UPOV
International Union for the Protection of New Varieties of Plants

E

Technical Working Party on Testing Methods and Techniques **TWM/3/10**

Third Session
Beijing, China, April 28 to May 1, 2025

Original: English
Date: April 22, 2025

EXPLOITING CROP HAPLOTYPE-TAG POLYMORPHISMS MARKER FOR PEDIGREE IDENTIFICATION

Document prepared by an expert from China

Disclaimer: this document does not represent UPOV policies or guidance

The annex to this document contains a copy of a presentation “Exploiting Crop Haplotype-tag Polymorphisms Marker for Pedigree Identification”, to be made by an expert from China, at the third session of the TWM.

[Annex follows]

2025 UPOV/Technical Working Party on Testing Methods and Techniques (TWM/3)

Exploiting Crop Haplotype-tag Polymorphisms Marker for Pedigree Identification

Presented by Yikun Zhao

Maize Research Institute,
Beijing Academy of Agriculture and Forestry Sciences
Beijing Key Laboratory of Maize DNA Fingerprinting and Molecular Breeding

April 28 to May 1, 2025, Beijing, China

1

Content

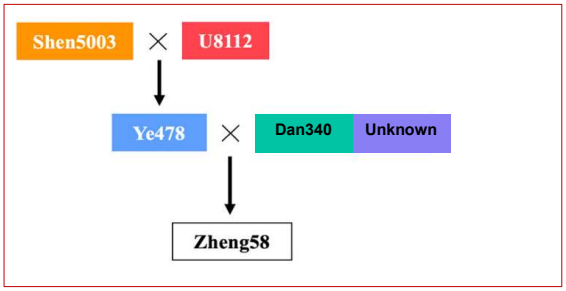
2

Part I
Background

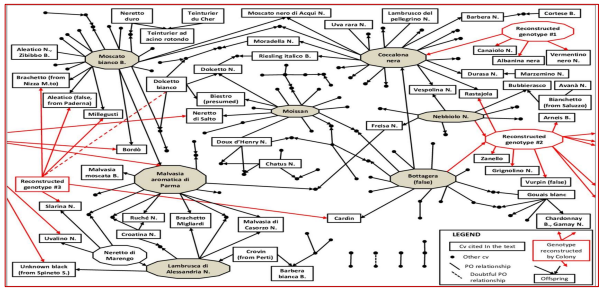
3

1.1 Crop Pedigree

- **Significance:** Crop pedigrees documentation serves as a **foundational data** repository for intellectual property protection in agriculture, delineating genetic ancestry, evolutionary trajectories, and material provenance within breeding programs.
- **Incomplete Pedigree Issues:** Systematic gaps in germplasm pedigree annotation persist due to historical breeding practices, anthropogenic interventions, and archival, resulting in **fragmented**, **incomplete**, or **unreliable documentation**.
- **Technical requirements:** Conventional pedigree mapping methodologies relying on tabular representations and simplified kinship diagrams lack comprehensive representation of **genomic architecture**. Precise pedigree identification and reconstruction can effectively safeguard the rights and interests of breeders and farmers.



Unknown Maize pedigree identification. (Zhang *et al*., 2024)

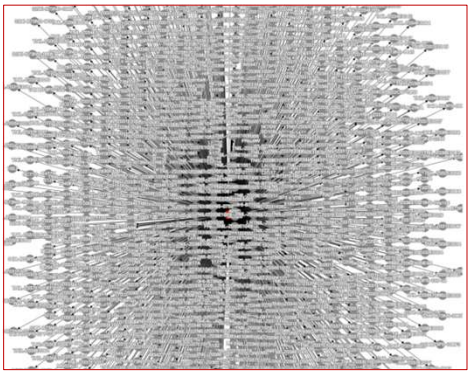


Kinship network of the traditional grape varieties. (Raimondi *et al*., 2020)

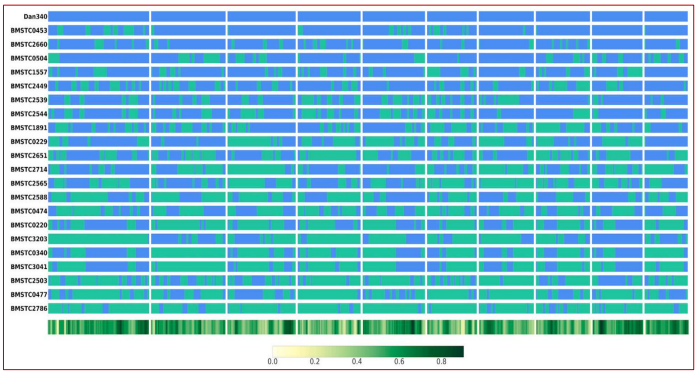
4

1.1 Crop Pedigree

- **Complexity of Pedigree Inference and Reconstruction:** Crop reconstruction confronts significant technical challenges stemming from **historical data fragmentation**, **inconsistent data annotation**, and **polygenic inheritance patterns in contemporary cultivars** — a predicament particularly evident in species like *Zea mays L.* with multigenerational selective breeding histories.



Known Maize Variety-B73 Pedigree Diagram. (Braun *et al.*, 2019)

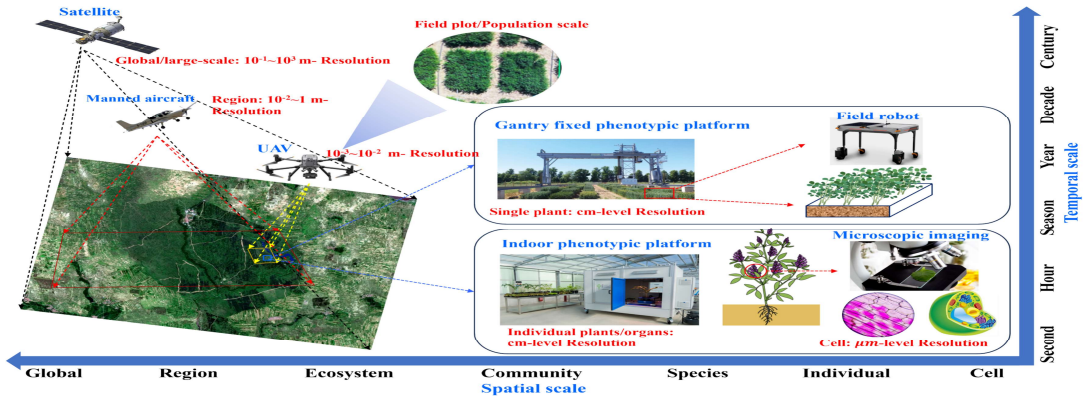


Known Maize Variety-Dan340 derived lines Diagram. (Zhang *et al.*, 2024)

5

1.2 Common Pedigree Identification Technologies

- **Phenotypic Identification**: Although phenotypic identification remains predominantly employed in plant variety characterization, it exhibits some limitations in achieving precise pedigree identification and reconstruction for varieties.



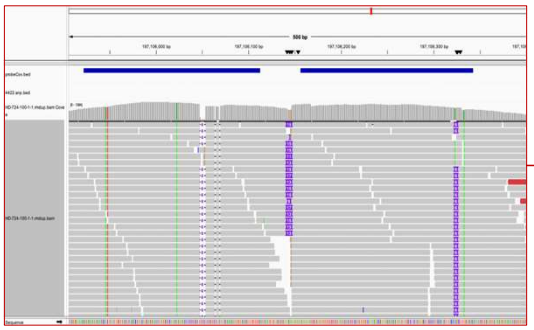
High-throughput phenotyping techniques. (Cheng *et al.*, 2025)

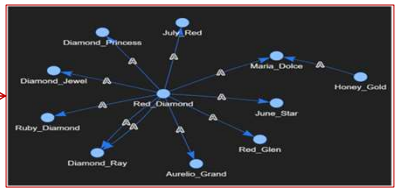
6

1.2 Common Pedigree Identification Technologies

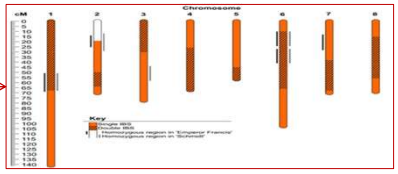
- **Molecular Marker**: The identification of crop pedigrees currently relies on employing molecular markers directly to calculate genetic similarity or detect long genomic segments for inference.

a. **SNP** (Single Nucleotide Polymorphism);
b. **InDel** (Insertion/Deletion Polymorphism);
c. **SSR** (Simple Sequence Repeat).





Jurado-Ruiz *et al*., 2022

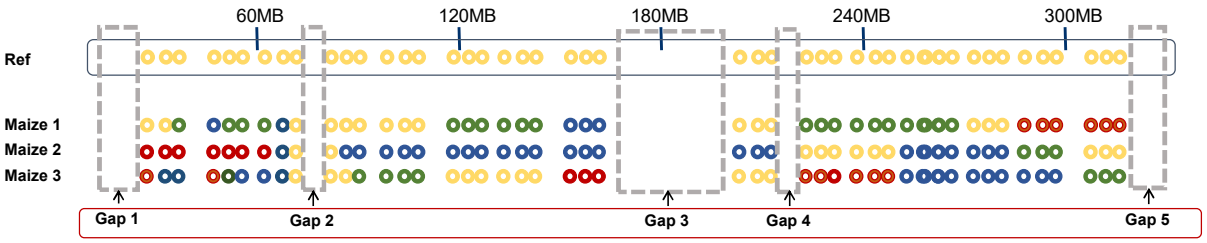


Howard et al., 2021

7

1.2 Common Pedigree Identification Technologies

- **Point Markers** (SNP/InDel/SSR): From a whole-genome perspective, SNP/InDel/SSR can be classified as 'point markers'. They may leave gaps in the genome, preventing seamless coverage.



These gaps may hold breed-specific genetic information, which could directly affect the accuracy of pedigree inference. Therefore, is there a type of marker that can fill these gaps while also possessing very high polymorphism?

8

1.3 Block/Bin Map

- **Block/Bin:** Currently, in plant breeding and genetics research, Block/Bin methodologies are utilized for genomic analysis and functional studies.

a. LD-based block:

Through evaluating allelic association patterns across single nucleotide polymorphisms (SNPs), genomic regions exhibiting elevated linkage disequilibrium are identified as conserved haplotype blocks, representing stable genetic transmission units.

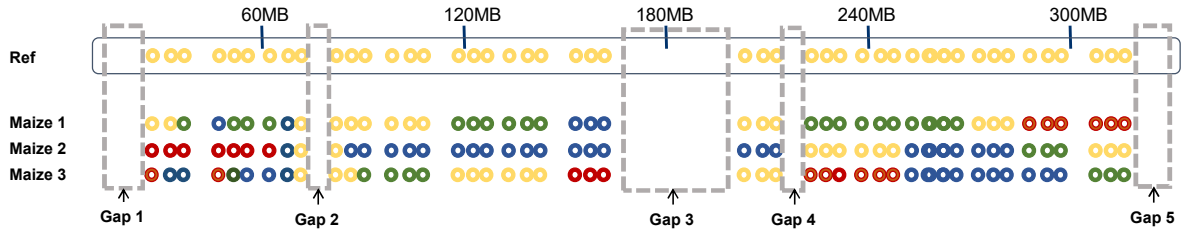
b. Recombination-based block:

Standardized genomic intervals facilitate the consolidation of co-segregating SNPs into recombination blocks, thereby streamlining marker selection. Genotype assignment is subsequently achieved through employing computational.

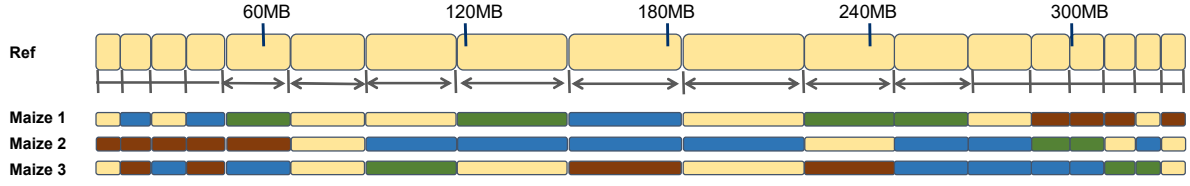
9

1.4 Seamless Integration vs. Dispersed Points

Point Markers:



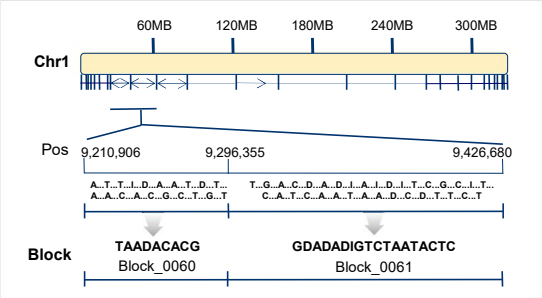
- **Block Markers:** If all blocks are developed into 'Block Markers', this novel marker type may partition and integrate polymorphic markers to fill the gaps, thus enabling precise identification of crop pedigrees.

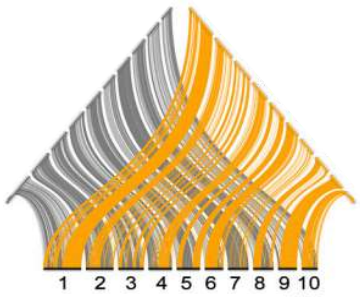


10

1.5 Block Marker-Based Pedigree Reconstruction in Crops

- **Technical Design Scheme：** Designed as contiguous blocks that span entire chromosomes, integrating multiple polymorphic loci (e.g., SNP, InDel, SSR) within each block, achieving full-genome continuous coverage. Acquiring more detailed and extensive variety information enhances the **precision of pedigree identification**.





Block markers are poised to become a more suitable type of marker for variety identification, especially in the context of pedigree identification.

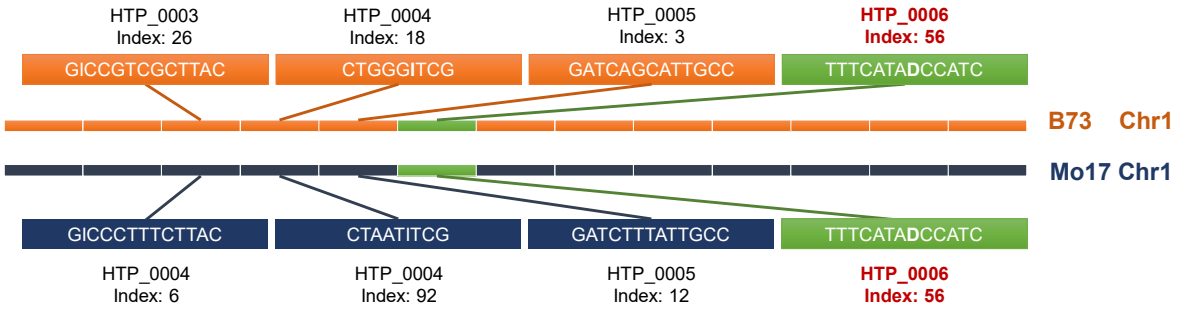
11

Part II
R&D Process

12

2.1 Haplotype-tag polymorphisms(HTP) Marker

☐ **Description:** Based on recombination exchange patterns and block map approaches, a haplotype-tag polymorphism (HTP) block marker has been developed. Each block seamlessly covers the entire genome, with co-segregating haplotypes within blocks and recombination exchanges occurring between blocks.

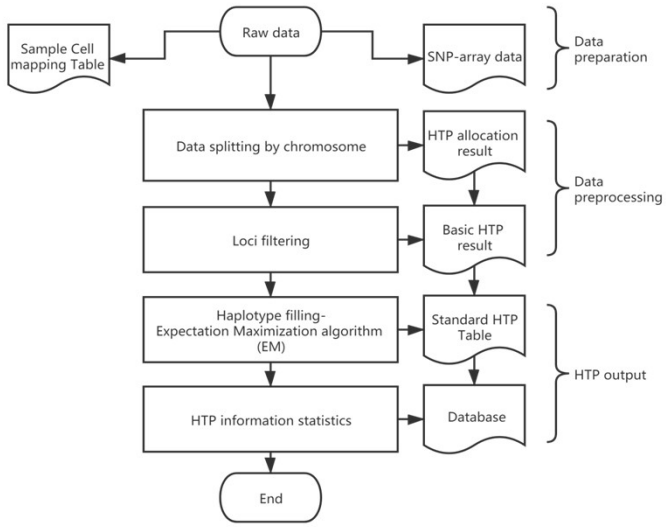


a. Continuous blocks are partitioned based on recombination exchange patterns, each with a unique ID (e.g., HTP_0006);
b. These blocks contain multiple variants, including SNPs and InDels. Each haplotype sequence within a block is assigned a genotype index (Index), representing a specific genotype;
c. If the haplotype sequences in the same block are identical across different varieties, their HTP genotypes are considered the same; otherwise, differences are identified.

13

2.1 HTP Development

☐ **HTP development workflow:**



Haplotype-tag Generation with maize Illustration:

I. SNPs generated by genotyping-by-sequencing using 7,000 recombinant inbred lines (RILs) were converted into **6,164** effective recombination blocks.

II. The Maize6H-60K array was employed to collect genotype data from **3,587** maize inbred lines. Each of these lines was genotyped at **66,905** loci.

III. Cosegregating SNP and InDel markers were combined in a recombination block to be used as an effective haplotype tag. Finally, **6,164** HTPs (6,164 loci) were obtained.

14

2.1 HTP Development

☐ HTP Genotype Generation

How to solve the comparison analysis of more than 10,000 samples if haplotypes are sequences?

Strategy:

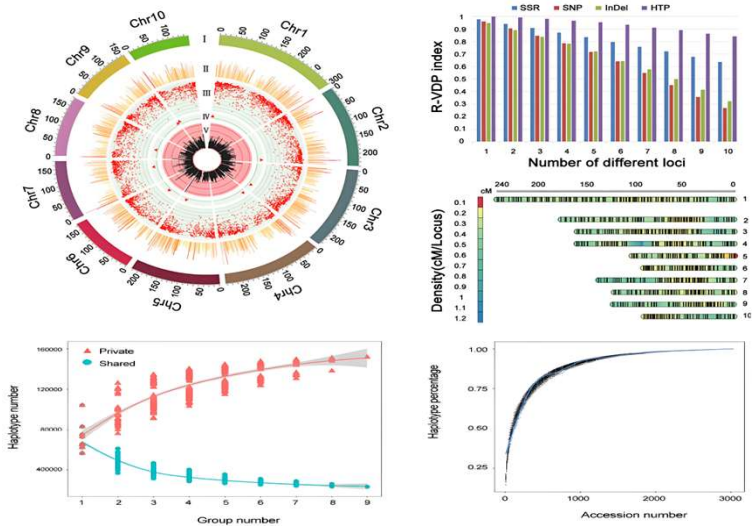
1. Each complete haplotype within HTP is indexed, ensuring that each allele variation in HTP has a **unique Index**.
2. Adopt the SSR marker recording mode, treating each allele variation as a length variation, and genotype information is recorded as integers.
3. Each sample's genotype is stored and analyzed in a manner similar to SSR mode.
4. A haplotype sequence similarity comparison algorithm has been developed, which supports custom similarity parameters.

Index	Chr	HTP ID	Allelic Variations Sequence/Haplotype_tag Sequence	Rate
1	1	HTP_0001	GICCGTCGCTTACTTTCATADCCATCCTGGGGTCCGATCAGCATTGCC	0.267
2	1	HTP_0001	GDTAACTGTCTACTCTTATAICAATTGCCGAACAAAGTCCGCCGTTGTC	0.101
3	1	HTP_0001	GDTAACTATCTACTCTTATAICAATTGCCGAACAAAGTCCGCCGTTATA	0.08
4	1	HTP_0001	ADTCATCATCCGTACCCCGGICCATTGCCAGCTAGCGCAAAATGCTGTC	0.079
5	1	HTP_0001	ADTCACTATCTACTTCCCGADCCATCCTGGGCTCCGATCAGCATTGTC	0.068
6	1	HTP_0001	ADTCATCATTCGTACCCCGAIGCCCCCTAGGGTAGGGTCCGCCGCCCTC	0.065
7	1	HTP_0001	ADTCATCATCCGTACCCGGGIGCCCCCCCGAGGTAGGGTAAGCGTTGTC	0.056
8	1	HTP_0001	ADTCATCATCCGTACCCAGAICCATCCTGGGGTCCGATCAGCACCATA	0.031
9	1	HTP_0001	ADTCATCATCCGTACTCAGADCCATTGCCAGCTAGCGCAAAATGCTGTC	0.027
10	1	HTP_0001	ADTCATCATCCGTACCCCGGGIGCCCCCCCGAGGTAGGGTAAGCGTTGCA	0.027
11	1	HTP_0001	GDTCATCATCTACTTTCATAICAATCCTAGAACAAAGTCCGCCGTTATC	0.025
12	1	HTP_0001	ADTCATCATTCGTACCCCGAIGCATCCTAGCGTAGCGTCCGCCGCCATA	0.01
13	1	HTP_0001	ADTCATCATTCGTACCCCGAIGCATCCTAGGGTAGGGTCCGCCGCCGTC	0.006
14	1	HTP_0001	ADTCATCATCCGTACCCCGGGIGCCCCCCCGAGGTAGGGTAAGCGTTATA	0.006
15	1	HTP_0001	ADTCATCATCCGTACCCCGGGIGCCCCCCCGAGGTAGGGTAAGCGTTATC	0.006
16	1	HTP_0001	GDTCATCATCTACTTTCATAICAATCCTAGAACAAAGTCCGCCGTTGTC	0.005
17	1	HTP_0001	ADTCATCATCTACTTTCATAIGCATCCTAGAACAAAGTAGCTGTTGTC	0.005

Call_code	P016-P54H11_WF631.CEL	P010-P34B06_WG24.CEL
Sample_name	Jing2416	5237
chr1_HTP0001	1	5
chr1_HTP0002	1	5
chr1_HTP0003	1	1
chr1_HTP0004	1	1
chr1_HTP0005	1	7
chr1_HTP0006	0	0
chr1_HTP0007	1	6
chr1_HTP0008	1	2
chr1_HTP0009	1	3
chr1_HTP0010	1	3
chr1_HTP0011	1	2
chr1_HTP0012	1	4
chr1_HTP0013	1	5
chr1_HTP0014	1	2
chr1_HTP0015	2	3
chr1_HTP0016	1	4
chr1_HTP0017	1	3

15

2.2 HTP Assessment



Evaluation of the HTPs. (Zhao *et al.*, 2022)

Among the 6,164 HTP markers:

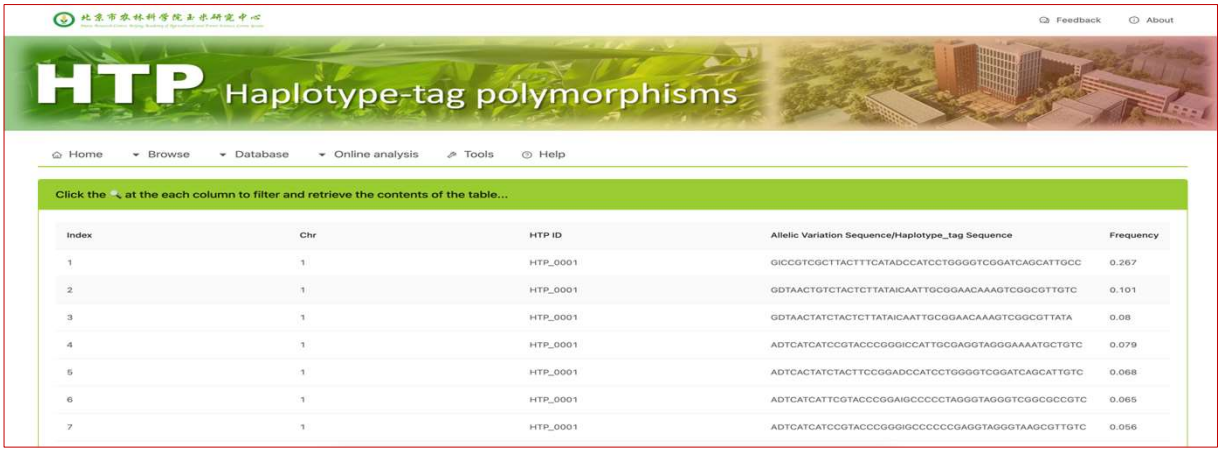
1. 2,811 had more than 10 variations (SNPs and InDels combined);
2. 5,649 of the markers had a PIC value exceeding **0.5**, and 871 had a PIC value greater than **0.9**;
3. The single HTP marker in the chloroplast genome had a PIC value of **0.5128** and 23 haplotypes.
4. The sample size reached 820, a plateau was approached and at least **90%** of the haplotype sets were included.

16

2.3 HTPdb

- **Online database:** An online database of haplotype tag polymorphisms based on maize inbred line data was constructed and named HTPdb. This database comprises 172,921 non-redundant HTP allelic variations from 3,587 accessions.

- **Available at:** https://htp.plantdna.site/database/nucleus-haplotype ; *DOI: 10.1016/j.xplc.2022.100331*



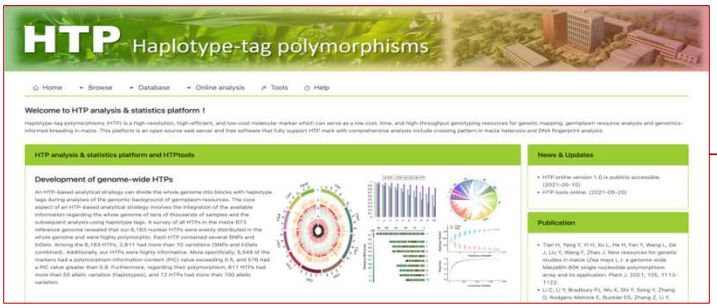
17

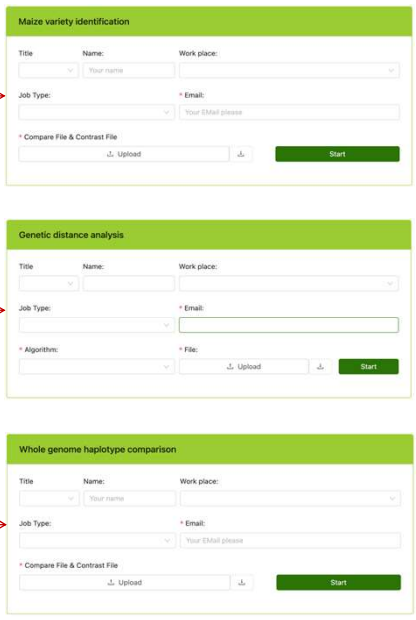
2.4 HTPTools

- **Platform：** An online platform dedicated to HTP marker generation and analysis based on HTPdb has also been built. The basic functions of this platform include **maize variety identification**, **genetic distance analysis**, and **whole genome haplotype comparison**.

1. The platform is constructed using the Next.js framework;
2. The frontend is composed of Ant Design components supported by Rect.js;
3. ESLint is used to eliminate obvious errors in the code; Babel maintains compatibility between ordinary JS and ES6.

- **Available at:** https://htp.plantdna.site





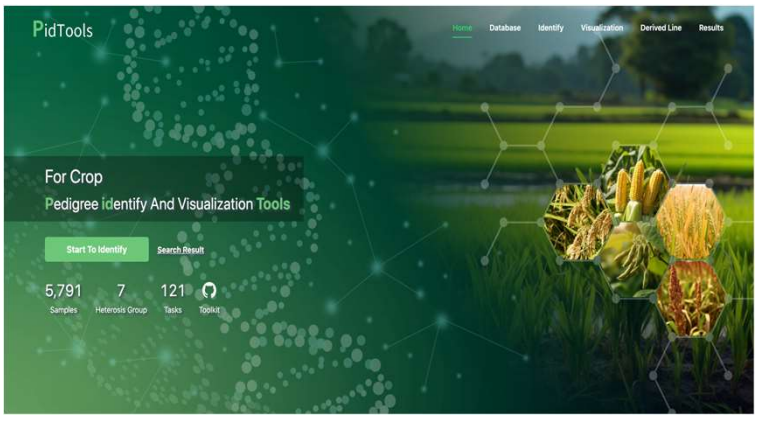
18

Part III
Application Case

19

3.1 Crop Pedigree Identification and Reconstruction

- **PidTools:** Based on HTP markers, a dedicated tool named PidTools has been developed. The algorithms and associated tools are suitable for all crops for the **reconstruction** and **visualization** of a complete pedigree.
- **Database:** A genotype database was constructed using Maize6H-60K array data from **5791** maize inbred lines.
- **Available at:** https://pidtools.plantdna.site/home ; *DOI: 10.1016/j.csbj.2024.07.004*

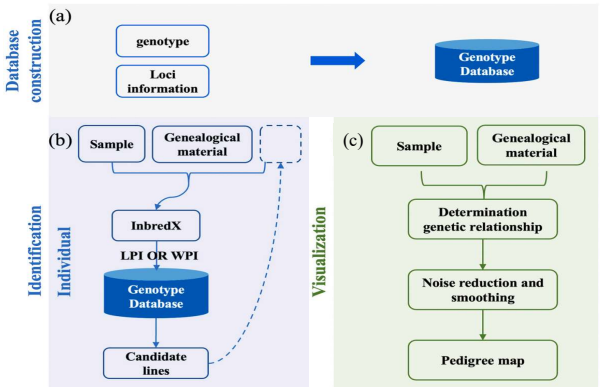


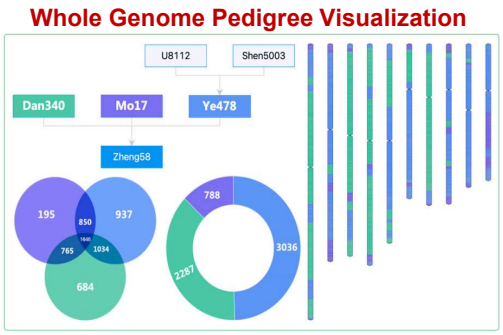
20

3.2 Workflow

☐ **PidTools workflow:**

a. Construction of the genotype database;
b. Pedigree identification workflow, which involves the Whole-genome Pedigree Identification (WPI) and Long-contiguous-fragment Pedigree Identification (LPI) algorithms;
c. Pedigree visualization workflow, including data noise reduction and smoothing.



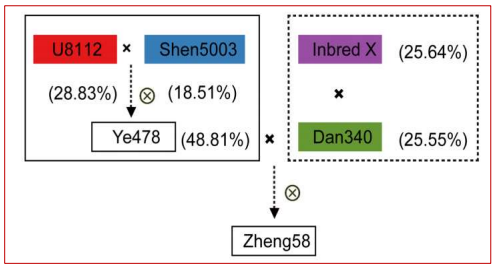


21

3.3 Application Case

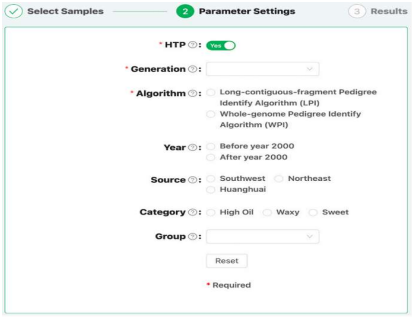
☐ **Maize pedigree identification:** Based on the developed PidTools platform, pedigree reconstruction were conducted for Zheng58 (maternal parent of Zhengdan958, which was China's most extensively cultivated commercial maize hybrid).

- **Inbred X identification:** Researchers have discovered the presence of unknown parental components in the genome of Zheng58, but the specific parent remains unidentified.



Inbred X remains unidentified (Zhang et al., 2018)

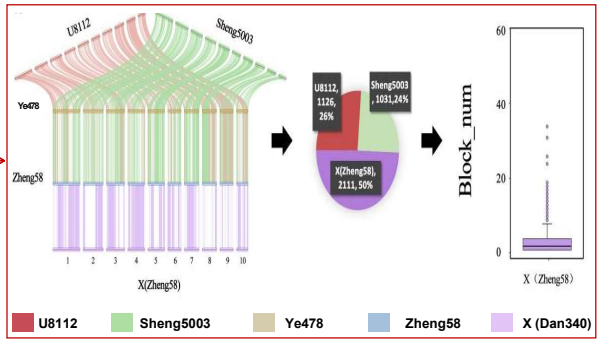
- **Step 1-Parameter Settings:** First, the genotype data need to be uploaded into PidTools, the WPI algorithm selected, and the required information about Zheng58 populated.



22

3.3 Application Case

- **Step 2-Genomic Anchoring:**
  Using known pedigree information, the algorithm automatically anchored Ye478's genomic segments to the Zheng58 genome.
- **Step 3-Primary Identification:**
  For non-Ye478 regions, the pedigree tracing algorithm deduced parental contributions, identifying Dan340 as the second parent and 667 uncharacterized HTPs (10.8% of the genome).



- **Step 4-Final Identification:** The uncharacterized HTPs were extracted and screened against the genotype database (repeat the Primary Identification workflow), revealing the unknown parent as Mo17.

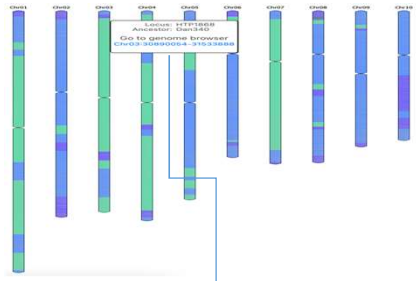
23

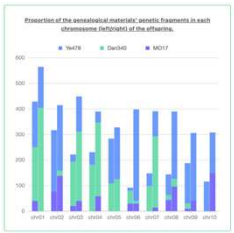
3.4 Genomic Pedigree

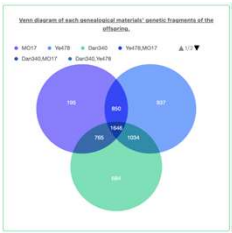
☐ **Pedigree Visualization:** Chromosome-level visualization of the pedigree of Zheng58 based on genotype data for Ye478, Dan340, and Mo17 was performed using the pedigree visualization algorithm.

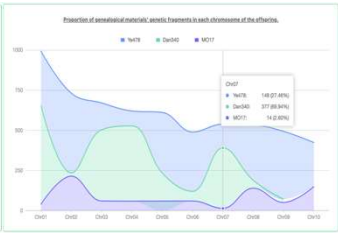
a. Distribution of the genealogical materials' genetic fragments in each chromosome of the offspring;
b. Venn diagram of each genealogical materials' genetic fragments of the offspring;
c. Proportion of the genealogical materials' genetic fragments in each chromosome (left/right) of the offspring.

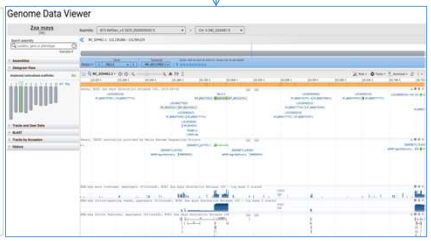
✓ All raw data can be downloaded for free.
✓ Interactive charts reveal details via clicks, including HTP ID and position.
✓ All positions are linked to the Genome Data Viewer.











24

Summary

- **HTPdb:** An online database of haplotype tag polymorphisms based on maize inbred line data was constructed and named HTPdb. This database comprises **172,921** non-redundant HTP allelic variations from **3,587** maize accessions.

  Available at: https://htp.plantdna.site/database/nucleus-haplotype ; *DOI: 10.1016/j.xplc.2022.100331*

- **HTPTools Platform:** The basic functions of this platform include **maize variety identification**, **genetic distance analysis**, and **whole genome haplotype comparison**.

  Available at: https://htp.plantdna.site

- **PidTools:** The algorithms and associated tools are suitable for all crops for the **reconstruction** and **visualization** of a complete pedigree. A genotype database was constructed using Maize6H-60K array data from **5791** maize inbred lines.

  Available at: https://pidtools.plantdna.site/home ; *DOI: 10.1016/j.csbj.2024.07.004*

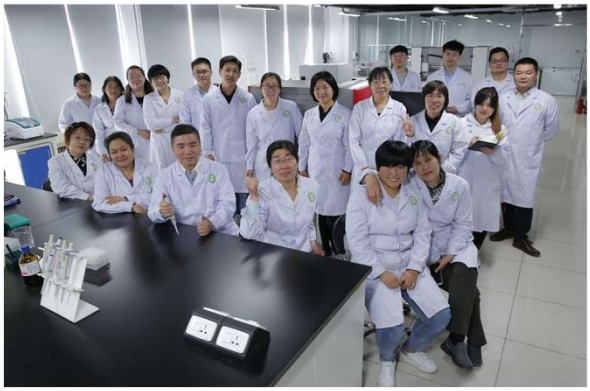
25

Thank you for your attention!

26

Beijing Maize Seed Testing Center

- ✓ **E-mail**: fenggewangmaize@126.com
- ✓ **Phone:** +86 010 51508608
- ✓ **Address:** Maize Research Center, Beijing Academy of Agriculture & Forestry Sciences (BAAFS), Beijing Key Laboratory of Maize DNA Fingerprinting and Molecular Breeding, Shuguang Garden Middle Road No. 9, Beijing 100097, China



27

[End of Annex and of document]