Deliverable D8.5

Report on federated learning technologies

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP8: Application and Innovation Solutions		
WP Leaders	Alfonso Valencia (37. BSC) Salva Capella-Gutierrez (37. BSC) Marc Van Den Bulcke (9. SC)		
Deliverable Lead Beneficiary	BSC, DKFZ		
Contractual delivery date	31/05/2024	**Actual delivery date**	31/05/2024
Delayed	No		
Partner(s) contributing to deliverable	VIB, P4DAB		
Authors	Carles Hernandez-Ferrer (BSC) Luiz Gadelha (DKFZ)		
Contributors	Dilza Campos (VIB), Laura Portell Silva (BSC), Francisco Fernandes (P4DAB), Miguel Santos (P4DAB)		
Acknowledgements	All members of Pillar II, all members of WP7, all members of WP8		
Reviewer	Dylan Spalding (CSC)		

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Log of changes

Date	Mvm	Who	Description
22/05/2024	0V1	Carles Hernandez-Ferrer (BSC), Luiz Gadelha (DKFZ)	Sent to reviewers
22/05/2024	0V2	Mercedes Rothschild Steiner (ELIXIR)	Sent to GDI-MB
29/05/2024	0V3	Carles Hernandez-Ferrer (BSC), Luiz Gadelha (DKFZ	Updated according reviewers and sent to Coordination
05/06/2024	1V0	Mercedes Rothschild Steiner (ELIXIR Hub)	Final version submitted to EC Portal

Table of contents

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

1. Executive Summary

GDI Pillar III aims to explore use cases and innovative applications for analysing genomic and clinical data, ideally supported by the infrastructure being deployed at the national nodes within Pillar II. Two recent trends can be observed in this area: the application of the FAIR principles computational workflows for enabling scalable and reproducible analyses[1,2] and the increasing application of artificial intelligence techniques[3,4]. *Federated learning* is described[5] as a distributed machine learning technique in which multiple participants, which provide remote devices or siloed data centres, collaboratively train a shared machine learning model while keeping their data locally, better supporting data privacy. As the model is trained locally by each participant on its own data, only model updates (e.g. gradients or weights) are sent to a central server. The central server aggregates these updates to improve the global model, which is then sent back to the participants for further iterative training rounds. Therefore, federated learning enables collaborative learning from distributed data sources without sharing the original data, thus reducing privacy concerns and leveraging the aggregate knowledge available to the multiple participants. In this report, we provide a brief background on recent work on federated learning applied to genomics and health and how they are aligned to demonstrations performed in GDI, report the results of surveys conducted within the GDI participants regarding workflows and federated learning technologies, and discuss and evaluation of different possible scenarios for integrating these technologies into the GDI infrastructure.

[1] Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M. R., Peters, K., & Schober, D. (2020). FAIR Computational Workflows. Data Intelligence, 2(1–2), 108–121. https://doi.org/10.1162/dint_a_00033

[2] Niehues, A., de Visser, C., Hagenbeek, F. A., Kulkarni, P., Pool, R., Karu, N., Kindt, A. S. D., Singh, G., Vermeiren, R. R. J. M., Boomsma, D. I., van Dongen, J., 't Hoen, P. A. C., & van Gool, A. J. (2024). A multi-omics data analysis workflow packaged as a FAIR Digital Object. GigaScience, 13, giad115. https://doi.org/10.1093/gigascience/giad115

[3] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. Npj Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

[4] Kolobkov, D., Mishra Sharma, S., Medvedev, A., Lebedev, M., Kosaretskiy, E., & Vakhitov, R. (2024). Efficacy of federated learning on genomic data: A study on the UK Biobank and the 1000 Genomes Project. Frontiers in Big Data, 7, 1266031. https://doi.org/10.3389/fdata.2024.1266031

[5] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

2. Contribution towards project outcomes

With this deliverable, the project has reached or the deliverable has contributed to the following project outcomes:

	Contributed
Outcome 1 Secure federated infrastructure and data governance needed to enable sustainable and secure cross border linkage of genomic data sets in compliance with the relevant and agreed legal, ethical, quality and interoperability requirements and standards based on the progress achieved by the 1+MG initiative.	Yes
Outcome 2 Platform performing distributed analysis of genetic/genomic data and any linked clinical/phenotypic information; it should be based on the principle of federated access to data sources, include a federated/multi party authorisation and authentication system, and enable application of appropriate secure multi-party and/or high-end computing, AI and simulation techniques and resources.	Yes
Outcome 3 Clear description of the roles and responsibilities related to personal data and privacy protection, for humans and computers, applicable during project lifetime and after its finalisation.	No
Outcome 4 Business model including an uptake strategy explaining the motivation, patient incentives and conditions for all stakeholders at the different levels (national, European, global) to support the GDI towards its sustainability, including data controllers, patients, citizens, data users, service providers (e.g., IT and biotech companies), healthcare systems and public authorities at large.	No

Outcome 5 Sustained coordination mechanism for the GDI and for the GoE multi-country project launched in the context of the 1+MG initiative.	No
Outcome 6 Communication strategy – to be designed and implemented at the European and national levels.	No
Outcome 7 Capacity building measures necessary to ensure the establishment, sustainable operation, and successful uptake of the infrastructure.	Yes
Outcome 8 Financial support to the relevant stakeholders to enable extension, upgrade, creation and/or physical connection of further data sources beyond the project consortium or to implement the communication strategy and for capacity-building.	No

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

3. Background

In the past few years, federated learning has been increasingly applied to the domain of genomics and health,particularly in cases of rare diseases or homogeneous populations. Choudhury et al.[6] survey the challenges faced by healthcare organisations in training machine learning models due to limited and heterogeneous data, and to address this, distributed machine learning has evolved, leveraging multiple nodes to enhance performance and efficiency by utilising larger datasets and computational resources. Federated learning has emerged as a technique tailored to distributed settings, addressing challenges such as statistical heterogeneity and data imbalance across nodes. Various federated learning topologies have been explored, including centralised and decentralised approaches, as well as Horizontal, Vertical, and Transfer Federated Learning, each suited to different data characteristics. Federated learning was applied to genomics, demonstrating comparable performance to centralised approaches in tasks like gene expression analysis and survival prediction. For instance, Flimma[7] implements a federated version of the limma voom workflow for differential expression analysis in transcriptomics, preserving data privacy across distributed hospital sites. Lastly, Choudhury et al. recommend a broader exploration of federated learning applications beyond image-based tasks, emphasising the importance of tasks such as genomic expression analysis, precision medicine, and patient prognosis using multi-modal data in cancer research. Rieke et al.[8] discuss the challenges in leveraging existing medical data for machine learning due to privacy concerns and data silos, emphasising the importance of federated learning as a solution for digital health. Kolobkov et al.[9] investigate the use of federated learning as a privacy-preserving method for training machine learning models on individual-level genomic data, particularly focusing on phenotype and ancestry prediction. The authors highlight that combining data from multiple sources enhances model accuracy by increasing sample size and reducing bias but note challenges in healthcare, where direct data pooling is restricted due to privacy concerns. They show the feasibility of federated learning on genomic data through experiments using UK Biobank and 1000 Genomes Project datasets. The results show that federated models achieve performance comparable to centralised models (in which all the data is directly accessible during training), even across nodes

[6] Chowdhury, A., Kassem, H., Padoy, N., Umeton, R., & Karargyris, A. (2022). A Review of Medical Federated Learning: Applications in Oncology and Cancer Research. In A. Crimi & S. Bakas (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (Vol. 12962, pp. 3–24). Springer International Publishing. https://doi.org/10.1007/978-3-031-08999-2_1

[7] Zolotareva, O., Nasirigerdeh, R., Matschinske, J., Torkzadehmahani, R., Bakhtiari, M., Frisch, T., Späth, J., Blumenthal, D. B., Abbasinejad, A., Tieri, P., Kaissis, G., Rückert, D., Wenke, N. K., List, M., & Baumbach, J. (2021). Flimma: A federated and privacy-aware tool for differential gene expression analysis. Genome Biology, 22(1), 338. https://doi.org/10.1186/s13059-021-02553-2

[8] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. Npj Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

[9] Kolobkov, D., Mishra Sharma, S., Medvedev, A., Lebedev, M., Kosaretskiy, E., & Vakhitov, R. (2024). Efficacy of federated learning on genomic data: A study on the UK Biobank and the 1000 Genomes Project. Frontiers in Big Data, 7, 1266031. https://doi.org/10.3389/fdata.2024.1266031

with significant heterogeneity. The technical implementation for federated learning was based on the Flower and PyTorch frameworks using Lasso regression. Factors like communication frequency are found to influence model accuracy, with recommendations provided to optimise computational efficiency. Overall, this work underscores the potential of federated learning to facilitate global data collaboration in healthcare, supporting the development of less biassed models that account for diverse genetic ancestries. Further research is suggested to address privacy concerns and optimise model performance in real-world data collaborations. The federated analysis demonstrator conducted in GDI Pillar III for computing polygenic risk scores for infectious diseases[10] (D8.8) is aligned with work described in this section, especially with Kolobkov et al.. However, in a production setting the technical framework for federated learning would need to support data access control and encryption standards implemented by the GDI Starter Kit[11], which come mainly from the Global Alliance for Genomics and Health[12] (GA4GH).

4. Methods for Evaluation of Federated Learning Technologies

In this section, the methodology implemented to ensure effective and goal-centric technology decision-making and the consolidation of knowledge regarding workflow managers for the GDI are discussed.

4.1 Fostering Effective and Goal-Centric Technology Decision-Making

A dedicated task force was established to cross-reference the developed use cases from WP7 (MS26) with the technology advancements explored thus far in WP8 (D8.8). Initially, a collaborative spreadsheet[13] was circulated within the task force, followed by wider distribution within WP8. This spreadsheet aimed to gather collective knowledge and experiences. Subsequently, the compiled results were meticulously analysed and synthesised into an internal report[14]. This report underwent rigorous scrutiny within WP8 before being shared across Pillar III to ensure clarity and accuracy, mitigating the risk of erroneous conclusions or miscommunications.

4.2 Consolidating Knowledge and Experience on Workflow Managers for GDI

The selection of a technology to serve as a Workflow Execution Service[15] (WES) / Task Execution Service[16] (TES) within the GDI Starter KitS, garnered significant attention within Pillar III. In collaboration with T8.3, a poll was orchestrated using Google Forms to harness the collective insights of Pillar II and Pillar III on this critical matter.

[10] GDI Deliverable D8.8 - Evaluation of distributed analysis and federated learning infrastructure solutions and recommendations for adoption
[11] https://github.com/GenomicDataInfrastructure/starter-kit
[12] https://www.ga4gh.org/
[13] Mapping of prototypical questions to federated solutions
[14] GDI] WP7/WP8 Flattening Federated Processing Technologies to Prototypical Questions
[15] https://github.com/ga4gh/workflow-execution-service-schemas
[16] https://github.com/ga4gh/task-execution-schemas

The poll was meticulously curated, focusing on a subset of well-established workflow managers within the bioinformatics domain (Nextflow[17], CWL[18], Snakemake[19], and Galaxy[20]). Additionally, participants were encouraged to suggest alternatives beyond the predefined subset through a brief open-text option. Furthermore, a comprehensive open-text section was provided to solicit diverse comments and opinions, fostering an inclusive dialogue on the subject matter.

5. Results

5.1 Fostering Effective and Goal-Centric Technology Decision-Making

These technologies serve distinct purposes: Galaxy is a workflow manager, while DataSHIELD[21] and Flower[22] are frameworks for data analysis under different data architecture, respectively. To clarify their specifics:

Feature	Galaxy	DataSHIELD	Flower
Purpose	Platform for biomedical research and analysis	Infrastructure for distributed data analysis under data preserving premises	Framework for federated learning
Functionality	Data access, analysis orchestration (workflow manager), data retrieval and data visualisation	Statistical analysis, privacy preservation, data visualisation	Federated model training, privacy preservation
Data handling	Centralised	Decentralised	Decentralised
Data sharing	Data shared within the platform	Statistical analysis are sent to data nodes, aggregated results shared with user	Models are sent to data nodes, encrypted model updates shared with central node

The results compacted into summary tables are stated as:

✓	The tool can provide an answer to this question
✓	The tool can provide the answer to this question using a third party tool
···	Unsure if the tool can answer this question or not
×	The tool cannot answer this questions

[17] https://www.nextflow.io/
[18] https://www.commonwl.org/
[19] https://snakemake.readthedocs.io/en/stable/
[20] https://galaxyproject.org/
[21] https://www.datashield.org/
[22] https://github.com/adap/flower

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5.1.1 Genome of Europe

Question		Galaxy	DataSHIELD	Flower
1	*Lookup of individual genetic variants*	✓	×	×
2	*Recalibration of polygenic risk scores*	…	✓	✓
3	*Ancestry-specific imputation*	…	×	×

5.1.2 1+MG/B1MG

Question		Galaxy	DataSHIELD	Flower
1	*Why do people with certain disease-specific genes not develop the disease?*	✓	✓	✓
2	*Why do some gene variants cause adverse side effects for medications?*	✓	✓	✓

5.1.3 Infectious diseases

Question		Galaxy	DataSHIELD	Flower
1	*GWAS (validation): risk variants of severe COVID-19*	✓	✓	✓
2	*Variants that may guide prognosis and/or treatment*	✓	✓	✓

5.1.4 Data-driven models for Cancer Research

Question		Galaxy	DataSHIELD	Flower
1	*Variant-gene enrichment + treatment discovery/enrichment*	✓	…	×
2	*Compare short tandem repeats between normal and cancerous tissue*	✓	✓	✓

5.2 Consolidating Knowledge and Experience on Workflow Managers for GDI

Out of the total of 24 responses gathered, the predominant workflow manager emerged as Nextflow, with 15 mentions, constituting 31.9% of the total (see ***Figure 1***). Following closely, there was a near tie between Snakemake, garnering 14 mentions (29.8%), and Galaxy with 13 mentions (27.7%).

Furthermore, contributors offered alternative suggestions beyond the predefined subset of workflow managers. Among these, Cromwell[23] was mentioned twice. Sarek[24], a variant calling workflow implemented in Nextflow, was suggested once.

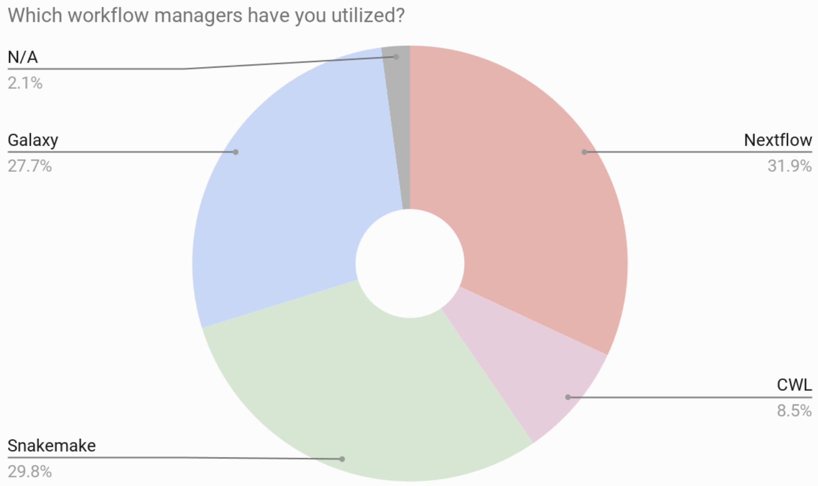

Figure 1. Results in form of pie chart of the mentions to each of the workflow managers considered in the poll created and distributed by T8.3.

We requested contributors to indicate their origin within GDI to help contextualise the demographics influencing decision-making based on the results. ***Figure 2*** illustrates this distribution.

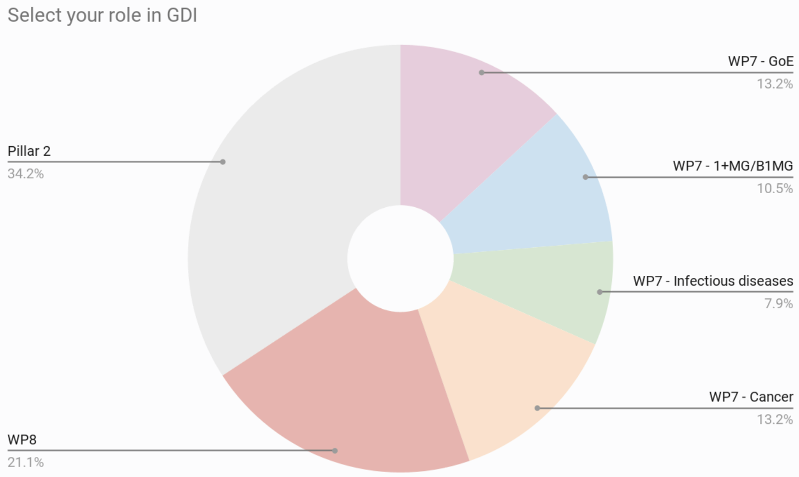

Figure 2. Demographics of the participants of the poll, distributed in Pillar II and the multiple work packages from Pillar III.

[23] https://cromwell.readthedocs.io/en/latest/
[24] https://github.com/nf-core/sarek

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

The survey featured an open-text field where contributors, including members of Pillar II and Pillar III, could share their opinions regarding the selection of workflow managers. To distil key insights for discussion, we conducted a simple sentiment analysis on the responses.

The following are the mentions of each of the proposed workflow managers:

- **Galaxy**: 5 mentions
- **Nextflow**: 11 mentions
- **CWL (Common Workflow Language)**: 2 mentions
- **Snakemake**: 7 mentions
- **Cromwell**: 4 mentions

The next table shows the count of sentiments for each workflow manager.

WM	*Positive*	*Neutral*	*Negative*
Galaxy	3	2	0
Nextflow	6	3	2
CWL (Common Workflow Language)	1	0	1
Snakemake	4	1	2
Cromwell	2	1	1

6. Discussion & Conclusion

6.1 Fostering Effective and Goal-Centric Technology Decision-Making

It is evident that each technologies presents unique considerations regarding compliance with GDI standards:

- **Galaxy**, while robust for data-intensive biomedical research, currently falls short of GDI standards due to its reliance on data residing on data nodes' infrastructure, necessitating further development to align with GDI requirements.
- **DataSHIELD** offers privacy-preserving analysis across distributed data, aligning well with most GDI standards. However, its deployment complexity (in terms of components like Opal, Mica, Agate and others) and potential challenges in granting and managing data access require careful consideration and additional discussion to ensure seamless integration with GDI user requirements.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

- **Flower** emerges as a promising option, fully compliant with GDI's data-privacy standards. However, its exclusive focus on federated model training limits its utility to this specific use case within the GDI framework.

Deferred datasets in Galaxy could potentially contribute to privacy preservation efforts by allowing users to control when and how their data is uploaded and stored within the Galaxy platform. By deferring the upload of sensitive data sets until they are needed for analysis, users can mitigate the risk of unauthorised access to their data. Unfortunately, deferred datasets are not a feature implemented in the core functionality of Galaxy at the current time.

In conclusion, each technology presents distinct advantages.

6.2 Consolidating Knowledge and Experience on Workflow Managers for GDI

The comments provided offer insights into various workflow management systems used in bioinformatics and genetics data analysis with the members of the Pillar II and Pillar III.

- **Galaxy**: Praised for its usability, especially for inexperienced users, but some find it cumbersome due to server management overhead. It's noted for its mature and user-friendly interface. Some preliminary support for GA4GH standards was included[25].
- **Nextflow**: Considered robust, reproducible, and suitable for High-Performance Computing (HPC). It's preferred for its versatility, support for different computing backends, and wide adoption, especially in clinical settings. Some experimental support for the GA4GH TES standard is available[26].
- **CWL (Common Workflow Language)**: Theoretically praised for its independence from specific implementations, but its effectiveness depends on tool support. Some prefer it for its potential standardisation benefits.
- **Snakemake**: Widely used among bioinformaticians, appreciated for its simplicity, ease of installation, and elegant backtracking feature. However, opinions vary on its documentation and comparison with other tools like Cromwell.
- **Cromwell**: Considered a good alternative, tested extensively, but criticised for its documentation and dependency on database management.

The community considers factors like ease of deployment, community support, integration with other tools (such as TES/WES), and ability to handle data securely as key elements before taking any decision on which of the workflow management systems to select for GDI.

Integrating workflow systems into the GDI Starter Kit has challenges related to the need for supporting the genomics and health standards adopted for data access control, encryption, and task execution. A session[27] for discussing the Crypt4GH[28] file encryption standard, workflows, and key

[25] Galaxy and the Global Alliance for Genomic Health
[26] Executors — Nextflow documentation
[27] W LSG - Crypt4GH: developments and demo - Connect 2024 - Agenda.docx
[28] https://crypt4gh.readthedocs.io/en/latest/

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

management took place during GA4GH Connect in April 2024. An in-depth overview was provided comprising Crypt4gh's capabilities and challenges in securing files at rest, and how to integrate it with other parts of the GA4GH ecosystem - for example, htsget[29], the Data Repository Service[30] (DRS), and the Task Execution Service (TES)[31] specifications, which are also adopted by the GDI Starter Kit. The session also discussed key management in different scenarios, from personal devices to TES, proposing solutions such as agent-based approaches and secure processing environments to safeguard keys and ensure controlled access and removal after use.

7. Next steps

Two possible approaches for selecting a technical solution that would support federated learning and workflow management need to be further discussed:

- A tightly-coupled approach would place workflow system components directly along other GDI Starter Kit components for providing easier access to services like the Secure Data Archive. In this case, a decisive choice between Galaxy, Nextflow, or Snakemake as the primary option, with a secondary backup, is crucial. Each option should be evaluated based on factors like usability, scalability, and compatibility with project goals. This selection will result in a formal proposal to Pillar II, outlining the selected primary workflow management option and the contingency plan.
- A loosely-coupled approach would focus on integration around existing GA4GH standards for managing workflow and task executions, namely WES and TES (Kanitz et al. 2024), with the advantage of being independent of the workflow system. The current containerized computation component of GDI Starter Kit follows this approach using the Funnel[32] implementation of TES. This approach would depend on better client support for the TES standard in workflow management systems, which is work in progress[33,34]. As Galaxy also supports TES on the server side with Pulsar[35], it could be an interesting option to evaluate.

As mentioned in the previous section, integration with the data access control and encryption standards and protocols adopted by GDI would be another challenge for seamless integration of workflow management functionality into the Starter Kit.

In both approaches, to support federated learning within workflow systems, it is necessary to integrate federated learning libraries, like Flower, into their application catalogues. Optionally the

[29] http://samtools.github.io/hts-specs/htsget.html
[30] https://ga4gh.github.io/data-repository-service-schemas/preview/release/drs-1.2.0/docs/
[31] Kanitz, A., McLoughlin, M. H., Beckman, L., Malladi, V. S., & Ellrott, K. P. (2024). The GA4GH Task Execution API: Enabling Easy Multi Cloud Task Execution (arXiv:2405.00013). arXiv. http://arxiv.org/abs/2405.00013
[32] https://ohsu-comp-bio.github.io/funnel/
[33] Galaxy and the Global Alliance for Genomic Health
[34] Executors — Nextflow documentation
[35] Pulsar's documentation!

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

federated learning frameworks could be integrated directly to the GA4GH-compliant execution services, independently of the workflow system.

The proposal will emphasise the rationale behind the choice, highlighting benefits, potential challenges, and strategies for mitigation. Input from Pillar II members will be sought to ensure alignment with pillar's objectives.

Training and education initiatives tailored to the selected option should be developed to equip project members with the necessary skills and knowledge. Encouraging collaboration and knowledge sharing among project members will foster a culture of continuous learning and improvement in workflow management practices.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.