NextGen

Genome-Centric Multimodal Data Integration in Personalised Cardiovascular Medicine.

Deliverable 3.4

Privacy assessment of federated approaches

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]

Author(s)	Marco Scutari, Daniele Malpetti
Editor	Francesca Mangili
Participating partners	SUPSI
Version	1.0
Status	Final
Deliverable date	M24
Dissemination Level	PU - Public
Official date	2025-12-18
Actual date	2025-12-15

Disclaimer

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
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	ASSOCIACAO DO INSTITUTO SUPERIOR TECNICO PARA A INVESTIGACAO E O DESENVOLVIMENTO	IST-ID	PT

Document Revision History

DATE	VERSION	DESCRIPTION	CONTRIBUTIONS
15/12/2025	1.0	Complete draft.	SUPSI

Authors

AUTHOR/EDITOR	ORGANISATION
Marco Scutari	SUPSI
Daniele Malpetti	SUPSI
Francesca Mangili	SUPSI

Reviewers

REVIEWER	ORGANISATION
Aaron Lee	QMUL
Philippe Page	HCF
Isabelle Hering	DPOA

Table of contents

1 Summary

This document provides an overview of the open issues and best practices in privacy assessments of ML/AI models, as described in the literature and with a focus on the types of data and models of interest to NextGen.

The first part of this document (Sections 3 to 5) focuses on general privacy considerations for genetic data, complemented by considerations for specific types of genetic data and methods for reducing privacy risks.

The second part (Sections 6 and beyond) examines the privacy-enhancing mechanisms relevant to the federated learning (FL) tools currently being developed within the NextGen project. These tools include a federated implementation of the PLIER algorithm for bulk transcriptomics, as well as the scVI and scANVI models from the scvi-tools library [Xu21] for single-cell transcriptomics. All methods are implemented using the Flower FL framework [Beu22]. Because these approaches differ substantially in their model classes, data modalities, and learning dynamics, we analyse the specific privacy risks associated with each method, together with the corresponding mitigation strategies. Before turning to the individual algorithms, we briefly review the main privacy threats that may arise in realistic deployments and the two families of privacy-preserving techniques relevant to their secure implementation: secure aggregation, which protects the confidentiality of per-client updates during training, and differential privacy, which limits the information that can be inferred from the final released model both intermediate and final federated global models.

For a more comprehensive introduction to FL and its applications in bioinformatics, we refer the reader to our earlier NextGen Deliverable D3.1, now published as a peer-reviewed article [Mal25a].

2 Glossary of terms

- **Man-in-the-middle (MitM) attack:** An attacker secretly intercepts and potentially alters data being transmitted between two parties.
- **Data breaches:** Sensitive information is stolen from storage systems, accessed without authorisation, or not meant to be accessed by the person accessing it.
- **Insider threats:** Individuals with authorised access to data misuse it for inappropriate purposes.
- **Data linkage attack:** Combining data with publicly available information.
- **DNA phenotyping:** Predicting physical characteristics from genetic data, potentially revealing sensitive personal traits.
- **Pedigree analysis:** Inferring genetic relationships and potentially identifying individuals based on family history data.

- **Genotype imputation:** Predicting missing genetic information, potentially revealing previously private data.
- **Data poisoning attacks:** Introducing false data to corrupt analysis results and potentially mislead researchers.
- **Gradient construction attack:** Reconstructing sensitive genetic data from the updates shared during machine learning model training.
- **Backdoor attacks:** Implanting hidden behaviours into a machine learning model that are activated by a specific trigger, causing the model to behave incorrectly. Such attacks typically require an active adversary.
- **White-box inference attacks:** Extracting sensitive information from a trained model when the attacker has full access to its internal parameters, architecture, or gradients.
- **Membership inference attacks:** Determining whether a specific individual's genetic data was used to train a machine learning model.
- **Model inversion attacks:** Reconstructing characteristics of individuals from a trained machine learning model using their data.
- **Adversarial attacks:** Introducing subtle changes to data to cause a machine learning model to make incorrect predictions.
- **Inference data leakage:** Sensitive information about individuals is revealed through the outputs or results of analyses performed on data.
- **Reidentification:** Reconstructing the identity of an individual from data that were originally thought to be anonymised.
- **Individuation:** Detecting whether an individual is part of a study based using external data on that individual, without necessarily uncovering their identity (that is, re-identification).

3 Threat model

Throughout this deliverable, we adopt a unified threat model for all three methods under consideration. We assume an *honest-but-curious* setting: participating institutions (clients) and the central server follow the prescribed FL protocol as intended but may attempt to infer additional information from the messages and model parameters to which they legitimately have access during training. Under this assumption, we do not consider active adversarial behaviours such as model poisoning or targeted backdoor insertion, which are outside the scope of this deliverable. We regard this scenario as realistic, as the models we consider are expected to be deployed by relatively small consortia of research institutions that share a common scientific objective.

For all algorithms, the final trained model is assumed to be made publicly available, for example as part of a repository accompanying a scientific publication. This reflects a plausible usage scenario in which several institutions jointly train a model that is subsequently shared with the broader community. Public release of the model requires protection against white-box inference attacks, including membership inference and model inversion.

These considerations lead to two distinct classes of privacy risk. The first concerns the potential exposure of sensitive information during the training process within the consortium. The second concerns the risk that sensitive information may be extracted from the final model once it is released externally. We describe how these two risks may be [addressed/mitigated] through the complementary use of secure aggregation and differential privacy, in latter sections.

4 Privacy in genetics data: general considerations

Genetic data can provide information about an individual's sensitive details (ethnic background; eye, hair and skin colour; height; susceptibility to diseases) [Tho24] and, when paired with other data, could conceivably determine an individual's identity with some reliability. Therefore, privacy is a relevant issue even if personal identifiers are removed. As publicly available genetic genealogy databases continue to expand, and direct-to-consumer sequencing kits become increasingly popular, the reidentification of genetic data records using publicly available databases should be considered a growing threat to privacy. As an additional complication, the leakage of an individual's genetic data can compromise the privacy of entire families, as DNA sequences are heritable.

The privacy risks associated with genetic data vary significantly across different data sets [Tho24]. Considering all genetic data at all times as information relating to an identifiable natural person is not correct, and it is becoming apparent that reidentification risk in genetic data must be assessed on a case-by-case basis and under the consideration of all the means reasonably likely to be used [Sha19]. While completely eliminating the possibility of reidentification is rarely achievable, a more practical approach of risk minimisation is warranted [Mar19,Wr19], accompanied by organisational and technical measures to safeguard genetic data from reidentification attack attempts and transparent communication of the remaining risks to data subjects.

Overall, a decade of research on genetic privacy from the GetPreCiSe Center at Vanderbilt University shows that the risk of reidentification in properly maintained databases is overstated: it is currently not significant, but at the same time, it is non-trivial. For instance, the 95th percentile expected risk of re-identifying an All of Us participant is below the threshold used by various U.S. state and federal agencies. Furthermore, public concern about genetic privacy may be no greater than concern about financial and other types of privacy [Slo25]. At present, re-identification seldom causes significant harm to an individual, even for those who disclose their participation in a data collection program. More generally, the research suggested that privacy experts may underestimate the extent to which the public values utility over privacy risk. GetPreCiSe also shows that the reidentification risk associated with real-world, feasible attack scenarios is often significantly lower than under the worst-case assumption across all case studies and different types of sequence data, regardless of the data set size [Slo25]. Hence, they suggest that substantially more data could be shared for biomedical research, notwithstanding the moratorium on using genetic data for insurance purposes in countries where this is applicable.

4.1 Geneal privacy considerations

[Tho24] presents a general overview of privacy concerns in genetic data sets. It distinguished between:

1. General categorisations of genomic data sets, providing a very rough estimate of the amount of privacy-critical information they include.
2. High-risk genomic features, whose assessment is critical for estimating the reidentification risk.
3. Low-risk genomic features that have not been exploited for privacy attacks yet, but could pose a risk in some circumstances.

This categorisation is explained and reproduced visually from [Tho24] in Table 1. The general categorisation of the data spans modality and coverage, noting that higher genotyping densities are inherently high risk; however, no modality other than high-density DNA data has been successfully used in reidentification attacks. Note that genome coverage alone is not a reliable risk indicator; how sequenced variants are spread across the genome and whether they are close to well-studied variants. The amount of data preprocessing also plays a role. On the one hand, raw data contain information that is less relevant to scientific research and more challenging to use, but is higher-risk. On the other hand, pre-processed data contain less information but are more readily usable for cross-referencing in the attacks listed in the glossary. Finally, [Tho24] draws a broad distinction between germline data modalities, which are common to all tissue types and remain mostly stable throughout an individual's lifetime (apart from mutations and epigenetic modifications), and somatic data modalities, which are tissue-specific and vary significantly with the individual's condition at the time of measurement. The former obviously present a higher risk to privacy, as their use is less time-sensitive.

Table 1 Overview of the privacy-critical features of genetic data sets, with exemplary values and key points to consider for risk assessment. From [Tho24].

General assessment	Biological modality	Protein expression (low) -> RNA -> Methylation -> DNA (high)
	• Do the data contain DNA sequence information directly (e.g., DNA sequencing reads)? If yes, could the data be processed such that sequence information is no longer available (e.g., report DNA methylation levels in percentage instead of providing raw sequencing read files)? • Could DNA sequence information be inferred from the data (e.g., via biological correlations such as expression or methylation quantitative trait loci)? • What sensitive information could be inferred from the data (e.g., age, sex, diseases, or physical traits)?	
	Experimental assay	PCR (low) -> Microarray -> targeted genes -> SNP panel -> Whole exome -> Whole genome (high)
	• Which method was used to generate the data? Does this method produce rich or sparse data? (What percentage of all base pairs or loci of the genome are covered by the method?) • How do the data produced with this method cover the genome (i.e., genome-wide vs targeted approach)? • How likely is it that data generated with the same method are present in publicly available databases (i.e., commercial assay vs custom)?	
	Data format or level of preprocessing	Tabular (medium) -> VCF/MAF -> BAM/SAM -> FASTA/FASTQ (high)
	• If the data are in a raw or semi processed format, do the data contain any information that is not directly relevant for their intended use?	
	Germline versus somatic variation content	Somatic (low) -> Germline (high)
	• Was germline or somatic variation of primary interest when generating or processing the data? • If somatic variation was of primary interest, was germline variation removed from the data?	
High-risk components	Single-nucleotide polymorphisms (SNPs)	1-10 (low) -> 30-100 (medium) -> >500 (high) associated with sensitive attributes (medium) -> common assays (high)
	Short tandem repeats (STRs)	1-5 (low) -> 5-10 -> 10-15 (medium) -> >15 (high)
	Aggregated sample measures	p-values (low) -> odds ratios -> allele frequencies (medium)
Low-risk components	Rare single-nucleotide variations (SNVs)	-
	Structural variants (SVs)	-

Some of these considerations are also present in [Slo25]. They highlight how an attacker may be able to cross-reference data sets or construct a chain of intermediate connections between the two data sets by purchasing data from a broker. The availability of such data sets varies greatly between modalities, with high-density DNA being the easiest to access.

Their key findings are summarised in Table 2.

Table 2 Key findings on the risks posed by the collection of genetic data from [Slo25].

- Estimates about the risk of re-identification of genomic data that are derived from attacks that assume adversaries already know a great deal about the target likely over-estimate real-world conditions. Overestimation of risks can induce unnecessary editing of shared data, making data repositories less useful.
- Nonetheless, the risk of re-identification is non-trivial, especially with the advent of data brokers that can piece together disparate pieces of personal information. Long-term security protection may also be difficult, given the possibility of advances in decryption methods.
- Other actions, such as a person's disclosure that they are in a genetic database, may also increase risk.
- Several strategies (some of which are discussed in more detail below) may reduce the risk of re-identification and other undesired disclosures:
    - The use of multipronged data protections strategies.
    - Requiring users of a genetic database to enter into data-use agreements.
    - Increasing penalties for privacy breaches.
    - Cautioning participants in research cohorts not to publicize their membership on social media or in other ways.

4.2 Legal considerations

[Bra24] presents an up-to-date, comprehensive summary of the legal aspects of privacy-enhancing technologies in genome-wide association studies. It highlights the need for privacy-by-design (i.e., privacy measures built into technical and organisational processes) to comply with the GDPR and proposes federated learning as a possible solution. Despite advancements like federated GWAS, which offer improved privacy, legal uncertainties significantly hinder genomic research, particularly concerning cross-border data transfers, family member rights, and the validity of consent. Surprisingly, even publicly available GWAS summary statistics can potentially reveal individual participation, risking unwanted disclosure of sensitive health information. Addressing these issues requires a robust legal framework aligned with regulations like GDPR, alongside continued implementation of Privacy Enhancing Technologies (PETs) and localised data processing to minimise risks and uphold privacy by design principles.

[Slo25] summarises the legal methods of reducing risks found by the research at the GetPrecise Center, which are reported in Table 3 below.

Table 3 Regulatory/legal methods of reducing risks, from [Slo25].

- Legal protections, such as those provided by the Health Insurance Portability and Accountability Act (HIPAA), the Genetic Information Nondiscrimination Act (GINA), and the Fourth Amendment's restrictions on unwarranted searches are inadequate.
- Existing legal regimes all recognize numerous exceptions that permit third party access to identifiable health information.
- Policies of DTC-GT companies are often woefully inadequate at preventing dissemination of genetic data.
- Searches of DNA profiles in third-party databases should require judicial authorization as a constitutional matter.
- Potential participants should be informed not only of the non-negligible risks of re-identification but also of the potential for third-party access and data-sharing across companies and government entities.
- Efforts to assert individual control over one's data have proven to be of limited utility in protecting genetic privacy.
- Although it is generally recognized that some form of consent for the collection and use of genetic data is important, the precise nature of that consent, which can range from dynamic consent to simple assent, is highly contested, and in fact the law allows many uses of genetic data without consent.
- Creating secure databases for specific purposes and tailoring privacy-protecting tools and rules that are appropriate for each situation may be superior to granting granular control over data in larger, general-purpose databases.

[Mal25] also provides an in-depth discussion of how federated learning can help satisfy the requirements of the GDPR for genetic data. Data providers, which have a complete control over and a more intimate knowledge of the data they collected as well as a direct connection to data subjects, act in a "data controller" role (Articles 4 and 24), taking "appropriate technical and organisational measures" (Article 25) to ensure privacy and security, thus minimising the risk of data breaches. Therefore, they can directly scrutinise their use, notify data subjects about it to request consent (Article 9); allow them to withdraw their data (Article 7); ensure lawful, fair and transparent processing (Articles 12–15); and directly assess risks to data subjects and minimise them through appropriate legal agreements. Consortium parties, which include both data controllers (as clients) and data processors (as servers, compute facilities, as defined in Article 4), are also required to use the techniques described in Section 3 to ensure data security and privacy beyond what is provided by base FL (privacy

by design, Articles 24, 25 and 32). For the same reasons, FL facilitates compliance with the EU AI Act. Some of its requirements strengthen those in the GDPR, such as data minimisation, localisation, transparency, auditability, security, and data quality. Additionally, the EU AI Act requires efforts to mitigate bias, ensure the robustness of models, implement human oversight, and assess high-risk systems. Data controllers are in the best position to ensure these requirements are met. Collectively, they can provide more representative samples that are less prone to bias and fairness issues. Finally, the presence of multiple data controllers in the consortium also implies that these requirements are verified by several independent parties.

5 Privacy in genetics data, by type

Moving beyond general considerations that apply broadly to several types of genetic data, [Tho24] notes that true reidentification requires a further step beyond getting hold of genetic data: matching them to databases with identifying or quasi-identifying information. Such a linkage attack has only been demonstrated on high-density DNA sequence data, for which such databases are available from direct-to-consumer services like 23andMe and genealogy databases like GEDmatch. Other data types should, therefore, be considered less risky until similar databases become available.

Privacy risks are also mitigated by the ephemeral nature of certain types of genetic data, such as sequence or expression data from cancer tissues. A linkage attack would require a matching data record of the same tissue, ideally taken at a similar time in life, to match with such data. Hence, the risk of reidentification attacks is negligible.

4.1 Single nucleotide polymorphisms (SNPs)

[Tho24] identifies the following guiding questions in assessing the privacy risks of SNP data.

- How many SNPs do the data contain (directly or indirectly)?
- Are the SNPs in close proximity or spread across the genome (nearby SNPs are more likely to be correlated and thus often contain less information than statistically independent SNPs)?
- Are the interrogated SNPs frequently assessed in research or by direct-to-consumer providers (ie, how likely is it that they can be linked to publicly available, identifying data sets)? The study by [Lu21] presents an overview of genotyping arrays commonly used by direct-to-consumer companies.
- Are all SNPs relevant to the intended use of the data, or could some be removed from the data?
- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with STRs or other)?

SNPs must be considered a high risk for privacy, and data sanitisation efforts should be used in any genetic data set containing more than 20 SNPs. A small number of SNPs (30-80, [You18]) can, theoretically, uniquely identify an individual, and SNPs are widely available in public databases, along with identifying and quasi-identifying information. In practical situations, however, the risk of

reidentification is less severe than reported in such literature [Slo25] because of the unrealistic operating conditions assumed in academic publications [Wai22].

[Erl18] could only identify the family name of an anonymous study participant with a much larger (700,000) set of SNPs via its relatives' sequence data on the genetic genealogy website GEDmatch. They could not identify the individual itself. They estimated that a genetic database needs to cover "only 2% of the target population to provide a third-cousin match to nearly any person" in such a matching attack. As of 2018, the probability for such a match was estimated to be 60% for the platform GEDmatch. In other words, their precision was only six degrees of relatedness from the target individual, and even that with less than a 2-out-of-3 chance of success.

Completion and inference attacks, which involve inferring genomic regions, modalities, or physical attributes that were not originally studied [9,90-96], have a linkage attack success rate of only 5% (i.e., the proportion of correctly matched individuals) and are likely to perform worse in more realistic scenarios. Both [Tho24] and [Slo25] use extracting information for photos, concluding that it does not represent a viable attack vector given the low quality of publicly available photos.

4.2 Other variant data (STRs, SVNs and Svs)

[Tho24] identifies the following guiding questions in assessing the privacy risks of short tandem repeats (STRs), rare single-nucleotide variations (SNVs) and structural variants (SV) data.

- Do the data directly or indirectly (eg, STRs in raw data and STRs imputable from SNPs) contain >10 STR loci?
- Are these STR loci either (1) part of the CODIS system or (2) on the Y chromosome (ie, high linkability)?
- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with SNPs or other)?
- Are there any databases that could be used to cross-link the data to identifiable data, and how accessible are the databases?

Knowing the repeat numbers of as few as 10 to 30 STRs can suffice for individual identification. Due to their high identifiability, STRs are utilised to determine identity and kinship in forensic science, law enforcement, paternity testing, and genetic genealogy. In particular, STRs on the Y chromosome are included in several direct-to-consumer kits to identify relatives along the paternal ancestry. They pose high privacy risks because they are also included in several databases with quasi-identifying information. Studies on membership attacks have successfully achieved a 50% reidentification rate of patients in the 1000G data [Gym13] and successfully matched STRs with the corresponding sequence data with 98% accuracy. Linking STRs with individuals via linking attacks has a significantly lower success rate of 12% [Gym13] in recovering correct family names, and even less in recovering personal identities. The reconstruction of identity from a family name is not trivial and can take months to complete, as others have pointed out [Gue21]. Furthermore, linking is only possible for men.

Rare variants greatly increase the risk of reidentification for the small subpopulation of variant carriers. However, they are generally not included in research studies and direct-to-consumer kits. It is also unlikely that a set of SNVs could be linked to any database with quasi-identifying information. Therefore, they can currently be viewed as a low risk for reidentification, despite the theoretical potential for identifiability.

Structural variants, of which the best-known is copy-number variations, are not used for genetic genealogy analyses, and many SVs that are somatic, that is, nonhereditary, are not present in all cells of the body, not stable, and thus not strongly associated with identity. Human CNV databases are also very scarce [Ho19]. Therefore, the risk of reidentification is very low.

4.3 Aggregated sample values

[Tho24] identifies the following guiding questions in assessing the privacy risks of aggregated sample values, that is, summary statistics.

- What sensitive information could an attacker gain from ascertaining the membership of an individual to the data set (eg, geographic information, sex, disease, and age)?

The limited information content in these summary statistics usually only allows for membership attacks [Im12]. Their power depends on the size and quality of the actual and reference cohorts, the number of reported SNP allele frequencies, prior knowledge of the attacker, and the fulfilment of several underlying assumptions, many of which are likely not fulfilled in practice. No identity tracing attack based on aggregate data has been demonstrated yet.

4.4 Single-cell and bulk transcriptomics

[Kra25] comments on the privacy risks of single-cell genomics data. Potential privacy breaches could occur if a publicly accessible eQTL data set contains summary statistics associating specific SNPs with gene expression changes across tissues. Simultaneously, a disease-specific scRNA-seq data set (e.g. from patients with cardiomyopathy) is privately shared among collaborators. An attacker could correlate these gene expression profiles from the single-cell data set with those predicted by SNPs in the eQTL data set, potentially inferring the presence of specific genetic variants in individual patients and uncovering sensitive genetic information, including an individual's predisposition to diseases like cardiovascular or kidney disorders.

[Kra25] refers back to [Tho24] for an examination of the privacy risks associated with bulk transcriptomics. [Tho24] does not provide specific guidance for this type of data, instead referring to the older [Scha12]. [Scha12] used gene expression data of individuals (40,000 transcript counts) to infer genetic variants (1000 SNPs), which allowed them to determine with high certainty whether individuals with known SNPs were members of a gene expression study cohort (N=378). They also assessed the success rate of matching gene expression records to SNP records in a simulated cohort of 300 million individuals, correctly matching 97.1% of the records, which demonstrates the feasibility of cross-linking these data types.

6 Methods for reducing risks to genetic privacy

[Slo25] collects a comprehensive list of technical approaches to reduce privacy risks in this setting. Beyond the methods covered in other references, it highlights:

- Blockchain technology, which records and manages all data access, may help reinforce participants' autonomy. The decentralised nature of the technology removes reliance on biobanks and grants participants complete control over their consent data, as well as over its modification and withdrawal, making them the true data owners. However, here too there are challenges, particularly those that arise from the broadcast property of blockchains, which may leak sensitive facts about participants.
- Beacon services, which are developed to broaden accessibility to genomic data by enabling users to query for the presence of a particular minor allele in a data set; that information, in turn, helps care providers determine if genomic variation is spurious or has some known clinical indication. Principled algorithms can guarantee both privacy and, in some cases, worst-case utility.

[Bra24] notes that complete data protection and fully exploiting the full potential of the data can rarely be reconciled; it provides the following recommendations:

1. **Data collection:** If genotype data collection involves consolidating electronic health records from multiple institutions, privacy-preserving record linkage techniques could be employed. This is not yet practised in most GWAS but is likely to be of interest in the future as recent GWAS studies are increasingly based on electronic health records and related biobanks of patient collectives.
2. **Data storage:** Federated data platforms are emerging as essential resources to facilitate the secure exchange of data without the need to physically move the data outside of its organisational or legal boundaries. For this purpose, for instance, multi-party TREs have been developed to provide a safe space for data analysis.
3. **Quality control:** Apply quality control tools locally. If this is not possible, state-of-the-art privacy practices should be observed, and additional safety measures should be implemented (e.g., secure private clouds). So-called cookbooks can be helpful in this context, and a generation of specially coded and aggregated statistics (so-called partial derivations) for secure, predefined association tests can already take place, so that no person-level information can be obtained from genotype matrices after coding.
4. **Imputation:** Apply local tools if possible or federated solutions. In many cases, this is infeasible because some imputation reference panels are only accessible through the use of proprietary servers. Especially in cases where data that falls under the scope of the GDPR is processed and the server is located outside the EU, this raises privacy concerns. Researchers must strike a balance between the privacy and effectiveness of each option. One option for data processing that is governed by the GDPR is the use of GDPR-compliant phasing and imputation web services, which are also now being implemented in the EU.

5. **SNV association testing and analysis:** If the desired statistical tests are already implemented as a federated version, use federated GWAS tools.
6. **Visualisation:** privacy-enhancing techniques are not required if only summary statistics are used, allowing the use of standard tools on the statistical model output (e.g., for producing Manhattan plots).

[Zho24] provides a comprehensive taxonomy of privacy attacks, their associated risks, and technical solutions to counter them. It suggests the most secure setup to span several of the following: controlled access, anonymisation, cryptographic approaches including homomorphic encryption (HE), trusted execution environments (TEEs), secure multiparty computation (SMPC), federated learning (FL), differential privacy (DP) and blockchain. Its considerations are summarised in Table 4.

Table 4 Scenarios and solutions using privacy technologies across three stages of AI-driven omics method development, from [Zho24].

Stage of development of AI omics methods	Attack	Privacy risk	Privacy solution
Data sharing	Man-in-the-middle (MitM) attack	Intercepting communication between data centres to steal or manipulate omics data	Encryption protocols or digital signatures
	Data breaches	Unauthorized access to sensitive omics data in databases	Controlled access
	Insider threats	Malicious actions by authorized individuals leading to omics data leakage or misuse	Restriction of access to sensitive data on a need-to-know basis
	Data linkage attack and re-identification	Combining seemingly anonymized omics data sets enabling re-identification of individuals	Anonymization, DP, or k-anonymity techniques to prevent re-identification from linked data sets
	DNA phenotyping	Prediction of physical traits, such as facial features, hair colour, and skin colour, from DNA samples	Controlled access
	Pedigree analysis	Disclosure of sensitive family health information or potential identification of individuals through familial connections	Anonymization, pseudonymization, or encryption techniques to safeguard privacy
	Genotype imputation	Prediction of missing genetic variations in a data set and increased risk of re-identification due to expanded genetic data	Controlled access, encryption, or use of DP techniques to mitigate the disclosure risk associated with imputed genotypes
Model training	Data poisoning attacks	Injection of malicious data into omics data sets leading to biased model training or exposure of sensitive genetic information	Encryption, SMPC, or FL approaches to protect raw data during model training
	Gradient construction attack	Exploitation of gradients to reverse engineer or reconstruct sensitive input omics data	Employment of SMPC or FL for secure gradient aggregation or DP for gradient perturbation
Model release	Membership inference attacks	Determination of whether an individual's genomic data was used in the model's training process, leading to privacy breaches	Application of DP mechanisms to prevent inference of membership in the training data set
	Model inversion attacks	Extraction of sensitive genomic information from trained models, compromising individual privacy	Utilization of DP techniques to limit the leakage of sensitive information from model outputs
	Adversarial attacks	Manipulation of model outputs to induce misclassifications or extract sensitive genomic information	Training of robust models with adversarial training techniques to resist adversarial attacks
	Inference data leakage	Risk of input omics data being leaked when using online inference services such as Machine Learning as a Service (MLaaS)	Implementation of encrypted online inference techniques

7 Privacy enhancing methods

7.1 Secure aggregation

Secure aggregation is based on secure multiparty computation (SMC), which enables multiple mutually distrustful parties to compute a shared result without exposing their private data. In the context of FL, secure aggregation ensures that the server can compute the sum of client updates without ever seeing any individual update in plaintext.

Assume that N clients hold model updates $\{g_1, \ldots, g_N\}$. A secure aggregation protocol allows the server to recover only their sum

$$G = \sum_{i=1}^{N} g_i$$

while guaranteeing that no information about any single g_i is leaked.

A classical technique to achieve this is *additive secret sharing*. For illustration, we describe a simplified version of the mechanism. Each client i generates random shares $\{s_1, \ldots, s_N\}$ such that

$$g_i = \sum_{j=1}^{N} s_{ij}$$

Client i sends share s_{ij} privately to client j. No individual share reveals meaningful information about g_i, since each share appears uniformly random. After receiving shares from all peers, each client i computes

$$h_i = \sum_{j=1}^{N} s_{ji}$$

which is its masked contribution. Each client then sends h_i to the server, which can recover the global sum because

$$\sum_{i=1}^{N} h_i = \sum_{i=1}^{N}\sum_{j=1}^{N} s_{ji} = \sum_{j=1}^{N}\sum_{i=1}^{N} s_{ji} = \sum_{j=1}^{N} g_j = G,$$

without observing any individual update g_i.

Modern secure aggregation protocols such as *SecAgg+* [Bel20] extend this principle with improved robustness and scalability. *SecAgg+* employs a secure multiparty coordination scheme, in which each client communicates with only a subset of peers, and adds encryption to protect the exchanged information. As a result, it is:

- **Privacy preserving**: the server observes only the aggregated update and never any individual client contribution;
- **Scalable**: communication costs grow linearly with the dimension of the model update [Li21];

- **Robust to client dropout**: correct aggregation is preserved even if some clients disconnect. While this property is less critical in our cross-silo biomedical setting with a limited number of participants, it remains important for domains such as cross-device FL.

Importantly, the Flower framework includes a ready-to-use implementation of *SecAgg+* that can be used directly in FL workflows.

7.2 Differential privacy

Differential privacy (DP) [Dwo06] provides a rigorous mathematical framework to ensure that the contribution of any single individual's data has only a limited effect on the outcome of a computation. Intuitively, a differentially private algorithm produces outputs that are nearly indistinguishable regardless of whether any particular record is included in the data set. The main goal of differential privacy is to enable useful statistical analysis while guaranteeing that participation does not significantly increase the risk of disclosing private information. Chapter 1 of [Fio25] provides a detailed introduction to the mathematical foundations of differential privacy, while Chapter 8 discusses its application in FL. For a broader, non-technical overview, [Woo18] offers an accessible introduction.

We now introduce several basic concepts needed to state the formal definition of (ε, δ)-DP. As an illustrative example, we consider a computation that takes a medical data set, calculates the average age, and adds a small amount of random noise to the output. The added noise ensures that the same data set does not always produce exactly the same output, which is essential for the privacy guarantees offered by DP.

Before presenting the definition of DP, it is useful to describe a few key elements precisely:

- **Mechanism** M: DP models computations as mechanisms, namely complete randomized algorithms that take a data set as input and produce an output. In our example, M consists of two steps. It first computes the average age from a table of patient records and then adds a small amount of random noise to the result. The combination of both steps, calculation and noise addition, constitutes the mechanism.
- **Domain** D: the set of all data sets on which the mechanism may operate, for instance, all possible tables of patient records.
- **Adjacency**: two data sets D and D' are adjacent if they differ in the data of exactly one individual, for example by adding or removing a single patient record. In the noisy-average-age example, adjacency captures the requirement that the mechanism's output must look nearly the same whether or not a particular person's age is included. This prevents an observer from determining an individual's participation.
- **Range** R: the set of all possible outputs that the mechanism can produce, such as all possible noisy average-age values.
- **Event** $S \subseteq R$: any collection of outputs to which we may assign a probability. For example, all outputs where the noisy average age exceeds a chosen threshold. Events allow us to reason about privacy in probabilistic terms, because instead of focusing on a single output, we consider the likelihood that the noisy output falls within a specified set of values.

Formally, a mechanism $M : D \longrightarrow R$ satisfies (ε, δ)-DP if, for every pair of adjacent data sets $D, D' \in D$ and for every eve$S \subseteq R$nt that could be observed as an outcome of the mechanism, with Pr denoting probability over the randomness of the mechanism,

$$Pr[M(D) \in S] \leq Pr[M(D) \in S] + \delta$$

In this formula, ε and δ are the fundamental parameters of (ε, δ)-DP. The parameter ε is called the privacy budget and bounds how different the output distributions for D and D' may be. Smaller values of ε correspond to stronger privacy guarantees, because they force the outputs obtained from any two adjacent data sets to be very similar. The parameter δ is introduced for practical and mathematical reasons. Many useful mechanisms cannot satisfy pure $(\varepsilon, 0)$-differential privacy. Instead, δ allows for a rare event, occurring with probability at most δ, in which the privacy guarantee may fail and the mechanism could reveal more information than permitted by ε. When $\delta = 0$, the mechanism satisfies the original strict definition of DP introduced in [Dwo06].

Unlike the average-age example, in machine learning the quantity to be privatized is typically a vector computed as a function of high-dimensional data rather than a single scalar statistic. A standard privacy-preserving approach in this setting is the Gaussian mechanism [Dwo14], which adds Gaussian noise to the vector before it is released. To ensure that the required amount of noise is finite, the vector is first clipped to a maximum norm C. If a vector's norm exceeds C, it is scaled down so that its norm becomes C; otherwise, it remains unchanged. After clipping, a calibrated amount of Gaussian noise is added to the vector, producing an (ε, δ)-differentially private mechanism.

In a single (non-iterative) computation, we can directly select the amount of Gaussian noise required to ensure (ε, δ)-DP using a simple closed-form expression. The Gaussian mechanism satisfies -DP when the stand(ε, δ)ard deviation σ of the added Gaussian noise is set to

$$\sigma = \frac{C\sqrt{2}\ln(1.25/\delta)}{\varepsilon}$$

A larger value of σ provides stronger privacy (smaller ε), while a smaller value provides weaker privacy (larger ε).

The single-round expression for σ provided above cannot be used directly in FL because FL is iterative: each communication round incurs a privacy cost, and the total loss accumulates across all rounds. This cumulative cost depends on the number of rounds T, the clipping threshold C, the noise scale σ, and the pattern of client participation (although in our setting we can assume full participation at each round). In practice, σ is treated as a tunable hyperparameter, and the privacy loss is computed using a differential privacy accountant, such as the one implemented in Opacus [You21], which composes the privacy cost across rounds of Gaussian noise addition and reports the resulting overall (ε, δ) guarantee.

The accountant computes the expected privacy loss purely from the training hyperparameters and therefore does not access any data or consume privacy budget. This makes it possible to tune σ offline:

one evaluates the privacy budget for different noise levels and selects the smallest σ for which the cumulative privacy loss remains within the target privacy budget ε.

In this process, one must also specify the parameter δ. Following standard practice in differentially private machine learning, δ is taken to be a very small value, typically an inverse-polynomial function of the total number of training examples N (for example $\delta = 1/N$ or $\delta = 1/N^2$).

At present, Flower provides preview-stage support for DP, including tools for clipping and noise injection as well as an example integration with Opacus. As DP is becoming a standard requirement for FL deployments, these components are expected to mature in forthcoming releases.

7.3 Complementarity of secure aggregation and differential privacy

It is important to distinguish the guarantees provided by secure aggregation from those offered by differential privacy, since the two mechanisms address different aspects of the overall risk profile. Secure aggregation protects client individual updates during aggregation: the server receives only the combined update and cannot inspect any individual contribution. This prevents an honest-but-curious server from inferring information about any specific participant's local data.

However, secure aggregation alone does not defend against white-box inference attacks on the aggregated model, whether during training by parties within the consortium or after training on the publicly released model by external attackers. Differential privacy fills this gap by ensuring that the parameters of both intermediate and final global models do not reveal sensitive information, even to an adversary with full access to the model's internal weights.

The combination of secure aggregation and differential privacy is therefore particularly relevant in biomedical applications, where both confidentiality during training and privacy of the publicly released model must be addressed simultaneously.

8 PLIER

The Pathway-Level Information Extractor (PLIER) [Mao19] is a matrix factorisation method designed for bulk transcriptomic data. Its goal is to produce a lower-dimensional representation of each sample, not in terms of individual genes but in terms of a set of latent variables (LVs), a subset of which can be interpreted using prior biological knowledge. This prior knowledge is expressed through selected gene sets, such as pathways or single-cell signatures. In the latter case, PLIER can be viewed as providing an approximate deconvolution of bulk expression profiles into contributions associated with specific cell types or cell states [Mal25b].

Given a gene expression matrix $Y \in R^{p \times n}$ with p genes and n samples, PLIER incorporates prior biological information via a binary matrix $C \in \{0,1\}^{p \times m}$ that encodes genes membership in selected gene sets. An entry C_{gm} indicates that gene g belongs to gene set m. Using this prior structure, PLIER learns a decomposition

$$Y \approx ZB, \qquad Z \approx CU,$$

with $Z \in R^{p \times k}$, $Y \in R^{k \times n}$, and $U \in R^{m \times k}$, where k denotes the number of LVs.

In this framework, the matrices Z, B, and U are learned during training. The matrix Z contains gene loadings associated with the LVs, while B represents each sample in terms of these LVs. Biological interpretability arises from the matrix U, which encodes the association between the gene sets and the LVs, thereby allowing selected components of U to be interpreted in terms of known biological processes. Contrary to unconstrained factorisations such as singular value decomposition (SVD), PLIER incorporates structural assumptions that guide the latent space toward biologically meaningful patterns. As a result, the representation learned by PLIER is both low-dimensional and partially interpretable.

Because PLIER is not a deep learning model, developing a federated variant requires a detailed examination of its algorithmic structure to determine which computations must remain local to each client and which must be executed in a federated manner. In deep learning, this analysis is largely unnecessary: standard federated training schemes such as FedAvg [McM17] provide a generic and well-tested training loop that can be applied with minimal modification. By contrast, PLIER relies on explicit matrix operations whose distributed execution must be designed explicitly.

To avoid overloading the main text with technical detail, we present this in-depth analysis in the appendix and provide only a compact description of PLIER and federated PLIER here. The appendix includes (i) a complete mathematical exposition of the PLIER training procedure, (ii) additional technical notes on the federated implementation, and (iii) further details on privacy-enhancing technologies for federated PLIER.

8.1 Federated implementation

In a federated setting, the matrices Y and B are partitioned across clients.

Let $i = 1, \ldots, N$ index the participating clients; the matrices Y and B are distributed across clients as $Y = [Y_1 \quad Y_2 \quad \ldots \quad Y_N]$ and $B = [B_1 \quad B_2 \quad \ldots \quad B_N]$ where $Y_i \in R^{p \times n_i}$ and $B_i \in R^{k \times n_i}$ contain the n_i local samples stored at client i, and remain private throughout the entire process. The matrices Z, C, and U are global and therefore known to the clients.

The federated workflow involves calculating the matrices

$$Y Y^T = \sum_{i=1}^{N} Y_i Y_i^T, \quad Y B^T = \sum_{i=1}^{N} Y_i B_i^T, \quad B B^T = \sum_{i=1}^{N} B_i B_i^T.$$

where the first one is computed only once.

All of these quantities are obtained in the same way: each client computes its local matrices $Y_i Y_i^T$, $Y_i B_i^T$, or $B_i B_i^T$ and sends them (preferably through secure aggregation) to the server.

Our experiments indicate that federated PLIER converges to a decomposition equivalent to that obtained in a centralised setting. At the time of writing, these federated implementations have been

evaluated without additional privacy-enhancing mechanisms.

8.2 Privacy-enhancing strategies

We first consider the sensitivity of the information exchanged during federated PLIER training. The only quantities communicated across sites are the aforementioned matrices $Y_i Y_i^T$, $Y_i B_i^T$, $B_i B_i^T$ and their sums. Among these, the covariance-like terms $Y_i Y_i^T$ (and their sum $Y Y^T$) are the most sensitive, as they directly reflect the structure of the raw expression profiles. The other matrices contain substantially less information: the terms $B_i B_i^T$ depend only on the LVs learned by PLIER, and the mixed products $Y_i B_i^T$ entangle gene expression with latent variables, making it considerably more difficult to recover meaningful expression patterns. Consequently, $Y_i Y_i^T$ and $Y Y^T$ are the objects requiring the most careful protection.

To our knowledge, the privacy sensitivity of such covariance-like matrices has not been fully characterised in the literature. However, a study by Zari et al. [Zar22], discussed in more detail in the appendix, is directly relevant to our setting. Their main conclusion is that matrices of this form are sensitive when they are constructed from a small number of samples relative to the number of variables (genes, in our case), but become substantially less sensitive as the cohort size increases. In practical federated deployments, this implies that individual client-level matrices $Y_i Y_i^T$ should never be shared, as some sites may hold only a small number of samples. Instead, secure aggregation must be used to ensure that only the fully aggregated matrix $Y Y^T$ ever becomes visible to the server.

It is also worth noting that bulk transcriptomic data are generally considered less sensitive than other biomedical data types, as discussed in the first part of this deliverable. This further mitigates the attack risk. If stronger guarantees were required, one could additionally apply differential privacy (DP) mechanisms, as also discussed by Zari et al.

Finally, the release of the final PLIER model presents only limited additional risk. The objects shared with downstream users are the matrices C, U, and Z (with $Z \approx CU$), which contain only publicly defined gene sets and their associations with LVs. Moreover, as discussed in the appendix, the non-orthogonal nature of the PLIER decomposition further reduces the risk of inference attacks.

9 scVI and scANVI

scVI and scANVI [Xu21] are deep generative models for single-cell transcriptomics built on variational autoencoders and implemented in the *scvi-tools* ecosystem [Gay22]. Both models rely on a Bayesian hierarchical formulation that explicitly models the uncertainty and technical noise characteristic of scRNA-seq data. A central output of these models is a low-dimensional latent representation well suited for tasks such as data set integration, batch-effect correction, reference mapping, and visualisation.

In scVI, each cell is associated with a continuous latent vector that captures its underlying biological state, together with a "library-size" latent variable that models cell-specific variation in sequencing

depth. The latter is an important component of the generative model, as differences in total transcript counts constitute a major source of technical variation in single-cell RNA-seq data.

Gene expression counts are modelled using a Negative Binomial likelihood, which captures two key properties of scRNA-seq data: overdispersion (variance exceeding the mean) and the high frequency of zero counts often referred to as dropout. These probabilistic components are implemented through a neural encoder-decoder architecture. The encoder maps each observed gene expression vector to a distribution over the latent variables. At the same time, the decoder uses samples from this distribution, together with the library-size variable and any batch covariates, to generate the parameters of the Negative Binomial distribution for the observed counts. This architecture supports efficient minibatch training, allows for the inclusion of covariates in a principled way, and scales to data sets containing millions of cells.

scANVI extends scVI by adding a semi-supervised component for cell-type annotation. In addition to the continuous latent vector, it introduces a latent categorical variable representing the cell type. Labelled cells provide supervision for learning this categorical variable, while unlabeled cells contribute through the generative model. In this way, scANVI jointly learns (i) a latent space shared by labelled and unlabelled cells and (ii) a probabilistic classifier for cell-type prediction.

Unlike PLIER, we do not provide a detailed mathematical formulation of scVI and scANVI here, as their deep learning structure makes federation conceptually straightforward and does not require modifications to the underlying training algorithm.

9.1 Federated implementation

Adapting scVI and scANVI to a federated setting is straightforward, as both are deep learning models for which there is strong evidence that federated learning performs well. The overall design is identical for the two models and follows the standard FedAvg [McM17] paradigm: at each communication round, a subset of participating sites (clients) downloads the current global model, performs several local optimisation steps on their single-cell data sets, and then returns model updates that are averaged on the server to produce the next global model.

Our experiments indicate that the federated variants of scVI and scANVI achieve performance close to that of centrally trained models, both in terms of reconstruction error and the structure of the learned latent space (for example, clustering by cell type), and in label prediction accuracy. In line with standard practice in FL, we explored different numbers of local training epochs between communication rounds, as this parameter is known to influence both convergence speed and final model quality. At the time of writing, these federated implementations have been evaluated without additional privacy-enhancing mechanisms such as secure aggregation or differential privacy.

9.2 Privacy-enhancing strategies

The privacy considerations for federated scVI and scANVI are conceptually simpler than for PLIER, as both models are deep neural networks for which there is extensive prior work on differentially private federated training. During the federated workflow, the only sensitive information exchanged between

clients is the model updates produced at each round of FedAvg. Because scVI and scANVI are high-capacity variational autoencoders, these individual client updates may, in principle, be vulnerable to white-box inference attacks, including model inversion and membership inference.

To mitigate this risk, we plan to combine secure aggregation and differential privacy. Secure aggregation, implemented through the SecAgg+ protocol in Flower, ensures that the server receives only aggregated updates and is never exposed to any client-specific contributions. This prevents an honest-but-curious server from inferring information from individual client updates.

Secure aggregation alone, however, does not prevent white-box inference attacks on the global model, either during training or after training. During training, the periodically updated global model is sent back to all participating sites, so any honest-but-curious party could attempt model inversion or membership inference on the global parameters it receives. After training, once the final model is publicly released, external adversaries with full access to the model weights could attempt the same forms of attack. For this reason, we will incorporate differential privacy into the training pipeline, providing protection both against honest-but-curious parties within the consortium and against external attackers after public release of the final model.

In line with the discussion presented earlier in this deliverable, we adopt the CDP setting. Clients send unnoised updates through secure aggregation, and the server adds calibrated noise to the aggregated update at each communication round. In this setting, the server sees only the securely aggregated client updates, while clients see only their own local computations; because the server is responsible for adding the calibrated noise, CDP assumes a certain degree of trust in the server, which is reasonable in our setting. CDP is known to provide a substantially better balance between privacy and model utility than local differential privacy, which would require each client to add noise independently and would severely degrade model accuracy in high-dimensional settings such as single-cell transcriptomics [Nas22].

We will use the Gaussian mechanism with per-round clipping of aggregated gradients and server-side noise injection. The noise level will be chosen using an offline differential privacy accountant (Opacus), which composes the privacy loss across communication rounds and reports the resulting (ε, δ) guarantee. During model development, we will evaluate the federated workflow on publicly available single-cell data sets and explore a range of privacy budgets, selecting the largest value of ε that yields negligible performance degradation.

10 Genome-wide association studies (GWAS)

Genome-wide association studies use regression models to explain a given phenotype using sequence data (mainly SNPs) after adjusting for environmental factors, experimental design (e.g., stratification) and relatedness between individuals [Bal06]. Thus, they can be broadly denoted as

$$y = Z\alpha + X\beta + \varepsilon, \varepsilon \sim N(0, K)$$

where Z is the design matrix of the factors being adjusted for; α are their effects on y; X is the matrix containing the SNP allele counts (0 = minor allele homozygous, 1 = heterozygous, 2 = major allele homozygous); β are the SNP effects on on y; and ε is an exogenous error term that also captures the kinship relationships between individuals. In modern practice, these relationships are estimated from the sequence data themselves using the covariance matrix K, which is known as the *genetic relatedness* or *kinship* matrix. This basic regression formulation is replaced by an equivalent logistic regression when the trait is binary, such as a disease indicator.

Their key output is a set of estimated effect sizes that quantify how each SNP affects the phenotype, with associated p-values to establish statistical significance. These p-values are typically computed using single-SNP statistical independence tests between each SNP and the phenotype for performance reasons.

10.1 Federated implementation

Federated implementations in the literature rely on two approaches:

- Using secure aggregations to directly learn the global model, aggregating model updates in the form of gradients from clients. sPLINK [Na22] is a notable example.
- Decomposing model learning into simpler operations, such as linear algebra matrix-matrix and matrix-vector operations, and leveraging federated implementations of these operations. SF-GWAS [Cho25] is a notable example.

These approaches represent different trade-offs in terms of algorithmic design. The former requires reformulating existing models in terms of gradients; regression models are typically not off-the-shelf, but rather constrained or penalised optimisation problems. The latter requires a higher level of mathematical sophistication because it involves working with low-level mathematical operations. For instance, federated implementations of linear operations require careful partitioning of matrices into tiled submatrices such that the results of federated operations can later be collated correctly.

10.2 Privacy-enhancing strategies

We first consider the sensitivity of the information exchanged during federated PLIER training. The only quantities communicated across sites are the aforementioned matrices $Y_i Y_i^T$, $Y_i B_i^T$, $B_i B_i^T$ and their sums. Among these, the covariance-like terms $Y_i Y_i^T$ (and their sum $Y\,Y^T$) are the most sensitive, as they directly reflect the structure of the raw expression profiles. The other matrices contain substantially less information: the terms $B_i B_i^T$ depend only on the LVs learned by PLIER, and the mixed products $Y_i B_i^T$ entangle gene expression with latent variables in a way that makes recovering meaningful expression patterns considerably more difficult. Consequently, $Y_i Y_i^T$ and $Y\,Y^T$ are the objects requiring the most careful protection.

To our knowledge, the privacy properties of covariance-like matrices have not been systematically analyzed in the literature. Nonetheless, a study by Zari et al. [Zar22], discussed in more detail in the

appendix, is directly relevant to our setting. Their results suggest that matrices of this form are most sensitive when constructed from a small number of samples relative to the number of variables, whereas sensitivity decreases markedly as cohort size increases. In practical federated deployments, this implies that client-level matrices $Y_i Y_i^T$ should never be exposed, as individual sites may contribute only a limited number of samples. Secure aggregation is therefore essential to ensure that only the fully aggregated matrix $Y Y^T$, which in our practical federated PLIER scenarios we expect to be built from a sufficient number of samples, is accessible to the server.

The release of the final PLIER model introduces only limited additional risk. Although attacks conceptually similar to the ones discussed in [Zar22] could, in principle, be attempted using the released model components, dimensionality considerations substantially mitigate this risk too. In particular, when the number of samples exceeds both the latent dimensionality and the gene dimension, the influence of any single individual on the learned matrix becomes negligible. This effect is further strengthened when incorporating large publicly available transcriptomic datasets into training, which increases the effective sample size and further dilutes individual contributions. We plan to adopt this strategy, as prior work has shown that PLIER benefits from training on larger datasets, even when some data originate from related but not perfectly matched domains, thereby simultaneously improving model quality and strengthening privacy (note that this sample size increase also reduces the sensitivity of the matrices exchanged during training). As an additional safeguard, we plan to complement this approach with simple empirical checks based on local holdout testing, allowing clients to assess membership inference risk prior to model release.

Finally, it is worth noting that bulk transcriptomic data are generally considered less sensitive than many other biomedical data types, as discussed earlier in this deliverable. This further reduces the practical risk associated with potential inference attacks.

Appendix

Standard PLIER training

PLIER's training is conducted through an iterative optimization scheme with initialization based on an SVD. Specifically, an SVD of $Y\,Y^T$ is computed and, together with an "elbow" heuristic on the singular values, is used to select an appropriate number of LVs k, to set the regularization parameter λ_1, and to initialize the matrices $B^{(0)}$ and $Z^{(0)}$ for the subsequent updates.

Subsequently, the method proceeds by alternating updates of Z, B, and U.

At iteration $l+1$, the updates take the following form.

Update of Z

$$Z^{(l+1)} \leftarrow \left(Y\,B^{(l)T} + \lambda_1 C\,U^{(l)}\right)\left(B^{(l)}\,B^{(l)T} + \lambda_1 I\right)^{-1}.$$

Update of U

$$U^{(l+1)} \leftarrow arg\ \ min_U\ \ ||Z^{(l+1)} - CU||_F^2 + \lambda_3||U||_1\ \ ,$$

where $||\cdot||_F$ and $||\cdot||_1$ denote the Frobenius and L^1 norms, respectively, and λ_3 controls the sparsity of U.

Update of B

$$B^{(l+1)} \leftarrow \left(Z^{(l+1)T}\,Z^{(l+1)} + \lambda_2 I\right)^{-1} Z^{(l+1)T}\,Y\,,$$

with $\lambda_2 = \lambda_1 / 2$.

These alternating updates continue until convergence or until a maximum number of iterations is reached. At the end of the procedure, the quantities that constitute the "PLIER model" are the matrices Z, C, and U. Together, they define a biologically structured latent space in which new gene expression profiles can be projected, allowing users to represent their samples in terms of the learned LVs.

Federated implementation details

The initialization step of PLIER requires computing

$$Y\,Y^T = \sum_{i=1}^{N} Y_i\,Y_i^T$$

which can be obtained in a federated manner by having each client compute its local contribution $Y_i\,Y_i^T$ and sending it to the server (ideally via secure aggregation). The aggregated matrix is then used to perform the initial SVD, from which the number of LVs k, the regularization parameter λ_1, and the initial factors $B^{(0)}$ and $Z^{(0)}$ are derived.

In practice, in our experiments, we found that even client-specific initializations lead to convergence, so this federated SVD step is primarily useful for selecting k and tuning λ_1.

Among the three update steps of PLIER, only the update of Z requires federation, whereas the updates of U and B are performed centrally and locally, respectively:

7. B **update (local):** each client updates its own matrix B_i using its local data Y_i and the global matrix Z;
8. U **update (server):** the server updates U using the current global Z and the fixed prior matrix C;
9. Z **update (federated):** requires the aggregation of contributions from all clients, involving the calculation of both

$$Y B^T = \sum_{i=1}^{N} Y_i B_i^T, \quad B B^T = \sum_{i=1}^{N} B_i B_i^T.$$

The server aggregates the contributions, performs the update of Z, and broadcasts the updated Z back to the clients, enabling the next round of local B updates.

Extended analysis of privacy-enhancing techniques for federated PLIER

We discuss here more in detail the work by Zari et al. mentioned in the main text and its implications for federated PLIER. The study analyzes membership inference when an adversary has access to a subset of the principal components of a data matrix X (using a convention in which instances are rows and variables are columns). They show that principal components can leak membership information and that attacks become more effective as more components are revealed.

Let $X \in R^{N \times d}$ be a data matrix with N samples (rows) and d variables (columns). Consider the SVD

$$X = U \Sigma V^{\top},$$

where:

- $U \in R^{N \times r}$ has orthonormal columns,
- $\Sigma \in R^{r \times r}$ is diagonal with singular values,
- $V \in R^{d \times r}$ has orthonormal rows and contains the (right) singular vectors, which correspond to the principal component directions,
- $r = rank(X)$.

Assume the adversary knows the top k principal component directions $V_k \in R^{d \times k}$. From these, they can form the orthogonal projector onto the corresponding subspace:

$$\Pi_k = V_k V_k^{\top}.$$

For any vector $x \in R^d$, the residual after projection is

$$\| x - \Pi_k x \|_2^2.$$

Zari et al. consider test statistics based on comparing this residual for a candidate record against the behavior expected for samples drawn from the training set. Intuitively, samples that were used to construct the PCA subspace tend to be represented "too well" by that subspace, resulting in systematically smaller residuals than for non-members. As k increases, Π_k captures more of the structure of X, and the residual becomes a more discriminative signal, meaning that the attack becomes stronger when more principal components are revealed.

A key conclusion of Zari *et al.* is that attack performance depends strongly on the dataset. For a fixed number of variables and revealed principal components, attacks are effective when the sample size is small relative to the number of variables but rapidly deteriorate as the cohort size increases. This occurs because, in most settings, when a data matrix is not full rank, adding samples increases its effective rank, such that a larger number of principal components is required to capture the underlying structure.

Connection to covariance-like matrices

A key parallel with our setting is that access to covariance-like matrices reveals essentially the same information as a full principal component analysis. Indeed, for a data matrix $X = U \Sigma V^\top$,

$$X^\top X = (U \Sigma V^\top)^\top (U \Sigma V^\top) = V \Sigma^\top \Sigma V^\top = V \Lambda V^\top ,$$

where $\Lambda = \Sigma^\top \Sigma$ is diagonal and contains the squared singular values. Consequently, access to $X^\top X$ is equivalent to knowing all principal component directions V and their associated variances. Thus, releasing $X^\top X$ corresponds to the "maximal information" regime in the threat model of Zari et al.

Importantly, X and $X^\top X$ have the same rank, and the same rank-related considerations as above apply in this setting too. This implies that lower-rank covariance-like matrices are more vulnerable to membership inference attacks.

Membership inference attack carried out by the server

These observations have direct implications for our setting. Let us focus on the matrices $Y_i Y_i^\top$, which are the most sensitive among the covariance-like quantities and the most natural to attack, as already commented in the main text. In principle, an adversary could attempt a membership inference attack by performing an eigendecomposition and applying the same methodology described by Zari *et al.*

Individual clients typically hold only a small number of samples relative to the number of genes, implying that their local matrices $Y_i Y_i^\top$ are inherently low rank and consequently vulnerable to membership inference. Secure aggregation is therefore essential to ensure that only fully aggregated matrices are ever exposed to the server. Since each global matrix is obtained as the sum of local contributions, secure aggregation (in our case, SecAgg+) can be applied directly, ensuring that the server observes only the final aggregate and never any client-level terms.

In contrast to the client-level matrices, the aggregated matrix is constructed from a substantially larger number of samples. In our bulk transcriptomic setting, the data do not exhibit structured low-dimensional artifacts, such as repetitive background patterns sometimes observed in computer vision

datasets. As a result, additional samples are expected to contribute largely independent information, at least until the sample size approaches the dimensionality of the feature space. Under typical training conditions, this leads to a high-rank aggregated matrix, which substantially limits the effectiveness of membership inference attacks.

Membership inference attack on the released model

A separate question is whether membership inference can be performed after training, once the final PLIER model is released. The released objects are the matrices Z, U, C (with $Z \approx CU$), and the regularization parameter λ_2. Suppose an adversary has the gene expression profile $y \in R^p$ of an individual and wants to test whether that individual was included in the training cohort.

Given the released matrix Z, the adversary can compute the individual's latent representation by projecting y onto the learned latent space, using the same procedure employed for standard downstream use:

$$b = \left(Z^\top Z + \lambda_2 I \right)^{-1} Z^\top y .$$

The adversary can then compute the reconstruction residual

$$\| y - Zb \|_2^2,$$

which quantifies how well the released model represents the sample, as $Y \approx ZB$. By analogy with PCA-based membership inference attacks, one might hypothesize that samples included in training could exhibit systematically smaller reconstruction errors than non-training samples, since the model was fitted to them.

To empirically assess this risk, we plan to evaluate the sensitivity of the decomposition through explicit membership inference tests. During training, each client will hold out a subset of its local samples. After training is complete, clients will attempt to distinguish held-out samples from training samples using residual-based statistics of the form above. If these attacks are found to be ineffective across clients, the model can be released with increased confidence that membership leakage is negligible.

In practice, we expect membership inference attacks on the released PLIER model to be weak, based on dimensionality considerations involving the latent space dimension, the number of genes, and the cohort size. In realistic training scenarios within a federated consortium, the number of samples is expected to exceed both the latent dimensionality and the gene dimension. Under these conditions, similarly to what happens in the PCA setting, the influence of any individual sample on the learned Z decreases as cohort size grows. Consequently, Z encodes progressively less sample-specific information, reducing sensitivity to membership inference.

In addition, we plan to adopt a simple mitigation strategy that further reduces membership inference risk while simultaneously improving model quality. Prior work has shown that PLIER benefits from training on larger and more diverse transcriptomic datasets, even when some additional data originate from domains that are not perfectly matched to the target application. Several large, publicly available bulk transcriptomic datasets are suitable for this purpose.

These datasets can be distributed across participating clients, with each client instructed to train using only a randomly selected subset (for example, 90%) of its assigned data. The subset selection is performed locally and is not disclosed. This approach offers two advantages. First, it increases the effective sample size, further diluting the contribution of any single individual and reducing the power of membership inference attacks. Second, because neither the server nor other clients know which specific samples were used by each participant, it becomes infeasible to isolate or subtract the contribution of these records from aggregated quantities. This added randomness thus serves as a practical, lightweight privacy-enhancing mechanism that also improves the robustness and quality of the learned PLIER representation. Note that this mechanism also reduces the sensitivity of the matrices exchanged during training.

References

[Bra24] Brauneck, A.; Schmalhorst, L.; Weiss, S.; Baumbach, L.; Volker, U.; Ellinghaus, D.; Baumbach, J.; Buchholtz, G.: Legal Aspects of Privacy-Enhancing Technologies in Genome-Wide Association Studies and Their Impact on Performance and Feasibility. In: Genome Biology 25 (2024), Nr. 1, S. 154.

[Fio25] Fioretto, F.; Van Hentenryck, P.: Differential Privacy in Artificial Intelligence: From Theory to Practice. (2025)

[Kra25] Kramann, R.; Kuppe, C.; Luyckx, V.; van Biesen, W.; Steiger, S.: Unveiling the Risks: Protecting Privacy in Single-Cell Genomics Data. In: Nephrology Dialysis Transplantation 40 (2025), Nr. 6, S. 1077–1080.

[Mye25] Myers, C. T.; Kumar, R. D.; Pilgram, L.; Bonomi, L.; Thomas, M.; Griffith, O. L.; Fullerton, S. M.; Gibbs, R. A.: Genomic Data and Privacy. In: Clinical Chemistry 71 (2025), Nr. 1, S. 10–17.

[Slo25] Slobogin, C.; Tellis, K.; Clayton, E. W.; Clayton, J. ; Eilmus, A. ; Malin, B. A.: A Decade of Research on Genetic Privacy: The Findings of the Getprecise Center at Vanderbilt University. In: Frontiers in Genetics 16 (2025), S. 1629386.

[Tho24] Thomas, M. ; Mackes, N. ; Preuss-Dodhy, A. ; Wieland, T. ; Bundschus, M.: Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review. In: JMIR Bioinformatics and Biotechnology 5 (2024), S. e54332.

[Woo18] Wood, A.; Altman, M.; Bembenek, A.; Bun, M.; Gaboardi, M.; Honaker, J.; Nissim, K.; O'Brien, D. R.; Steinke, T.; Vadhan, S.: Differential privacy: A primer for a non-technical audience. In: Vand. J. Ent. & Tech. L. 21 (2018), S. 209.

[Zhi25] Zhi, D.; Jiang, X.; Harmanci, A.: Proxy Panels Enable Privacy-Aware Outsourcing of Genotype Imputation. In: Genome Research (2025).

[Zho24] Zhou, J.; Huang, C.; Gao, X.: Patient Privacy in AI-Driven Omics Methods. In: Trends in Genetics 40 (2024), Nr. 5, S. 383–386.

[Sha19] Shabani, M.; Marelli, L.: Re-Identifiability of Genomic Data and the GDPR: Assessing the Re-Identifiability of Genomic Data in Light of the EU General Data Protection Regulation. In: Embo Reports 20 (2019), Nr. 6, S. e48316.

[Mar19] Martinez-Martin, N.; Magnus, D.: Privacy and Ethical Challenges in Next-Generation Sequencing. In: Expert Review of Precision Medicine and Drug Development 4 (2019), Nr. 2, S. 95–104.

[Wr19] Wright Clayton, E.; Evans, B. J.; Hazel, J.; Rothstein, M. A.: The Law of Genetic Privacy: Applications, Implications, and Limitations. In: Ssrn Electronic Journal (2019)

[Lu21] Lu, C.; Greshake Tzovaras, B.; Gough, J.: A Survey of Direct-to-Consumer Genotype Data, and Quality Control Tool (Genomeprep) for Research. In: Computational and Structural Biotechnology Journal 19 (2021), S. 3747–3754.

[You18] Yousefi, S.; Abbassi-Daloii, T.; Kraaijenbrink, T.; Vermaat, M.; Mei, H.; van 't Hof, P.; van Iterson, M.; Zhernakova, D. V.; Claringbould, A.; Franke, L.; Hart, L. M.; Slieker, R. C.; van der Heijden, A.; de Knijff, P.; 't Hoen, P. A. C.: A Snp Panel for Identification of DNA and RNA Specimens. In: BMC Genomics 19 (2018), Nr. 1, S. 90.

[Erl18] Erlich, Y.; Shor, T.; Pe'er, I.; Carmi, S.: Identity Inference of Genomic Data Using Long-Range Familial Searches. In: Science 362 (2018), Nr. 6415, S. 690–694

[Gym13] Gymrek, M.; McGuire, A. L.; Golan, D.; Halperin, E.; Erlich, Y.: Identifying Personal Genomes by Surname Inference. In: Science 339 (2013), Nr. 6117, S. 321–324

[Gue21] Guerrini, C. J.; Wickenheiser, R. A.; Bettinger, B.; McGuire, A. L.; Fullerton, S. M.: Four Misconceptions About Investigative Genetic Genealogy. In: Journal of Law and the Biosciences 8 (2021), Nr. 1, S. lsab001

[Ho19] Ho, S. S.; Urban, A. E.; Mills, R. E.: Structural Variation in the Sequencing Era. In: Nature Reviews Genetics 21 (2019), Nr. 3, S. 171–189.

[Im12] Im, H.; Gamazon, E.; Nicolae, D.; Cox, N.: On Sharing Quantitative Trait GWAS Results in an Era of Multiple-Omics Data and the Limits of Genomic Privacy. In: The American Journal of Human Genetics 90 (2012), Nr. 4, S. 591–598.

[Mal25] Malpetti, D.; Scutari, M.; Gualdi, F.; van Setten, J.; van der Laan, S.; Haitjema, S.; Lee, A. M.; Hering, I.; Mangili, F.: Technical and Legal Aspects of Federated Learning in Bioinformatics: Applications, Challenges and Opportunities. In: Frontiers in Digital Health (2025).

[Scha12] Schadt, E. E.; Woo, S.; Hao, K.: Bayesian Method to Predict Individual SNP Genotypes From Gene Expression Data. In: Nature Genetics 44 (2012), Nr. 5, S. 603–608

[Beu22] Beutel, D.; Topal, O.; Mathur, A.; Qiu, X.; Parcollet, T.; Lane, N. D.: Flower: A Friendly Federated Learning Research Framework. arXiv Preprint arXiv:2204.03008 (2022).

[Mal25a] Malpetti, D.; Scutari, M.; Gualdi, F.; van Setten, J.; van der Laan, S.; Haitjema, S.; Lee, A. M.; Hering, I.; Mangili, F.: Technical and Legal Aspects of Federated Learning in Bioinformatics: Applications, Challenges and Opportunities. In: Frontiers in Digital Health 7 (2025), S. 1644291.

[Bel20] Bell, J. H.; Bonawitz, K. A.; Gascón, A.; Lepoint, T.; Raykova, M.: Secure Single-Server Aggregation with (Poly) Logarithmic Overhead. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (2020), S. 1253–1269.

[Li21] Li, K. H.; de Gusmão, P. P. B.; Beutel, D. J.; Lane, N. D.: Secure Aggregation for Federated Learning in Flower. In: Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning (2021), S. 8–14.

[Dwo06] Dwork, C.: Differential Privacy. In: International Colloquium on Automata, Languages, and Programming (2006), S. 1–12.

[Fio25] Fioretto, F.; Van Hentenryck, P. et al.: Differential Privacy in Artificial Intelligence: From Theory to Practice. (2025).

[Woo18] Wood, A.; Altman, M.; Bembenek, A.; Bun, M.; Gaboardi, M.; Honaker, J.; Nissim, K.; O'Brien, D. R.; Steinke, T.; Vadhan, S.: Differential Privacy: A Primer for a Non-Technical Audience. In: Vand. J. Ent. & Tech. L. 21 (2018), S. 209.

[Dwo14] Dwork, C.; Roth, A.: The Algorithmic Foundations of Differential Privacy. In: Foundations and Trends in Theoretical Computer Science (2014).

[Nas20] Naseri, M.; Hayes, J.; De Cristofaro, E.: Local and Central Differential Privacy for Robustness and Privacy in Federated Learning. arXiv Preprint arXiv:2009.03561 (2020).

[You21] Yousefpour, A.; Shilov, I.; Sablayrolles, A.; Testuggine, D.; Prasad, K.; Malek, M.; Nguyen, J.; Ghosh, S.; Bharadwaj, A.; Zhao, J. et al.: Opacus: User-Friendly Differential Privacy Library in PyTorch. arXiv Preprint arXiv:2109.12298 (2021).

[Mao19] Mao, W.; Zaslavsky, E.; Hartmann, B. M.; Sealfon, S. C.; Chikina, M.: Pathway-Level Information Extractor (PLIER) for Gene Expression Data. In: Nature Methods 16 (2019), Nr. 7, S. 607–610.

[Mal25b] Malpetti, D.; Mangili, F.; Bolis, M.; Rinaldi, A.; Legouis, D.; Ruinelli, L.; Cippà, P.; Azzimonti, L.: Protocol for Interpretable and Context-Specific Single-Cell-Informed Deconvolution of Bulk RNA-Seq Data. In: STAR Protocols 6 (2025), Nr. 1, S. 103670.

[McM17] McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B. A.: Communication-Efficient Learning of Deep Networks From Decentralized Data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (2017), S. 1273–1282.

[Zar22] Zari, O.; Parra-Arnau, J.; Ünsal, A.; Strufe, T.; Önen, M.: Membership Inference Attack Against Principal Component Analysis. In: International Conference on Privacy in Statistical Databases (2022), S. 269–282.

[Xu21] Xu, C.; Lopez, R.; Mehlman, E.; Regier, J.; Jordan, M. I.; Yosef, N.: Probabilistic Harmonization and Annotation of Single-Cell Transcriptomics Data with Deep Generative Models. In: Molecular Systems Biology 17 (2021), Nr. 1, S. e9620.

[Gay22] Gayoso, A.; Lopez, R.; Xing, G.; Boyeau, P.; Pour Amiri, V. V.; Hong, J.; Wu, K.; Jayasuriya, M.; Mehlman, E.; Langevin, M. et al.: A Python Library for Probabilistic Analysis of Single-Cell Omics Data. In: Nature Biotechnology 40 (2022), Nr. 2, S. 163–166.

[Bal06] Balding, D. J.: A Tutorial on Statistical Methods for Population Association Studies. In: Nature Reviews Genetics 7 (2006), Nr. 10, S. 781–791.

[Wai22] Wainakh, A.; Zimmer, E.; Subedi, S.; Keim, J.; Grube, T.; Karuppayah J.; {Sanchez Guinea, A.; Muhlhauser, M: Federated Learning Attacks Revisited: A Critical Discussion of Gaps, Assumptions, and Evaluation Setups. Sensors 23 (2022), Nr 1, S 31.

[Na22] Nasirigerdeh, R.; Torkzadehmahani, R.; Matschinske, J.; Frisch, T.; List, M.; Spath, J.; Weiss, S.; Volker, U.; Pitkanen, E.; Heider, D.; Wenke, N. K.; Kaissis, G.; Rueckert, D.; Kacprowski, T.; Baumbach, J.: sPLINK: A Hybrid Federated Tool as a Robust Alternative to Meta-Analysis in Genome-Wide Association Studies. In: Genome Biology 23 (2022), S. 32.

[Cho25] Cho, H.; Froelicher, D.; Chen, J.; Edupalli, M.; Pyrgelis, A.; Troncoso-Pastoriza, J. R.; Hubaux, J.; Berger, B.: Secure and Federated Genome-Wide Association Studies for Biobank-Scale Datasets. In: Nature Genetics 57 (2025), S. 809–814.

[Me23] Mendelsohn, S.; Froelicher, D.; Loginov, D.; Bernick, D.; Berger, B.; Cho, H.: SFkit: A Web-Based Toolkit for Secure and Federated Genomic Analysis. In: Nucleic Acids Research 51 (2023), Nr. W1, S. W535–W541