Empowering breeding decisions with BIOFLOW: a transformative open-source community tool for data-driven agriculture research from CGIAR

Lorena Guimaraes Batista – Quantitative Geneticist

Accelerated Breeding Initiative (ABI)

What is the BIOFLOW about?

- Provide access to a **user-friendly Breeding Analytics service**
- **Funded through the Crops to End Hunger Initiative** to pay salaries, traveling and capacity building to develop and maintain analytics.
- The first phase is focused in providing access to semi-automated **genetic evaluation** (predictive analytics).

CGIAR

INITIATIVE ON

Breeding Resources

What is the BIOFLOW about?

BIOFLOW is a **tool to support selection decisions** for recycling and advancement **in breeding programs at scale**

- We do not intend it to be a cutting-edge tool for scientific analysis
- Nevertheless, statistical models underlying its analysis should be robust enough to drive breeding decisions

What is the BIOFLOW about?

- BIOFLOW as a one-stop shop for analytics

- Harmonized crop reports on KPIs harvested from BIOFLOW output

- Institutionalizing data-driven decisions with Breeding Analytics rightfully taking a seat at the table where pipeline decisions are made.

Who is part of the initiative?

CGIAR | INITIATIVE ON Breeding Resources

CGIAR

CIP INTERNATIONAL POTATO CENTER

CIMMYT

Alliance Bioversity & CIAT

IITA Transforming African Agriculture

Africa Rice Center Centre du riz pour l'Afrique AfricaRice

ICARDA Science for resilient livelihoods in dry areas

CGIAR | INITIATIVE ON Accelerated Breeding

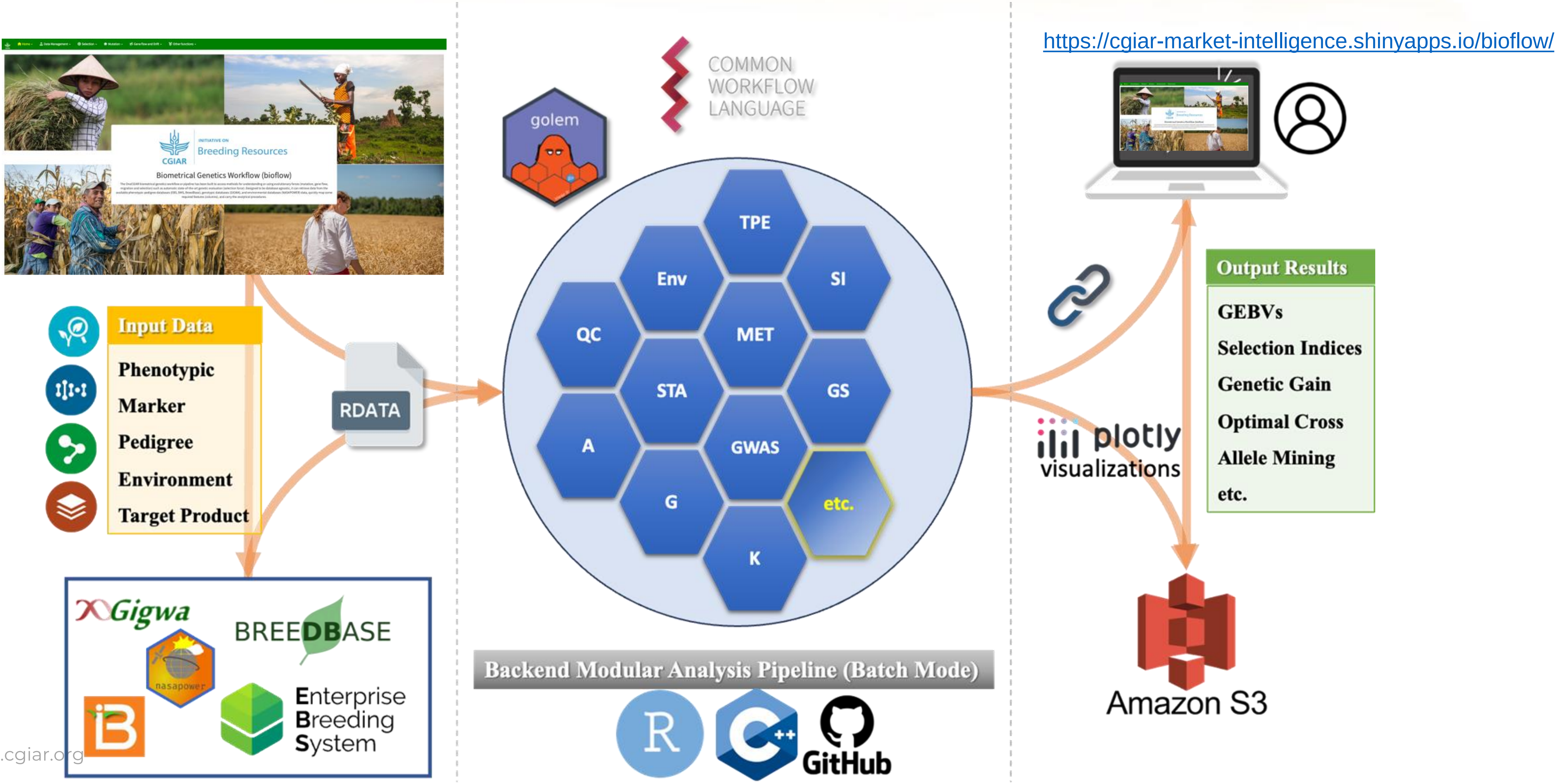
IRRI

Key wins

- Access to state-of-the-art **semi-automated breeding analytics** tools for OneCGIAR and NARES
- **Quick turnaround** of breeding decisions & decreased dependency on manual time-consuming requests
- Quick and easy access to **breeding data and metrics** to assess the performance of breeding programs
- Additional time for BA and QG experts to focus on new methods

The bioflow platform as an example of downstream analytics

https://cgiar-market-intelligence.shinyapps.io/bioflow/



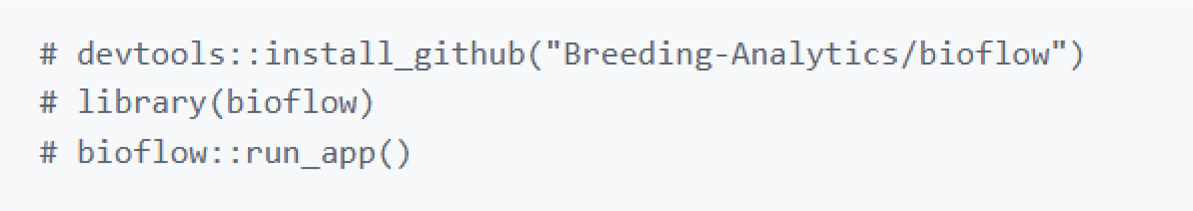
Input Data
- Phenotypic
- Marker
- Pedigree
- Environment
- Target Product

Backend Modular Analysis Pipeline (Batch Mode)

Output Results
- GEBVs
- Selection Indices
- Genetic Gain
- Optimal Cross
- Allele Mining
- etc.

Current available setups

- Bioflow Production Server: requires login (OneCGIAR and partners)
- Bioflow Guest Server: anyone can use (good to get to know the tool)
- Local installation:



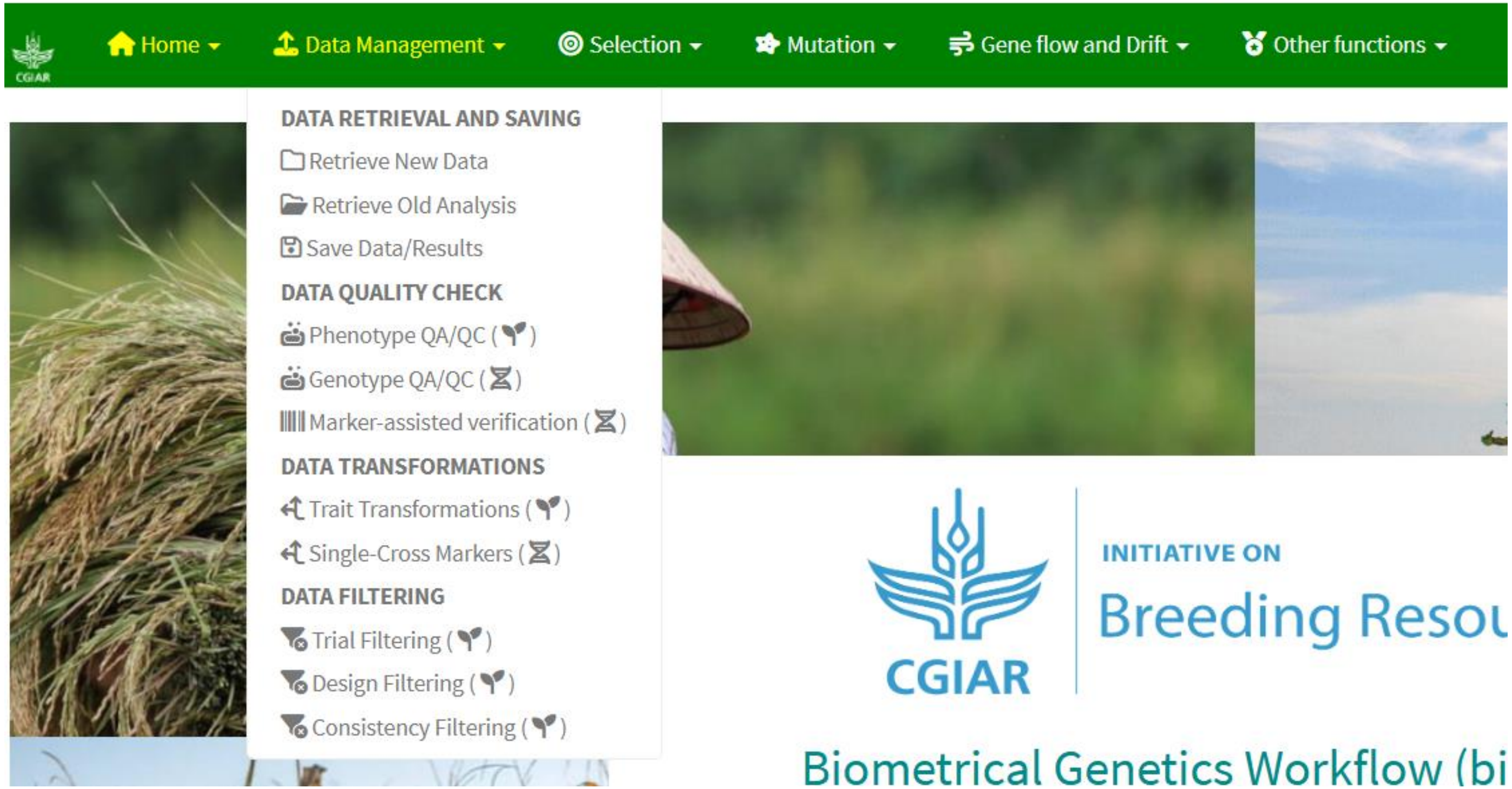
```
# devtools::install_github("Breeding-Analytics/bioflow")
# library(bioflow)
# bioflow::run_app()
```

Bioflow overall interface

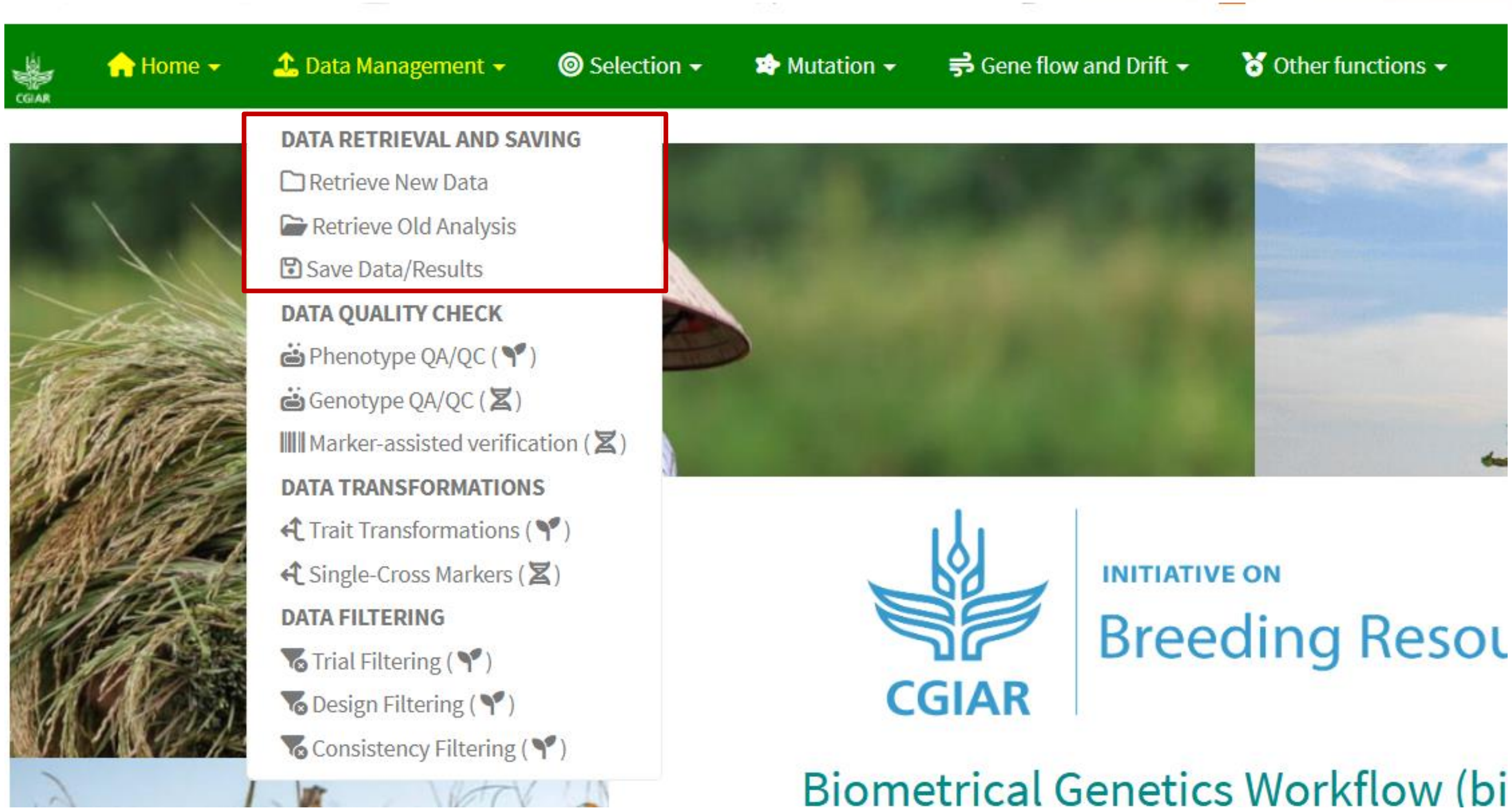
Biometrical Genetics Workflow (bioflow)

The OneCGIAR biometrical genetics workflow or pipeline has been built to access methods for understanding or using evolutionary forces (mutation, gene flow, migration and selection) such as automatic state-of-the-art genetic evaluation (selection force) in decision-making. Designed to be database agnostic, it can retrieve data from the available phenotypic-pedigree databases (EBS, BMS, BreedBase), genotypic databases (GIGWA), and environmental databases (NASAPOWER), and carry the analytical procedures.

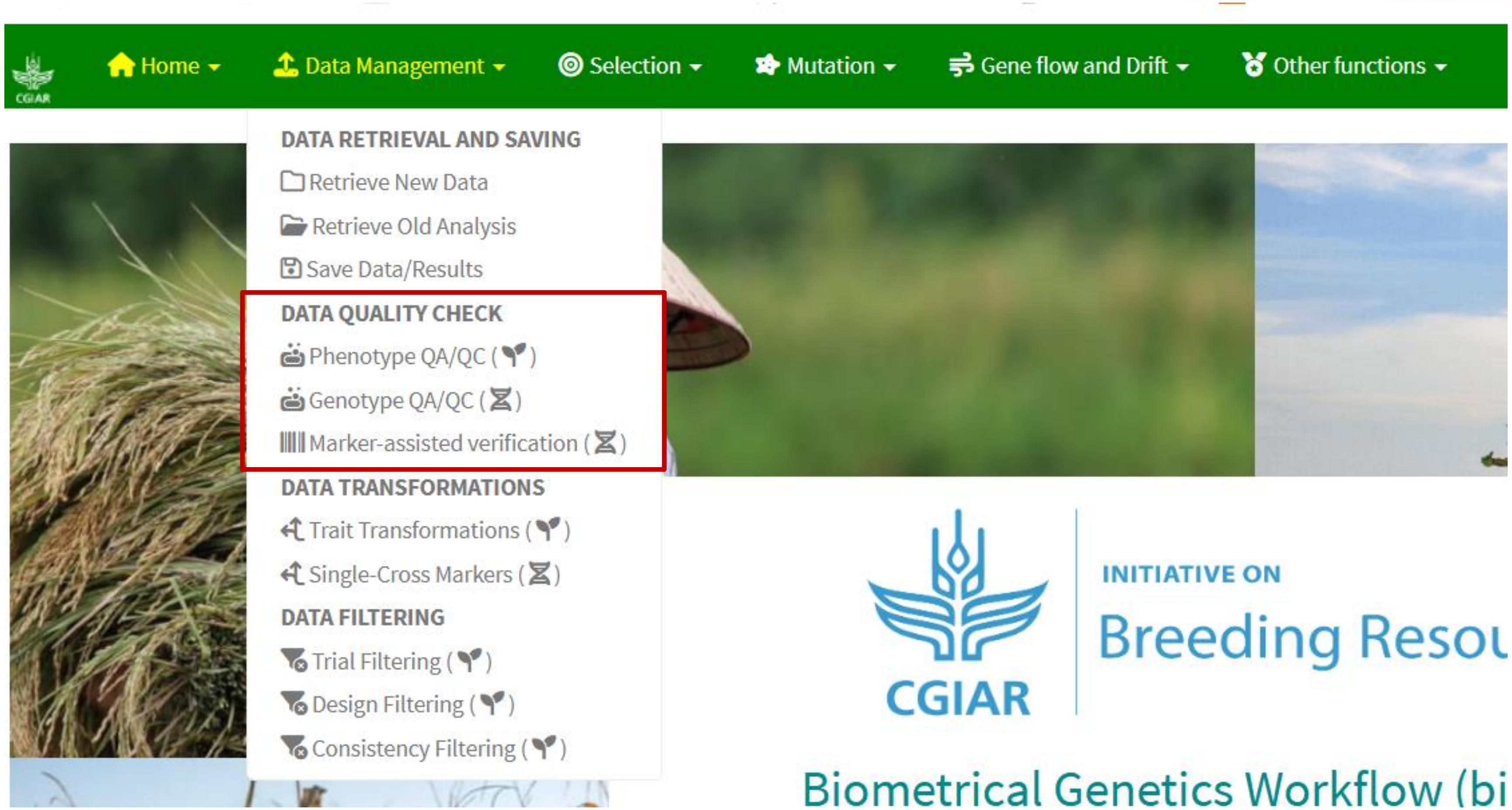
Bioflow overall interface



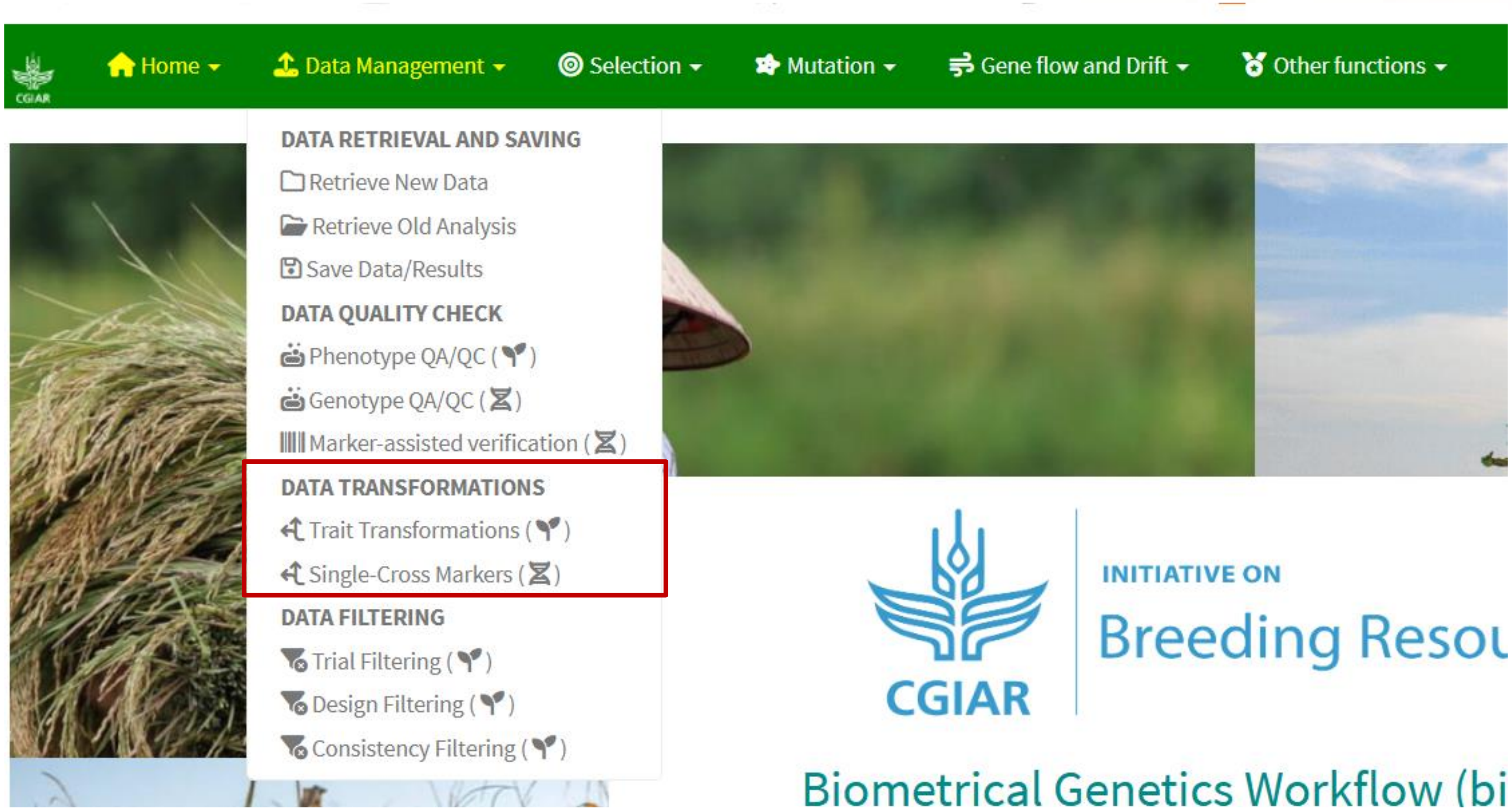
Bioflow overall interface



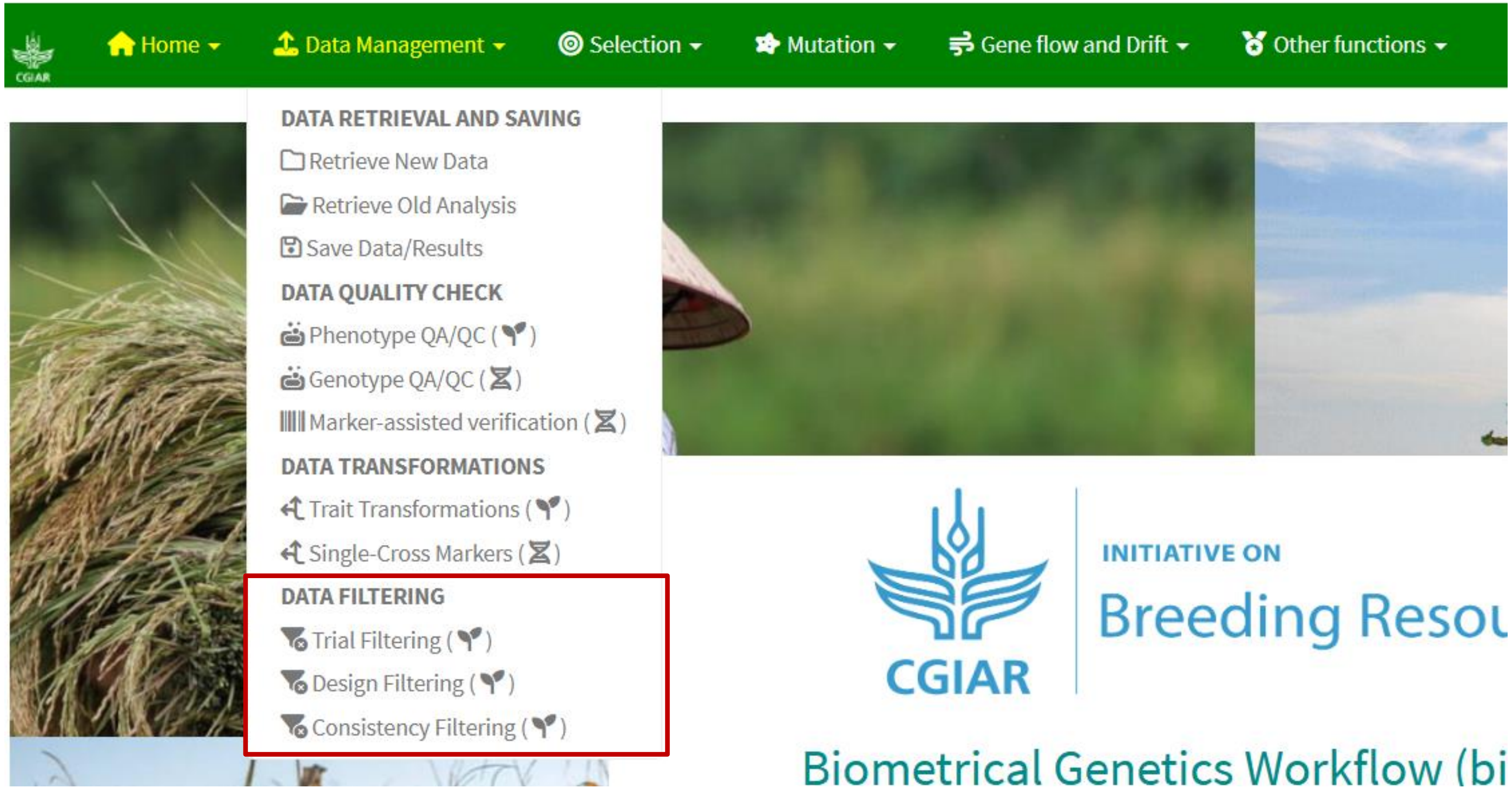
Bioflow overall interface



Bioflow overall interface



Bioflow overall interface



Home
Data Management
Selection
Mutation
Gene flow and Drift
Other functions

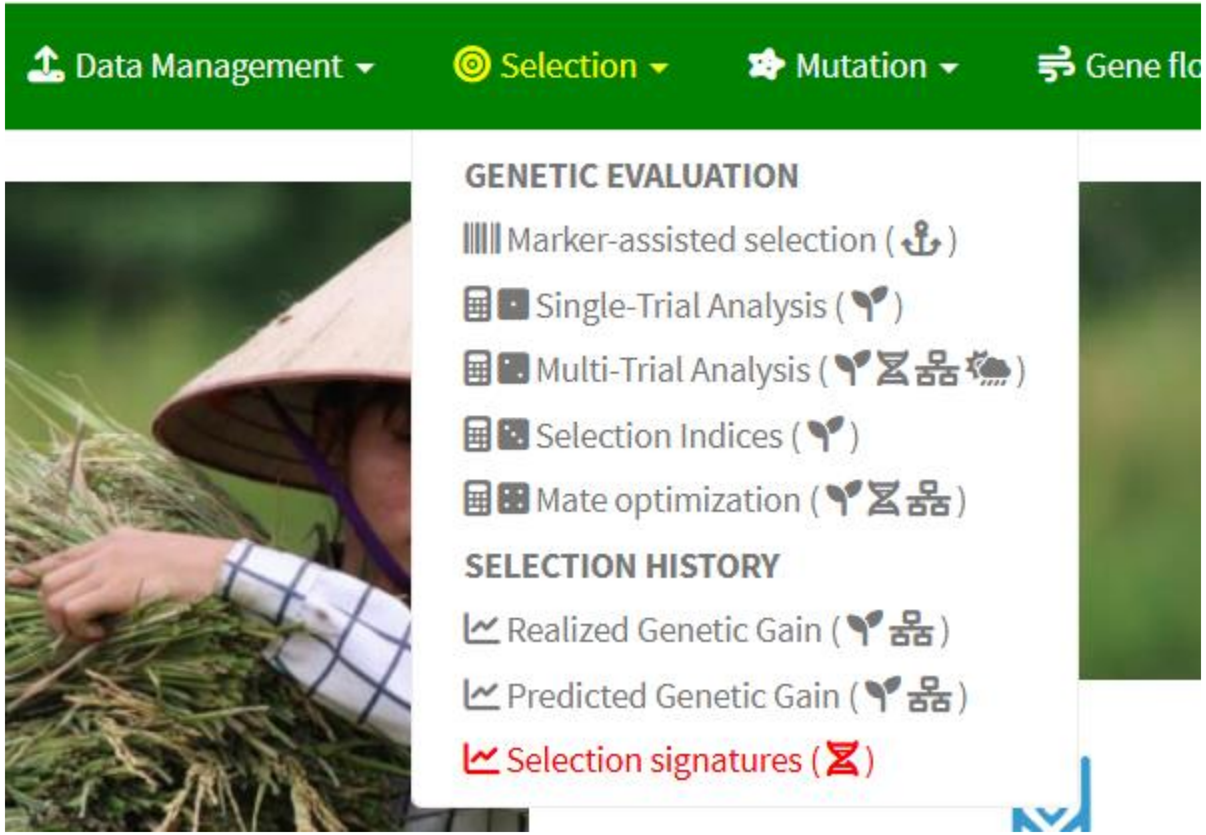
DATA RETRIEVAL AND SAVING
- Retrieve New Data
- Retrieve Old Analysis
- Save Data/Results

DATA QUALITY CHECK
- Phenotype QA/QC ( )
- Genotype QA/QC ( )
- Marker-assisted verification ( )

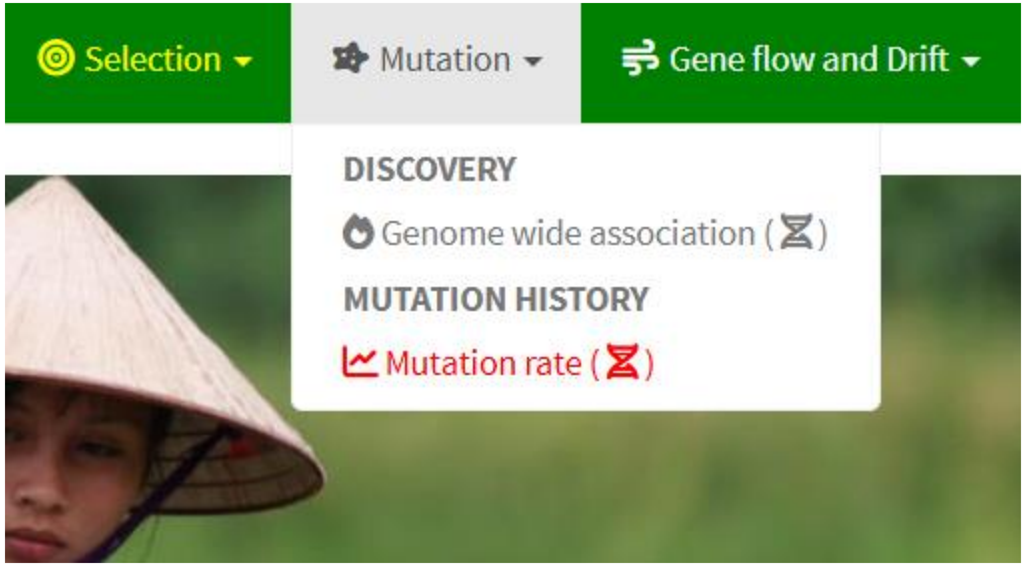
DATA TRANSFORMATIONS
- Trait Transformations ( )
- Single-Cross Markers ( )

DATA FILTERING
- Trial Filtering ( )
- Design Filtering ( )
- Consistency Filtering ( )

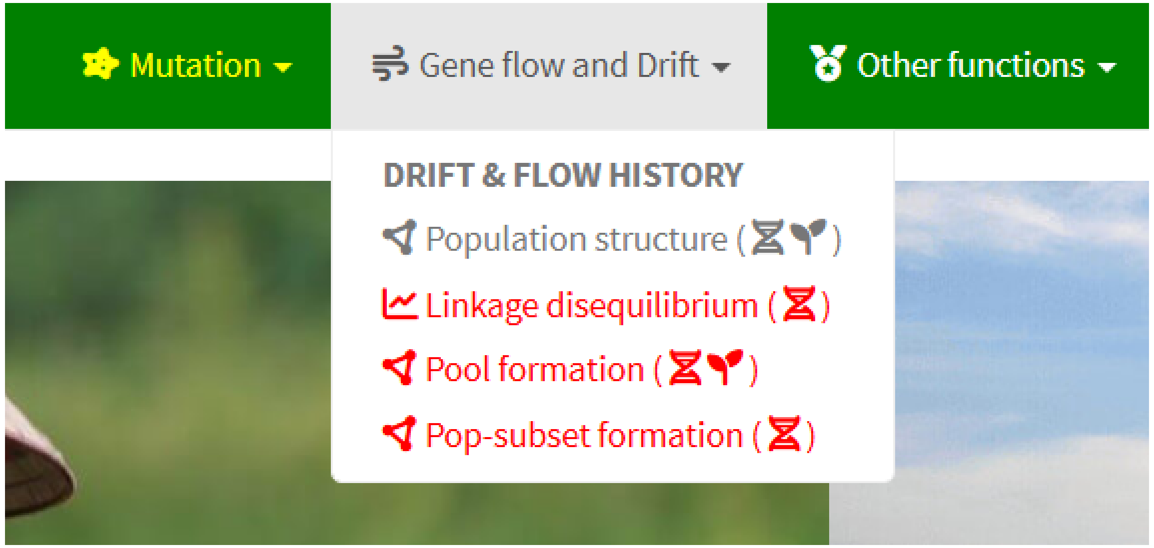
Bioflow overall interface



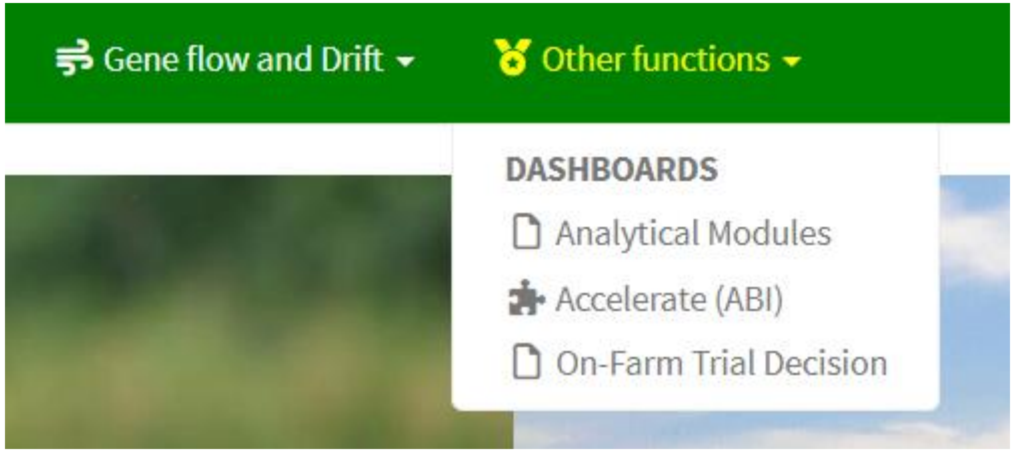
Bioflow overall interface



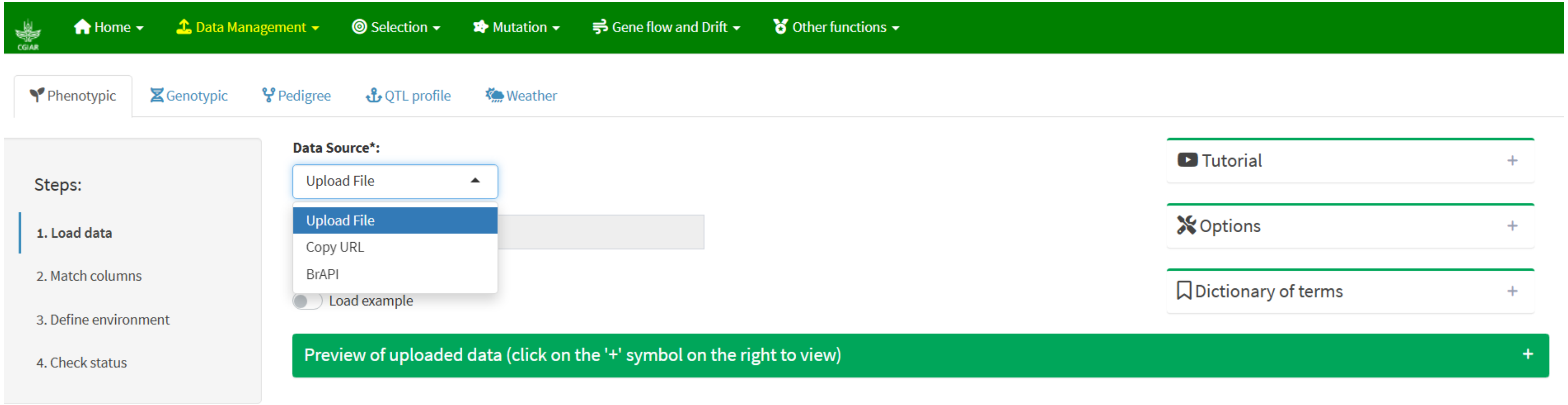
Bioflow overall interface



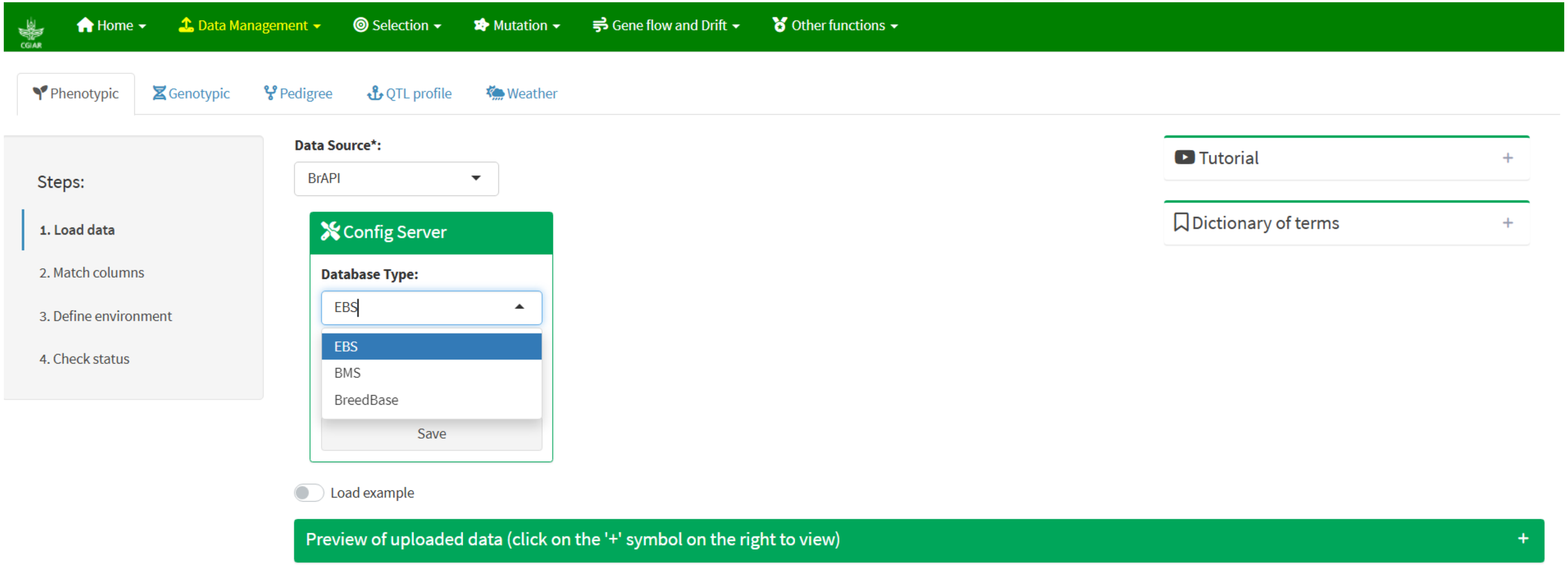
Bioflow overall interface



Data management > Retrieve new data



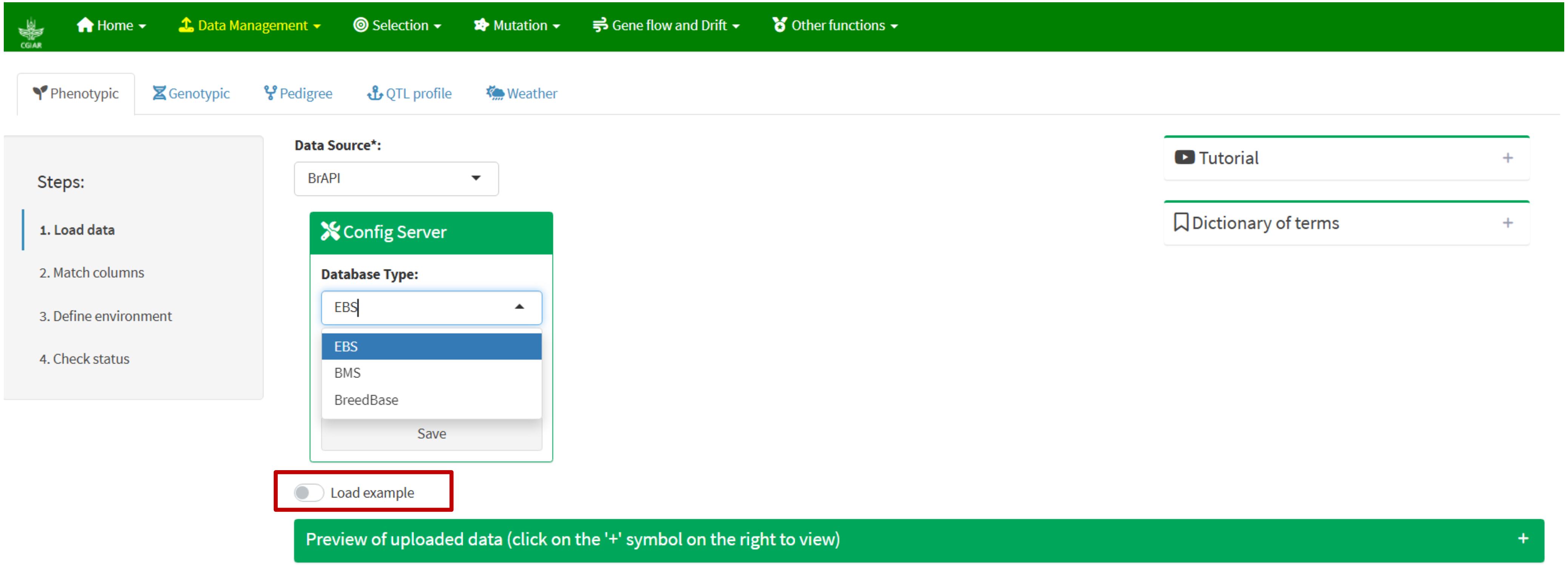
Data management > Retrieve new data



Data Source*: BrAPI

Database Type: EBS

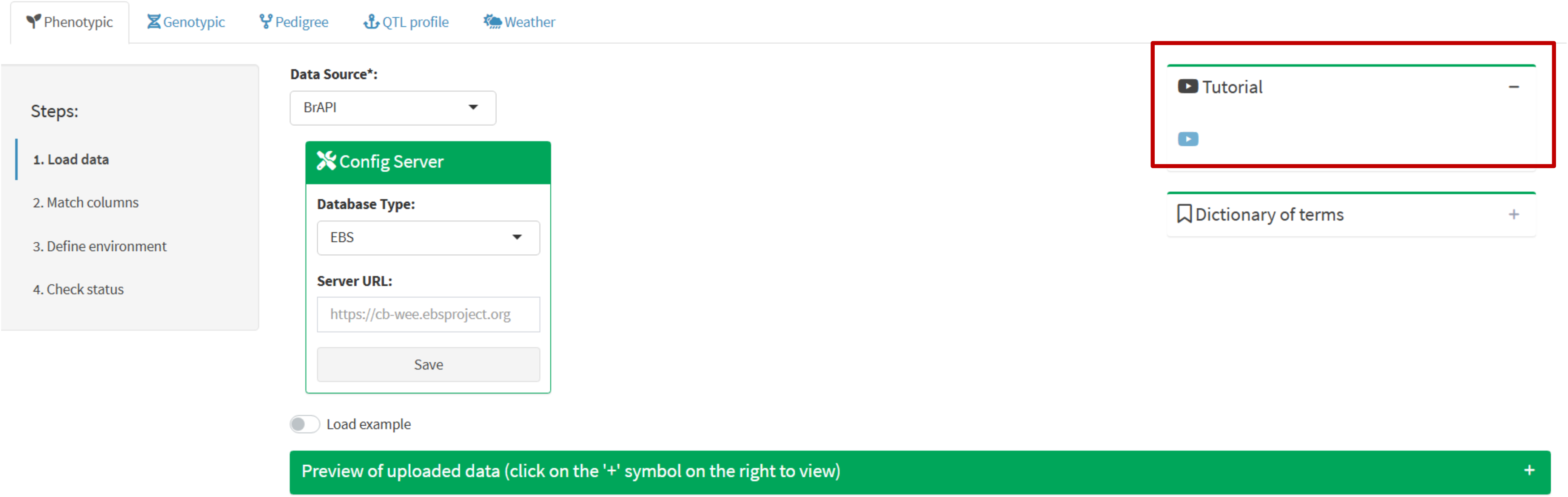
Data management > Retrieve new data



Data Source*: BrAPI

Database Type: EBS

Data management > Retrieve new data

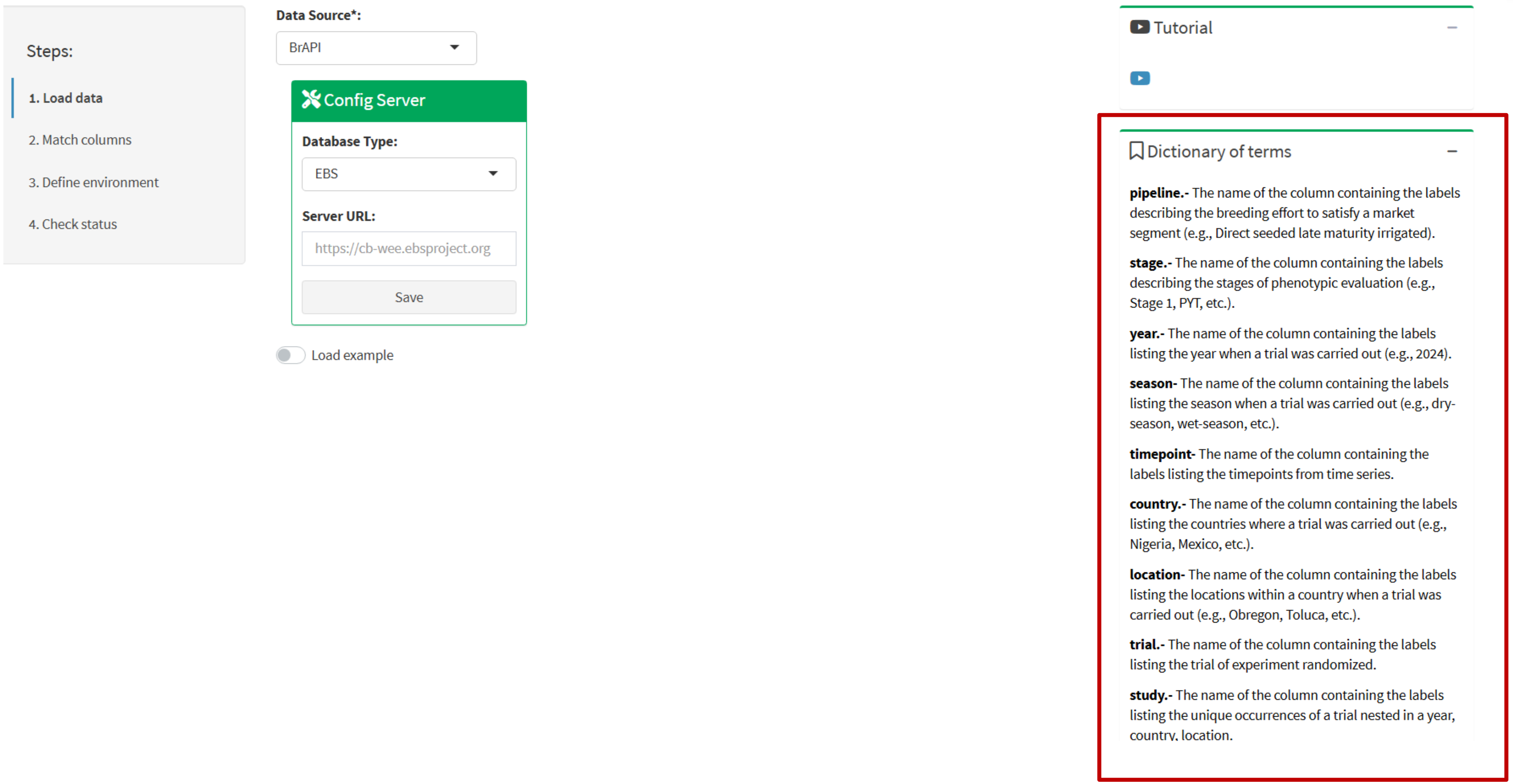


Data Source*: BrAPI

Database Type: EBS

Server URL: https://cb-wee.ebsproject.org

Data management > Retrieve new data



Steps:

1. Load data
2. Match columns
3. Define environment
4. Check status

Data Source*: BrAPI

Config Server

Database Type: EBS

Server URL: https://cb-wee.ebsproject.org

Save

○ Load example

Tutorial

Dictionary of terms

pipeline.- The name of the column containing the labels describing the breeding effort to satisfy a market segment (e.g., Direct seeded late maturity irrigated).

stage.- The name of the column containing the labels describing the stages of phenotypic evaluation (e.g., Stage 1, PYT, etc.).

year.- The name of the column containing the labels listing the year when a trial was carried out (e.g., 2024).

season- The name of the column containing the labels listing the season when a trial was carried out (e.g., dry-season, wet-season, etc.).

timepoint- The name of the column containing the labels listing the timepoints from time series.

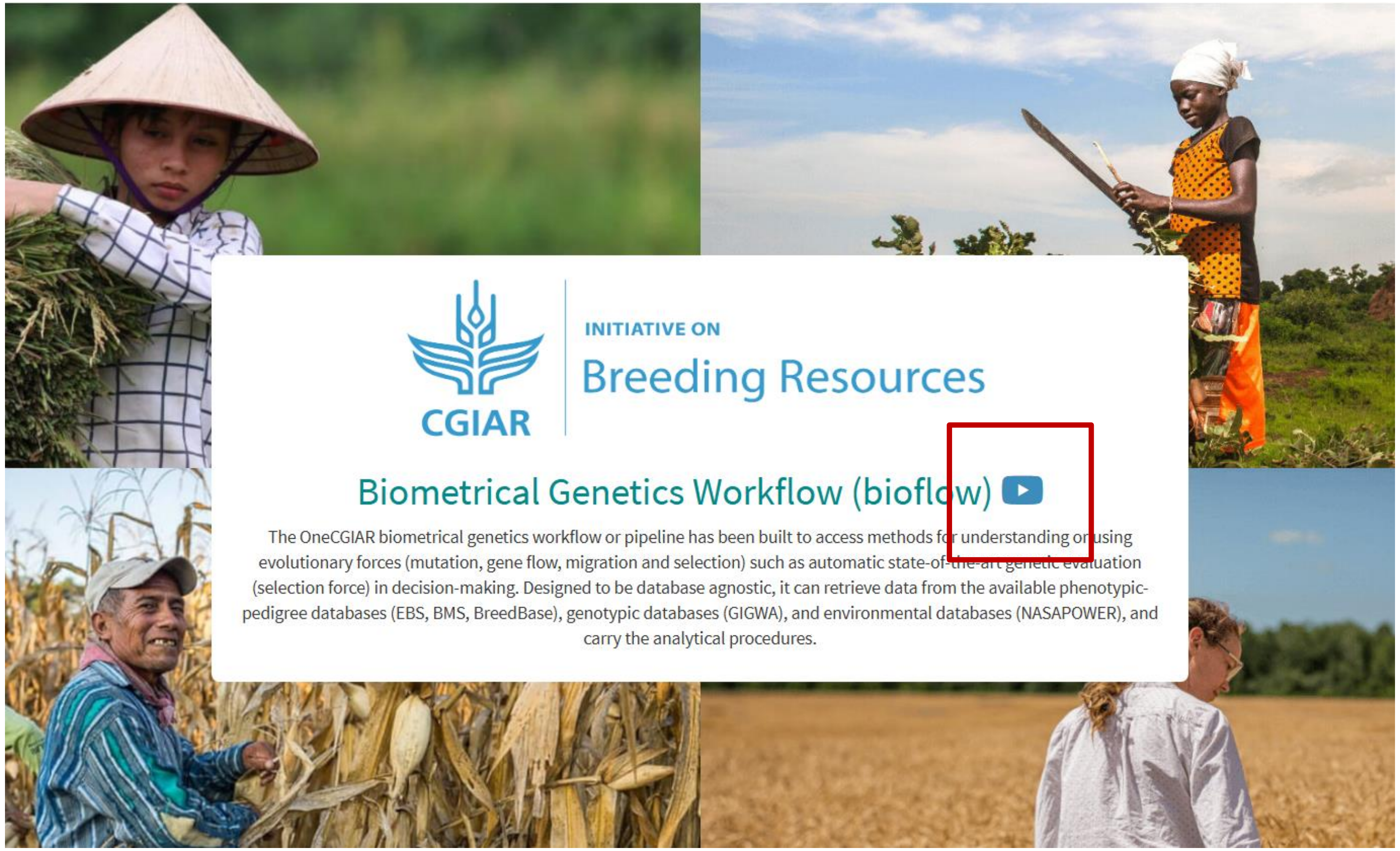
country.- The name of the column containing the labels listing the countries where a trial was carried out (e.g., Nigeria, Mexico, etc.).

location- The name of the column containing the labels listing the locations within a country when a trial was carried out (e.g., Obregon, Toluca, etc.).

trial.- The name of the column containing the labels listing the trial of experiment randomized.

study.- The name of the column containing the labels listing the unique occurrences of a trial nested in a year, country, location.

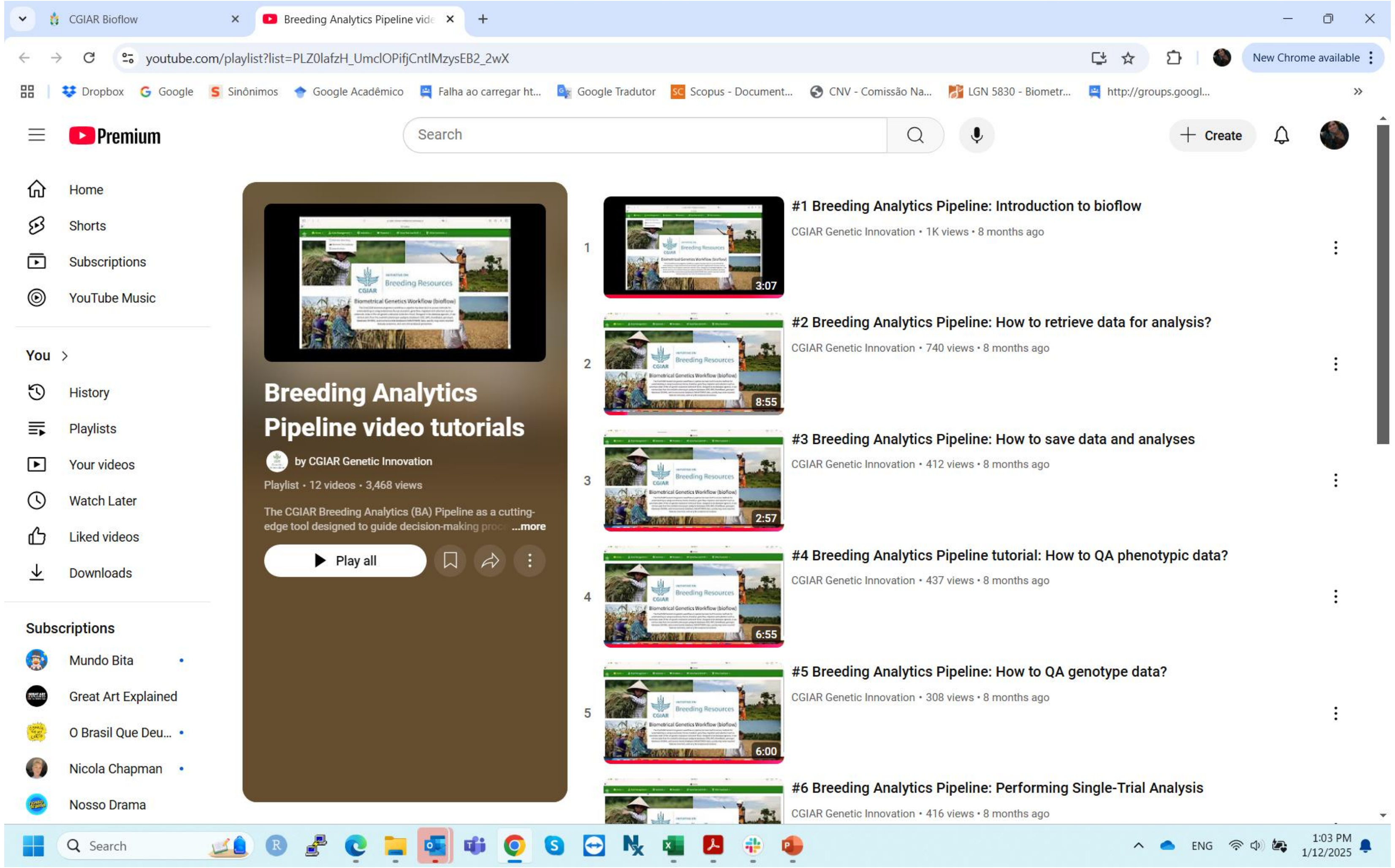
Help resources



Biometrical Genetics Workflow (bioflow)

The OneCGIAR biometrical genetics workflow or pipeline has been built to access methods for understanding or using evolutionary forces (mutation, gene flow, migration and selection) such as automatic state-of-the-art genetic evaluation (selection force) in decision-making. Designed to be database agnostic, it can retrieve data from the available phenotypic-pedigree databases (EBS, BMS, BreedBase), genotypic databases (GIGWA), and environmental databases (NASAPOWER), and carry the analytical procedures.

Help resources



Breeding Analytics Pipeline video tutorials

by CGIAR Genetic Innovation

Playlist • 12 videos • 3,468 views

The CGIAR Breeding Analytics (BA) Pipeline as a cutting-edge tool designed to guide decision-making proc ...more

▶ Play all

1 **#1 Breeding Analytics Pipeline: Introduction to bioflow** (3:07)
CGIAR Genetic Innovation • 1K views • 8 months ago

2 **#2 Breeding Analytics Pipeline: How to retrieve data for analysis?** (8:55)
CGIAR Genetic Innovation • 740 views • 8 months ago

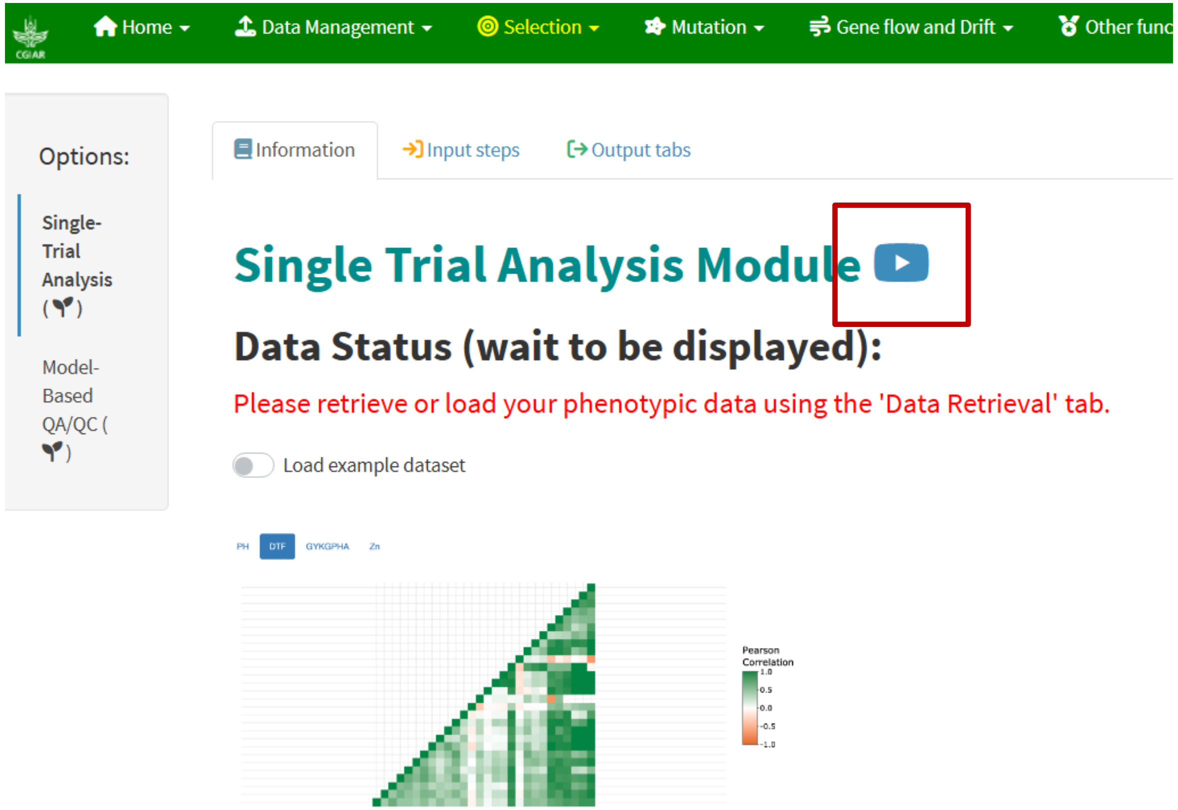
3 **#3 Breeding Analytics Pipeline: How to save data and analyses** (2:57)
CGIAR Genetic Innovation • 412 views • 8 months ago

4 **#4 Breeding Analytics Pipeline tutorial: How to QA phenotypic data?** (6:55)
CGIAR Genetic Innovation • 437 views • 8 months ago

5 **#5 Breeding Analytics Pipeline: How to QA genotype data?** (6:00)
CGIAR Genetic Innovation • 308 views • 8 months ago

#6 Breeding Analytics Pipeline: Performing Single-Trial Analysis
CGIAR Genetic Innovation • 416 views • 8 months ago

Help resources

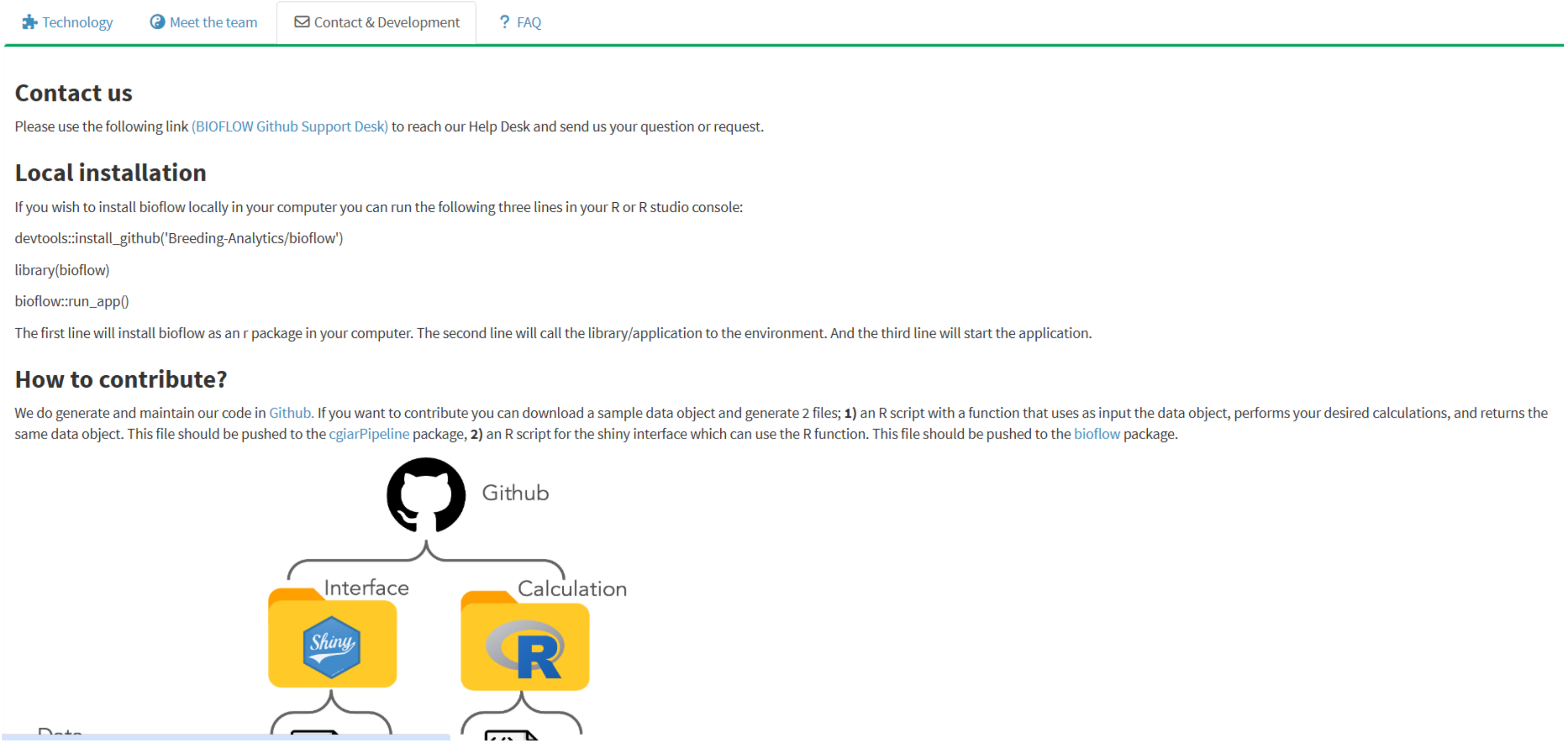


Single Trial Analysis Module

Data Status (wait to be displayed):

Please retrieve or load your phenotypic data using the 'Data Retrieval' tab.

Help resources



Contact us

Please use the following link (BIOFLOW Github Support Desk) to reach our Help Desk and send us your question or request.

Local installation

If you wish to install bioflow locally in your computer you can run the following three lines in your R or R studio console:

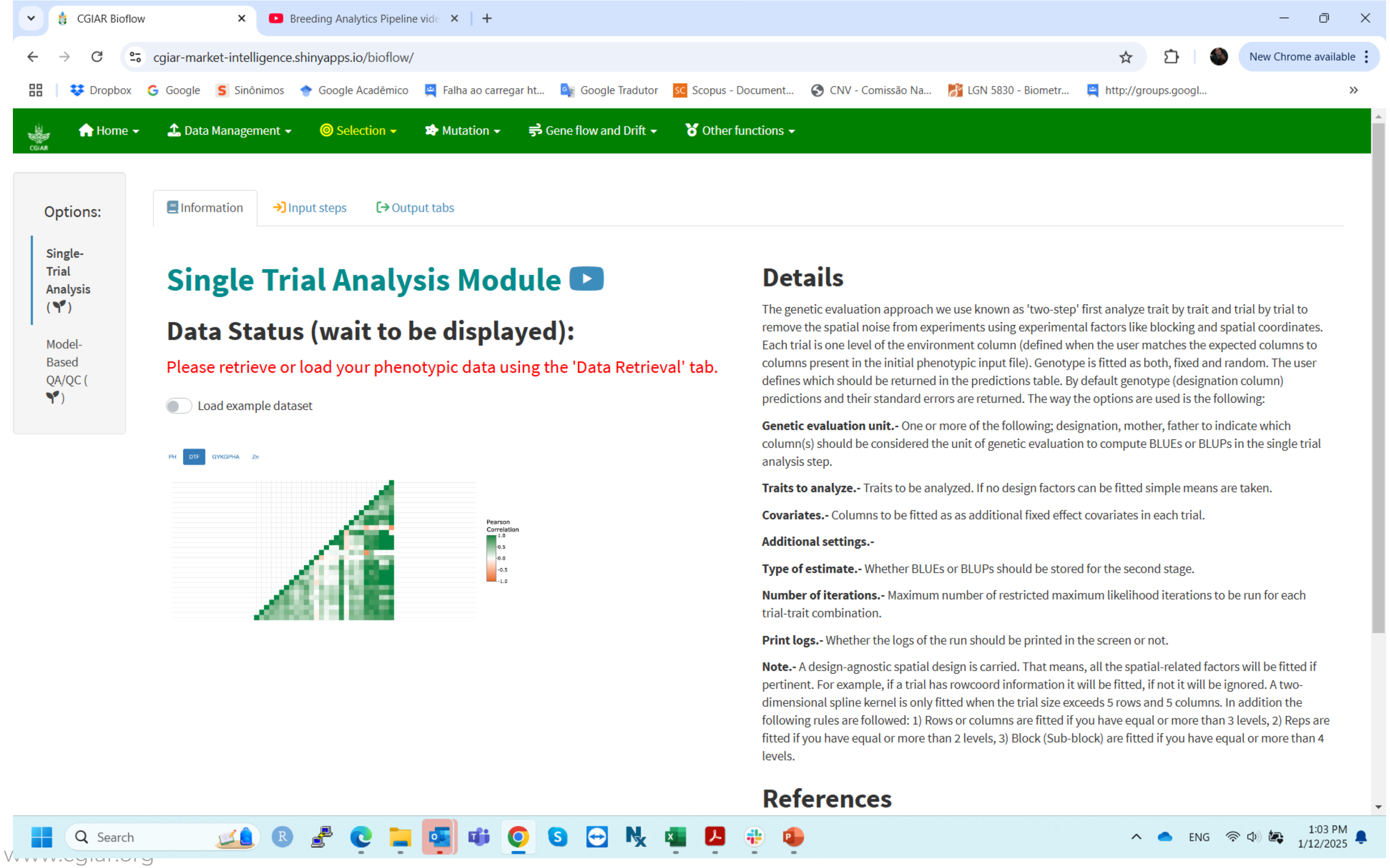
```
devtools::install_github('Breeding-Analytics/bioflow')

library(bioflow)

bioflow::run_app()
```

The first line will install bioflow as an r package in your computer. The second line will call the library/application to the environment. And the third line will start the application.

How to contribute?

We do generate and maintain our code in Github. If you want to contribute you can download a sample data object and generate 2 files; **1)** an R script with a function that uses as input the data object, performs your desired calculations, and returns the same data object. This file should be pushed to the cgiarPipeline package, **2)** an R script for the shiny interface which can use the R function. This file should be pushed to the bioflow package.

Selection modules



Single Trial Analysis Module

Data Status (wait to be displayed):

Please retrieve or load your phenotypic data using the 'Data Retrieval' tab.

○ Load example dataset

Details

The genetic evaluation approach we use known as 'two-step' first analyze trait by trait and trial by trial to remove the spatial noise from experiments using experimental factors like blocking and spatial coordinates. Each trial is one level of the environment column (defined when the user matches the expected columns to columns present in the initial phenotypic input file). Genotype is fitted as both, fixed and random. The user defines which should be returned in the predictions table. By default genotype (designation column) predictions and their standard errors are returned. The way the options are used is the following:

Genetic evaluation unit.- One or more of the following; designation, mother, father to indicate which column(s) should be considered the unit of genetic evaluation to compute BLUEs or BLUPs in the single trial analysis step.

Traits to analyze.- Traits to be analyzed. If no design factors can be fitted simple means are taken.

Covariates.- Columns to be fitted as as additional fixed effect covariates in each trial.

Additional settings.-

Type of estimate.- Whether BLUEs or BLUPs should be stored for the second stage.

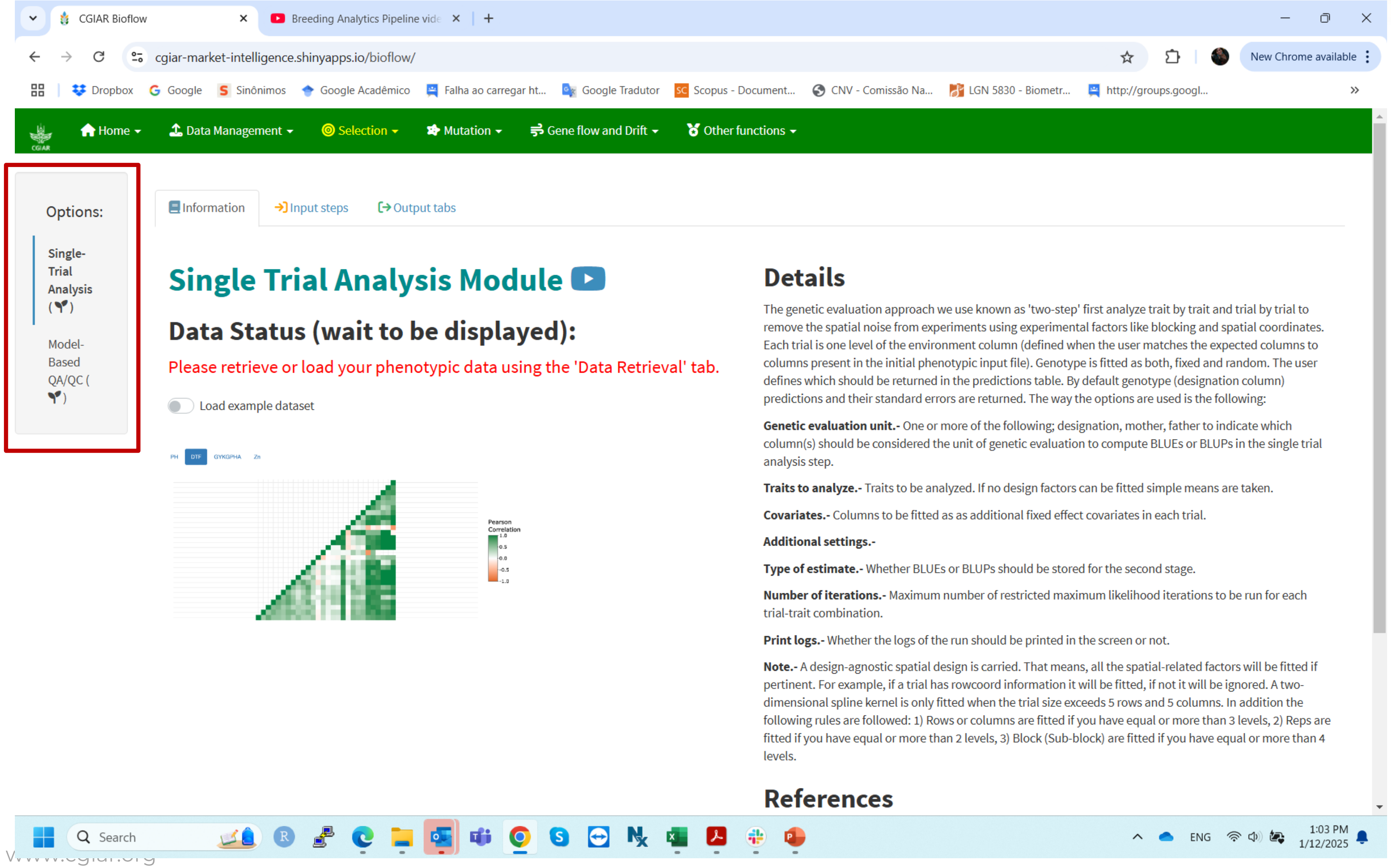
Number of iterations.- Maximum number of restricted maximum likelihood iterations to be run for each trial-trait combination.

Print logs.- Whether the logs of the run should be printed in the screen or not.

Note.- A design-agnostic spatial design is carried. That means, all the spatial-related factors will be fitted if pertinent. For example, if a trial has rowcoord information it will be fitted, if not it will be ignored. A two-dimensional spline kernel is only fitted when the trial size exceeds 5 rows and 5 columns. In addition the following rules are followed: 1) Rows or columns are fitted if you have equal or more than 3 levels, 2) Reps are fitted if you have equal or more than 2 levels, 3) Block (Sub-block) are fitted if you have equal or more than 4 levels.

References

Selection modules



Single Trial Analysis Module

Data Status (wait to be displayed):

Please retrieve or load your phenotypic data using the 'Data Retrieval' tab.

Details

The genetic evaluation approach we use known as 'two-step' first analyze trait by trait and trial by trial to remove the spatial noise from experiments using experimental factors like blocking and spatial coordinates. Each trial is one level of the environment column (defined when the user matches the expected columns to columns present in the initial phenotypic input file). Genotype is fitted as both, fixed and random. The user defines which should be returned in the predictions table. By default genotype (designation column) predictions and their standard errors are returned. The way the options are used is the following:

Genetic evaluation unit.- One or more of the following; designation, mother, father to indicate which column(s) should be considered the unit of genetic evaluation to compute BLUEs or BLUPs in the single trial analysis step.

Traits to analyze.- Traits to be analyzed. If no design factors can be fitted simple means are taken.

Covariates.- Columns to be fitted as as additional fixed effect covariates in each trial.

Additional settings.-

Type of estimate.- Whether BLUEs or BLUPs should be stored for the second stage.

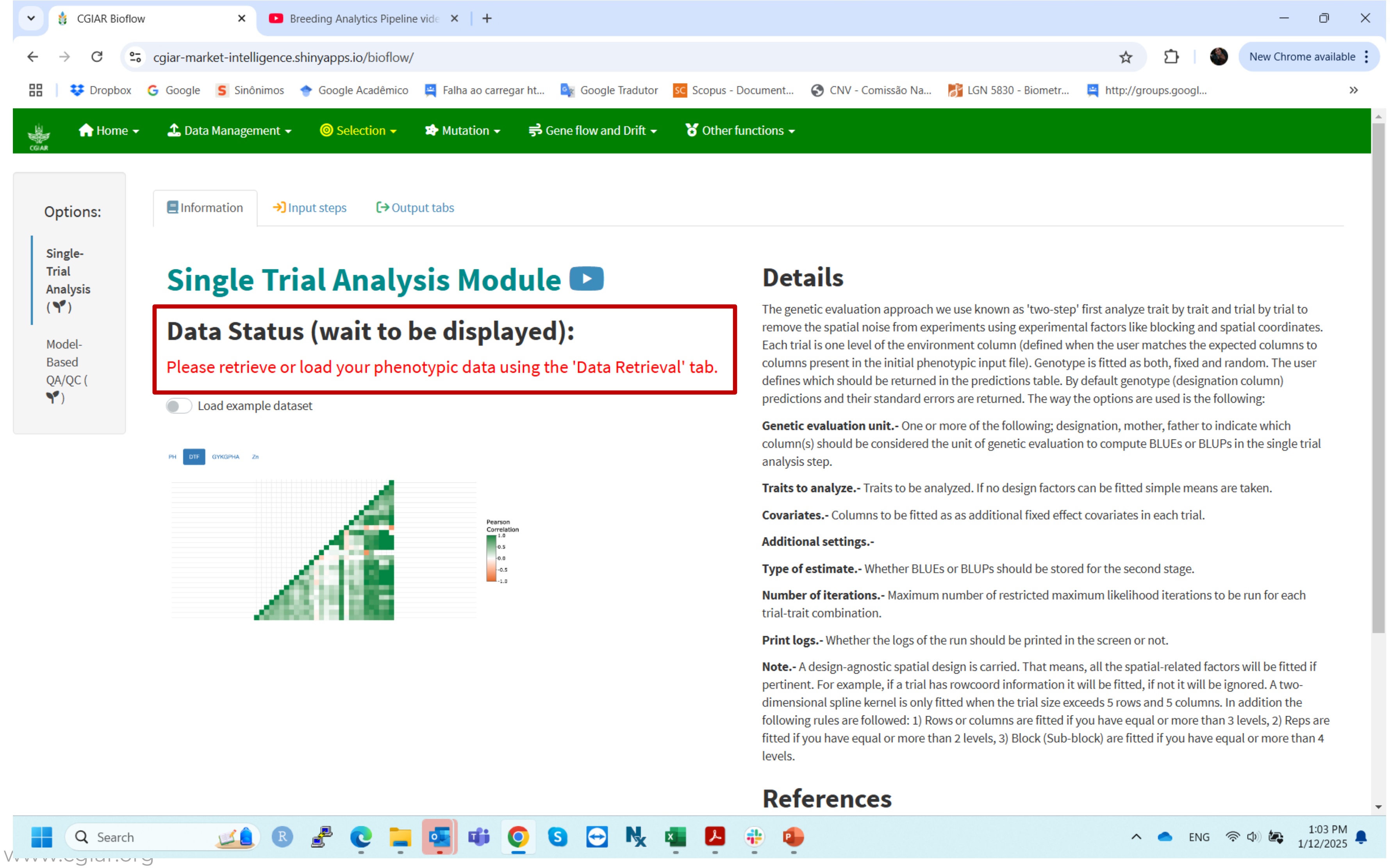
Number of iterations.- Maximum number of restricted maximum likelihood iterations to be run for each trial-trait combination.

Print logs.- Whether the logs of the run should be printed in the screen or not.

Note.- A design-agnostic spatial design is carried. That means, all the spatial-related factors will be fitted if pertinent. For example, if a trial has rowcoord information it will be fitted, if not it will be ignored. A two-dimensional spline kernel is only fitted when the trial size exceeds 5 rows and 5 columns. In addition the following rules are followed: 1) Rows or columns are fitted if you have equal or more than 3 levels, 2) Reps are fitted if you have equal or more than 2 levels, 3) Block (Sub-block) are fitted if you have equal or more than 4 levels.

References

Selection modules



Single Trial Analysis Module

Data Status (wait to be displayed):

Please retrieve or load your phenotypic data using the 'Data Retrieval' tab.

○ Load example dataset

Details

The genetic evaluation approach we use known as 'two-step' first analyze trait by trait and trial by trial to remove the spatial noise from experiments using experimental factors like blocking and spatial coordinates. Each trial is one level of the environment column (defined when the user matches the expected columns to columns present in the initial phenotypic input file). Genotype is fitted as both, fixed and random. The user defines which should be returned in the predictions table. By default genotype (designation column) predictions and their standard errors are returned. The way the options are used is the following:

Genetic evaluation unit.- One or more of the following; designation, mother, father to indicate which column(s) should be considered the unit of genetic evaluation to compute BLUEs or BLUPs in the single trial analysis step.

Traits to analyze.- Traits to be analyzed. If no design factors can be fitted simple means are taken.

Covariates.- Columns to be fitted as as additional fixed effect covariates in each trial.

Additional settings.-

Type of estimate.- Whether BLUEs or BLUPs should be stored for the second stage.

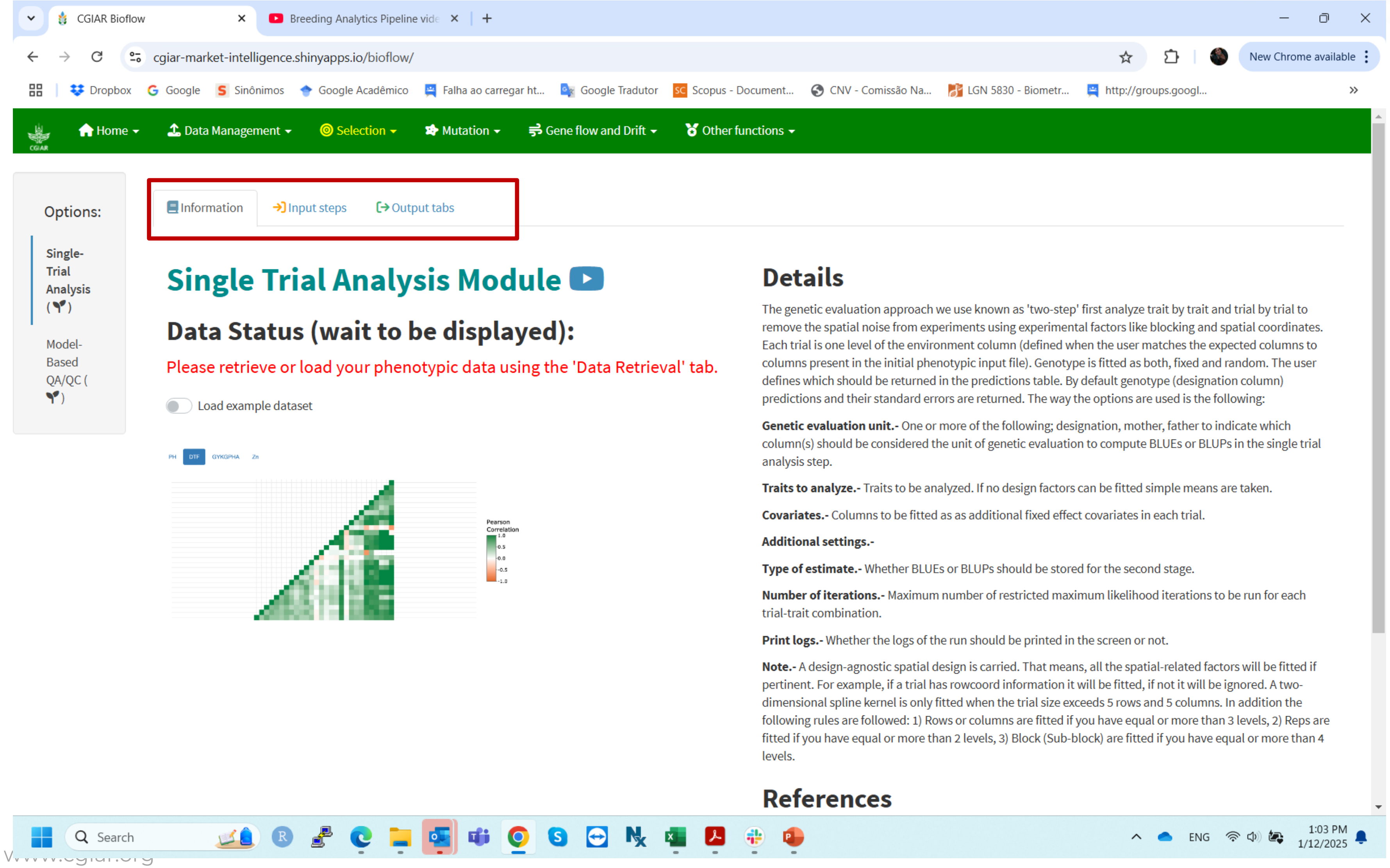
Number of iterations.- Maximum number of restricted maximum likelihood iterations to be run for each trial-trait combination.

Print logs.- Whether the logs of the run should be printed in the screen or not.

Note.- A design-agnostic spatial design is carried. That means, all the spatial-related factors will be fitted if pertinent. For example, if a trial has rowcoord information it will be fitted, if not it will be ignored. A two-dimensional spline kernel is only fitted when the trial size exceeds 5 rows and 5 columns. In addition the following rules are followed: 1) Rows or columns are fitted if you have equal or more than 3 levels, 2) Reps are fitted if you have equal or more than 2 levels, 3) Block (Sub-block) are fitted if you have equal or more than 4 levels.

References

Selection modules



Single Trial Analysis Module

Data Status (wait to be displayed):

Please retrieve or load your phenotypic data using the 'Data Retrieval' tab.

○ Load example dataset

Details

The genetic evaluation approach we use known as 'two-step' first analyze trait by trait and trial by trial to remove the spatial noise from experiments using experimental factors like blocking and spatial coordinates. Each trial is one level of the environment column (defined when the user matches the expected columns to columns present in the initial phenotypic input file). Genotype is fitted as both, fixed and random. The user defines which should be returned in the predictions table. By default genotype (designation column) predictions and their standard errors are returned. The way the options are used is the following:

Genetic evaluation unit.- One or more of the following; designation, mother, father to indicate which column(s) should be considered the unit of genetic evaluation to compute BLUEs or BLUPs in the single trial analysis step.

Traits to analyze.- Traits to be analyzed. If no design factors can be fitted simple means are taken.

Covariates.- Columns to be fitted as as additional fixed effect covariates in each trial.

Additional settings.-

Type of estimate.- Whether BLUEs or BLUPs should be stored for the second stage.

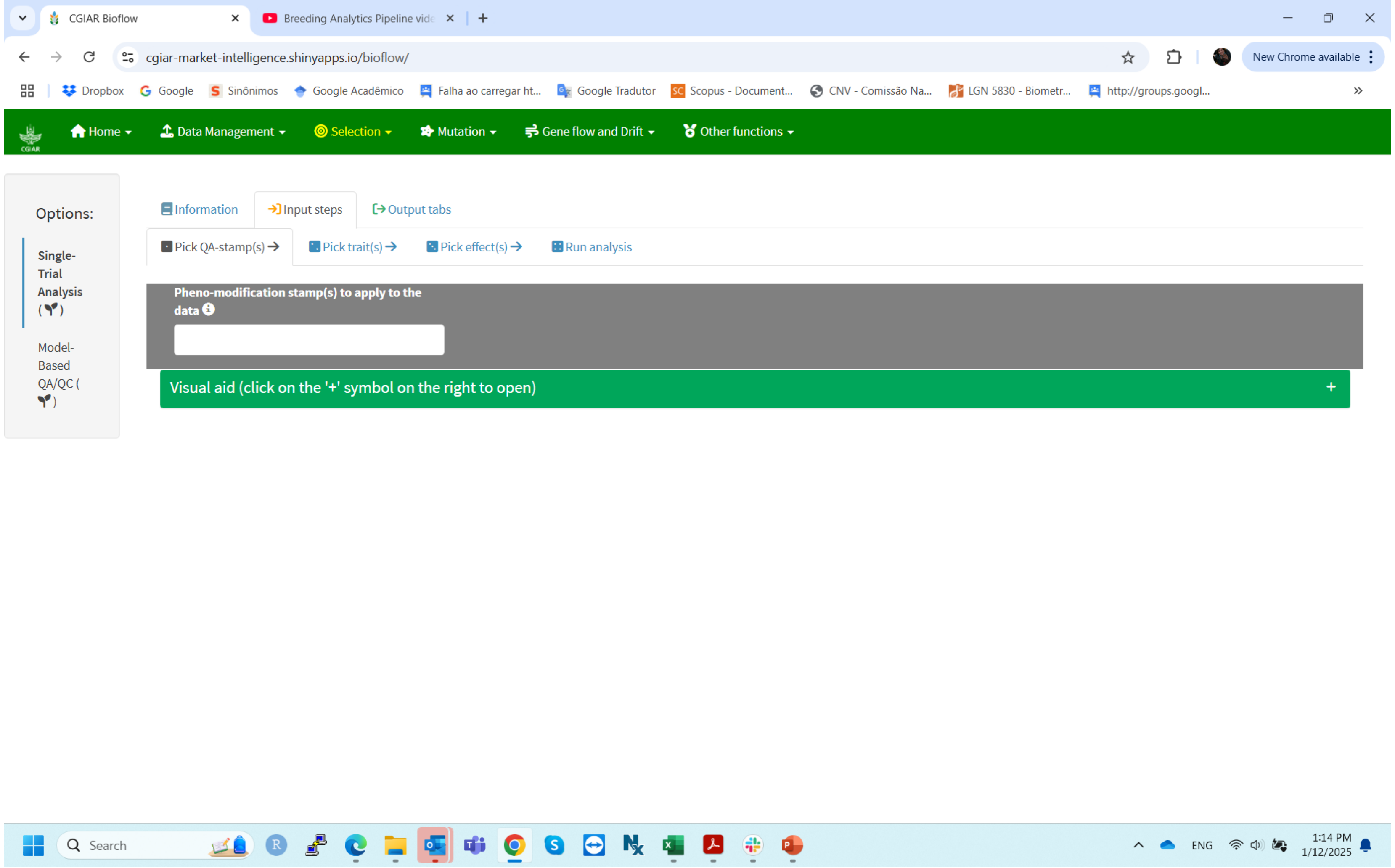
Number of iterations.- Maximum number of restricted maximum likelihood iterations to be run for each trial-trait combination.

Print logs.- Whether the logs of the run should be printed in the screen or not.

Note.- A design-agnostic spatial design is carried. That means, all the spatial-related factors will be fitted if pertinent. For example, if a trial has rowcoord information it will be fitted, if not it will be ignored. A two-dimensional spline kernel is only fitted when the trial size exceeds 5 rows and 5 columns. In addition the following rules are followed: 1) Rows or columns are fitted if you have equal or more than 3 levels, 2) Reps are fitted if you have equal or more than 2 levels, 3) Block (Sub-block) are fitted if you have equal or more than 4 levels.

References

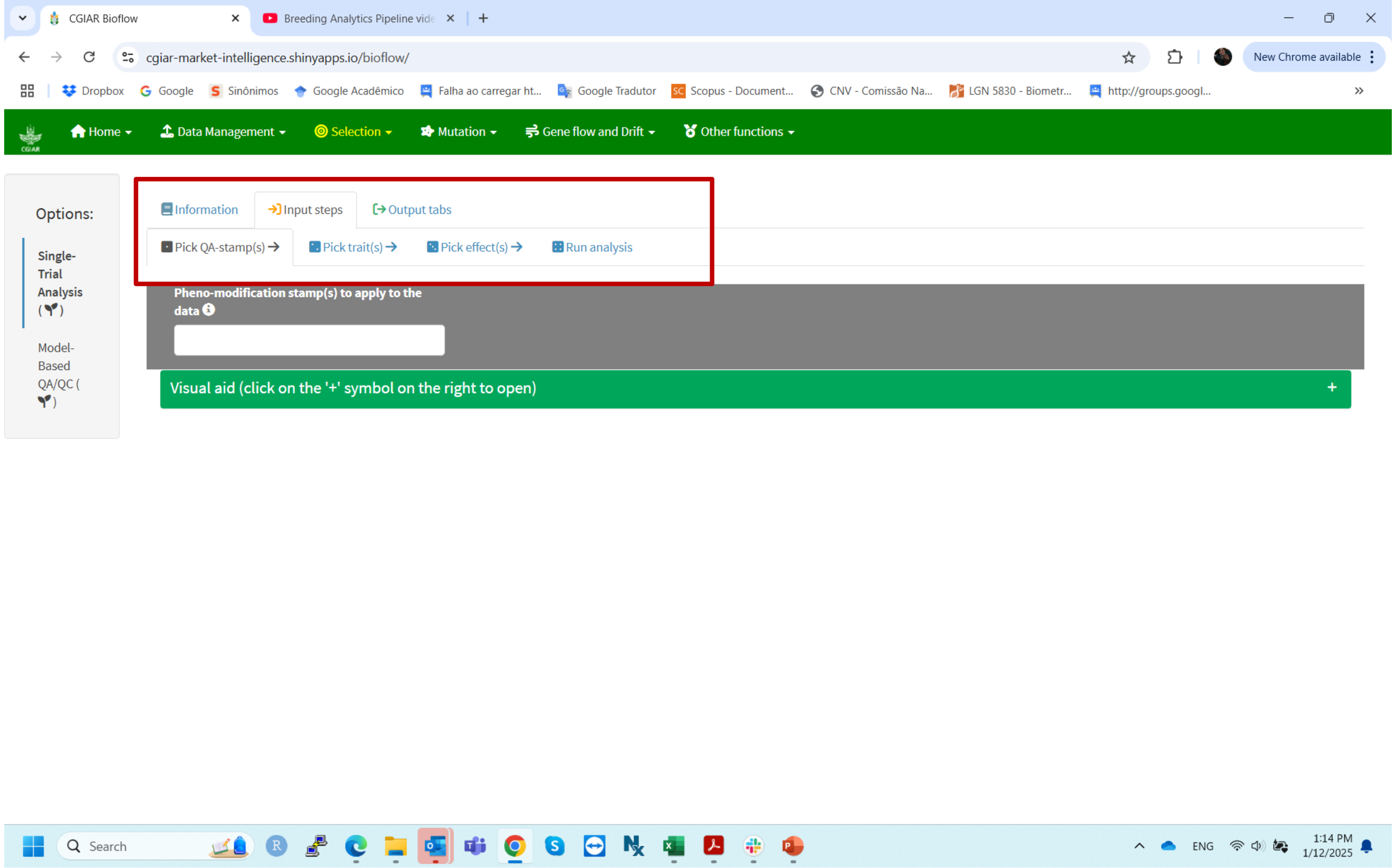
Selection modules



Pheno-modification stamp(s) to apply to the data

Visual aid (click on the '+' symbol on the right to open)

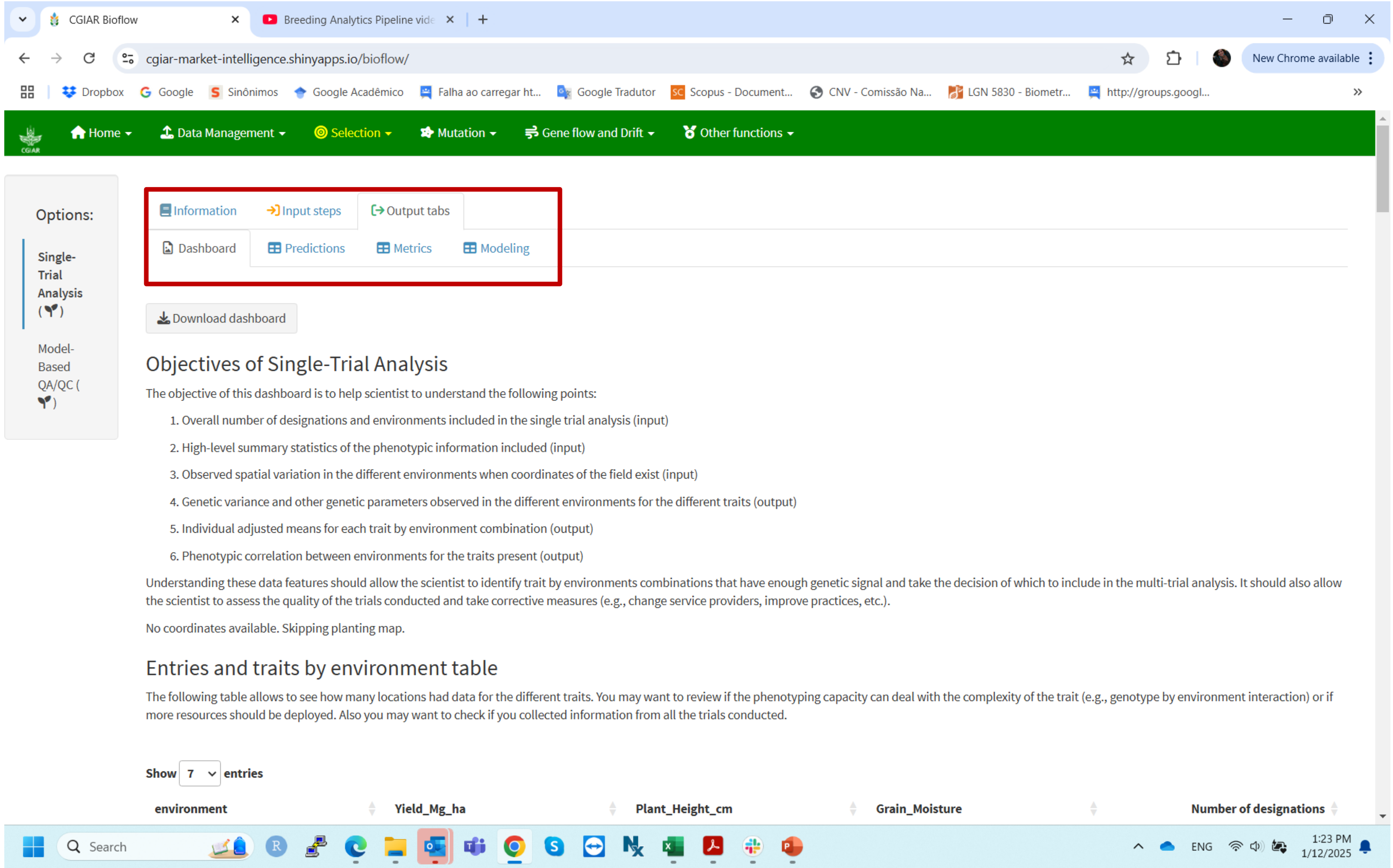
Selection modules



Pheno-modification stamp(s) to apply to the data

Visual aid (click on the '+' symbol on the right to open) +

Selection modules



Objectives of Single-Trial Analysis

The objective of this dashboard is to help scientist to understand the following points:

1. Overall number of designations and environments included in the single trial analysis (input)
2. High-level summary statistics of the phenotypic information included (input)
3. Observed spatial variation in the different environments when coordinates of the field exist (input)
4. Genetic variance and other genetic parameters observed in the different environments for the different traits (output)
5. Individual adjusted means for each trait by environment combination (output)
6. Phenotypic correlation between environments for the traits present (output)

Understanding these data features should allow the scientist to identify trait by environments combinations that have enough genetic signal and take the decision of which to include in the multi-trial analysis. It should also allow the scientist to assess the quality of the trials conducted and take corrective measures (e.g., change service providers, improve practices, etc.).

No coordinates available. Skipping planting map.

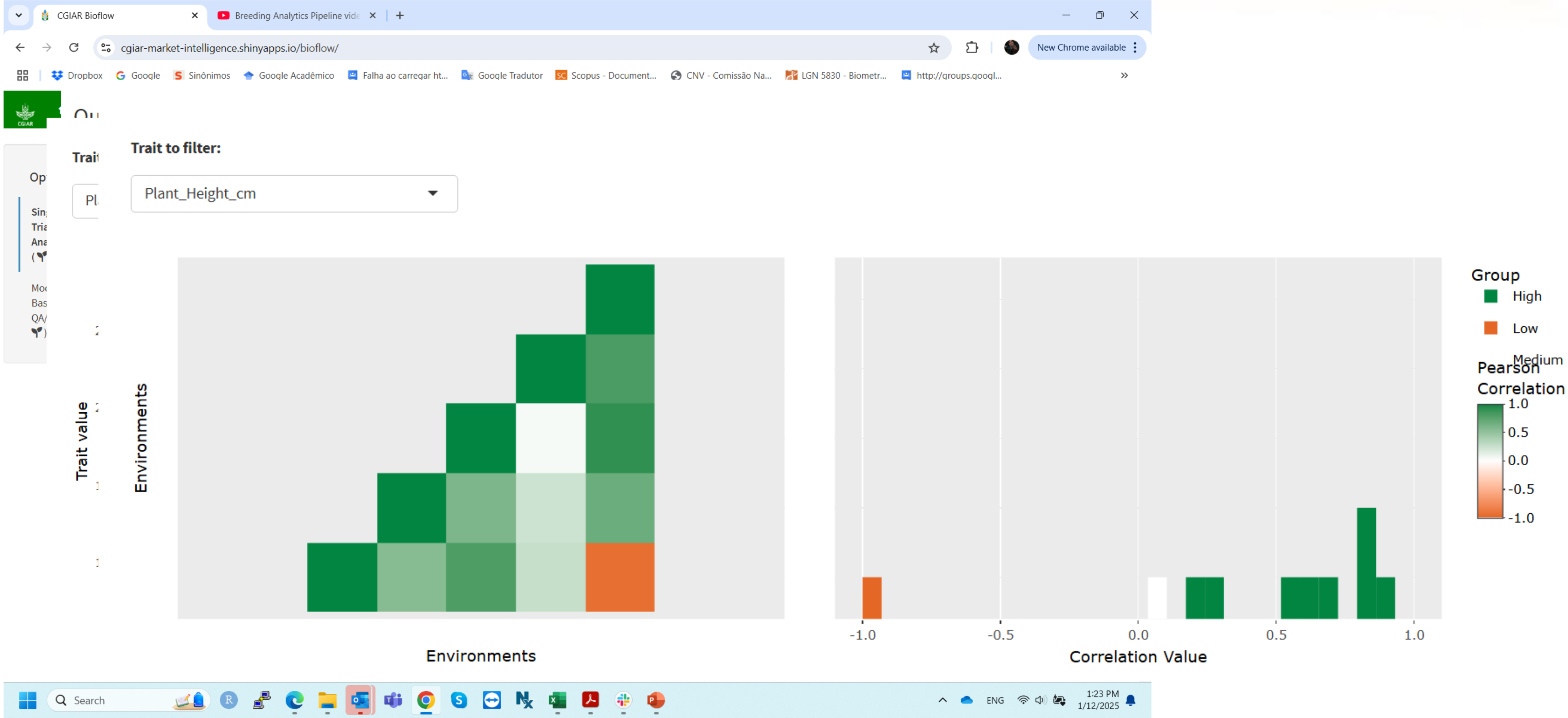
Entries and traits by environment table

The following table allows to see how many locations had data for the different traits. You may want to review if the phenotyping capacity can deal with the complexity of the trait (e.g., genotype by environment interaction) or if more resources should be deployed. Also you may want to check if you collected information from all the trials conducted.

Show 7 **entries**

environment	Yield_Mg_ha	Plant_Height_cm	Grain_Moisture	Number of designations

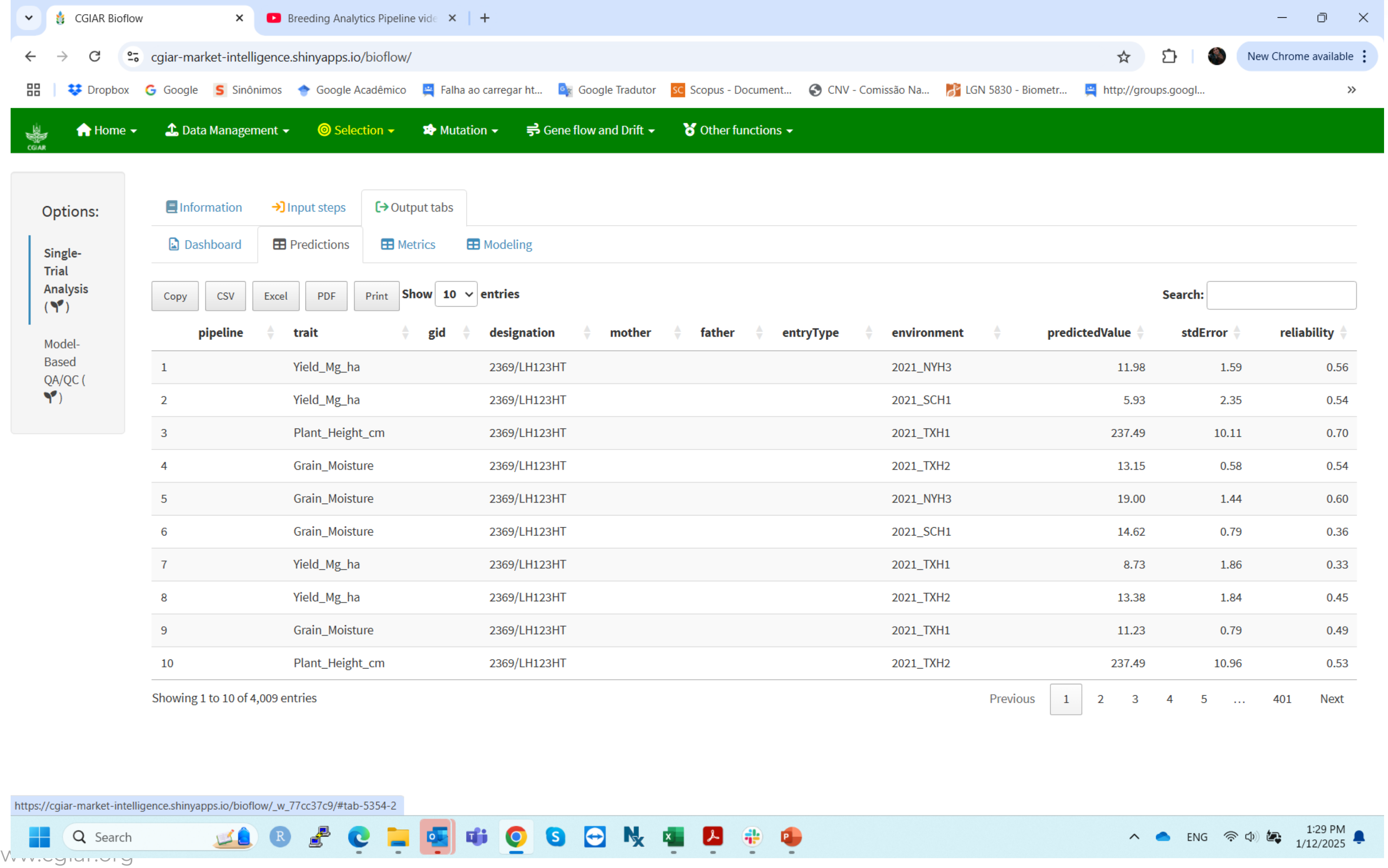
Selection modules



Trait to filter:

Plant_Height_cm

Selection modules



	pipeline	trait	gid	designation	mother	father	entryType	environment	predictedValue	stdError	reliability
1		Yield_Mg_ha		2369/LH123HT				2021_NYH3	11.98	1.59	0.56
2		Yield_Mg_ha		2369/LH123HT				2021_SCH1	5.93	2.35	0.54
3		Plant_Height_cm		2369/LH123HT				2021_TXH1	237.49	10.11	0.70
4		Grain_Moisture		2369/LH123HT				2021_TXH2	13.15	0.58	0.54
5		Grain_Moisture		2369/LH123HT				2021_NYH3	19.00	1.44	0.60
6		Grain_Moisture		2369/LH123HT				2021_SCH1	14.62	0.79	0.36
7		Yield_Mg_ha		2369/LH123HT				2021_TXH1	8.73	1.86	0.33
8		Yield_Mg_ha		2369/LH123HT				2021_TXH2	13.38	1.84	0.45
9		Grain_Moisture		2369/LH123HT				2021_TXH1	11.23	0.79	0.49
10		Plant_Height_cm		2369/LH123HT				2021_TXH2	237.49	10.96	0.53

Showing 1 to 10 of 4,009 entries

Hope I was able to entice you to try out BIOFLOW!

Demo session in the afternoon!

2pm-5pm

California 2

CGIAR
Science for a food-secure future

Thank you!