

# Introduction to SODAR



**Mikko Nieminen**

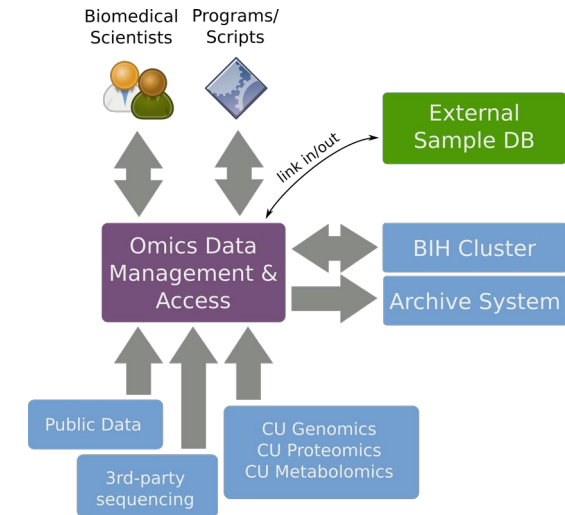
Senior Software Engineer, Core Unit Bioinformatics  
SODAR User Meeting, 2024-10-18

# Contents

1. Background
2. SODAR Features and Workflow
3. Deployment Status
4. Using SODAR
5. Related Software
6. Software Status and Roadmap
7. Resources

# Background

- **SODAR** (System for Omics Data Access and Retrieval) is specialized software for managing data in omics research projects
- **Feature Overview**
  - Project based access control and data encapsulation
  - Management of study design metadata
  - Providing access to large scale data storage
  - Management of file uploads
  - Linking files to metadata
  - Tools for aiding data management in research projects



The screenshot shows the SODAR Beta web interface. The top navigation bar includes 'SODAR Beta', a search bar, and links for 'Manual' and 'Help'. The breadcrumb trail is 'Home / BIH Projects / T-Cell CRG / T-Cell CRG WES/RNA-Seq Cohort'. The main content area is titled 'T-Cell CRG WES/RNA-Seq Cohort' and includes a 'ReadMe' section (no ReadMe is currently set), a 'Sample Sheets Overview' table, and a 'Landing Zones Overview' table.

Investigation/Study/Assay	Statistics	Links
Investigation	5248 files (10.1 TB)	> 📄 📄
Tcell2015	5248 files (10.1 TB)	> 📄 📄
Tcell2015 Exome Sequencing Nucleotide Sequencing	3852 files (7.8 TB)	> 📄 📄
Tcell2015 Transcription Profiling Nucleotide Sequencing	1396 files (2.4 TB)	> 📄 📄

Zone	Assay	Status Info	Status
------	-------	-------------	--------

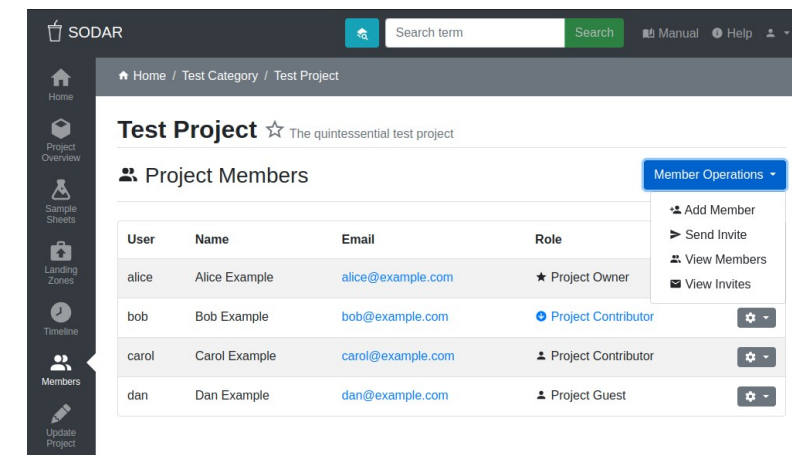
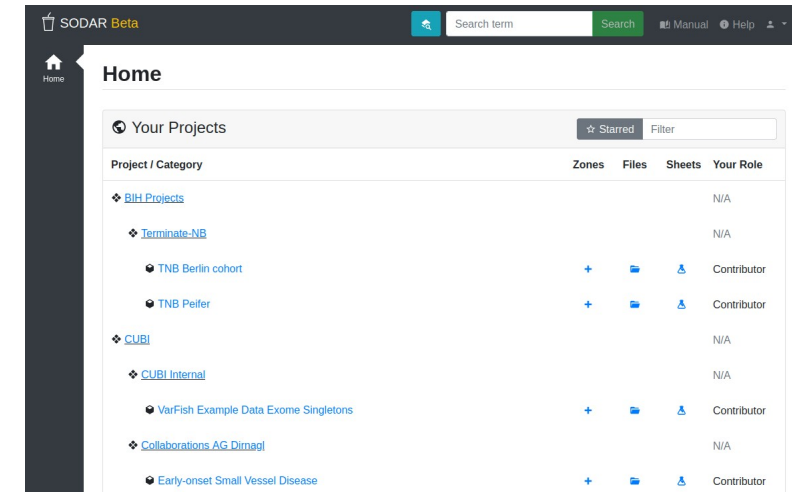
# SODAR Basics (1/2)

- **SODAR for the User**

- Web UI for user interaction in the browser
- REST APIs for scripts and software
- Access mass storage data by command line and web
- UUIDs and permanent URLs are provided for all relevant objects in the system
- Access with existing institute credentials, support for multiple organizations
- Integration with e.g. genome browsers

- **Projects**

- Data is organized in **projects** and **categories**
- Project-specific **roles** are assigned to users
- The same access control affects SODAR server and raw data in mass storage

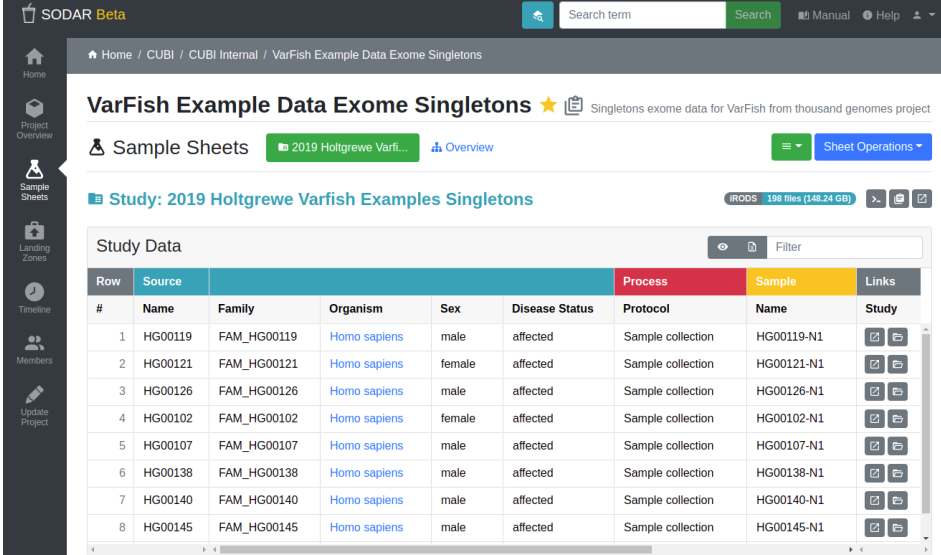


# SODAR Basics (2/2)

	Owner	Delegate	Contributor	Guest	Finder
Edit project metadata and settings	✓	✓			
Transfer project ownership	✓				
Modify project delegates	✓				
Modify other project members	✓	✓			
Modify project data	✓	✓	✓		
Read project data	✓	✓	✓	✓	
Create sub-categories and projects if role is given for category	✓	✓	✓		
Access nested categories and projects	✓	✓	✓	✓	
See nested categories and projects without access					✓

# Sample Sheets (1/3)

- **Sample sheets** contain the study metadata for each project
  - Modeled in the **ISA-Tab** format
    - <https://isa-tools.org>
  - ISA = Investigation, Study, Assay
  - Studies depict sample extraction from patients/donors
  - Assays represent processes and materials from samples in a study
- More about the ISA-Tab modeling in a separate session
- For uploading data, each project **must** have sample sheets



The screenshot shows the SODAR Beta web interface. The main content area displays a table titled 'Study Data' for the project '2019 Holtgrewe Varfish Examples Singletons'. The table has columns for Row, Source, Name, Family, Organism, Sex, Disease Status, Process, Sample, and Links. The data rows show sample information for 8 different individuals, all of whom are affected and from the same family (FAM\_HG00119 to FAM\_HG00145).

Row	Source	Name	Family	Organism	Sex	Disease Status	Process	Sample	Links
1	HG00119		FAM_HG00119	Homo sapiens	male	affected	Sample collection	HG00119-N1	[Icons]
2	HG00121		FAM_HG00121	Homo sapiens	female	affected	Sample collection	HG00121-N1	[Icons]
3	HG00126		FAM_HG00126	Homo sapiens	male	affected	Sample collection	HG00126-N1	[Icons]
4	HG00102		FAM_HG00102	Homo sapiens	female	affected	Sample collection	HG00102-N1	[Icons]
5	HG00107		FAM_HG00107	Homo sapiens	male	affected	Sample collection	HG00107-N1	[Icons]
6	HG00138		FAM_HG00138	Homo sapiens	male	affected	Sample collection	HG00138-N1	[Icons]
7	HG00140		FAM_HG00140	Homo sapiens	male	affected	Sample collection	HG00140-N1	[Icons]
8	HG00145		FAM_HG00145	Homo sapiens	male	affected	Sample collection	HG00145-N1	[Icons]

# Sample Sheets (2/3)

- Two ways to add sample sheets to SODAR projects
  - Creating sheets from templates
  - Uploading existing ISA-Tab TSV files
- **Creating from Template**
  - SODAR provides a selection of templates for different types of studies
  - Default values can be filled in a web-based form
- **Uploading ISA-Tab TSV Files**
  - Sample sheets can be created and edited elsewhere as ISA-Tab TSV files and uploaded into SODAR
  - Files are validated upon upload and must conform to the ISA specification

## Select ISA-Tab Template

Please select an ISA-Tab template for sample sheet creation. Alternatively, you can [import an existing ISA-Tab here](#).

Template

Mass cytometry ISA-tab template

Bulk RNA sequencing ISA-tab template  
Bulk RNA sequencing ISA-tab template from hiPSC for stem cell core projects  
Generic RNA sequencing ISA-tab template  
Germline DNA sequencing ISA-tab template  
MS Metabolomics Biocrates kit ISA-tab template  
**Mass cytometry ISA-tab template**  
Microarray ISA-tab template  
Single cell RNA sequencing ISA-tab template  
Single cell RNA sequencing ISA-tab template from hiPSC for stem cell core projects  
Somatic sequencing ISA-tab template

## Create from "Microarray ISA-tab template"

Directory Name\*

Example\_Project

Investigation directory name and assay prefix

investigation\_title\*

Investigation Title

sample\_names\*

alpha,beta,gamma

a\_measurement\_type\*

transcription profiling

# Sample Sheets (3/3)

- **Editing Sample Sheets**

- Editing cell values
- Configuring allowed values, value ranges, units etc.
- Adding/removing rows
- Ontology term lookup
- Version control for sheet changes
- Sheets can be replaced with new versions edited elsewhere

- **Exporting Sample Sheets**

- Can be exported in ISA-Tab compatible TSV files
- Simplified Excel export also available

Age At Sampling	Protocol
50 year	Sample collection
52 year	Sample collection
-	Sample collection
2 year	Sample collection
23 year	Sample collection

A001: Hpo Terms

Paste

limb muscle

HP

Sort by ontology

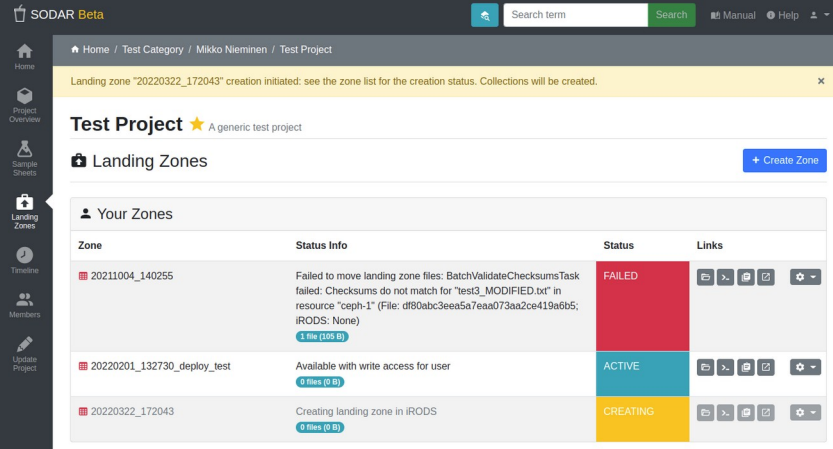
- [HP:0007156] Asymmetric limb muscle stiffness
- [HP:0009053] Distal lower limb muscle weakness
- [HP:0040267] Distal upper limb muscle hypertrophy
- [HP:0008959] Distal upper limb muscle weakness
- [HP:0030198] Fatigable weakness of distal limb muscles
- [HP:0030200] Fatigable weakness of proximal limb muscles
- [HP:0009055] Generalized limb muscle atrophy
- [HP:0009028] Generalized weakness of limb muscles

Name	Ontology	Accession	
Abnormal location of ears	HP	<a href="http://purl.obolibrary.org/obo/HP...">http://purl.obolibrary.org/obo/HP...</a>	↑ ↓ ✎ ✕
Mild expressive language delay	HP	<a href="http://purl.obolibrary.org/obo/HP...">http://purl.obolibrary.org/obo/HP...</a>	↑ ↓ ✎ ✕
obsolete Anaphylactoid purpura	HP	<a href="http://purl.obolibrary.org/obo/HP...">http://purl.obolibrary.org/obo/HP...</a>	↑ ↓ ✎ ✕



# Landing Zones (1/3)

- **Large scale study data** is stored in a distributed file system built on iRODS (<https://irods.org>)
- Uploading new sample data is done using **landing zones**
- **Landing Zone**
  - Temporary file area specific to project, user and assay
  - Directory in iRODS with access managed by SODAR
  - Sample sheets with at least one assay are required to create a landing zone in a project
  - User has full read/write access to their own zones
  - SODAR does not restrict the types of uploads, **any file is OK**
  - Every file must be accompanied with an .md5 file
  - Each landing zone enforces a directory structure depending on assay type or metadata



The screenshot displays the SODAR Beta web interface. At the top, there is a search bar and navigation links for 'Manual' and 'Help'. Below the search bar, a breadcrumb trail shows 'Home / Test Category / Mikko Nieminen / Test Project'. A yellow notification banner at the top states: 'Landing zone "20220322\_172043" creation initiated: see the zone list for the creation status. Collections will be created.' The main content area is titled 'Test Project' and includes a 'Create Zone' button. Below this, the 'Landing Zones' section is visible, showing a table of 'Your Zones'.

Zone	Status Info	Status	Links
20211004_140255	Failed to move landing zone files: BatchValidateChecksumsTask failed: Checksums do not match for "test3_MODIFIED.txt" in resource "ceph-1" (File: df80abc3eea5a7eaa073aa2ce419a6b5; iRODS: None) <a href="#">1 file (105 B)</a>	FAILED	[Icons]
20220201_132730_deploy_test	Available with write access for user <a href="#">0 files (0 B)</a>	ACTIVE	[Icons]
20220322_172043	Creating landing zone in iRODS <a href="#">0 files (0 B)</a>	CREATING	[Icons]

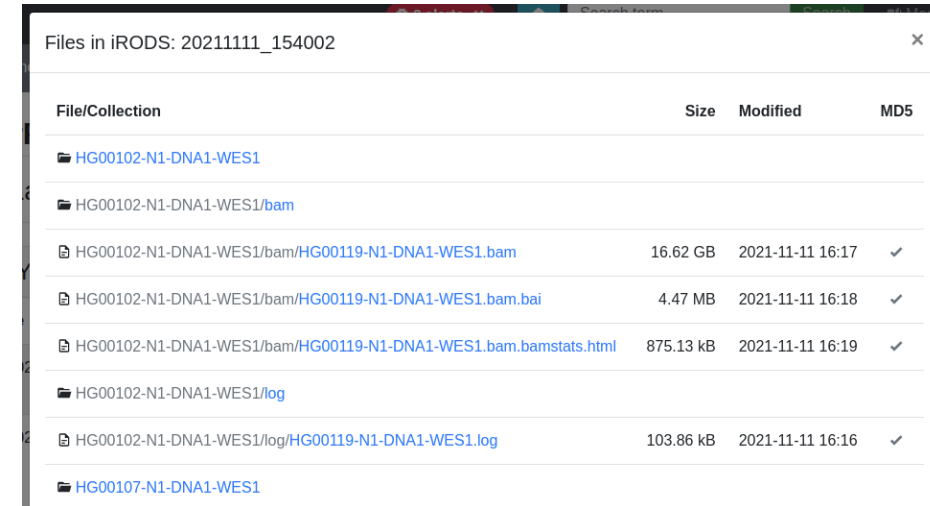
# Landing Zones (2/3)

- **Uploading Data in iRODS**

- Typically through command line from terminal, using the iRODS **iCommands** package
  - Commands similar to Unix shell, e.g. „icp“, „irsync“
  - SODAR auto-generates a user-specific client environment file for accessing iRODS

- **Other iRODS Access Options**

- CUBI Toolkit command line package for generic operations
  - Python iRODS client or the new iBridges wrapper
  - Mounting iRODS as a network drive via WebDAV
- More about iRODS access and file uploads in the feedback & tutorial session



The screenshot shows a web interface for iRODS with the title "Files in iRODS: 20211111\_154002". It displays a table of files and collections. The table has columns for "File/Collection", "Size", "Modified", and "MD5".

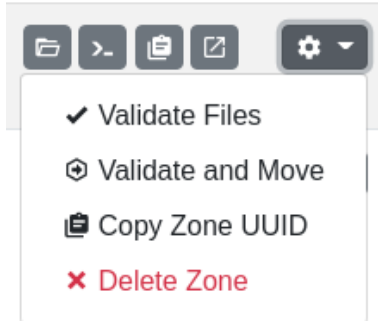
File/Collection	Size	Modified	MD5
📁 HG00102-N1-DNA1-WES1			
📁 HG00102-N1-DNA1-WES1/bam			
📄 HG00102-N1-DNA1-WES1/bam/HG00119-N1-DNA1-WES1.bam	16.62 GB	2021-11-11 16:17	✓
📄 HG00102-N1-DNA1-WES1/bam/HG00119-N1-DNA1-WES1.bam.bai	4.47 MB	2021-11-11 16:18	✓
📄 HG00102-N1-DNA1-WES1/bam/HG00119-N1-DNA1-WES1.bam.bamstats.html	875.13 kB	2021-11-11 16:19	✓
📁 HG00102-N1-DNA1-WES1/log			
📄 HG00102-N1-DNA1-WES1/log/HG00119-N1-DNA1-WES1.log	103.86 kB	2021-11-11 16:16	✓
📁 HG00107-N1-DNA1-WES1			

# Landing Zones (3/3)

## • Validation of Uploads

- Once uploading is completed, user starts validation in SODAR
  - Checksum validation against uploaded .md5 files
  - Checking against overwriting of existing files
  - (Possibility to add other types of validation in the future)
  - User is informed by SODAR once the operation completes
- Once validation succeeds, **SODAR moves the files into the read-only sample data repository**
  - There the files can be read by project members via the sample sheets
  - The landing zone is removed and a new one must be created for further uploads
- If validation fails, the operation is rolled back with an error

### Links



Successfully moved 2 files, landing zone removed

[Browse files in sample sheet](#)

MOVED

Failed to move landing zone files: BatchMoveDataObjectsTask failed: CAT\_NAME\_EXISTS\_AS\_DATAOBJ (Target file already exists: /omicsZone/projects/f6/f6641bb5-6dc2-41c9-8593-bd81a300eae9/sample\_data/study\_6d08d8dc-da1f-4776-8aea-28eaec379b40/assay\_0edb51f0-6936-4a6d-a08a-351e27453553/MiscFiles/pdf-sample.pdf)

1 file (7.9 kB)

FAILED











# File Access (1/2)

- **Browsing Files in Sample Sheets**





- Links to directories and files are generated in the sample sheets
- Logic for linking is based on the study and assay type or ISA-Tab metadata
- For studies, shortcuts are displayed for specific studies
- For assays, there can be one or more of the following:
  - Shortcuts for assay-level directories
  - Row-specific directory links
  - Inline file or directory links





- **Accessing Files in iRODS**


- Links in the samplesheets point to WebDAV, allowing access directly in the web browser
- Also provided are iRODS paths for iCommands and other iRODS tools to access files from your workstation or on the cluster

Sample	Links
Name	Study
HG00119-N1	 
HG00121-N1	 
HG00126-N1	 
HG00102-N1	 
HG00107-N1	 

### Assay Shortcuts

Results and Reports    





Misc Files    



This is a directory listing of the CUBI iRODS server.

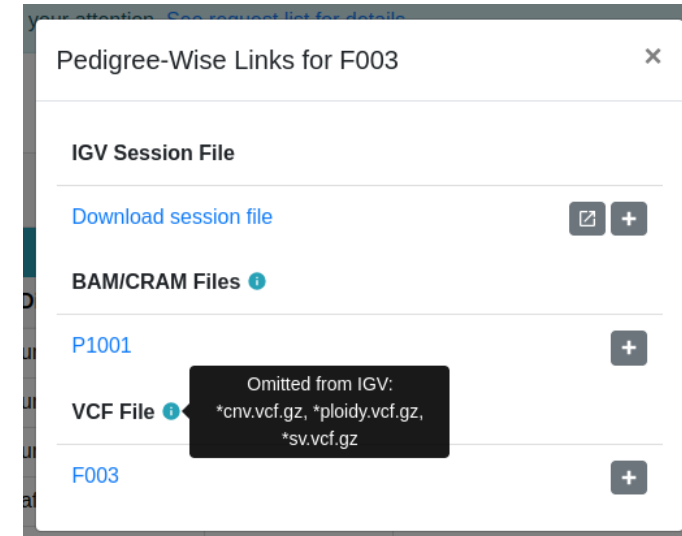
**Index of /sodarZone/projects/a1/a12164ff-9753-4d32-a520-7f499f862910/sample\_data/study\_34dcf3de-e410-4e53-b717-38f6bf6407f0/assay\_a435a0b1-9ef0-4fc7-b69f-faff954a260e/HG00119-N1-DNA1-WES1/raw\_data/2019-09-18/ on sodarZone**

Parent collection

Name	Size	Owner	Last modified
 SRR099967_1.fastq.gz	7.4G	holtgrem@CHARITE	2019-09-23 08:27
 SRR099967_1.fastq.gz.md5	55	holtgrem@CHARITE	2019-09-23 08:27
 SRR099967_2.fastq.gz	7.5G	holtgrem@CHARITE	2019-09-23 08:27
 SRR099967_2.fastq.gz.md5	55	holtgrem@CHARITE	2019-09-23 08:27

# File Access (2/2)

- **Automated Study Shortcut Generation**
  - Bam/VCF files for germline/cancer cases
- **IGV Integration**
  - Automated generation of session files linking to BAM/VCF files
  - Links for merging files with one click
- **UCSC Genome Browser Integration**
  - Track hub generation and management in iRODS
- **iRODS Access Tickets**
  - Read-only access tickets can be created for integration with other software
- **iRODS Delete Requests**
  - Used to delete files from the read-only sample repository
  - Owner or delegate must approve the request



## Test Project ★ The quintessential test project

### iRODS Access Tickets

Project Sheets [+ Create Ticket](#)

Name	Ticket	User	Created	Expires
hub1 / Track Hub Ticket	1o2426oyq5at6p17	alice	2023-08-07 16:43	Never
MiscFiles / 2023-08-07 16:07	jjho1b0r laxacdvv	alice	2023-08-07 16:07	Never
P1001-N1-DNA1-WGS1 / Example Ticket	jza2vk8qr6c9q7eg	alice	2023-08-07 16:06	2024-10-08

# Other Notable Features

- **Site-Wide Search**

- Supports searching for project metadata, files in iRODS, alternative names of patients and donors, etc.
- Displays results in projects the user can access

- **Timeline Application**

- Enhanced event logging for changes in project data
- Provides audit trails

- **Project Archiving**

- Finished projects can be made read-only

- **SODAR Core**

- Separate package of core components
- Allows development of external systems based on SODAR
- Enables project and access management between multiple system

The screenshot shows the SODAR Beta search interface. The search term 'HG00119' is entered in the search bar. The results are divided into two sections: 'Sources and Samples (1)' and 'Sample Files in iRODS (20)'. The 'Sources and Samples' section shows a single result for 'HG00119' as a 'Source' in the 'VarFish Example Data Exome Singletons' project. The 'Sample Files in iRODS' section lists 20 files, including 'bwa.gatk\_hc.HG00119-N1-DNA1-WES1.conda\_info.txt', 'bwa.gatk\_hc.HG00119-N1-DNA1-WES1.conda\_list.txt', 'bwa.gatk\_hc.HG00119-N1-DNA1-WES1.HG00119-N1-DNA1-...', 'bwa.gatk\_hc.HG00119-N1-DNA1-WES1.log', and 'bwa.gatk\_hc.HG00119-N1-DNA1-WES1.txt'. Each file is associated with the 'VarFish Example Data Exome Singletons' project and the '2019 Holtgrewe Varfish Examples Singletons...' assay. The interface includes navigation controls like 'Page', 'Filter', 'Prev', and 'Next'.

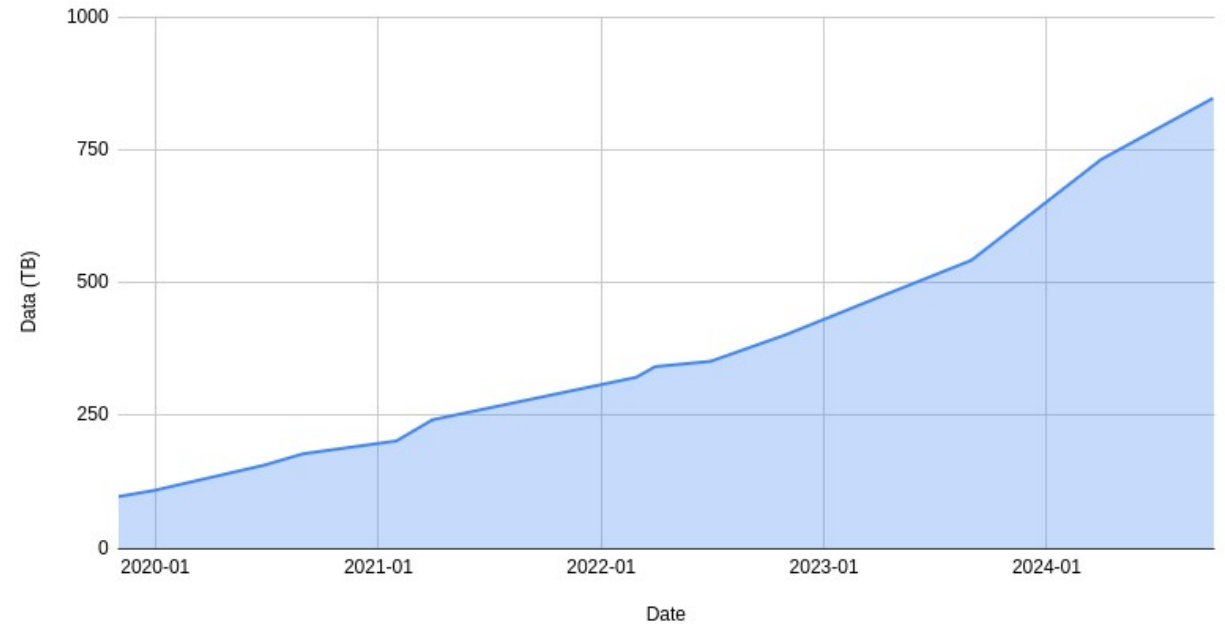
The screenshot shows the SODAR Beta project timeline for 'VarFish Example Data Exome Singletons'. The project is described as 'Singletons exome data for VarFish from thousand genomes project'. The timeline table lists events with their timestamps, users, descriptions, and statuses.

Timestamp	Event	User	Description	Status
2022-01-19 16:56:44	role_update	mikkopen_admin	update role to "project delegate" for mikkopen@CHARITE	OK
2022-01-19 15:09:08	project_star	mikkopen@CHARITE	star project	INFO
2022-01-19 14:27:06	role_create	mikkopen_admin	create role "project contributor" for mikkopen@CHARITE	OK
2022-01-19 14:18:06	project_star	mikkopen_admin	star project	INFO
2021-11-03 17:49:12	sheet_export	mriemin@MDC-BERLIN	export 2019 Holtgrewe VarFish Examples Singletons as ISA-Tab	OK
2021-11-02 11:52:17	project_star	mriemin@MDC-BERLIN	star project	INFO
2021-11-02 09:32:50	role_create	mikkopen_admin	create role "project contributor" for mriemin@MDC-BERLIN	OK
2020-07-17 15:19:40	role_create	holtgrem@CHARITE	create role "project guest" for cfsche@MDC-BERLIN	OK
2019-10-17 10:59:48	sheet_replace	holtgrem@CHARITE	replace previous investigation with 2019 Holtgrewe VarFish Examples Singletons	OK
2019-09-23 15:23:03	update_remote	stolpeo_admin	update remote access for site Varfish Staging to READ_ROLES	OK
2019-09-23 10:28:02	sheet_cache_update	holtgrem@CHARITE	update cache for project sheets	OK

# Deployment Status

- SODAR is in use for a large number of projects at BIH
- CUBI SODAR instance can be found at <https://sodar.bihealth.org>
- ~845TB of data stored in iRODS
- ~680 projects
- ~600 users

Project Data in SODAR  
(Very unscientific!)



# Using SODAR

- **Access the CUBI SODAR Instance**


- Point your web browser to <https://sodar.bihealth.org>
- The site is accessible from the Charité and MDC networks
- You can log in with your Charité or MDC credentials
- Firefox and Chromium-based browsers are supported
- Project access must be granted to you by an authorized user
- REST APIs accessible by generating an access token on the website

- **Access Data in iRODS**

- SODAR iRODS server is located in the same url ([sodar.bihealth.org:1247](https://sodar.bihealth.org:1247))
- Generate your personal client environment file on the website
- Use your preferred iRODS tools to upload/download iRODS data
- ..and/or use the WebDAV links via sample sheets in your browser

## Login

Please log in using your Charité or MDC account. Enter your user name as `username@CHARITE` or `username@MDC-BERLIN`.

 Login



# Related Software

- **Varfish**

- System for quality control, filtering, prioritization, analysis, and user-based annotation of DNA variant data with a focus on rare disease genetics
- Uses SODAR for project access management and file storage
- <https://www.cubi.bihealth.org/software/varfish/>
- <https://varfish.bihealth.org/>

- **Kiosc**

- System for running data exploration and visualization tools as containers
- Uses SODAR for project access management and file storage
- <https://kiosc.bihealth.org/>

- **SCelViz**

- Visualization of pre-processed single-cell data
- Used via Kiosc
- <https://www.cubi.bihealth.org/software/scelvis/>

- **Seapiper**

- Standard processing of bulk RNA-Seq or pseudo-bulk scRNA-Seq data combined with gene set enrichment analysis
- Used via Kiosc
- <https://www.cubi.bihealth.org/software/seapiper/>

- **CUBI Toolkit**

- Command-line tools for data delivery and access using SODAR, Varfish, Kiosc et al.
- <https://www.cubi.bihealth.org/software/cubi-tk/>

# Software Status and Roadmap

- **v0.15 (Current Release)**

- Generic assay plugin
- Isatemplates app for custom templates
- iRODS checksum verification stability improvements

- **v1.0 (Nov/Dec 2024)**

- iRODS v4.3 support
- **Single sign-on support via OIDC**
  - For e.g. LS AAI (formerly Elixir)
- Standardized REST API versioning
- Project access sync controls for owner/delegate
  - For Kiosc, Varfish etc.
- Admin alerts as email
- Opt out settings for email alerts
- A large number of critical backend updates

- **v1.1 (Spring 2025)**

- Improved project list
- Project deletion
- Landing zone access control updates
- Landing zone status display improvements
- User-facing updates from the backlog
- Minor sample sheet browser/editor improvements prior to rewrite

- **v1.2 (Autumn 2025)**

- Rewrite sample sheet browser/editor
- Improved search
- Etc. Etc. (TBA)

- Schedules and features subject to change

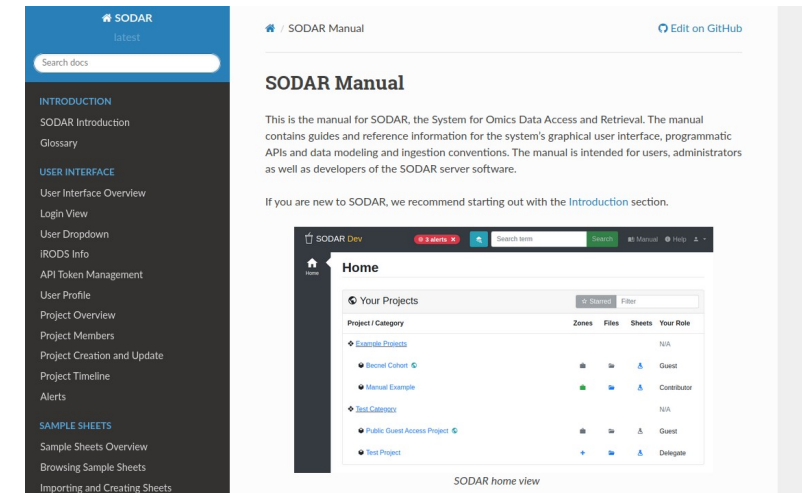
# Resources (1/2)

## • Web Resources

- SODAR manual: <https://sodar-server.readthedocs.io>
  - Or click the Manual link on the SODAR website
  - Manual also links to an introductory YouTube video
- SODAR software on the CUBI website: <https://www.cubi.bihealth.org/software/sodar/>
- CUBI SODAR resource: <https://www.cubi.bihealth.org/resources/sodar/>

## • Publication

- Mikko Nieminen, Oliver Stolpe, Mathias Kuhring, January Weiner, Patrick Pett, Dieter Beule, Manuel Holtgrewe, **SODAR: managing multiomics study data and metadata**, GigaScience, Volume 12, 2023, giad052, <https://doi.org/10.1093/gigascience/giad052>



## SODAR: managing multiomics study data and metadata

Mikko Nieminen<sup>1\*</sup>, Oliver Stolpe<sup>1</sup>, Mathias Kuhring<sup>1</sup>, January Weiner, III<sup>1</sup>, Patrick Pett<sup>1</sup>, Dieter Beule<sup>1†</sup> and Manuel Holtgrewe<sup>1†</sup>

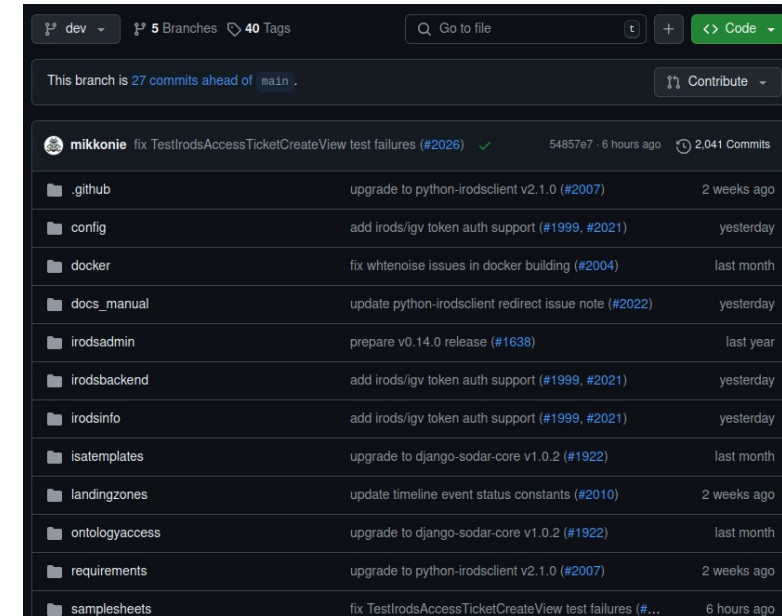
Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Core Unit Bioinformatics (CUBI), Berlin 10117, Germany  
\*Correspondence address: Mikko Nieminen, Berlin Institute of Health, Core Unit Bioinformatics, Chariteplatz 1, 10117 Berlin, Germany. E-mail: [mikko.nieminen@bih-charite.de](mailto:mikko.nieminen@bih-charite.de)  
†These authors contributed equally.

**Abstract**  
Scientists employing omics in life science studies face challenges such as the modeling of multiassay studies, recording of all relevant parameters, and managing many samples with their metadata. They must manage many large files that are the results of the assays or subsequent computation. Users with diverse backgrounds, ranging from computational scientists to wet-lab scientists, have dissimilar needs when it comes to data access, with programmatic interfaces being favored by the former and graphical ones by the latter. We introduce SODAR, the system for omics data access and retrieval. SODAR is a software package that addresses these challenges by providing a web-based graphical user interface for managing multiassay studies and describing them using the ISA (Investigation, Study, Assay) data model and the ISA-Tab file format. Data storage is handled using the iRODS data management system, which handles large quantities of files and substantial amounts of data. SODAR also offers programmable APIs and command-line access for metadata and file storage. SODAR supports complex omics integration studies and can be easily installed. The software is written in Python 3 and freely available at <https://github.com/bihealth/sodar-server> under the MIT license.

**Keywords:** scientific data management, ISA-Tab, iRODS

# Resources (2/2)

- SODAR source code, issue trackers and milestones are available at <https://github.com/bihealth/>
- **Get Involved**
  - Bug reports, feature requests and pull requests are welcome
  - Project and data specific issues should go to [cubi-helpdesk@bih-charite.de](mailto:cubi-helpdesk@bih-charite.de)
- **Repositories**
  - **sodar-server**
    - The Django server for the main SODAR system logic, GUI and REST APIs
  - **sodar-core**
    - Server framework with project-based access control, common UI and tools
    - Used to build SODAR, Varfish, Kiosc et al.
  - **sodar-docker-compose**
    - Docker-compose network for deploying, developing and trying out SODAR
  - **irods-docker**
    - Our SODAR-compatible iRODS Docker image



# Thank You

[www.cubi.bihealth.org](http://www.cubi.bihealth.org)

**BIH** Berlin Institute  
of Health  
@Charité