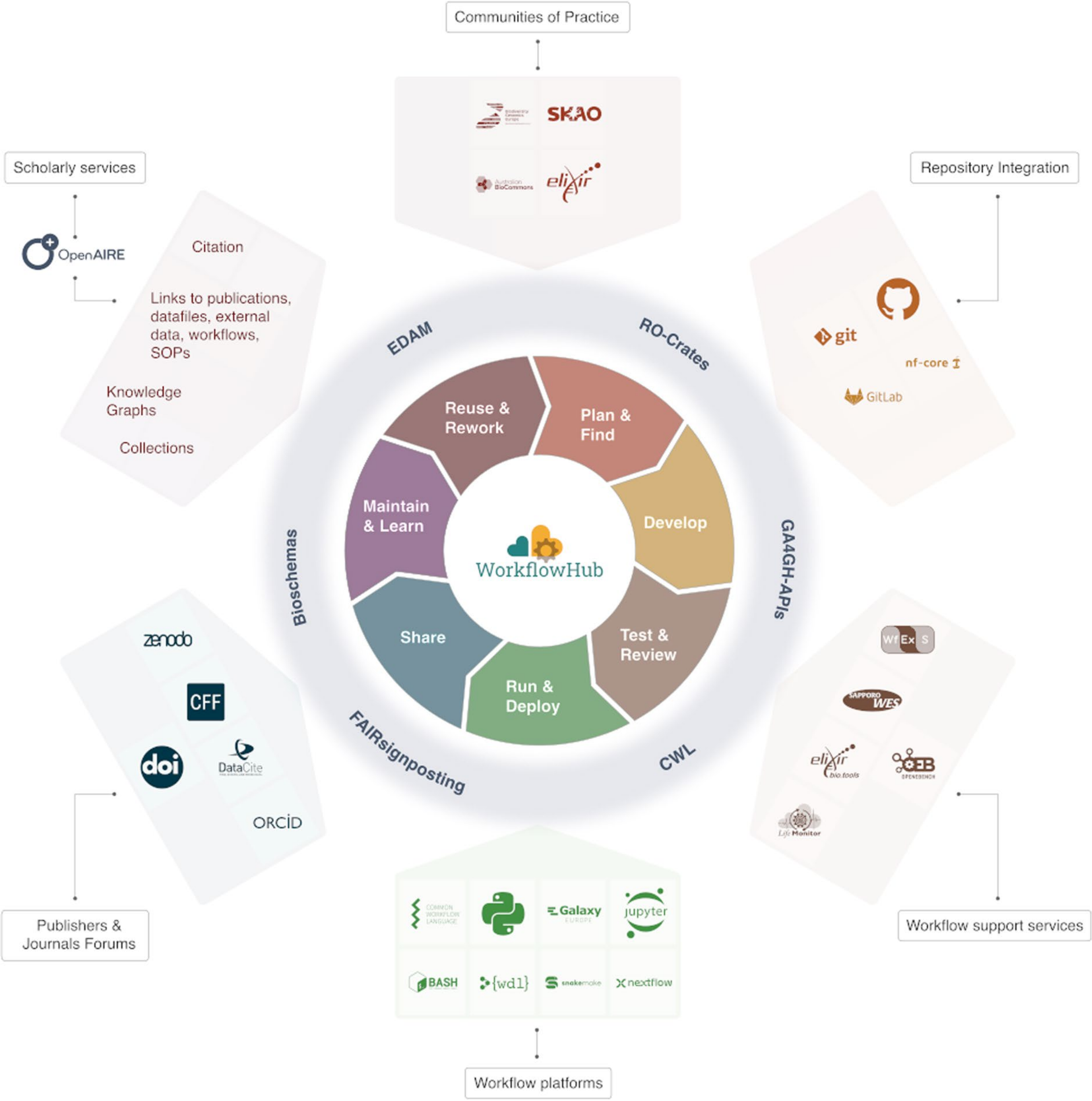


Fig. 1 WorkflowHub connects to platforms, services, and resources that support a workflow's life cycle[47]. A researcher initially needs to **Plan & Find**, where they either plan for a particular analysis and find existing workflows (i.e. using a registry), or **Develop** a new workflow. WorkflowHub integrates with Git repositories (e.g. GitHub, GitLab), and Git-supported communities (e.g. nf-core), to support development. A workflow requires **Test & Review** to **Run & Deploy**, and here WorkflowHub connects to support services (e.g. LifeMonitor, bio.tools, Sapporo WES, WfExS) and welcomes diverse workflow platforms that aid deployment (e.g. Common Workflow Language CWL, Snakemake, Galaxy, Jupyter, Python, BASH, WDL, Nextflow). A creator needs to **Share** a workflow and can benefit from WorkflowHub's use of citation infrastructures and standards (i.e. CITATION.cff, Zenodo, DataCite, DOI and ORCID). In the **Maintain & Learn** stage, maintenance, and also understanding of a workflow by other researchers, becomes critical as it impacts workflow **Reuse & Rework**, where a workflow is either reused, or adapted, by other researchers to suit their requirements. WorkflowHub supports these stages through registration of digital objects that enrich a workflow (e.g. documents, publications, SOPs), the ability to create Collections and workflow citations based on DOIs, and ultimately through the connections created to knowledge graphs. WorkflowHub also enables communities of practice to benefit from all its integrations and connections, ensuring that they can reuse or rework workflows from across the globe. The entire support framework is enabled by the implementation of standards that allow WorkflowHub to interact with the ecosystem and truly act as a "Hub": EDAM, Research Object Crates (RO-Crates), Global Alliance for Genomics and Health (GA4GH) application programming interfaces (APIs), abstract CWL, FAIR Signposting, and Bioschemas.

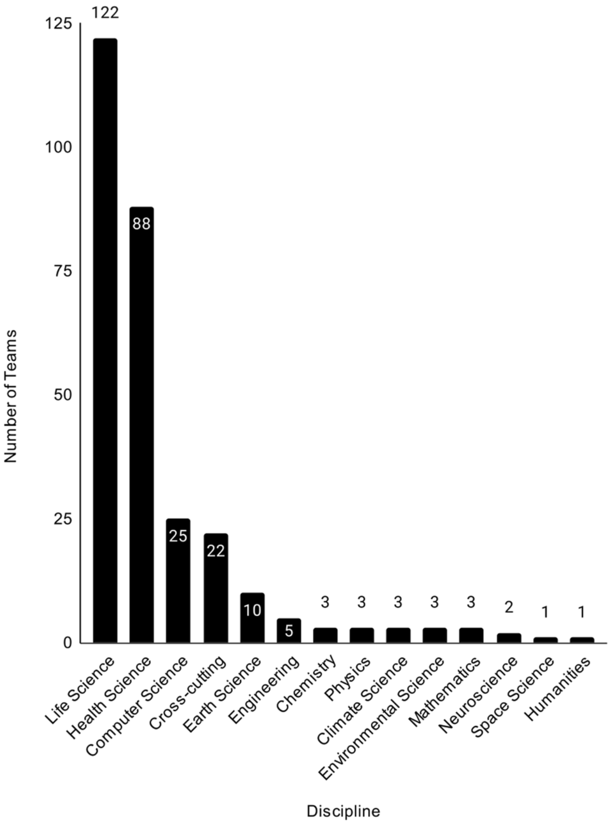


Fig. 2 Breakdown of discipline origins for the groups (also known as Teams) that have registered with WorkflowHub. Teams were categorised by the registry maintainers.

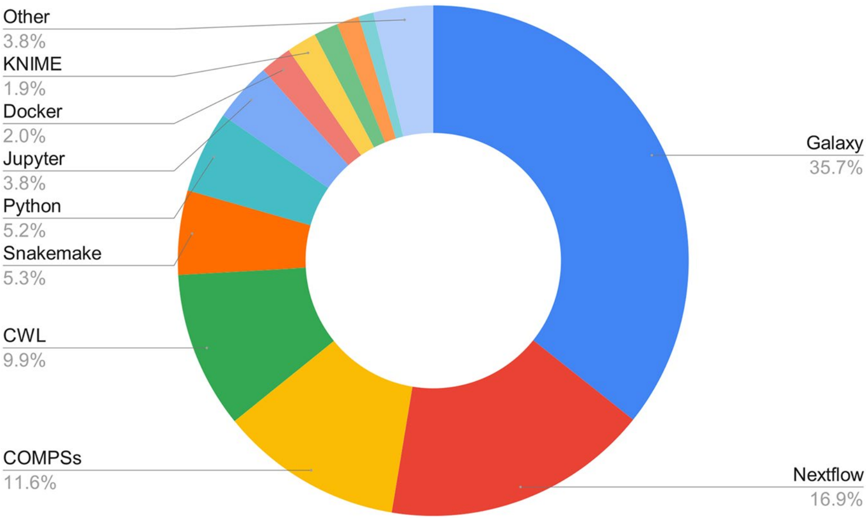


Fig. 3 Workflow types registered with WorkflowHub. See https://workflowhub.eu/workflow_classes for more details.

workflows can be constructed to help a community manage specific sets of relevant workflows, and other outputs, that may span many scientific applications, workflow languages and research programs (e.g. Threatened Species Initiative https://threatenedspeciesinitiative.com/about-tsi/ annotation workflow https://workflowhub.eu/collections/23). To streamline the onboarding process for consortia and larger projects, a generic set up guide[32], and multiple guides for specific consortia[33–35], have been created. These guides describe the steps required to get set up on WorkflowHub and use the platform effectively.

Platform / service /standard	URL / link
bio.tools	https://bio.tools/
Common Workflow Language (CWL)	https://www.commonwl.org/
DARE-UK	https://dareuk.org.uk/
DataCite	https://datacite.org/
EDAM	https://edamontology.org/
EuroScience Gateway	https://eosc.eu/eu-project/eurosciencegateway/
Galaxy	https://galaxyproject.org/
Git (i.e GitHub, GitLab)	https://github.com/, https://about.gitlab.com/
LifeMonitor	https://app.lifemonitor.eu/
LS-inc	https://lifescience-ri.eu/ls-login/
Research Object (RO)-Crate	https://www.researchobject.org/ro-crate/
Research Software Ecosystem (RSEc)	https://research-software-ecosystem.github.io/
Sapporo	https://github.com/sapporo-wes
Tools Registry Service (TRS)	https://www.ga4gh.org/product/tool-registry-service-trs/
Workflow Execution Service (WfExS)	https://github.com/inab/WfExS-backend

Table 1. List of platforms, services, and standards that are used by, or integrate with, WorkflowHub.

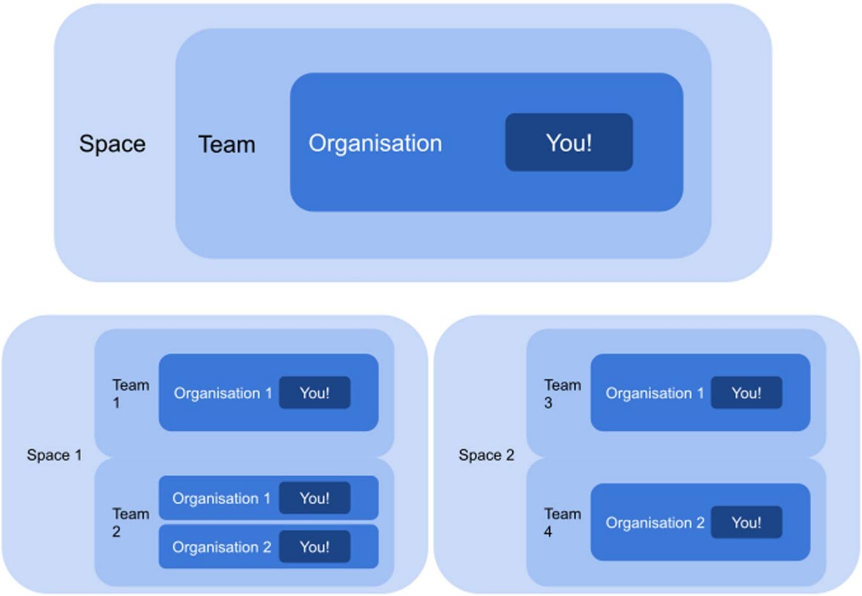


Fig. 4 A guide to the structures in WorkflowHub. You, the user, belong to one or more Organisations (i.e. affiliations). You can also belong to one or more Teams, each of which also needs to belong to a single Space (top). You can nominate which Organisations you wish to use for the different Teams that you have created or joined, and you can belong to multiple Teams in the same Space, as well as multiple Teams in other Spaces (bottom). Image reused with permission from WorkflowHub documentation.

A Hub for workflows. As the name of the registry suggests, its underlying aim is to act as a "Hub", and specifically one that helps to connect the workflows ecosystem and ease the interoperation of its constituent platforms and services. To realise this, WorkflowHub relies on a web-friendly metadata framework that simultaneously supports data representation within the registry itself, while also acting as a foundation for exchange of data and metadata between the platforms and services that are described in Fig. 1. Here, we provide more details for four core areas that underpin this "Hub" functionality[14].

WorkflowHub participates in the EOSC federated Authentication and Authorization Infrastructure through Life Science Login (LS-Login, https://lifescience-ri.eu/ls-login/) and the OAuth2 protocol, in addition to also supporting authentication via GitHub. This feature allows systems interacting with WorkflowHub to identify users as the same individual across different systems - even with their identity provided by their participating institutional accounts - enabling single-sign on and smart authorization decisions on accessing and operating on workflows, their metadata and other resources across the ecosystem forming around WorkflowHub.

To help describe workflows and their components, the registry uses three profiles from bioschemas[36]: Computational Tool (https://bioschemas.org/profiles/ComputationalTool/1.0-RELEASE), Computational

Workflow (https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE), and Formal Parameter (https://bioschemas.org/profiles/FormalParameter/1.0-RELEASE). Even though these are a part of the bioschemas effort, the profiles support a discipline independent and standardised way of describing workflows and their components. In addition, and owing to close collaboration with the Common Workflow Language (CWL)[37,38] community, the CWL workflow specification is encouraged as a workflow language independent way of describing a WorkflowHub entry. This is the so-called "Abstract CWL". This format can even hold semantic annotations, which WorkflowHub leverages to extract the typing of workflow inputs and outputs, as well as EDAM ontology concepts for Topics and Operations[39].

Research Object Crate (RO-Crate) is a standard for FAIR Research Objects[40]. It was developed by the community to package a workflow with the components required to understand and execute that workflow. The additional packaged components may include test data, Abstract CWL, diagrams, publications, and SOPs, as well as the flat metadata file that provides the context for all these assets (https://www.researchobject.org/ro-crate/)[40]. The implementation of RO-Crate is central to the ability of WorkflowHub to interoperate and exchange Workflow-RO-Crate (https://w3id.org/workflowhub/workflow-ro-crate/1.0) data with the ecosystem in Fig. 1 (e.g. Zenodo archiving) and for exposing workflows as FAIR Digital Objects[41,42]. RO-Crate therefore provides users with the option to easily download structured WorkflowHub content, export this content to other repositories, and even integrate additional platforms or services with the registry to automatically exchange content.

WorkflowHub can also interoperate with workflow execution platforms (see Fig. 1) through its implementation of the Global Alliance for Genomics and Health (GA4GH) Tools Registry Service (TRS) application programming interface (API, https://ga4gh.github.io/tool-registry-service-schemas/, https://www.ga4gh.org/product/tool-registry-service-trs/). This means that a user of a TRS enabled analysis platform, like Galaxy, is able to search for and retrieve workflows, without leaving the platform. It also means that WorkflowHub can trigger the import of workflows to these platforms, ready for execution.

The impact of these features is twofold. Firstly, users are able to access this ecosystem "Hub" using multiple authentication mechanisms and leverage multiple standards when contributing workflows. Secondly, WorkflowHub is able to support machine actionability, as described by the FAIR principles[15]. This underpins the registry's ability to connect to services and platforms that are in use day-to-day by workflow users and developers.

Register, find and launch workflows.

The primary purpose of WorkflowHub is to allow researchers to register and share workflows. WorkflowHub is therefore designed to store metadata about workflows, provide opportunities and encouragement to enrich this metadata (i.e. make workflows more FAIR), and then link to the native repositories that are used to develop, version, and store workflows long term. In other words, WorkflowHub seeks to faithfully represent and make visible the versioning and storage locations used by contributing developers and communities. This is also a key reason why manual upload and storage of files is important: the registry seeks to be inclusive, even when workflows are not supported by an underlying version control system. In these cases, WorkflowHub can act as a repository for files, and manages versioning on behalf of the user.

Existing public workflows can be viewed, downloaded and launched without needing to register as a user. By extension, the contribution of open access and publicly accessible workflows is encouraged. However, workflows may be registered privately, or be embargoed. This functionality supports cases where users would like to (i.e. a workflow is still being developed), or need to, limit access to a specific group of users.

User authentication (i.e. login) is required to register content with WorkflowHub, and enables the registry to assign credit and enable citation. To contribute content an individual user needs to:

1. **Register** and indicate the Organisation(s) to which they are affiliated. A user can add the following to their profile: a description, their Open Researcher and Contributor ID (ORCID, https://orcid.org/), contact details that are visible to those in shared Teams and Spaces, as well as knowledge and expertise. More advanced configurations are also available via a user profile, including the management of OAuth sessions, authorised applications, API applications, and API tokens.
2. **Decide which "Space" to use:** a Space is a user-administered section of WorkflowHub for managing the Teams required for consortia, institutes, or other large research activities. WorkflowHub administers a single default Space called "Independent Teams" for users that only need to create and manage a single Team. All other Spaces are created upon request and administered by those who requested the Space.
3. **Create or join at least one Team:** Teams are one or more people working on a particular research activity involving workflows. Every workflow in WorkflowHub is owned by at least one Team. WorkflowHub users must therefore belong to at least one Team, and this Team must belong to a Space. In addition to supporting the correct assignment of credit to workflow developers, contributors, and submitters, the Team also enables its members to further describe the context for their workflow development (i.e. background, project description), and serves to promote their contributions to other registry users.

Once these steps are complete, a user has the option to register:

- **Core resources** such as workflows and Collections. As workflows do not exist in a vacuum, Collections allow a WorkflowHub user to bring together workflows with any of their other resources and activities (see below) to create a visible and holistic resource that can support workflow reuse.
- **Other resources**, including publications, documents, data files, and SOPs.
- **Activities**, including presentations and events.

Registration of workflows can be carried out by manually uploading a workflow file, importing either a RO-Crate or Git repository, and by submission through a representational state transfer (REST) API (https://about.workflowhub.eu/developer/ro-crate-api/, https://about.workflowhub.eu/docs/adding-files/). Although WorkflowHub is primarily a registry, manual upload and storage of files is also supported, and in these cases WorkflowHub acts as a repository. RO-Crate is used by WorkflowHub as a fundamental unit that underpins upload, download, import and export. A user does not need to know how to create, or work with, RO-Crates as the registry automatically builds a crate when a workflow is registered. Wizards are used to streamline these processes, and at each stage users are prompted to carry out the next step (see https://about.workflowhub.eu/docs/registering_workflows/registering-a-workflow/). For example, after registration with WorkflowHub, a user is prompted to join or create a new Team. After workflow registration, the user is prompted to both indicate a main workflow file (e.g. **.nf* for nextflow workflows) and complete the metadata suggested for creating a well described workflow entry.

Two example WorkflowHub entries are provided, including the *dna-seq-varlociraptor* Snakemake workflow (Fig. 5, https://workflowhub.eu/workflows/686)[43], and the *Find transcripts - TSI* Galaxy workflow (Fig. 6, https://workflowhub.eu/workflows/877)[44]. The top of the *dna-seq-varlociraptor* entry emphasises the WMS (Fig. 5A) and workflow title (Fig. 5B), along with quick links to the development repository, requesting contact with the authors, subscribing to notifications about changes to the workflow, downloading a workflow RO-Crate, and adding the workflow to a collection (Fig. 5C). In addition, the workflow creator also has access to administrative options for the workflow entry in this section, which include adding new documents or presentations connected to the workflow, and workflow actions such as adding new versions, requesting a digital object identifier (DOI) for a specific workflow version, editing the workflow metadata, managing workflow contributors and visibility, and deleting the workflow entry. The main panel for the entry has three tabs (Fig. 5D) that provide a workflow overview (including descriptions in Fig. 5E, as well as the version information, metadata, critical annotations, and activity analytics described below), access and view capability for the files that were registered, and links to related items (e.g. people, Spaces, Teams, Collections, other workflows). The examples also use additional features of WorkflowHub, including those enabled by the registration wizard, which prompts inclusion of metadata: these include creators (Fig. 5G), component tools (Fig. 5H), license information (Fig. 5I) and ontology annotations (Fig. 5K). The *dna-seq-varlociraptor* workflow also used the WorkflowHub Git integration to ingest the repository README file (complete with badges), as well as providing the link to the development repository (Fig. 5C), the ability to access and view the repository file list natively in WorkflowHub (Fig. 5D), and an annotated version history (including commit IDs, Fig. 5F). Finally, WorkflowHub generates activity metrics for a workflow entry (Fig. 5J).

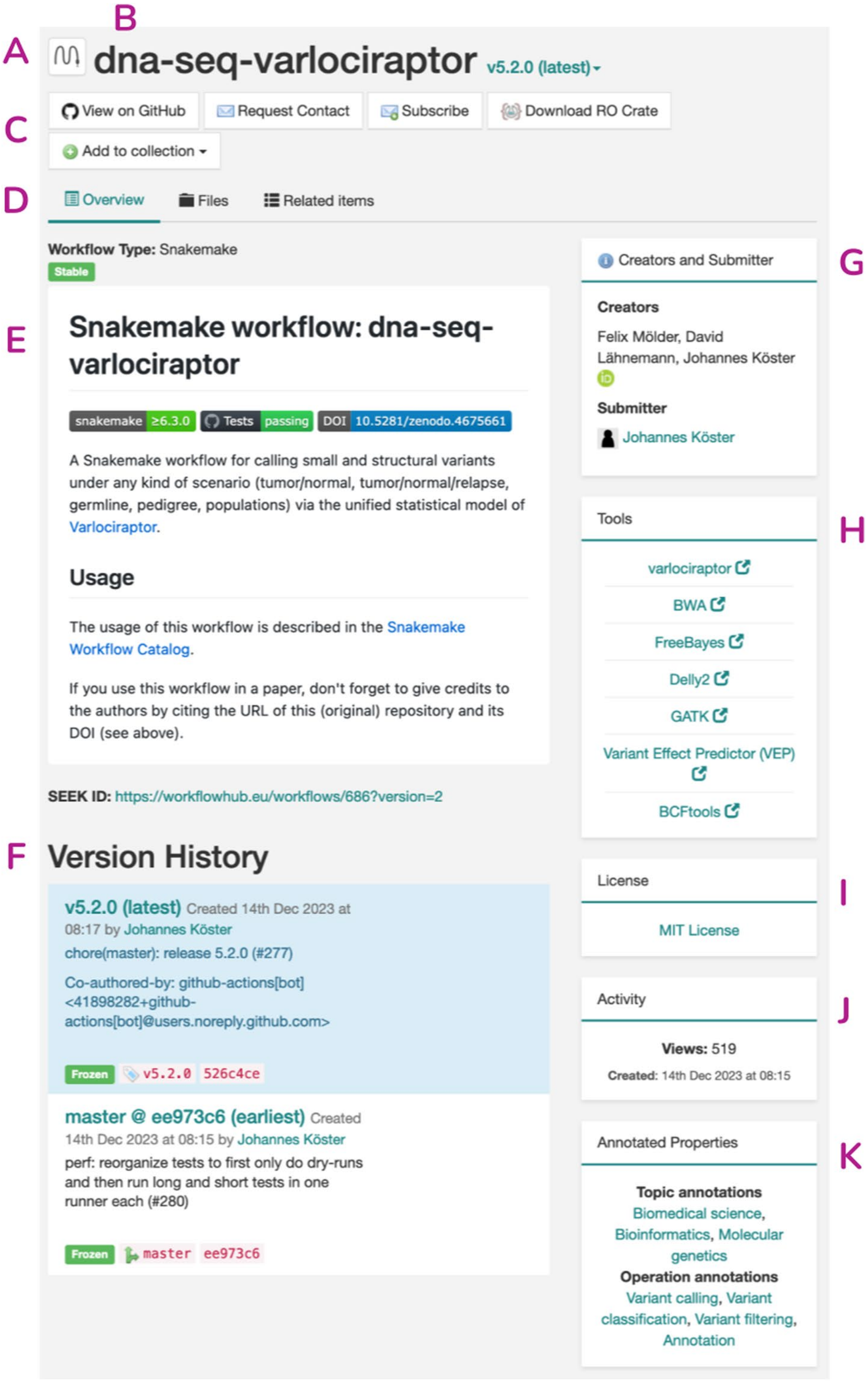
The *Find transcripts - TSI* entry illustrates additional features that can enrich a workflow, including adding a workflow diagram (Fig. 6A), as well as an annotated list of workflow inputs, outputs and steps (Fig. 6B). A citation can also be generated for a workflow. This can either use information from a minted DOI, or a custom citation such as a workflow publication (Fig. 6C). Finally, a user can add custom tags (Fig. 6D) to supplement any ontology annotations (see example of this in Fig. 5K), Collections that the workflow belongs to can be specified (Fig. 6E), and it is possible to include buttons for opening a workflow on an execution platform, in this case Galaxy (Fig. 6F). For a given entry a user can also access administrative drop down menus such as *Add new* (e.g. document, standard operating procedure SOP) and *Actions* (e.g. edit or manage the workflow, including versions and minting a digital object identifier, DOI).

To find these workflows within the registry, a user can apply a text-based search, or visit the complete workflow listing where they can filter by type (e.g. Galaxy, Snakemake), tools used (i.e. bio.tools identifiers[45]), creators, Organisations, Teams, Spaces, and more. Researchers can also refine their searches by making use of faceted browsing and filtering on tags and other annotations, and are able to sort by titles, dates, views and downloads.

Galaxy's integration with WorkflowHub leverages the GA4GH TRS API, enabling seamless workflow exchange. Researchers can discover, import, and run workflows from WorkflowHub directly within Galaxy. From the WorkflowHub interface users can utilise the "Run on Galaxy" button (see Fig. 6F), which redirects them to a Galaxy instance and a workflow run form. This interoperability facilitates immediate application and further development of workflows. The use of RO-Crate specification ensures that workflow metadata and components remain accessible and interoperable, aligning with FAIR principles.

Design that supports the workflow life cycle. To support the workflow life cycle[6,46,47], WorkflowHub integrates with services that workflow creators use for development, execution, maintenance, testing, citation, and ultimately archiving. These integrations initially form part of the workflow registration wizard, which guides a workflow creator through the process of registering their workflow for the first time. However, they are also accessible during a workflow's maintenance phase, when the workflow may be updated to modify, improve, or repair its function.

Ease of access. A user of WorkflowHub begins by accessing the service via LS-Login, using credentials from their specific institutions or even other identity-providing platforms (e.g. GitHub, Google, Apple, ORCID). Authentication via a local WorkflowHub account is also supported. Authentication enables WorkflowHub administrators to manage user access rights, and to create a custom combination of access levels that are suitable for specific user groups (e.g. research groups, consortia, international projects). For example, a contributor may simply want to register a single workflow to make it findable, and the contributor may be the only user that requires edit access for the workflow, or for the Team to which the workflow belongs. Access rights can be set accordingly. Conversely, a community of practice (e.g. nf-core), or consortium (e.g. BGE) may have multiple contributors, from multiple institutes, that also belong to multiple WorkflowHub Teams. In this case, granular permissions for edit rights can be set at the workflow, Team and Space levels.



dna-seq-varlociraptor v5.2.0 (latest)

Workflow Type: Snakemake

Snakemake workflow: dna-seq-varlociraptor

A Snakemake workflow for calling small and structural variants under any kind of scenario (tumor/normal, tumor/normal/relapse, germline, pedigree, populations) via the unified statistical model of Varlociraptor.

Usage

The usage of this workflow is described in the Snakemake Workflow Catalog.

If you use this workflow in a paper, don't forget to give credits to the authors by citing the URL of this (original) repository and its DOI (see above).

SEEK ID: https://workflowhub.eu/workflows/686?version=2

Version History

v5.2.0 (latest) Created 14th Dec 2023 at 08:17 by Johannes Köster

chore(master): release 5.2.0 (#277)

Co-authored-by: github-actions[bot] <41898282+github-actions[bot]@users.noreply.github.com>

master @ ee973c6 (earliest) Created 14th Dec 2023 at 08:15 by Johannes Köster

perf: reorganize tests to first only do dry-runs and then run long and short tests in one runner each (#280)

Creators and Submitter

Creators

Felix Mölder, David Lähnemann, Johannes Köster

Submitter

Johannes Köster

Tools

varlociraptor
BWA
FreeBayes
Delly2
GATK
Variant Effect Predictor (VEP)
BCFtools

License

MIT License

Activity

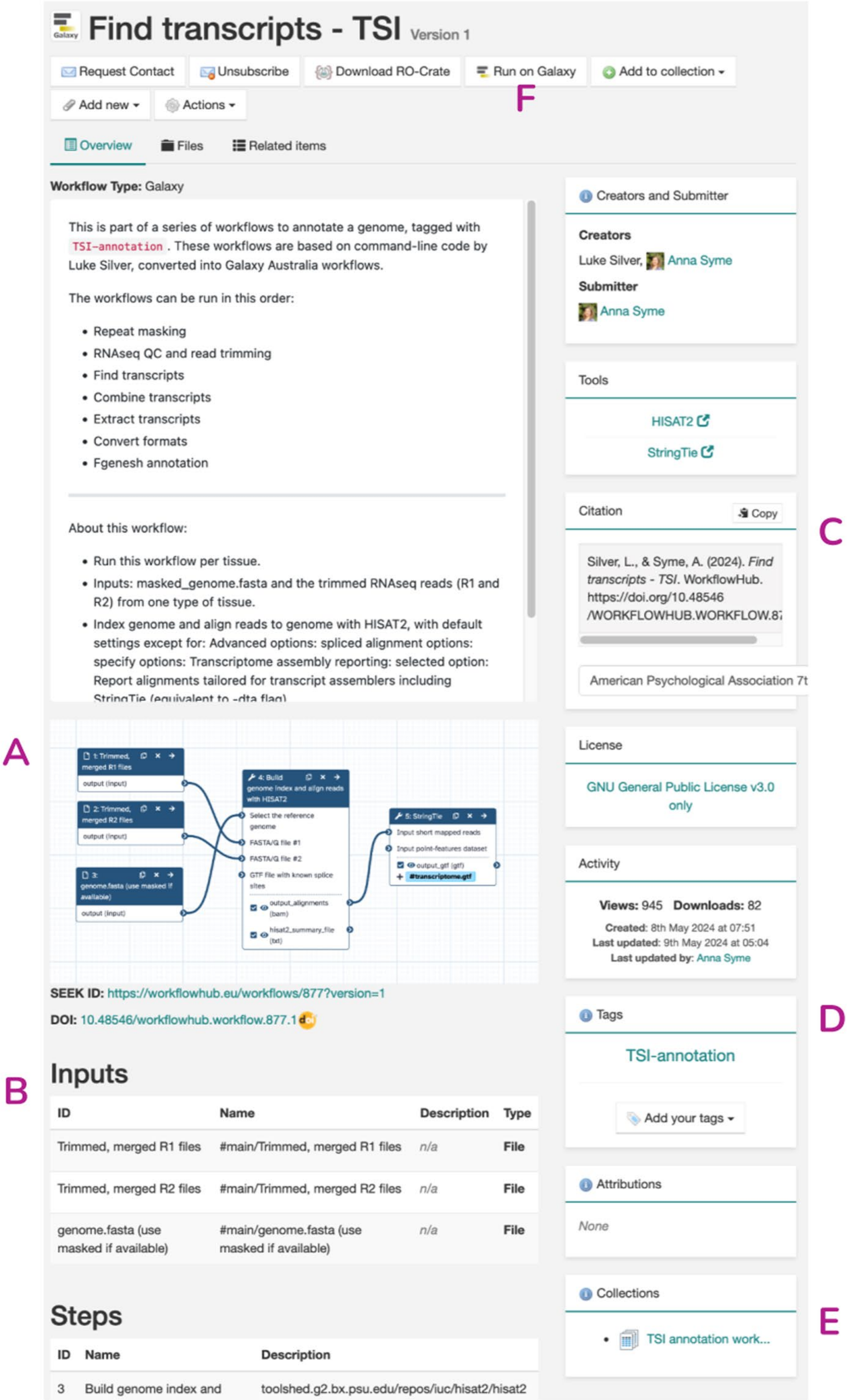
Views: 519

Created: 14th Dec 2023 at 08:15

Annotated Properties

Topic annotations
Biomedical science, Bioinformatics, Molecular genetics
Operation annotations
Variant calling, Variant classification, Variant filtering, Annotation

Fig. 5 WorkflowHub entry for the *dna-seq-varlociraptor* Snakemake workflow[43] with annotations for different sections of the user interface. Features include **(A)** workflow type, **(B)** title, **(C)** access panel with links to the source repository (e.g. GitHub), requests to contact the creators, subscribe / unsubscribe, download Research Object Crate (RO-Crate), and add to a Collection, **(D)** tabs for navigation between the entry overview, the list of files in the entry, and lists of items related to the workflow, including people, Teams, Spaces, Organisations, and other digital objects (e.g. publications, documents, SOPs, other workflows), **(E)** description, which can be imported from Git (*if available*), **(F)** version history, including Git commits, (*if available*), **(G)** creator and submitter information, **(H)** links to more information about tools that comprise the workflow (i.e. bio.tools registry entries), **(I)** license information, **(J)** activity metrics (i.e. downloads and views), and **(K)** ontology concept annotations (i.e. EDAM).



Find transcripts - TSI Version 1

Workflow Type: Galaxy

This is part of a series of workflows to annotate a genome, tagged with `TSI-annotation`. These workflows are based on command-line code by Luke Silver, converted into Galaxy Australia workflows.

The workflows can be run in this order:

- Repeat masking
- RNAseq QC and read trimming
- Find transcripts
- Combine transcripts
- Extract transcripts
- Convert formats
- Fgenesh annotation

About this workflow:

- Run this workflow per tissue.
- Inputs: masked_genome.fasta and the trimmed RNAseq reads (R1 and R2) from one type of tissue.
- Index genome and align reads to genome with HISAT2, with default settings except for: Advanced options: spliced alignment options: specify options: Transcriptome assembly reporting: selected option: Report alignments tailored for transcript assemblers including StringTie (equivalent to -dta flag)

SEEK ID: https://workflowhub.eu/workflows/877?version=1

DOI: 10.48546/workflowhub.workflow.877.1

Inputs

ID	Name	Description	Type
Trimmed, merged R1 files	#main/Trimmed, merged R1 files	n/a	File
Trimmed, merged R2 files	#main/Trimmed, merged R2 files	n/a	File
genome.fasta (use masked if available)	#main/genome.fasta (use masked if available)	n/a	File

Steps

ID	Name	Description
3	Build genome index and	toolshed.g2.bx.psu.edu/repos/iuc/hisat2/hisat2

Creators

Luke Silver, Anna Syme

Submitter

Anna Syme

Tools

HISAT2

StringTie

Citation

Silver, L., & Syme, A. (2024). *Find transcripts - TSI*. WorkflowHub. https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.87

American Psychological Association 7t

License

GNU General Public License v3.0 only

Activity

Views: 945 **Downloads:** 82

Created: 8th May 2024 at 07:51
Last updated: 9th May 2024 at 05:04
Last updated by: Anna Syme

Tags

TSI-annotation

Attributions

None

Collections

- TSI annotation work...

Fig. 6 WorkflowHub entry for the *Find transcripts - TSI* Galaxy workflow[44] with additional entry specific annotations. Features unique to this entry include **(A)** a workflow diagram, **(B)** parsed workflow inputs, outputs and steps for the registered Galaxy workflow, **(C)** a citation for the workflow, **(D)** custom tags, **(E)** Collections that include the current workflow entry, and **(F)** a button for launching the workflow on Galaxy.

Development and versioning. Integration with the Git version control system is a key aspect of WorkflowHub that supports workflow creation, development and maintenance. If the workflow registration wizard is provided with a Git repository URL, WorkflowHub will automatically import and parse its metadata. In this case, a workflow creator only needs to review, and potentially update, metadata fields prior to completing the registration. Git integration also supports workflows to remain in their native development environment, avoiding any impact of registration on the workflow's development and management process. Moreover, it allows for automation of tasks like updating WorkflowHub entries when workflows are versioned (e.g. using LifeMonitor (https://lifemonitor.eu/lm_wft_best_practices_github_app)[14]). For use cases like Galaxy, where workflows can be created via graphical user interface, WorkflowHub provides the option to manually upload a workflow file and step through the wizard manually to enter metadata. In this case, WorkflowHub itself manages the versioning of the uploaded workflow files. As described earlier, version history is provided at the bottom of each WorkflowHub entry, regardless of whether versioning is being managed externally (e.g. Git) or by the registry itself. Note that WorkflowHub does not store, or version, components of workflows that may be required for execution and which would typically be stored by a repository (e.g. BioConda[48], quay.io https://quay.io/). However, depending on the WMS being used, metadata about the tool versions and their dependencies may be recorded and included in WorkflowHub entries as part of the workflow file set (e.g. CWL).

WorkflowHub welcomes all workflows. Workflows come in all shapes and sizes, and may even be composed of multiple subworkflows. They may start off as a set of scripts and evolve into a heavily standardised and portable workflow[12]. They may use one of the many known WMS[2]. And, of course, workflows can span virtually every field in the sciences and beyond. In short, the workflow and WMS ecosystems are diverse. One role of WorkflowHub is to make these ecosystems transparent, and it does this by being agnostic to workflow language, maturity, source, structure, and even scientific quality. Contributions of every workflow type are encouraged (see https://workflowhub.eu/workflow_classes). Workflows at any development stage (i.e. work-in-progress) are encouraged, not just those workflows considered to be mature and stable. An indication of the maturity of a workflow can be assigned by the creators, and naturally, this metadata is presented to workflow consumers in the registry entry to help identify workflows that are suitable for reuse.

Annotating workflow purpose (i.e. adding metadata). The registration wizard guides users in provision of metadata. Although Bioschemas metadata profiles are used, the only mandatory metadata fields are the workflow Title and the contributing Team(s). In addition to describing the workflow itself, the metadata wizard can be used to associate a workflow to other workflows, presentations, publications, documents, SOPs and data files.

Two key metadata integrations are in place that allow users to search for and add standard identifiers when editing workflow metadata. This functionality is available for bio.tools software identifiers (https://bio.tools/) and the EDAM ontology concept identifiers for both Topics (e.g. genomics) and data transformation Operations (e.g. genome assembly)[39]. A user can therefore annotate their workflow with persistent links to registry metadata about software that the workflow contains, and add standardised shorthand terms that describe its application area and function. In the case of Galaxy, the standard structure of the workflow file is used to extract software tool identifiers and map these automatically to bio.tools[49]. A user can build on these integrations by manually including custom tags and keywords.

Discovery and understanding. When a workflow is ready to be shared and reused, a creator can update their workflow maturity from "work-in-progress" to "stable" in the WorkflowHub entry metadata. It is at this point that a workflow needs to be discoverable in multiple ways, and remain accessible in its original location. The reason for this is that not all researchers will necessarily seek to discover workflows in the same way. As a result, WorkflowHub is flexible in its approach. As mentioned earlier, within the registry itself a user can start by applying direct search and filtering to find workflows. External to the registry, its machine-readable, standardised format (i.e. Bioschemas) increases the search engine visibility of the metadata in a workflow entry. You can interactively explore the impact of workflow registration using evaluator tools such as FAIR-checker[50] and FAIRsoft[51].

Finding a workflow is step one for a potential user. Once found, the WorkflowHub entry metadata supports a user to understand the workflow, including its design, content, and purpose[7]. For example, "this is a Galaxy type workflow, containing these tools (i.e. links to bio.tools), which are capable of these types of data transformations (i.e. EDAM annotations)". From a workflow entry that has been annotated with tools, a user can access links to navigate directly to the referenced bio.tools entries to explore and further understand the components that comprise the workflow. Users can also view the files in the source Git repository (if the workflow was imported from Git), and with a single click visit the repository. It is even possible to subscribe and be notified of changes to entries, removing the need for constant monitoring.

Execution / reuse. As a registry, WorkflowHub does not guarantee long term re-executability of workflows, as this is dependent on the design of a workflow and the other elements of the workflow ecosystem that it uses (see Fig. 1). However, WorkflowHub's design still aims to support the reuse of workflows in multiple ways.

Firstly, WorkflowHub actively supports and develops integrations with workflow execution platforms and services. A key example is the GA4GH TRS API (https://github.com/ga4gh/tool-registry-service-schemas). If an execution platform or system adopts the TRS standard, it can search WorkflowHub for suitable workflows, retrieve those workflows, and execute them, without the need to develop custom integrations with the workflow's native repository. Examples include, Galaxy, Sapporo[52] (DNA Data Bank of Japan, DDBJ, https://ddbj.nig.ac.jp/), and the Workflow Execution Service (WfExS, https://github.com/inab/WfExS-backend)[53], all of which implement the TRS API, either as providers or consumers. As execution platforms (e.g. Galaxy) can make use of the TRS, they are also able to provide inbuilt search interfaces that connect to WorkflowHub and support platform users to find and import a specific workflow. However, WorkflowHub does not currently record requirements for deployment beyond that contained within a WMS standard and the associated files and descriptions provided by a workflow creator.

WorkflowHub also provides mechanisms to connect workflows to data. For example input data types can be connected, either by direct upload or by reference in the entry metadata. Workflows may also be linked to external data repositories as "Related items". Registering a Workflow Run RO-Crate[54], an extension of RO-Crate, also provides a mechanism to share a workflow instance, whereby a specific execution and its provenance are captured (see example here: https://workflowhub.eu/workflows/1072). This provides the provenance of derived data created by a workflow, which enhances reusability when 1) the workflow specifically supports a publication or 2) the workflow provides test data to run validations for benchmarking.

The connection of WorkflowHub to the LifeMonitor service (https://www.lifemonitor.eu/), through the LifeMonitor GitHub app, allows workflow function and status to be reported to maintainers and users through regular automated tests driven by continuous integration (CI) based monitoring (e.g. Planemo automated workflow testing using Galaxy[55]). In these cases, WorkflowHub will also include a badge that shows if the tests are passing or failing. An embedded link in the badge takes a user to the LifeMonitor page for the workflow, providing information on the reliability of the workflow over time and the timeliness of the workflow maintainers in solving issues as they arise. In addition, the app can automatically suggest changes to the workflow Git repository that will bring its metadata and structure in line with best practices. To streamline the inclusion of reuse conditions, licensing information is included in the metadata for workflows registered with WorkflowHub, default sharing and license conditions can be specified for Teams, and these rules can be updated by Team administrators.

Finally, WorkflowHub focuses on enabling workflow re-implementation. This is in contrast to re-execution, which can be challenging for a workflow, particularly in the long term. The idea is that even when a workflow can no longer be executed, or when it is no longer accessible via its original source, its metadata is available and can be both reviewed and understood. The workflow can thus be reimplemented if necessary. WorkflowHub enables this FAIR aligned approach by emphasising the collection of metadata (e.g. using in-built wizards, standards, and WMS integrations). However, it should be noted that the ability of the registry to aid re-implementation depends on the richness of the metadata annotations provided by users.

Attribution and citation. It is important to properly attribute contributions to workflows, and this includes provenance of the entire workflow development process. The RO-Crate format used by WorkflowHub allows for provenance tracking of metadata, ensuring that workflow creators are given the credit they deserve, while also adding accountability. Workflows can be linked to each other using WorkflowHub metadata (i.e. the "attribution" metadata can be used to indicate that a workflow is based on another workflow). The Git integrations described above also support the citation standard CITATION.cff (https://citation-file-format.github.io/)[56], so that WorkflowHub can import this file and use its contents to populate workflow entry metadata for creators, including their ORCID. This approach simplifies citing a workflow according to the wishes of the workflow developer. Once a workflow is registered and credit is established within the metadata framework of WorkflowHub, the registry can also, at the push of a button, use DataCite (https://datacite.org/) to mint persistent DOIs for workflows and contribute to the DataCite persistent identifier (PID Graph (https://support.datacite.org/docs/datacite-graphql-api-guide). This is the first step to ensuring that workflows can be cited effectively, increasing their visibility and potential impact, and supporting inclusion in scholarly knowledge graphs.

User engagement and training.

WorkflowHub engages and supports a broad set of use cases, including numerous projects and consortia of global significance. Major projects are supported so that their members are able to use WorkflowHub in a way that aligns with the expectations of their project and its funders. Workflow communities (e.g. nf-core, CWL, Snakemake) are directly engaged by the WorkflowHub team, and supported to make their best practice workflows available in the registry.

The main mechanism through which engagement happens is the fortnightly open format WorkflowHub Club meeting (https://about.workflowhub.eu/#community), where anyone can join the conversation, learn more about the registry, ask questions, and even contribute to the on-going development and evolution of the registry. There are also indirect ways through which workflow creators and users interact with WorkflowHub and its resources. The WorkflowHub Club team creates and presents content for the registry at conferences, in webinars and workshops, and as part of Ask-Me-Anything forum events (https://about.workflowhub.eu/project/outreach/). Registered users of WorkflowHub are also able to ask questions and provide direct feedback via the registry interface. These communications are sent directly to the administrators of WorkflowHub for review and response. Finally, WorkflowHub operates a documentation site where users can access information on how to use the registry (https://about.workflowhub.eu/docs/).

Clear and practical guidelines are required to support users of WorkflowHub. For example, some annotations are consistently missing from workflow entries, and useful features are sometimes overlooked (e.g. Git integration, parsing of CITATION.cff files, linking data to workflows, and linking workflows to each other). This highlights that the rich feature set of WorkflowHub is not necessarily immediately clear, and that guidance in leveraging these features is absolutely critical to support users in achieving best practice. It is important that the provided guidelines also extend to recommendations for how to organise a workflow development repository (i.e. Git repositories). This will enable WorkflowHub to extend the process currently in place for Galaxy IWC to more community repositories: integrating with these repositories in a more standard way to automatically manage the update of workflow versions in WorkflowHub, such that they are in sync with Git releases. As a result, and as highlighted earlier, a general onboarding and set up guide for projects and consortia has been developed[32], as have multiple consortia specific guides[33–35]. Workflow resources have also been developed for the Galaxy Training Network (GTN[57]) Smörgåsbord events (https://gallantries.github.io/video-library/modules/ro-crate) and specific GTN tutorial sets (https://training.galaxyproject.org/training-material/topics/fair/).

Finally, the WorkflowHub team actively identifies opportunities to engage with peer infrastructures to grow the user base of the registry, investigate and create integrations that are of enduring value, and further improve the function of the registry. For example, WorkflowHub is actively fostering a conversation with publishers and journals focused on how to make workflows citable objects in the literature (https://galaxyproject.org/news/2024-08-03-workflow-publisher-forum/). WorkflowHub is also being further developed to fully align with the best practice guidelines of the SciCodes Consortium[58], implement the FAIR Principles for Research Software (FAIR4RS)[16] and contribute to the CodeMeta specification for software (https://w3id.org/codemeta/v3.0). As described earlier, a registry like WorkflowHub enables the creation of workflows that follow the FAIR principles, from the perspective of data[15], software[16], and the unique features of workflows (e.g. abstraction, composition). WorkflowHub is central to discussions in the FAIR Computational Workflows working group for the Workflows Community Initiative (WCI, https://workflows.community/groups/fair/). This working group has engaged across the global workflows ecosystem (i.e. workflow developers, communities, platforms and services) to develop FAIR principles for workflows[18]. The ultimate aim is to use the outcomes of these engagements to guide the evolution of WorkflowHub as a registry that encourages and supports workflows to be more FAIR.

Use Cases. To effectively support the sharing of workflows, WorkflowHub supports collaborations and communities of practice within the sciences. WorkflowHub contributions span domains such as cancer, COVID, genomics, rare diseases, geosciences, climate, physics, and more. WorkflowHub users span the globe and 36 countries are represented in the registered user list (https://workflowhub.eu/people).

Research consortia & infrastructures. WorkflowHub is an integral platform for consortia and projects. Here we provide details for three specific use cases, EOSC-Life, BGE, and Australian BioCommons.

EOSC-Life was a key use case driver, as it supported the implementation of FAIR computational workflows in the EU by seeking to develop a cloud-based Workflow Collaboratory[27] that ultimately resulted in the creation of WorkflowHub. The aim was to create a platform that would support community collaboration on the development, use, and reuse of FAIR computational workflows[19], and to do so in a way that bridges research domains and infrastructures[14,27]. WorkflowHub accommodates the diversity of EOSC-Life and ensures the visibility of workflows applied across its many established research infrastructures as they are created and registered[27].

The BGE project is a coming together of two communities of researchers with a common goal of cataloguing biodiversity through genomic resources: the European Reference Genome Atlas (ERGA)[59] and the European node of the International Barcode of Life consortium (iBOL Europe, https://iboleurope.org/). Providing reference-quality genomes to the community (i.e. ERGA), and monitoring biodiversity through DNA barcoding (iBOL Europe), requires the management and processing of vast amounts of data, in an accessible and distributed fashion, relying on input from multiple individuals and institutes. The combination of a BGE WorkflowHub Space (https://workflowhub.eu/programmes/25), Teams (https://workflowhub.eu/projects/163) and a Collection (https://workflowhub.eu/collections/10) allows individuals to contribute as needed to the projects across the consortium. As workflows have been collected and curated by the community, they in effect also come with a "seal of approval" for external users that wish to replicate the work of ERGA or iBOL Europe. All together, such a structure of publishing and maintaining workflows facilitates BGE in achieving their ambitious goals of cataloguing biodiversity in Europe and bringing together researchers from the biodiversity genomics community.

Australian BioCommons (https://www.biocommons.org.au/) is a national infrastructure project that actively supports life science research communities with community scale digital infrastructure[60]. Rather than building anew, BioCommons aims to adopt fit-for-purpose international platforms and services that, in the case of workflows, can assist with the provision of sophisticated software, analysis capabilities, and digital asset stewardship. WorkflowHub is the primary workflow registry for BioCommons (see https://workflowhub.eu/programmes/8), and it is the focal point for sharing the collaborative workflow efforts of BioCommons and its infrastructure partners together with Australian life science researchers (e.g. https://workflowhub.eu/collections/6).

Workflow management systems. WorkflowHub accepts all workflow types, and includes Galaxy[24], Snakemake[25], Nextflow[22], job schedulers (e.g. Python COmputational Modeling Platform Service, PyCOMPSs[61], application-specific types like SCIPION[62], notebooks (e.g. Jupyter, https://jupyter.org/), and even scripting languages (e.g. R[63] and Python (https://www.python.org/)[64] (see also Fig. 3). WorkflowHub provides customised support for WMS that are critical for specific domain communities (e.g. bioinformatics). This is currently the case for Galaxy, CWL and Nextflow, and additional significant and popular WMS may be supported with relevant features, when appropriate. As an example, the registry development team, together with Galaxy, have created functionalities for WorkflowHub that include 1) semi-automated registration of new and updated Galaxy IWC workflows, 2) integration with LifeMonitor to further support semi-automated registration but also Planemo workflow test monitoring, and 3) automatic mapping of tool identifiers to the bio.tools registry[14]. As the CWL community has been closely involved in registry development, WorkflowHub also has in-built functions for parsing CWL, and makes use of Abstract CWL. Most recently, engagement between nf-core and the WorkflowHub team resulted in the creation of automatic registration and metadata parsing functions for this community's Nextflow workflows (https://elixiruknode.org/news/2024/workflowhub-nf-core-workflow-accessibility/). These workflows can now be found in WorkflowHub (https://workflowhub.eu/workflows?filter[project]=15).

Individual workflow developers. WorkflowHub supports large international science missions and consortia. However, the registry welcomes and encourages contributions from any workflow developer, regardless of their research domain, application areas, or the size of their research group. In fact, the largest Space on WorkflowHub is currently Independent Teams (https://workflowhub.eu/programmes/3).

Discussion

We have presented WorkflowHub, a registry that enriches the scientific workflows ecosystem by being a hub for discovery and sharing of workflows from across multiple languages, communities, consortia, and scientific domains. The registry connects this community of users and contributors to workflow development, support, and scholarly services that support the requirement for workflows to be shared and credited, as well as the various requirements developers encounter during the workflow life cycle (see Fig. 1). Workflows continually evolve to keep pace with changing research questions, data types and practices. Similarly, it is intended for WorkflowHub to evolve in lock step with the changing requirements of both the developers and users of computational workflows. The roadmap for WorkflowHub over the next few years can be broadly broken down into four aspects: improving the support and resources available to users of the registry, onboarding new communities and domains, contributing thought leadership for workflow best practices that directly impact aspects such as workflow visibility and quality, and aligning to FAIR principles for computational workflows.

Improving support. As WorkflowHub and its integrated services intend to support the workflow life cycle, and not simply the registration and publication of workflows, the registry will aim to improve processes that align to supporting this life cycle. This includes improving search and the inbuilt wizards that assist with onboarding and resource registration, but also improving overall ease-of-use and guidance. Additions and improvements for the user interface (UI) will continue, adding those features that are required by end users. Particular attention will be paid to tracking workflow cloning, as well as supporting workflow collections, sub-workflows, and nested workflows.

Continued work on the onboarding guidelines created for WorkflowHub will improve the quality of workflow organisation for contributing groups (i.e. big projects and consortia), in part by including guides that cover instructions for the structure and content of workflow repositories as well as how to make the most impact with LifeMonitor, but also by contributing to training. To date, WorkflowHub has been included in existing training, such as the GTN Smörgåsbords (https://training.galaxyproject.org/) and workflow registration workshops[65]. As a next step, this effort should be extended to include WorkflowHub lessons in Software Carpentry developed by the Edinburgh Data Science training programme for Health and Biosciences (Ed-DASH, https://edcarp.github.io/Ed-DaSH/) for Nextflow (https://carpentries-incubator.github.io/workflows-nextflow/) and Snakemake (https://carpentries-incubator.github.io/snakemake-novice-bioinformatics/).

Given the intertwined nature of data and workflows, it will be important to explore how to encourage developers to make use of the mechanisms available (e.g. registering data resources, Workflow Run RO-Crate) in WorkflowHub to contribute information about the data requirements of their workflows (i.e. reference, inputs, intermediates, and outputs). This would impact the ability of users to find and reuse workflows based on their data requirements, and if expanded may even support greater interoperability between WorkflowHub and sources of FAIR data. Additional options to explore alongside our developer communities may include expanding ontology annotations to also encompass EDAM data and format concepts, enriching the data annotation options in the registration wizard, and updating the UI to streamline annotation processes.

As the number of contributors and workflow assets grow, it will be critical to also explore the use of automated mechanisms that allow scaling of support. This could include streamlined metadata annotation approaches that use Large Language Models (LLMs) to populate metadata fields for review by a workflow creator, automatic reporting of issues to workflow creators and their Teams, or integration with additional added value services like APICURON[66].

Reflecting their importance to the operation of WorkflowHub, integrations with platforms and services that support the workflow life cycle will be added, improved, and updated. The RO-Crate format used by WorkflowHub will be central to this effort. Planned updates include the ability to natively configure automated synchronisation between Git repositories and WorkflowHub. Setting this up to match the expectations of the community of WorkflowHub users will be essential, and will depend on registry and community co-development, in particular for those cases where work involves the repositories used by WMS communities. In addition, the ability to "launch" workflows on additional instances that implement TRS will be explored (e.g. Seqera Platform, https://seqera.io/platform/), and benchmarking service integrations for workflow entries will be added through collaboration with OpenEBench (https://openebench.bsc.es/)[67].

Finally, the ability to connect to additional registries and repositories that maintain versioned archives of code and other software used in workflows will be explored. In the first instance, these additional connections could include BioConda and BioContainers[68], with integrations being dependent on consultation with the domain communities that use WorkflowHub.

Onboarding. WorkflowHub has already supported the onboarding of WMS communities, including Galaxy's IWC and Nextflow's nf-core. A similar integration is currently being finalised for the GTN. In total, these community-centric engagements will result in the addition of hundreds of workflows to the registry and this number will grow as a result of the integrations with the workflow source repositories. WorkflowHub will also engage with additional scientific domains beyond the life sciences, to encourage and support the sharing of all computational workflows. Support provided to new communities may entail 1) integrations with community repositories, facilitating semi-automated ingestion of workflows, and synchronised registration of workflow releases, and 2) supporting communities to make the best use of WorkflowHub, including guidance on contributing to the registry, standard repository structures and the adoption of both standards (e.g. RO-Crate, Abstract CWL, TRS) and services (e.g. LifeMonitor). This level of support is currently being explored for the Snakemake community. Onboarding new domains, disciplines, and WMS communities will be critical to increasing the visibility of the broader workflows ecosystem, and ensuring that the user requirements for WorkflowHub are collected from a diverse stakeholder set.

The registry also plans, through discussions with our users, to introduce multi-value fields for tagging; for example, primary (astronomy, biology), and secondary (genomics, transcriptomics). Including this level of metadata will allow WorkflowHub to measure how effectively different domains and disciplines are being onboarded.

Workflow visibility and recognition. WorkflowHub already has the capacity to import citation, author and contributor credit metadata from CITATION.cff files. In future, this could be extended to other popular standards, including codemeta.json (https://w3id.org/codemeta/v3.0). Through DataCite, WorkflowHub has the means to mint DOIs, and incorporate workflows with DOIs into OpenAIRE (https://www.openaire.eu/). WorkflowHub has thus actively worked to increase the visibility of workflows in a standardised and streamlined fashion. This approach is already bearing fruit, with examples of WorkflowHub formatted citations appearing in the published literature[69,70]. In addition, there are now examples of journals making computational workflows the focus of published works. A critical outcome here is to ensure that WorkflowHub becomes a recommended registry for journals and publishers. This includes providing workflow creators and users with a set of best practice recommendations for how to properly document and ultimately cite workflows in published research. A forum has already been established with multiple publishers and journals, and these continuing conversations will aim to address the challenges (i.e. citation formats, complexity of workflow citations, impact on publishing system, recommended practices for peer review of workflows) and opportunities (i.e. developer recognition, tracking, reproducibility) presented by formal citation of computational workflows.

Hand-holding for FAIR principles. As frequently alluded to in other sections, WorkflowHub is strongly connected to the FAIR principles at numerous levels. One of these levels that is of particular importance for users is the way that it "holds users' hands" during the process of sharing workflows and related research artefacts. The FAIR principles are only guidelines, and even the most well-intentioned attempts to follow them can go awry in unexpected ways[71]. WorkflowHub aims to help the user to follow best practices, including following the FAIR principles, by making them convenient, but not imposing them as requirements.

For example, WorkflowHub helps users make their workflows Findable - easy to find for both humans and machines - by assigning globally unique and persistent identifiers for the workflows and their different versions (i.e. WorkflowHub identifiers and DOIs). WorkflowHub also guides users to describe their workflows with rich metadata, including the identifier for the workflow, and these metadata are automatically exposed for indexing through WorkflowHub's use of Bioschemas.

WorkflowHub similarly helps users make their workflows Accessible - available to humans and machines over open protocols that provide optional access control - by providing multiple APIs over hypertext transfer protocol secure (HTTPS). These APIs include the JSON-based FAIRDOM-SEEK API (https://workflowhub.eu/api), an RO-Crate Submission API (https://about.workflowhub.eu/developer/ro-crate-api/), and the TRS API. Workflow DOIs can also be minted, which means the workflow and its metadata will be accessible even if the workflow itself is no longer available or if the workflow itself cannot be shared openly.

Workflows, after being found and accessed, should ideally be Interoperable - able to be used by humans and machines as part of a wider computational ecosystem. Interoperability often requires a lot of "plumbing" that WorkflowHub provides for users automatically through the use of open-source standards (i.e. RO-Crate as the primary data exchange format) and domain- and tool-specific integrations (i.e. Galaxy IWC and Nextflow nf-core). By guiding users through the process of inputting metadata, WorkflowHub reduces complexity and tedium, making it significantly easier to create interoperable workflows.

Finally, WorkflowHub helps users ensure their workflows are Reusable - allowed to be used in part or in entirety by other humans and machines. In particular, it does this by providing opportunities to specify a clear and accessible license, qualified references to other software, and detailed provenance. WorkflowHub also collaborates closely with prominent communities in the computational workflows space (see the Use Cases section) so that the registry can accommodate and incorporate domain-relevant community standards.

Methods

Governance. WorkflowHub is an ELIXIR service supported by the United Kingdom (UK) and Belgium ELIXIR Nodes (https://elixir-europe.org/)[72] as well as Australian BioCommons (https://www.biocommons.org.au/)[60]. The registry is part of both the ELIXIR Tools Platform (https://elixir-europe.org/platforms/tools) and Research Software Ecosystem (https://research-software-ecosystem.github.io/), and forms part of both EuroScienceGateway (https://esciencelab.org.uk/projects/eurosciencegateway/) and Data and Analytics Research Environments UK (DARE-UK) TRE-FX (Trusted Research Environments Five Safes, https://esciencelab.org.uk/projects/tre-fx/). Governance is coordinated and managed by the WorkflowHub Club (https://about.workflowhub.eu/project/community/): an inclusive community that meets biweekly online and consists of workflow developers/creators and workflow users, as well as WorkflowHub developers and product owners. The minutes for these meetings are open and a GitHub organisation is used to manage documentation (https://github.com/workflowhub-eu/about). Club members include representatives from ELIXIR, WCI, Australian BioCommons, and more. Over 60 people are listed as contributors (https://about.workflowhub.eu/project/acknowledgements/#workflowhub-club), and any new contributors are invited to join. WorkflowHub has secured funding for sustainability through Horizon Europe projects, ELIXIR Europe and national UK funds.

Technical infrastructure. WorkflowHub is developed openly, and largely virtually, using open software development practices, hackathons, and virtual communication channels. It has both a roadmap (https://about.workflowhub.eu/project/roadmap/) and regular release cycle (i.e SEEK release cycle, https://github.com/

seek4science/seek/). WorkflowHub requests for Team registration are supported for the American and Asia Pacific time zones by Oak Ridge National Laboratory (https://www.ornl.gov/) and Australian BioCommons, respectively. WorkflowHub is built on the FAIRDOM-SEEK software[73]. FAIRDOM-SEEK was originally developed as a data-management platform for the systems biology community, but has been generalised over time to support a wide variety of use cases, and now has numerous deployments across the world supporting many different communities. Development of FAIRDOM-SEEK is a collaborative activity with contributors from institutions in the UK, Germany, Belgium, Sweden and elsewhere. WorkflowHub is currently hosted on the University of Manchester's Research IT cloud. WorkflowHub makes use of Git to store workflows as repositories. This enables the addition, modification and deletion of files for workflows that are uploaded directly to the registry by a user, as well as the ability to freeze "snapshots" of specific workflow versions. If a workflow has been added either via Git import or through submission of an RO-Crate, it then needs to be explicitly versioned and uploaded again. Curation of registered workflows is the responsibility of the workflow creators and / or submitters.

Data availability

WorkflowHub (https://workflowhub.eu) is long-term supported and sustained for the international community by the ELIXIR European Research Infrastructure for Life Science Data and the Australian BioCommons. It is a service of the European Open Science Cloud. Should WorkflowHub reach its end of service, in accordance with Trusted Repository best practice, the published contributions and metadata will be archived as RO-Crates and made available through the public repository Zenodo, and DOI registrations (https://about.workflowhub.eu/docs/citable/) will in this case be updated to link to the archived deposits (https://about.workflowhub.eu/#retention-and-end-of-life-policy). A knowledge graph of registered Workflow RO-Crates as of August 2024 is also published on Zenodo[74].

Code availability

WorkflowHub code is available as part of the FAIRDOM-SEEK project (https://github.com/seek4science/seek)[75].

Received: 5 November 2024; Accepted: 7 March 2025;

References

1. Ferreira da Silva, R. *et al*. Workflows Community Summit 2022: A Roadmap Revolution https://doi.org/10.5281/zenodo.7750670 (2023).
2. Amstutz, P., Mikheev, M., Crusoe, M. R., Tijanić, N. & Lampa, S. Existing Workflow systems. Common Workflow Language wiki. https://s.apache.org/existing-workflow-systems.
3. Maier, W. *et al*. Ready-to-use public infrastructure for global SARS-CoV-2 monitoring. *Nature Biotechnology* **39**, 1178–1179, https://doi.org/10.1038/s41587-021-01069-1 (2021).
4. Freudling, W. *et al*. Adaptive Data Reduction Workflows for Astronomy – The ESO Data Processing System (EDPS) https://doi.org/10.48550/ARXIV.2311.03822 (2023).
5. McClure, J. E. *et al*. in *Toward Real-Time Analysis of Synchrotron Micro-Tomography Data: Accelerating Experimental Workflows with AI and HPC* (eds Nichols, J. *et al*.) *Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI*, Vol. 1315 226–239 (Springer International Publishing, Cham, 2020). https://doi.org/10.1007/978-3-030-63393-6_15.
6. Gil, Y. *et al*. Examining the Challenges of Scientific Workflows. *Computer* **40**, 24–32, https://doi.org/10.1109/MC.2007.421 (2007).
7. Cohen-Boulakia, S. *et al*. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems* **75**, 284–298, https://doi.org/10.1016/j.future.2017.01.012 (2017).
8. Hernández, J. A. & Colom, M. Repeatability, reproducibility, replicability, reusability (4r) in journals' policies and software/data management in scientific publications: A survey, discussion, and perspectives https://arxiv.org/abs/2312.11028 (2023).
9. Lariviére, D. *et al*. Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nature Biotechnology* **42**, 367–370, https://doi.org/10.1038/s41587-023-02100-3 (2024).
10. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. *et al*. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93, https://doi.org/10.1038/s41586-020-1969-6 (2020).
11. Reiter, T. *et al*. Streamlining data-intensive biology with workflow systems. *GigaScience* **10**, giaa140, https://doi.org/10.1093/gigascience/giaa140 (2021).
12. Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods* **18**, 1161–1168, https://doi.org/10.1038/s41592-021-01254-9 (2021).
13. Patel, R. *et al*. Reproducibility of the First Image of a Black Hole in the Galaxy M87 from the Event Horizon Telescope (EHT) Collaboration https://doi.org/10.48550/ARXIV.2205.10267 (2022).
14. Goble, C. *et al*. EOSC-Life Implementation of a mechanism for publishing and sharing workflows across instances of the environment https://doi.org/10.5281/zenodo.7886545 (2023).
15. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
16. Barker, M. *et al*. Introducing the FAIR Principles for research software. *Scientific Data* **9**, 622, https://doi.org/10.1038/s41597-022-01710-x (2022).
17. Huerta, E. A. *et al*. FAIR for AI: An interdisciplinary and international community building perspective. *Sci. Data* **10**, 487, https://doi.org/10.1038/s41597-023-02298-6 (2023).
18. Wilkinson, S. R. *et al*. Applying the FAIR principles to computational workflows. *Scientific Data* **12**, 328, https://doi.org/10.1038/s41597-025-04451-9 (2025).
19. Goble, C. *et al*. FAIR Computational Workflows. *Data Intelligence* **2**, 108–121, https://doi.org/10.1162/dint_a_00033 (2020).
20. Gil, Y. From data to knowledge to discoveries: Artificial intelligence and scientific workflows. *Scientific Programming* **17**, 167604, https://doi.org/10.3233/SPR-2009-0261 (2009).
21. Ewels, P. A. *et al*. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**, 276–278, https://doi.org/10.1038/s41587-020-0439-x (2020).
22. Di Tommaso, P. *et al*. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316–319, https://doi.org/10.1038/nbt.3820 (2017).
23. Stevens, F. Understanding how researchers find research software for research practice. Tech. Rep., [object Object] https://doi.org/10.5281/ZENODO.7340034 (2022).

24. The Galaxy Community *et al.* The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research* gkae410 https://doi.org/10.1093/nar/gkae410 (2024).
25. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33, https://doi.org/10.12688/f1000research.29032.2 (2021).
26. Yuen, D. *et al.* The Dockstore: enhancing a community platform for sharing reproducible and accessible computational protocols. *Nucleic Acids Research* **49**, W624–W632, https://doi.org/10.1093/nar/gkab346 (2021).
27. Goble, C. *et al.* Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory https://doi.org/10.5281/ZENODO.4605654 (2021).
28. Goble, C. *et al.* WorkflowHub - a FAIR registry for workflows https://doi.org/10.7490/F1000RESEARCH.1118984.1 (2022).
29. Silva, R. F. D., Pottier, L., Coleman, T., Deelman, E. & Casanova, H. *Workflowhub: Community framework for enabling scientific workflow research and development*, Vol. 2, 49-56 (IEEE, 2020). https://doi.org/10.1109/WORKS51914.2020.00012.
30. Coleman, T. A. *et al.* Wfcommons: A framework for enabling scientific workflow research and development. *Future Generation Computer Systems* **128**, 16–27, https://doi.org/10.1016/j.future.2021.09.043 (2022).
31. Baker, D. *et al.* No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics. *PLOS Pathogens* **16**, e1008643, https://doi.org/10.1371/journal.ppat.1008643 (2020).
32. Soiland-Reyes, S., Goble, C., Bacall, F., Gustafsson, J. & Andrade Buono, R. A guide to using WorkflowHub https://doi.org/10.48546/WORKFLOWHUB.SOP.13.4 (2024).
33. Soiland-Reyes, S. The BGE guide to using WorkflowHub https://doi.org/10.48546/WORKFLOWHUB.SOP.15.1 (2024).
34. Soiland-Reyes, S. BioDT Guide to using WorkflowHub https://doi.org/10.48546/WORKFLOWHUB.SOP.14.1 (2024).
35. Goble, C., Bacall, F. & Soiland-Reyes, S. The BY-COVID Guide to using WorkflowHub https://doi.org/10.48546/WORKFLOWHUB.SOP.10.1 (2024).
36. Gray, A., Goble, C. & Jimenez, R. Nikitina, N., Song, D., Fokoue, A. & Haase, P. (eds) *Bioschemas: From potato salad to protein annotation*. (eds Nikitina, N., Song, D., Fokoue, A. & Haase, P.) ISWC 2017 Posters & Demonstrations and Industry Tracks, CEUR workshop proceedings (RWTH Aachen University, Germany, 2017). https://iswc2017.semanticweb.org/paper-579/. The 16th International Semantic Web Conference 2017, ISWC 2017 ; Conference date: 21-10-2018 Through 25-10-2018.
37. Crusoe, M. R. *et al.* Methods included: standardizing computational reuse and portability with the Common Workflow Language. *Communications of the ACM* **65**, 54–63, https://doi.org/10.1145/3486897 (2022).
38. Amstutz, P. *et al.* Common Workflow Language, v1.0https://doi.org/10.6084/M9.FIGSHARE.3115156.V2 (2016).
39. Ison, J. *et al.* EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* **29**, 1325–1332, https://doi.org/10.1093/bioinformatics/btt113 (2013).
40. Soiland-Reyes, S. *et al.* Packaging research artefacts with RO-Crate. *Data Science* **5**, 97–138, https://doi.org/10.3233/DS-210053 (2022).
41. Soiland-Reyes, S. *et al.* EuroScienceGateway D2.1: Reproducible FAIR Digital Objects for Workflows https://doi.org/10.5281/zenodo.13225792 (2024).
42. De Smedt, K., Koureas, D. & Wittenburg, P. FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. *Publications* **8**, 21, https://doi.org/10.3390/publications8020021 (2020).
43. Köster, J. *et al.* snakemake-workflows/dna-seq-varlociraptor: v5.0.2https://doi.org/10.5281/zenodo.8421328 (2023).
44. Silver, L. & Syme, A. Find transcripts - TSI https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.877.1 (2024).
45. Ison, J. *et al.* The bio.tools registry of software tools and data resources for the life sciences. *Genome Biology* **20**, 164, https://doi.org/10.1186/s13059-019-1772-6 (2019).
46. Deelman, E. & Gil, Y. *Managing Large-Scale Scientific Workflows in Distributed Environments: Experiences and Challenges*, 144–144 https://doi.org/10.1109/E-SCIENCE.2006.261077 (2006).
47. Courbebaisse, G. *et al.* Research Software Lifecycle https://doi.org/10.5281/ZENODO.8324828 (2023).
48. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods* **15**, 475–476, https://doi.org/10.1038/s41592-018-0046-7 (2018).
49. Lamothe, L. *et al.* An evaluation of EDAM coverage in the Tools Ecosystem and prototype integration of Galaxy and WorkflowHub systems. preprint, BioHackrXiv https://doi.org/10.37044/osf.io/79kje (2023).
50. Rosnet, T., Gaignard, A., Devignes, M.-D. & Frikha, S. FAIR-checker https://github.com/IFB-ElixirFr/fair-checker (2024).
51. del Pico, E. M., Gelpi, J. L. & Capella-Gutiérrez, S. Fairsoft - a practical implementation of fair principles for research software. *bioRxiv*https://doi.org/10.1101/2022.05.04.490563 (2022).
52. Suetake, H. *et al.* Sapporo: A workflow execution service that encourages the reuse of workflows in various languages in bioinformatics. *F1000Research* **11**, 889, https://doi.org/10.12688/f1000research.122924.1 (2022).
53. Fernández, J. M., Rodríguez-Navas, L. & Capella-Gutiérrez, S. Secured and annotated execution of workflows with WfExS-backend https://doi.org/10.7490/F1000RESEARCH.1119198.1 (2022).
54. Leo, S. *et al.* Recording provenance of workflow runs with ro-crate. *PLoS one* **19**, e0309210, https://doi.org/10.1371/journal.pone.0309210 (2024).
55. Bray, S. *et al.* The Planemo toolkit for developing, deploying, and executing scientific data analyses in Galaxy and beyond. *Genome Research* **33**, 261–268, https://doi.org/10.1101/gr.276963.122 (2023).
56. Druskat, S. *et al.* Citation File Format https://doi.org/10.5281/zenodo.5171937 (2021).
57. Hiltemann, S. *et al.* Galaxy Training: A powerful framework for teaching! *PLOS Computational Biology* **19**, e1010752, https://doi.org/10.1371/journal.pcbi.1010752 (2023).
58. Garijo, D. *et al.* Nine best practices for research software registries and repositories. *PeerJ Computer Science* **8**, e1023, https://doi.org/10.7717/peerj-cs.1023 (2022).
59. Mazzoni, C. J., Ciofi, C. & Waterhouse, R. M. Biodiversity: an atlas of European reference genomes. *Nature* **619**, 252–252, https://doi.org/10.1038/d41586-023-02229-w (2023).
60. Francis, R. & Christiansen, J. H. Australian BioCommons Strategic Plan 2023 - 2028 https://doi.org/10.5281/zenodo.13626350 (2024).
61. Tejedor, E. *et al.* PyCOMPSs: Parallel computational workflows in Python. *The International Journal of High Performance Computing Applications* **31**, 66–82, https://doi.org/10.1177/1094342015594678 (2017).
62. De La Rosa-Trevín, J. *et al.* Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *Journal of Structural Biology* **195**, 93–99, https://doi.org/10.1016/j.jsb.2016.04.010 (2016).
63. Team, R. C. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.
64. Van Rossum, G. & De Boer, J. Interactively testing remote servers using the python programming language. *CWI quarterly* **4**, 283–303 (1991).
65. Gustafsson, J. & Samaha, G. WORKSHOP: Make your bioinformatics workflows findable and citable https://doi.org/10.5281/zenodo.7787488 (2023).
66. Hatos, A., Quaglia, F., Piovesan, D. & Tosatto, S. C. E. APICURON: a database to credit and acknowledge the work of biocurators. *Database* **2021**, baab019, https://doi.org/10.1093/database/baab019 (2021).
67. Capella-Gutierrez, S. *et al.* Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking https://doi.org/10.1101/181677 (2017).

68. da Veiga Leprevost, F. *et al*. Biocontainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **33**, 2580–2582, https://doi.org/10.1093/bioinformatics/btx192 (2017).
69. Hall, M. B. & Coin, L. J. M. Pangenome databases improve host removal and mycobacteria classification from clinical metagenomic data. *GigaScience* **13**, giae010, https://doi.org/10.1093/gigascience/giae010 (2024).
70. Roach, M. J. *et al*. Hecatomb: an integrated software platform for viral metagenomics. *GigaScience* **13**, giae020, https://doi.org/10.1093/gigascience/giae020 (2024).
71. Wilkinson, S. R. *et al. F*** workflows: when parts of FAIR are missing*, 507–512 https://doi.org/10.1109/eScience55777.2022.00090 (IEEE, Salt Lake City, UT, USA, 2022).
72. ELIXIR. ELIXIR Annual Report 2023 https://doi.org/10.7490/F1000RESEARCH.1119751.1 (2024).
73. Wolstencroft, K. *et al*. SEEK: a systems biology data and model management platform. *BMC Systems Biology* **9**, 33, https://doi.org/10.1186/s12918-015-0174-y (2015).
74. Hambley, A., Chadwick, E., Woolland, O., Soiland-Reyes, S. & Savchenko, V. WorkflowHub Knowledge Graph https://doi.org/10.5281/zenodo.13362051 (2024).
75. Owen, S. *et al*. seek4science/seek: FAIRDOM-SEEK v1.15.0 https://doi.org/10.5281/zenodo.11209855 (2024).

Acknowledgements

The authors acknowledge the contributions of the WorkflowHub community (https://about.workflowhub.eu/project/acknowledgements/), and are grateful to Sehrish Kanwal of the University of Melbourne for feedback and suggestions about the draft, and are grateful to Xenia Perez-Sitja for helping to create Figure 1. The authors would also like to thank the following individuals for their contributions to WorkflowHub: Bert Droesbeke, Timothy Booth, Alan Williams, Eli Chadwick, Rafael Buono, Helena Roche, Florian Hoedt, Paul Brack, Simon Bray. workflowhub.eu was founded as part of EOSC-Life (WP2 Tools Collaboratory), funded by the European Union's Horizon 2020 programme under grant agreement H2020-INFRAEOSC-2018-2 824087. This work was further supported by funding from European Commission's Horizon Europe programme and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee: EuroScienceGateway (HORIZON-INFRA-2021-EOSC-01-04 101057388, UKRI 10038963), BY-COVID (HORIZON-INFRA-2021-EMERGENCY-01 101046203), FAIR-IMPACT (HORIZON-INFRA-2021-EOSC-01-05 101057344, UKRI 10038992), BioDT (HORIZON-INFRA-2021-TECH-01-01 101057437, UKRI 10038930), PREP-IBISBA (H2020-INFRADEV-2019-2 871118), AgroServ (HORIZON-INFRA-2021-SERV-01-02 101058020, UKRI 10038927), BIOINDUSTRY 4.0 (HORIZON-INFRA-2022-TECH-01 101094287, UKRI 10048146). This work is supported by Australian BioCommons which is enabled by NCRIS via Bioplatforms Australia funding. This work is also funded by the Sardinian Regional Government through the XData Project, and by the LIFEMAP project (Italian Ministry of Health, POS T3). The work supported by the ELIXIR-DE node was supported by the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).

Author contributions

Here, we follow the Contributor Role Taxonomy (CRediT, https://credit.niso.org). O.J.R.G. contributed through: Writing – original draft, Writing – review & editing, Project administration, Data curation, and Supervision. S.R.W. contributed through: Writing – original draft, Writing – review & editing, and Project administration. F.B. contributed through: Conceptualization, Data curation, Investigation, Project administration, Software, Methodology, and Resources. S.S.R. contributed through: Conceptualization, Funding acquisition, Investigation, Supervision, and Writing – review & editing. S.O. contributed through: Conceptualization, Investigation, Project administration, Software, and Methodology. N.J. contributed through: Data curation and Writing – review & editing. F.C. contributed through: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, and Supervision. C.G. contributed through: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, and Writing – review & editing. All other authors contributed to the manuscript through: Writing – review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.G.