Galaxy for Genomics-enabled Breeding

Star Yanxin Gao

yg28@cornell.edu

Introduction

Star Yanxin Gao, Ph.D., PMP

- **Application Specialist, IT**
  - Enterprise Breeding System (2020+)
  - GOBii (2015-Present)
- **Breeding and Genetics**
  - Corn, DAS (2008-2015)
  - Soybean, VT (2004-2008)
  - Vegetables, Cornell (1999-2004)

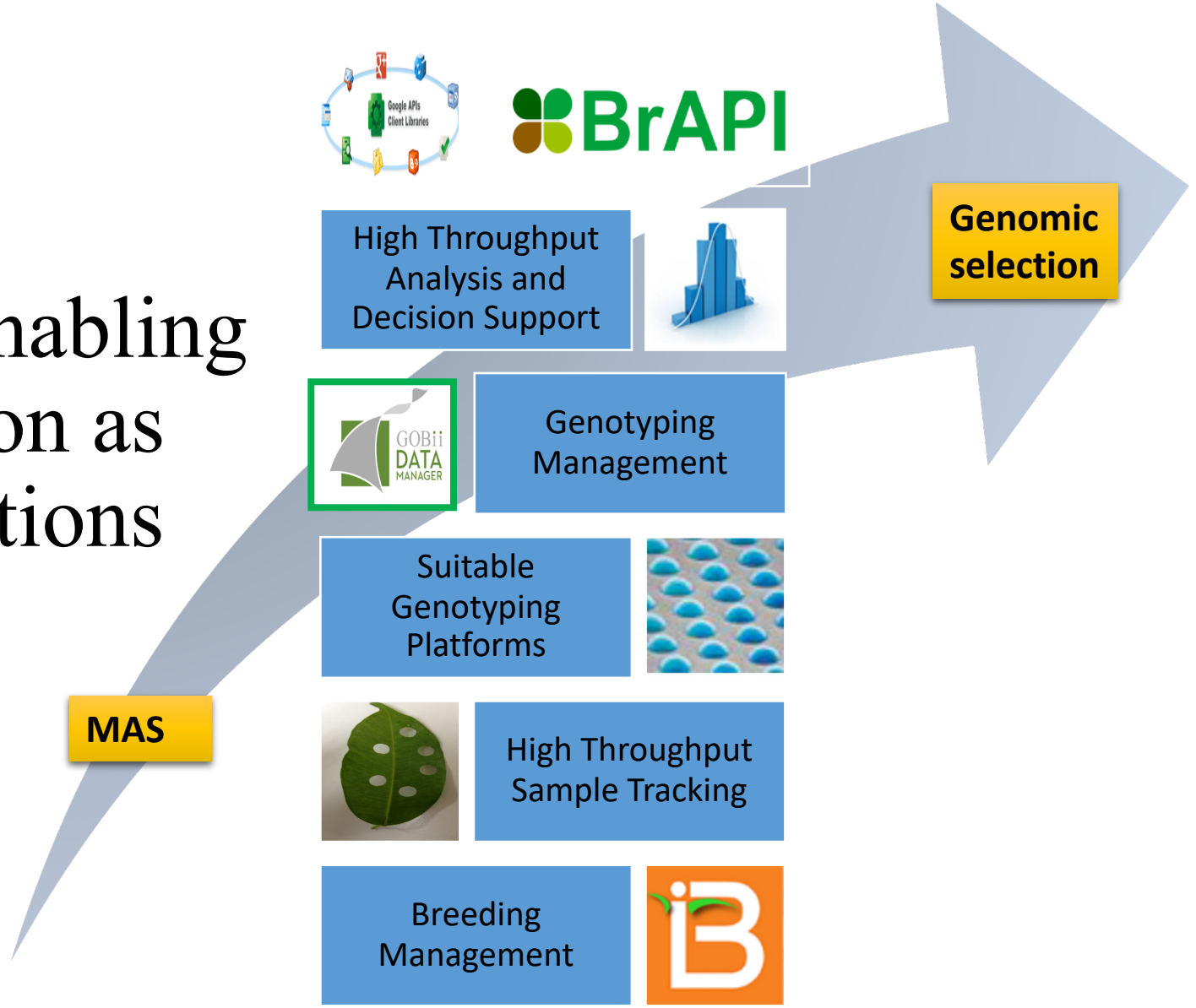
Outline

1. Project
2. Products
3. Partnership

GS-Galaxy Project (1st *P*)

GOBii Mission

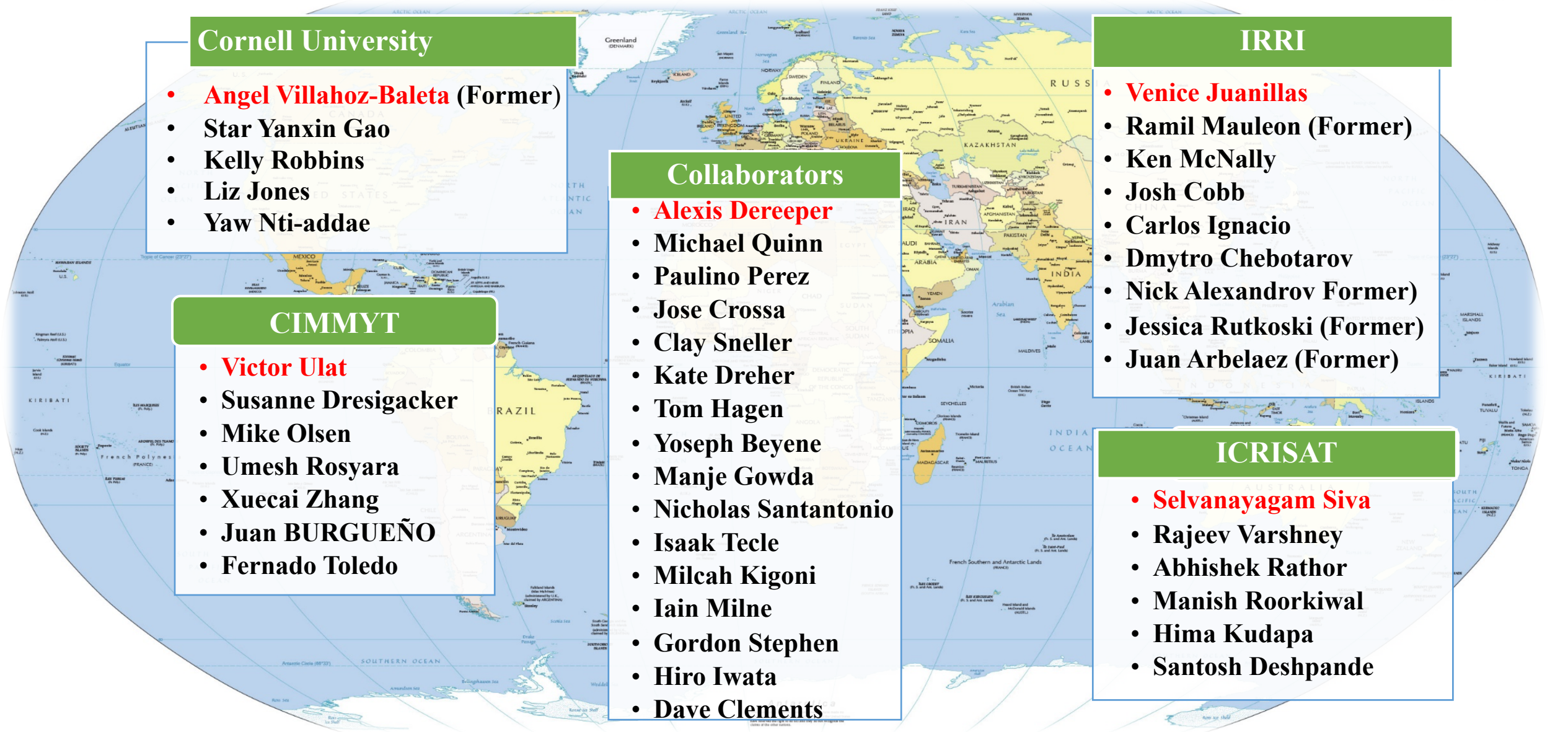
Transform breeding by enabling genomic–assisted selection as **routine** breeding applications



Genomic-Assisted Selection Platform

GS-Galaxy Project (1st *P*)

People



Cornell University

- Angel Villahoz-Baleta (Former)
- Star Yanxin Gao
- Kelly Robbins
- Liz Jones
- Yaw Nti-addae

CIMMYT

- Victor Ulat
- Susanne Dresigacker
- Mike Olsen
- Umesh Rosyara
- Xuecai Zhang
- Juan BURGUEÑO
- Fernado Toledo

Collaborators

- Alexis Dereeper
- Michael Quinn
- Paulino Perez
- Jose Crossa
- Clay Sneller
- Kate Dreher
- Tom Hagen
- Yoseph Beyene
- Manje Gowda
- Nicholas Santantonio
- Isaak Tecle
- Milcah Kigoni
- Iain Milne
- Gordon Stephen
- Hiro Iwata
- Dave Clements

IRRI

- Venice Juanillas
- Ramil Mauleon (Former)
- Ken McNally
- Josh Cobb
- Carlos Ignacio
- Dmytro Chebotarov
- Nick Alexandrov Former)
- Jessica Rutkoski (Former)
- Juan Arbelaez (Former)

ICRISAT

- Selvanayagam Siva
- Rajeev Varshney
- Abhishek Rathor
- Manish Roorkiwal
- Hima Kudapa
- Santosh Deshpande

GS-Galaxy Project (1st *P*)

Milestones

1st milestone: Minimum set of GS tools installed in Galaxy- June 2018:

- GS workflow mapped for each crop and common minimum desirable tools and features identified.
- Pipeline developed with minimum desirable features.
- Basic and common tool components put in place.
- Tested with well curated test datasets for each crop by product owners and testers

2nd milestone: Production server with published GS workflows-June 2019

- Customized workflows for each crop.
- v1 for fully functioning GS pipeline.
- Working pipeline available to centers.
- Training and workshop to users other than product owners and more users start using pipeline and tools

3rd milestone: GOBII integrated as data source for Galaxy-June 2020

- Deliver a complete pipeline with no data manipulation required.
- Access data extract for phenotypes and genotypes from Galaxy pipeline using remote Galaxy web servers.
- Found a solution for storing the output data long term so it can easily be accessed.
- Pipelines widely used within CG and outside CG in future.

GS-Galaxy Products (2nd *P*)

Galaxy / EiB-demo

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

Genomic Selection

- BLUP/BLUE calculator (multiple traits)
- Hapmap filter Filter per SNPs based on user-specified criteria
- Numeric matrix encoder encodes genotypes to numeric
- Sample Matching compares samples between genotype and phenotype files
- Parallel Genotype-Phenotype Matching genotype and phenotype files
- Parallel Genotype-Phenotype Matching genotype and phenotype files
- GEBV calculator
- Cross validation within and across groups

Cluster Analysis

- Calculate Distance Matrix from Genotype Data
- SNPRelate GRM calculator
- Get Distance Matrix from Genotype Data
- Principal Component Analysis
- Hybrid K-Means Clustering
- K-Means Clustering

GWAS tools

- Calculate Kinship from Genotype Data
- Run PCA (Principal Component Analysis) on genetic data
- Association using GLM (General Linear Model)
- Association using MLM (Mixed Linear Model)

Imputation

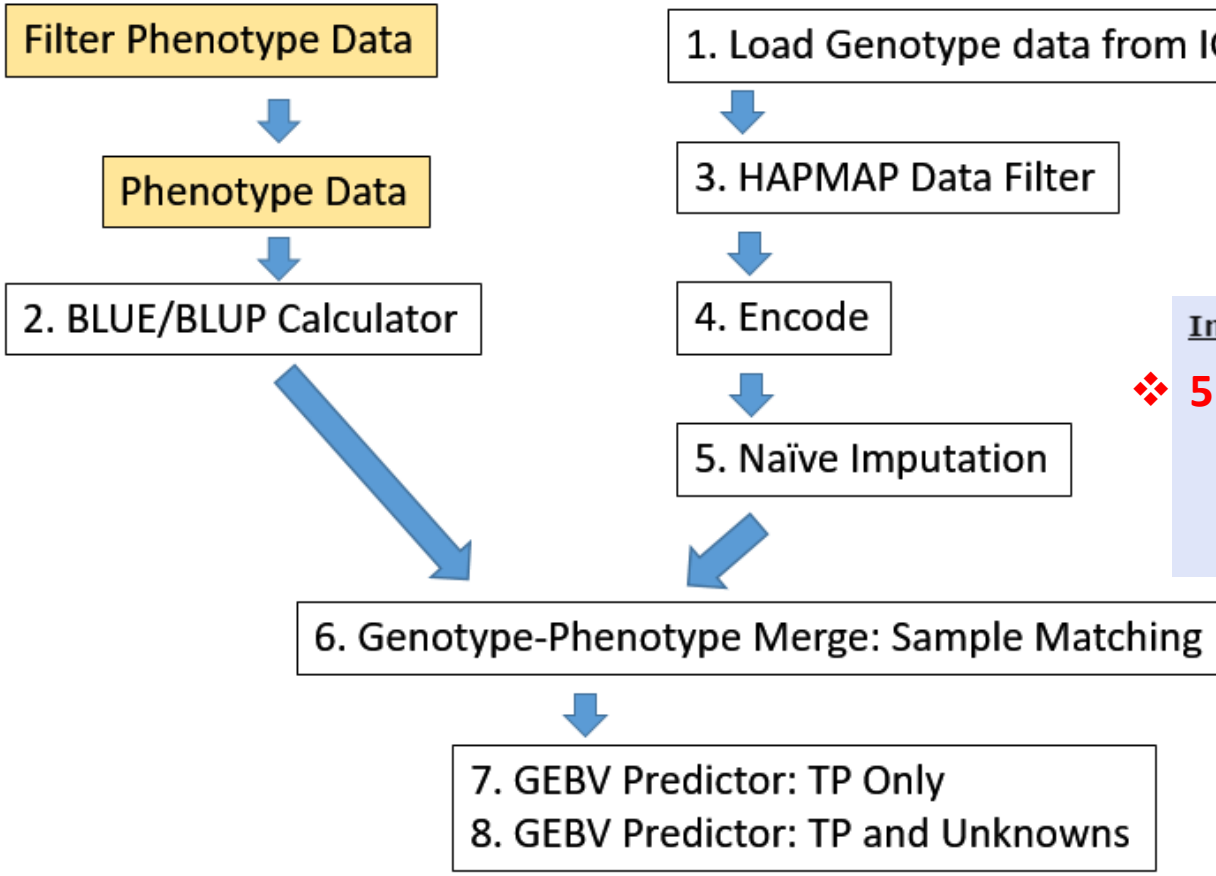
- Naive imputation using population mean or mode
- Beagle imputation
- Impute2 imputation

GALAXY

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

GS-Galaxy Products (2nd *P*)

Enable routine genomic selection analysis



Imputation

❖ 5 Naive imputation using population mean or mode

Beagle imputation

Impute2 imputation

Genomic Selection

❖ 2 BLUP/BLUE calculator (multiple traits)

❖ 3 Hapmap filter Filter per SNPs based on user-specified criteria

❖ 4 Numeric matrix encoder encodes genotypes to numeric

❖ 6 Sample Matching compares samples between genotype and phenotype files

Parallel Genotype-Phenotype Matching genotype and phenotype files

Parallel Genotype-Phenotype Matching genotype and phenotype files

❖ 7 GEBV calculator

❖ 8 Cross validation within and across groups

Adopted Slide from Clay Sneller

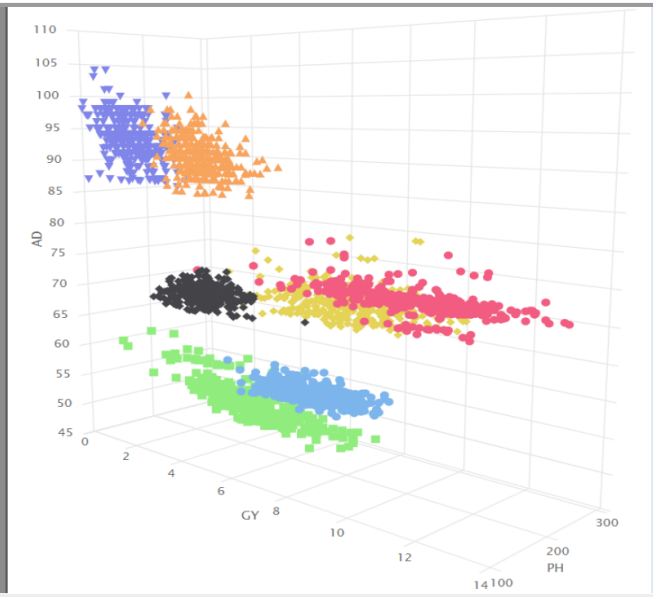
http://galaxy-demo.excellenceinbreeding.org/

GS-Galaxy Products (2nd *P*)

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

Try our
Genomic Selection workflows

GS Workflow1



Workflow1

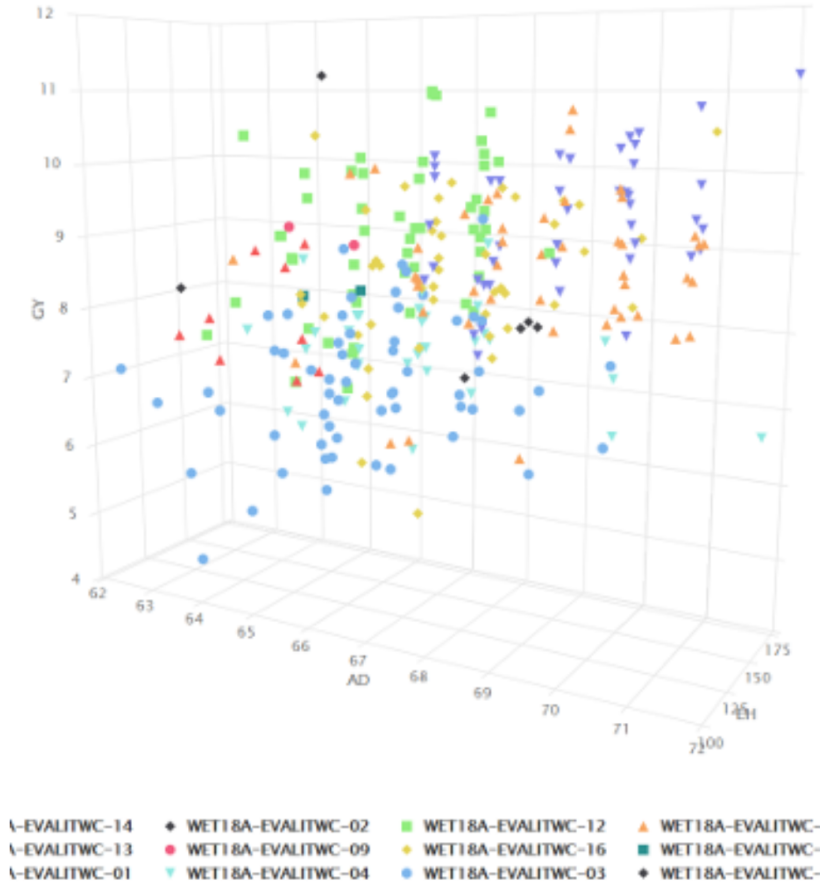
Prediction of genomic breeding values
Well defined training and prediction datasets, prediction within groups.
The prediction is based on the **GEBV** (Genomic Breeding Value Estimator) method implemented in BGLR R package.

Input: Genotyping matrix + Phenotyping file
Output: Table of Predicted values for each trait

Access workflow

GS Workflow2
GS Workflow3
GWAS

Examples of graphical outputs



Workflow 1: Predict GEBVs in untested lines

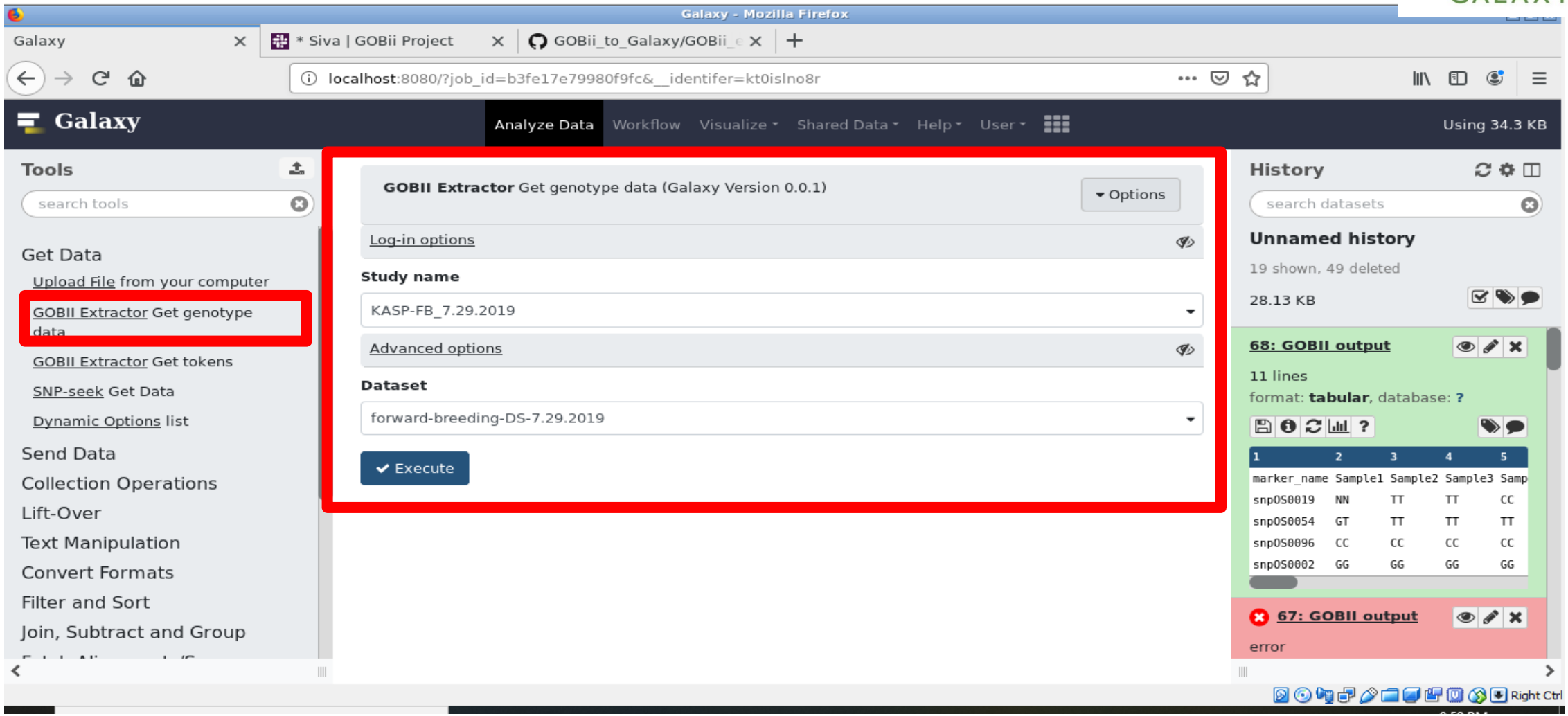
Workflow 2: Clustering/population structure

Workflow 3: Cross validation

Workflow 4: Genome-wide association study

GS-Galaxy Products (2nd *P*)

GOBii-Galaxy Integration



GS-Galaxy Partnership (3rd *P*)

Partnership Models

Adopt products

- Open-source and free
- Test with own datasets and use cases
- Download from toolshed

Community development

- Implement new tools
- Develop crop-specific workflows
- Connect with databases

GS-Galaxy Partnership (3rd *P*)

Adoption Metrics

- Number of registered users: **214**
- Number of different crops/programs: **3** (rice, wheat, maize)
- Number of analyses/jobs performed: **14200**
- Total job file sizes stored: **148 Gb**
- Average file size: **4.2982 Mb**

GS-Galaxy Partnership (3rd *P*)

BrAPI

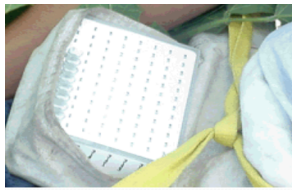
Breeding Management

B4R

PHENOAPPS

Sample Tracking

Sample Management



Service Tracking

Vendors

intertek

Diversity Arrays Technology Pty Ltd

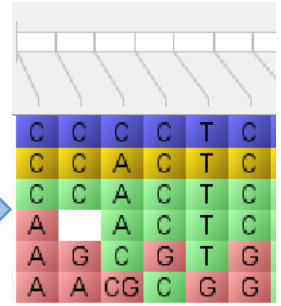
Data Loading

Genotyping Management

GOBii DATA MANAGER

Data Extract

Analytics



GALAXY

Decision Support

EBS ENTERPRISE BREEDING SYSTEM

Breeding Data Extract

Invite to Partner with GS-Galaxy

1. Project
2. Products
3. Partnership

EiB Demo Server:

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

Publish tools to Galaxy main toolshed

Galaxy workshop at Cornell/BTI April, 2020


GS
GALAXY


GOBii


Excellence
in Breeding
PLATFORM

Thanks


Galaxy
PROJECT


BILL & MELINDA
GATES foundation


CORNELL UNIVERSITY
FOUNDED A.D. 1865


BTI
BOYCE
THOMPSON
INSTITUTE


USDA


CIMMYT
International Maize and Wheat
Improvement Center


ICRISAT


IRRI


The James
Hutton
Institute


Diversity
Arrays
Technology
IT'S A BULLSEYE WITH DArT