Bioinformatics tools for identifying traits in PGR

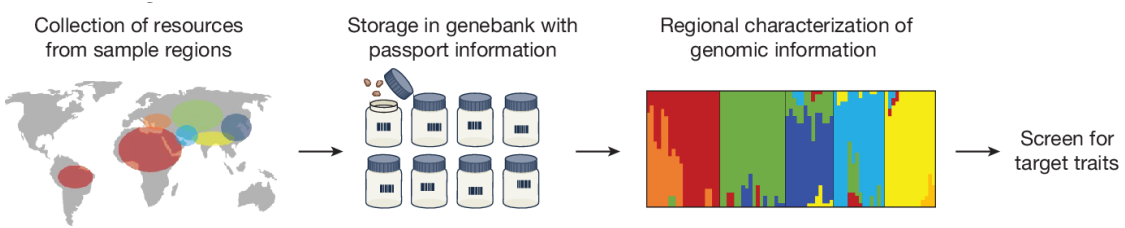
Luciana Gaccione

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

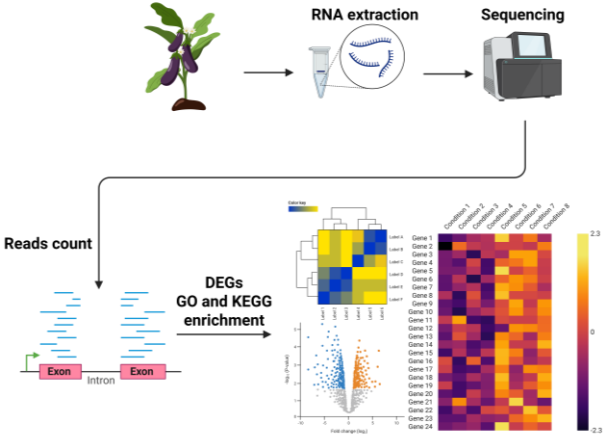
UNIVERSITÀ DI TORINO

Unlocks the diversity in PGRs

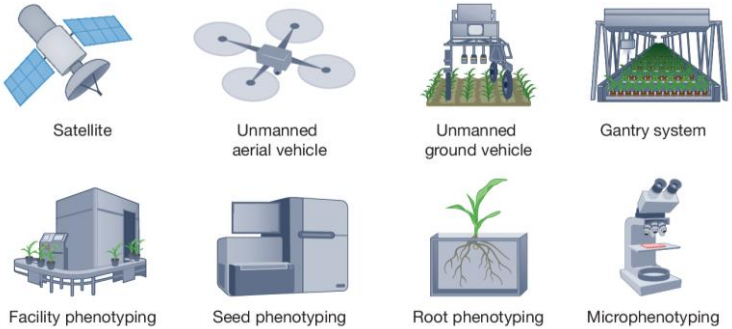
Genomics

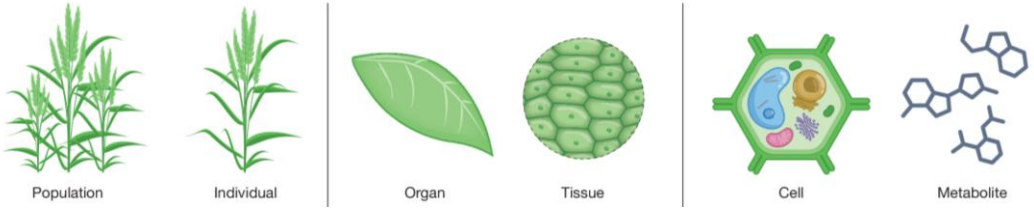


Transcriptomics

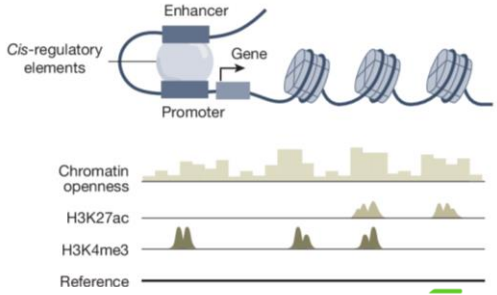


Phenomics - Metabolomics



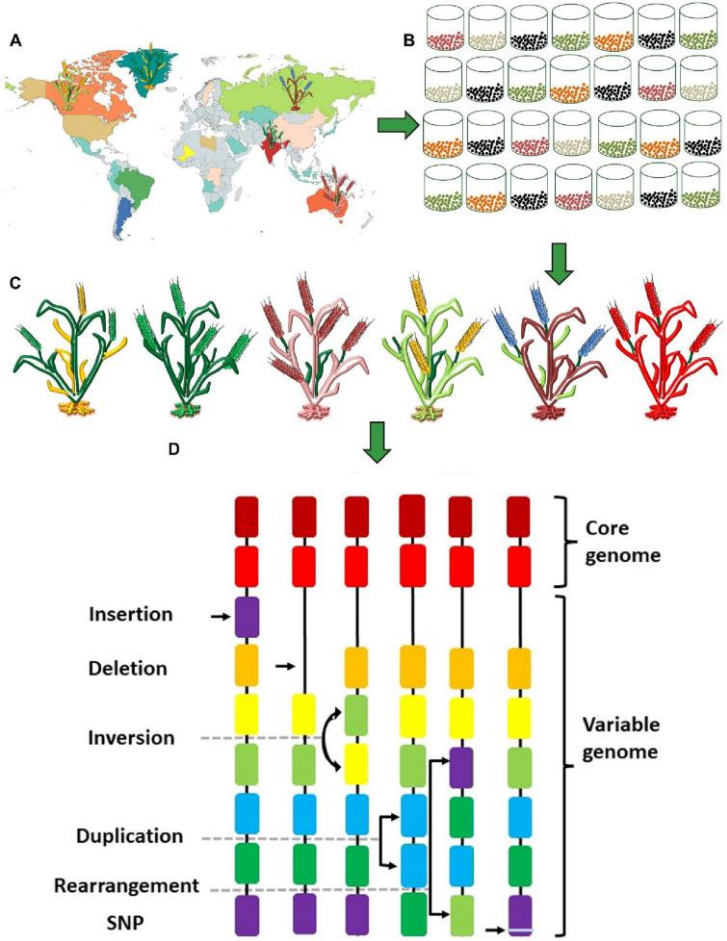


Epigenomics



PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Unlocks the diversity in PGRs



Omics technologies represent a leap forward for the conservation, management and characterization of PGRs.

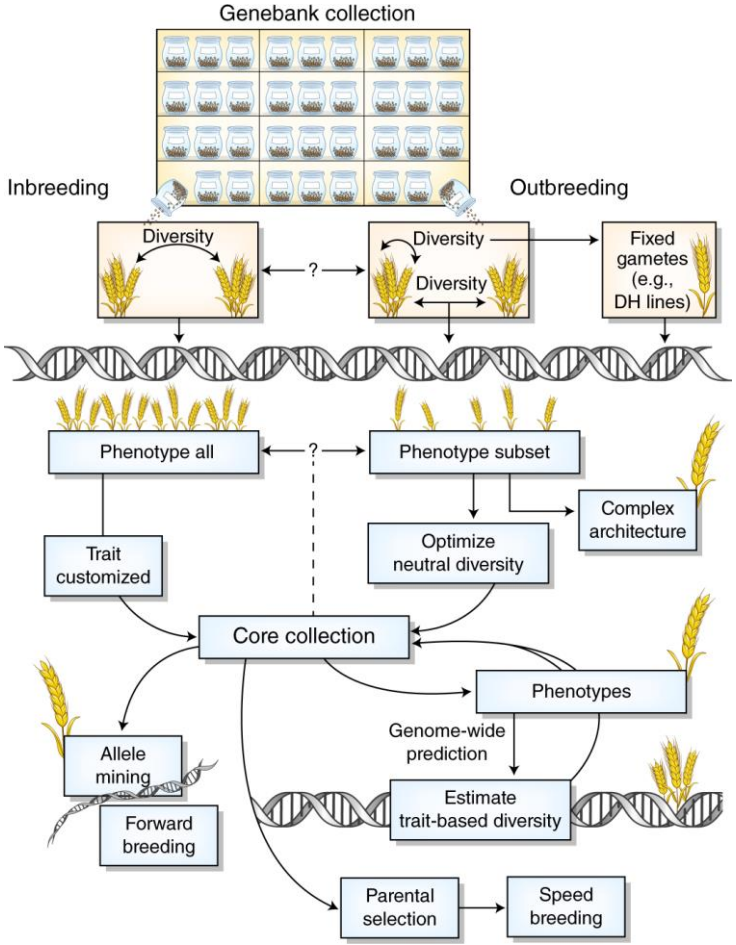
Current PGR management does not involve the routine use of omic tools to trace accessions during seed regeneration or vegetative propagation and identifying traits.

A gap between what is technically possible with modern omics, and what is actually implemented in routine PGRs conservation practices

Bioinformatics pipelines can help bridge the gap

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Informed selection for PGRs pre-breeding programs



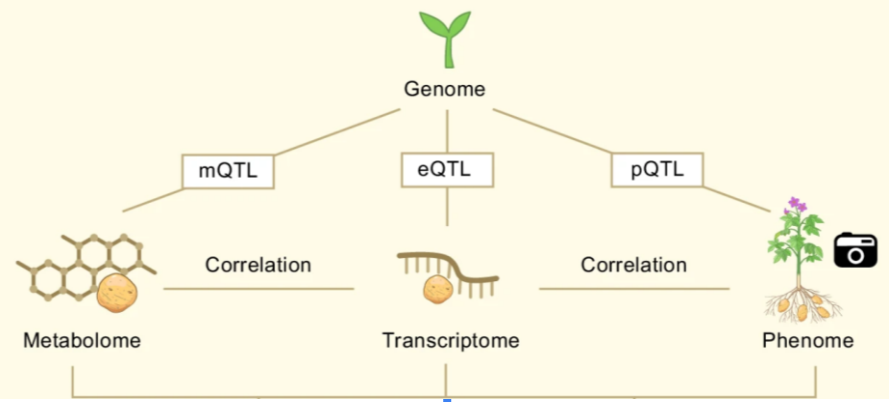
Genotypic and phenotypic information enables the informed selection of the most promising genetic resources.

Depending on the genetic architectures of the traits, **entire collections** can be phenotyped, or **core collections** maximizing genetic diversity.

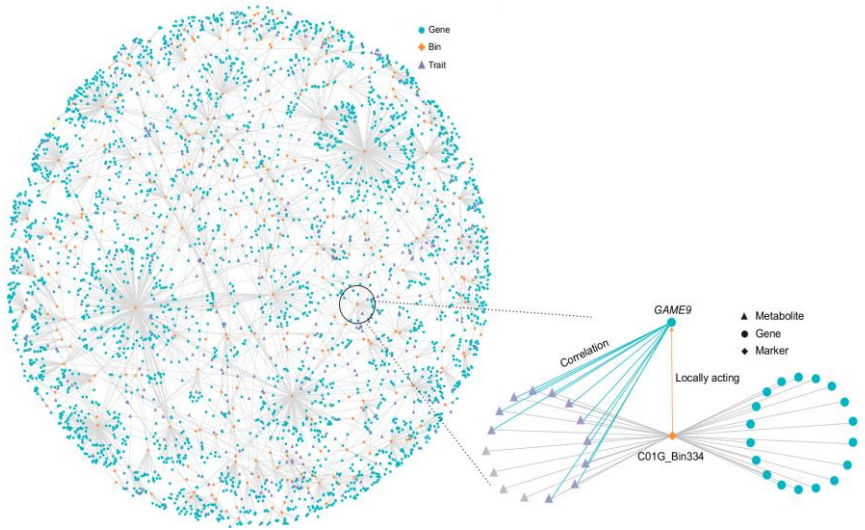
Genotypes with the highest breeding values enter pre-breeding programs and are used in genetic studies to elucidate the molecular basis of beneficial traits of PGRs.

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Integrated multi-omics is more than the sum of its parts



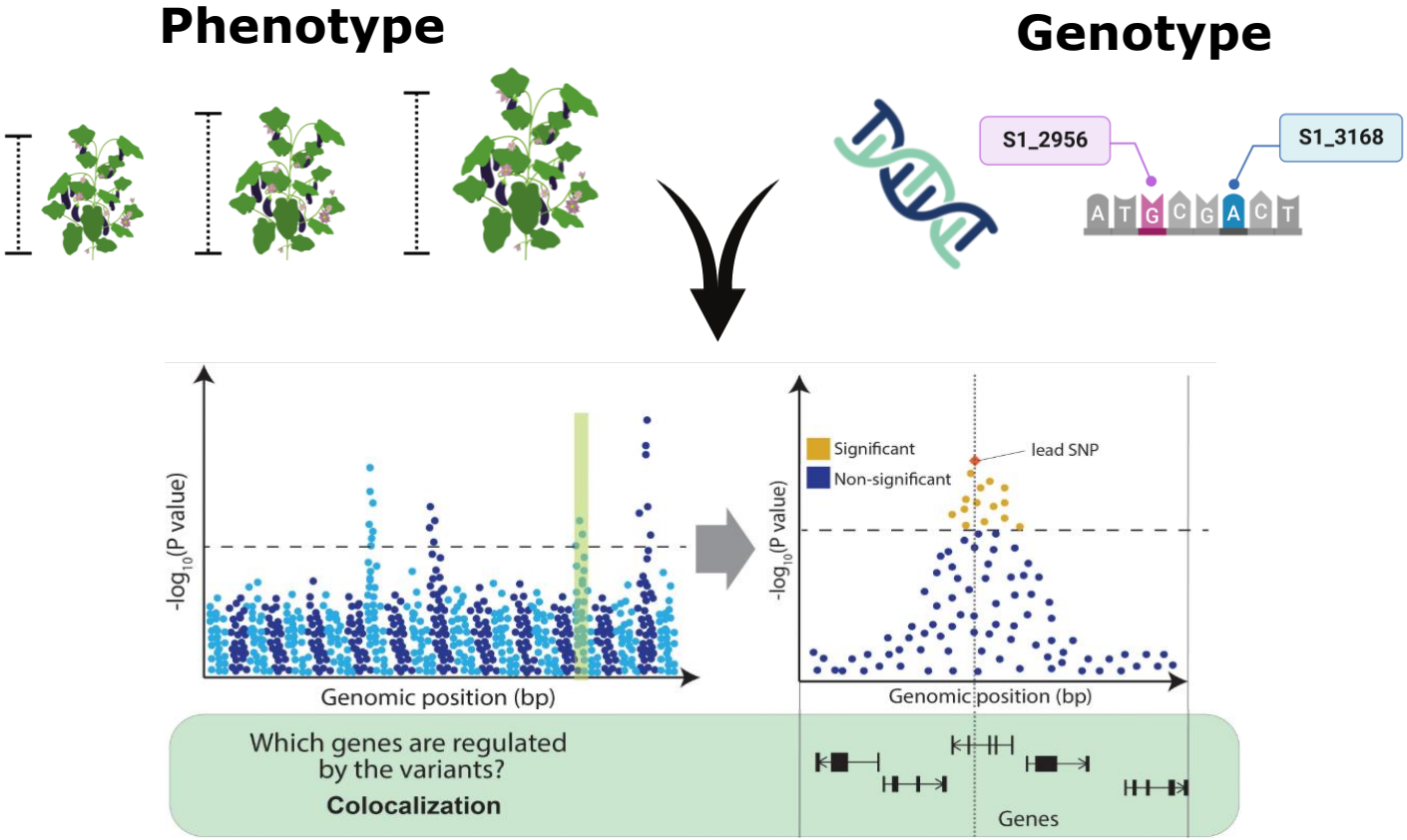
Multi-omics refers to the integrated study of various "omics" fields, including genomics, transcriptomics, proteomics, metabolomics, and epigenomics.



This approach is particularly valuable for unraveling intricate **networks and pathways associated with PGRs traits**

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Traditional GWA approaches



PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Linkage disequilibrium decay

What size should my sliding window be to define QTLs?

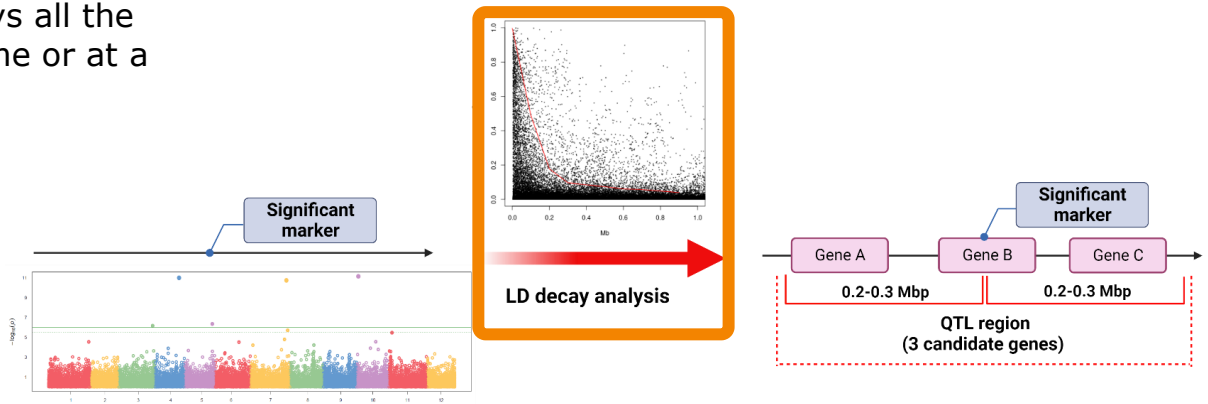
A simple rule of thumb is to set the window where linkage disequilibrium decays to the genome-wide background

Linkage equilibrium: haplotype frequencies in a population have the same value that they would have if the genes at each locus were combined at random.

Linkage disequilibrium: Non-random association of alleles at different loci in a given population

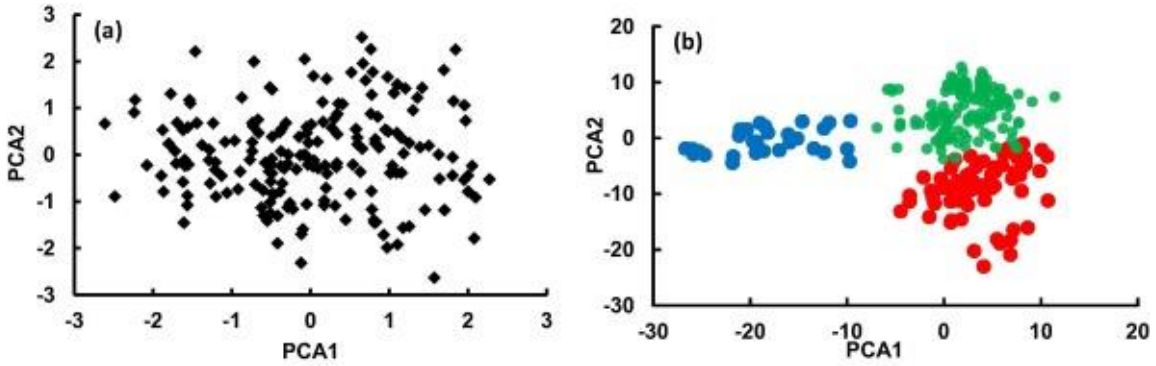
How LD decay is calculated?

1. **Get R^2 value** for each marker vs all the others (on the same chromosome or at a certain distance) using **plink**
2. **Use PopLDdecay**



Population Structure and spurious associations

- The problem arises when cases disproportionally represent a genetic subgroup, **resulting (apparently) associated with the trait**.
- Geographic or growth habit, etc. lead to having a specific genetic variation and an effect on the end-use of association analysis



Main methods:

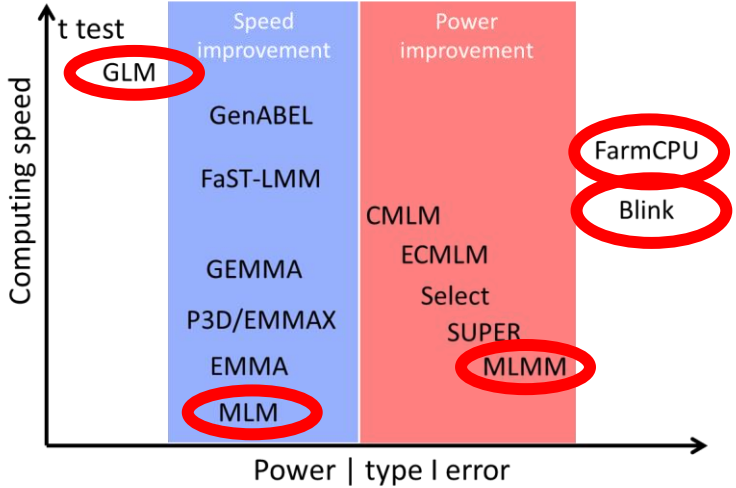
- Principal component analysis (PCA)
- Structured association (SA)
- Kinship analysis

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

GWA models

Multi-locus mixed models

- **MLMM** (Multi-Locus Mixed Model) is an extension of the standard mixed linear model (MLM) used in GWAS. The goal is to better handle **polygenic traits** by including **multiple associated loci** as cofactors, rather than testing SNPs one by one. The most significant SNPs as **pseudo-QTNs** (covariates), then removes non-significant ones.

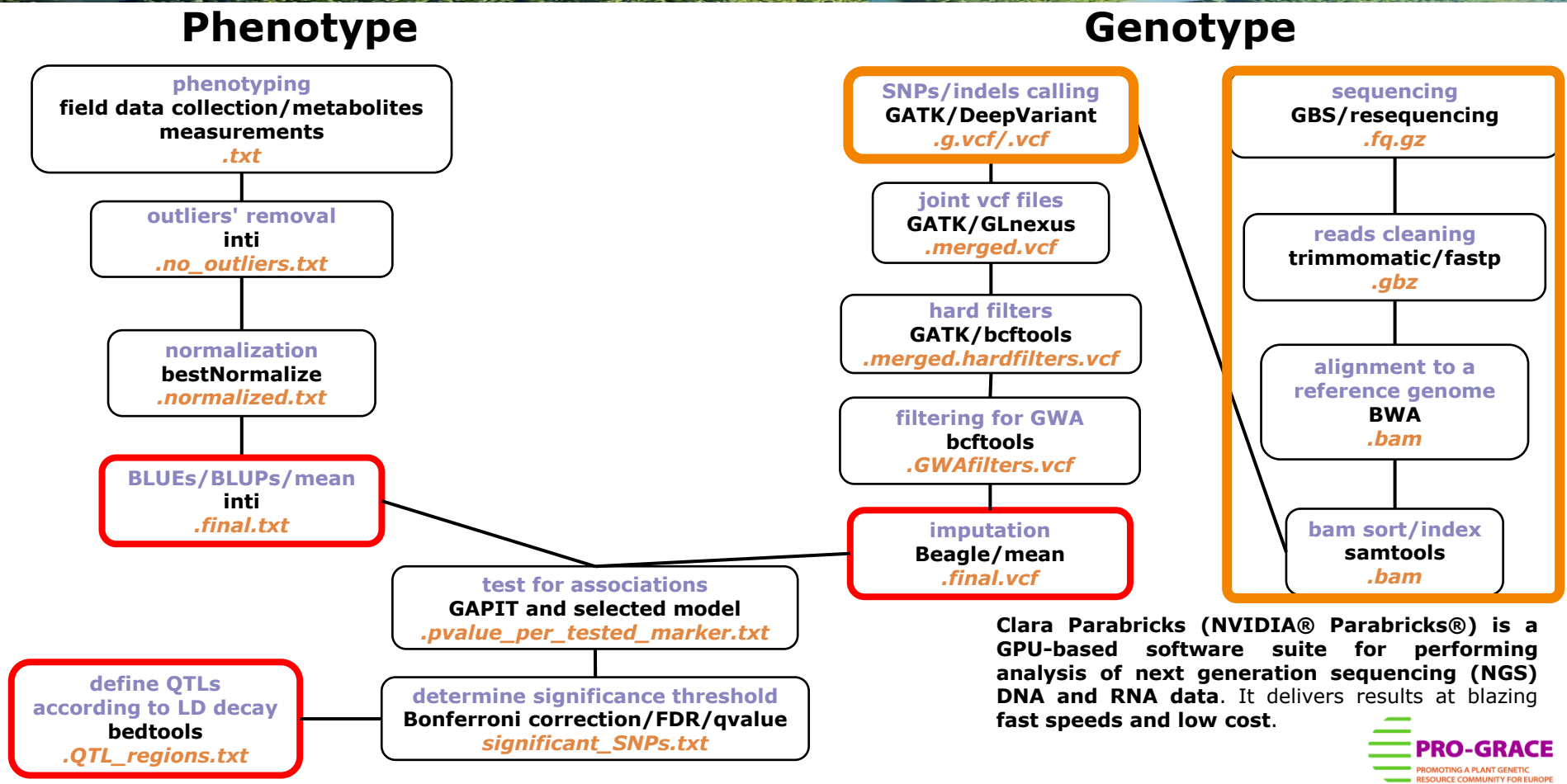


- **FarmCPU** (Fixed and Random Model Circulating Probability Unification) **separates marker testing (fixed effect model) from population structure/kinship control (random effect model)**, iteratively updating pseudo-QTNs to reduce confounding. Iteratively updates both parts, so pseudo-QTNs and tested SNPs keep adjusting each other.

- **BLINK** (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway) improves over FarmCPU by **replacing the random effect with LD-based clustering** and BIC model selection, making it faster and more powerful for large datasets.

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

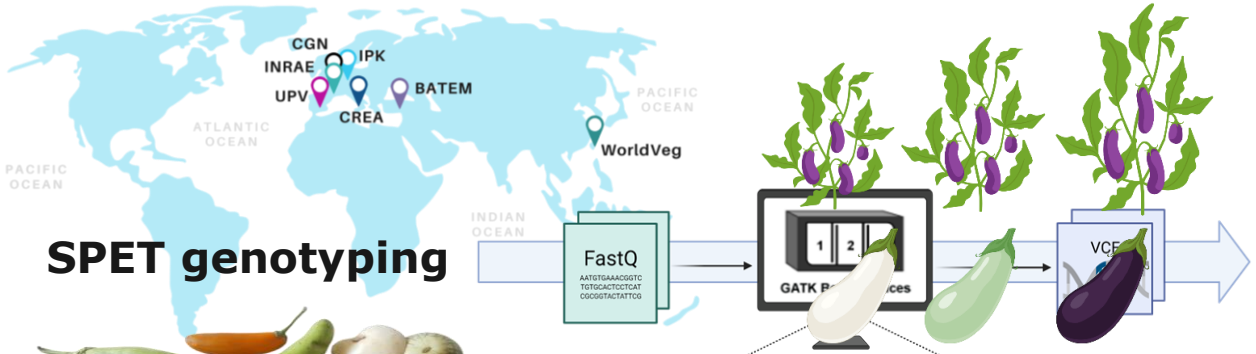
Traditional GWA workflow



Clara Parabricks (NVIDIA® Parabricks®) is a GPU-based software suite for performing analysis of next generation sequencing (NGS) DNA and RNA data. It delivers results at blazing **fast speeds and low cost**.

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

GWA: Case study - SPET data and historical GWAS in eggplant



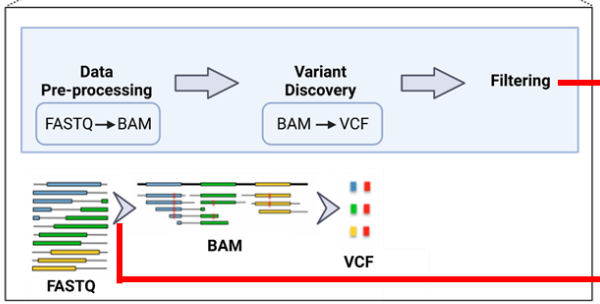
SPET genotyping

Historical phenotypic data collected by genebanks:

23 qualitative/pseudoqualitative descriptors for plant, flower, and fruit trait categories

Total number of target SNPs: 4.3K

Total number of SNPs: 120K



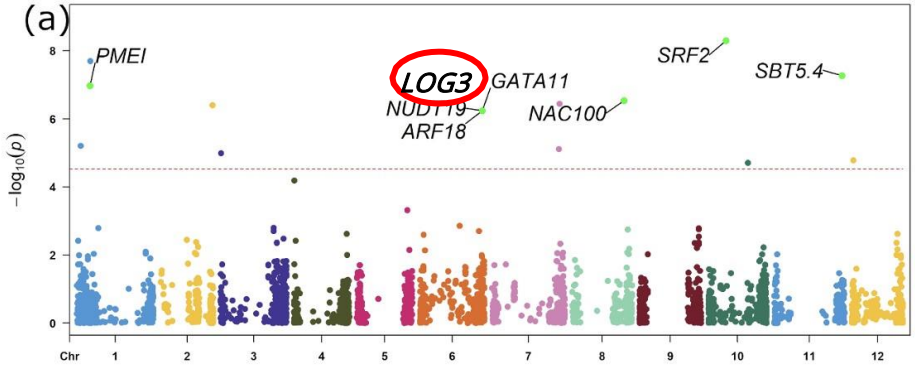
SNPs filter parameters:
Mean DP 15
Max missing count (at SNP level) 5%

Reference line '67/3' (GPE001970)
Version 4.1 (Barchi et al., 2021)

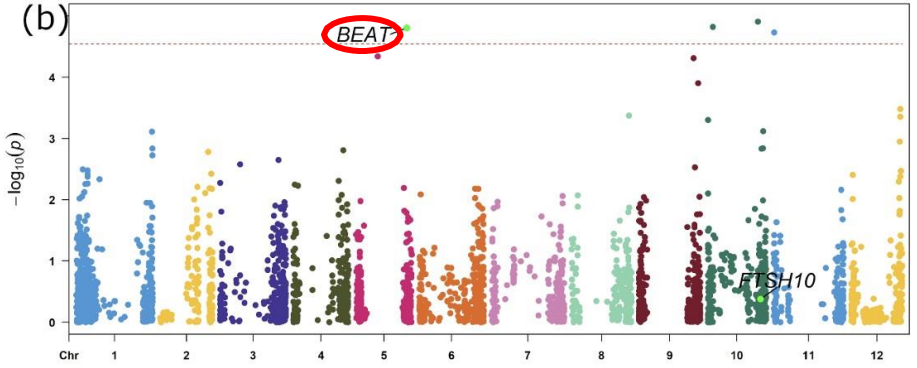
PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

GWA: Case study – SPET data and historical GWAS in eggplant

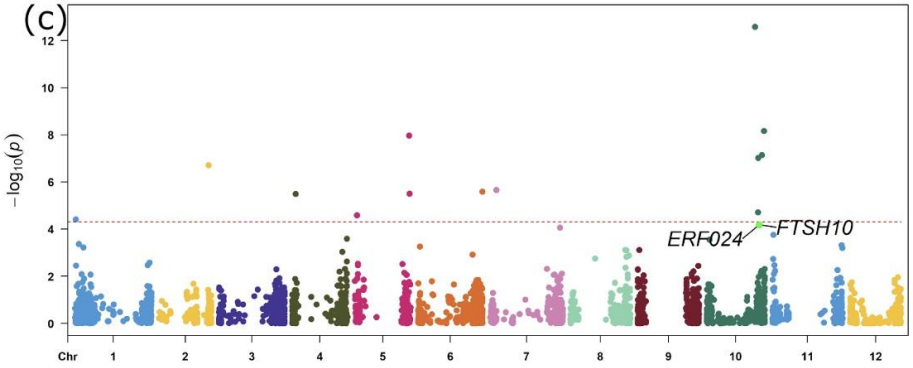
Leaf prickles



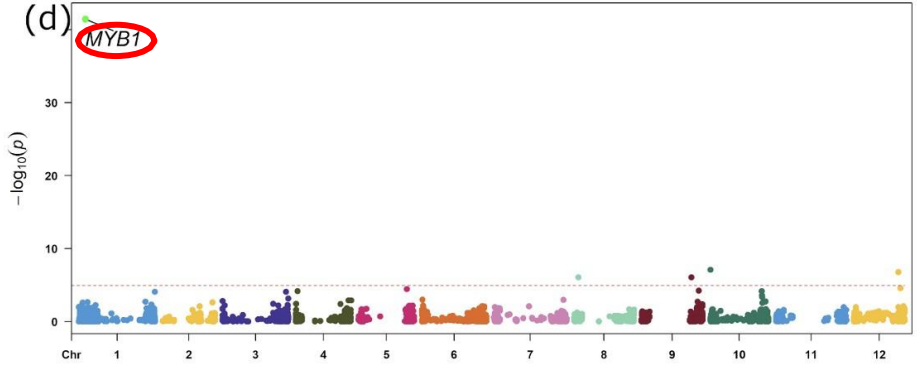
Corolla color



Fruit color



Anthocyanin distribution

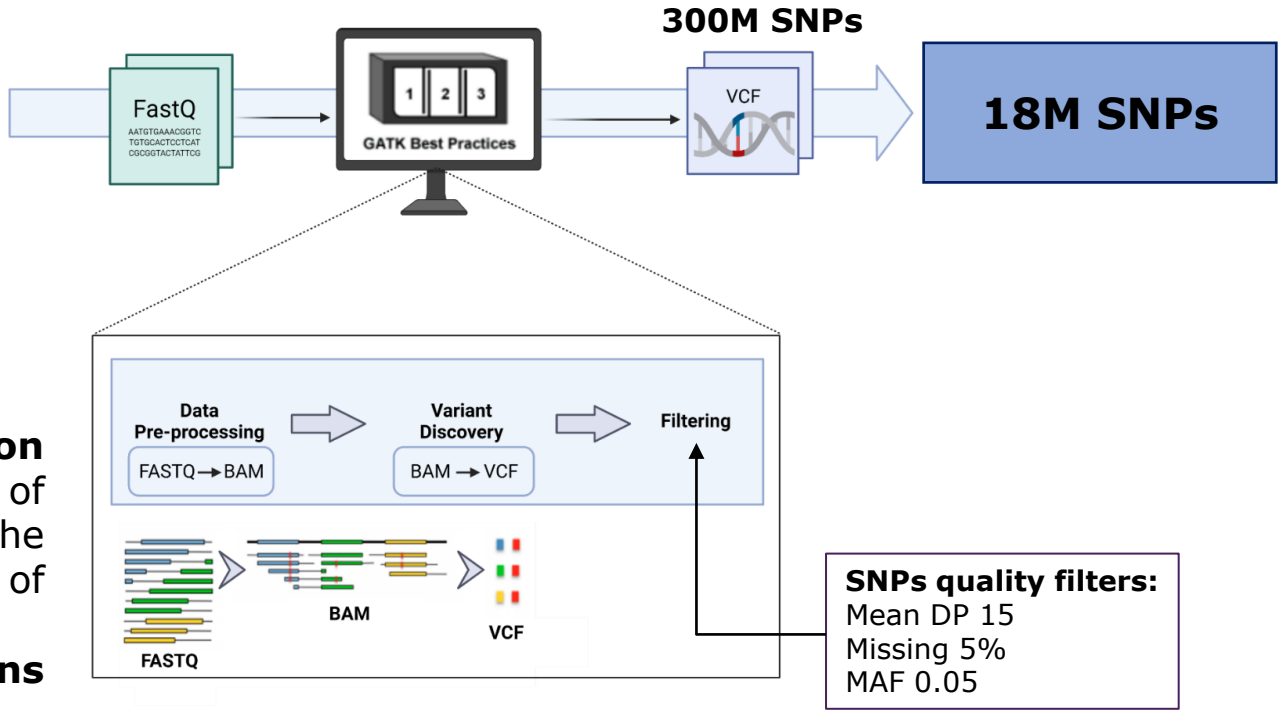


Candidate genes circled in red have been confirmed by different models, confirming the quality of genebank data

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

GWA: Case study – Resequencing and GWA in pepper

- The **G2P-SOL core collection of *Capsicum* spp.**, consisting of 423 accessions representing the genetic variability of a panel of 10,083 accessions.
- **393 *C. annuum* accessions used for GWAS**



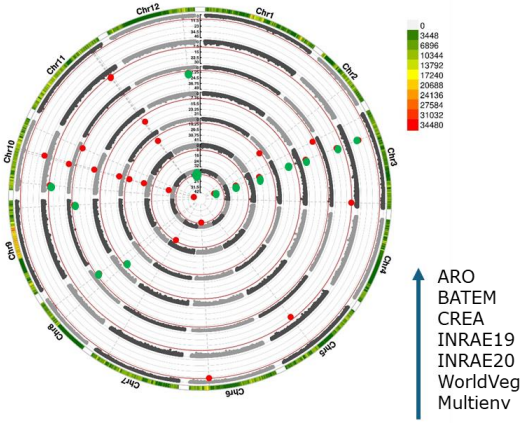
The genome sequence of the *C. annuum* var. *annuum* Zhangshugang was used as reference

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

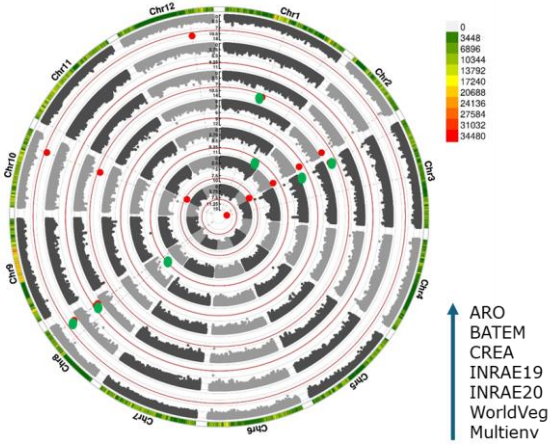
GWA: Case study – Pepper: Agronomic traits

- **207 significant markers** based on pepper descriptors from the International Plant Genetic Resources Institute (IPGRI)
- **112 QTL regions** based on the linkage disequilibrium (LD) decay
- **Up to 6 independent field trials (2–4 different years and locations**: ARO in Israel (2020), BATEM in Türkiye (2020–2021), CREA in Italy (2019), INRAE in France (2019–2020), WorldVeg in Taiwan (2020–2021)

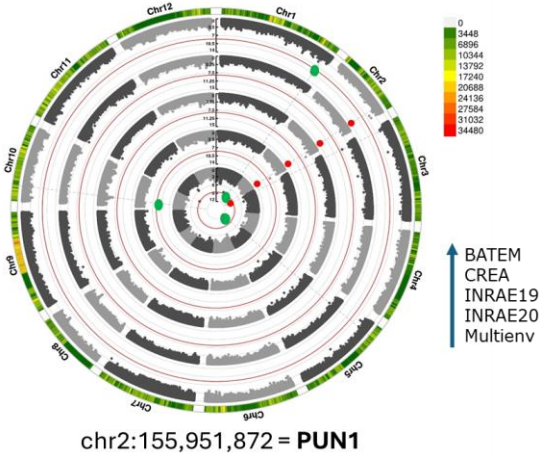
Fruit shape index (FShI)



Fruit weight (FWe)



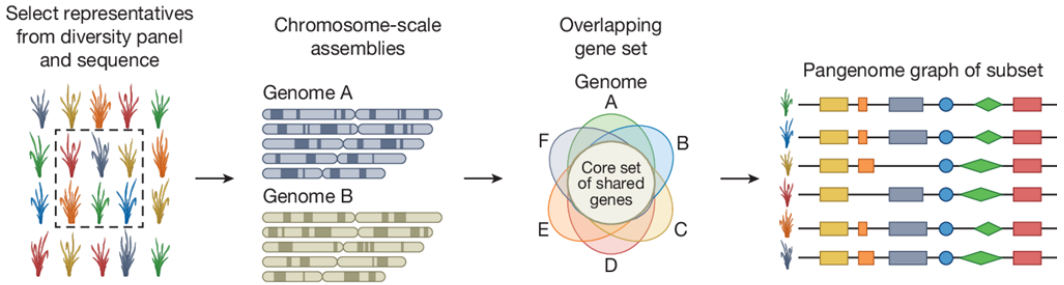
Fruit pungency (FP)

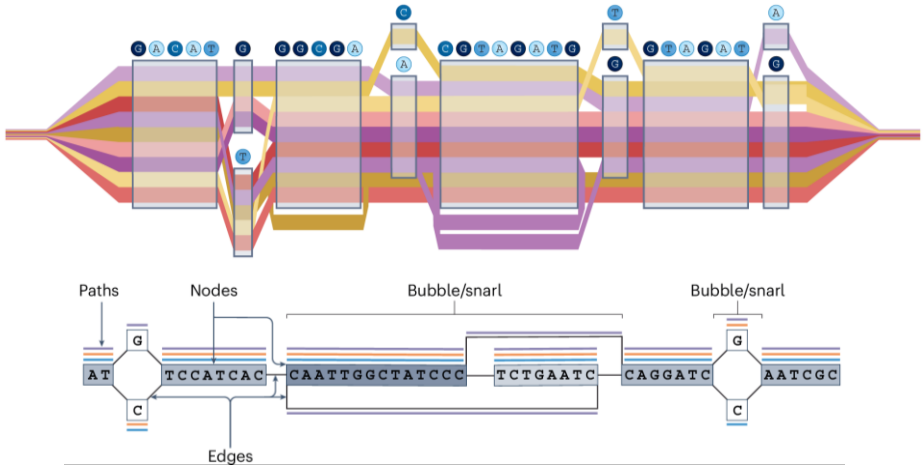


chr2:155,951,872 = **PUN1**

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

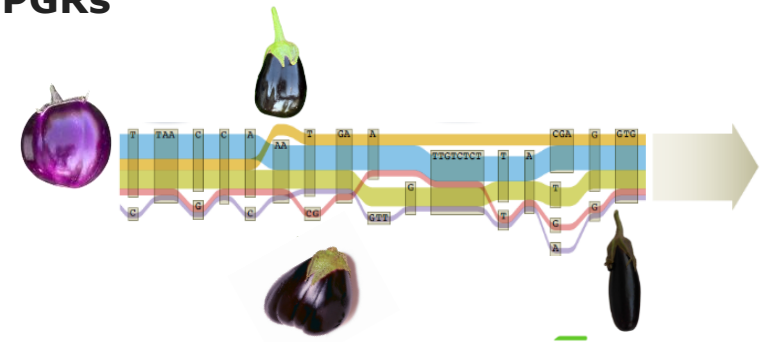
Pangenomics and Pangenome-wide association studies (Pan-GWA)





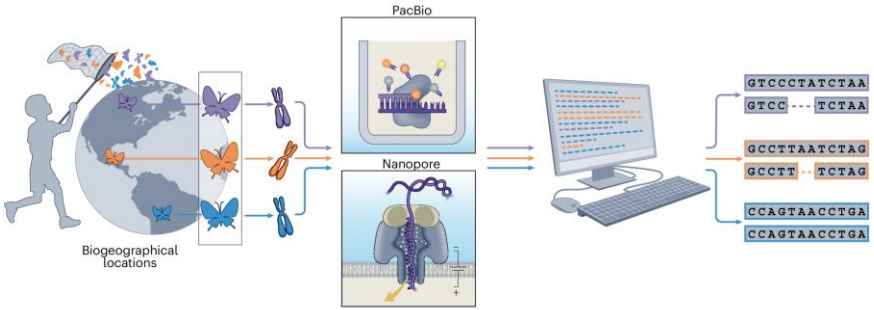
Pangenomics and pangenome graphs have emerged as transformative tools, reshaping the way we study genetic diversity and uncover the complexity of genomes across populations.

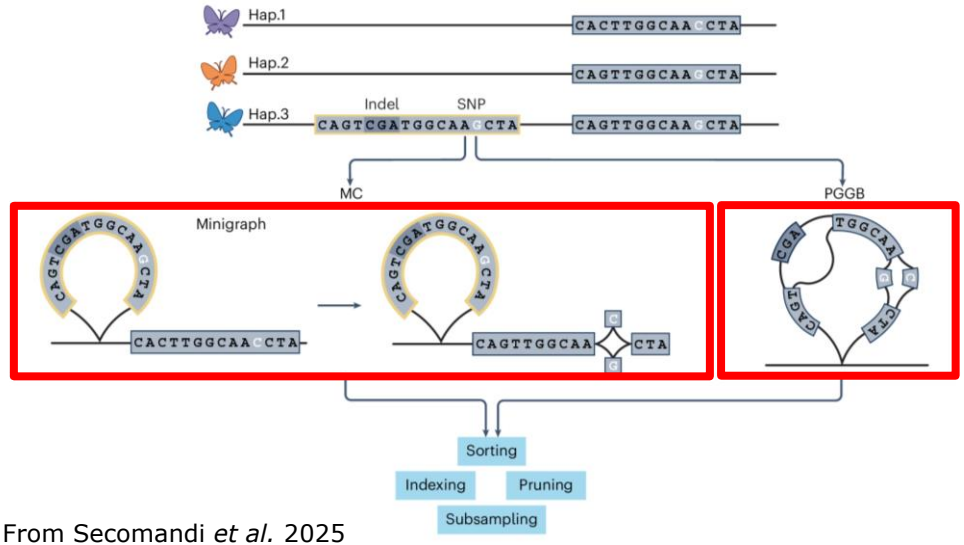
There is a need of **new tools and pipeline to handle graph structures** for **pangenome-wide associations analysis for identifying traits in PGRs**



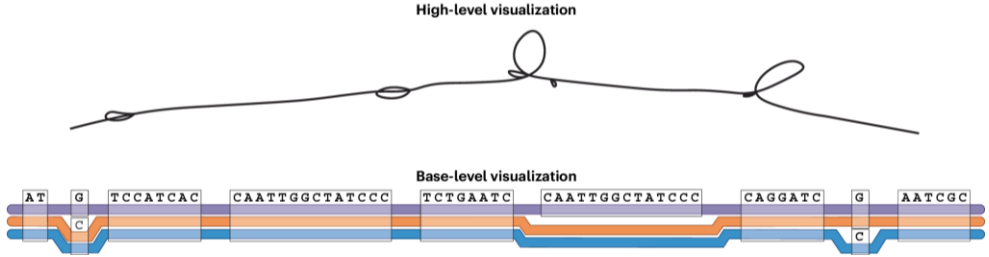
PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Pangenome graphs building pipeline





From Secomandi *et al.* 2025



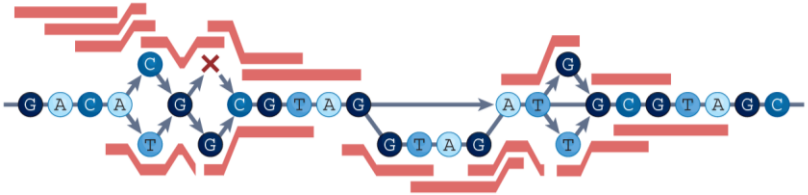
	PG-SMA	*PG-SM*
Tool	**PGGB**	**Minigraph-Cactus**
Length	2,934,734,179	2,540,649,469
Node count	141,826,052	33,186,357
Edge count	194,712,053	45,701,072
Average degree	2.75	2.75
Minimum degree	1	1
Maximum degree	1,073	29
Paths count	480	396
Steps count	3,473,585,945	635,992,212

Nodes maximum degree are the number of edges connected to a node.
The higher the maximum degree, the more complex the graph structure, indicating highly connected nodes.

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Pangenome-wide association studies (Pan-GWAS)

graph
Minigraph-Cactus/PGGB
.gfa



gbz indexing
VG index
.gbz

map
VG giraffe
.gam

convert to bam
VG surject
.bam

graph snarls
VG snarls
.snarls

support
VG pack
.pack

filtering
VG filter
.filtered.gam

phasing
Whatshap phase
.phased.vcf

merge vcf files
bcftools merge
.merged.vcf

variants
VG call
.vcf

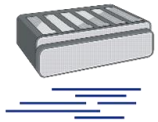
SNPs/SVs GWA filtering
bcftools
.SNPs.vcf/.SVs.vcf

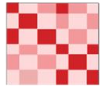
PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Pan-GWAS: Case study - Eggplant

S. melongena* pangenome graph *"PG-SM"

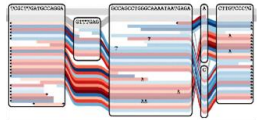
Oxford Nanopore sequencing





Assembly + Scaffolding with Hi-C data (Yahs and Juicebox) or **reference-based** (RagTag)

33 *S. melongena* chromosome-scale genome



Align short reads to the graph (vg giraffe)

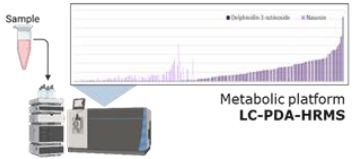
46 agronomic traits

10 biotic/abiotic stress

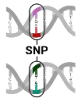
Pangenome graph construction (Minigraph-Cactus)

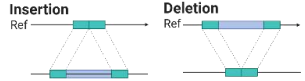
S. melongena pangenome graph

162 semi-polar fruit metabolites



Variant calling (vg call)

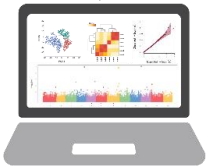




GWA analyses

Multi-environment approach

CREA (Italy) UPV (Spain) BATEM (Turkey)



GAPIT

Genomic Association and Prediction Integrated Tool

(Version 3)

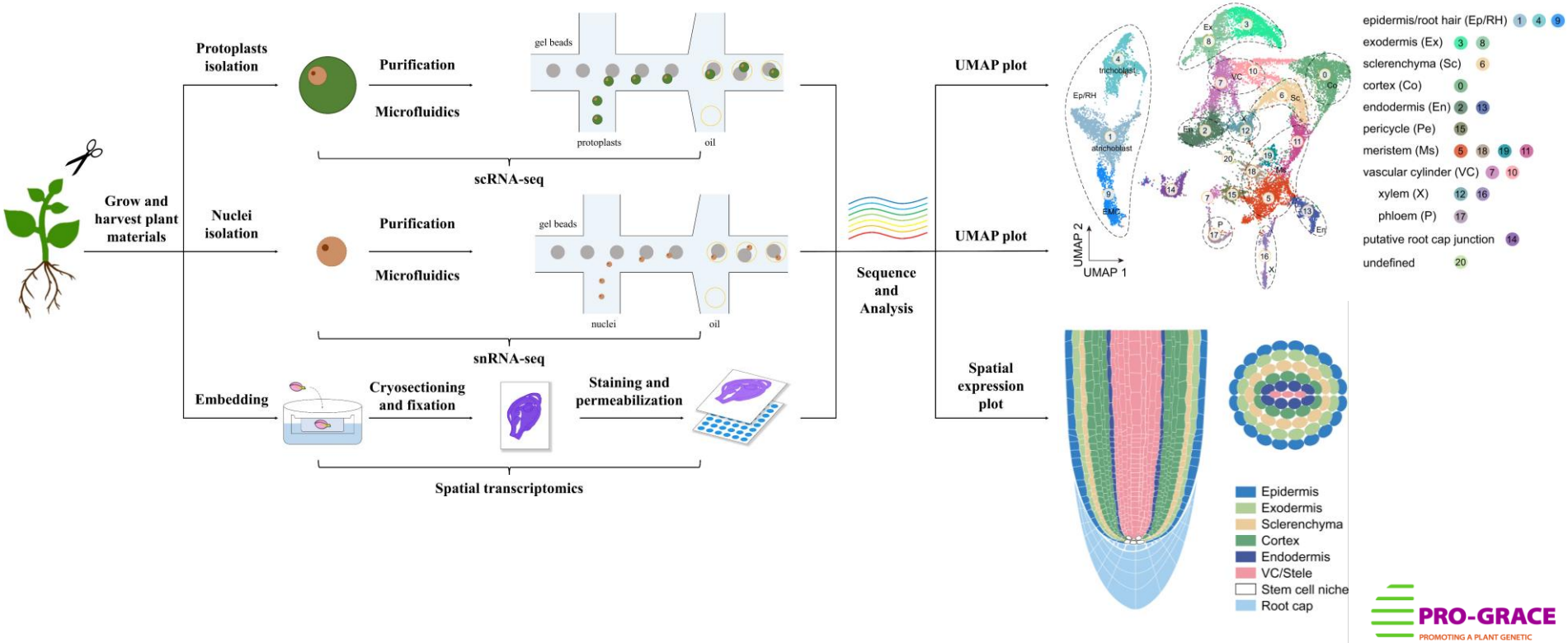
Blink model

PRO-GRACE

PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Single cell and spatial omics

Single cell omics allow to study **cell type-specific transcriptomic** and **chromatin-accessible landscapes**

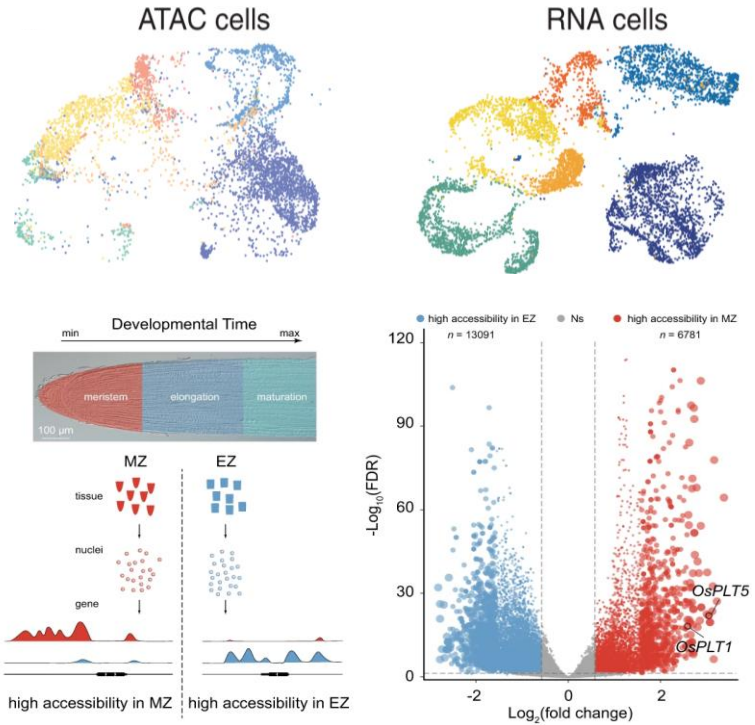


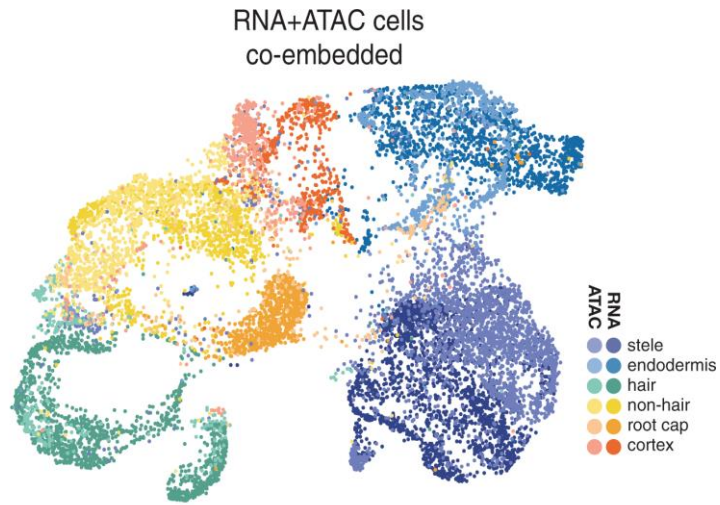
PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Single cell and spatial omics

Integration of single cell omics possible with different tools:

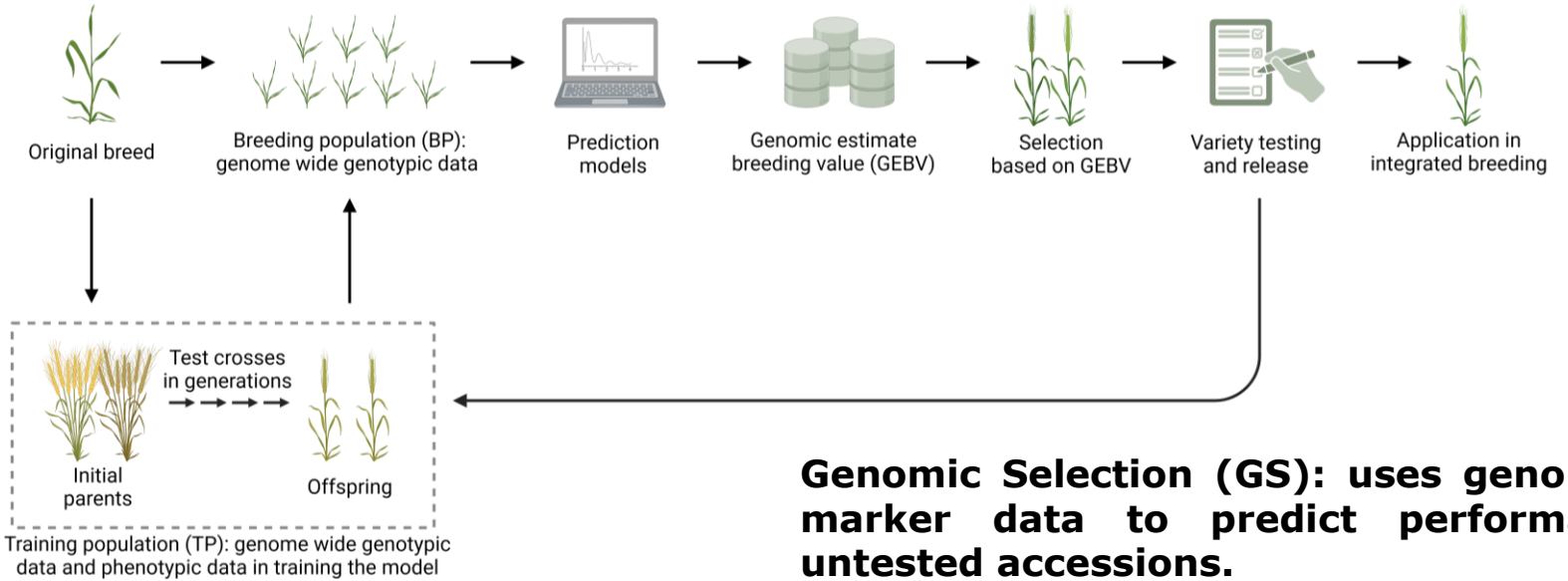
- **Signac** and **Seurat** R packages
- **Sanpy** - Single-Cell Analysis in Python





Beyond GWAS: Unlocking the Power of Genomic Selection

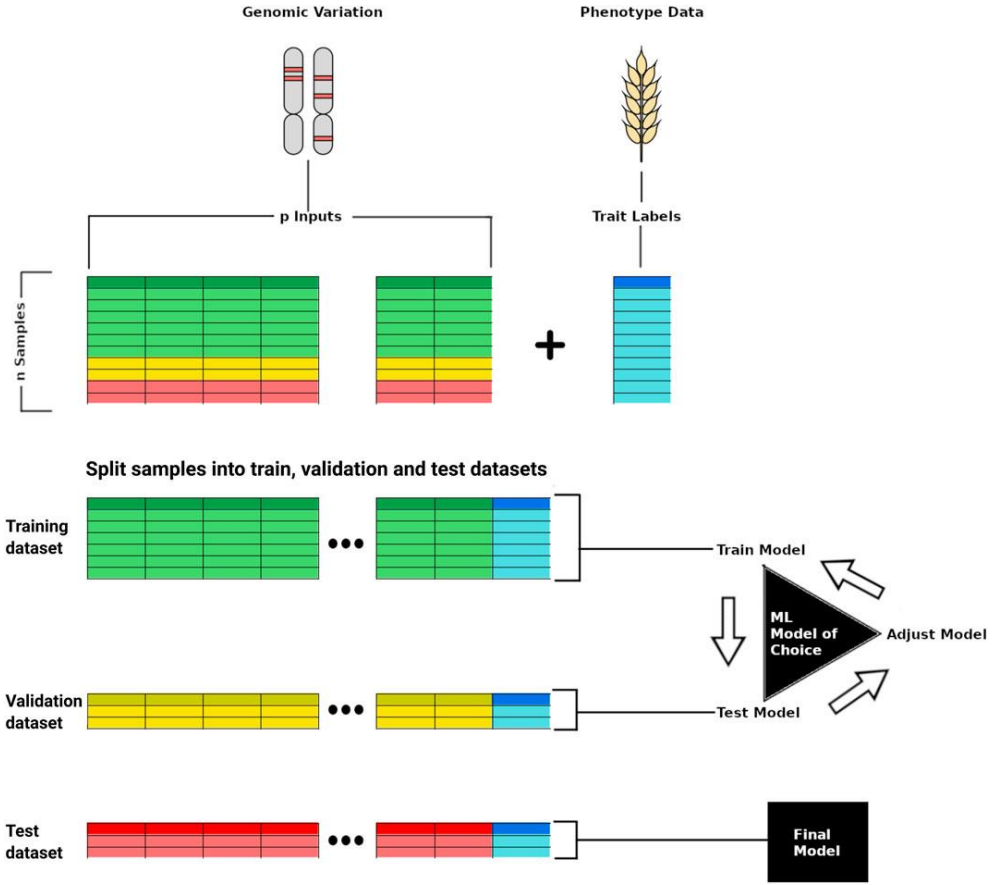
- GWAS and Pan-GWAS identify trait–marker associations across diverse accessions.
- Single-cell omics helps us understand in which cell types and contexts those variants act.
- **Translating markers into predictive breeding** power requires new strategies.



Genomic Selection (GS): uses genome-wide marker data to predict performance of untested accessions.

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Genomic Selection Workflow in Practice

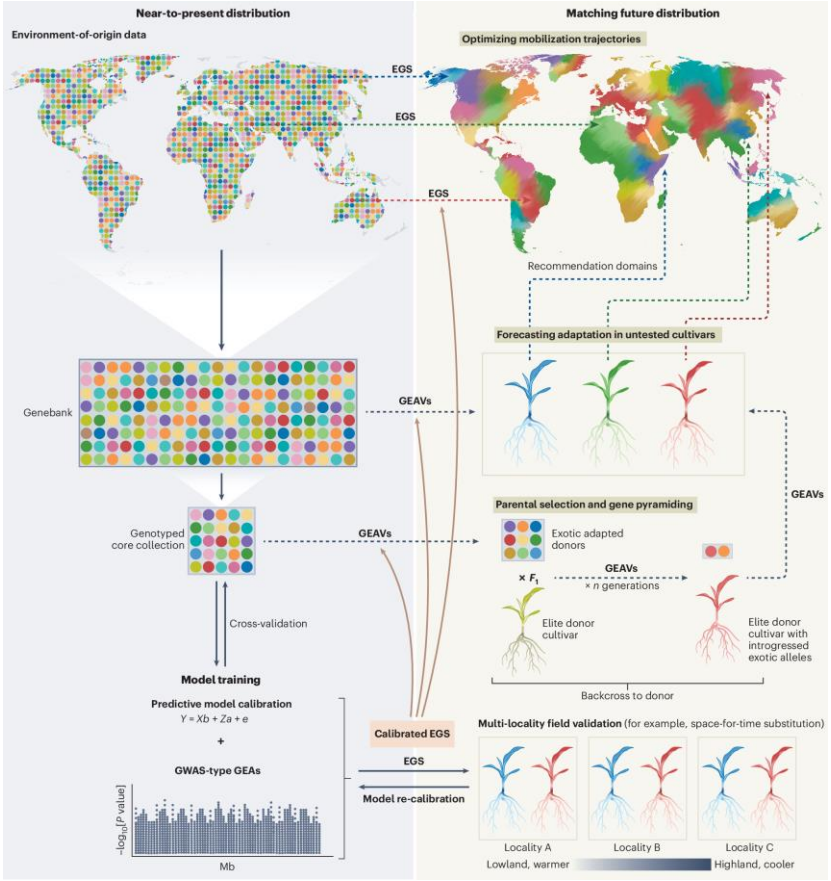


- Train Genomic and phenotypic data is split into training, validation, and test datasets.
- and adjust ML models to optimize prediction.

Select the final model for genomic prediction.

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Take home messages



- **PGRs are reservoirs of untapped alleles** for resilience, adaptation, and nutrition.
- **Multi-omics, GWAS, Pan-GWAS and single-cell omics expand our power to identify and interpret trait associations.**
- **Genomic Selection** leverages genome-wide variation but **predicts breeding values early**.

The future of PGRs pre-breeding: **an AI-enabled multi-omics system that turns PGRs diversity into predictive power for future crops.**


Thank you
for your attention