

The Swedish Reference Genome Portal:

A new service facilitating access and discovery of
genome data studied in Sweden

Launch Event

2024-11-08

Data Science Node in Evolution and Biodiversity (DSN-EB)

SciLifeLab Data Centre & NBIS

Uppsala University, Swedish Natural History Museum

Email: dsn-eb@scilifelab.se

Agenda



14:00 - 14:10	Welcome and Introduction Henrik Lantz, NBIS
14:10 - 14:40	Overview of the Swedish Reference Genome Portal <ul style="list-style-type: none">• Introduction and live Demo, Daniel Brink, SciLifeLab Data Centre• Technical implementation, Rory Crean, SciLifeLab Data Centre• Features that boost and facilitate researchers' work, Angela P. Fuentes-Pardo, SciLifeLab Data Centre
14:40 - 15:00	Q&A



Welcome and Introduction

Henrik Lantz, NBIS

The Data-Driven Life Science (DDLs) Program



Aims:

- To **develop and operate data services and resources** with high impact for data-driven life science in Sweden.
- To lead cutting edge **technology development** to strengthen Swedish competence in the area.
- SciLifeLab has appointed **SciLifeLab Data Centre** to organize the data support and databases operational area.

Vision and mission:

- Develop and coordinate **national data services and resources** for the corresponding research area
- Provide these services in coordination with the **SciLifeLab Data Platform** (<https://data.scilifelab.se>).



The Data-Driven Life Science (DDLs) Program



DDLs Fellows (incl.
PhDs, post docs)

39 DDLs Fellows
78 PhDs and 78 postdocs



PhDs and
Industry PhDs

140 PhDs in academia, 45 industry PhDs



Post docs and
Industry post docs

90 postdocs, 45 industry postdocs



WASP
WASP-HS

210 MSEK WASP
35 MSEK WASP- HS



WABI
(incl. Cryo-EM)

235 MSEK



Data support,
data bases

670 MSEK



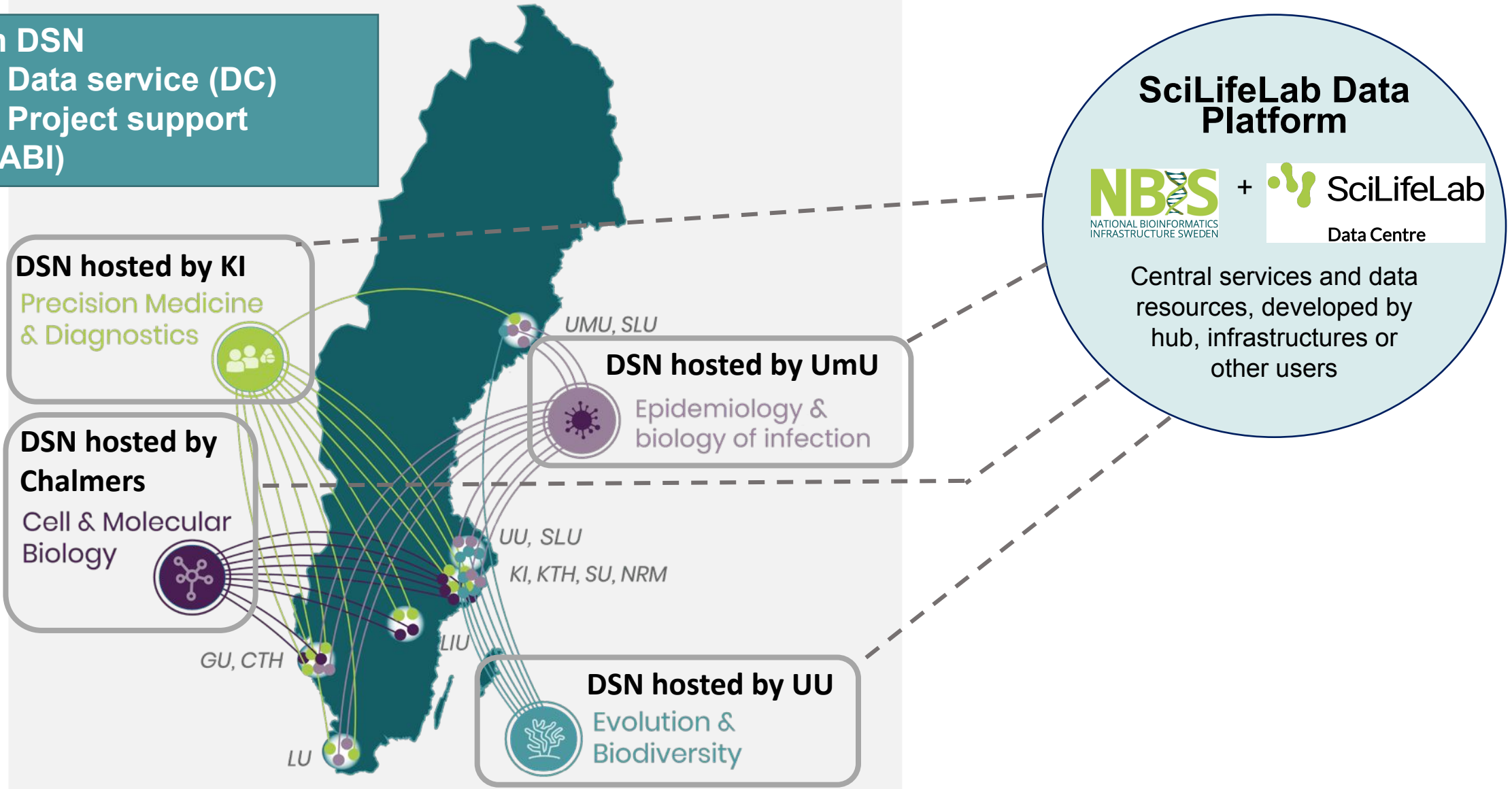
*Knut och Alice
Wallenbergs
Stiftelse*

DDL S Data Science Nodes (DSNs)



For each DSN

- 3 FTEs Data service (DC)
- 2 FTEs Project support (NBIS/WABI)



DDLs Data Science Node in Evolution and Biodiversity (DSN-EB)



Henrik Lantz
NBIS, UU
Scientific lead



Angela Fuentes
SciLifeLab DC, UU
Data steward,
Coordinator



Quentin Ågren
SciLifeLab DC, NRM
Systems developer



Rory Crean
SciLifeLab DC, UU
Data engineer



Daniel Brink
SciLifeLab DC, UU
Data steward



Hanna Kultima
SciLifeLab DC, UU
Vice-head of DC



Johan Rung
SciLifeLab DC, UU
Head of DC



Bengt Persson
NBIS, UU
Director



Cormac Kinsella
NBIS (WABI), NMR
Bioinformatician



Miguel Angel Redondo
NBIS (WABI), UU
Bioinformatician

The Genome Portal team



Henrik Lantz
NBIS, UU
Scientific lead



Angela Fuentes
SciLifeLab DC, UU
Data steward,
Coordinator



Quentin Ågren
SciLifeLab DC, NRM
Systems developer



Rory Crean
SciLifeLab DC, UU
Data engineer



Daniel Brink
SciLifeLab DC, UU
Data steward

Earth Biogenome Project



Genome portals collect and help visualise data

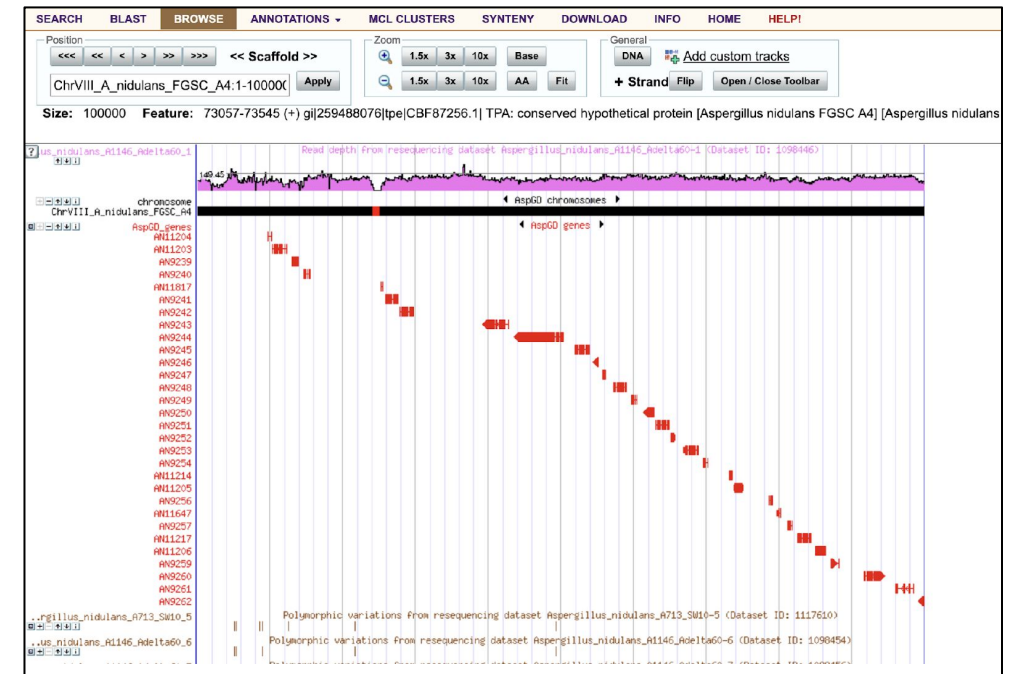


Typically consist of:

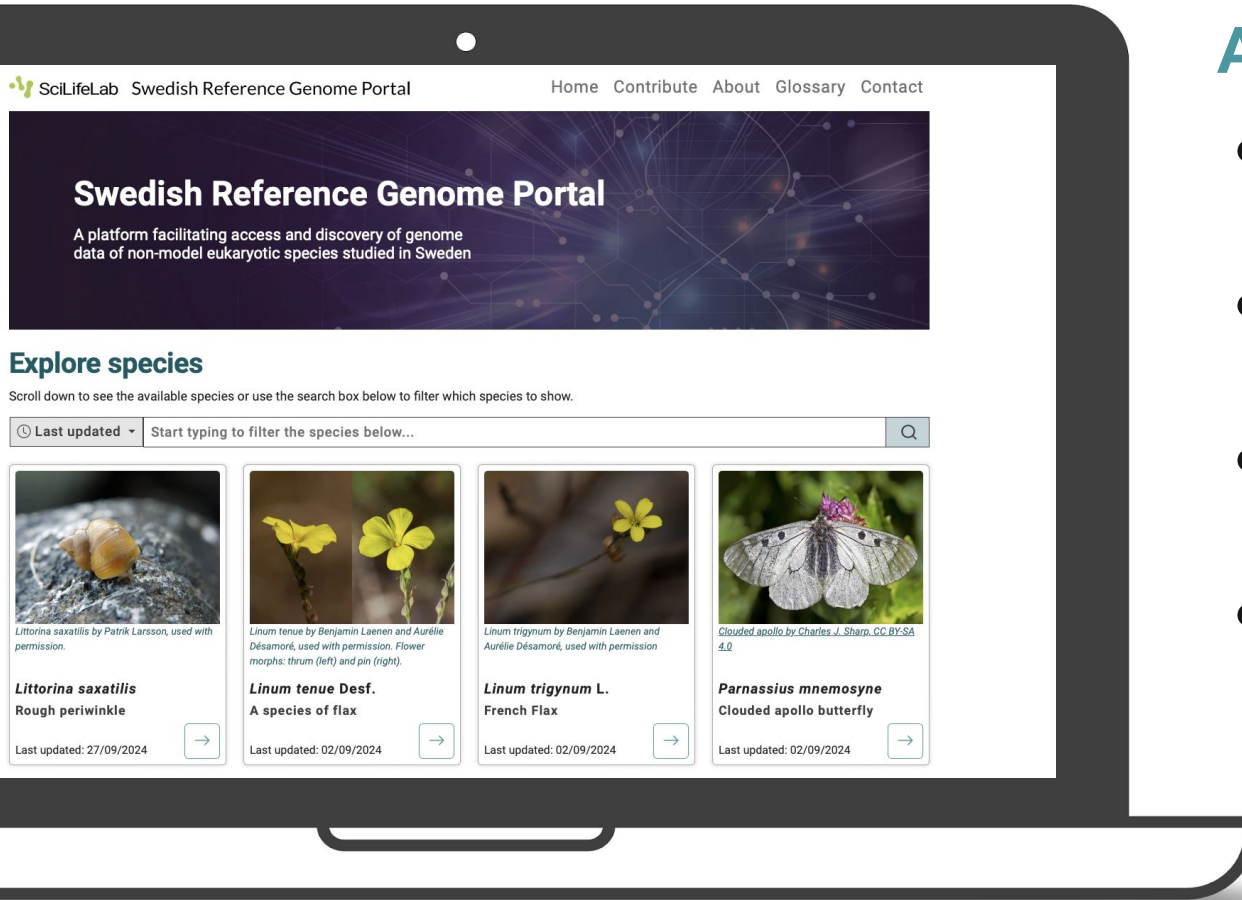
- A set of web pages with information and links for the species and its data
- A genome browser facilitates the visualisation of genome data

The screenshot shows the MycoCosm website interface. At the top, there's a navigation bar with links like JGI HOME, GENOME PORTAL, MYCOCOSM, PHYCOCOSM, and LOGIN. Below this, a search bar and a list of tabs (SEARCH, BLAST, BROWSE, ANNOTATIONS, MCL CLUSTERS, SYNTENY, DOWNLOAD, INFO, HOME, HELP) are visible. The main content area is titled 'Home • Aspergillus nidulans'. It features a microscopic image of the fungus on the left and text on the right stating that the genome was obtained from AspGD. A 'Genome Reference(s)' section provides a list of publications, including one by Arnaud MB et al. (2012) and another by Galagan JE et al. (2005). The footer contains contact information, a disclaimer, and copyright details for 1997-2024.

<https://genome.jgi.doe.gov/portal/>



The Swedish Reference Genome Portal



Aims:

- Highlight and **showcase genome research** performed in **Sweden**.
- Make it easier to access, visualise, and interpret genome data by **lowering the barriers to entry**.
- Promote **sharing of annotations of genomic features** that rarely get published.
- Ensure all data shown on the Genome Portal is aligned with the **FAIR principles** and available in public repositories.

Overview of the Swedish Reference Genome Portal

Introduction and live demo

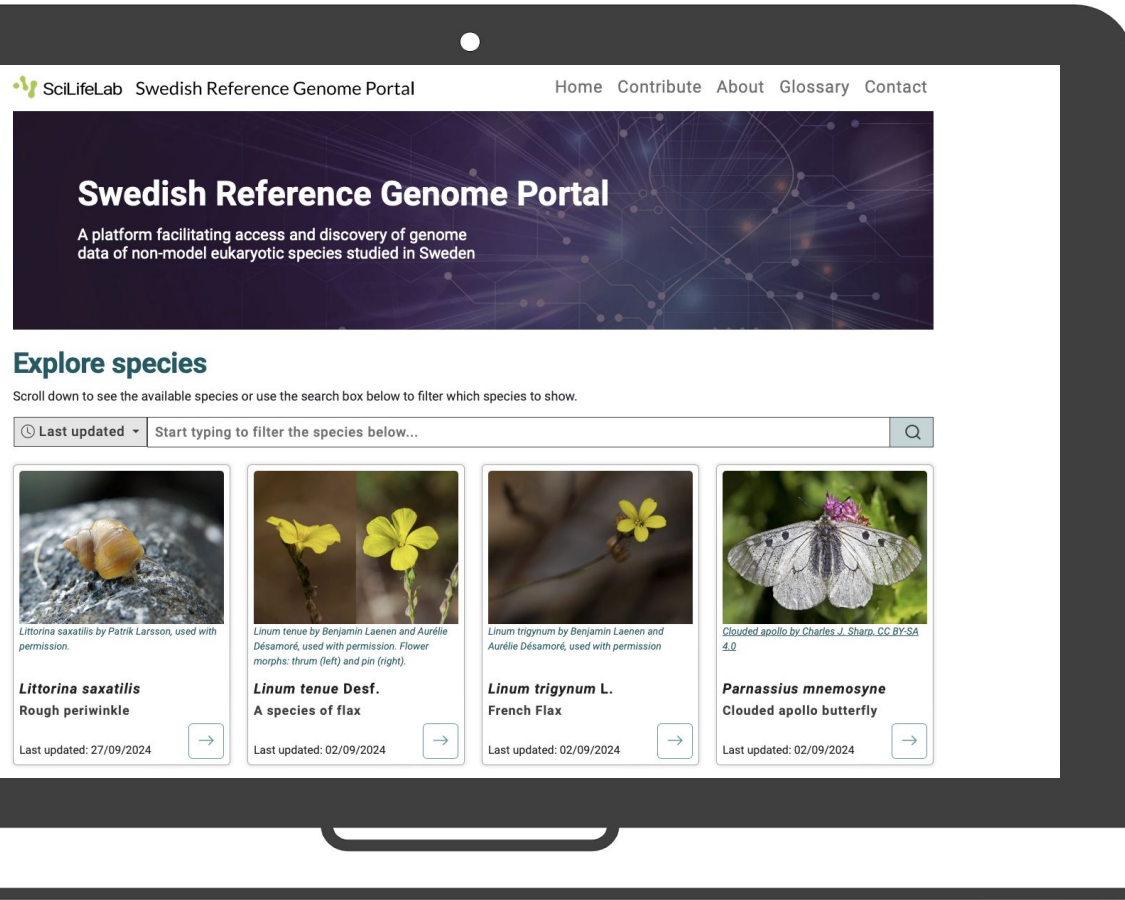
Daniel Brink, SciLifeLab Data Centre

The Swedish Reference Genome Portal



Scope:

- **Non-human eukaryotic species.**
- Data (co-)produced by researchers with affiliation to a **Swedish institution.**
- **Genome assembly (FASTA) and protein-coding genes (GFF)** needs to be available.
- **Data publicly accessible** in external repositories (e.g., ENA, NCBI).



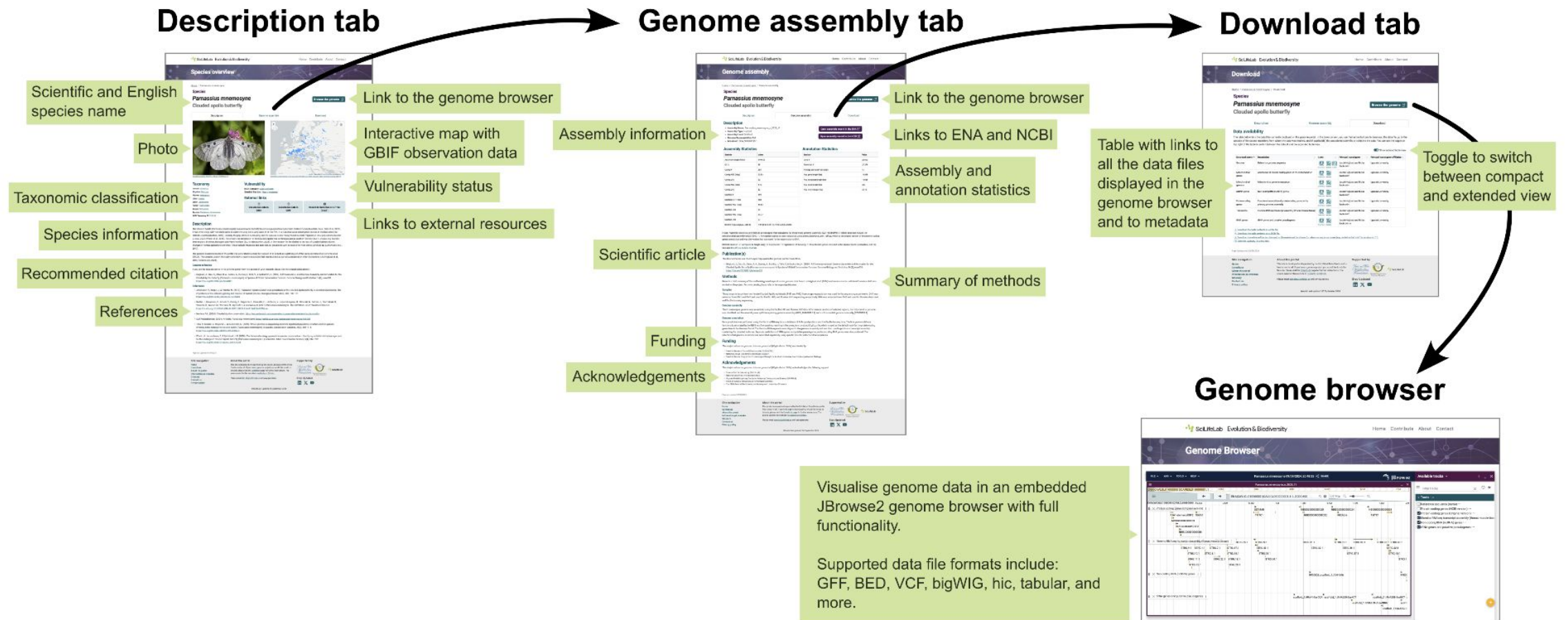
Have unpublished data? No problem!

We offer **FREE** support for submitting data to the SciLifeLab Data Repository



Web-design philosophy

Multi-species portal, with a common structure for each species.





The embedded genome browser: JBrowse 2

- Open source (<https://jbrowse.org/>)
- Designed to be embedded in websites or apps
 - also has a desktop client
- Highly customisable
- Powerful features

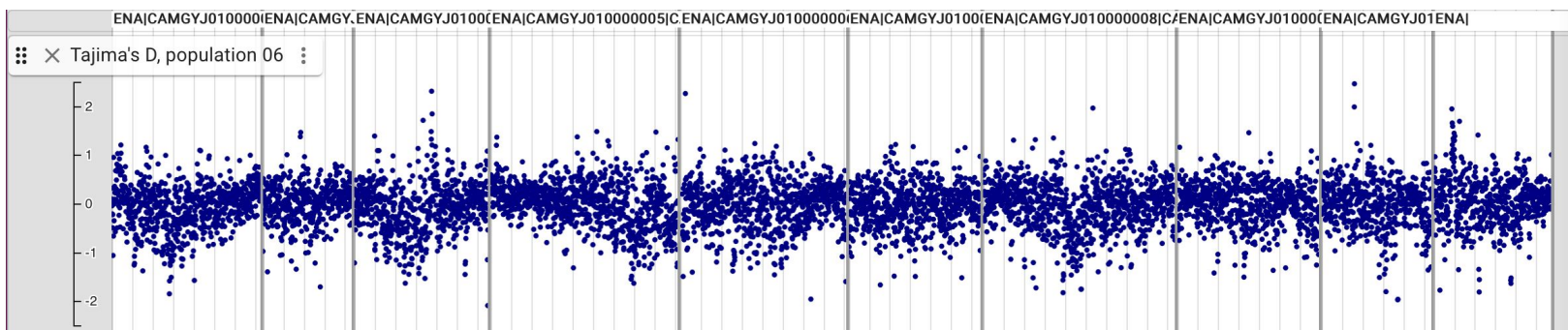


Image source: Diesh, C., Stevens, G.J., Xie, P. et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol* 24, 74 (2023).
<https://doi.org/10.1186/s13059-023-02914-z>



Supported file formats in the genome portal

- FASTA, GFF3, VCF, BED, BigBed, BigWig, PAF, 2bit, .hic
- BED-like tabular data, for instance for score values used in population genomics data (gwas)

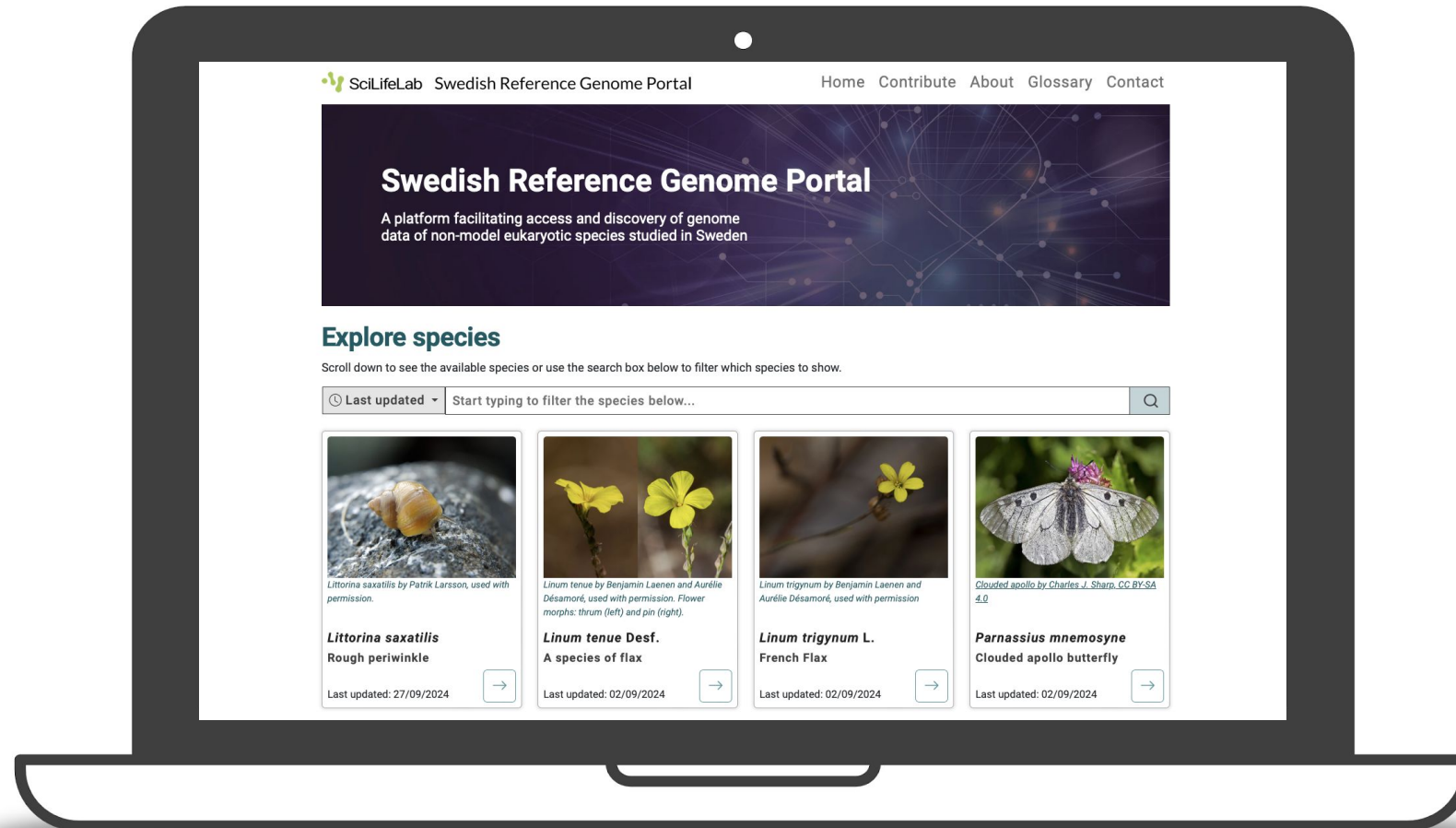


- BAM files are not supported at the moment, for performance reasons
 - However, BAM files *can* be loaded in the JBrowse web instance by the user
 - In the future, reduced BAM files are planned to be supported for visualisation of chromosomal inversions

Live demo



<https://genomes.scilifelab.se/>

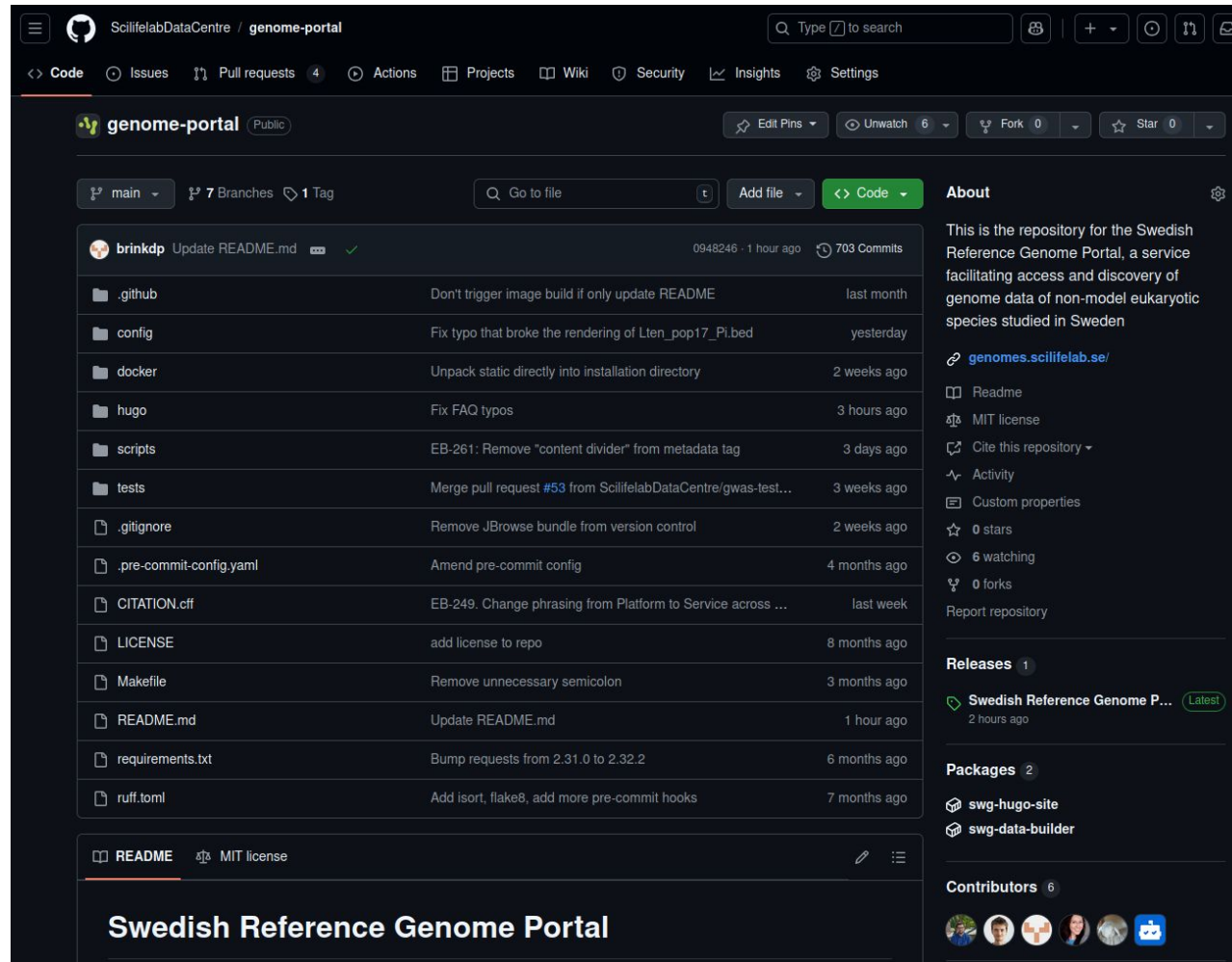


Overview of the Swedish Reference Genome Portal

Technical implementation

Rory Crean, SciLifeLab Data Centre

The source code is publically available on Github



Everything welcome!

- Suggestions, new features
- Bugs, typos, corrections
- Contributions

How to get in touch:

- Email us, dsn-eb@scilifelab.se
- Contact form on the website
- Directly through GitHub

<https://github.com/ScilifelabDataCentre/genome-portal>

The Genome Portal is built using a range of open source technologies



Website



Genome Browser



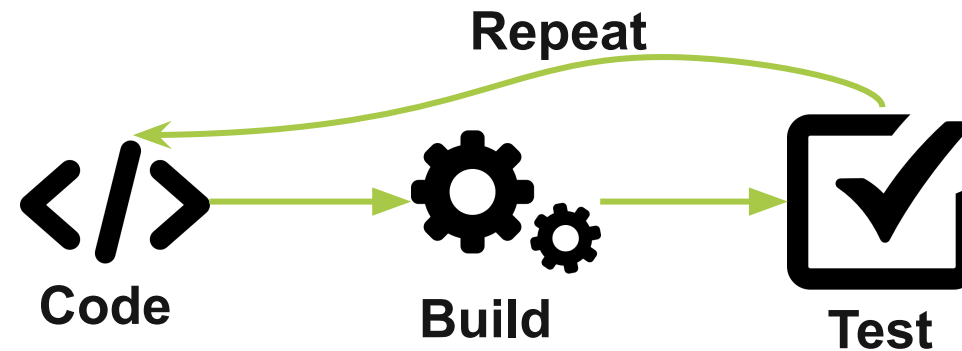
Deployment



Software development with large datasets



- Software development relies on rapid cycles of *continuously integrating* code changes.



- Re-preprocessing genomic datasets after every change would be slow.



But, we need to handle updating/making changes...

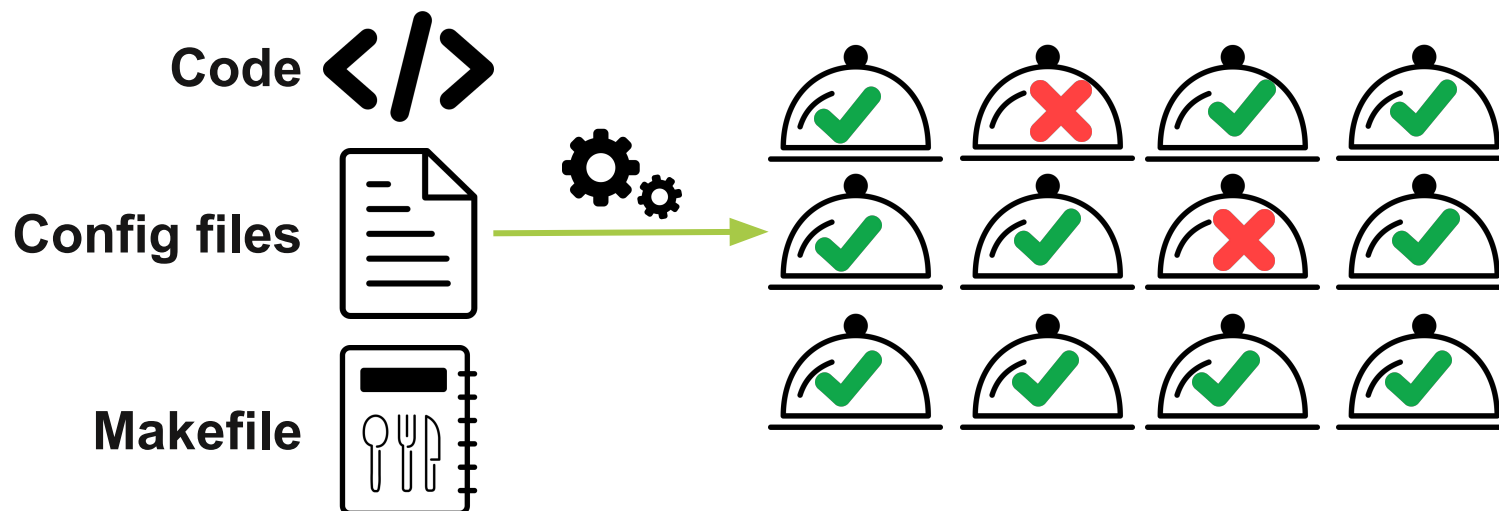


Northern krill case:

- ~5 GBs (compressed) download
- Many hours to pre-process



We use “Make” to avoid repetition

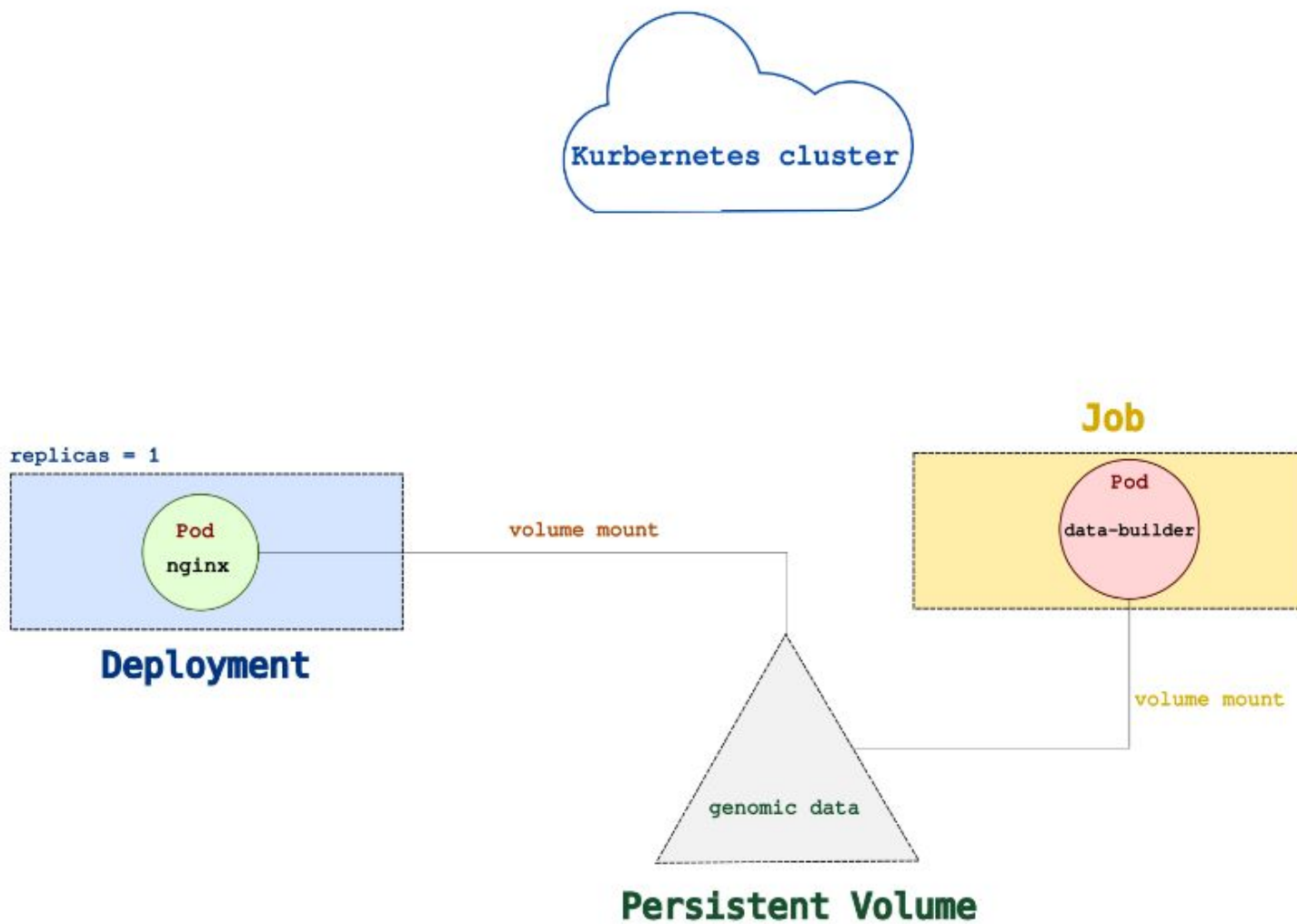


- If the meal (data) is already made, don't make it again.
- If the ingredients have changed, update the meal.

An example config file
(1 file for each species)

```
1 organism: "Linum trigynum"
2 assembly:
3   name: Ltrigynum_genome
4   displayName: "L. trigynum genome assembly GCA_964030455.1"
5   accession: GCA_964030455.1
6   url: "https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/964/030/455/GCA_964030455.1_Ltrigynum_genome_assembly.fasta.gz"
7   aliases: "https://raw.githubusercontent.com/ScilifelabDataCentre/genome-portal/main/scripts/linum_trigynum_genome_assembly.fasta.gz"
8 tracks:
9   - name: "Protein-coding genes"
10     url: "https://figshare.scilifelab.se/ndownloader/files/48853129"
11     fileName: "Ltrigynum_v1_genes.gff.gz"
12   - name: "Repeats"
13     url: "https://figshare.scilifelab.se/ndownloader/files/48879658"
14     fileName: "L_trigynum_v1_rep.bed.gz"
15   - name: Ltri_pop01_TD.bed"
16     url: "https://figshare.scilifelab.se/ndownloader/files/50074032"
17     fileName: "Ltri_pop01_TD.bed.gz"
18   addTrack: false
```

What does that look like in practice



Overview of the Swedish Reference Genome Portal

Features that boost and facilitate researchers' work

Angela P. Fuentes-Pardo, SciLifeLab Data Centre

Genome data visualisations play a key role in life science research



- Visualisations are essential for:
 - **Data interpretation**
 - **Hypothesis generation**
 - **Communicating discoveries**
- Examples of use cases:
 - Comparison of normal versus diseased individuals
 - Phenotype-genotype association studies

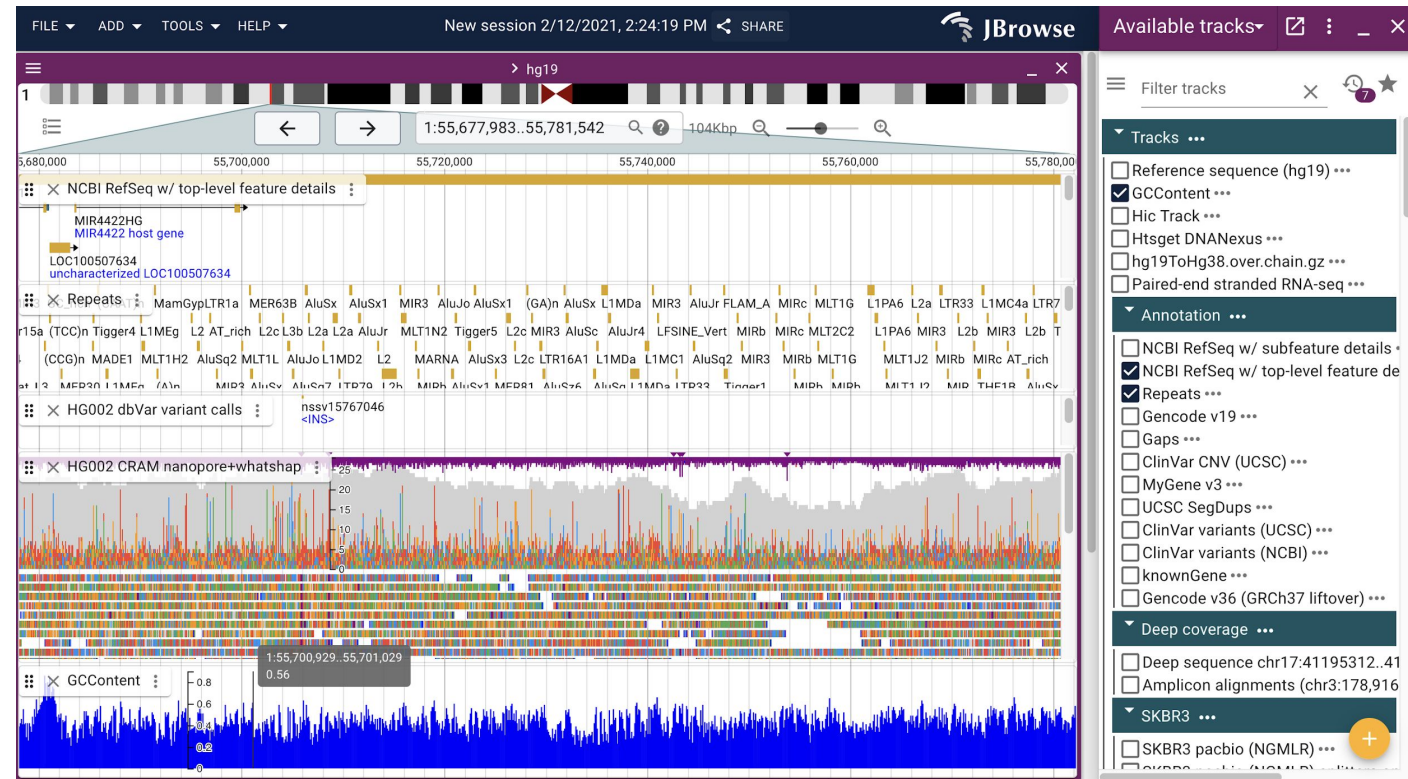
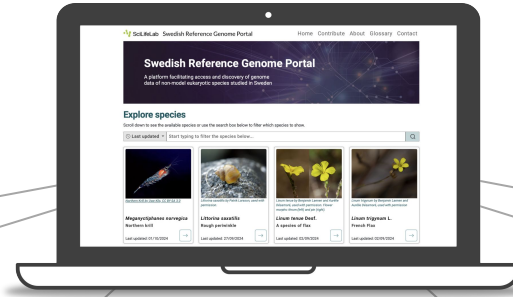


Image source: Diesh, C., Stevens, G.J., Xie, P. et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol* 24, 74 (2023).
<https://doi.org/10.1186/s13059-023-02914-z>

Benefits for researchers



Greater work exposure

Increased visibility of your work can lead to more citations and foster collaborations.



Unlock genome data visualisations

Accessible to everyone, regardless of technical expertise.

URL to your data displayed on the Genome Portal.

Add this URL in the Data Availability Statement of your manuscript.



Save yourself valuable time

No need to install software or download large files on your computer.

Short waiting time to have your genomic data displayed on the Genome Portal.



Publication-quality images

Export genomic visualisations as an SVG file.

Easily edit this vector file in any graphic design software, if needed.



FAIR data sharing

Publish genomic annotations.

Support with submissions to the SciLifeLab Data Repository.

Promote data reusability.



Enhanced teamwork

Share a session.

Bookmark genomic regions and easily share them with colleagues.

<https://genomes.scilifelab.se>

Advanced features of JBrowse 2



- HiC contact maps
- Quantitative tracks (e.g., depth of coverage for copy number variation, CNV, profiling)
- Structural Variant (SV) inspector view, examine breakpoint splits
- Variant widget that contains a table indicating the calls made in a multi-sample VCF
- And more!
- Features extended via plugins
- <https://jbrowse.org/jb2/>

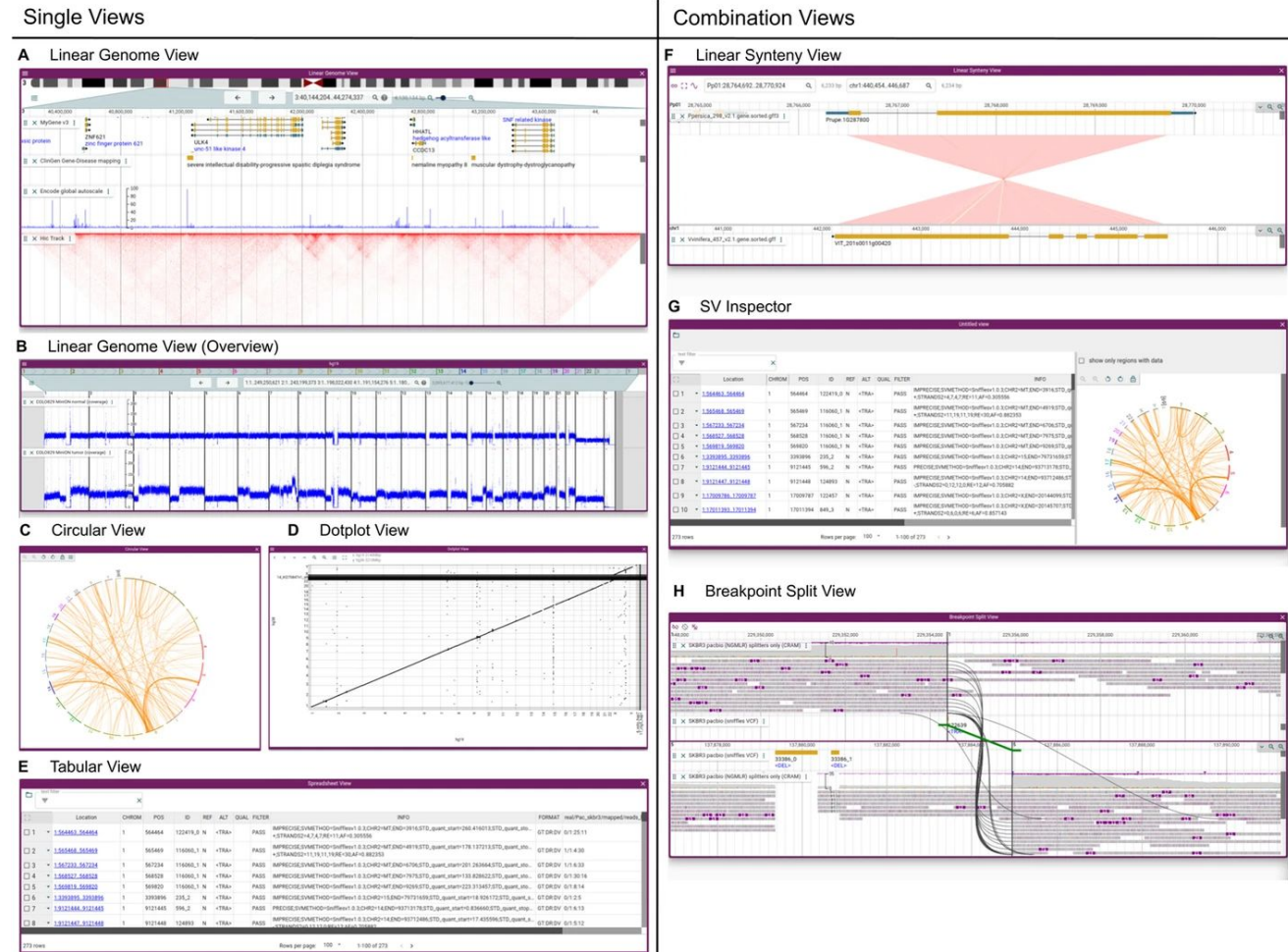


Image source: Diesh, C., Stevens, G.J., Xie, P. et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol* 24, 74 (2023).
<https://doi.org/10.1186/s13059-023-02914-z>

Interested? Adding your data to the portal is easy!



1

Express your interest

Send us an email to dsn-eb@scilifelab.se

2

Tell us more about your data

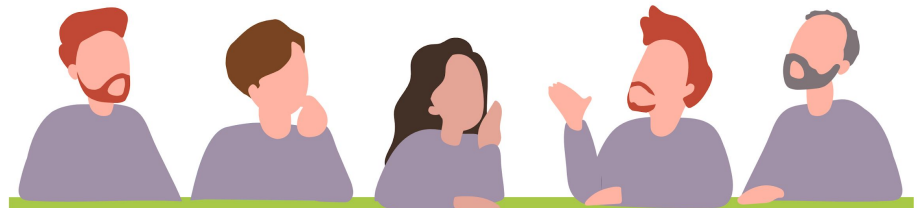
Help us find the data and its associated metadata by filling out a brief form.

3

Let us set up your data in the portal

Our developers and data stewards will upload the data to the portal for you.

Contact us



dsn-eb@scilifelab.se

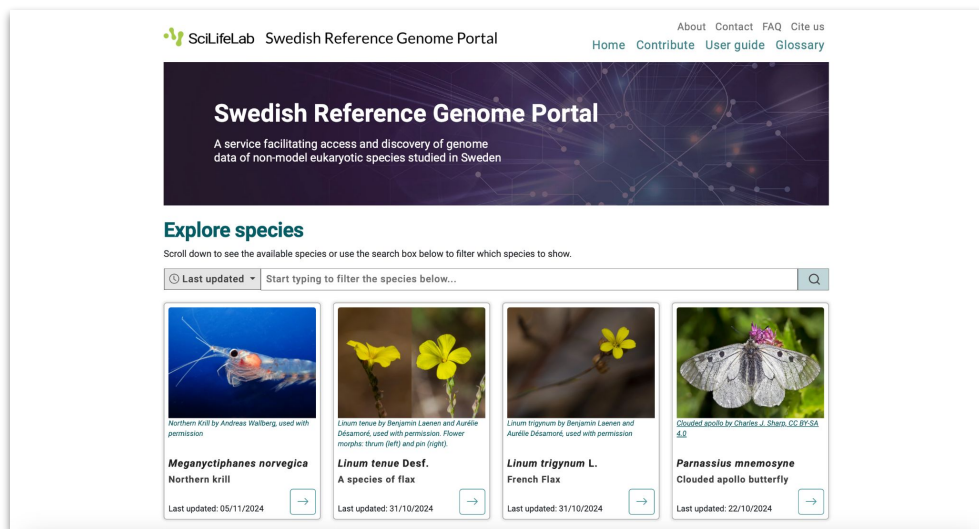
We are here to help you maximize
the impact of your data!



Visit our website! Two entry points



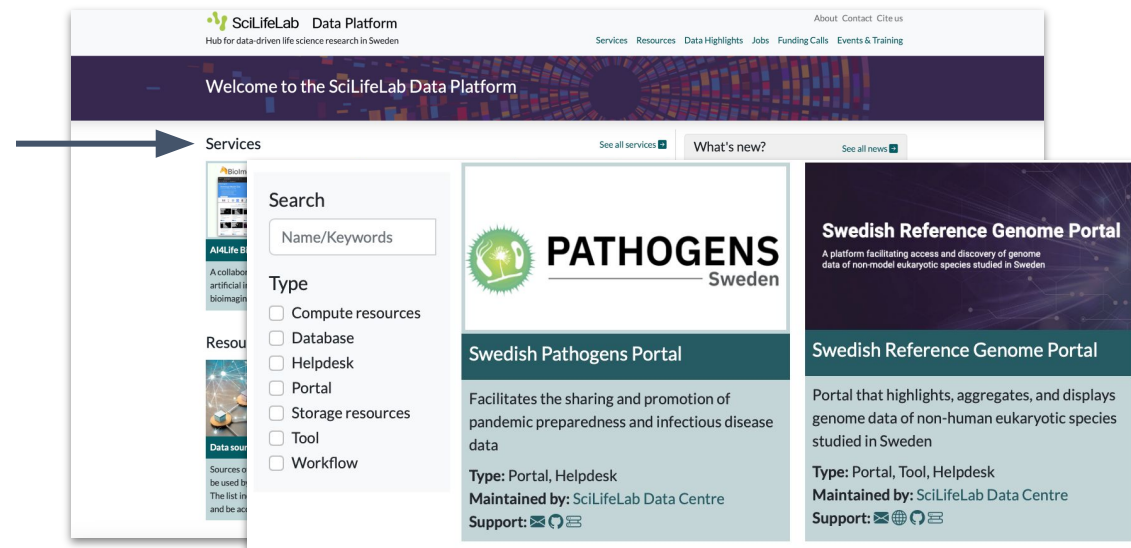
Direct URL



<https://genomes.scilifelab.se/>



SciLifeLab Data Platform



<https://data.scilifelab.se/>



Acknowledgements



Special thank you to all the researchers and PIs that participated in the pilot implementation of the Genome Portal!

- ***Linum tenue* Desf., and *Linum trigynum* L. (two species of flax)**
Tanja Slotte, Marco Fracassetti, and Zoé Postel, Stockholm University
- ***Parnassius mnemosyne* (Clouded Apollo butterfly)**
Jacob Höglund and Niclas Backström, Uppsala University
- ***Meganyctiphanes norvegica* (Northern krill)**
Andreas Wallberg, Uppsala University
- ***Skeletonema marinoi* (a diatom species)**
Mats Töpel, University of Gothenburg
- ***Littorina saxatilis* (Rough periwinkle)**
Kerstin Johannesson, University of Gothenburg

Thanks to the staff at NBIS who provided us with assistance with the datasets
Guilherme Dias, Tomas Larsson, and Stephan Nylinder

Thanks to our funders and host institutions.

Thank you for listening!



Data Centre



UPPSALA
UNIVERSITET



SWEDISH FOUNDATION for
STRATEGIC RESEARCH



Naturhistoriska
riksmuseet

Funders

Host institutions



Q&A session

You feedback is very important to us to keep improving our services.

What does make the Genome Portal unique, and different from existing genome browser/portal initiatives?



	Swedish Reference Genome Portal	Global initiatives (e.g., UCSC, Ensembl)
Target community	Affiliated to a Swedish research institution.	Global.
Taxonomic coverage	Non-human eukaryotic species.	Everything.
Data incorporation	Relies on researchers' submissions.	Automatic indexing.
Primary goal	Facilitate access, visualisation, and interpretation of genomic data. Focus on offering a powerful genome browser. Facilitates access and discovery by aggregating links to datasets associated with each genome assembly in a single page.	Own specific goals and features besides their genome browser capabilities, such as support for comparative genomics analyses.
Waiting time	Once minimum requirements are met, researchers have the possibility to decide what, when, and how their genomic data is displayed on the Genome Portal. Shorter processing times with the SRGP, as it is maintained by a local team.	While global genome portals also address researchers' inquiries, users can generally expect much longer waiting times as responses will depend on staff availability and priorities.
On genomics annotations	Displays unique genomic annotations (annotation tracks) that are rarely published.	Global genome portals often index and retrieve data from public genomic repositories such as NCBI or ENA, which means they may overlook the specific genomic annotations contributed by researchers in Sweden.