NextGen

Genome-Centric Multimodal
Data Integration in Personalised
Cardiovascular Medicine.

Deliverable D3.2
Implementation of federated genomic analysis methods

Grant Agreement Number: 101136962

Funded by
the European Union

UK Research and Innovation

The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No. 10098097, No. 10104323]

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

NextGen	
Project full title	Next Generation Tools for Genome-Centric Multimodal Data Integration In Personalised Cardiovascular Medicine
Call identifier	HORIZON-HLTH-2023-TOOL-05-04
Type of action	RIA
Start date	01/01/2024
End date	31/12/2027
Grant agreement no	101136962

Funding of associated partners
The Swiss associated partners of the NextGen project were funded by the Swiss State Secretariat for Education, Research and Innovation (SERI). The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No 10098097, No 10104323]

D3.5 – Synthetic datasets for testing and piloting-1	
Author(s)	Marco Scutari, Daniele Malpetti
Editor	Marco Scutari
Participating partners	SUPSI
Version	1.0
Status	Final
Deliverable date	M12
Dissemination Level	PU - Public
Official date	2025-06-13
Actual date	2025-06-13

Disclaimer

This document contains material, which is the copyright of certain **NextGen** contractors, and may not be reproduced or copied without permission. All **NextGen** consortium partners have agreed to the full publication of this document if not declared "Confidential". The commercial use of any information contained in this document may require a licence from the proprietor of that information. The reproduction of this document or of parts of it requires an agreement with the proprietor of that information., according to the provisions of the Grant Agreement nr. 101136962 and the Consortium Agreement.

Furthermore, a disclaimer will be included, indicating that: "Any communication or publication related to the action, whether produced collectively by the beneficiaries or individually in any format and through any medium, represents solely the views of the author, and the Commission bears no responsibility for any use that may be made of the information contained therein."

The NEXTGEN consortium consists of the following partners:

No	PARTNER ORGANISATION NAME	ABBREVIATION	COUNTRY
1	UNIVERSITAIR MEDISCH CENTRUM UTRECHT	UMCU	NL
2	HIRO MICRODATACENTERS B.V.	HIRO	NL
3	EURECOM GIE	EURE	FR
4	JOHANN WOLFGANG GOETHE-UNIVERSITAET FRANKFURT AM MAIN	GUF	DE
5	KAROLINSKA INSTITUTET	KI	SE
6	HUS-YHTYMA	HUS	FI
7	UNIVERSITY OF VIRGINIA	UVA	US
8	KLINIKUM RECHTS DER ISAR DER TECHNISCHEN UNIVERSITAT MUNCHEN	TUM-Med	DE
9	HL7 INTERNATIONAL FOUNDATION	HL7	BE
10	MYDATA GLOBAL RY	MYDTA	FI
11	DATAPOWER SRL	DPOW	IT
12	SOCIETE EUROPEENNE DE CARDIOLOGIE	ESC	FR
13	WELLSPAN HEALTH	WSPAN	US
14	LIKE HEALTHCARE RESEARCH GMBH	LIKE	DE
15	NEBS SRL	NEBS	BE
16	THE HUMAN COLOSSUS FOUNDATION	HCF	CH
17	SCUOLA UNIVERSITARIA PROFESSIONALE DELLA SVIZZERA ITALIANA	SUPSI	CH
18	DRUG INFORMATION ASSOCIATION	DIA	CH
19	DPO ASSOCIATES SARL	DPOA	CH
20	QUEEN MARY UNIVERSITY OF LONDON	QMUL	UK
21	EARLHAM INSTITUTE	ERLH	UK
22	ASSOCIAÇÃO DO INSTITUTO SUPERIOR TÉCNICO PARA A INVESTIGAÇÃO E DESENVOLVIMENTO	IST-ID	PT

Document Revision History

DATE	VERSION	DESCRIPTION	CONTRIBUTIONS
13/06/2025	1.0	Complete draft.	SUPSI
27/06/2025	2.0	Complete document.	SUPSI

Authors

AUTHOR/EDITOR	ORGANISATION
Marco Scutari	SUPSI
Daniele Malpetti	SUPSI

Reviewers

REVIEWER	ORGANISATION
Francesca Mangili	SUPSI

List of terms and abbreviations

ABBREVIATIO N	DESCRIPTION
AI	Artificial intelligence
EDHS	European Health DataSpace
FL	Federated (machine) learning
GWAS	Genome-wide association study
KPI	Key Performance Indicator
ML	Machine learning
MMIO	Multi-model integration object
PGS	Polygenic (risk) score
VCF	Variant call format
WP	Work Package

Table of contents

1 Summary

This document presents the implementation of two distinct federated genomics methods, both demonstrated live at the NextGen project meeting held in Heraklion (Greece) on June 10 and 11, 2025:

- **Federated GWAS** – a federated implementation of genome-wide association studies.
  https://gitlab.com/idsia/nextgen-deliverable-sfgwas-docker
- **Federated PLIER** – a federated implementation of the PLIER algorithm for dimensionality reduction and deconvolution of bulk RNA-seq data.
  https://gitlab.com/idsia/nextgen-deliverable-federated-plier

The source code, along with additional materials for each method, is available in the dedicated Git repositories linked above. Both approaches facilitate secure, distributed analysis across multiple institutions without requiring the exchange of raw data, thus addressing key challenges in privacy-preserving genomic research.

2 Federated GWAS

2.1 Overview

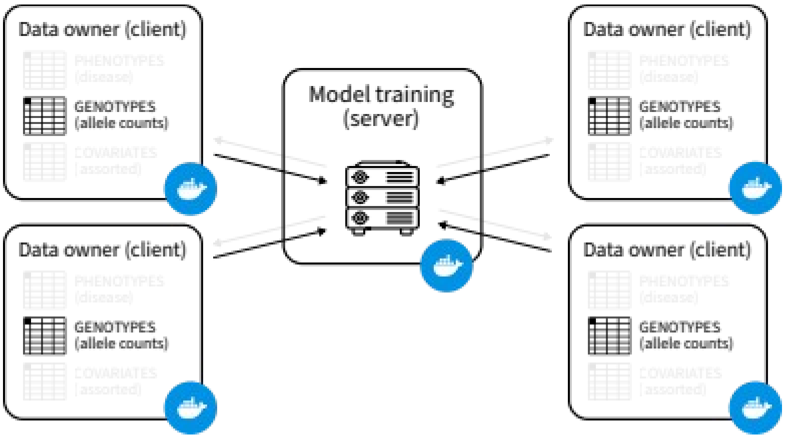
This section describes the anatomy of federated genomewide association studies (GWAS) and its implementation using the SF-GWAS software. In particular, we consider the more mature SF-GWAS PCA from the GitHub repository

https://github.com/hhcho/sfgwas

and the newer SF-GWAS LMM software from the GitHub repository

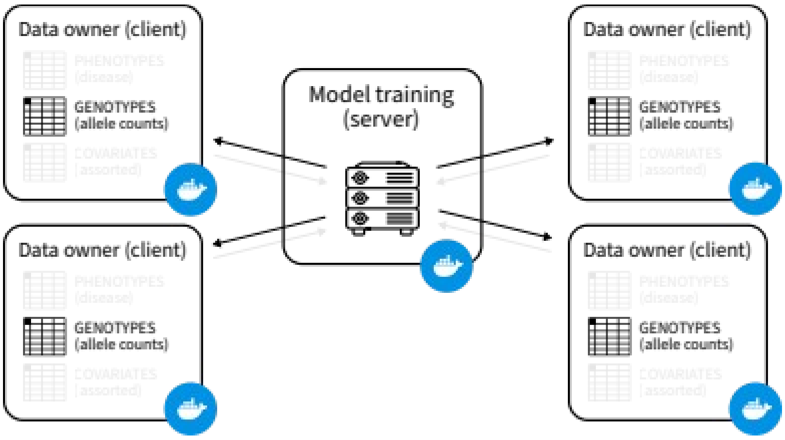
https://github.com/hhcho/sfgwas-lmm

from the same authors. Both pieces of software are introduced in Cho et al. (2025). SF-GWAS PCA SF-GWAS LMM follow the same general data analysis workflow patterned after that of the PLINK software (Purcell et al., 2007). Firstly, each client gathers summary statistics from the local genotype data and sends it to the server.



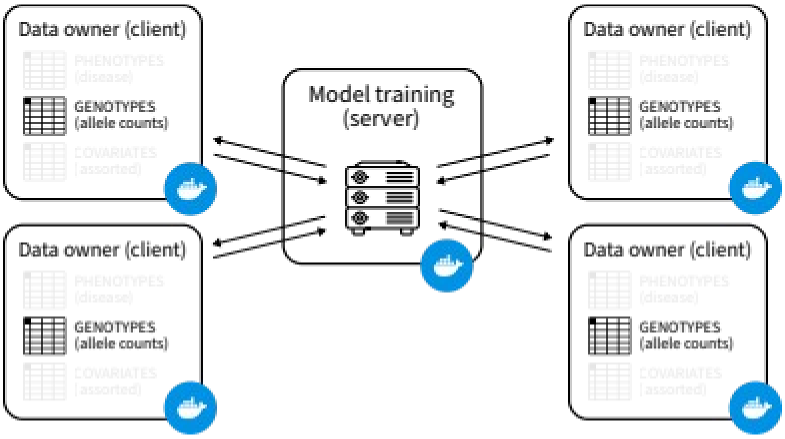
Then, the server tells each client which SNPs to retain for further analysis because they

pass all quality checks (maximum proportion of missing data, minimum minor allele frequency, maximum value of the Hardy-Weinberg disequilibrium statistic).



The server and the clients then compute the genetic relatedness between the individuals to adjust for population structure, either using PCA (for SF-GWAS PCA) or ridge regression (for SF-GWAS PCA). Among non-federated software, the former approach is used in PLINK and the latter in REGENIE (Mbatchou et al., 2021).

The selected number of PCA components is used, along the covariates, to compute the statistical association tests between each marker in the data and the phenotype.



The resulting test statistics are then saved in each of the client, but not in the server.

Further details on the workings of federated learning GWAS can be found in Deliverable D3.1.

2.2 Benchmarks

We benchmark these software against each other and against PLINK and REGENIE using the first chromosome of the 1000G (One Thousand Genomes) data. The raw data are available here:

https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/

The characteristics of the benchmark data are as follows.

- Genotypes: One Thousand Genomes
  - Sample size: 3202 individuals.
  - Variables: 5,013,617 SNPs from chromosome 1 (the largest).
  - Filtering: only SNPs with rsIDs and minor allele frequency > 0.05.
- Phenotypes: Cardiovascular diseases, synthetic binary phenotype.
  - Generated from EFO_0000319 in the GWAS Catalog as detailed in the Deliverable D3.5.
  - 36 causal SNPs, 7 of which are in chromosome 1.
  - Assumed heritability 0.85 to increase signal.

Both SF-GWAS software are configured with a topology with 1 server and 3 clients, with the data split in equal parts among them. Each client's data is provided a set of PVAR, PGEN, PSAM files (PLINK format). No covariates other than the first five principal components were used to adjust the statistical tests used to estimate association.

The benchmark results are as follows.

Software	# Threads	Memory use	Disk use	Running Time
PLINK	30	650MB	400MB	8m
REGENIE	30	4GB	660MB	9m
SF-GWAS PCA	6 for the server, 8 for each client	16GB for each client/server	5GB per client	2h 25m
SF-GWAS LMM	6 for the server, 8 for each client	>60GB for each client, 2GB for the server	20GB per client	-

SFGWAS-LMM did not require too much memory (>200GB in total) and did not complete after running out of memory. Since memory use is tied to the number of threads and the size of the data, we reduced the data to 100,000 SNPs and allocated only two threads for each client. This allowed the benchmark to complete, but only after using 38.8GB of memory for each client and more than 9 days of running time. This suggests that running SF-GWAS LMM on the whole chromosome 1 would require an order of magnitude more time and memory than SF-GWAS PCA.

To further understand the scalability of SF-GWAS PCA, we increased the complexity of different aspects of the federated analysis and recorded the results.

- Adding 5 covariates: running time increases to 3h 6 m, memory use increases to 17GB per client/server.
- Adding 5 more principal components: running time increases to 3h 24m, memory use to 19GB per client/server.

- Adding chromosome 2 (+6,088,709 SNPs) from the 1000G data: running time increases to 4h 33m, memory use increases to 28GB.
- Using 6 clients instead of 3: running time increases to 4h 36m, memory use remains 16GB per client/server.

From these benchmarks, we conclude the following.

- SF-GWAS LMM REGENIE is not mature enough to use.
    - It lacks basic features.
    - It uses too many resources.
    - It does not even have an interface to run it, it's just a library.
- SF-GWAS is more mature, it can be refined to make it actually usable.
- Realistically, each client needs server-class hardware to run either SF-GWAS software on real data.
- SF-GWAS LMM is still under development, so it might possibly improve with time.

3 Federated PLIER

3.1 Overview

The Pathway-Level Information ExtractoR (PLIER) algorithm (Mao et al., 2019) is a method designed for dimensionality reduction and deconvolution of transcriptomic data. PLIER requires a training phase that uses a large gene expression dataset along with a set of gene sets, such as pathways or single-cell signatures, with the latter serving specifically to incorporate prior biological knowledge. This training phase produces a transformation (also referred to as model) that can subsequently be used to reduce the dimensionality of gene expression data. As the method is unsupervised, the learned model can be applied both to the training data and to new, unseen data. In this way, PLIER converts a high-dimensional gene expression matrix into a lower-dimensional representation based on new variables, called latent variables (LVs), some of which are biologically interpretable through their association with specific gene sets. Thanks to the high correlation among genes, PLIER performs dimensionality reduction effectively without significant loss of information.

In this work, we propose a federated implementation of PLIER, enabling multiple data holders to collaboratively learn a shared model without exchanging raw data. The federated learning approach follows a centralized architecture, analogous to that described in the previous section, involving a central server and several clients (data holders). The training proceeds iteratively: each client computes intermediate quantities based on local data and transmits them to the server, which aggregates the contributions and sends updated information back to the clients, continuing until a pre-defined stopping criterion is reached.

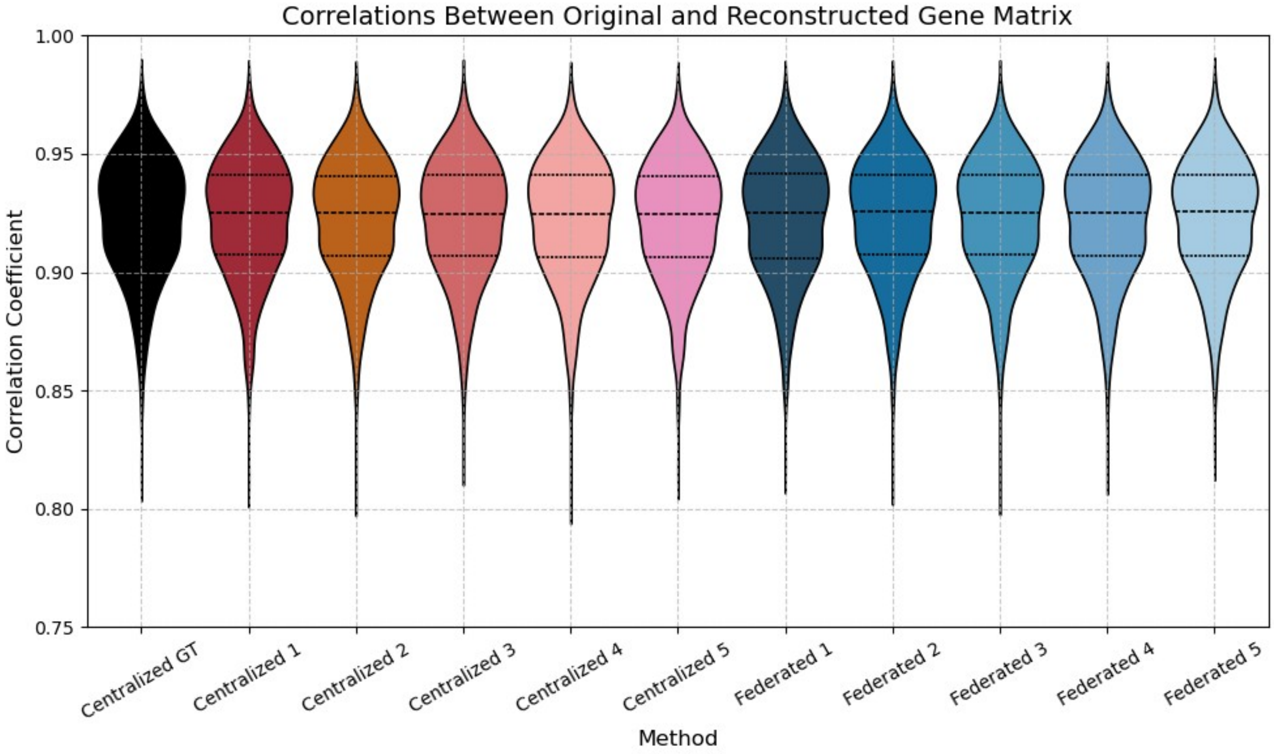
The implementation of the algorithm was carried out using Flower, a federated learning Python framework identified in recent literature as one of the most effective and flexible frameworks for federated computing (Riedel et al., 2024). The live demonstration in Heraklion involved four laptops, with one acting as the central server and the remaining three simulating independent data holders. To facilitate understanding and discussion, the demo was structured in two distinct phases. The first phase involved a standardization of the datasets across clients, a necessary preprocessing step for PLIER. This phase served as a didactic opportunity to walk through and explain each component of the code and process. The second phase focused on the actual federated training of the PLIER algorithm.

Please note that the plots presented in the following sections are not extracted directly from the live demo. Due to time constraints, the live session used a reduced dataset, while the figures shown here are based on results from multiple full-dataset simulation runs.

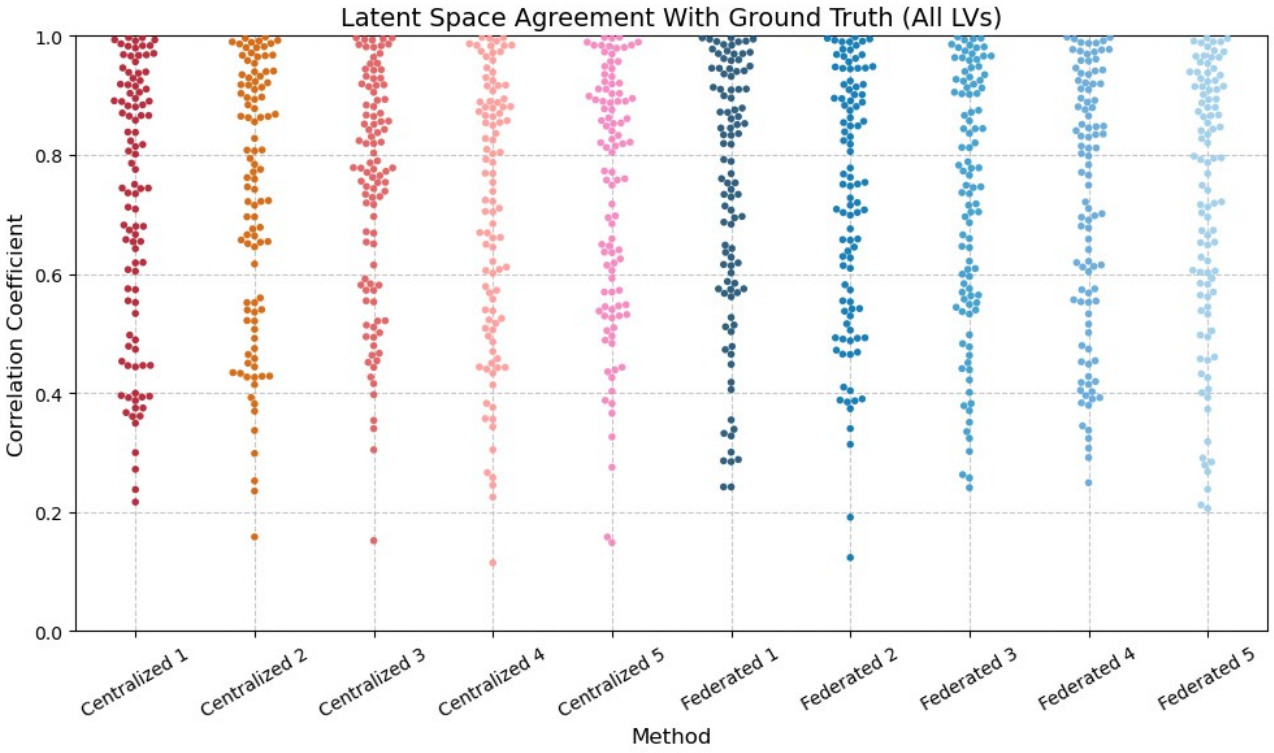
3.2 Benchmarks

To assess the effectiveness of Federated PLIER, we used a dataset comprising kidney samples from The Cancer Genome Atlas Project (TCGA, https://www.cancer.gov/ccg/research/genome-sequencing/tcga), along with a gene set matrix extracted from a study by Legouis et al. (2024), which applied PLIER in the kidney domain. We compared multiple models: a centralized model (i.e., trained storing the entire dataset at a single location) considered as the ground truth and trained using the R PLIER package; five additional centralized models that differed from the ground truth only in their random seed initialization; and five federated models trained using the Flower framework, each based on a different distribution of the dataset across three clients.

To evaluate model quality, we examined how well each model could reconstruct the original high-dimensional gene expression data from its lower-dimensional latent representation, an operation analogous to reversing the transformation. Specifically, we reconstructed the gene expression matrix and computed gene-wise correlations with the original data, following a procedure similar to that used in the original PLIER publication. The resulting distribution of correlation values, presented in the following figure, shows no appreciable difference among the different models.



Beyond reconstruction accuracy, we also examined the consistency of the latent spaces produced by the different implementations. This involved comparing the latent spaces from each run against those produced by the ground truth model. Using a correlation-based approach proposed by Taroni et al. (2019), we observed that the latent space generated by the federated decomposition was as similar to the ground truth as other centralized runs differing only by random seed initialization, as shown in the following image.



These results demonstrate that the federated implementation of PLIER produces models that are not only effective but also consistent with those obtained through centralized computation, confirming that Federated PLIER can achieve results equivalent to centralized PLIER while preserving data privacy. For future evaluations aimed at publication, we plan to perform a similar analysis to train, in a federated manner, a model equivalent to the MultiPLIER model developed by Taroni et al., which leverages a substantially larger training corpus. This comparison will help validate our results and further establish the applicability of our federated implementation in real-world, multi-institutional settings.

4 References

H. Cho et al. (2025). Secure and federated genome-wide association studies for biobank-scale datasets. Nature Genetics 57, 809–814.

J. Mbatchou et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nature Genetics 53, 1097–1103.

S. Purcell et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81(3), 559–575.

W. Mao et al. (2019). Pathway-level information extractor (PLIER) for gene expression data. Nature Methods 16(7), 607-610.

P. Riedel et al. (2024). Comparative analysis of open-source federated learning frameworks - a literature-based survey and review. International Journal of Machine Learning and Cybernetics 15(11), 5257-5278.

D. Legouis et al. (2024). A transfer learning framework to elucidate the clinical relevance of altered proximal tubule cell states in kidney disease." Iscience 27(3).

J. N. Taroni et al. (2019). MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. Cell systems 8(5), 380-394.