



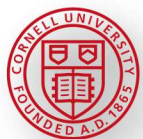
# *Integrated Genomic Selection Galaxy Analysis Pipeline and Workflows*

**Star Yanxin Gao**  
**yg28@cornell.edu**  
**<http://gobiiproject.org>**

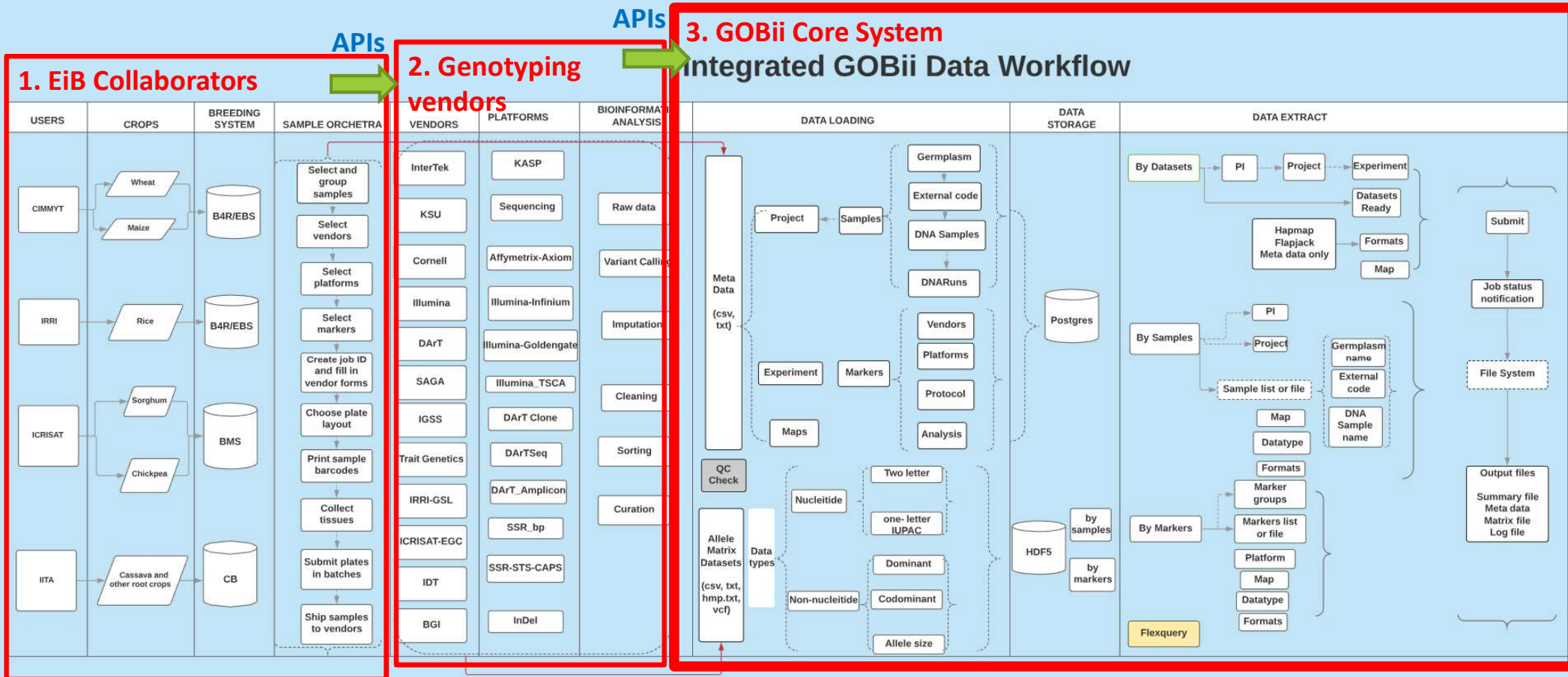
BILL & MELINDA  
GATES foundation

# GOBii: A Global Team

<http://gobiiproject.org>



# What is GOBii?



# What is GOBii (Cont'd)?

Plug-ins

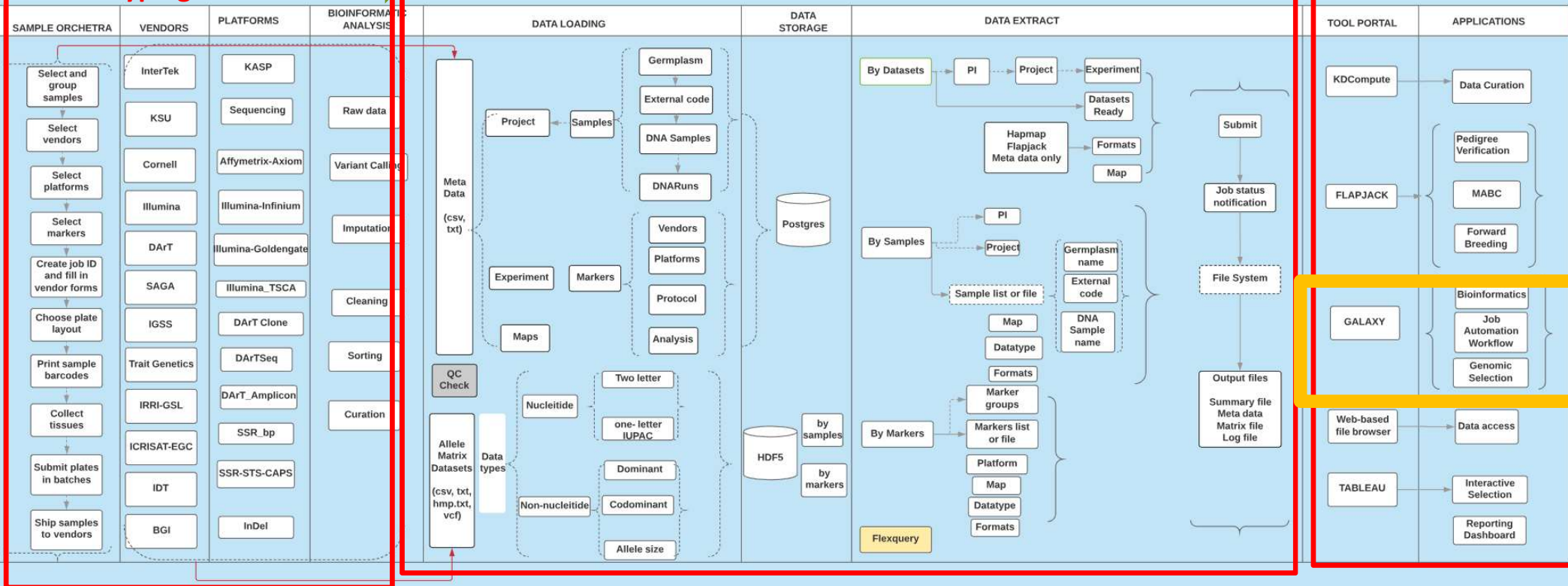
## 4. GOBii Analysis and decision support tools

APIs

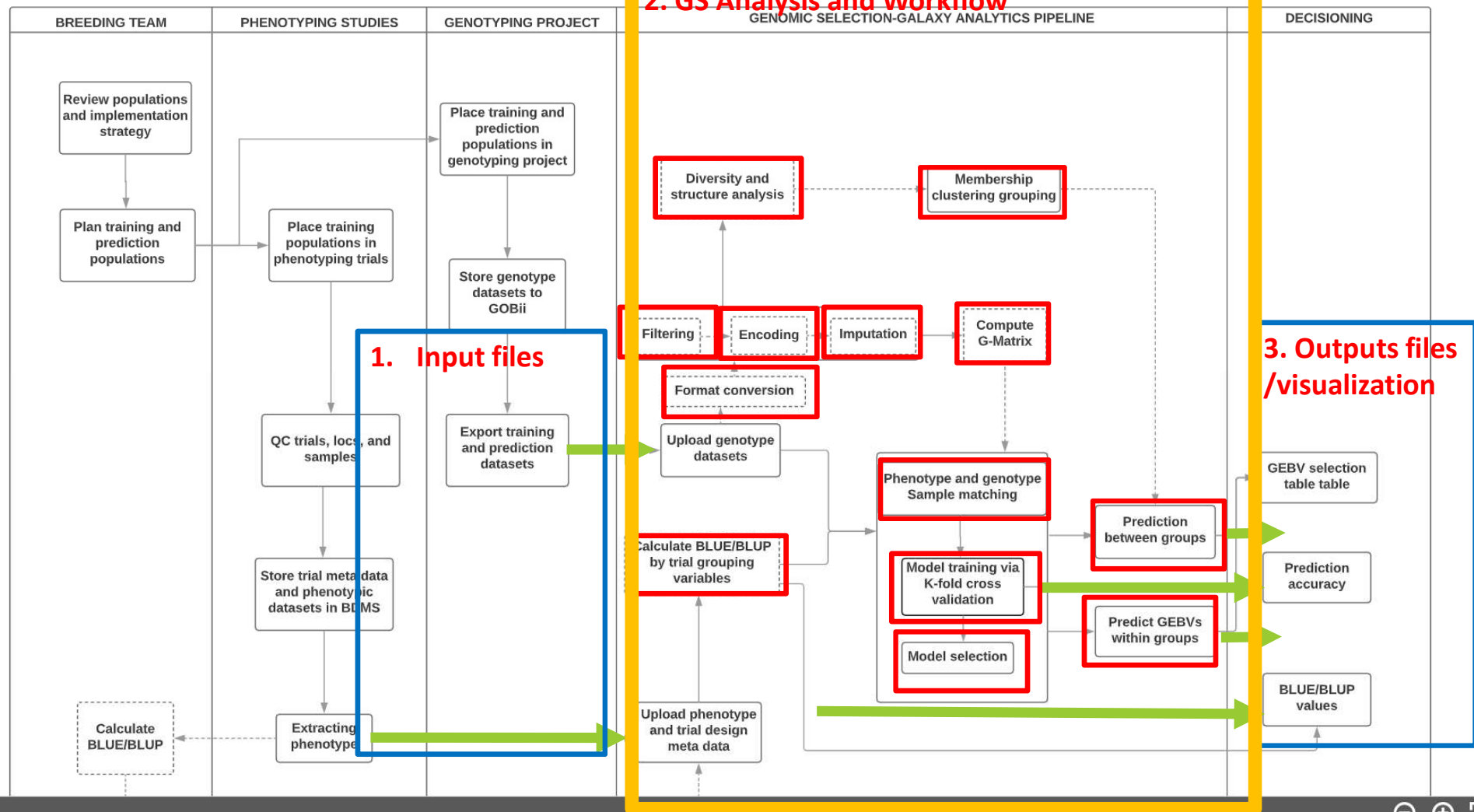
## 3. GOBii Core System Integrated GOBii Data Workflow

APIs

## 2. Genotyping vendors



# Genomic Selection? Why Galaxy?



# Importing Workflows and Test Datasets

This first workflow is using Beagle for Imputation. It connects BLUP calculator, Imputation and GEBV prediction.

 Galaxy Workflow | Genomic selection (workflow1)   

## Input Datasets

Genotyping dataset in hapmap format




Import datasets

 Galaxy Dataset | test2.hmp   

Phenotyping dataset

 Galaxy Dataset | Phenotype test file 1015 GS-Galaxy.tab   

## Complete history for workflow1

 Galaxy History | Genomic selection workflow1  

## Your workflows

Run workflow

search for workflow...  

Name	Tags	Owner	# of Steps	Published	Show in tools panel
 SNIPlay diversity workflow		You	17	Yes	<input type="checkbox"/>
 Genomic selection (workflow1)		You	8	Yes	<input type="checkbox"/>
 G) - Flapjack Analysis		You	4	No	<input type="checkbox"/>

Edit

Run

Share

Download

Copy

Rename

View

Delete

# Define datasets and parameters

Send results to a new history

**1: Convert genotyping data (Galaxy Version 1.0)**

**Input genotyping file**  
1: imported: test2.hmp  
Genotyping file can be in VCF, hapmap, hapmap with IUPAC

☒ **Format for output file**  
Hapmap

**Job Post Actions**  
Hide output 'output'.

**2: BLUP/BLUE (Galaxy Version 1.11.0)**

**Encoded Data**  
5: imported: Phenotype\_test\_file\_1015 GS-Galaxy.tab  
(required) Must be the TAB as both column delimiter and file type

**Design**  
RCBD

☒ **Replication Column**  
1

☒ **Genotype Column**  
18

☒ **Y Column**  
13

☒ **Summarize By**  
false

☒ **Summarize By Column**  
3

☒ **Variable 1, Factor**  
false

☒ **Variable 1, Factor Column**  
6

☒ **Variable 2, Factor**  
false

☒ **Variable 2, Factor Column**

1. Users select input files: uploading your own or importing shared datasets

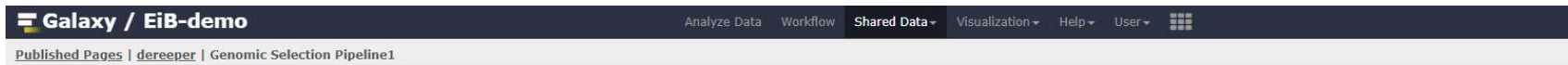
2. Users to define based on input file

5 shown  
4.39 MB

5: imported: Phenotype\_test\_file\_1015 GS-Galaxy.tab

1: imported: test2.hmp

# Genomic Selection (GS) demo-workflow1



## GOBii Genomic Selection Galaxy Workflows

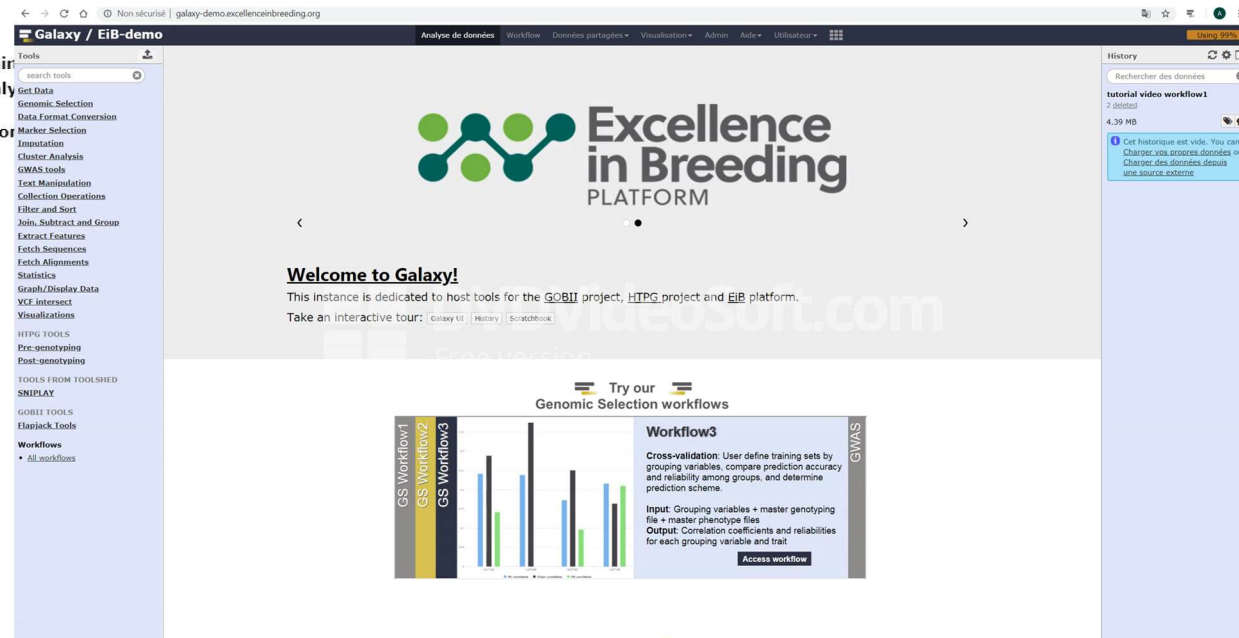
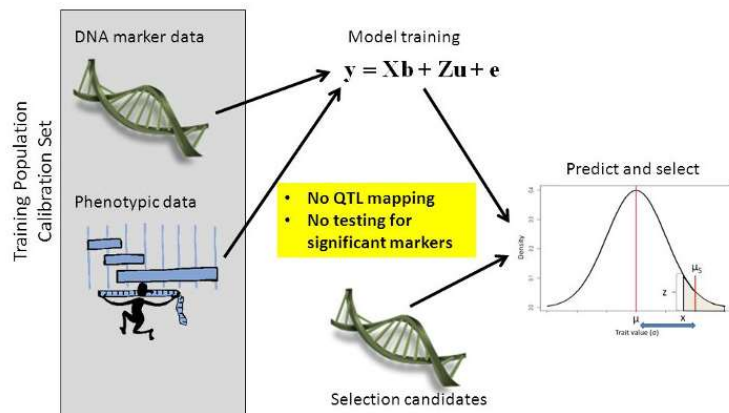
### Pipeline 1: Prediction of genomic breeding values (GEBV)

Prediction of genomic estimated breeding values (GEBV) for untested individuals within a training dataset.  
Calculate the genomic estimated breeding values (GEBVs) for individuals that have only a training dataset.

If sub-structure or groups exist within the test dataset, run Workflows 2 or 3 first before Pipeline 1.

- Input: *Genotyping matrix + Phenotyping file*
- Output: *Table of observed and predicted values for each trait*

## Genomic selection



# Genomic Selection (GS) demo-workflow 2

Galaxy / EiB-demo

Analyze Data Workflow Shared Data Visualization Help User

Published Pages | dereeper | Genomic Selection Pipeline2

## GOBii Genomic Selection Galaxy Workflows

### Pipeline 2: Population structure, clustering and d

Allows the grouping of genotyped samples using genetic information

- Input: *Genotyping matrix*

- Output: *Membership file*

Workflow2.1 : Based on Hybrid K-Means Clustering : Hybrid K-Means

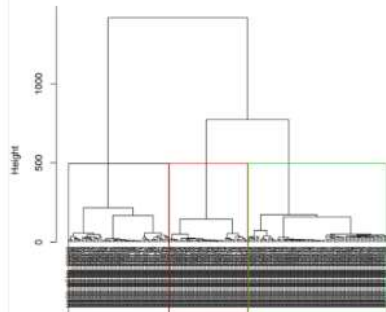
This workflow is based on the *hkmeans* solution implemented in *factoextra* R package

The final k-means clustering solution is very sensitive to the initial random selection

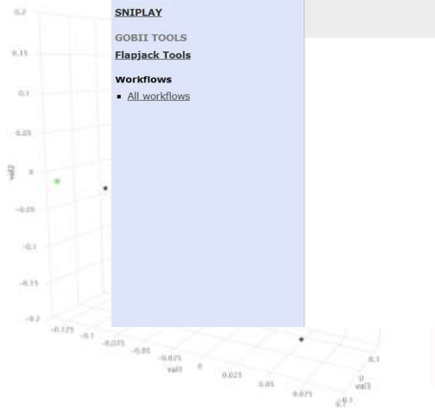
Input datasets: Genotyping dataset in hapmap format

History using workflow2.1

Examples of graphical outputs



Cluster Plot from Hybrid K-means clustering



MDS plot from plink, 3D visualization (samples colored by cluster)

Phylogenetic distance tree (samples colored by cluster)

Galaxy / EiB-demo

Analyse de données Workflow Données partagées Visualisation Admin Aide Utilisateur

Tools  
search tools  
Get Data  
Genomic Selection  
Data Format Conversion  
Marker Selection  
Imputation  
Cluster Analysis  
GWAS tools  
Text Manipulation  
Collection Operations  
Filter and Sort  
Join, Subtract and Group  
Statistics  
Graph/Display Data  
VCF intersect  
HTPG TOOLS  
Pre-genotyping  
Post-genotyping  
TOOLS FROM TOOLSHED  
SNIPLAY  
GOBII TOOLS  
Flapjack Tools  
Workflows  
All workflows



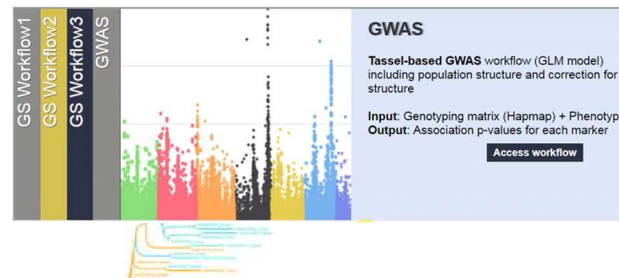
Excellence  
in Breeding  
PLATFORM

### Welcome to Galaxy!

This instance is dedicated to host tools for the [GOBII](#) project, [HTPG](#) project and [EiB](#) platform.

Take an interactive tour: [Galaxy UI](#) | [History](#) | [Scratchbook](#)

Try our  
Genomic Selection workflows



History  
Rechercher des données  
demo video  
42 selected, 28 hidden  
(empty)  
Cet historique est vide. You can  
Charger vos propres données or  
Charger des données depuis  
une source externe

# Genome Wide Association Analysis workflow demo (GWAS)



Galaxy / EiB-demo

Analyze Data Workflow Shared Data Visualization Help User

Published Pages | dereeper | GWAS workflow

## GOBii Genomic Selection Galaxy Workflows

### GWAS workflow

This workflow is based on TASSEL v5 software for Genome Wide Association Study (GWAS) analysis, using GLM model with a correction by population structure performed by sNMF software.

The screenshot displays the Galaxy web interface for a GWAS workflow. On the left, a sidebar lists various tools categorized under 'Genomic Selection', 'Phenotyping', and 'Computation'. The main workspace shows a 'Welcome to Galaxy!' message and a 'Try our Genomic Selection workflows' section with three options: 'GS Workflow1', 'GS Workflow2', and 'GS Workflow3' (labeled 'GWAS'). The 'GS Workflow3' option is highlighted. On the right, a 'History' panel shows a list of datasets, including 'tutorial video gwas' and '92.12 MB'. The central part of the interface displays a workflow diagram with several steps: 'Convert genotyping data', 'Input genotyping file', 'output (tab)', 'Trait file', 'output', 'sNMF', 'VCF file', 'best\_k\_output (txt)', 'best\_k\_groups (txt)', 'best\_k\_logfile (txt)', 'outputs (txt)', 'logs (txt)', 'Tassel', 'HapMap file', 'Trait file', 'Structure file', 'output1 (txt, png)', 'output2 (txt)', 'output3 (txt)', and 'log (txt)'. The workflow is connected by yellow lines, indicating the flow of data between steps.

# GOBii GS-Galaxy pipeline

<http://galaxy-demo.excellenceinbreeding.org/>



- Free and open to the public
- Teaching and training platform to the next generation molecular breeders
- GS testing team collaboration playground to share datasets, history, workflows, and parameters
- Data exploration for samples genotyped but not phenotyped vice versa and study planning- **Sample matching**
- Handling bioinformatics and file **conversions-encoding, filtering, format conversion**
- Phenotype analysis and visualization- **BLUP/BLUE calculator**
- Batch analysis with reproducible parameters for multiple traits and variables, factors- **BLUP/BLUE and GEBV calculators**
- Sub-structure analysis- **Clustering and diversity analysis tools**
- Well established customer-defined GS and GWAS workflows and training methods- **Workflows/Pipelines**

## GOBii GS-Galaxy pipeline

<http://galaxy-demo.excellenceinbreeding.org/>

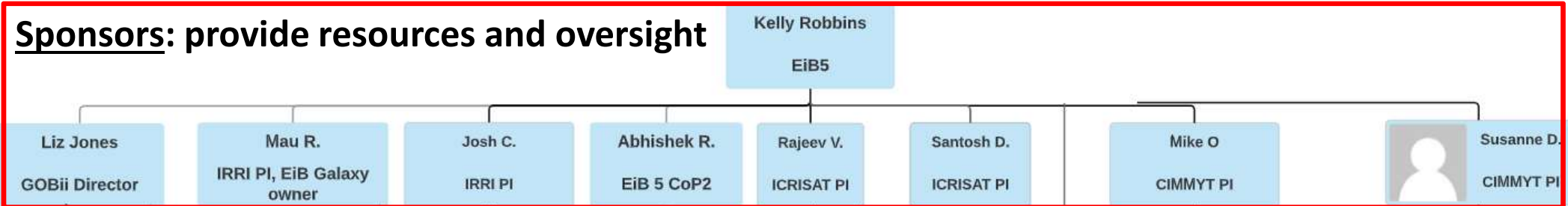


### ARE NOT

- GS statisticians' model development, who are proficient with command-line implementation
- Free testing site of large datasets-need to download tools and workflows onto your own server
- For producing publication without a sound understanding or data interpretation

# Acknowledgements

## Sponsors: provide resources and oversight



## Project management team: Lead requirements, testing, training, and adoption



## Development team: develop tools and local administration support



## EiB contractor: optimize and integrate tools, set up workflows and visualization, enable toolshed and system admin training

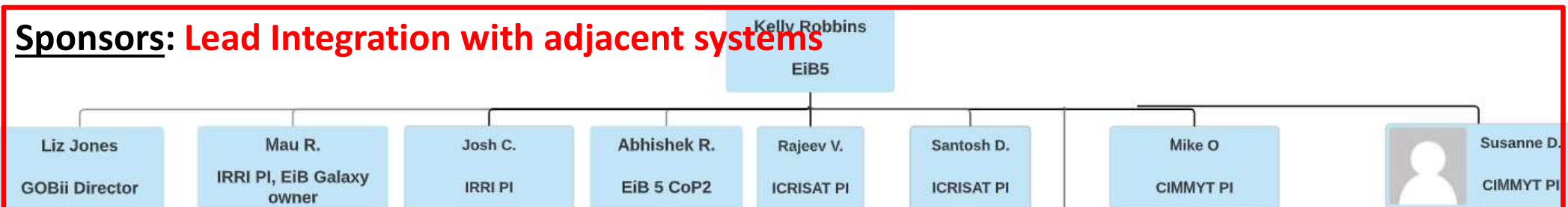
**Alexis DEREPPER**

## Contributors:

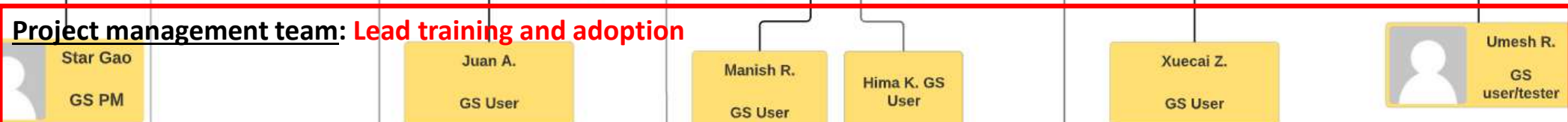
**Paulino Perez, Pancho Jose Crossa**, Fernando Toledo, Juan BURGUEÑO, Jessica Rutkoski, Demya Chebotarov, Jean-Luc Jannink, Isaak Yosief Tecle, Nicholas Santantonio, **Kate Dreher**, Claudio Ayala, Félix SanVicente, Mark Sorrells, **Yoseph Beyene**, Manje Gowda, Yaw Nti-Adde

# What's next?

## Sponsors: Lead Integration with adjacent systems



## Project management team: Lead training and adoption



## Development team: Deployment and local administration support



EiB contractor(s): Toolshed, server, and system admin training

Alexis DEREPPER

Community-based development and support

Contact us at <http://gobiiproject.org> for collaboration, training, and customization