Sattanathan *et al. Genome Medicine* (2026) 18:37
https://doi.org/10.1186/s13073-026-01627-9

Genome Medicine

SOFTWARE Open Access

WiNGS-API: a federated genome/phenome data sharing platform enabling gene discovery and variant classification for rare diseases

Nishkala Sattanathan[3†], Benjamin Huremagic[2†], Joris Robert Vermeesch[2], Yves Moreau[1†] and Geert Vandeweyer[3,4*†]

Abstract

Background Genomic research on rare diseases requires large-scale data sharing to identify pathogenic variants. Traditionally, this is achieved through centralized databases that require the physical transfer of genomic data to third-party servers. However, this approach presents significant challenges under data protection frameworks like the General Data Protection Regulation (GDPR), which considers genomic data inherently identifiable and imposes strict limitations on data sharing.

Results To address these limitations, WiNGS-API introduces a federated, automated, and privacy-compliant platform that enables variant-level data sharing across institutions without requiring centralized data transfer. WiNGS-API has been deployed in seven centers, managing over 6,400 samples with single-nucleotide variants (SNV) calls, 1,500 with copy number variants (CNVs), and 500 with structural variants (SVs) from long-read sequencing. The system supports sample-based analysis for variant interpretation and automated trio-based filtering, which reduces candidate *de novo* variants for clinical review. Federated querying, enables the identification of additional carriers and false positives, as well as improved understanding of variants of unknown significance—outcomes not readily achievable with data from a single institution.

Conclusions WiNGS-API bridges the gap between data privacy regulations and genomic collaboration, offering automated variant filtering, annotation, and genotype–phenotype correlation. It provides a scalable solution for multi-center genomic research. WiNGS represents a significant step toward secure, large-scale genomic data sharing, advancing the study of rare diseases while maintaining compliance with ethical and legal standards. WiNGS-API is publicly accessible at https://wings.esat.kuleuven.be/rest-api/api-docs/.

Keywords Federated querying, Variant-level data sharing, Rare diseases, Variant filtering, Clinical genomics

†Nishkala Sattanathan and Benjamin Huremagic are shared first authors

†Yves Moreau and Geert Vandeweyer are shared last authors.

*Correspondence:
Geert Vandeweyer
geert.vandeweyer@uza.be

[1]ESAT-STADIUS, KU Leuven, Leuven 3000, Belgium
[2]Department of Human Genetics, KU Leuven, Leuven 3000, Belgium
[3]Department of Medical Genetics, University of Antwerp, Antwerp 2000, Belgium
[4]Department of Medical Genetics, University Hospital of Antwerp, Edegem 2650, Belgium

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Advancements in genomics and bioinformatics have greatly enhanced our understanding of genetic variations and their role in human diseases. The widespread adoption of whole-genome sequencing (WGS) has exponentially increased the ability to identify diverse genomic alterations that may contribute to a patient's phenotype [1]. However, many patients remain undiagnosed, and this diagnostic gap is even more pronounced in rare genetic disorders. For instance, in neurodevelopmental disorders (NDD), a definitive diagnosis is typically established when the implicated genetic variant is well-characterized, widely acknowledged, and readily detectable. Consequently, due to the high genetic heterogeneity, approximately 50% of the patients remain undiagnosed [2]. Bridging this gap requires the interpretation of complex and nonlinear relationships between variants and phenotypes [3]. Additionally, genetic modifiers may influence clinical presentations by functioning as enhancers or suppressors of disease-associated genes [4].

Traditional gene discovery studies for rare diseases are based on collecting patients with similar or overlapping phenotypic ascertainment followed by shortlisting shared genetic lesions. This approach inherently limits gene identification to higher frequency and core genes and may overlook certain genotype–phenotype associations in case of variable clinical presentation. In contrast, genomic-first analysis groups individuals by mutations affecting the same gene or genes in the same pathway. This genetic similarity-based analysis can provide more subtle insights into how genetic variations influence phenotypic expression [5], but typically requires vast genomic datasets, often beyond the capacity of individual sites. This challenge necessitates data sharing between institutions, which, given the sensitive and personal nature of genomic data, remains particularly difficult to accomplish. Privacy regulations, such as the GDPR in Europe and Health Insurance Portability and Accountability Act (HIPAA) in the USA, restrict the extent of data sharing to protect individual privacy. Full data disclosure is prohibited without explicit patient consent, complicating the aggregation of data across institutions. While high-level statistical summaries can often be shared, they lack the detailed granularity needed for complex genomic analyses. This creates a fundamental dilemma in balancing the need for comprehensive data to advance genomic research with the imperative to protect individual privacy and comply with legal standards.

Federated data sharing offers a promising solution to navigate the regulatory landscape and address the challenges of sharing sensitive genomic data. In a federated model, sensitive data remains under the control of the original custodian, such as a hospital or research institution, while analyses are performed within the custodian's infrastructure. This approach enhances data privacy and ensures compliance with regulations like GDPR and HIPAA, while facilitating access to larger and more diverse datasets. By enabling collaboration without compromising privacy, federated data sharing supports more comprehensive and accurate genomic research.

The concept of federated data sharing in genomics has gained significant traction in recent years, and different implementations and approaches have emerged. One example is the GenoMetric Query Language (GMQL), which provides a flexible framework for analyzing omics data [6]. Originally designed with a centralized system architecture, GMQL introduced a structured query language, similar to SQL, to analyze diverse omics datasets using genomic location as the basic metric for correlation. A federated extension of this model now allows administrators to share predefined datasets between GMQL instances under bilateral data-sharing agreements. However, depending on the query strategy, substantial data may need to be transferred to a processing client, and the sequential execution of complex queries across multiple databases can lead to performance issues. Alternatively, the Genomic Research and Innovation Network (GRIN) [7] is based on a federated setup focused on handling the 'phenotypic bottleneck'. GRIN seeks to establish a definitive and representative reference cohort for rare diseases, which involves collecting and integrating comprehensive genomic and phenotypic data from diverse populations. Another widely adopted approach in the rare-disease community is the Matchmaker Exchange (MME), a federated network of clinical and research databases that enables case-based matchmaking across institutions [8]. Through a standardized API, participating nodes submit structured patient records—typically comprising phenotype annotations and candidate genes or variants—and receive matching patient cases from other connected nodes, while underlying data remain stored locally. The Beacon Project [9] is an example of a platform designed to provide access to anonymized genomic data. Launched as part of the Global Alliance for Genomics & Health (GA4GH) initiative, it aims to facilitate the sharing of clinical and genomic data across federated networks. Initially, the platform offered simple yes/no responses about the presence of specific variants within a cohort, along with basic metadata. Over time, its functionality has expanded to include filtering by phenotypes and disease codes, as well as guidance on how to access relevant cohorts for more detailed exploration.

Federated GMQL, GRIN, MME, and the Beacon Project all operate within a federated context and adhere to FAIR principles. All four approaches share the common intent to expand data exploration capabilities while maintaining strict data privacy boundaries. However, each approach comes with limitations. Federated GMQL,

which relies on a specialized query language, poses a steep learning curve for users unfamiliar with its syntax and nuances. GRIN lacks support for patient-agnostic variant- or gene-based analysis strategies, which can hinder its utility in variant-disease correlation analyses. MME supports case-based matchmaking, where queries originate from an individual patient record and participating nodes return matching patient cases. As a patient-centric system, it does not extend to systematic, variant-level analysis across sites. Beacon v2 introduces enhanced filtering strategies, but does not support initiating automated federated filtering queries across all Beacons. Additionally, combined support for SNV and SV data remains uncommon across these solutions.

Here, we present the WiNGS-API, a federated solution focused on analysis, automation and sharing of SNVs, Insertions/Deletions (InDels), CNVs and SVs. WiNGS-API offers a framework that complies with patient privacy constraints while enabling collaborative genomics by facilitating data analysis and sharing through a straightforward Representational State Transfer (REST) Application Programming Interface (API) approach. By focusing on automation, the integration into clinical processes and pipelines is simplified compared to user interface-based analysis. Guided by the FAIR principles, our approach enhances interoperability and reusability—both essential in the context of rare disorders.

Implementation

WiNGS [10] is designed to facilitate federated computing [11]. To ensure compliance with privacy regulations, analyses of sensitive data are performed by on-premise deployments, referred to as the WiNGS-Client, fully within the client infrastructure (Fig. 1). To accomplish this, all requests are authenticated by and routed through a centralized server, referred to as the WiNGS Central Infrastructure, or WiNGS-CI, towards the WiNGS-Clients. Technical details, including version information concerning this setup, are available in Additional file 1: Table S1.

WiNGS-client infrastructure

Sensitive genomic and phenotypic data is stored and managed on-premises, within the network and infrastructure of the enrolled institutions. This is enabled through a containerized installation of WiNGS-Clients (see Fig. 2). Each WiNGS-Client hosts a Node.js API, referred to as the Client-API, that handles requests from the WiNGS-CI. On-premises data storage is managed using MongoDB, a NoSQL-based solution [12]. MongoDB was selected for its ability to accommodate flexible

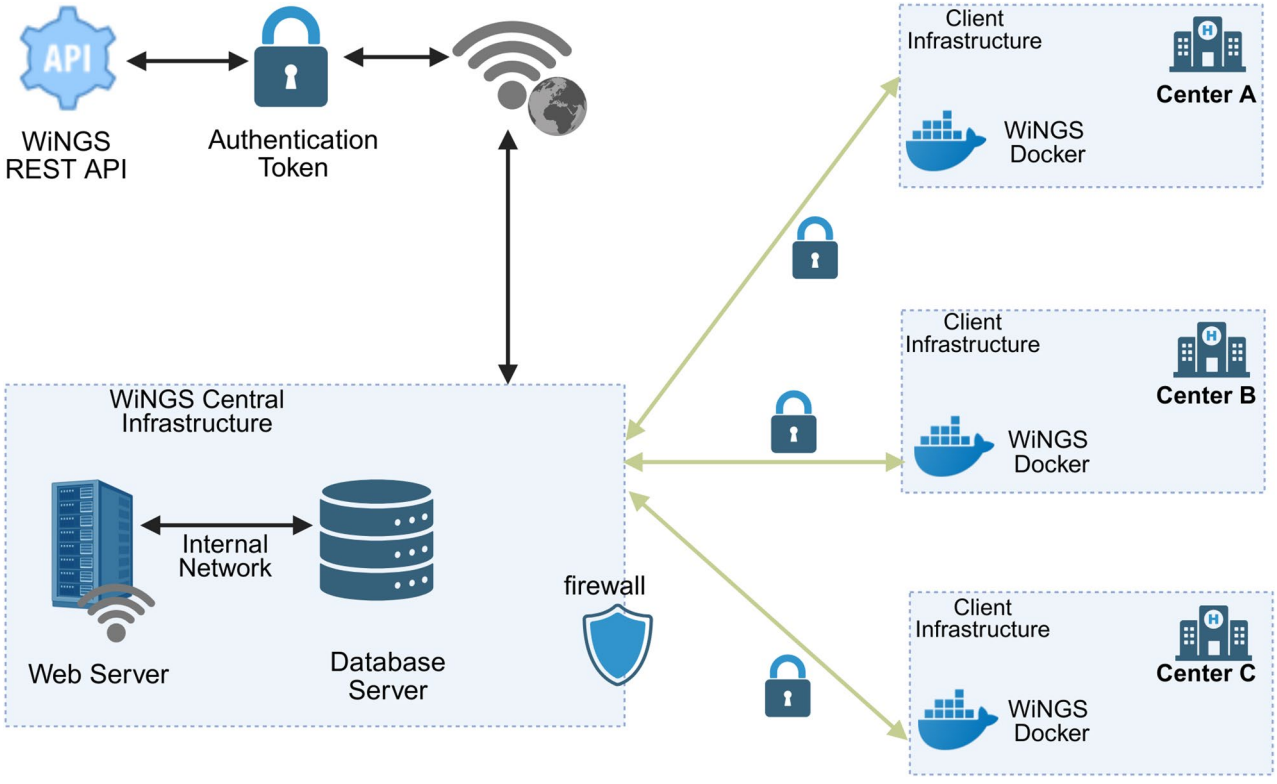


Fig. 1 Overview of the WiNGS architecture: Client infrastructure comprises WiNGS dockers to perform analysis of sensitive data within the center. Authorized users can access the WiNGS interface. Central infrastructure manages access control and forwards authenticated requests to client instances

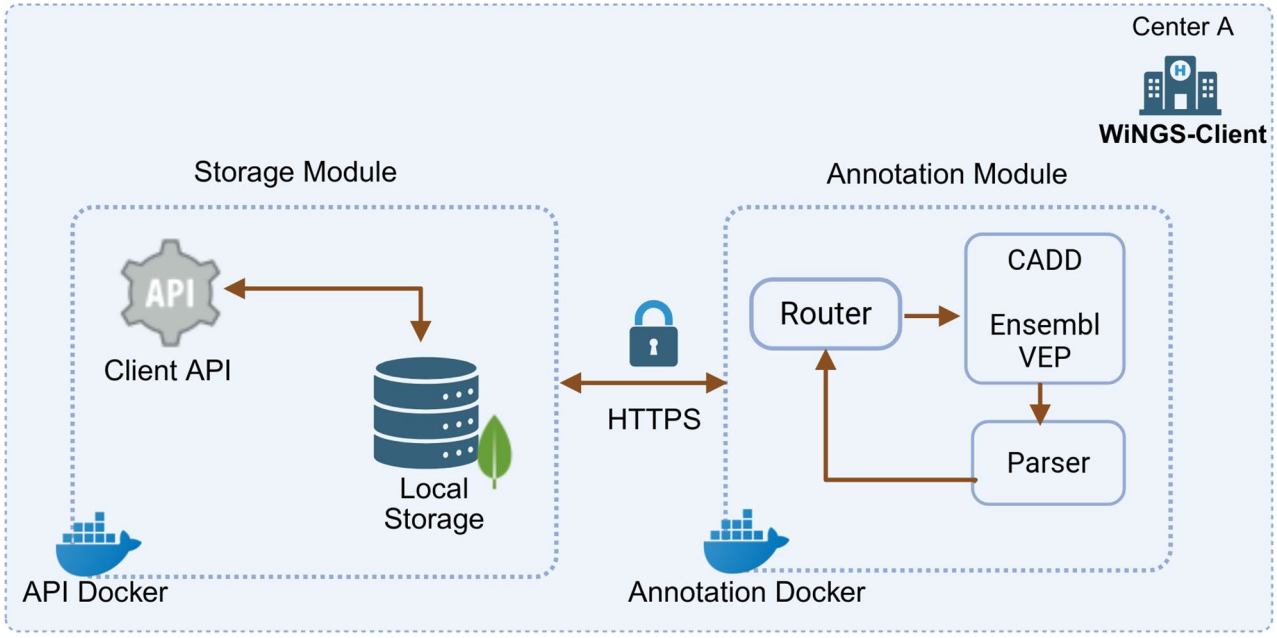


Fig. 2 WiNGS-Client: Components within WiNGS dockers

schemas, support complex queries, and optimize storage efficiency. Configuration of data access and querying criteria is fully handled by the WiNGS-CI, so that no direct user access to WiNGS-Clients is required. Next to the Client-API, WiNGS-Clients can host an annotation module (see below).

SNV and InDel storage

Communication between the Client-API and the WiNGS-CI is secured through JSON Web Token (JWT) authentication and encrypted through standard SSL protocols. The Client-API provides functionality to validate incoming requests from the central infrastructure, interact with the local storage, and provide appropriate responses. Efficient MongoDB-based querying functionalities allow WiNGS-Clients to run on modest hardware, requiring 4 CPU threads and 32GB of RAM. However, it is important to consider the significant storage requirements, which scale linearly at approximately 4 GB per WGS/SNV sample (see Additional file 1: Table S1).

Technically, variants are stored as documents in the database, indexed on a variant key consisting of the chromosome, position, and alleles. Each document contains all annotations and, if samples are assigned to families, a precomputed inheritance vector to allow efficient selection of specific inheritance patterns.

Variant annotation

Variant annotation is a *de facto* standard in genomics data interpretation. With a few million small variants (SNVs and InDels) identified per WGS sample, this can be a significant bottleneck. WiNGS offers a dedicated annotation module, for which the installation inside a WiNGS-Client is optional. Private annotation prevents sending any variant information off-premise. Alternatively, small variants can be annotated by a centrally hosted annotation deployment. The choice depends on the infrastructure capabilities, network policies, and data protection regulations. In either case, variants are only annotated once, when they are first seen by the annotation server. For subsequent samples, the information is taken from a variant cache and only the remaining fraction of newly seen variants will be annotated to reduce overall time.

Technically, the annotation stack for small variants consists of a router module, multiple annotation services, and a parser module (Fig. 2), deployed using Docker Swarm. The router forwards JSON requests to individual annotation tools (Table 1). Results are outputted through JSON files and stored in the MongoDB backend. The annotation files require approximately 2 TB of storage, primarily due to the precomputed CADD scores. In addition to annotating small variants, structural variants are annotated using a Dockerized version of AnnotSV.

Public knowledge in medical genetics evolves rapidly. Therefore, WiNGS provides annotation versioning, enabling the reannotation of all variants, and access to previous annotation versions on demand. The update procedure is semi-automated, involving stopping the Docker Swarm, replacing Docker images with the latest versions (including updated plugins), updating VEP caches to the latest release, and restarting the Docker Swarm. Once completed, all variants at the client site are automatically reannotated. Depending on the

Table 1 Supported Annotations within the WiNGS Annotation module

Annotation Service	Version	Annotation type	Supported annotations
Ensembl VEP [13]	112.0	Gene annotations	Entrez gene ID, RefSeq Transcript
		Functional Effect	consequence
		Population frequency gnomAD(v2) - GRCh37 gnomAD(v4) - GRCh38	AC, AN and other fields from gnomAD
		Non-Coding	Regulatory regions, regulatory features within cell types
			RNA type [14]
		Splicing predictions	MaxEntScan scores
CADD [15]	1.6	Pathogenicity prediction	Phred score
AnnotSV [16]	3.4	Structural Variant	ACMG Class, AnnotSV Ranking, Gene start and end positions

number of samples, this procedure may take several days to complete.

Structural variant storage

Structural variant comparison and merging for trio analysis is performed using an implementation of the Sniffles2 merging algorithm [17]. In trio analysis, variants from the proband, father, and mother are analyzed, generating a SUP_VEC value for each variant, similar to trio-preprocessing for small variants. This binary vector consists of three positions, indicating the presence of the variant in the proband (first position), father (second position), and mother (third position). For instance, a SUP_VEC value of '100' signifies the presence of the variant in the proband but not in the parents. Alternatively, a similar approach can be applied to precompute sample data to exclude population variants. The same algorithm is then employed for population filtering, where the selected sample is compared with a predefined population dataset to identify exclusively sample-specific variants.

WiNGS central infrastructure

WiNGS adopts the Controlled Access model, relying on a centralized access point called the WiNGS Central Infrastructure (WiNGS-CI), to regulate data sharing and access through predefined rules. Although the main functionalities of WiNGS are delivered via a Node.js-based REST API, certain features such as user management, access control, and variant filter configuration are supported through a C#-based web interface. Hardware requirements for the combined API and UI services in WiNGS-CI are modest. The public instance operates on a system with 64 GB RAM and 16 cores, hosting both the Microsoft SQL Server database and a Microsoft Server 2019 virtual machine for the web interface.

Access control

The access control module of WiNGS, defined as WiNGS-ACL, manages the granularity of data access for individual users within the federated system. We apply a hierarchical structure for both users and data (Fig. 3). For every newly registered WiNGS-Client, a Local Administrator (LA) is registered to oversee data imports and user registrations within their organization. Next, Principal Investigators (PIs) are the primary data custodians, and newly imported data are assigned to them. Regular users are created under a specific PI, inheriting full access to the data assigned to that PI. For collaborative purposes, a LA can associate a user with additional PIs (blue dashed line in Fig. 3a) to provide access to additional data. Alternatively, users can be grouped into user groups (green shaded area), which can later be granted access to specific datasets.

Data is also structured hierarchically, as shown in Fig. 3b. Each client center can operate up to two WiNGS-Client installations, for example, an on-premises and a cloud-based instance. To facilitate data sharing with external parties, custom subsets of samples (red area in Fig. 3b) can be created to represent specific cohorts of interest. Within WiNGS, individuals serve as the top-level data entity and require only basic metadata, such as sex (female/male/unspecified). One or more experimental samples can then be linked to each individual. These samples represent the sequencing type (e.g., whole-genome, whole-exome, or panel resequencing). Ultimately, analysis files are associated with samples. Different files represent different types of analysis, for example, analyses against different genome builds or distinguishing between SNV and SV analyses.

Through the user interface, PIs can monitor and manage access rights to their data, including sharing subsets with users or user groups from the same or different institutions. However, processes like obtaining ethical approval remain outside the scope of WiNGS and are the responsibility of the PI.

Data import

Data import is automated through the use of sample sheets, populated by local administrators. Periodic requests are sent by the WiNGS-CI to all WiNGS-Clients to process these CSV-based files (see Fig. 4). Templates for these sample sheets are available within the WiNGS-Client and Table 2 provides a brief overview of key fields. As previously explained, imported datasets are assigned to registered PIs, and if no metadata is available to link

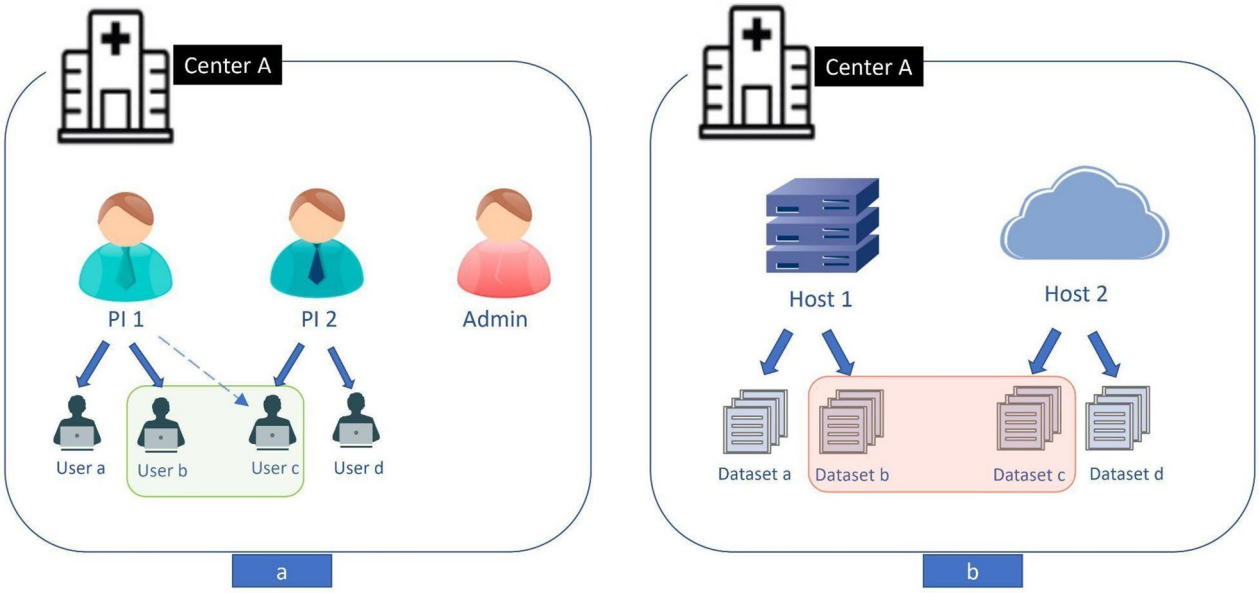


Fig. 3 Reproduced from our preprint [10] describing the WiNGS platform. User (**a**) and data (**b**) hierarchy layout underlying the WiNGS Access Control List. User hierarchy is defined per client center, restricting user groups (green area) to single centers. Data hierarchy is defined per PI in a center, as illustrated on the right for one PI. Datasets (red area) are thus restricted to data from a single PI in a single center. Access to datasets can also be granted across center users and user groups

data files together, each file provided in the sample sheet is treated as a new individual.

WiNGS USER-API

The main functionality of WiNGS is exposed to end users as a REST API, referred to in this manuscript as the User-API, to distinguish between other internal APIs. As a security measure, users have to authenticate through an API key; additionally, each user is limited to a maximum of 360 requests per minute. Exceeding this rate will result in rejected requests. Users can log in to the platform at https://wings-platform.org to generate or revoke their API key through the user menu. Only one active API key is permitted per user at any time. To authenticate requests, users must include their API key in the request headers (see Table 3).

The WiNGS User-API is designed to facilitate and automate large-scale or routine genomic analyses and, therefore, delivers results in a standardized JSON format. The stages of genomic data analysis supported by the API are outlined in Table 4. As illustrated in Fig. 5, end-to-end variant analysis using the User-API is provided through linked automated sample sheet generation (step 1), running variant filters (step 4) using predefined settings (step 3), followed by downstream parsing of JSON files into PDF reports or tabular formats

Privacy safeguards

WiNGS incorporates several privacy safeguards directly into the query-processing layer and the central aggregation layer. Phenotype filtering uses an ANY-term matching strategy: individuals are included in results if they possess *at least one* of the queried HPO terms. To reduce the risk of exposing rare phenotype–genotype combinations, all aggregated outputs are subject to a minimum cohort-size threshold of five individuals; queries returning fewer than five matches are suppressed. This follows small-cell suppression practices used in European secure-data infrastructures, such as the European Medicines Agency (EMA) DARWIN EU (Data Analysis and Real-World Interrogation Network) program, which applies a minimum cell count of five in reported analyses [18].

To further mitigate probing attacks or iterative narrowing of cohorts—risks previously highlighted in privacy analyses of genetic data [19, 20],—WiNGS implements several User-API–level controls. Per-phenotype query limits restrict users to 15 phenotype–genotype association queries for the same HPO profile within a defined time window. Next, the repeated execution of identical queries from the same user is capped at 10. Additionally, WiNGS performs enhanced logging of suspicious or systematic query patterns to support audit and review by participating centers. These measures complement a strict cap of 1,000 variants per correlation query, and operate alongside the existing API authentication and rate-limiting mechanisms.

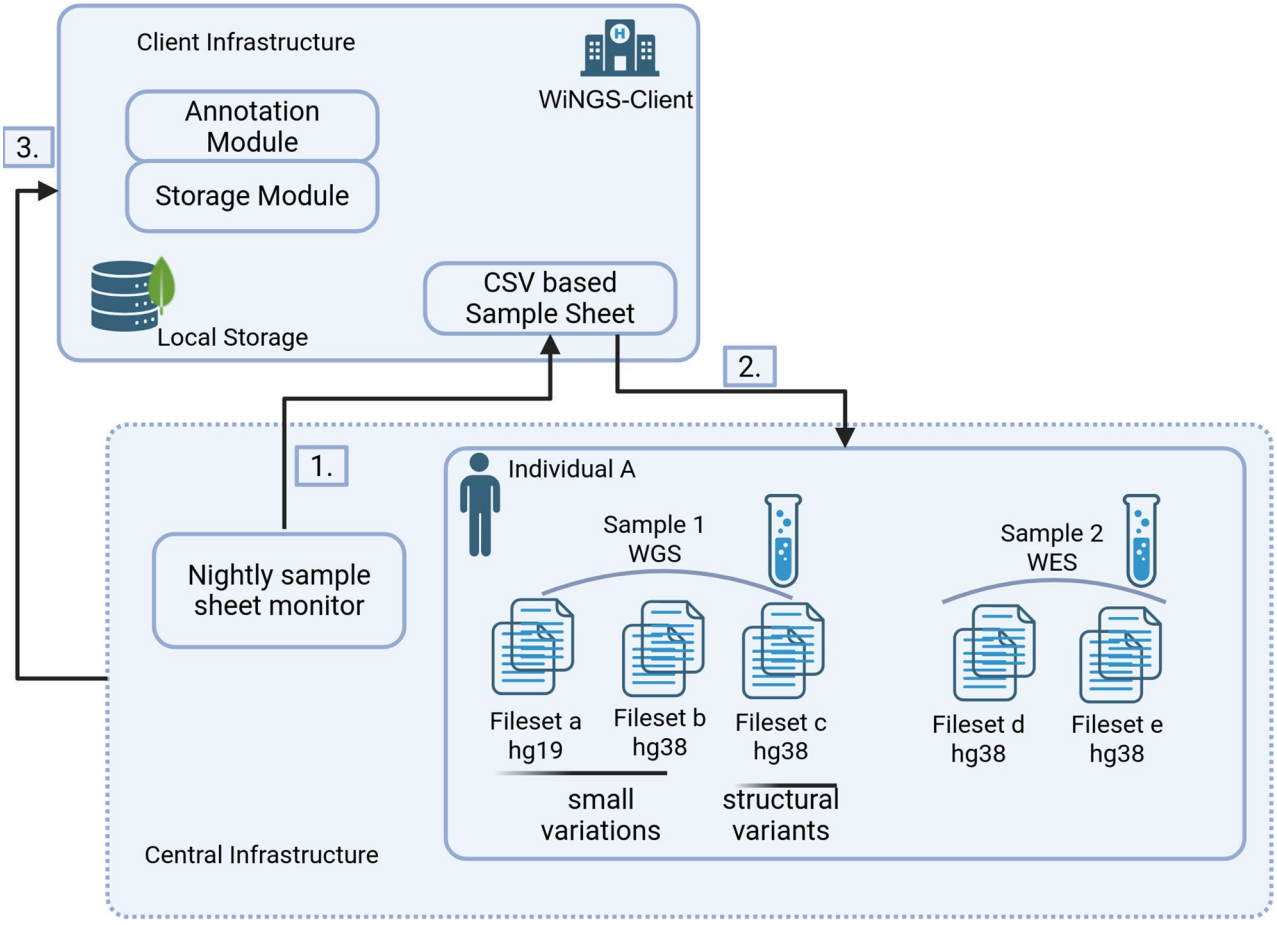


Fig. 4 Data Import process. (1) CSV-based files located at the WiNGS-Client, queried daily by the central infrastructure to fetch new samples to be imported in bulk. (2) Anonymised identifiers are automatically created for Individuals, samples, and datasets, which are internally linked (3) (g)VCF files are parsed, annotated, and stored within the WiNGS Client

Results

WiNGS has currently been deployed in seven centers, hosting SNVs from approximately 6000 short-read WGS and WES samples. The WiNGS infrastructure does not support extraction of variant numbers from remote clients. However, as an indication, we can report numbers from the centers with which we, as authors, are affiliated. One center hosts 130,765,336 SNVs, while another center has 28,709,852 SNVs. In addition to SNVs, CNVs were called from short-read sequencing in 1,500 samples and from long-read sequencing in 500. In terms of structural variation, 500,034 CNVs were detected from short-read WGS and 4,873,166 from long-read WGS. To provide a representative number of structural variants within the platform, the majority of long-read sequencing variants are extracted from publicly available lrWGS from the 1000 Genomes Project [21] .

We evaluated the performance through benchmarking three key areas: import, annotations, and querying. First, import time is significantly reduced with the implementation of annotation caching—from approximately 410 min (6.8 h) for the first run to about 186 min (3.1 h) on the second run, and down to just 84 min for subsequent samples (see Additional file 4: Table S2A). Single sample WGS filtering can be done in real time (see Additional file 4: Table S2), with average response times < 12 s. Even federated querying achieves nearly real time response rates of < 50 s, enabled by parallel execution across centers and indexed queries.

(Multi-Center) federated data analysis to identify pathogenic variants

We present a synthetic use case to demonstrate the key functionalities of WiNGS-API, in supporting rare disease diagnosis through secure, cross-institutional variant discovery and interpretation. This use case was evaluated using a stepwise genomic interpretation workflow combining sample discovery, trio analysis and federated genotype–phenotype association (see Additional file 5: Fig. S2).

Sample Discovery API is aimed to identify pathogenic SNVs/SVs in individual samples. To demonstrate this

Table 2 Data Import Fields with the corresponding metadata that can be used with the sample sheet. Variant information can be imported from both VCF and gVCF files. Pointers towards BAM, CRAM, and FASTQ files are also managed in WiNGS. SV_VCF is used for importing VCF files that contain structural variations. Only hg38 AssemblyType is supported for SV_VCF File type. Sample Sheet (CSV) template provides a detailed overview of all the available fields (see Additional file 2)

Data Entity	WiNGS Term	SampleSheet Field	Example
Principal Investigator	PI	PIName	FirstName LastName
Patient	Individual	IndividualSex	female/male/Unknown
Experiment	Sample	SequenceType	WES/WGS/Panel
		SequenceMachine	NovaSeqX
		Readlength	150
Data analysis Results	File	FileType	(g)VCF, BAM, CRAM, FASTQ
		FileType	SV_VCF
		FileLocation	/server/wings-client1/file1.vcf
		AssemblyType	hg19,hg38

Table 3 Example of an API request to retrieve the Individuals within the center of the authenticated user. Authentication is handled through the Json Web Token, which is coupled with the WiNGS-ACL. The API_KEY is provided as part of the request header

GET /rest-api /individuals HTTP/1.1
Host wings.esat.kuleuven.be
Authorization : Bearer $API_KEY
Accept: application/json

functionality for SNVs, a synthetic genome was constructed using the variants in the *PUF60* gene, which is associated with Intellectual Disability. Steps taken in this analysis are available in Additional file 2. A *de novo* variant in *PUF60* (NM_001271096.2:c.404G > T;p.Gly-135Val), which had not been described in the literature at that time was spiked into the Illumina paired-end genome sequencing data of HG002, the Ashkenazi Jewish NIST/GiaB Trio Index [22]. After importing the VCF files into WiNGS, phenotypic descriptors reflecting "Intellectual Disability" were also assigned to the proband. Variant filtering criteria included quality, variant frequency in control populations, and predicted functional impact by VEP and CADD. Using these criteria, Sample Discovery successfully retained 119 candidate variants in a dataset of 5.6 million.

Trio Analysis, which compares the variants of an index patient against the biological parents (father and mother), can further reduce the list of candidate variants. To assess the *trio analysis* functionality, parental genomes (HG003 and HG004) from the Ashkenazi Jewish NIST/GiaB Trio [22] were imported (see Additional

Table 4 Different Stages of WiNGS User-API Usage and Automation Options

Stage	Endpoint/Tool	Details
Data Collection	/samplesheet/status	Track import status based on a specific date
Data Collection	/individuals/phenotype	Manage patient phenotypes as Human Phenotype Ontology(HPO)
Data Collection	/families	Create families
Data Collection	/families/assign	Link Individuals to the family
Data Analysis	/sample/discovery	Filter SNV variants within a sample
Data Analysis	/SV_sample/discovery	Filter SV variants within a sample
Data Analysis	/trio/inheritance	Trio based SNV analysis using inheritance-based strategies
Data Analysis	/SV_samples/trio_discovery	Trio based SV analysis using inheritance-based strategies
Data Analysis	/families/filter/query	SNV analysis using generalized familial relations
Multi-Center Data Analysis	/variant/frequency	Frequency of a variant within the WiNGS eco-system
Multi-Center Data Analysis	/variant/association	Variant and Phenotype correlation query
Multi-Center Data Analysis	/variant/discovery/query	Query SNV variants based on a region or gene
Multi-Center Data Analysis	/SVvariant/discovery/query	Query SV variants based on a region or gene
Multi-Center Data Analysis	/variant/discovery/collaborate	Request for collaboration based on the result from the above analysis
Authentication	/revoke? type=admin	Admin specific API to disable API access for a specific user
Authentication	/reactivate? type=admin	Admin specific API to reactivate access for a specific user

Data Collection: This stage involves importing patient information and data files (such as VCF), adding phenotypes, and creating family records

Data Analysis: At this stage, the APIs enable users to perform analyses within their own centers

Multi-Center Data Analysis: These APIs provide anonymized results, as the analysis is conducted across all WiNGS clients

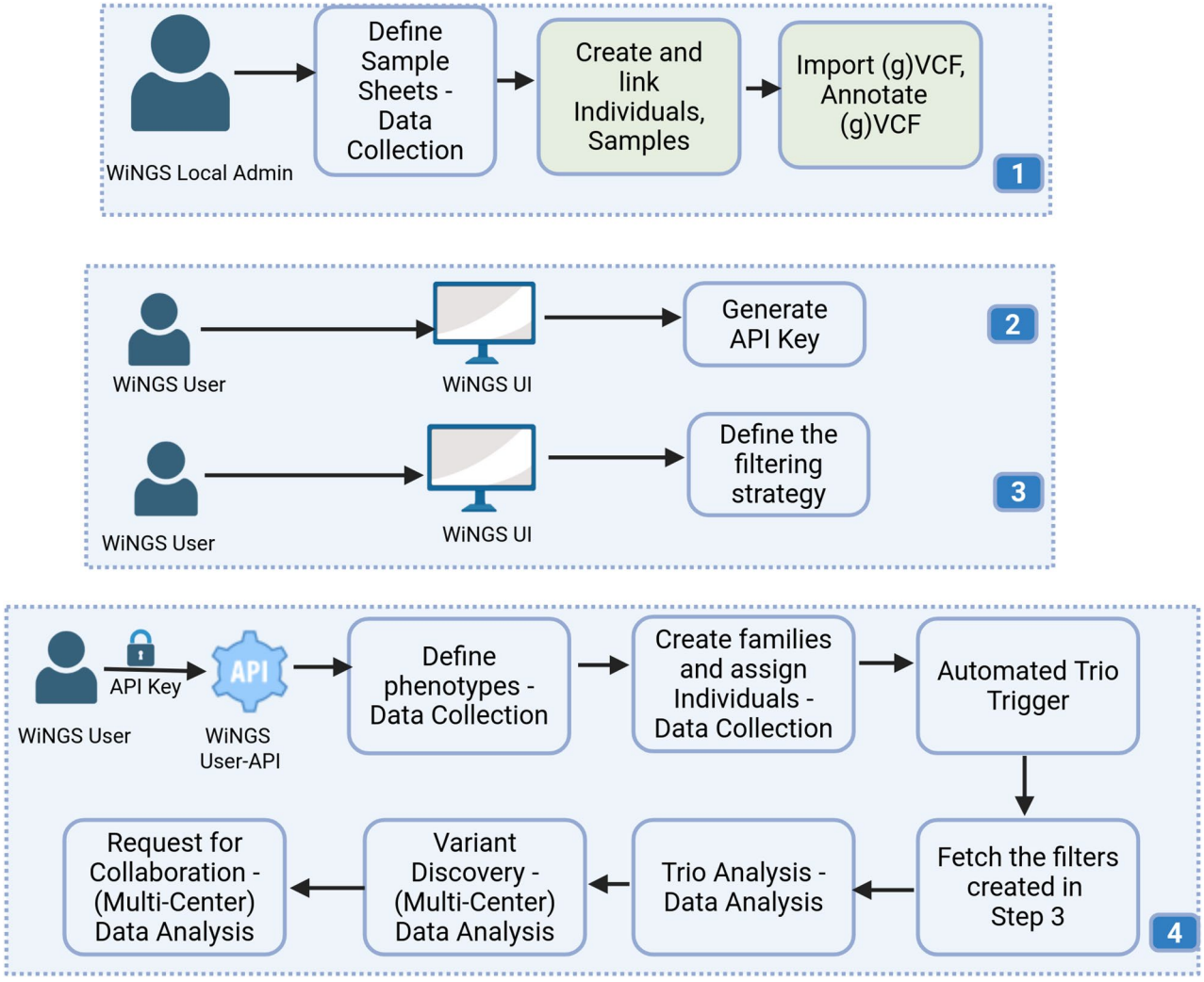


Fig. 5 Data Analysis Approach. Step 1: Sample sheets defined by WiNGS Local Admin are parsed. The sections highlighted in green are automated processes. Step 2: The authorized WiNGS User has to perform a one-time process to generate an API key using the user interface. Step 3: Define filtering strategies to be applied on imported data (see Additional file 2). Step 4: WiNGS Users can use the APIs corresponding to each stage (e.g., define phenotypes) to perform the specific action. Automated Trio triggers precomputation of variants shared between families (proband, father, mother) for further analyses

file 2). Precomputation identified 292,405 *de novo* variants. Out of these, the Trio Analysis API workflow (see Fig. 6) retained six variants under the hypothesis of a *de novo* mutation and passing the functional filtering criteria like allele frequency, predicted functional impact that were applied before. However, based solely on *de novo* inheritance and functional annotation, its clinical significance could not be confidently established. This reflects a typical diagnostic challenge where *de novo* findings, though suggestive, require additional evidence—such as phenotype association—to support pathogenicity.

Application of Sample Discovery and Trio Analysis on the synthetic data did not yield any definitive conclusion. Hence, we set out to demonstrate the potential of WiNGS to interpret the six remaining apparently *de novo* variants by mining anonymized and aggregated data from different clients. The following sections highlight how such aggregation can help in three interpretational challenges, i.e. promoting a VUS to a pathogenic variant, degrading a VUS to a benign class and excluding variants as technical artifacts.

First, we applied genotype–phenotype association on the *PUF60* variant to evaluate potential correlation with the phenotype (see Additional file 2). Single variant Fisher exact testing over all WGS samples, between individuals with and without a phenotype of "Intellectual Disability" identified two more carriers of the exact same variant, presenting with intellectual disability, and no unaffected carriers. Although the number of carriers is low, this results in a significant correlation between the phenotype "Intellectual Disability" (HP:0001249) and the variant (Fisher exact < 0.01). Disruptive variants in *PUF60* are associated with Verheij syndrome [23] which is in line with the phenotype of the synthetic case. The

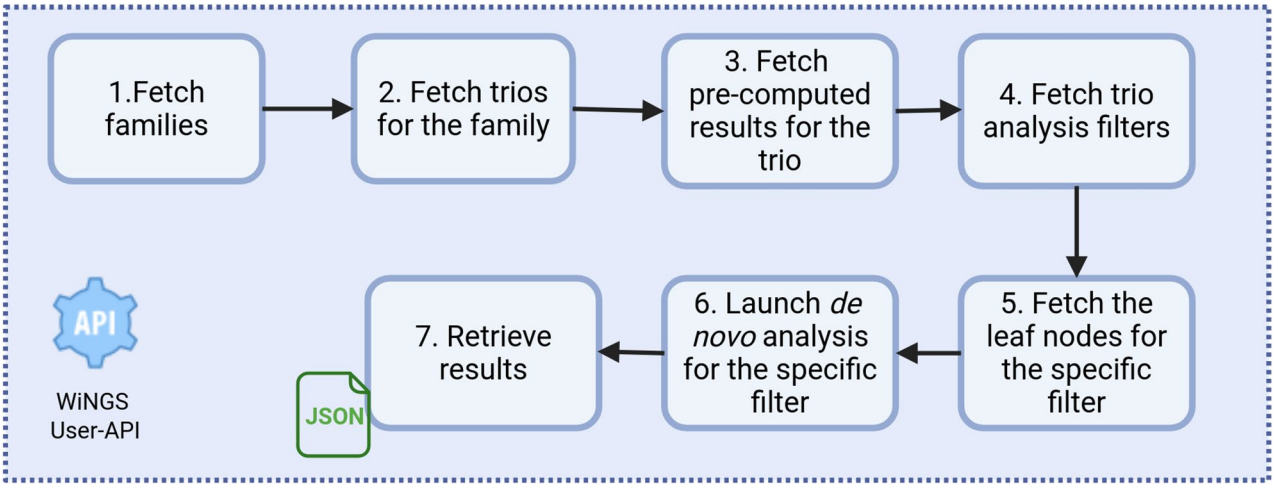


Fig. 6 Small variants trio analysis Workflow. (1) Families are returned based on the user access level. (2) Trios available for the family are listed. (3) Pre-computed segregation indicates the counts of variants shared between the family members based on different inheritance patterns. (4) Filters within the Trio Analysis modules are listed. (5) The leaf node id indicates the filter conditions corresponding to the specific node. (6) Launch a request to filter the de novo variants based on the created filter conditions. (7) Retrieve results for the specific request

identification of two more carriers of an identical mutation and similar phenotype, adds strength to the interpretation of this nonsynonymous variant as causative. As a final step, WiNGS-API allows users to contact the PI in charge of the data for the identified additional carrier samples. Up to this point, no additional information is shared about the second carrier, and it is up to PIs to engage in the collaboration and share additional data. As such, this feature complements inheritance-based methods and supports candidate variant prioritization at a population scale, without compromising data privacy.

(Multi-Center) data analysis to assign VUS as benign

The second analysis demonstrates how WiNGS-API can be used to readjust a variant class from VUS to benign, by incorporating genotype–phenotype evidence from federated analysis (Additional file 6: Fig. S3 for the workflow of the use case).

To evaluate *genotype–phenotype association functionality* at the variant level, we wanted to investigate the variant MYO6p.Gln918Thrfs*24, already reported in the *Sample Discovery* step of the previous use case. The variant is assigned as pathogenic by both ClinVar and CADD, and the population frequency is very low according to gnomADv2 (see Additional file 2). To determine the potential pathogenicity of the variant in our intellectual disability population, a Fisher's exact test was internally performed by WiNGS across the federated dataset. This analysis did not detect a statistically significant association with intellectual disability; however, given the sparse counts, this should be interpreted as a lack of detectable enrichment rather than definitive evidence of no association.

As mentioned earlier, *MYO6* has not been described to be related to intellectual disability, so non-surprisingly, we did not expect to identify single variant correlation. However the analysis is included here to demonstrate, assuming no prior knowledge on *MYO6*, the ability to exclude potentially damaging variants in the context of a phenotype when supported by sufficient data. To extend this analysis, we performed a gene-wide analysis on *MYO6* to correlate filtered (pathogenic-likely, low frequency) variants within the gene of interest with a broadened phenotype, including both first-level parents and children of intellectual disability in the HPO tree. In contrast to the single-variant analysis, gene-based analysis is conducted on a single harmonized cohort, in this case derived from WES data, across the federated nodes. Functional filtering criteria-such as low population frequency and possibly damaging effects were applied to restrict the analysis to potentially deleterious variants. Next, the resulting dataset of just over 5,500 variants, was submitted to the correlation analysis endpoint for Fisher's exact test. Again, as expected, the resulting Fisher's exact test did not detect a statistically significant association with the extended phenotype of interest (see Additional file 2). As with the variant-level test, this reflects a lack of detectable enrichment rather than confirmed absence of association. Nevertheless, these results highlight the platform's ability to support gene-level exploratory analyses across federated datasets while preserving data privacy.

(Multi-Center) data analysis to disentangle a genuine pathogenic variant from sequencing artifacts

MYO6 mutations are causative for a dominant form of progressive sensorineural hearing loss. As mentioned earlier, variant MYO6_Gln918Thrfs*24, is listed as

pathogenic on ClinVar and predicted as disruptive by CADD. Hence, we wondered whether this phenotype could be extracted from the federated HPO terms.

In this third use case we therefore aim to demonstrate how the WiNGS-API platform can allow users to explore genotype–phenotype associations while incorporating quality control to distinguish true biological signals from sequencing artifacts. It showcases the platform's utility in refining associations through iterative filtering and statistical validation.

We first applied the *variant discovery functionality* to assess a potential association between the identified variant in *MYO6*, and first level parents and children of "(progressive) sensorineural hearing loss" using the cross-center WES cohort, similar to the previous use case but using the known phenotypic relation. This time, the filtering criteria were restricted to only the variant of interest instead of the whole gene. The platform reported 579 WES samples presenting with this form of hearing loss and 138 WES samples having this variant (see Additional file 2). As most of the variant carriers (123 samples) did not present with a related phenotype, no statistically significant association was detected (P-value = 0.88), with phenotype rates being similar between carriers and non-carriers.

To assess whether the absence of statistical signal might be due to sequencing artifacts, we then extended the filtering *strategy* beyond the variant itself. Additional filters—such as minimal sequencing depth (20x) and a minimal allelic-balance (minimal 30% of alternate allele reads) were added to the analysis. After applying these quality filters, all 138 variants reported before were excluded, indicating that the variant calls did not meet quality thresholds (see Additional file 2). This step demonstrated that the original findings were likely false positives resulting from low-quality genotyping.

Finally, we evaluated *variant discovery functionality* on the broad phenotype description and the combined set of high quality, potentially damaging variants (CADD > 25) in *MYO6*, specifically excluding the potential false positive variant identified above. The association was re-evaluated across the same WES cohort and HPO term set. This refined analysis resulted in a statistically significant association (P-value = 0.008), with an estimated odds ratio of 1.93 and a relatively narrow 95% CI (1.23–3.05), indicating an enrichment of the phenotype among carriers of high-confidence *MYO6* variants. The outcome illustrates how the platform enables users to iteratively refine analysis results, filter out sequencing noise, and ultimately uncover genuine associations while safeguarding against false-positive interpretations.

Identification of pathogenic large structural variants

With the advent of long-read sequencing technologies, the detection of SVs has significantly improved. Long-read sequencing identifies approximately 25,000 SVs per sample [21]. Identifying the pathogenic variants remains a daunting challenge.

To demonstrate the platform's capabilities, we defined a trio within WiNGS consisting of a proband with an undiagnosed NDD and both parents, providing a representative use case for the identification of rare, potentially pathogenic structural variants. Variants identified from GRCh38 whole-genome Long-read Sequencing (WG-LRS) data, larger than 50 bp supported by over 3 supporting reads were filtered and uploaded for the proband, mother, and father. The proband alone presented 15,073 SVs. (see Additional file 7: Fig. S4, Additional file 8: Fig. S5). 503 candidate *de novo* SVs were retrieved following filtering out inherited SVs. To identify pathogenic events, variants classified as ACMG Class 4 (likely pathogenic) or Class 5 (pathogenic) were selected using annotation-based filters. A single variant was retrieved: a heterozygous ~ 3.5 Mb duplication at chr16:14,894,754–18,372,712, classified as ACMG Class 5 and involving several known NDD associated genes, including *NDE1*.

SV interpretation remains a major challenge. Hence, population incidences of SVs can help in the classification. WiNGS was designed to address this need by enabling federated queries across institutions, combining variant frequency data with phenotypic context to support clinical classification.

To demonstrate this functionality, we explored the potential role of a *CYP2A6* and *CYP2A7* deletion on chr19 (40847970–40882460, hg38) as a cause of DD in Patient X. This variant was not observed in the local control population, but was classified as ACMG Class 4, suggesting it could be of clinical significance. However, the potential role in Patient X's condition remained uncertain. The patient presented with global developmental delay and skeletal dysplasia (HPO terms: HP:0001263, HP:0002652). Since *CYP2A6* deletions and mutations are primarily associated with altered drug metabolism and not with developmental disorders, a population based analysis was instigated.

We queried the federated WiNGS database and identified 75 carriers out of 1375 individuals, yet none exhibited phenotypic overlap with our patient. The corresponding contingency table showed no individuals with the patient's phenotype in either the carrier or non-carrier groups (see Additional file 9), and the Fisher's exact test accordingly returned a non-significant p-value of 1.0. While a p-value of 1.0 does not provide direct evidence of absence of association, the absence of overlapping phenotypes among a relatively large number of carriers

supports the interpretation that this deletion is unlikely to be pathogenic in this clinical context.

Discussion

The identification of novel genetic causes for undiagnosed rare diseases which are not transmitted in families requires large cohorts of genomic and phenomic data from such patients. Hence, large-scale centralized databases have been established. However, both practical and regulatory aspects limit the data sharing paradigm. Here we present federated WiNGS-API, a platform enabling privacy-aware data-sharing and analysis across federated cohorts. This API-based approach supports large-scale analyses at both single and multi-center levels, while the federated analysis ensures that sensitive genetic and clinical data remain within the infrastructure and under the control of the original data controller.

The cornerstone of federated data analysis applications is formed by robust security and privacy protocols to prevent any data leakage. Homomorphic encryption (HE), secure multiparty computing (sMPC), differential privacy, and controlled access are some of the available technologies [11]. HE is a technology where questions are answered by federated data pools using encrypted data only. HE has been used successfully to perform federated GWAS analysis [24] but has the disadvantage of limited analysis options and heavily centralized computing requirements on the encrypted data. Secure Multiparty computing allows multiple parties to collaborate and perform computations without revealing their private inputs, ensuring data privacy and security [25]. The diversity of network technologies and complex protocols used among computing parties can create security loopholes that attackers may exploit, compromising the integrity, confidentiality, and availability of computations. Additionally, sMPC protocols normally involve complex cryptographic operations that are computationally intensive. Differential Privacy is a mathematical technique used to add controlled noise to the results of data analyses. However, adding noise can reduce the accuracy of the data or the results of analyses [26]. Controlled Access is a method that restricts access to sensitive data, explicitly authorizing access only to authenticated users. Controlled Access is relatively straightforward compared to the other privacy techniques, and the lower computational overhead makes it suitable for heavy data analysis applications. The Access Control List (ACL) implements Controlled Access for the WiNGS-API, allowing fine-grained control over access to specific datasets. Access rights are imposed internally by all WiNGS-API calls (Table 4) through the mandatory bearer token. Additionally, Admin-specific API endpoints can be used to revoke API keys in case of unforeseen user behavior. Considering security, accuracy of results, and performance on large genomic datasets, we can conclude that the implemented Controlled Access approach is both sufficient and suitable to provide the necessary safeguards and performance for federated analysis of genomic variants.

To protect patient privacy, the GDPR establishes data minimization as a core principle, favoring architectures that reduce data exposure whenever possible. When both centralized and federated models can achieve the same analytical purpose, GDPR encourages the use of federated models, which limit data movement and keep sensitive information under institutional control (Art. 89(1), *GDPR-Info.eu*). However, the architectural setup of federated platforms must then prevent misuse of distributed data and avoid reconstruction and tracing attacks [27]. Reconstruction attacks occur when sensitive data about most individuals within the dataset are revealed. It would happen if, from the aggregate statistical data, the data of underlying individuals could be reconstructed. Tracing attacks are more modest and refer to the ability to infer whether or not a specific individual is part of the dataset [28]. To perform a tracing attack, the adversary requires prior access to the genetic profile of the individual but could, for example, disclose additional medical information such as a particular diagnosis, based on the metadata of the attacked data source [28]. Mitigation strategies against re-identification attacks exist, including statistical noise injection for differential privacy or query monitoring to limit systematic data scraping [28–30]. We decided against the implementation of noise injections because it might negatively impact data accuracy and because the Controlled Access mechanisms authorize only registered users, who are always affiliated with, and under the authority of, a participating center. Upon registration of novel client centers, we request stated adherence to the European Code of Conduct for Research Integrity by the center's representative. These guidelines, together with enforced authentication, rate limiting, and additional safeguards such as minimum cohort-size thresholds, restrictions on repeated or phenotype-identical queries, and audit logging, provide a robust foundation for operating WiNGS within GDPR and similar regulatory frameworks. At the same time, the current minimum cohort-size threshold of five individuals—while aligned with European small-cell suppression practices [18] and offering a conservative privacy safeguard—also has practical limitations. Fixed thresholds can restrict the utility of aggregated outputs, particularly when analysing rare conditions or narrowly defined phenotypes. Recent work in public health surveillance [31] demonstrates that more nuanced, context-aware methods can preserve privacy while reducing unnecessary suppression. Incorporating similar risk-based approaches in future iterations of WiNGS may therefore enhance analytical flexibility while

remaining consistent with GDPR principles of proportionality and data minimization.

Unlike the research exemption, the GDPR statistical exemption is not restricted to research contexts, as long as data is processed exclusively for statistical purposes and aggregated in a way that prevents the (re-) identification of individuals. (Art. 9(2)(j), 89(1), Recital 162, *GDPR-Info.eu*). This means that statistical results should describe populations rather than individuals. For example, a mean, standard deviation, or predictive model trained on aggregated data qualifies as a statistical result, while a risk prediction score for a specific patient does not. Additionally, statistical data must never be used to support decisions about a particular individual, ensuring that the analysis remains purely exploratory rather than actionable at a personal level.

WiNGS was designed with these regulatory challenges in mind and integrates data minimization by default. Its federated model ensures that personal data never leaves its originating institution and only aggregated statistical outputs are shared across centers. Federated queries forwarded to individual institutions return summary counts rather than raw data, and these counts are further processed at the central hub to remove any identifying information linked to individual centers. While enabling institutions to perform cross-institutional frequency-based analyses, more complex collaborations can require additional agreements between users. The statistical exemption thus provides a compliant framework for routine and limited data sharing, and institutions pursuing broader research projects can still opt to operate under the research exemption when necessary.

WiNGS is different from other federated methods for genomic variant analysis in multiple aspects. First, it is a no/low-code end-to-end method tailored to the broad field of clinical genomics. REST API based interaction provides a uniform entry point towards integration into multiple programming languages. Other federated genomic frameworks can be used to query federated VCF data using command-line tools. However, understanding the queries and optimizing queries for distributed databases can be challenging, as is the case for GMQL. Second, the Beacon initiative has gained more clinical traction through phenotypic filtering added in version 2. However, it still lacks fine-grained insights into variant quality and overall sample phenotypes, which are essential in discriminating between causative variants and artifacts. Finally, WiNGS not only offers the analysis of SNVs/indels but also structural variants. This makes it the first federated platform ready for the integration of long-read data analysis.

Although the WiNGS-API shows a clear value for collaborative genomic projects, extensions can be envisioned to improve the system further and potentially integrate it with the other approaches. We envision that implementing bidirectional integration of WiNGS and the GA4GH Beacon project would expand the accessible data significantly for both parties. Next, more advanced genotype–phenotype association statistics can be implemented on top of the statistics endpoints. For example, gene burden analysis could be used to identify the cumulative effect of multiple rare variants more accurately, while automated phenotypic matching of individuals could provide better stratification of the analysis cohorts. Finally, the federated analytics system can be applied to train variant classification models using federated machine learning techniques [32]. As WiNGS evolves, we also aim to improve deployment flexibility; Trusted Research Environments (TREs) were not initially considered due to challenges in deploying containerized services and automated workflows [33]. In addition, many TREs restrict data export and inbound/outbound networking, which can limit both the transfer of genomic data into a WiNGS-Client and the establishment of external federated connections. One solution would be to deploy the full WiNGS stack within a TRE. Such an approach would allow multiple centers operating exclusively in TREs to participate in secure, internal federated analyses.

A general concern in routine analysis is performance and throughput. WiNGS addresses the annotation burden for small variations in genome samples through the recycling of annotations from previously encountered variants. This allows overall import times of approximately 84 min per genome (see Additional file 3: Fig. S1), which can be further optimized through parallelization. Second, the MongoDB-based storage engine within the WiNGS-Clients ensures adequate performance for near-real-time querying of the data (see Additional file 4: Table S2). As such, our results demonstrate that leveraging access to thousands of anonymous samples, extending a private cohort of a few hundred samples, allowed the discrimination between potentially causative variants and technical artifacts, while being able to highlight known causal genotype/phenotype relations.

Beyond performance optimization and privacy safeguards, the use cases highlight WiNGS as a versatile platform for a range of diagnostic and research applications. Trio analysis, population-level variant frequency estimation, and evidence-based genotype–phenotype association testing, are recurrently needed features. Particularly in rare disease diagnostics, the ability to securely query across multi-institutional datasets facilitates the identification of recurrent mutations and supports variant reclassification, even when only a few cases exist per site. Moreover, the integration of structural variant workflows extends this utility even for more recent long-read sequencing datasets, positioning WiNGS as a ready-to-use platform for future genomics pipelines.

Conclusions

Rare genetic disorders are complex, and as our understanding of high-penetrance disease genes expands, the required population size to accurately correlate genotypes to phenotypes increases. Furthermore, the rapid technological advances in long-read sequencing highlight the need for platforms offering integrated analysis of small and large-scale variants. These needs, combined with strict regulatory requirements related to privacy, necessitate a shift toward federated data-sharing and analysis frameworks. The presented platform, WiNGS-API addresses these needs using an API-based approach well-suited for automation and integration into existing workflows. The platform can be used for in-depth international research projects, while hospitals and clinical sequencing centers can adopt it for single-site clinical applications, such as SNV and SV trio analysis, leveraging anonymous cross-client baseline data for data filtering.

Availability and requirements

Project name: WiNGS API.

Project home page: https://wings.esat.kuleuven.be/rest-api/api-docs/#/.

Use Case Demonstration: *Collab Link*.

Operating system(s): Platform independent.

Programming language: NodeJS, MongoDB.

Other requirements: not applicable.

License: Affero general public license version 3.

Any restrictions to use by non-academics: e.g. licence needed.

Abbreviations

WiNGS	Widely integrated next generation sequencing
API	Application programming interface
SNV	Single nucleotide variant
CNV	Copy number variants
SV	Structural variant
WGS	Whole genome sequencing
NDD	Neuro development disorder
GMQL	GenoMetric query language
GRIN	Genomic research and innovation network
GA4GH	Global alliance for genomics & health
GDPR	General data protection regulation
InDel	Insertions/Deletions
REST	Representational state transfer
JSON	JavaScript object notation
WiNGS-CI	WiNGS central infrastructure
WiNGS-ACL	WiNGS access control list
LA	Local administrator
PI	Principal investigator
VCF	Variant caller format
HE	Homomorphic Encryption
sMPC	Secure Multiparty computing
HIPAA	Health insurance portability and accountability act
JWT	JSON web token
MME	Matchmaker exchange

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-026-01627-9.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Supplementary Material 5.

Supplementary Material 6.

Supplementary Material 7.

Supplementary Material 8.

Supplementary Material 9.

Acknowledgements

We gratefully acknowledge the contributions of Amin Ardeshirdavani, Haleh Chizari and Nasim Shabani to the early development of this work. Figures 1, 2, 4, 5, and 6, along with the figures in Additional files 4 through 8, were created using BioRender.com.

Authors' contributions

Conceptualization: NS, BH, GV, JRV; Data curation: NS, BH, GV; Formal analysis: NS, BH; Funding acquisition: GV, YM, JRV; Investigation: NS, BH; Methodology: NS, BH; Project administration: GV, YM, JRV; Resources: ; Software: NS, BH; Supervision: GV, JRV, YM; Validation: NS, BH, GV, JRV; Visualization: NS, BH; Writing – original draft: NS, BH, GV; Writing – review and editing: NS, BH, GV, JRV.All authors read and approved the final manuscript.

Funding

This work has been made possible by ELIXIR Belgium which was funded by FWO/IRI grants I002819N and I000323N. KU Leuven grants, C1-C14/18/092 and C14/22/125 to JRV. BH was supported by a Collaborative Doctoral Partnership Agreement of the Joint Research Center F.7/KU LEUVEN B&G 35332.

Data availability

WiNGS project readme and repositories (v1.0): [34, 35].
WiNGS REST API: [36].
Public datasets SNV: Publicly available whole-genome sequencing data from the GIAB Ashkenazi trio (HG002, HG003, and HG004) are used in the SNV use case [22] (https:/sciwheel.com/work/citation? ids=1498352&pre=&suf=&sa=0).
Structural Variants: The majority of long-read sequencing variants are extracted from publicly available IrWGS from the 1000 Genomes Project [21] (https:/sciwheel.com/work/citation? ids=17025552&pre=&suf=&sa=0).
Controlled-access datasets: The clinical genomic datasets used in this study cannot be shared because we do not own or control the data and are not permitted to redistribute them. These datasets remain under the custody of the originating clinical institutions, and WiNGS enables federated analysis without transferring any raw data outside those environments.
Platform access: The web server can be accessed via the following link: https:/ /wings-platform.org/.
Platform demo user: Data from the presented use case is available from the WiNGS system as a public data set. Login to the platform using the mentioned credentials and generate an API token: - User name: [wings-demo-user@ wings.be](mailto: wings-demo-user@wings.be) - Password: WiNGS2020!
Use Case Demonstration: [Google Collab Link] (https:/colab.research.google.com/drive/1PtKk6Q8pSl-HSS1d5hocJhd3fPmqu3RP).

Declarations

Ethics approval and consent to participate

This study was approved by the Ethical committee research UZ/KU Leuven (S65304 – S65991) and informed consent was obtained from all participants to participate. This study conformed with the principles of the Helsinki

Declaration. Secondary use of WES/WGS data is approved by EC approvals number 13/1/12 and by approval for Project ID 7746 - EDGE 004350 from the Antwerp University Hospital Ethical Committee.

Consent for publication
Written informed consent was obtained from all participants for publication.

Competing interests
The authors declare no competing interests.

Received: 16 June 2025 / Accepted: 13 March 2026
Published online: 26 March 2026

References

1. Marshall CR, Bick D, Belmont JW, Taylor SL, Ashley E, Dimmock D, et al. The Medical Genome Initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. Genome Med. 2020;12:48. https://doi.org/10.1186/s13073-020-00748-z.
2. Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet Med. 2019;21:2413–21. https://doi.org/10.1038/s41436-019-0554-6.
3. Katsanis N. The oligogenic properties of Bardet-Biedl syndrome. Hum Mol Genet. 2004;R65–71. https://doi.org/10.1093/hmg/ddh092. 13 Spec 1.
4. Rahit KMTH, Tarailo-Graovac M. Genetic modifiers and rare mendelian disease. Genes. 2020;11. https://doi.org/10.3390/genes11030239.
5. Wilczewski CM, Obasohan J, Paschall JE, Zhang S, Singh S, Maxwell GL, et al. Genotype first: Clinical genomics research through a reverse phenotyping approach. Am J Hum Genet. 2023;110:3–12. https://doi.org/10.1016/j.ajhg.2022.12.004.
6. Canakoglu A, Pinoli P, Gulino A, Nanni L, Masseroli M, Ceri S. Federated sharing and processing of genomic datasets for tertiary data analysis. Brief Bioinf. 2021;22. https://doi.org/10.1093/bib/bbaa091.
7. Mandl KD, Glauser T, Krantz ID, Avillach P, Bartels A, Beggs AH, et al. The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. Genet Med. 2020;22:371–80. https://doi.org/10.1038/s41436-019-0646-3.
8. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat. 2015;36:915–21. https://doi.org/10.1002/humu.22858.
9. Rambla J, Baudis M, Ariosa R, Beck T, Fromont LA, Navarro A, et al. Beacon v2 and Beacon networks: A lingua franca for federated data discovery in biomedical genomics, and beyond. Hum Mutat. 2022;43:791–9. https://doi.org/10.1002/humu.24369.
10. Chizari H, Sattanathan N, Ardeshirdavani A, Shabani N, Huremagic B, Vermeesch JR, et al. WiNGS: Widely integrated NGS platform for federated genome analysis. BioRxiv. 2022. https://doi.org/10.1101/2022.06.23.497325.
11. Casaletto J, Bernier A, McDougall R, Cline MS. Annu Rev Genomics Hum Genet. 2023;24:347–68. https://doi.org/10.1146/annurev-genom-110122-084756. Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal Primer.
12. Schulz WL, Nelson BG, Felker DK, Durant TJS, Torres R. Evaluation of relational and NoSQL database architectures to manage genomic annotations. J Biomed Inf. 2016;64:288–95. https://doi.org/10.1016/j.jbi.2016.10.015.
13. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17:122. https://doi.org/10.1186/s13059-016-0974-4.
14. RNAcentral C. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res. 2021;49:D212–20. https://doi.org/10.1093/nar/gkaa921.
15. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94. https://doi.org/10.1093/nar/gky1016.
16. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics. 2018;34:3572–4. https://doi.org/10.1093/bioinformatics/bty304.
17. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–80. https://doi.org/10.1038/s41587-023-02024-y.
18. European Medicines Agency. DARWIN EU® - Age - specific incidence rates of RSV - related disease in Europe. 2024 [cited 30 Nov 2025]. Available: https://catalogues.ema.europa.eu/system/files/2024-06/DARWIN_EU_D2.2.4_Report_P2-C1-011_RSV_v2.2_Public.pdf
19. Thomas M, Mackes N, Preuss-Dodhy A, Wieland T, Bundschus M. Assessing privacy vulnerabilities in genetic data sets: scoping review. JMIR Bioinform Biotech. 2024;5:e54332. https://doi.org/10.2196/54332.
20. Saleem K, Cicek AE, Sav S. Beacon Reconstruction Attack: Reconstruction of genomes in genomic data-sharing beacons using summary statistics. Bioinformatics. 2025;41. https://doi.org/10.1093/bioinformatics/btaf273.
21. Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, et al. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. Genome Res. 2024;34:2061–73. https://doi.org/10.1101/gr.279273.124.
22. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:160025. https://doi.org/10.1038/sdata.2016.25.
23. El Chehadeh S, Kerstjens-Frederikse WS, Thevenon J, Kuentz P, Bruel A-L, Thauvin-Robinet C, et al. Dominant variants in the splicing factor PUF60 cause a recognizable syndrome with intellectual disability, heart defects and short stature. Eur J Hum Genet. 2017;25:43–51. https://doi.org/10.1038/ejhg.2016.133.
24. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun. 2021;12:5910. https://doi.org/10.1038/s41467-021-25972-y.
25. Kumar AV, Monica K, Mandadi K. Data Privacy Over Cloud Computing using Multi Party Computation. 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). IEEE; 2023. pp. 262–267. https://doi.org/10.1109/IDCIoT56793.2023.10053427
26. Cummings R, Desfontaines D, Evans D, Geambasu R, Huang Y, Jagielski M, et al. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. Harv Data Sci Rev. 2024;6. https://doi.org/10.1162/99608f92.d3197524.
27. Dwork C, Smith A, Steinke T, Ullman J. Exposed! A survey of attacks on private data. Annu Rev Stat Appl. 2017;4:61–84. https://doi.org/10.1146/annurev-statistics-060116-054123.
28. Shabani M, Marelli L. EMBO Rep. 2019;20. https://doi.org/10.15252/embr.201948316. Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation.
29. Shringarpure SS, Bustamante CD. Privacy Risks from Genomic Data-Sharing Beacons. Am J Hum Genet. 2015;97:631–46. https://doi.org/10.1016/j.ajhg.2015.09.010.
30. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4:e1000167. https://doi.org/10.1371/journal.pgen.1000167.
31. Wilkinson K, Green C, Nowicki D, Von Schindler C. Less than five is less than ideal: replacing the less than 5 cell size rule with a risk-based data disclosure protocol in a public health setting. Can J Public Health. 2020;111:761–5. https://doi.org/10.17269/s41997-020-00303-8.
32. Angulo E, Márquez J, Villanueva-Polanco R. Training of classification models via federated learning and homomorphic encryption. Sensors. 2023;23. https://doi.org/10.3390/s23041966.
33. Kavianpour S, Sutherland J, Mansouri-Benssassi E, Coull N, Jefferson E. Next-Generation Capabilities in Trusted Research Environments: Interview Study. J Med Internet Res. 2022;24:e33720. https://doi.org/10.2196/33720.
34. Vandeweyer G, Sattanathan N, Huremagic B. WiNGS project readme and repositories v1.0 Github. https://github.com/wings-public
35. Vandeweyer G, Sattanathan N, Huremagic B. WiNGS project readme and repositories (v1.0). Zenodo. https://doi.org/10.5281/zenodo.18480125.
36. Sattanathan N, Huremagic B. wings-public/wings_rest_api: WiNGS REST API (v1.0). Zenodo. https://doi.org/10.5281/zenodo.18480351.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.