Reconstructing plant pangenomes

Ettore Riccucci
e.riccucci@santannapisa.it

Sant'Anna
School of Advanced Studies – Pisa

The complexity of plant genomes

Differences in genome size between plant species are much larger than between other eukaryotes

Dicotyledonous genome size

Genlisea margaretae
→ 63 Mb

Paris japonica
→ 150 Gb

The complexity of plant genomes

Differences in genome size between plant species are much larger than between other eukaryotes

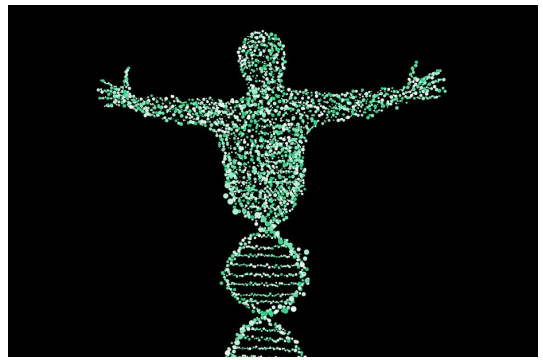
Dicotyledonous genome size

Genlisea margaretae
→ 63 Mb

Paris japonica
→ 150 Gb

Mammalian genome size

Carriker's round-eared bat → 1.6 Gb



Homo sapiens → 3.2 Gb

Red viscacha rat → 8 Gb

The complexity of plant genomes

Whole genome duplication (WGD)

Transposable elements (TEs)

The role of WGD

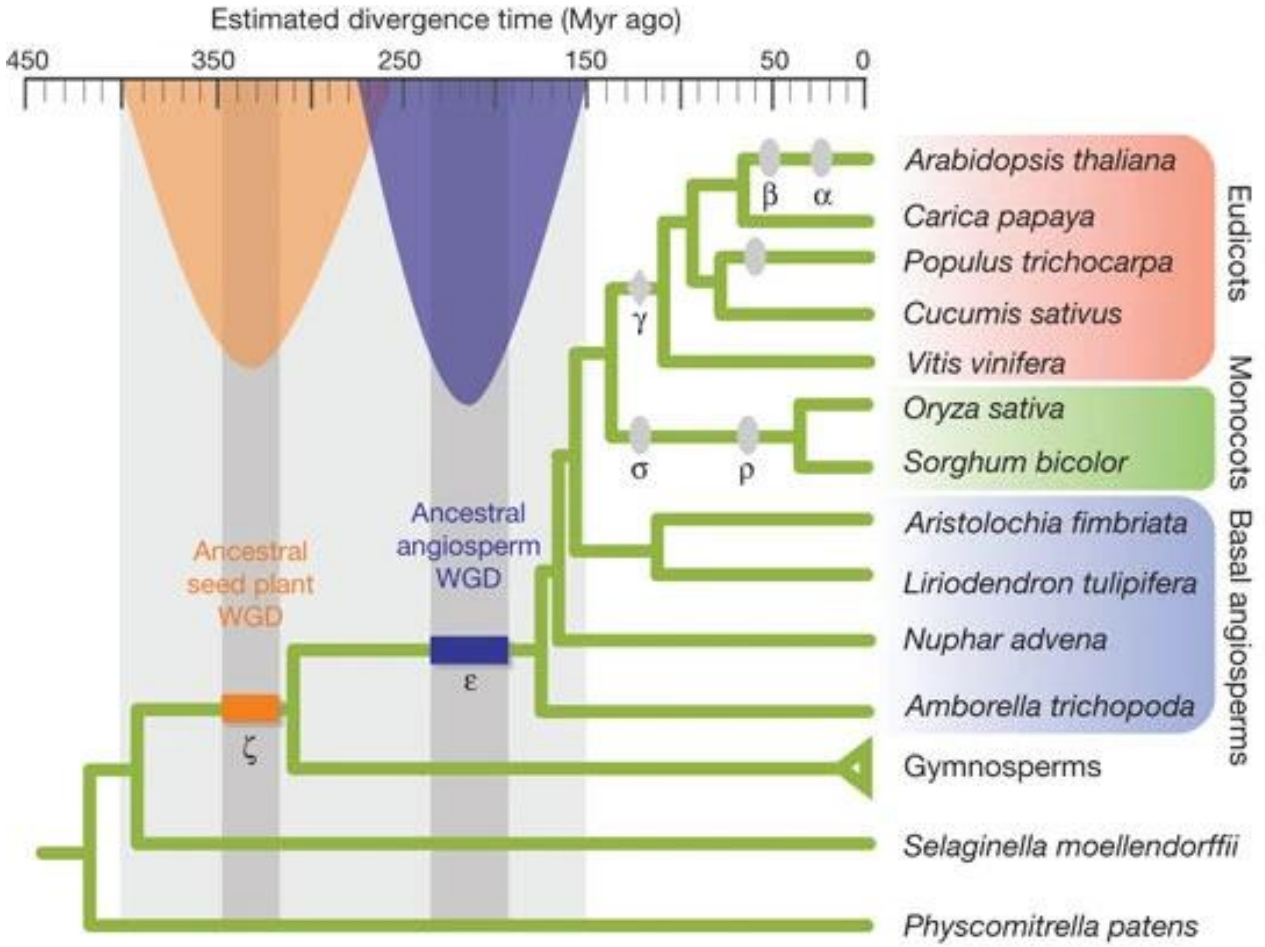
Multiple WGD over the past 200 million years

Most recent WGD event occurred approximately 450 MYA in the lineage leading to humans

The role of WGD

The phylogenetic history of **all** angiosperms abounds with WGD events

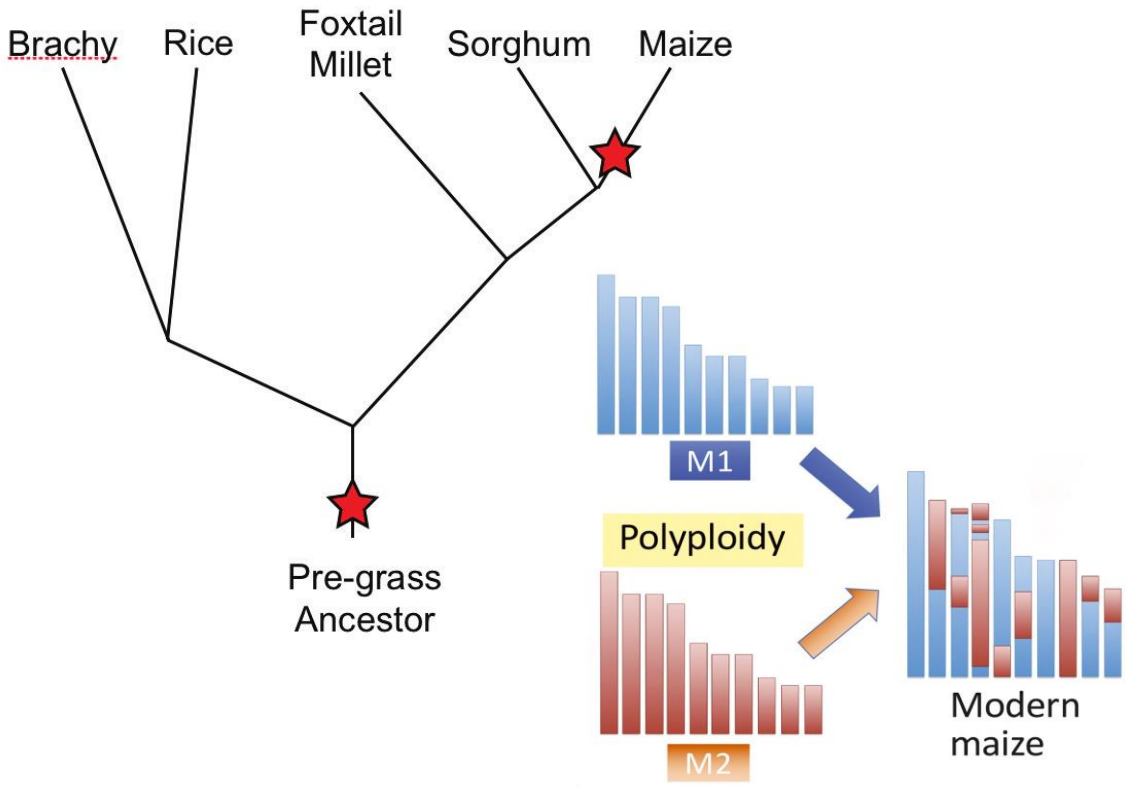
→ All seed plants are paleopolyploid



The role of WGD

The phylogenetic history of **all** angiosperms abounds with WGD events

→ All seed plants are paleopolyploid



Why current diploid plants?

- massive genome structural rearrangements
- reductions in chromosome numbers
- large-scale loss of repetitive sequences and duplicate genes

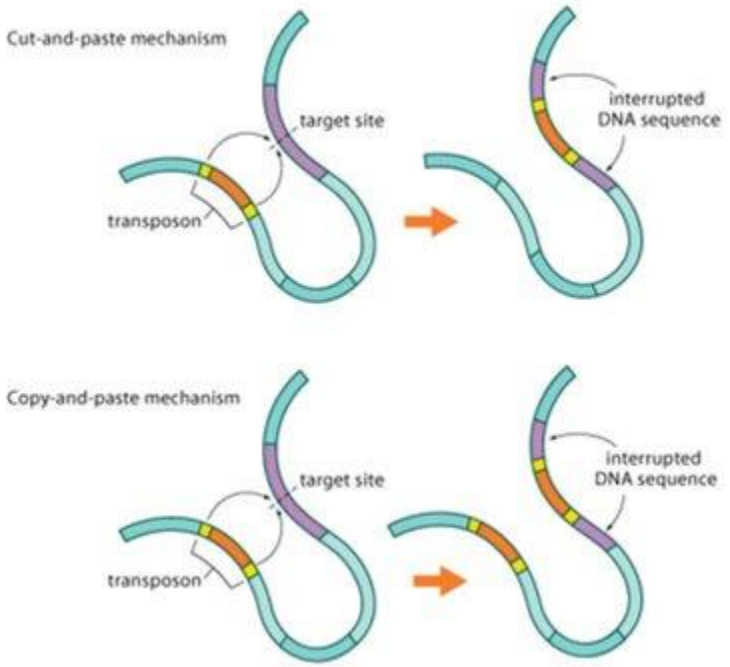
Maize genome

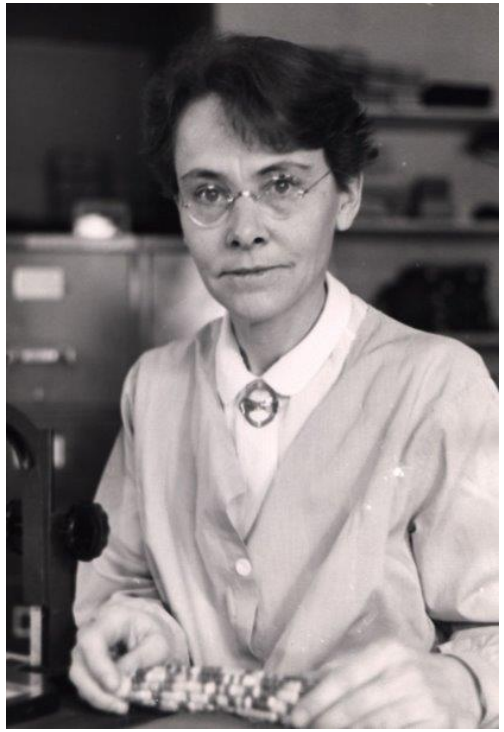
- WGD common to all angiosperms
- 5 – 20 MYA: allopoliploidy + fractionation → diploidy

The role of TEs proliferation

Sequences of DNA that move (or jump) from one location in the genome to another

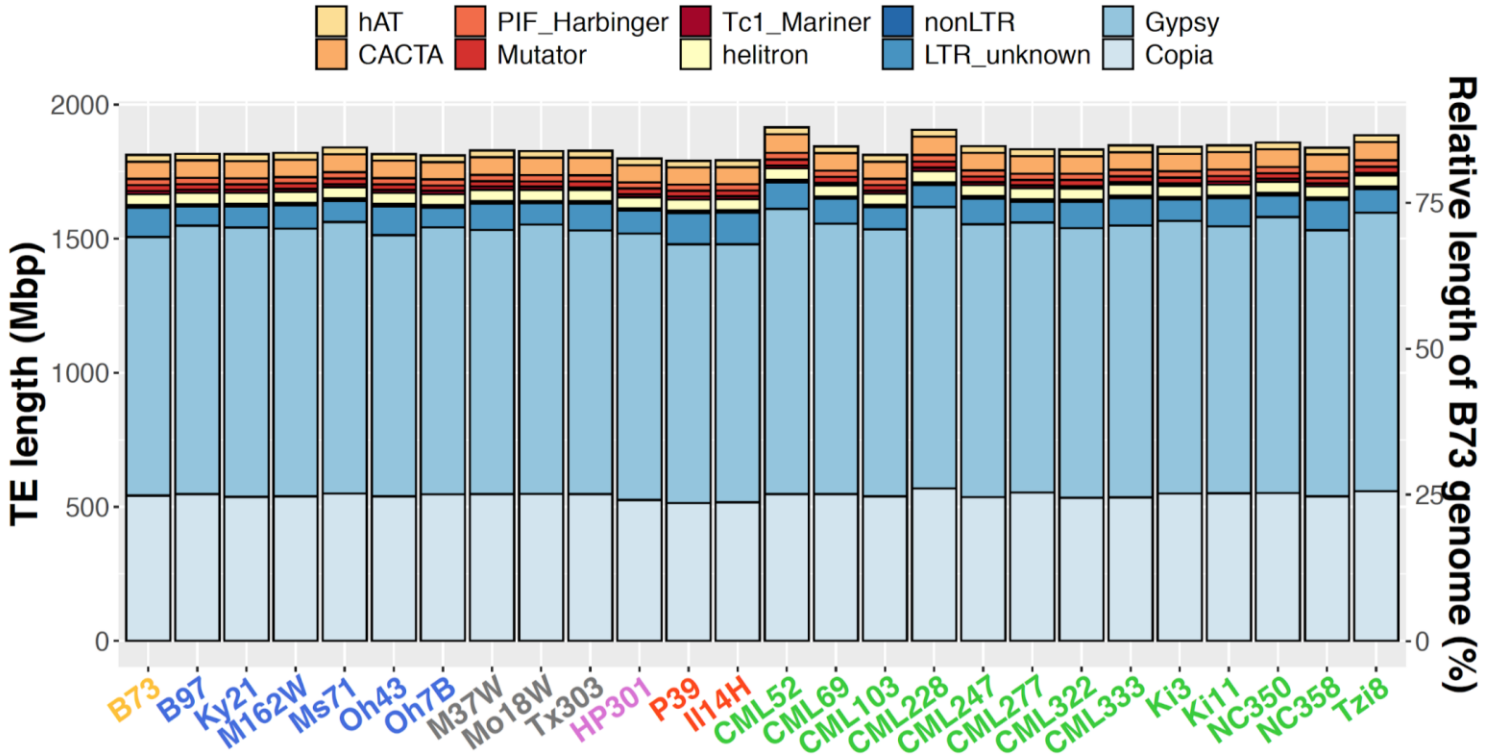
Infamously known as junk DNA (Ohno , 1972), parasitic DNA (Orgel and Crick, 1980)





Barbara McClintock, Nobel Prize for Physiology and Medicine 1983

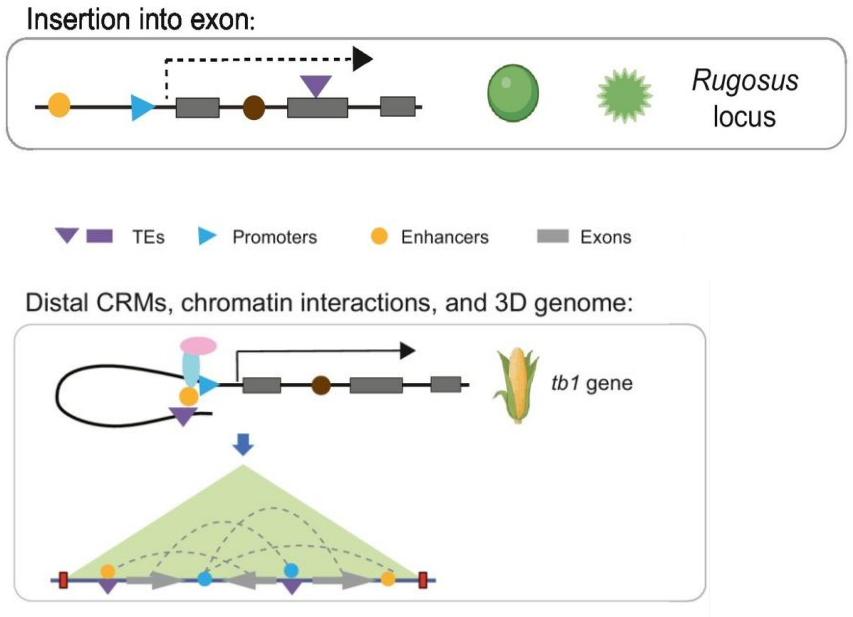
The role of TEs proliferation



- TEs: > 80 % of the maize genome
- TE insertions linked to dramatic phenotypic changes
- TEs are a source of diversity within the species

Sant'Anna
School of Advanced Studies – Pisa

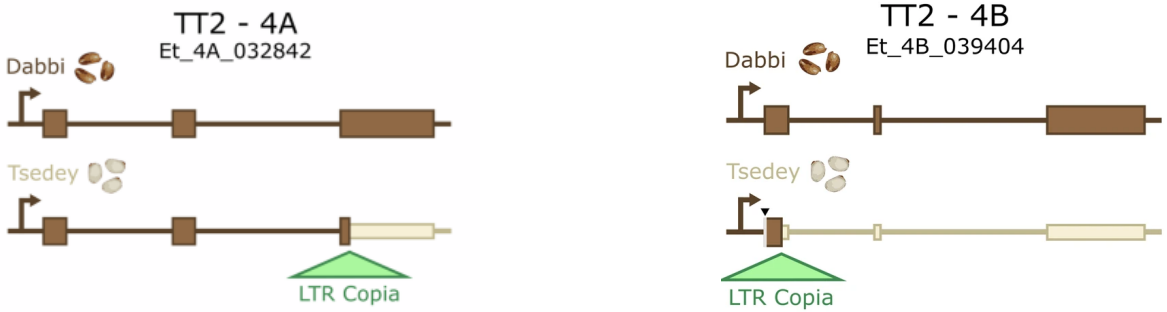
The role of TEs proliferation



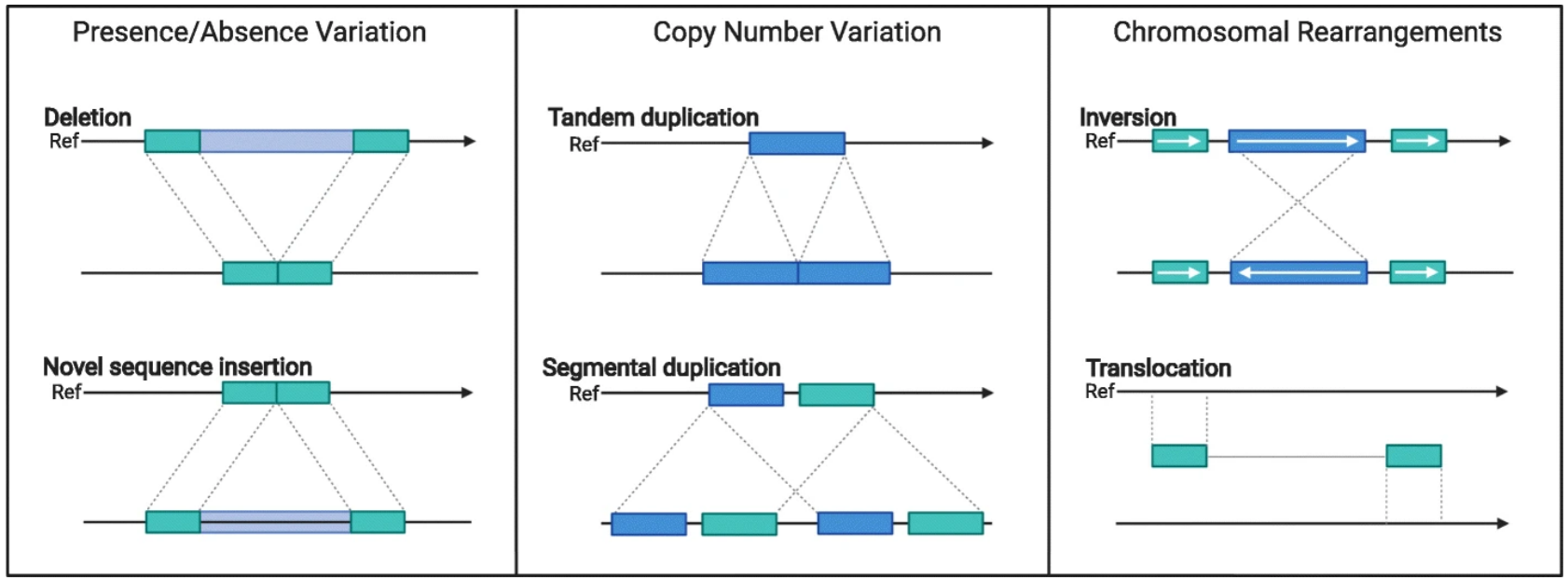
The wrinkled-seed phenotype in Mendel's pea plants is caused by a transposon-like insertion into the *SBEI* gene, which encodes starch-branching enzyme I

TE insertion in a distant upstream region (*cis-regulatory modules, CRM*) of *tb1* is the causative mutation in the *teosinte branched 1* (*tb1*) gene that account for the gene's role in the domestication of modern corn from its ancestor teosinte

TRANSPARENT TESTA 2 (TT2) is a candidate gene identified for teff seed color and highlight the putative effect of TE insertions in white seed phenotype

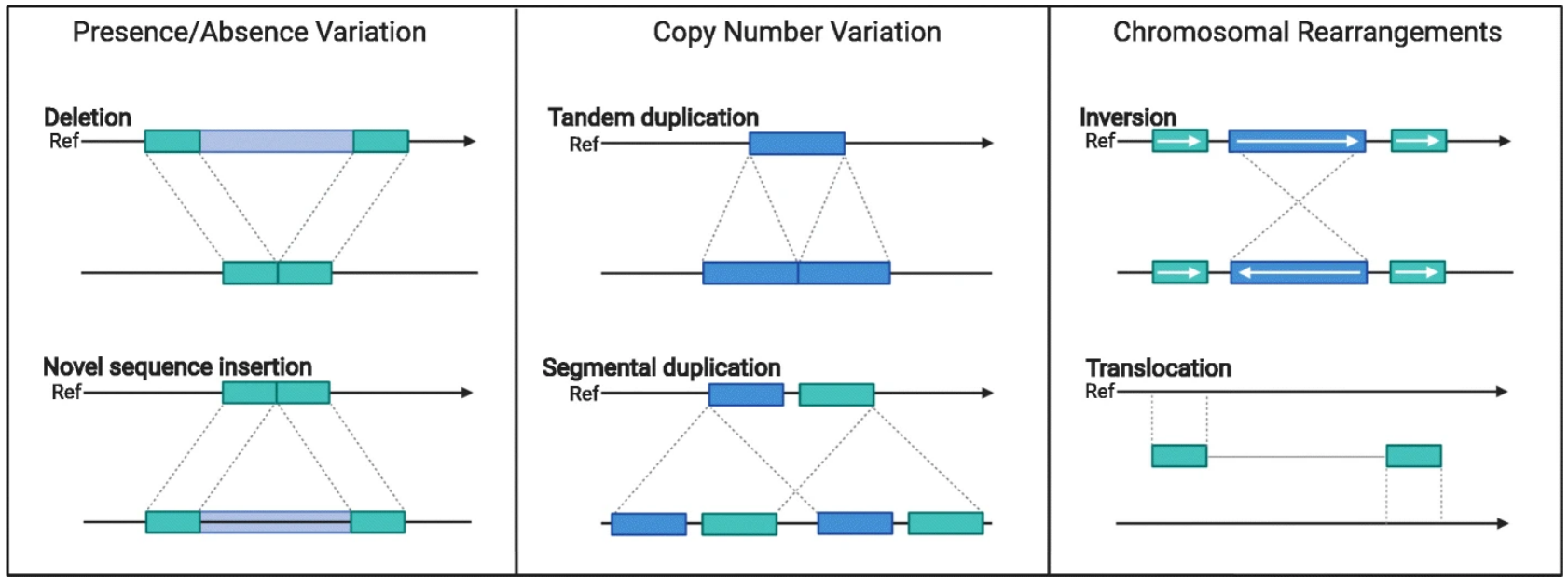


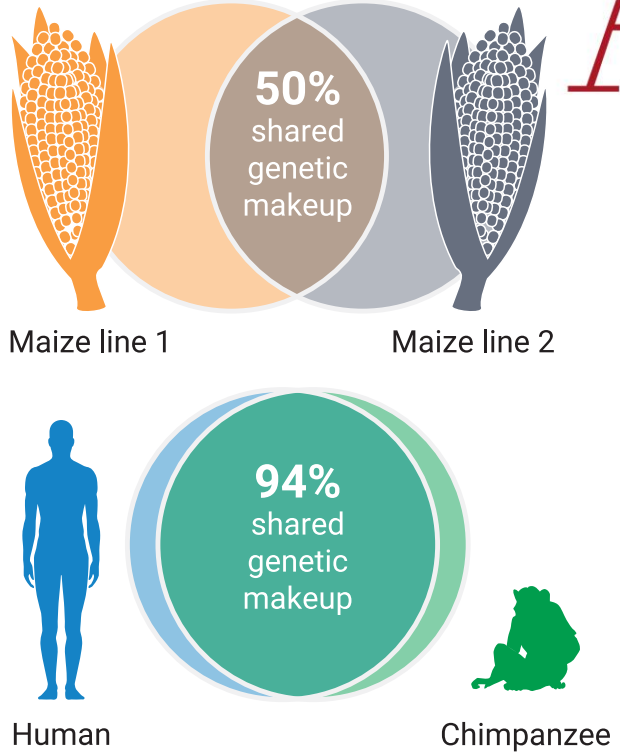
Intra-specific genetic diversity: structural variants (SVs)



- TEs-driven SVs
- Errors during meiotic recombination
- Differential genome fractionation across genotypes following WGD

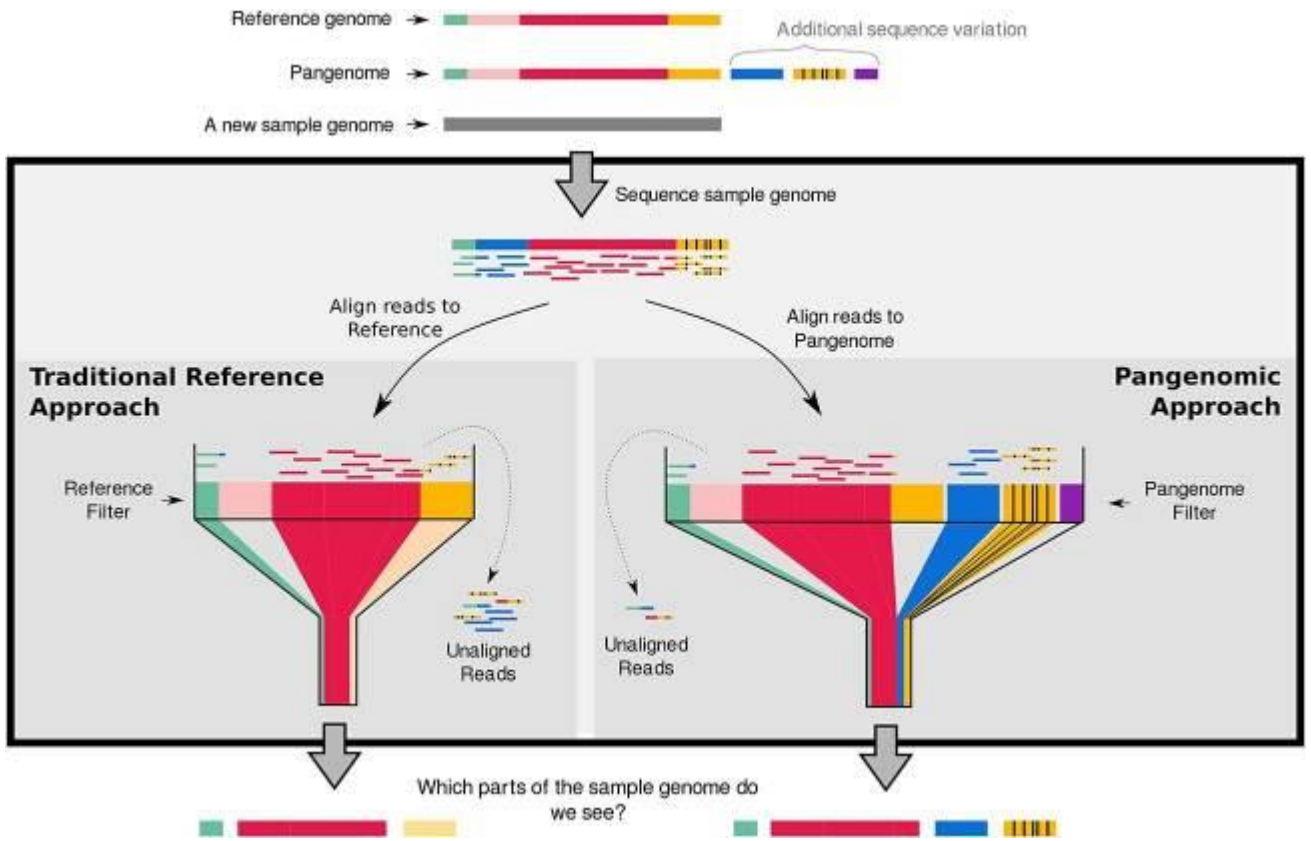
Intra-specific genetic diversity: structural variants (SVs)





- TEs-driven SVs
- Errors during meiotic recombination
- Differential genome fractionation across genotypes following WGD

A single reference genome is not enough

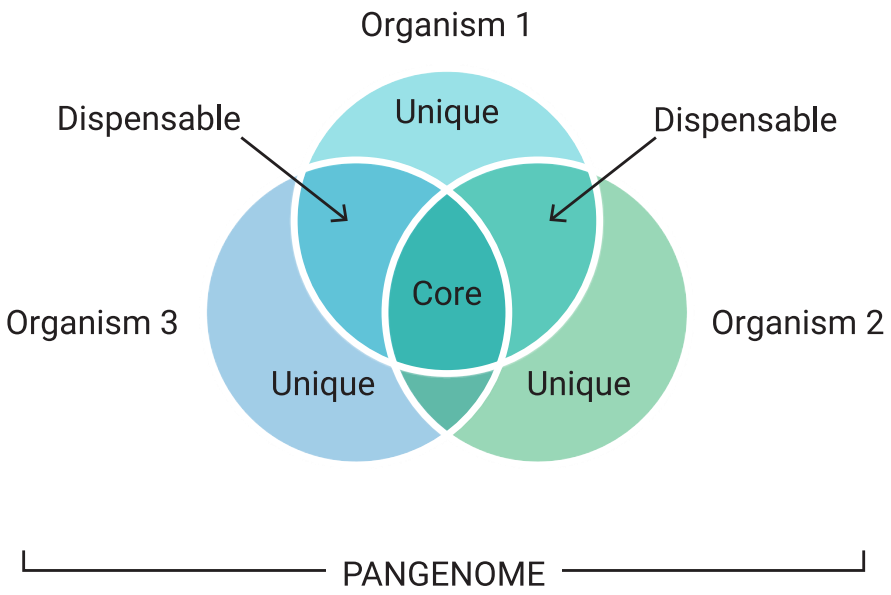


Traditional single linear reference genome cannot represent the whole genomic variation of a species

- Traditional bioinformatics analyses make comparisons between a new sample and the reference sequence through **read alignment**
- If the reference doesn't contain a genomic sequence that is similar to the sample sequence, reads from the sample will align poorly or won't align at all

→ **Reference bias**

Investigating intra-specific genetic diversity: pangenomics



- Pangenomes represent the genomic diversity of a species and includes core genes, found in all individuals, as well as dispensable genes, which are absent in some individuals
- Dispensable gene annotations often show similarities across plant species, with genes for biotic and abiotic stress commonly enriched within variable gene groups

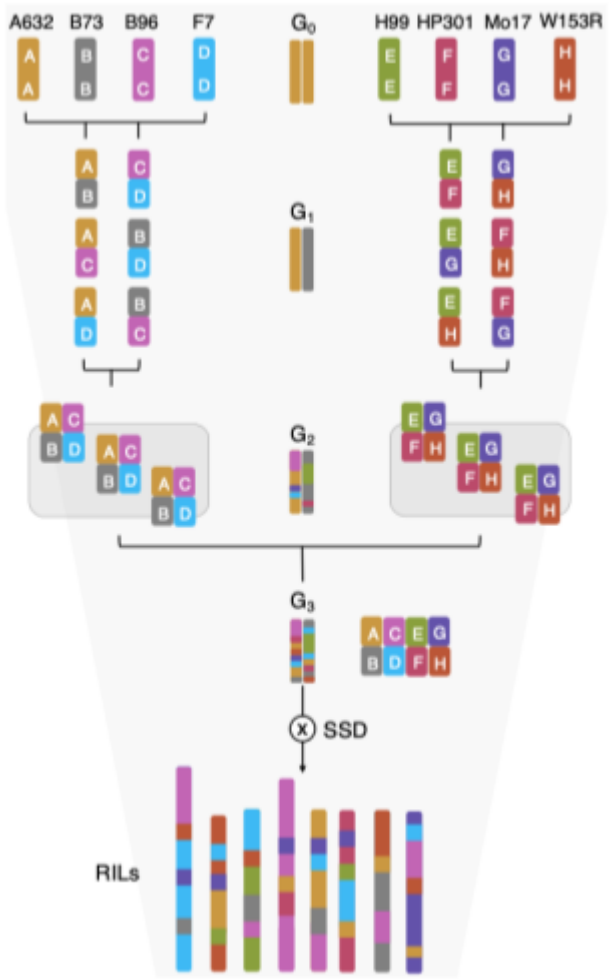
- 2005: first bacterial pan-genome (Tettelin et al.)
- 2025: Advances in sequencing technologies and bioinformatics → Pangenomics is the new standard

Combining multiparental mapping population and pangenomics

Mulitparental Advanced Generation InterCross (MAGIC) maize population

Teff Nested Association Mapping (NAM) population

The pangenome of the MAGIC maize founders

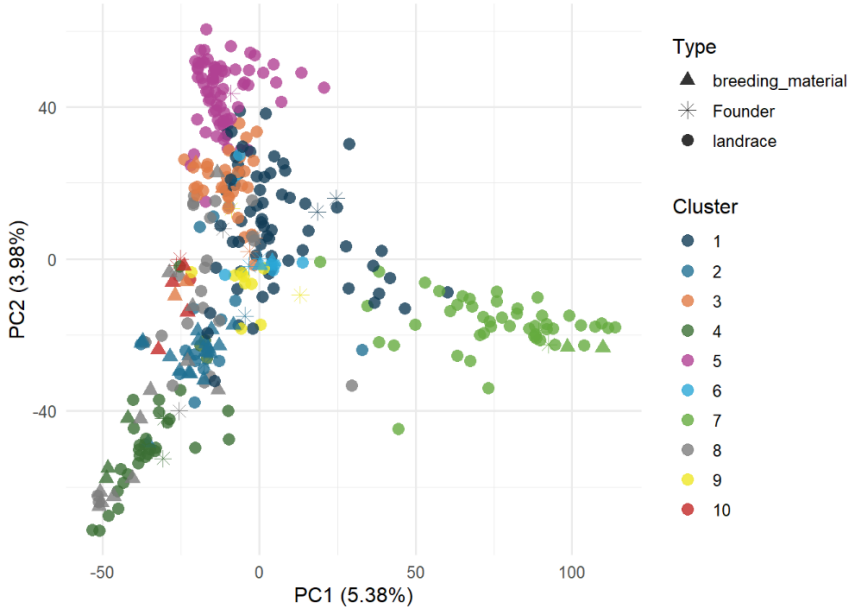


8 founders chosen to maximize genetic diversity

RILs genomes: evenly differentiated mosaic of the eight founders

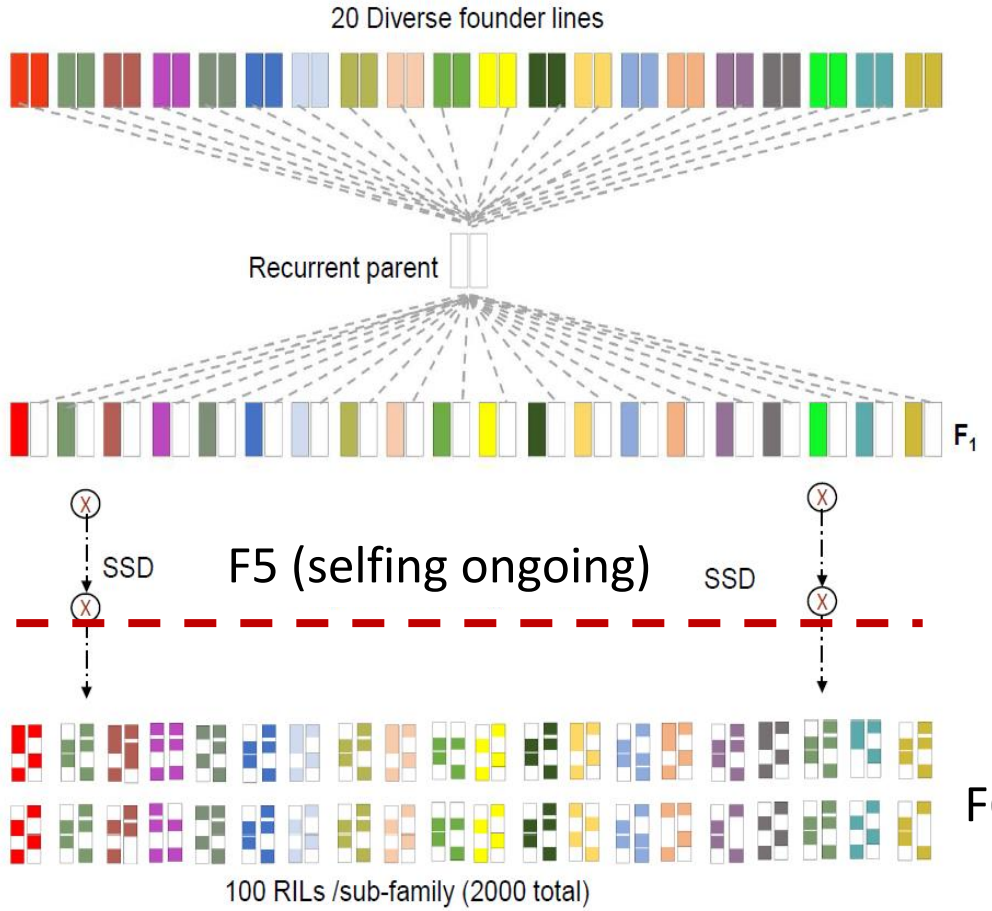
High diversity and dense recombination events to allow quantitative trait loci (QTL) mapping in maize

The pangenome of the teff NAM founders

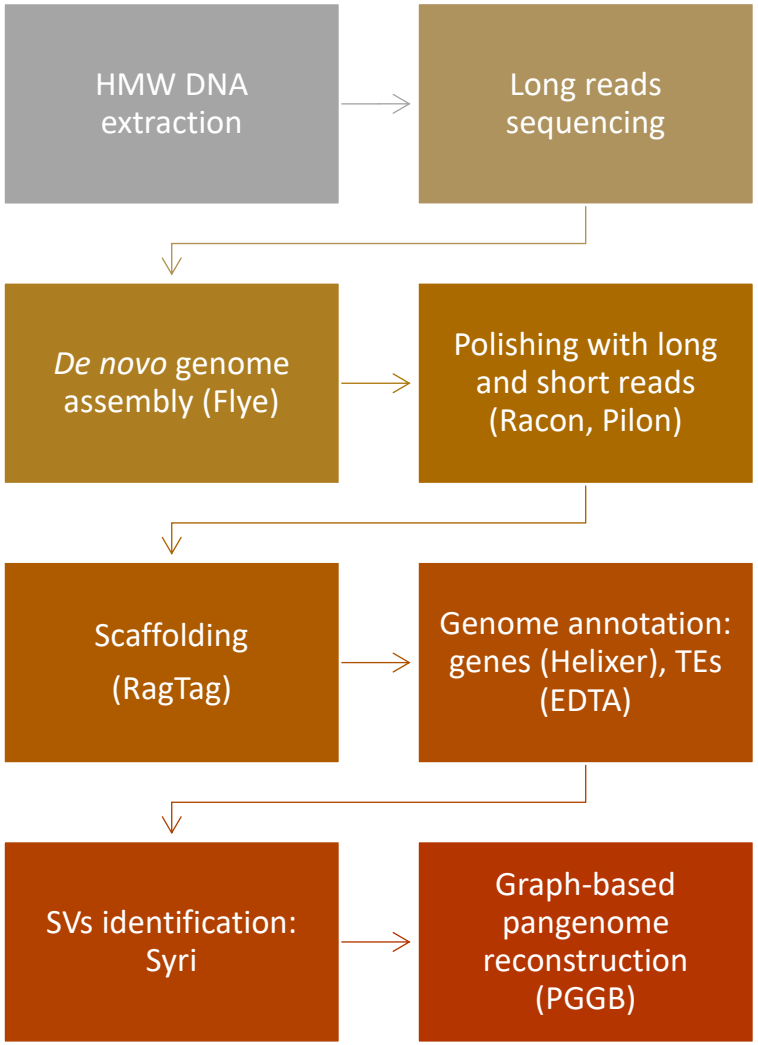


Ethiopian Teff Diversity Panel : landraces and improved varieties. Within each cluster, 2 accessions were selected as founders to maximize the diversity of the NAM population

→ 20 founders, including the recurrent founder Quncho



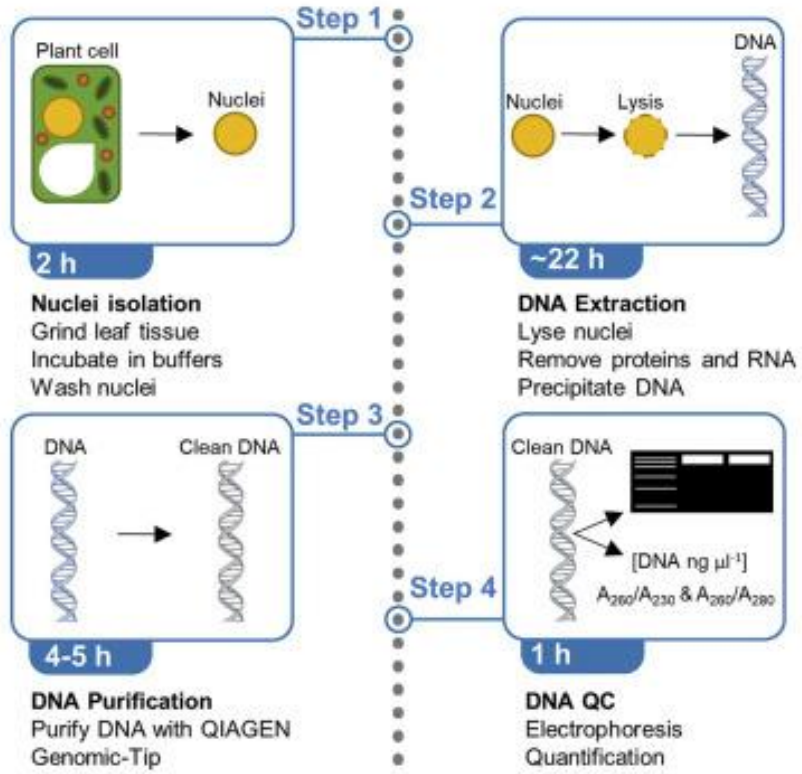
Workflow


HMW DNA extraction
Long reads sequencing
De novo genome assembly (Flye)
Polishing with long and short reads (Racon, Pilon)
Scaffolding (RagTag)
Genome annotation: genes (Helixer), TEs (EDTA)
SVs identification: Syri
Graph-based pangenome reconstruction (PGGB)

High Molecular Weight (HMW) DNA extraction

High Molecular Weight (HMW) DNA extraction
→ challenges due to rigid cell walls and secondary metabolites

1. **Nuclei isolation**
2. Nuclei lysis and cleaning by organic extraction
3. Genomic DNA precipitation with high concentration CTAB buffer



Li *et al. Plant Methods* (2020) 16:38
https://doi.org/10.1186/s13007-020-00579-4

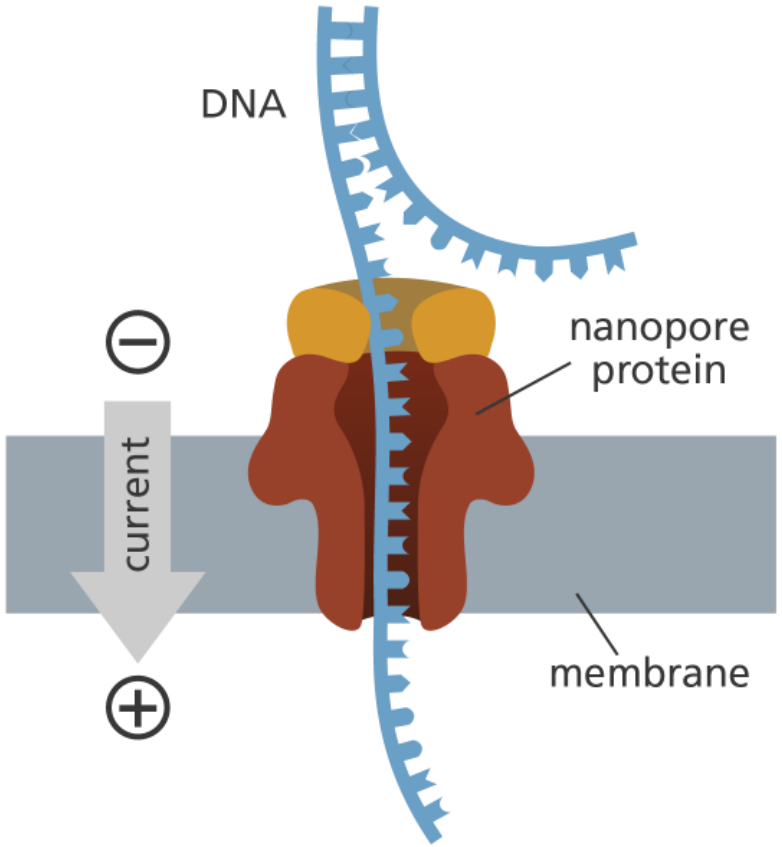
Plant Methods

METHODOLOGY Open Access

A simple plant high-molecular-weight DNA extraction method suitable for single-molecule technologies

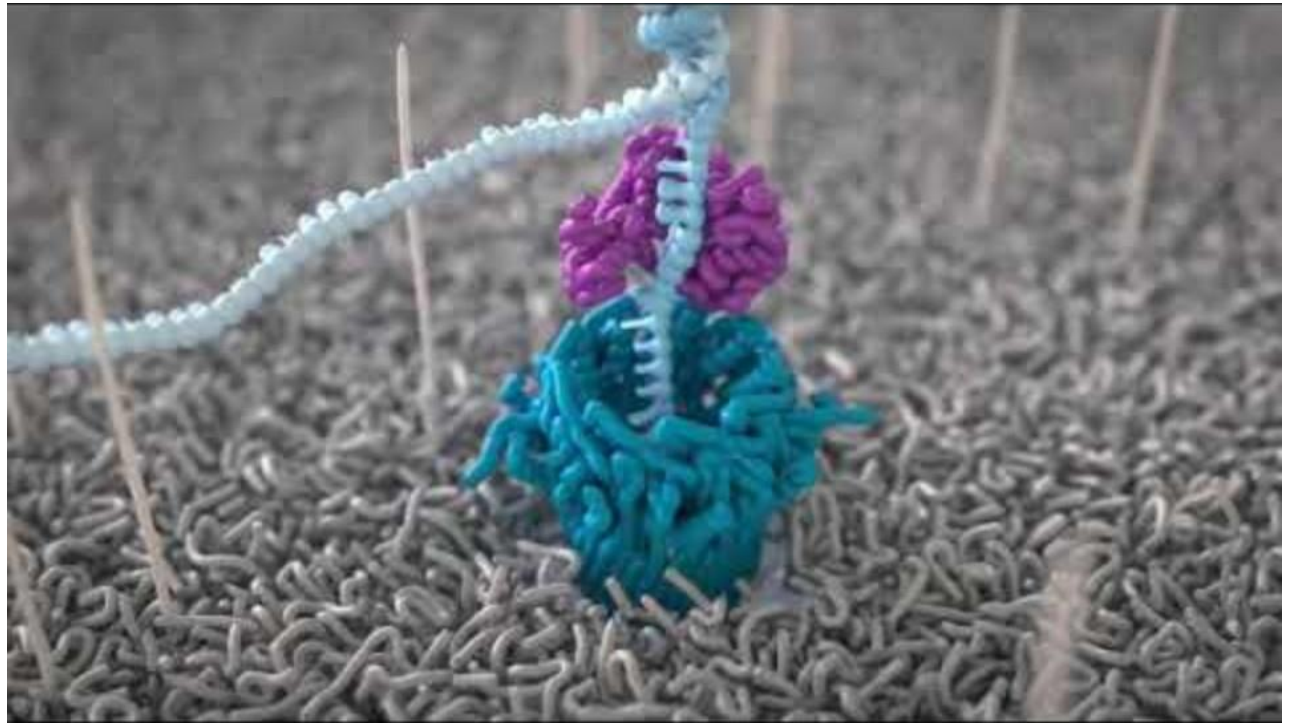
Zhigang Li, Stephen Parris and Christopher A. Saski*

Long read sequencing

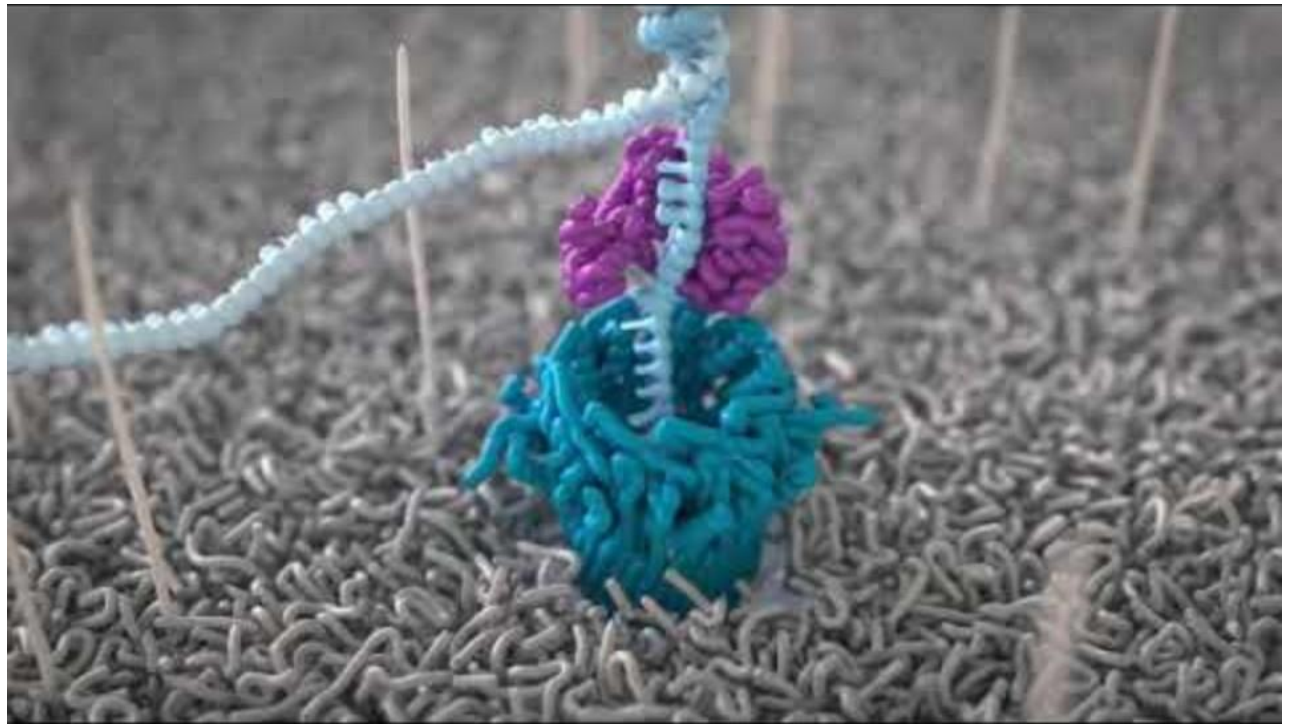


- DNA or RNA molecule through a tiny protein nanopore embedded in a membrane, which creates an ionic current across it
- As the molecule moves, each nucleotide causes a unique disruption in this current
- An electronic sensor measures these changes in real-time
- A base-calling algorithm interprets these signals to determine the exact DNA or RNA sequence.

Long read sequencing

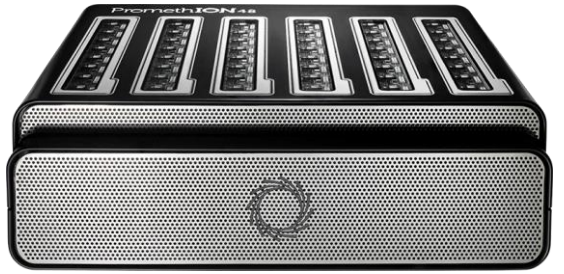


Long read sequencing



High troughput long read sequencing:

→ ***ONT PromethION***

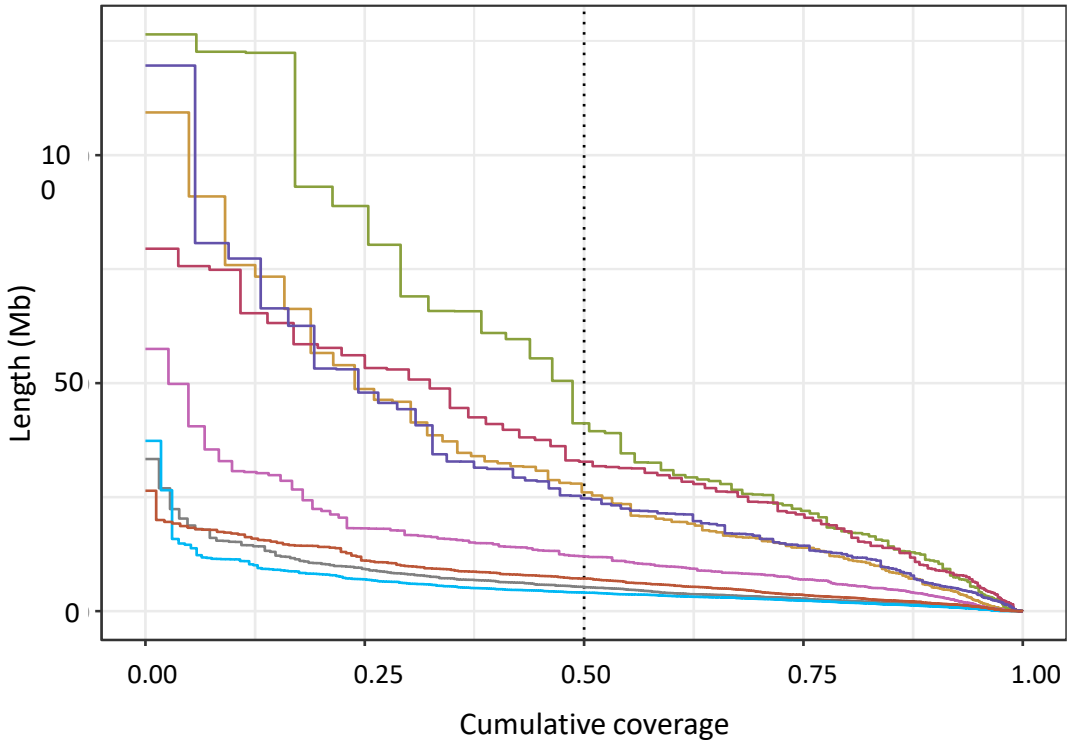


MAGIC maize: *de novo* genome assembly

Founder	N. Contigs	Length tot. (Gb)	Contigs N50 (Mb)	Sequencing protocol
A632	918	2.21	26.04	LR, **UL**
B73	2336	2.15	5.32	LR
B96	1093	2.19	11.99	LR, **UL**
F7	7089	2.11	4.10	LR
H99	637	2.18	41.20	LR, **UL**
HP301	977	2.12	32.77	LR, **UL**
Mo17	1136	2.11	24.72	LR, **UL**
W153R	4302	2.15	7.15	LR

ONT Long-Reads (LR): 10 to 100 Kb
ONT Ultra-Long (UL): > 100 Kb

Max read length: 746 Kb (H99)



MAGIC maize: chromosomes reconstruction

For each genome we choose the closer refence assembly available

FOUNDER	CLOSER REFERENCE AVAILABLE
A632	A632 - Wang et al. 2023
B73	B73 RefGen_v5
F7	F7 - Haberer et al. 2020
HP301	HP301 - Hufford et al. 2021
Mo17	Mo17 - Chen et al. 2023

* Founders for which the exact reference is not available
→ search for the closest one

MAGIC maize: chromosomes reconstruction

For each genome we choose the closer refence assembly available

FOUNDER	CLOSER REFERENCE AVAILABLE
A632	A632 - Wang et al. 2023
B73	B73 RefGen_v5
B96*	Ki 11 - Hufford et al. 2021
F7	F7 - Haberer et al. 2020
H99*	Oh43 - Hufford et al. 2021
HP301	HP301 - Hufford et al. 2021
Mo17	Mo17 - Chen et al. 2023
W153R*	A188 - Lin et al. 2021

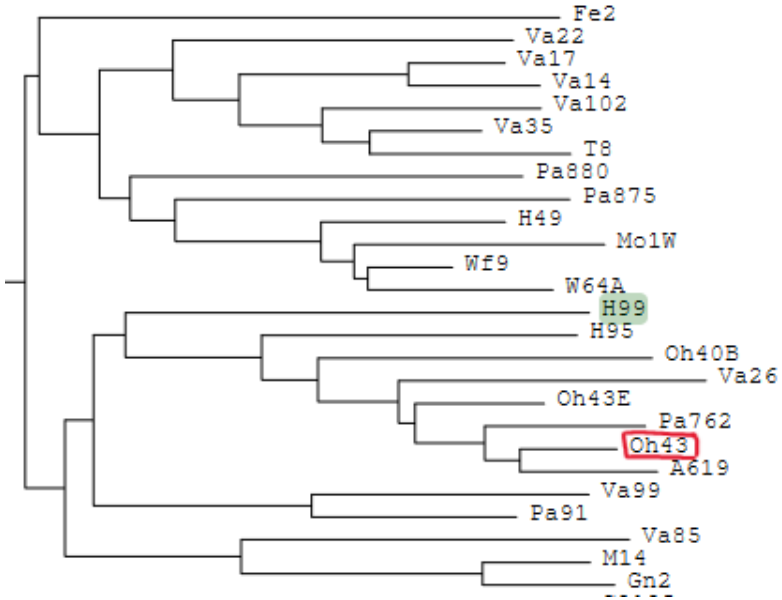
* Founders for which the exact reference is not available
→ search for the closest one

Genetic Structure and Diversity Among Maize Inbred Lines as Inferred From DNA Microsatellites

FREE

Kejun Liu, Major Goodman, Spencer Muse, J Stephen Smith, Ed Buckler, John Doebley

Genetics, Volume 165, Issue 4, 1 December 2003, Pages 2117–2128, https://doi.org/10.1093/genetics/165.4.2117



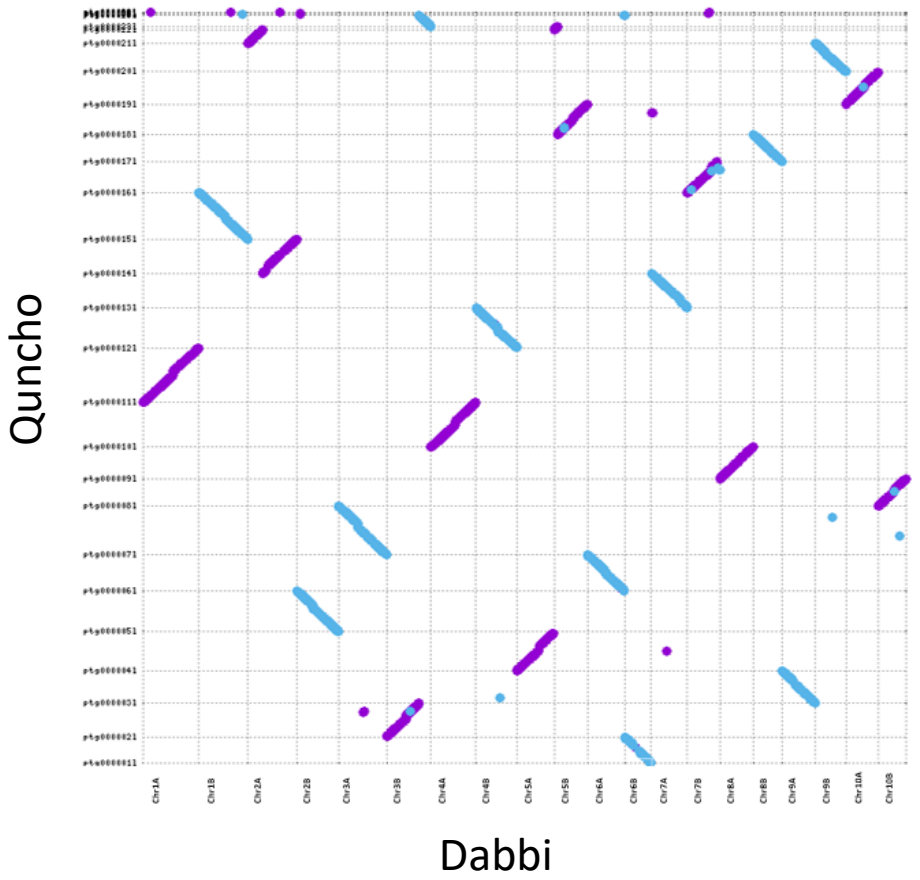
Sant'Anna
School of Advanced Studies – Pisa

Teff NAM: *de novo* genome assembly

Teff is an **allotetraploid** genome **(2n=4x=40)** chromosomes
→ Reference genome: Dabbi

Genome	contigs	Largest contig (Mb)	Total length (Mb)	N50 (Mb)
Dabbi	874	40.62	575.1	27.14
Quncho (T2T)	22	43.32	592	30.84

Assembled contigs almost reach chromosome level
→ Manual curation to merge few contigs



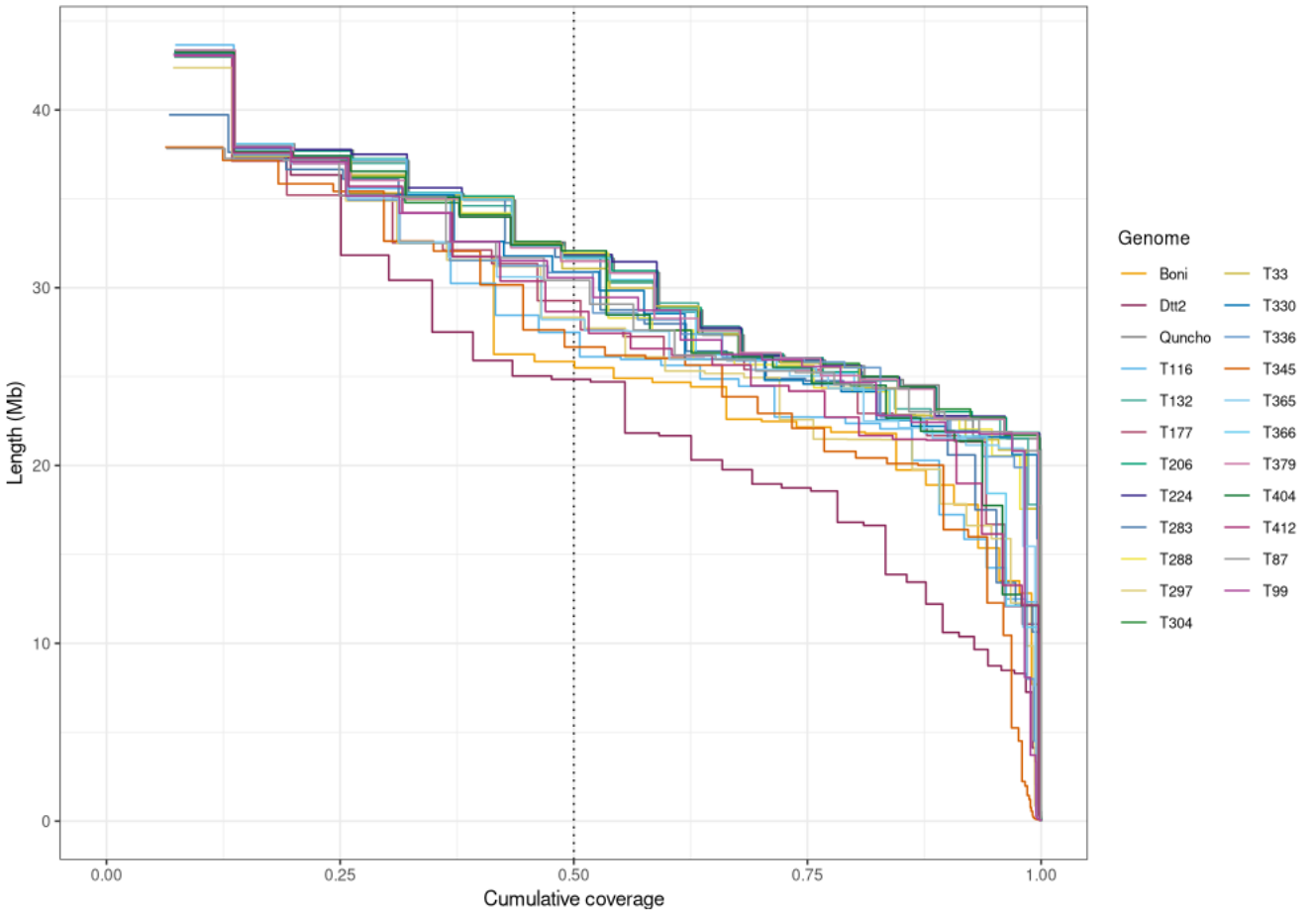
Teff NAM: *de novo* genome assembly

Teff is an **allotetraploid** genome **(2n=4x=40)** chromosomes
→ Reference genome: Dabbi

Genome	contigs	Largest contig (Mb)	Total length (Mb)	N50 (Mb)
Dabbi	874	40.62	575.1	27.14
Quncho (T2T)	22	43.32	592	30.84

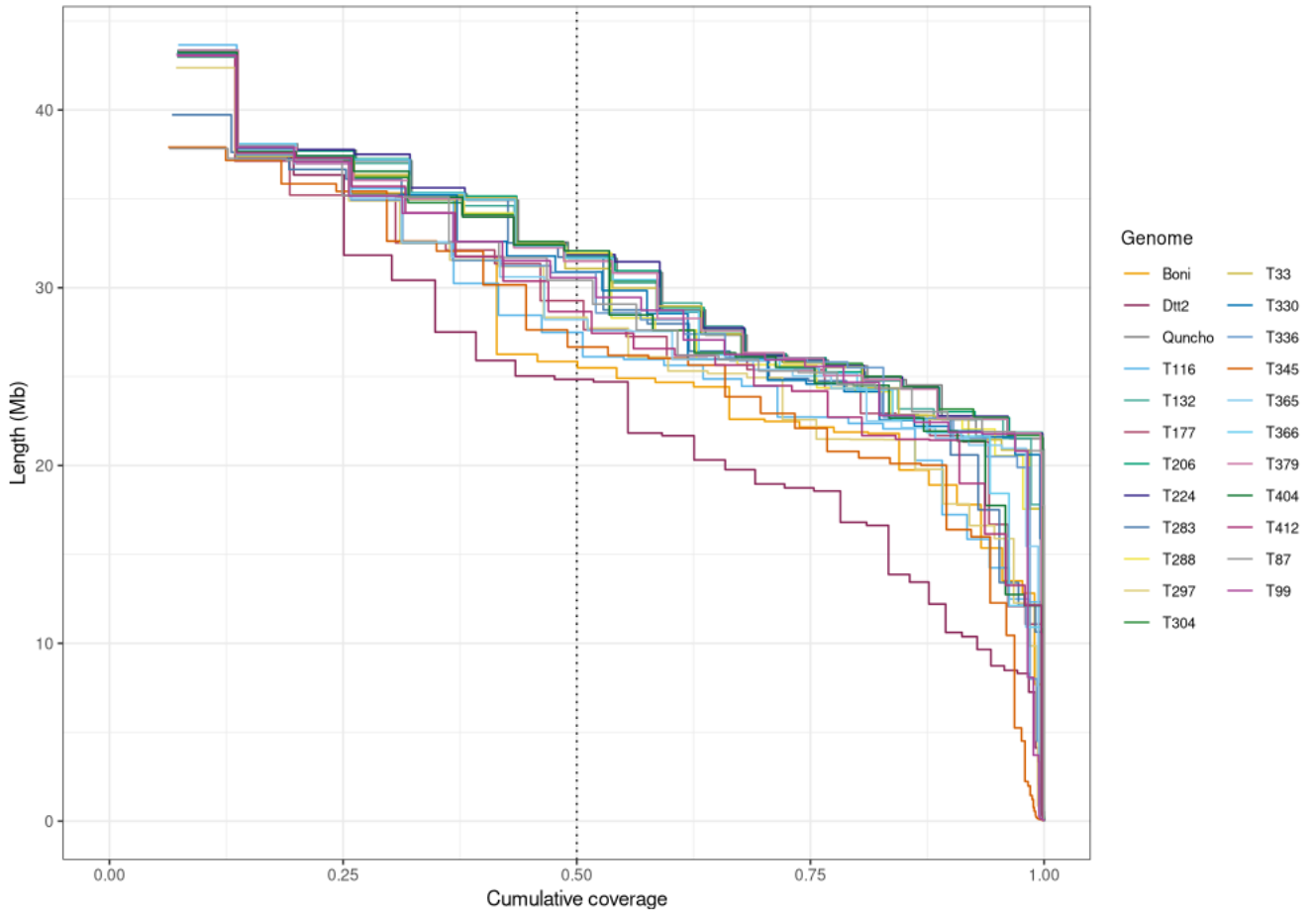
Assembled contigs almost reach chromosome level
→ Manual curation to merge few contigs

T2T level assembly for Quncho



NAM founders' assemblies

- Average assembly size: 593 Mb
- Average contigs N50: 29 Mb
- Number of contigs: from 22 to 60



NAM founders' assemblies

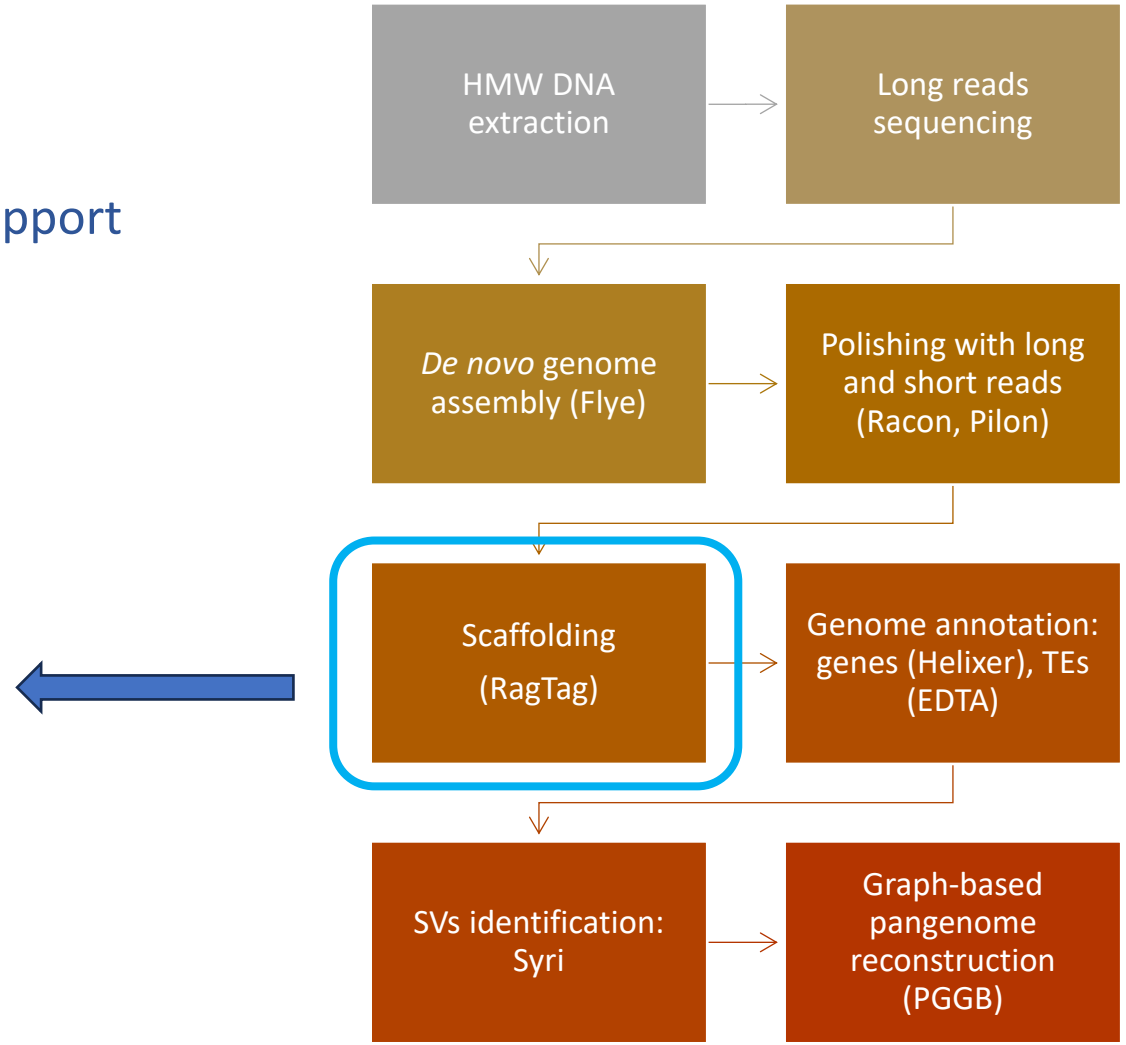
- Average assembly size: 593 Mb
- Average contigs N50: 29 Mb
- Number of contigs: from 22 to 60

Chromosomes reconstruction
→ Scaffolding was performed with RagTag using Quncho as reference to reconstruct pseudo-chromosomes

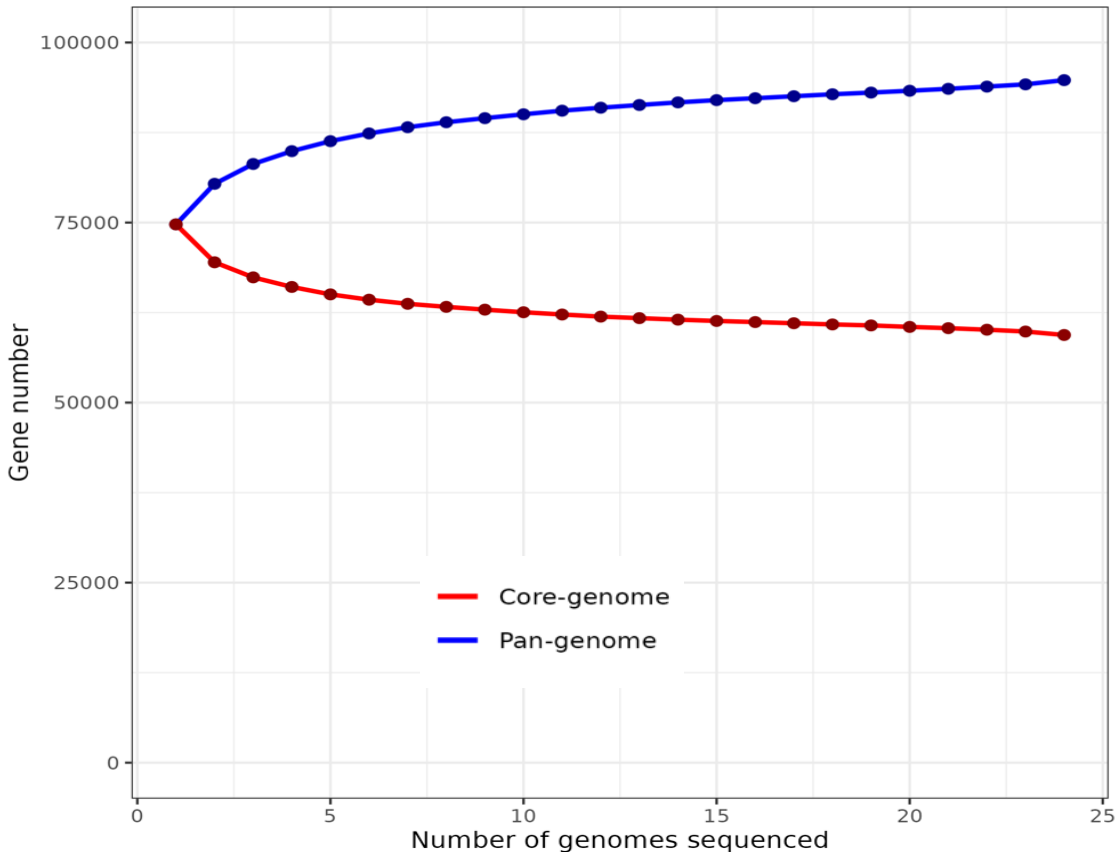
Both projects have the same objective:

Developing an advanced genomic resource to support QTL mapping with multiparental populations

- Teff NAM pangenome: population still under development → whole-genome comparison between genomes
- MAGIC maize pangenome: since the population is well established → focus on specific genomic loci identified by QTL mapping



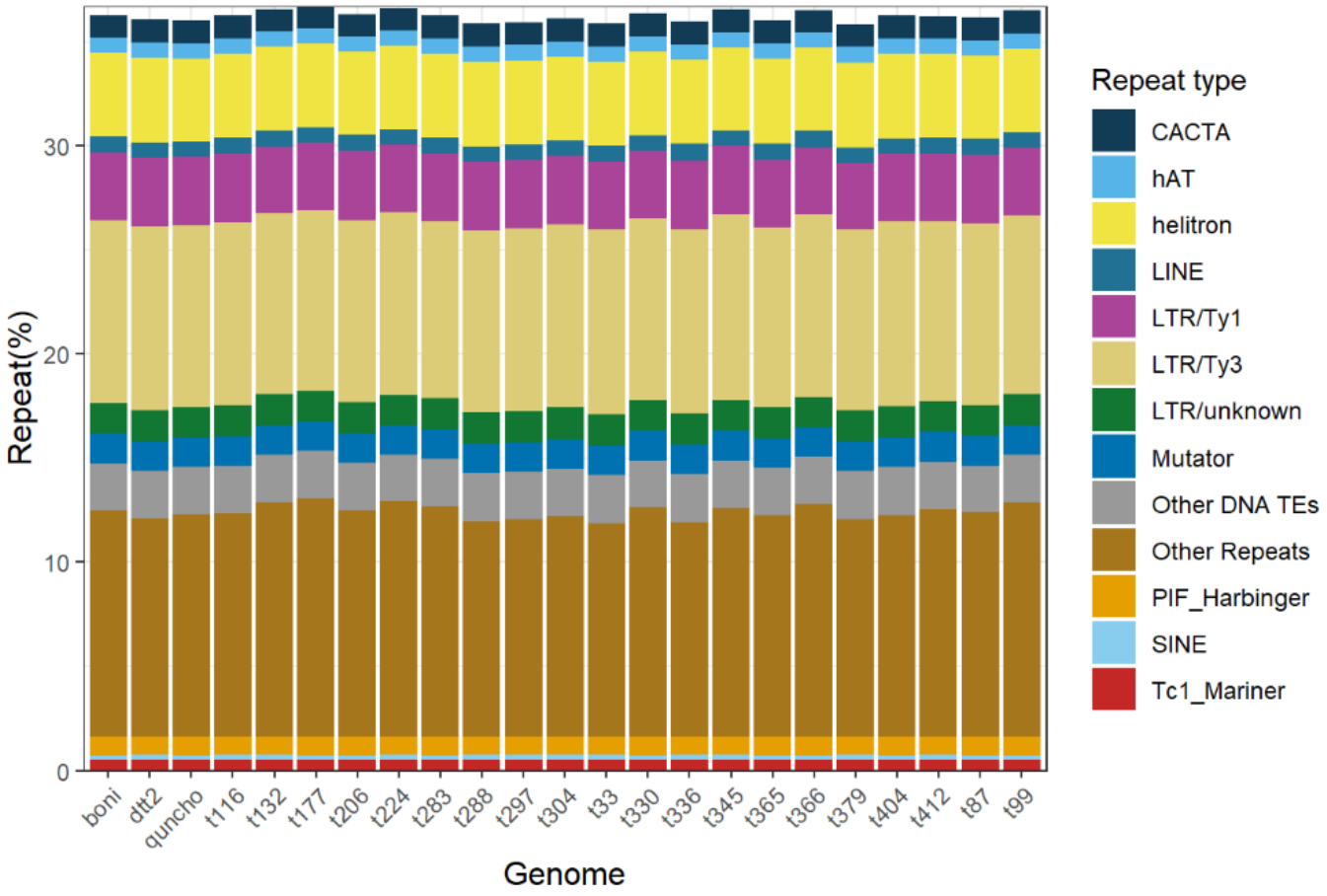
Teff NAM: gene annotation



Helixer: evidence-free *ab initio* predictions of gene models, based on Deep Neural Networks and Hidden Markov Model.

The number of genes shared among multiple founders ('Core', red curve) decreases as more genomes are included, while the number of dispensable genes ('Pan', blue curve) increases

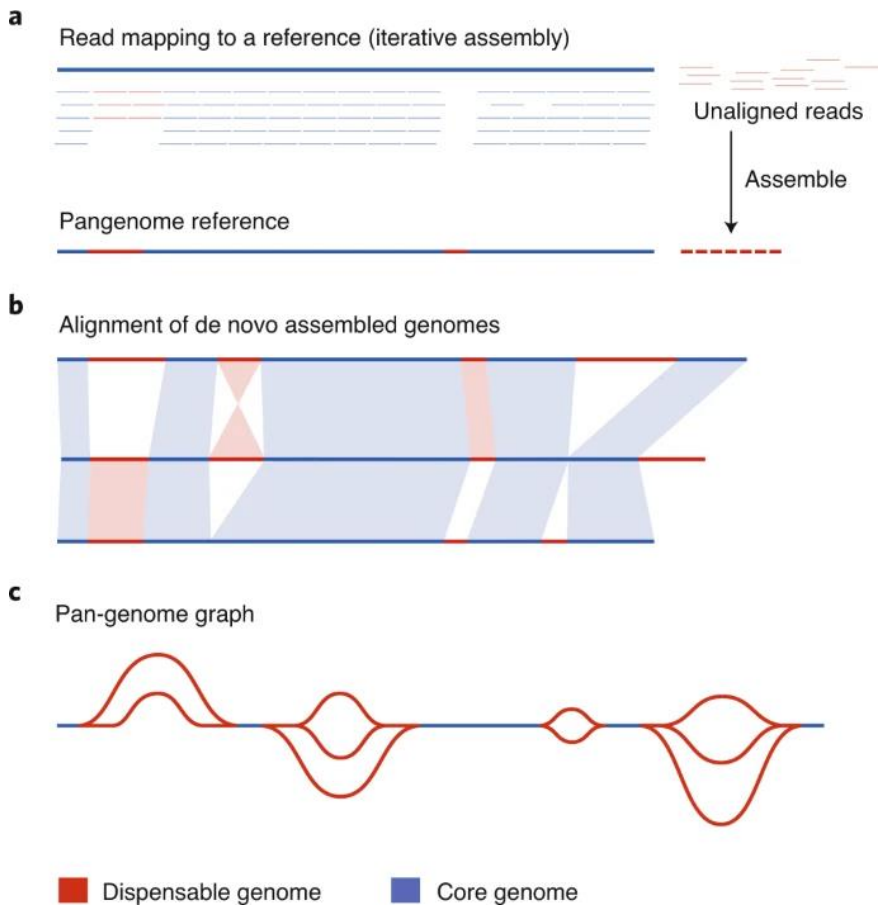
Teff NAM: TEs annotation



PanEDTA : *de novo* annotation of TEs and repetitive sequences

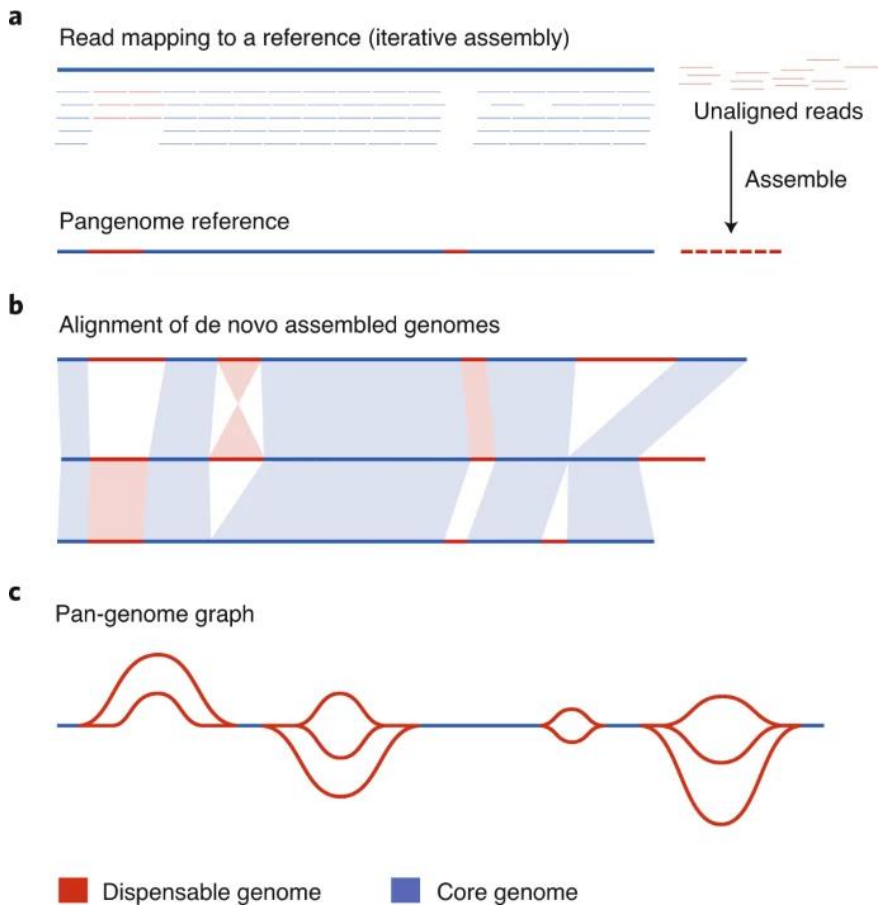
36 % of teff genome composed by TEs and repetitive sequence

Combining assemblies



a) Alignment of reads from multiple samples to a reference followed by assembly of unaligned reads into novel contigs

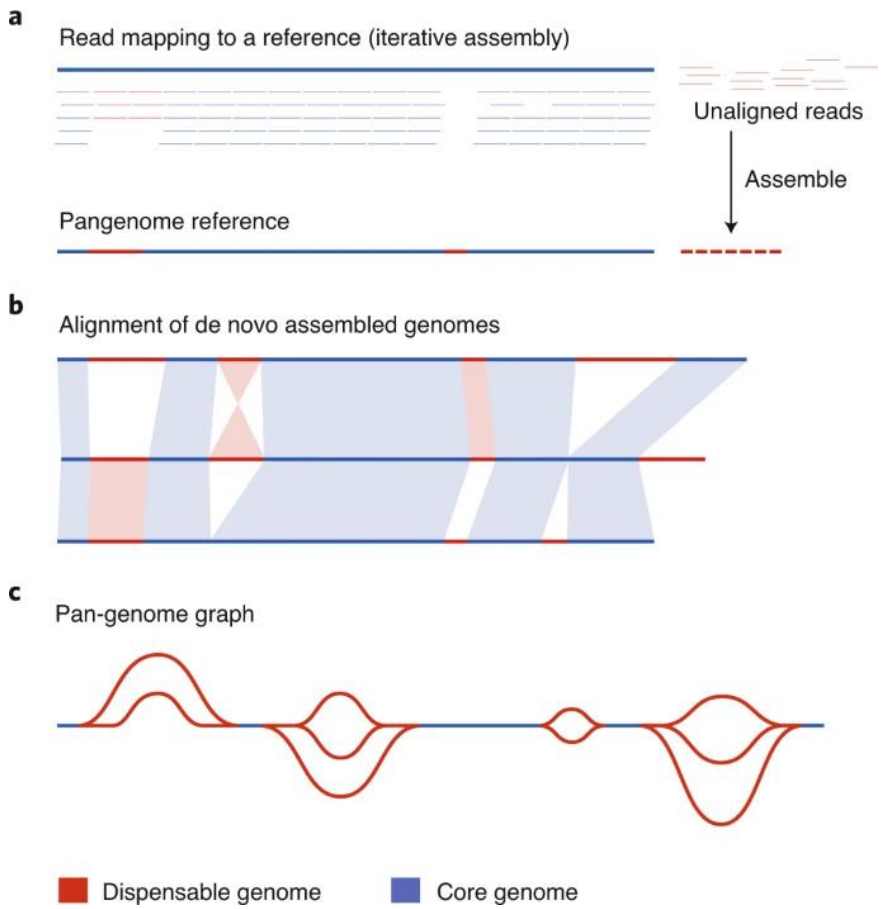
Combining assemblies



a) Alignment of reads from multiple samples to a reference followed by assembly of unaligned reads into novel contigs

b) De novo assembly of the genomes of multiple accessions allows whole genome alignment approaches to identify dispensable genomic regions

Combining assemblies

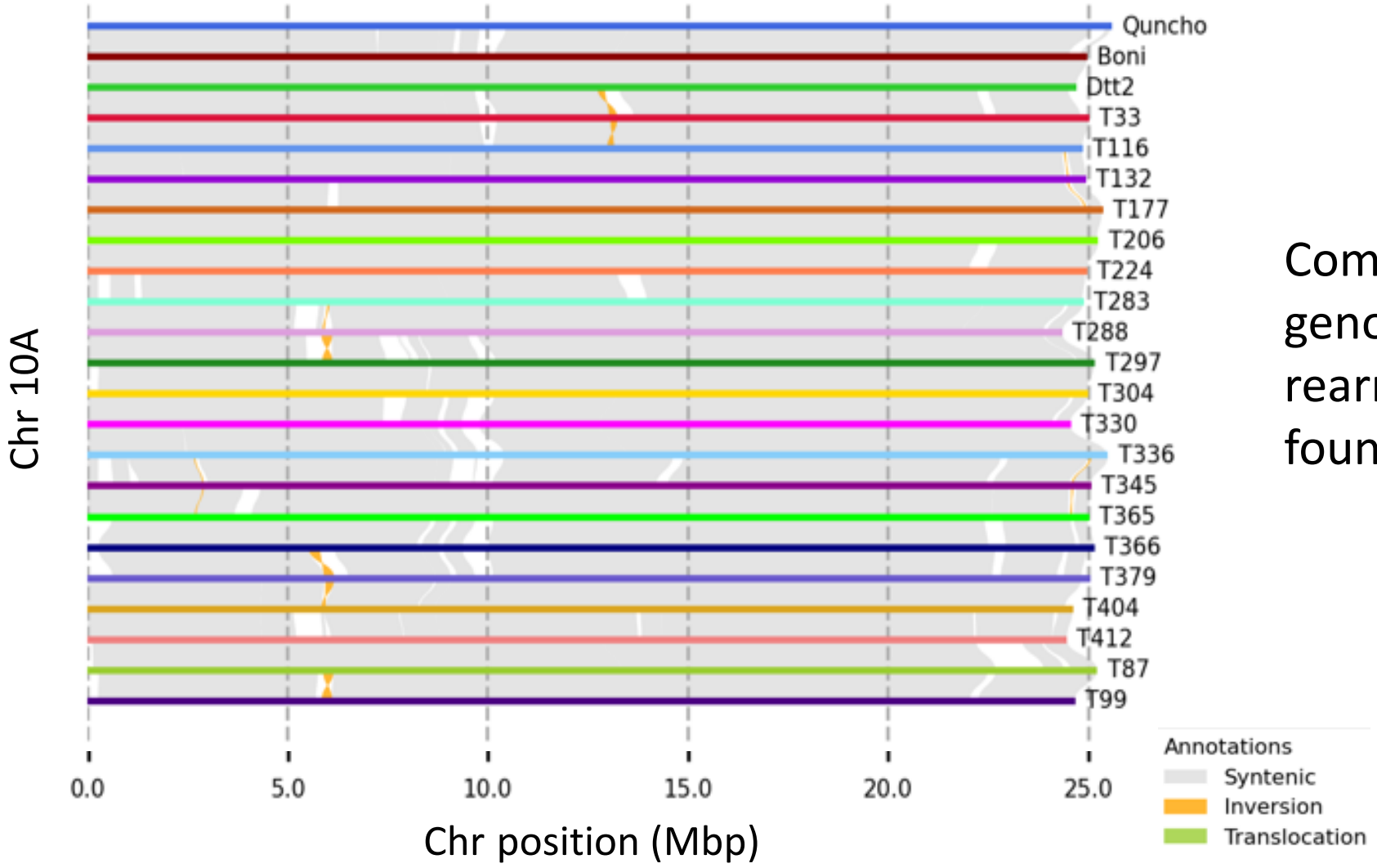


a) Alignment of reads from multiple samples to a reference followed by assembly of unaligned reads into novel contigs

b) De novo assembly of the genomes of multiple accessions allows whole genome alignment approaches to identify dispensable genomic regions

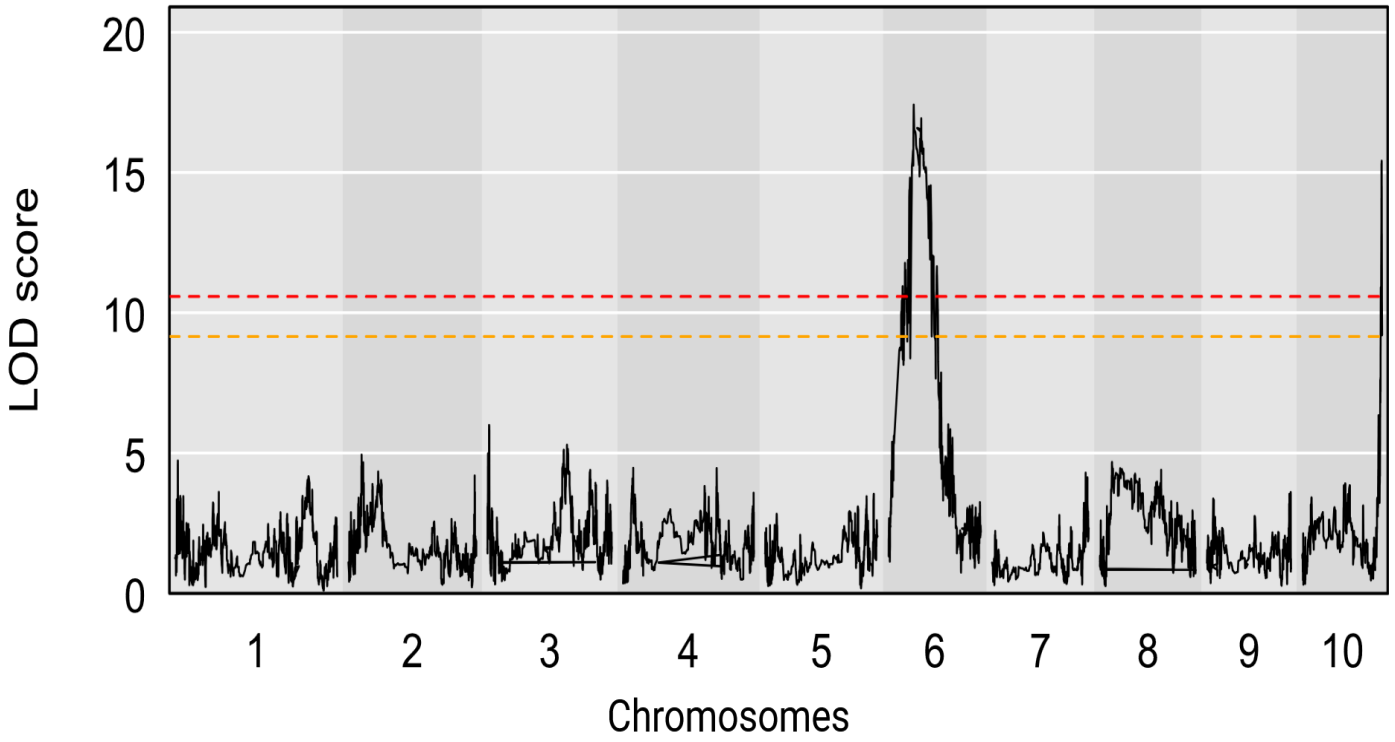
c) A pan-genome graph can be constructed from whole genome alignments, and efficiently stores variant information of dispensable regions as unique paths through the graph

SVs calling and synteny analysis, Teff



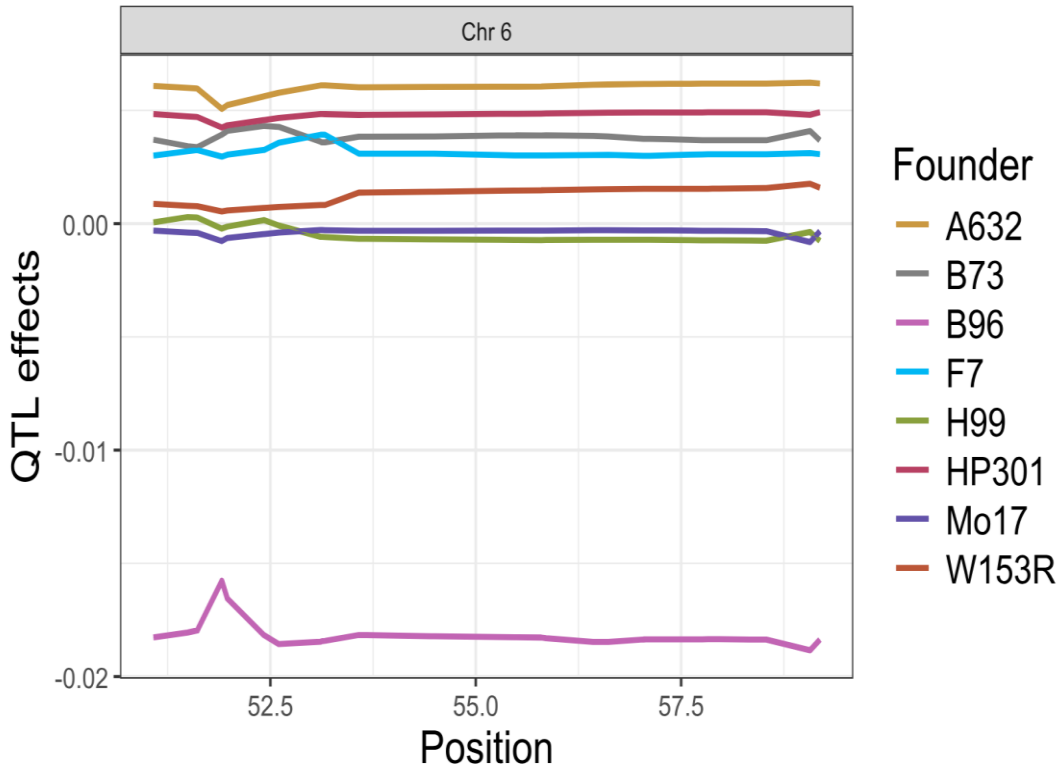
Comparative analysis of genome synteny and rearrangement across the founders' genomes

Leveraging the pan-genome to dissect complex traits: maize photosystem II (PSII) operating efficiency



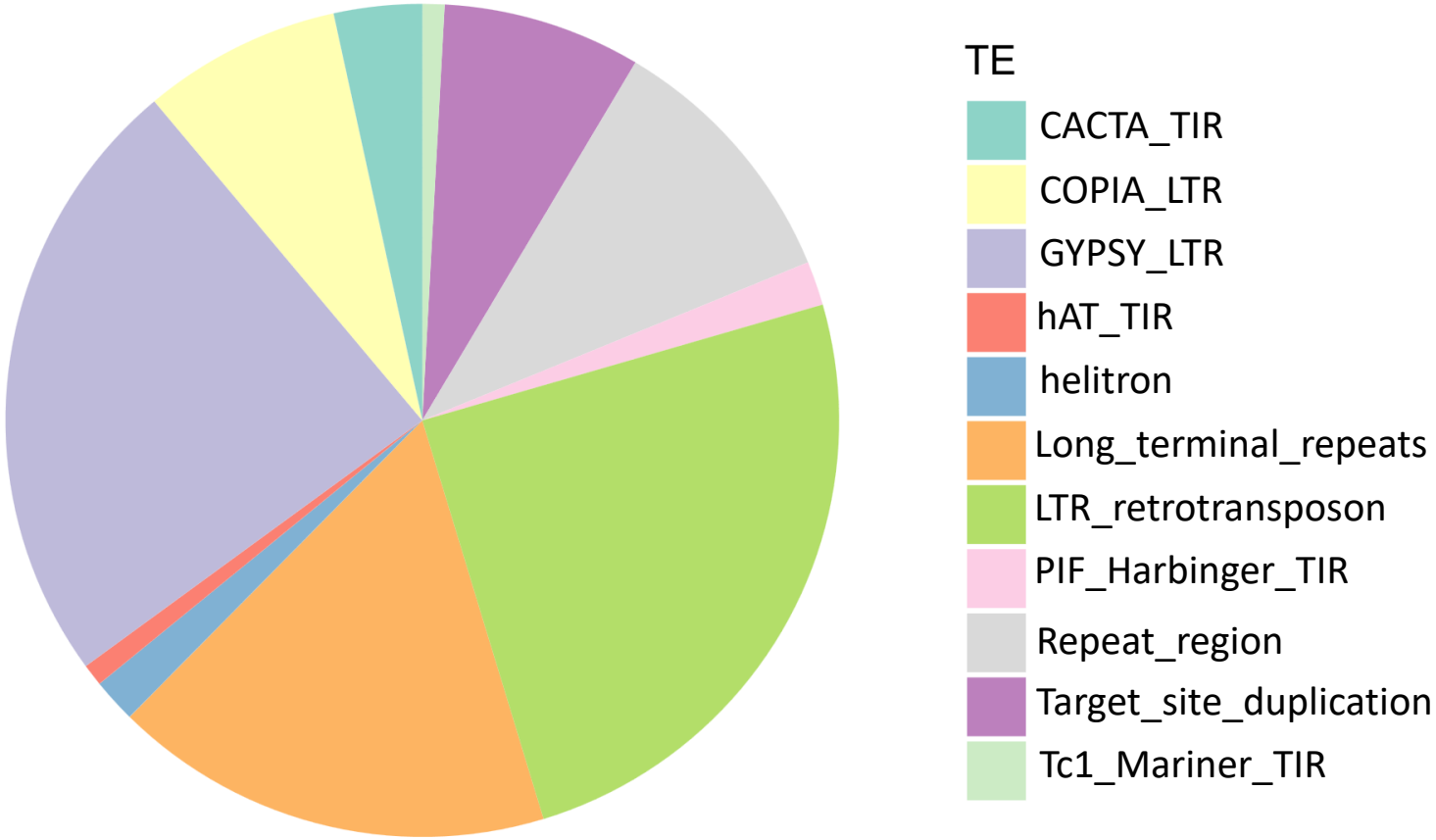
PSII operating efficiency shows a major QTL on the peri-centromeric region of Chr 6. Red line and yellow line represent stringent (< 1% false positive rate tolerated) and standard (< 5% false positive rate tolerated) thresholds

Probabilistic estimation of founder haplotypes:
SNP markers were used to probabilistically reconstruct the RILs genomes using a Hidden Markov Model

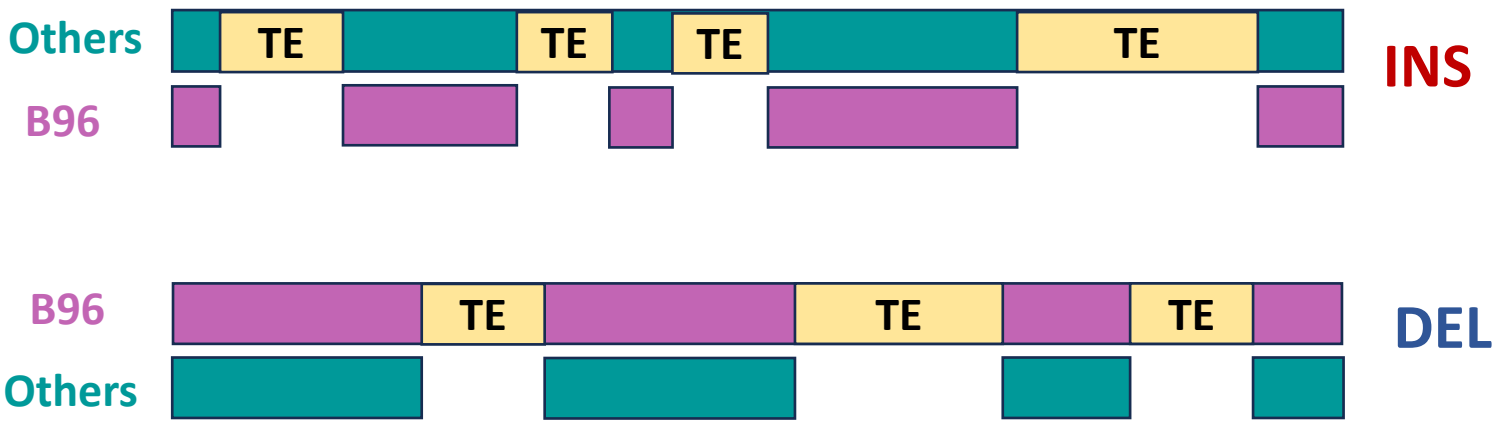


On the QTL on **chr 6** we identified **B96** as the founder haplotype exerting the strongest phenotypic effect

B96 at this chromosomic region, harbored **exclusive SVs** that consistently map over **annotated TEs**



B96 at this chromosomic region, harbored **exclusive SVs** that consistently map over **annotated TEs**

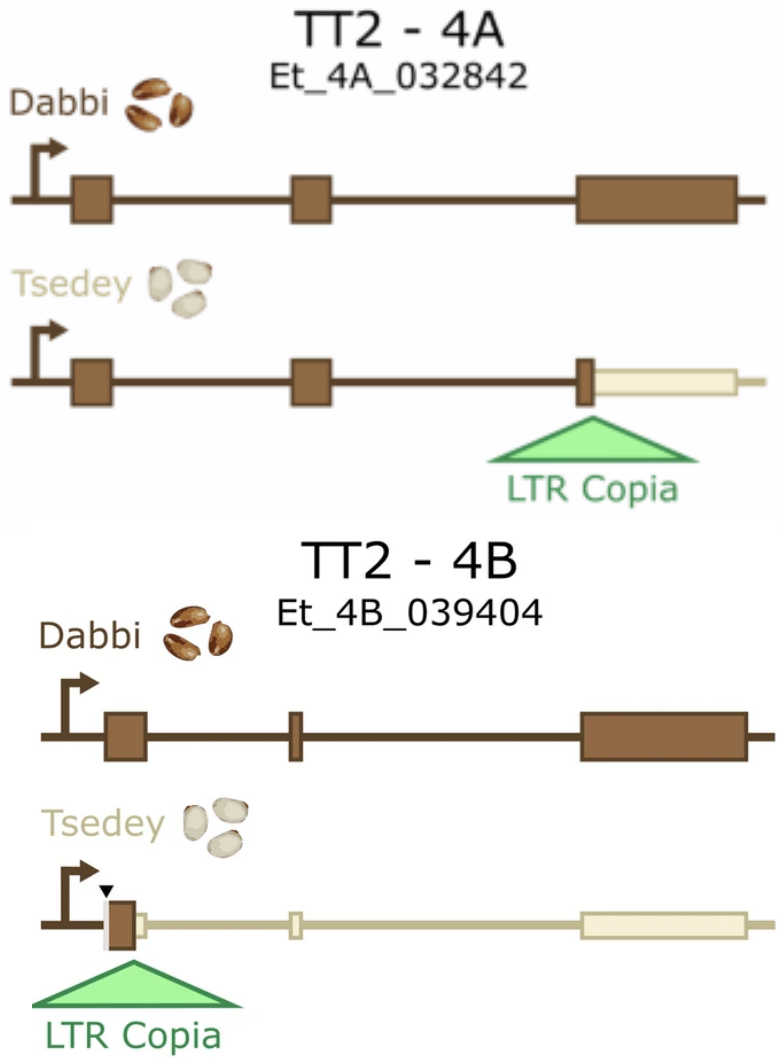


Looking at insertions (INS) shared among all founders we could pinpoint the chromosomic regions missing in B96. Chromosomic intervals unique to B96 were detected by identifying the regions deleted in all the others

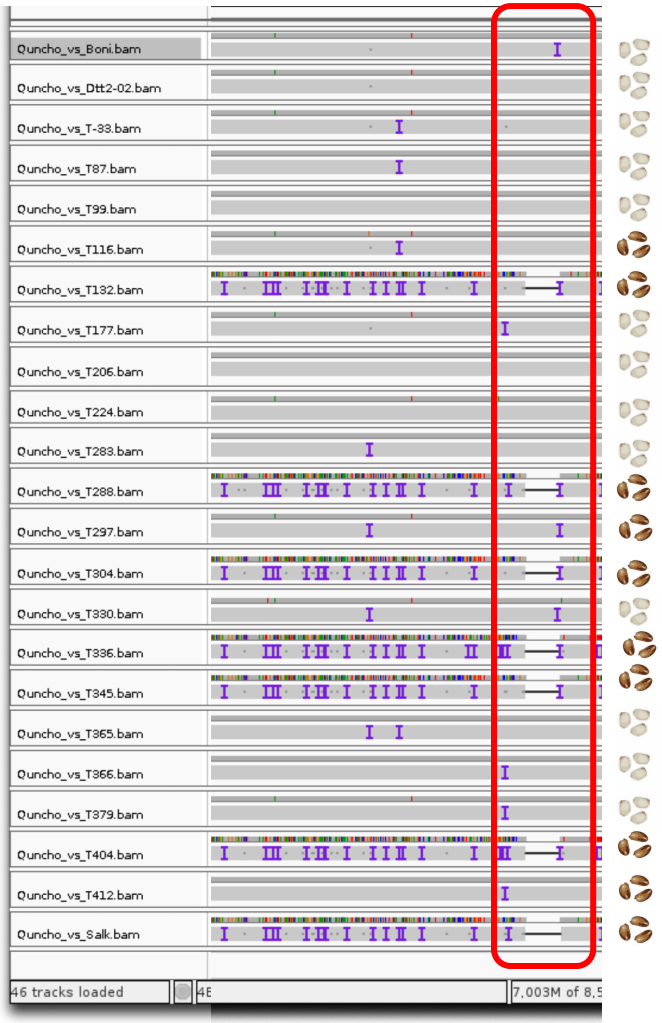
Trait dissection in Teff: seed color

NAM RILs are not available yet
→ we can only evaluate genomic differences between founders at loci previously discovered

TRANSPARENT TESTA 2 (TT2) is a candidate gene identified for teff seed color and highlight the putative effect of TE insertions in white seed phenotype

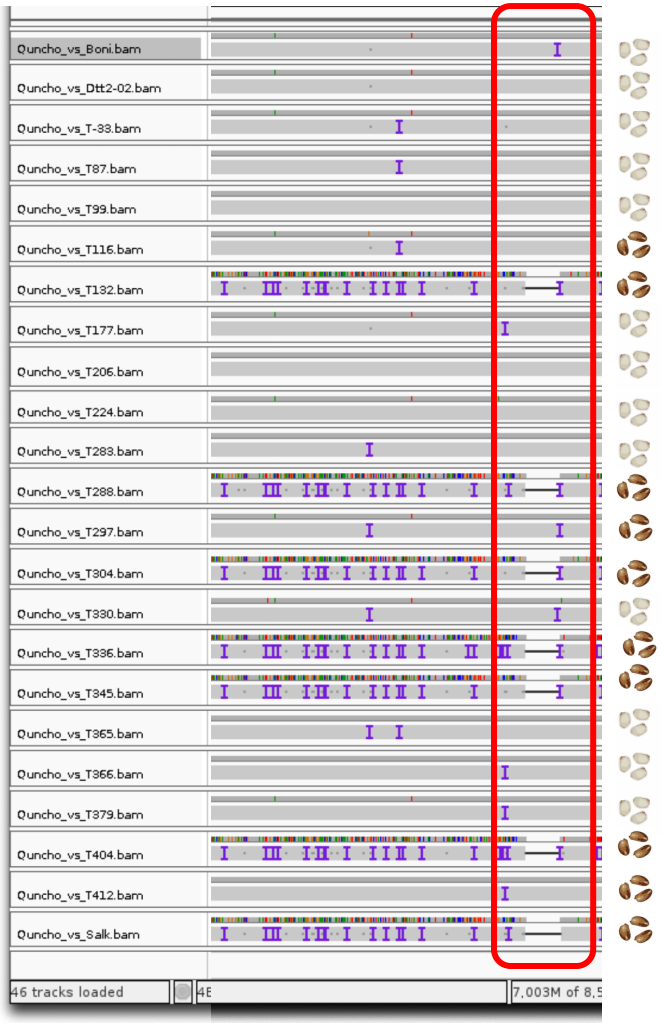


Trait dissection in Teff: seed color



Quncho and all white lines:
TE insertion of 5.141bp

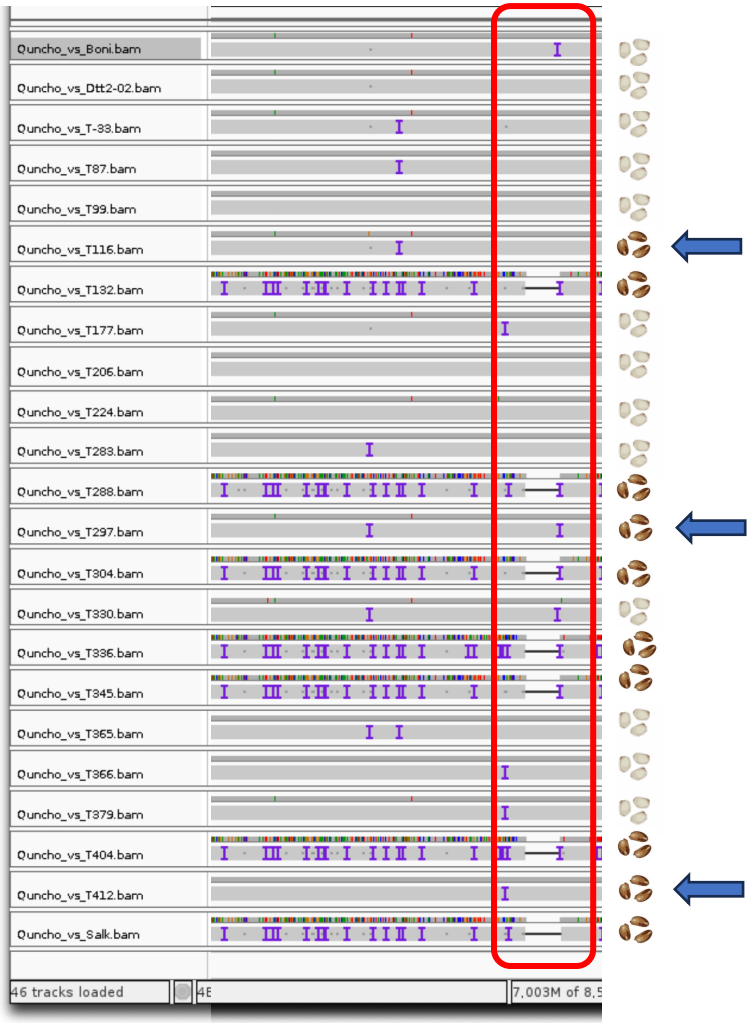
Trait dissection in Teff: seed color



Quncho and all white lines:
TE insertion of 5.141bp

Only brown lines do not have
the TE insertion

Trait dissection in Teff: seed color



Quncho and all white lines:
TE insertion of 5.141bp

Only brown lines do not have the TE insertion

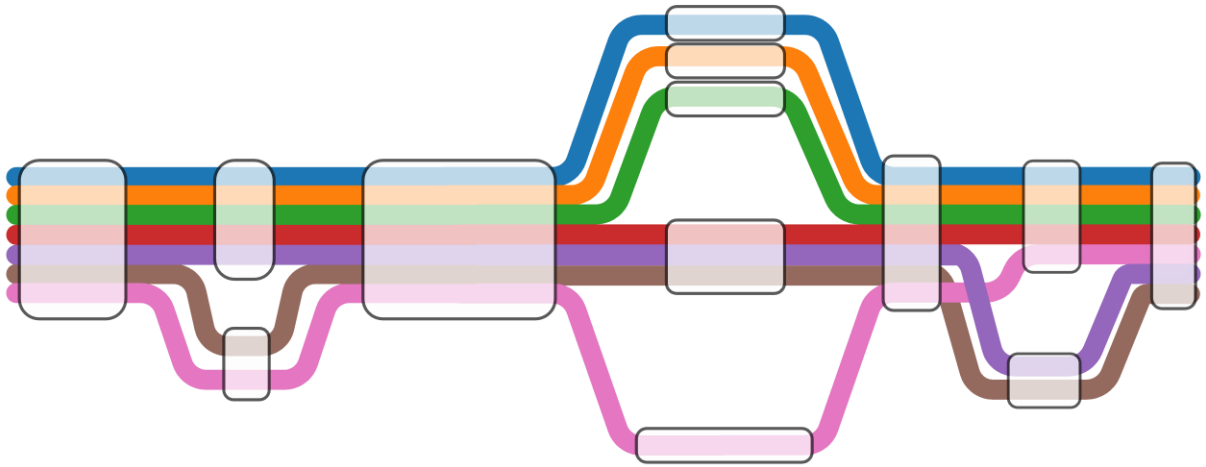
3 brown lines present the TE

Other loci involved?

Still to explore

Graph-based pangenome reconstruction

- Linear reference genomes → coordinate system to track the locations of many genomic features including genes and variants
- Graph-based pangenome → multiple different paths and different lengths of genomic sequences, this means that bases can no longer be numbered sequentially



Nodes: identical or similar sequences

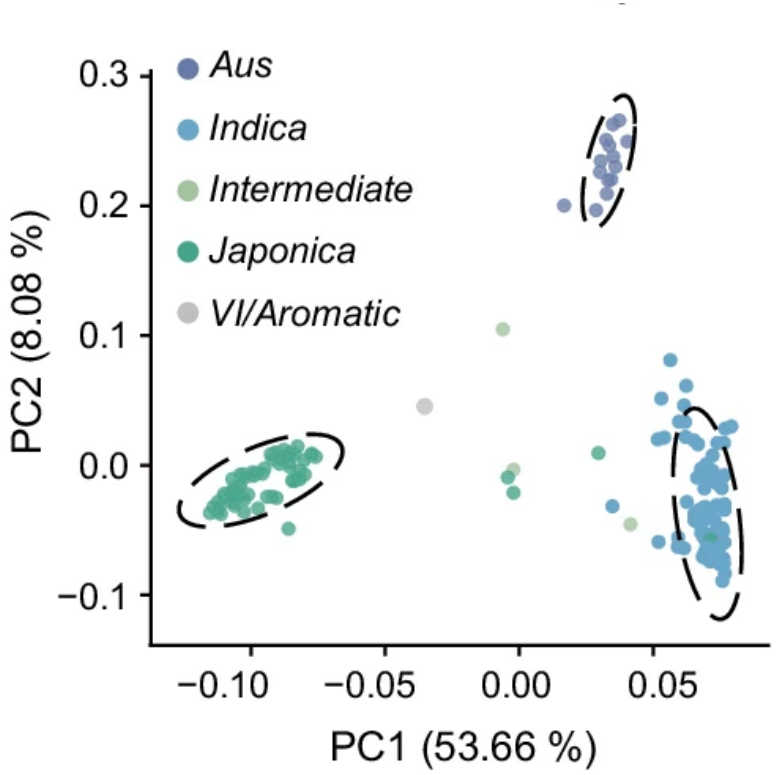
Edges: variation

Thank you

Sant'Anna
School of Advanced Studies – Pisa

The role of TEs proliferation

Lineage-specific amplification / deletion, of TEs is common in plants, even among closely related species, such as between subspecies of domesticated rice, *Oryza sativa* subsp. *indica* and subsp. *japonica*



Principal Component Analysis based on transposable element insertion polymorphisms reveals distinct clustering patterns, indicating that these polymorphisms contribute significantly to the genetic structure of the population.

Qian et al., 2025 **Pangenome** analysis of transposable element insertion polymorphisms reveals features underlying cold tolerance in rice