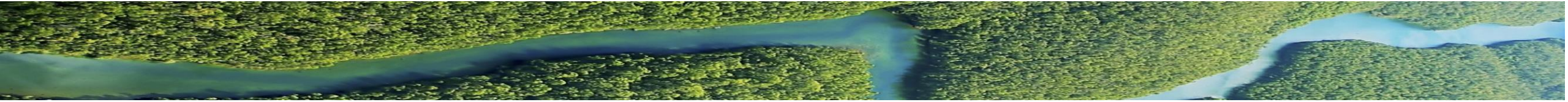


Molecular characterization of large collections using genomic tools

Lorenzo Barchi

2nd International Workshop on Plant Genetic Resources

UNIVERSITÀ DI TORINO

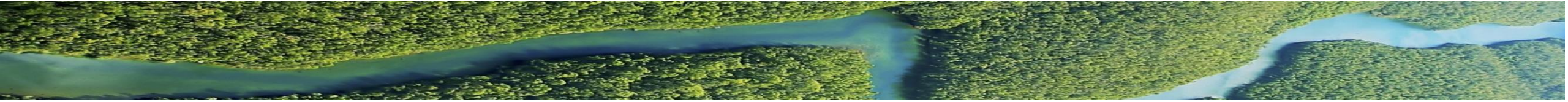
PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

- Crop breeding has to deliver adapted crops for future resilient agroecosystems
- This is particular important in the light of major challenges such as food security, climate change and loss of biodiversity

- The systematic genotyping and phenotyping of all conserved accessions in genebanks is hoped

The digitalization of collections would be an appropriate response to ongoing calls for improvement in the standardization of genebanks' sampling, storing and regeneration procedures

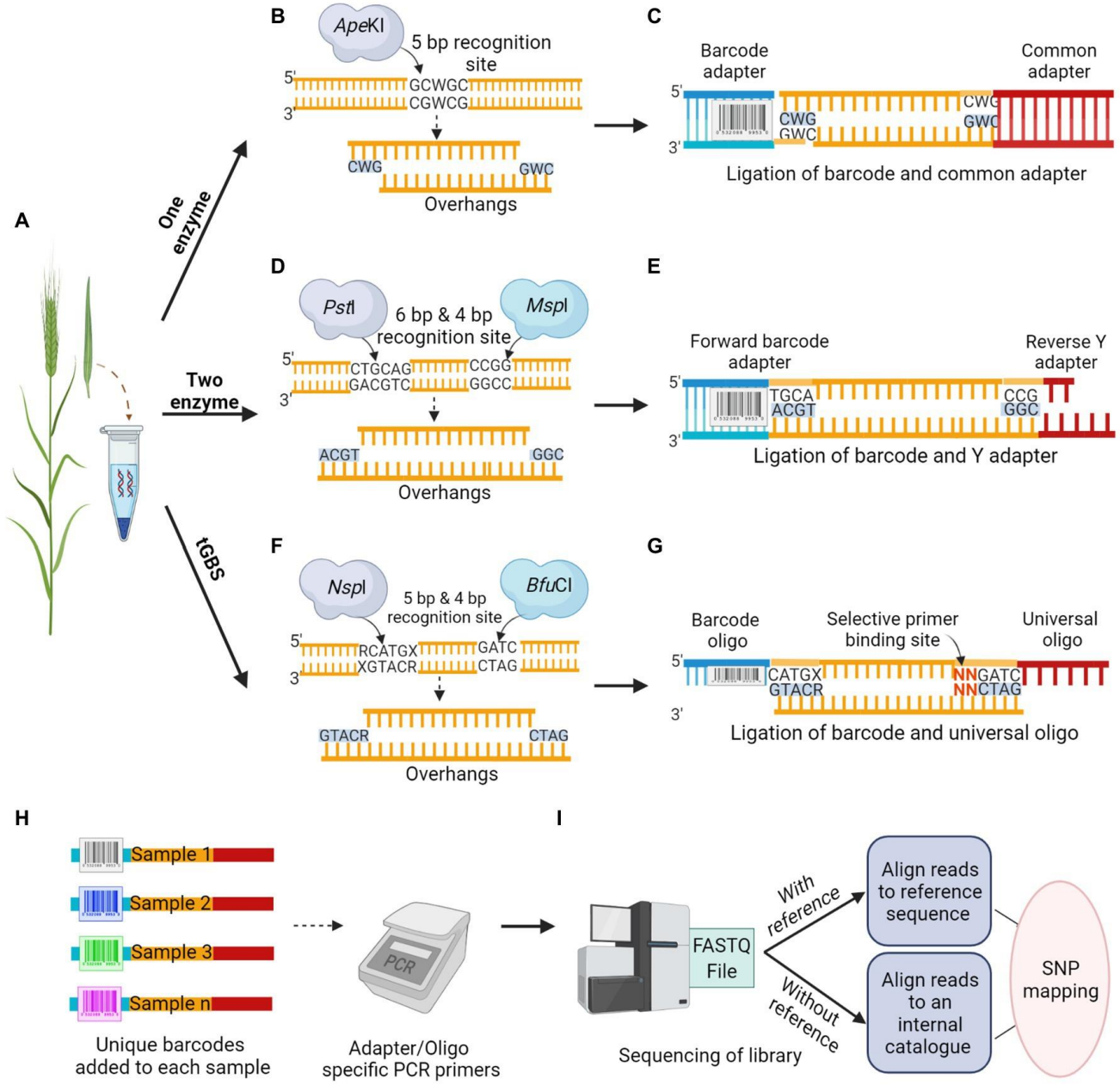
PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE



The traditional approaches

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

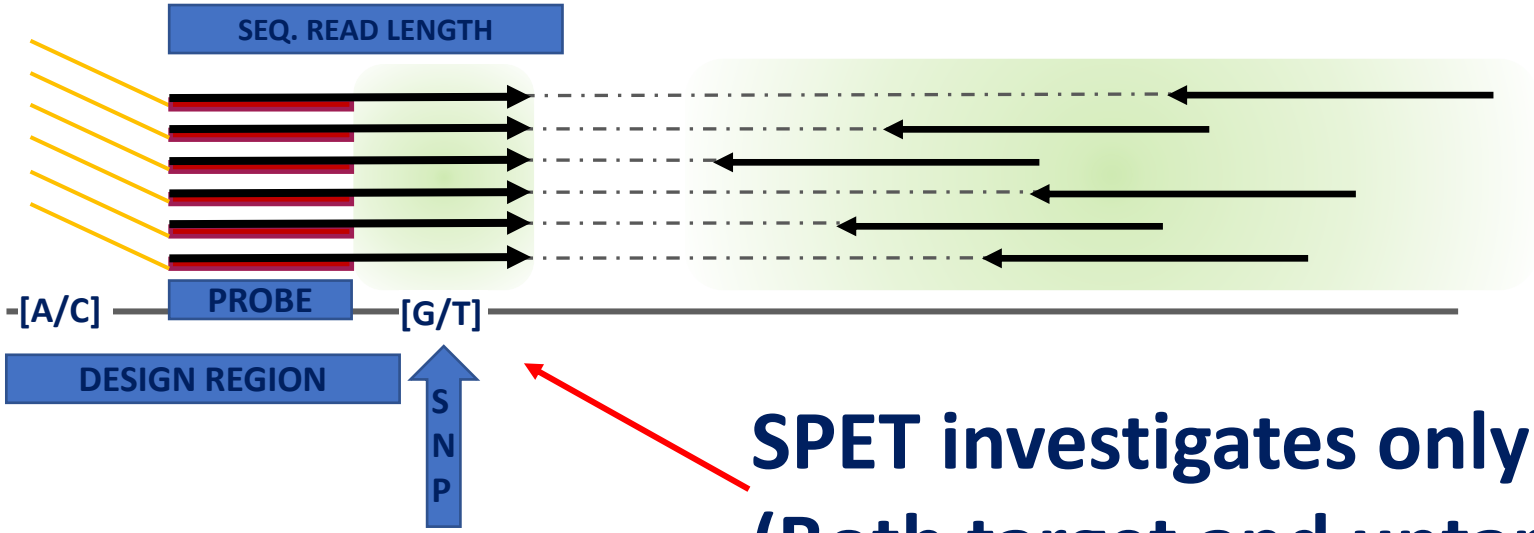
GBS: genotyping by sequencing



PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Reduced representation sequencing

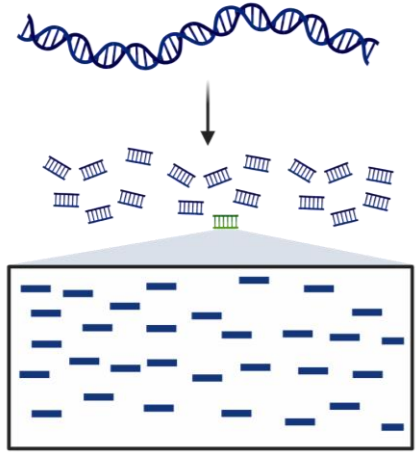
Target Enrichment Application: Genotyping By Sequencing

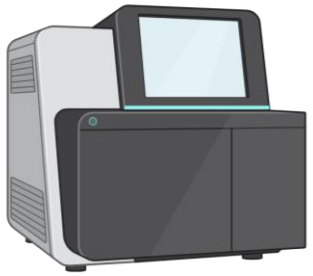


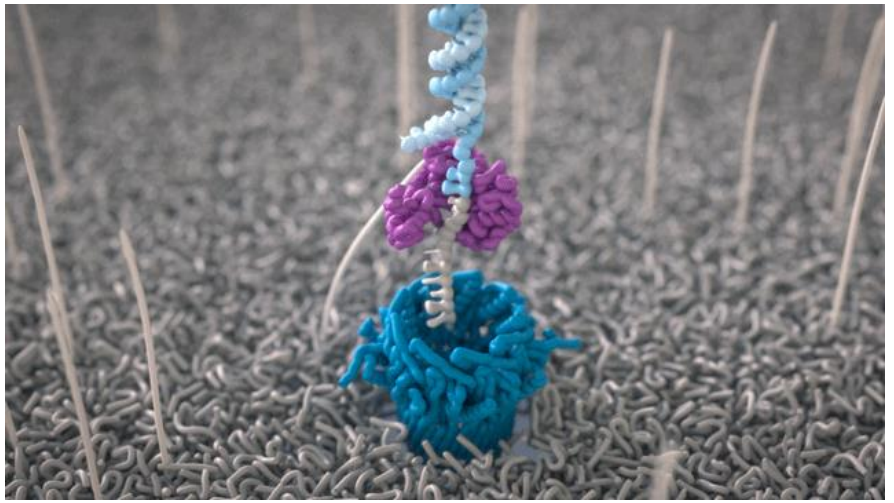
SPET investigates only <u>TARGET</u> loci (Both target and untargeted SNPs)

- Highly-multiplex genotyping (up to 3,092)
- Enables detection of a large number of known SNPs
- Every sequencing read is informative

WGS

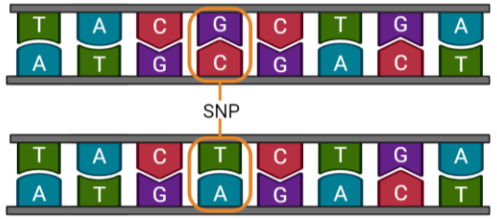




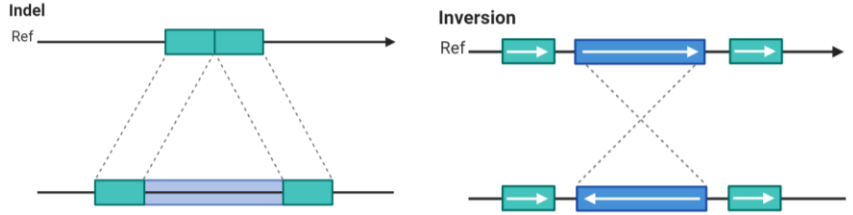


Sequencing data

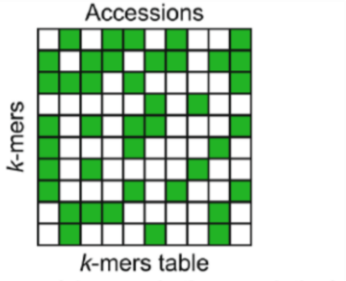
SNPs



Structural variations (SVs)



K-mers



6,574 accessions

3,532 genotyped accessions

Single Primer Enrichment Technology (SPET) genotyping

the plant journal

SEB
SOCIETY FOR EXPERIMENTAL BIOLOGY

Original Article | Open Access

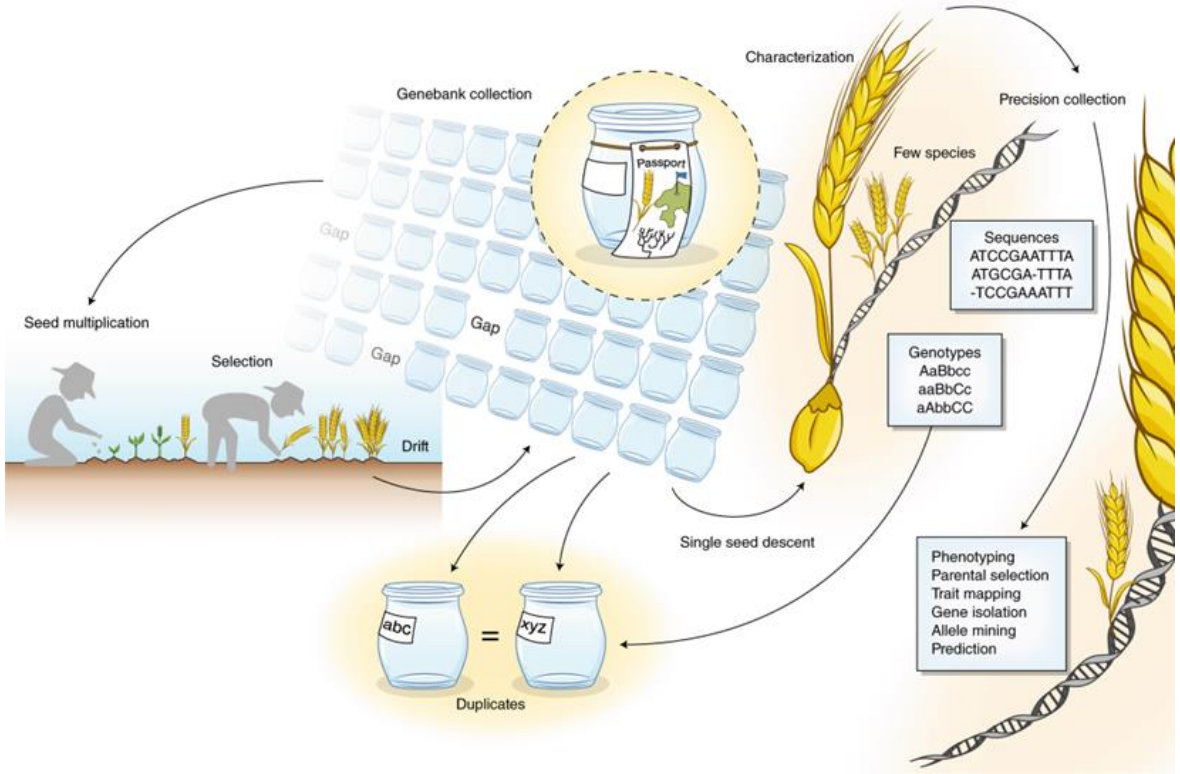
Analysis of >3400 worldwide eggplant accessions reveals two independent domestication events and multiple migration-diversification routes

Lorenzo Barchi, Giuseppe Aprea, M. Timothy Rabanus-Wallace, Laura Toppino, David Alonso, Ezio Portis, Sergio Lanteri, Luciana Gaccione, Emmanuel Omondi, Maarten van Zonneveld, Roland Schafleitner, Paola Ferrante, Andreas Börner, Nils Stein, Maria José Díez, Veronique Lefebvre, Jérémy Salinier, Hatice Filiz Boyaci, Richard Finkers, Matthijs Brouwer, Arnaud G. Bovy, Giuseppe Leonardo Rotino, Jaime Prohens, Giovanni Giuliano ... See fewer authors

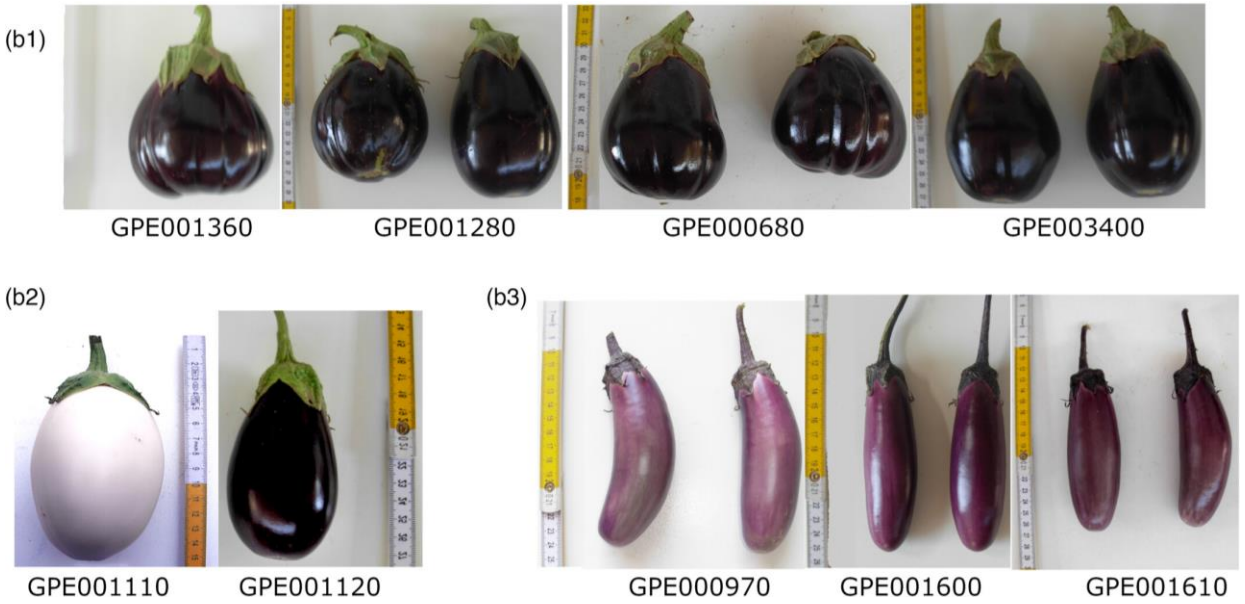
First published: 08 September 2023 | https://doi.org/10.1111/tpj.16455

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Evaluating genebank duplicates and misclassifications



A total of 591 non-control accessions



The majority of duplicates from *S. melongena* (425), followed by *S. aethiopicum* (105) and *S. macrocarpon* (43)

A similar approach used to spot **mislabelled accessions**

- **88 accessions were identified as putatively mislabeled**
- **Correct assignment** to a different species **for 31 accessions**

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Gaps

Historically, collecting efforts have focused on specific regions or species, leaving others unrepresented.

Without proper characterization and phenotyping, it's hard to understand what diversity is truly present and what is missing from collections.

Moreover, the diversity of crop wild relatives is poorly represented in gene banks

Procedure for population structure and kinship estimates are needed.
This will help to assist gap analysis and to reconstruct missing relations

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Moving ahead the traditional approaches

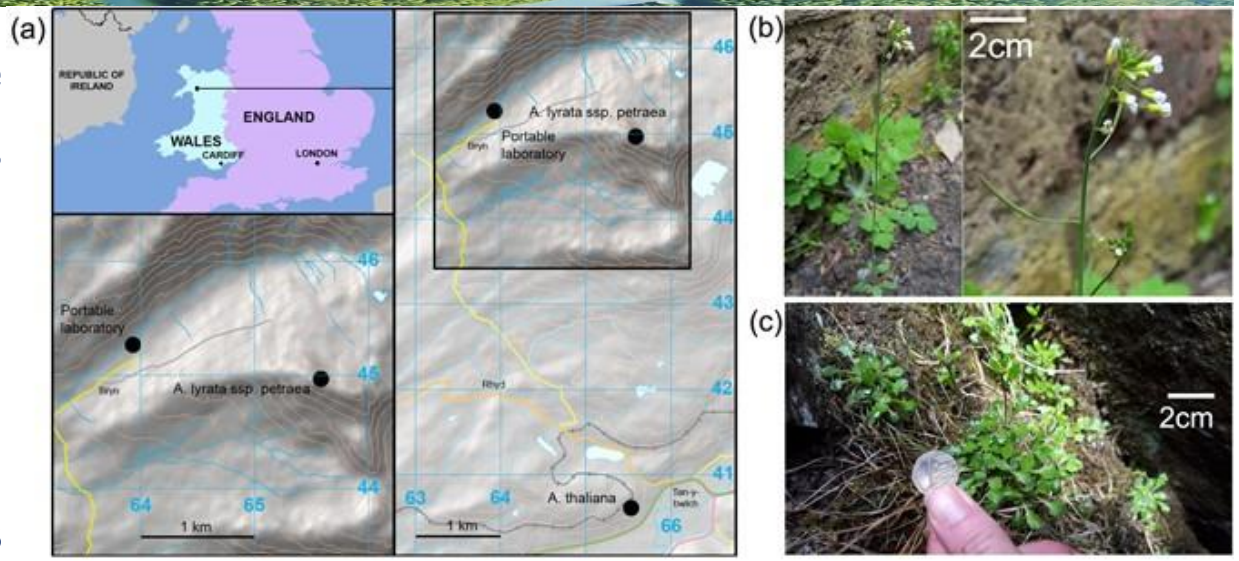
PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Molecular characterization using field-portable nanopore sequencing

- **Species Identification:** Portable devices like the Oxford Nanopore MinION enable species identification directly at the sample collection site.

This is valuable for fields such as:

- **Conservation:** Identifying endangered species or detecting invasive species (*in situ*).
- **Wildlife Crime:** Analyzing DNA from wildlife samples to track illegal trafficking.
- **Environmental Monitoring:** Assessing biodiversity or tracking specific genetic markers in environmental samples.

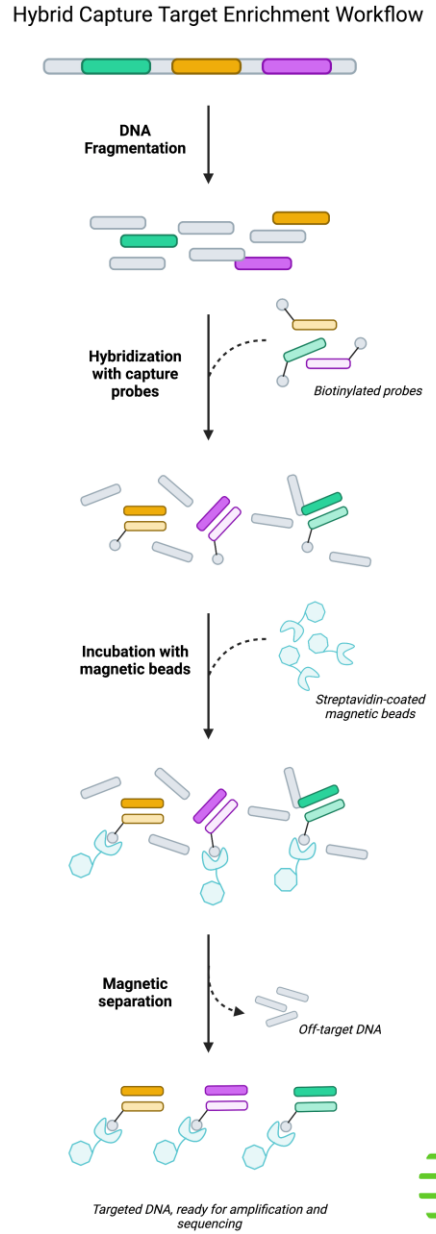


https://www.nature.com/articles/s41598-017-08461-5

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Target capture

- Target DNA enrichment, in combination with high-throughput sequencing technologies, represents a highly effective and cost-efficient strategy for the large-scale investigation and characterization of genomic loci.
- For many research objectives, obtaining a moderate number of loci (hundreds to thousands) across many individuals is more appropriate than sequencing entire genomes from a few individuals
- The hybrid capture-based approach encompasses a suite of molecular techniques that selectively increase the representation of specific DNA regions within next-generation sequencing (NGS) libraries through the application of oligonucleotide probes, commonly referred to as “baits”.
- Prices are relatively low per sample, allowing to genotype large collections



PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Gaps

Genebanks and curated collections aim to preserve the genetic richness of a species. We propose an empirical approach to assess **how much of the species' nucleotide diversity is represented** within a given collection.

To evaluate how comprehensively a genetic collection captures the diversity of a species, we adopted a rarefaction-inspired framework:

- Genotyping data were available for all or a representative set of accessions.
- Random subsamples of increasing size were generated, with multiple replicates.
- For each subset, nucleotide diversity (π) estimated across the genome/selected loci.
- Mean values and variation were computed and plotted.

Gaps

Interpretation

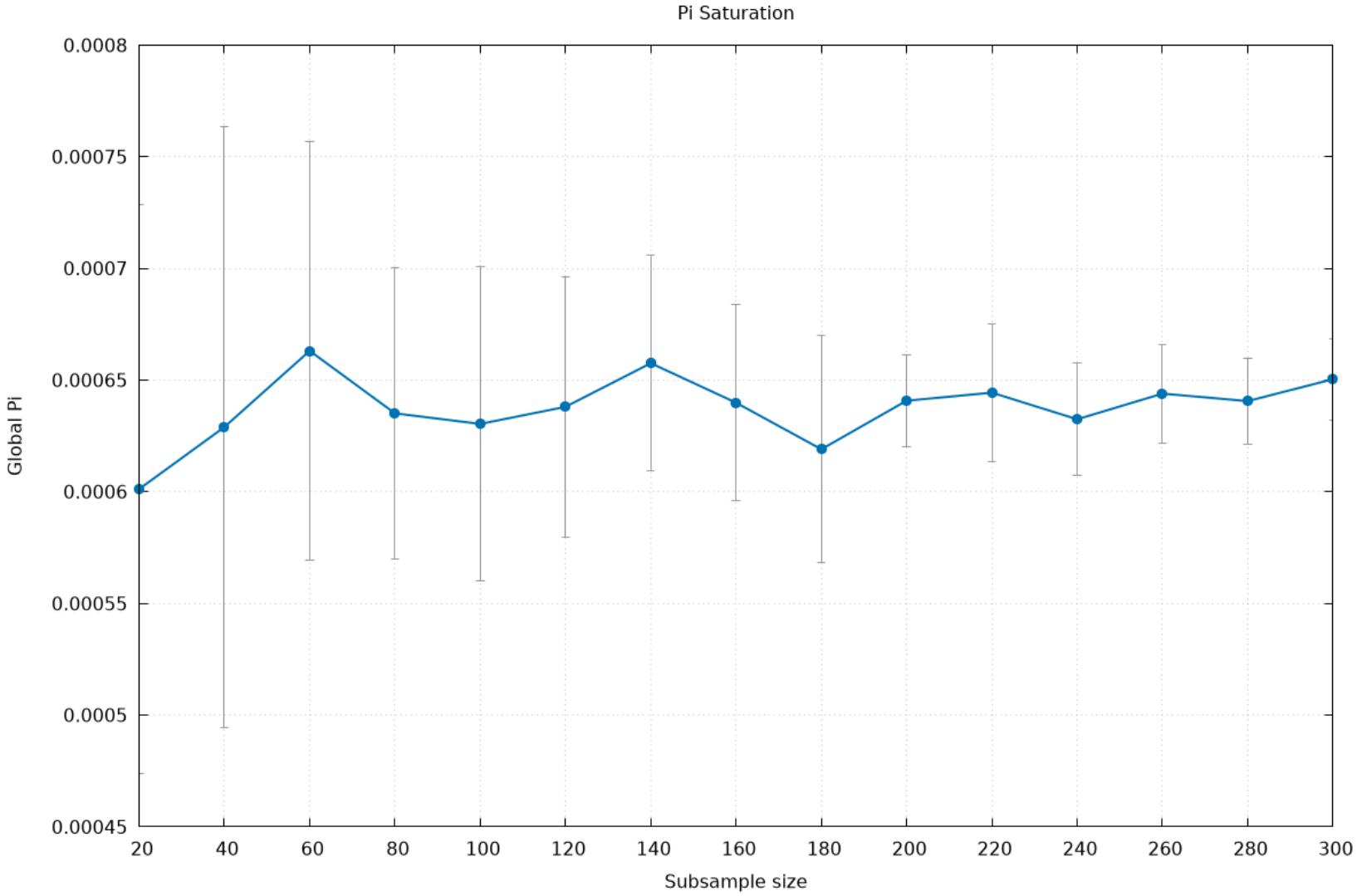
- If diversity increases rapidly with subset size and plateaus, the collection **captures most of the available diversity**.
- This **saturation curve** can be used to:
  - Assess completeness of a core collection,
  - Compare diversity coverage across genebanks,
  - Identify redundancy or gaps in conservation efforts.

Why π?

- Direct measure of average pairwise sequence diversity.
- Scalable and robust to missing data.
- Applicable to SNP-level VCFs across genomic regions.

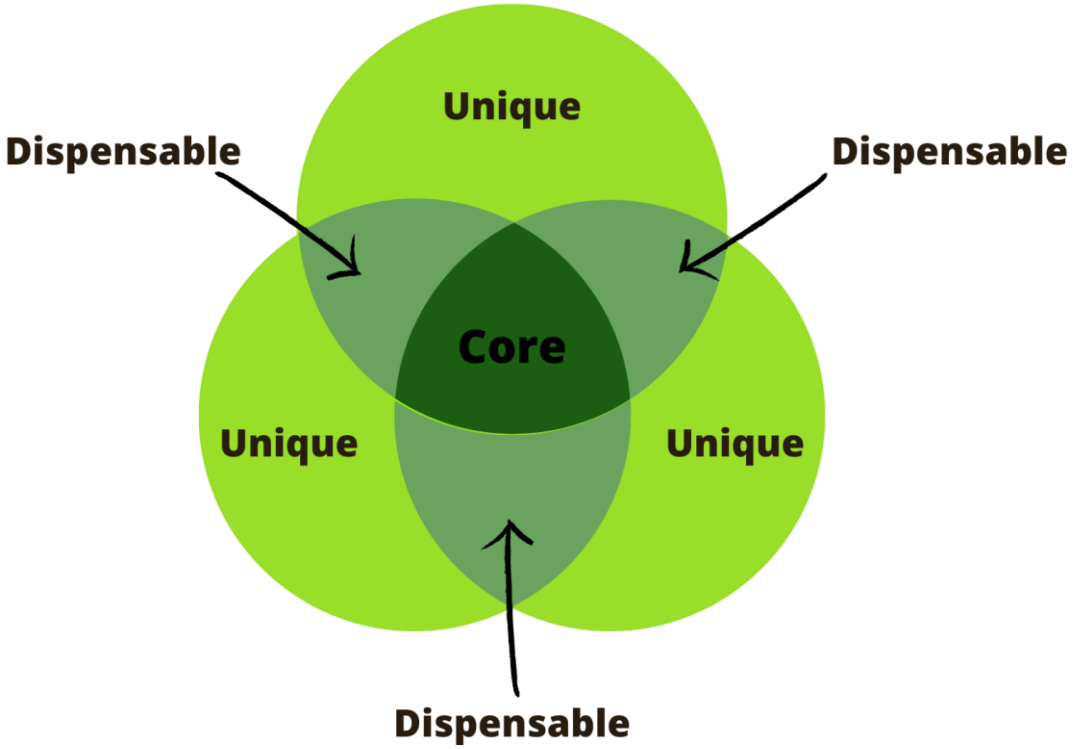
Gaps

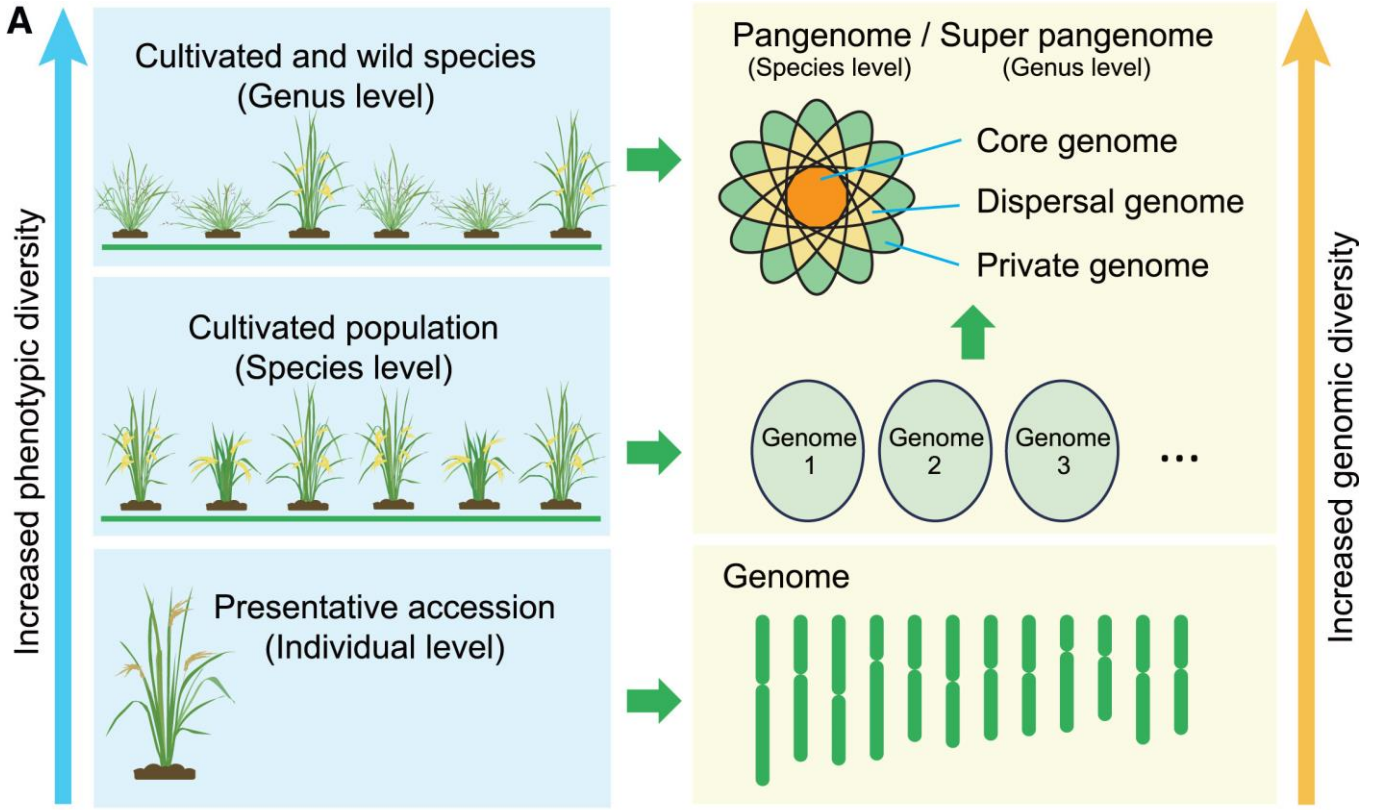
- Genotyping from High-coverage WGS data (20×) was available for a core collection of ~350 tomato accessions.
- 20 random sub-samplings of increasing size (from 20 to 300 accessions) were generated.



PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Pangenome and super pangenome concepts





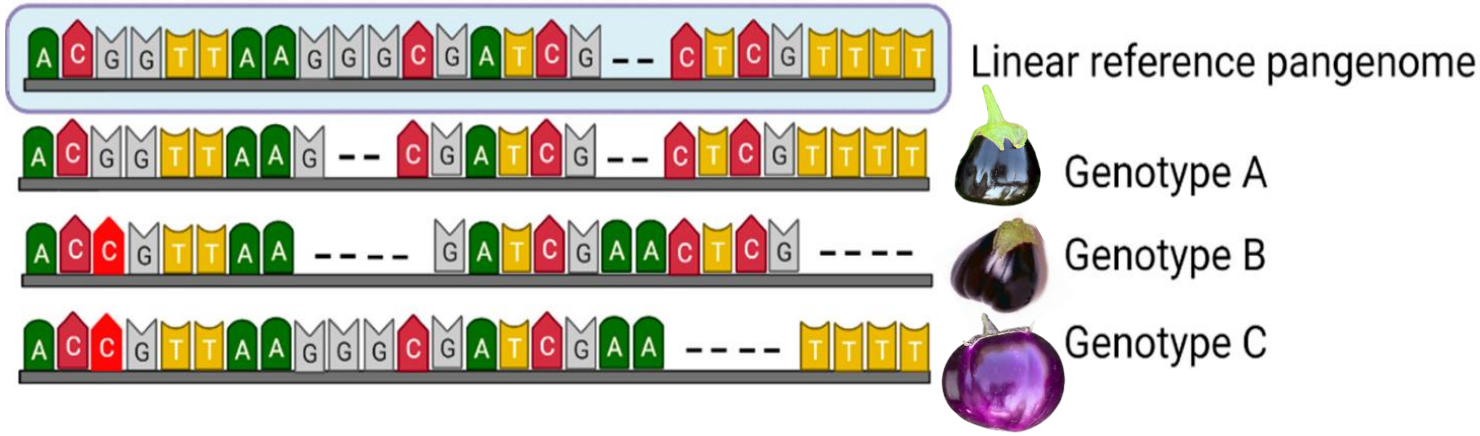
From He *et al.* 2025

PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

Eggplant pangenome

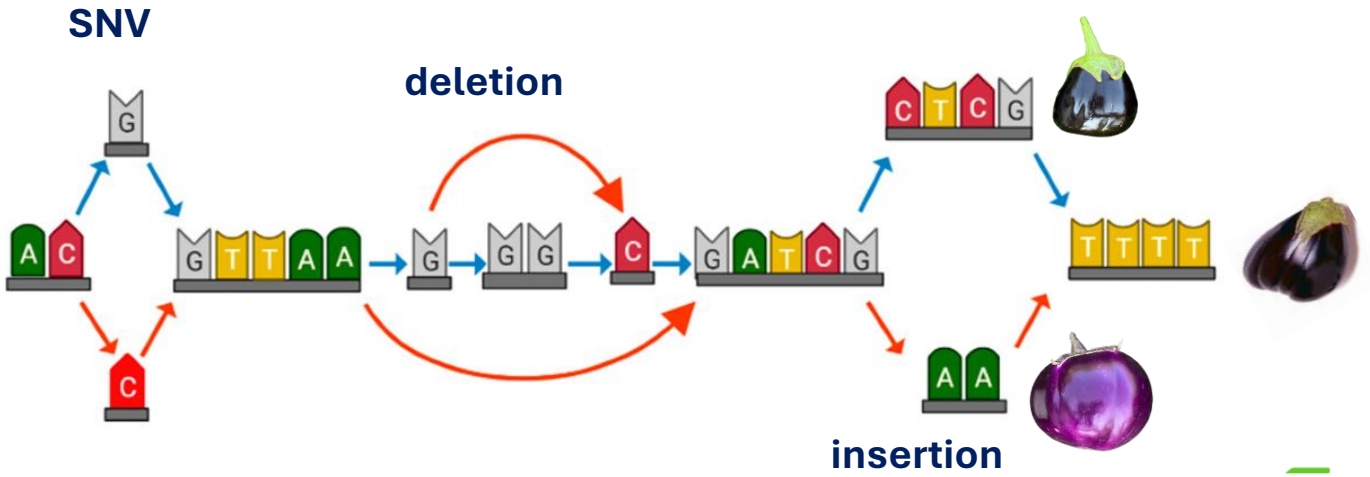
Linear pangenome

Already published
(24 *S.melongena, 1 S. incanum* and 1 *S. insanum* accessions*)*
Barchi et al., 2021



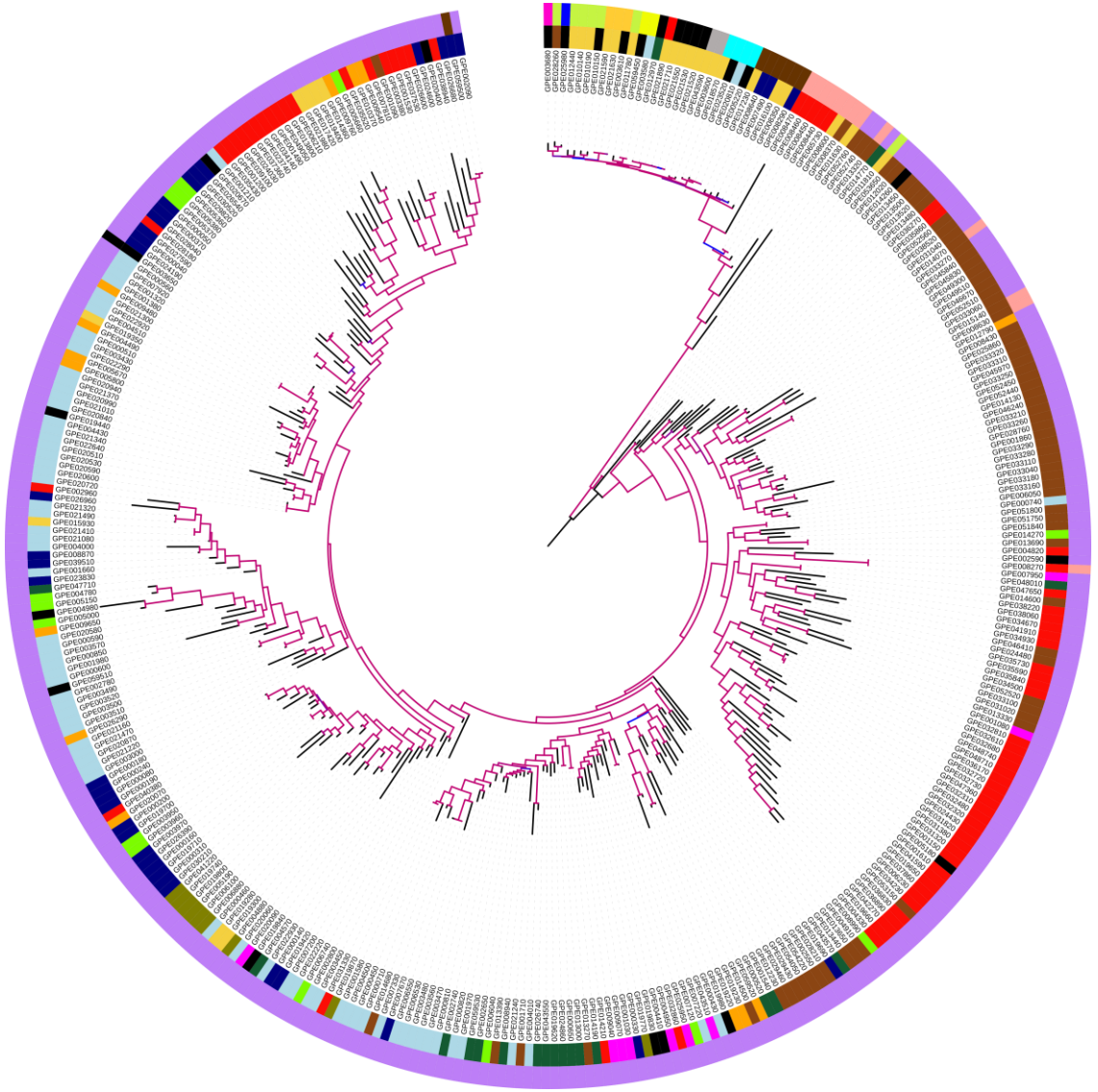
Graph-based pangenomes

33 *S.melongena, 4 S. incanum* and 3 *S. insanum*
(PG-SMA)



PRO-GRACE
PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

ML tree of *S. melongena* and wild relatives using SNPs genotyped on *PG-SMA*



Species

- *S.aethiopicum*
- *S.anguivi*
- *S.incanum*
- *S.linnaeanum*
- *S.macrocarpon*
- *S.melongena*
- *S.sisymbriifolium*
- *S.tomentosum*
- *S.torvum*
- *S.insanum*
- *S.campylacanthum*
- *OtherSpecies*

Origin

- *Middle East*
- *Southeast Asia*
- *South Europe*
- *Not available*
- *China*
- *India*
- *Japan/Korea*
- *Africa*
- *North-Central Europe*
- *Americas/Oceania*
- *East Europe*

Pangenome and super pangenome advantages

- Pangenome graphs improve the accuracy of variant genotyping with short reads (especially for large structural variants).
- This is because short reads span the entirety of small variants but are not able to span larger structural variants.
- By using a pangenome graph containing known structural variants, short-read coverage along the graph can be used to genotype structural variations with more accuracy than traditional short-read methods using a single linear reference genome

- Pangenomes represent the genomic content of a population much more completely than a traditional single linear reference genome.
- They are less restrictive than a traditional reference in that they can be used in the analysis of phylogenetic clades above the species level, and they improve research outcomes by reducing reference bias.


Thank you for attention!