UPOV

International Union for the Protection of New Varieties of Plants

E

Technical Working Party on Testing Methods and Techniques

Second Session
Virtual meeting, April 8 to 11, 2024

TWM/2/14

Original: English
Date: March 19, 2024

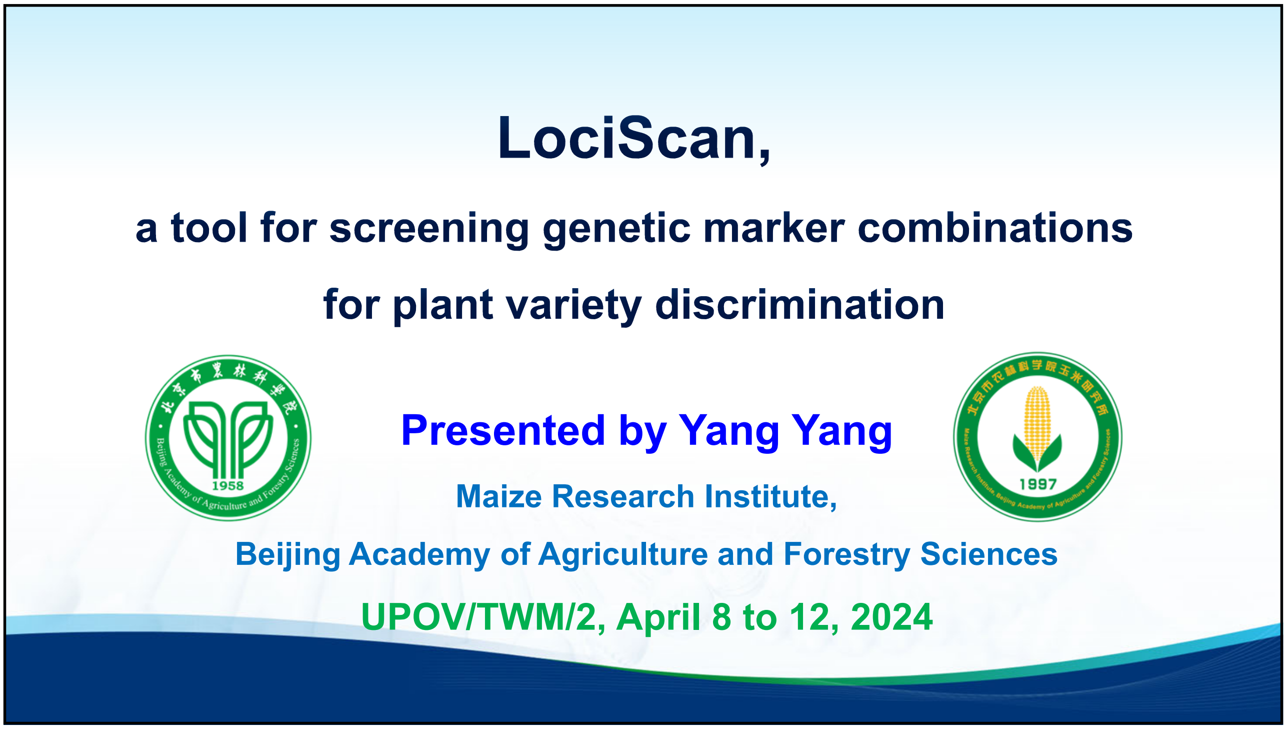
LOCISCAN, A TOOL FOR SCREENING GENETIC MARKER COMBINATIONS FOR PLANT VARIETY DISCRIMINATION

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 “LociScan, a tool for screening genetic marker combinations for plant variety discrimination”, to be made by an expert from China, at the second session of the Technical Working Party on Testing Methods and Techniques (TWM).

[Annex follows]



LociScan,

a tool for screening genetic marker combinations for plant variety discrimination

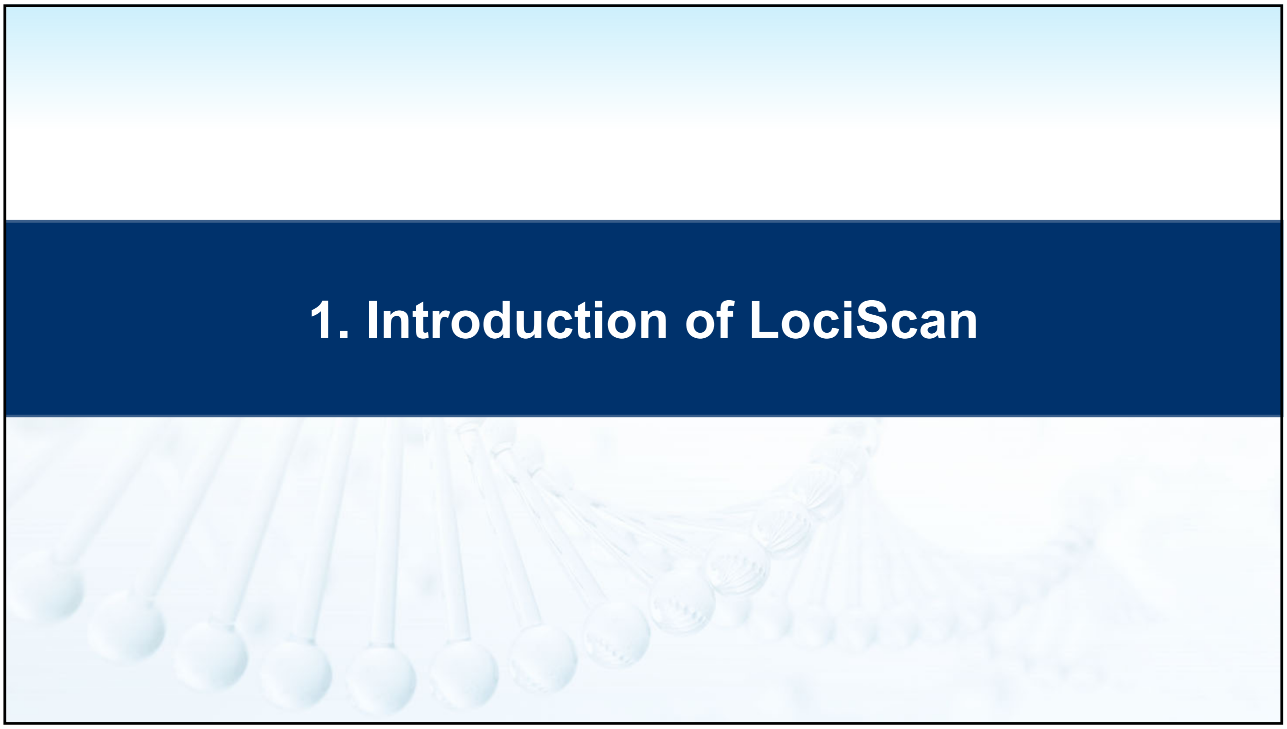
Presented by Yang Yang

Maize Research Institute,

Beijing Academy of Agriculture and Forestry Sciences

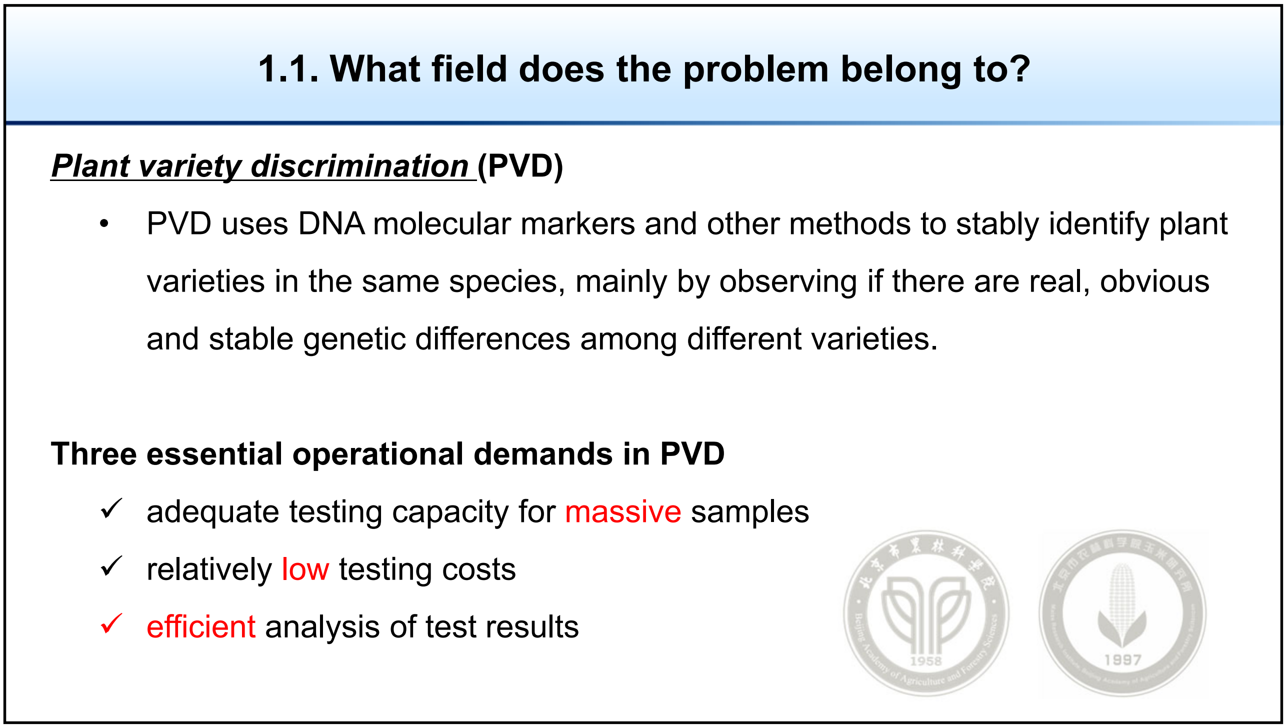
UPOV/TWM/2, April 8 to 12, 2024

1



1. Introduction of LociScan

2



1.1. What field does the problem belong to?

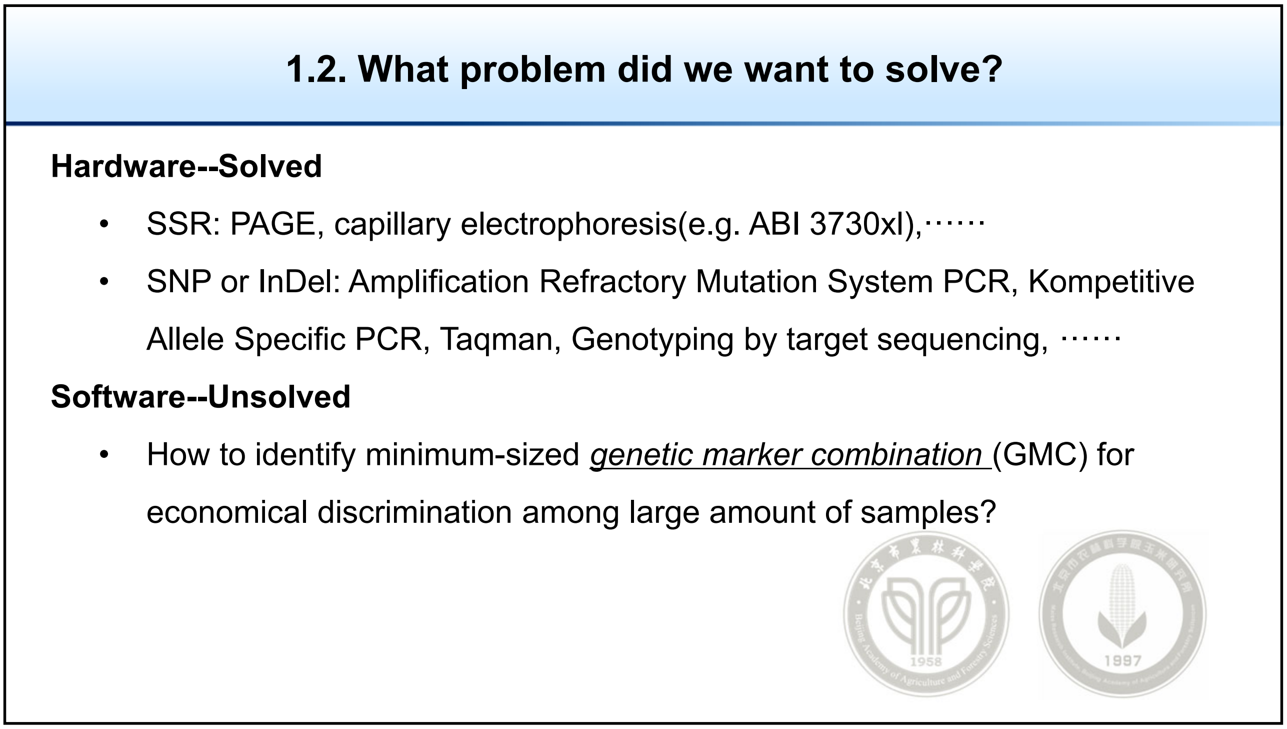
<u>Plant variety discrimination</u> **(PVD)**

- PVD uses DNA molecular markers and other methods to stably identify plant varieties in the same species, mainly by observing if there are real, obvious and stable genetic differences among different varieties.

Three essential operational demands in PVD

- ✓ adequate testing capacity for massive samples
- ✓ relatively low testing costs
- ✓ efficient analysis of test results

3



1.2. What problem did we want to solve?

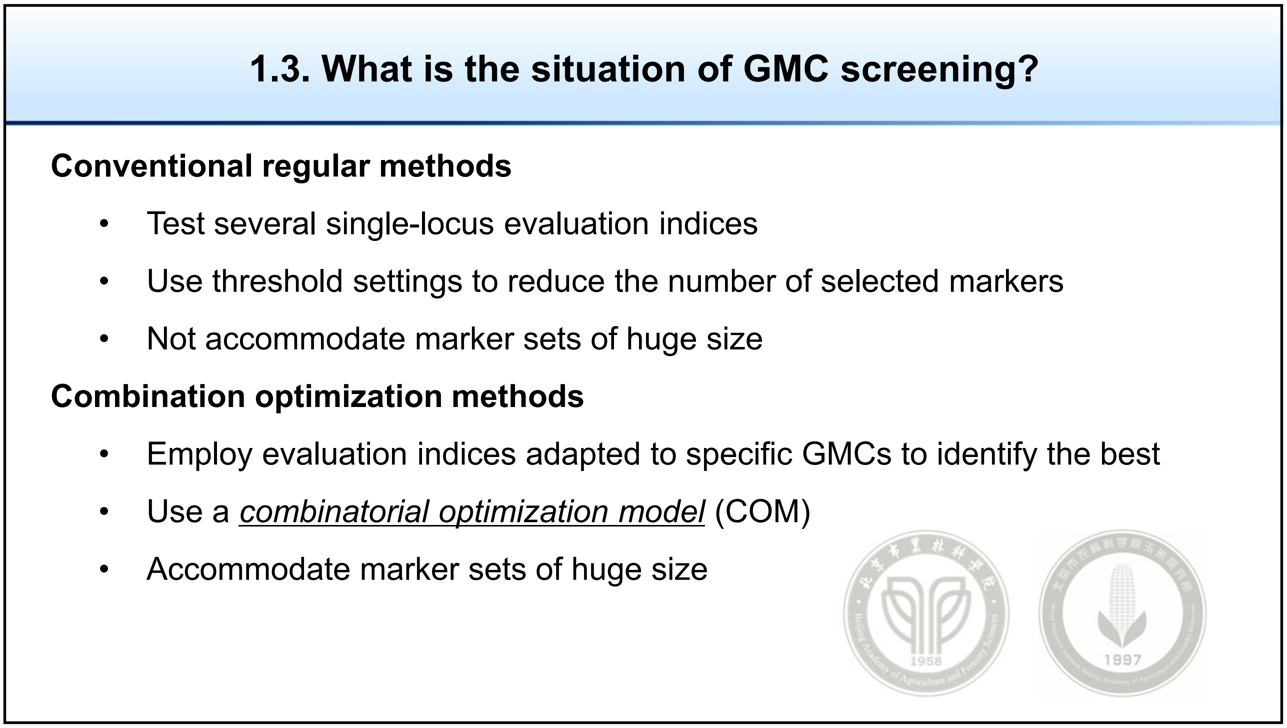
Hardware--Solved

- SSR: PAGE, capillary electrophoresis(e.g. ABI 3730xl),⋯⋯
- SNP or InDel: Amplification Refractory Mutation System PCR, Kompetitive Allele Specific PCR, Taqman, Genotyping by target sequencing, ⋯⋯

Software--Unsolved

- How to identify minimum-sized *<u>genetic marker combination</u>* (GMC) for economical discrimination among large amount of samples?

4



1.3. What is the situation of GMC screening?

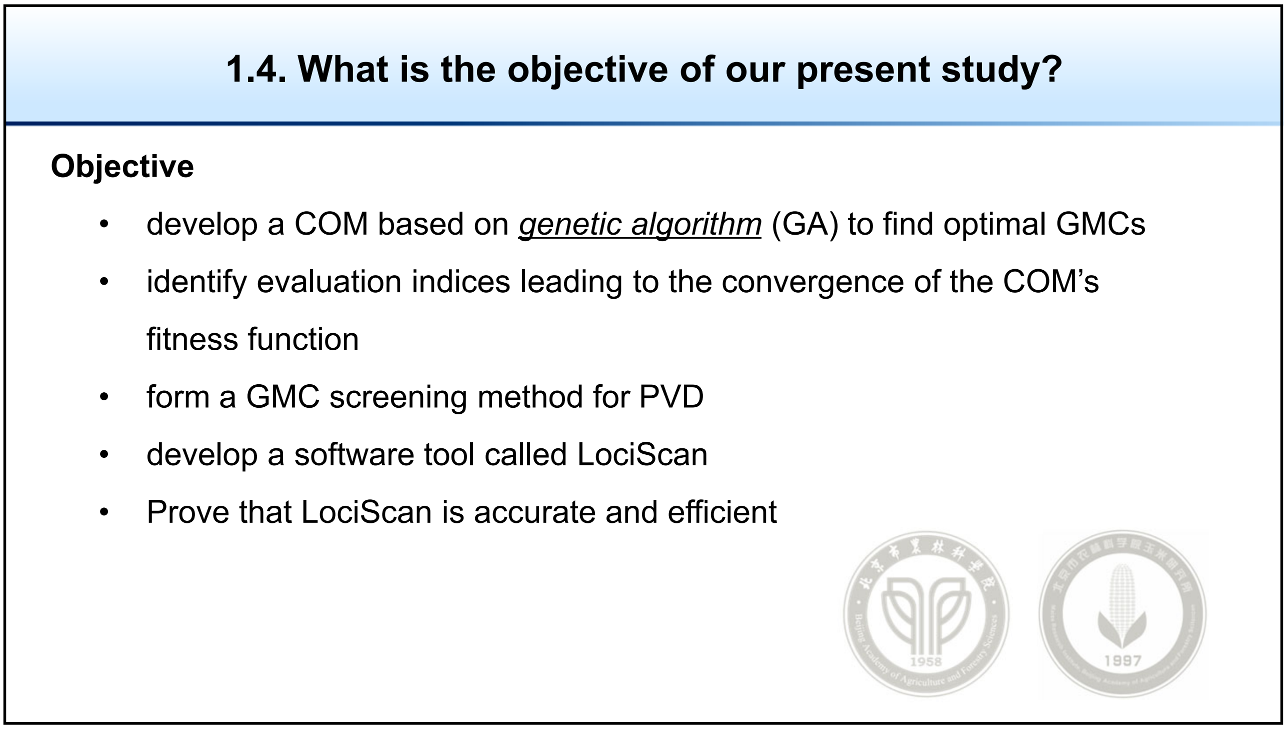
Conventional regular methods

- Test several single-locus evaluation indices
- Use threshold settings to reduce the number of selected markers
- Not accommodate marker sets of huge size

Combination optimization methods

- Employ evaluation indices adapted to specific GMCs to identify the best
- Use a *combinatorial optimization model* (COM)
- Accommodate marker sets of huge size

5



1.4. What is the objective of our present study?

Objective

- develop a COM based on *genetic algorithm* (GA) to find optimal GMCs
- identify evaluation indices leading to the convergence of the COM's fitness function
- form a GMC screening method for PVD
- develop a software tool called LociScan
- Prove that LociScan is accurate and efficient

6

1.5. What is the approach of our present study?

Approach

- **Optimize the parameters**: use genotype data for various plant species and marker types to evaluate the influence of evaluation indices and GA model parameters on the results.
- **Improve our method**: use evaluations of the performance of several algorithms and software tools in finding GMCs.
- **Expand our tools**: evaluate the performance of our method with simulated genotype datasets of diverse sizes.

7

2. Materials and Methods

8

2.1. Data sources and processing

Row genotype data (the publicly available real data)

- sorghum SSR data (Casa et al., 2005)
- wheat SSR data (Hao et al., 2011)
- maize SNP data (Tian et al., 2015)
- rice SNP data (Zhao et al., 2010)

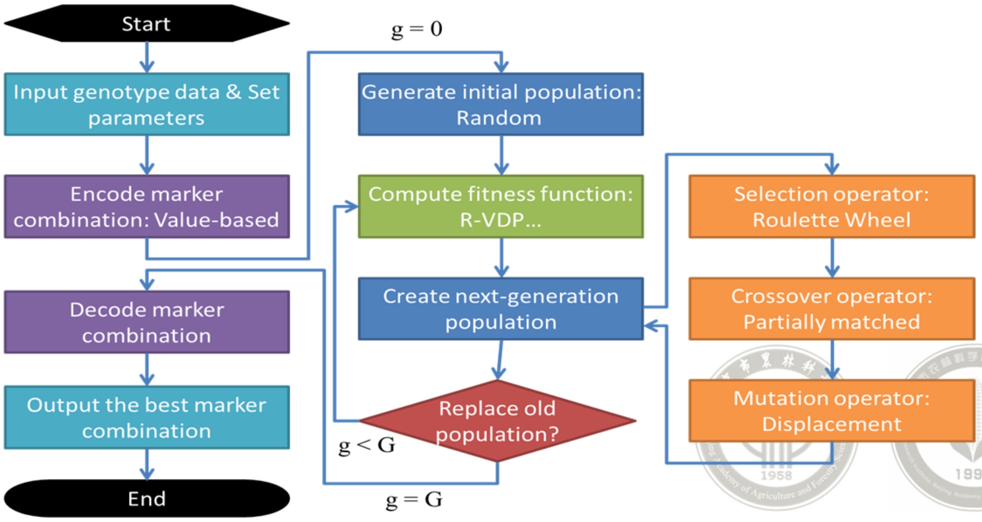
Dataset 1: preprocess the raw genotype data as with deleting samples unable to be discriminated by the full marker set, and deleting markers with 0 PIC.

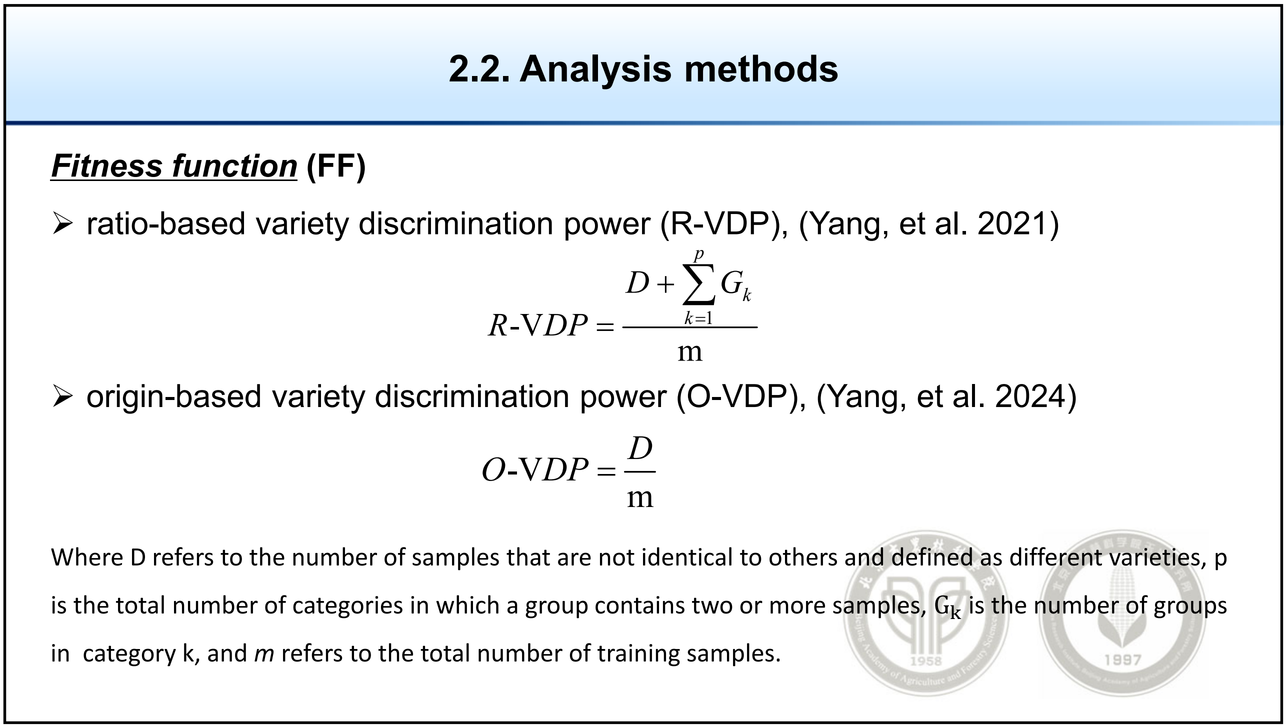
Dataset 2: Populate dataset 1 with no data missing.

Dataset 3: simulated SNP genotype data with 8 sizes.

2.2. Analysis methods

Process of GMC screening model based on GA


Start
g = 0
Input genotype data & Set parameters
Generate initial population: Random
Encode marker combination: Value-based
Compute fitness function: R-VDP…
Selection operator: Roulette Wheel
Create next-generation population
Crossover operator: Partially matched
Decode marker combination
Replace old population?
Mutation operator: Displacement
g < G
Output the best marker combination
End
g = G



2.2. Analysis methods

Fitness function **(FF)**

- ratio-based variety discrimination power (R-VDP), (Yang, et al. 2021)

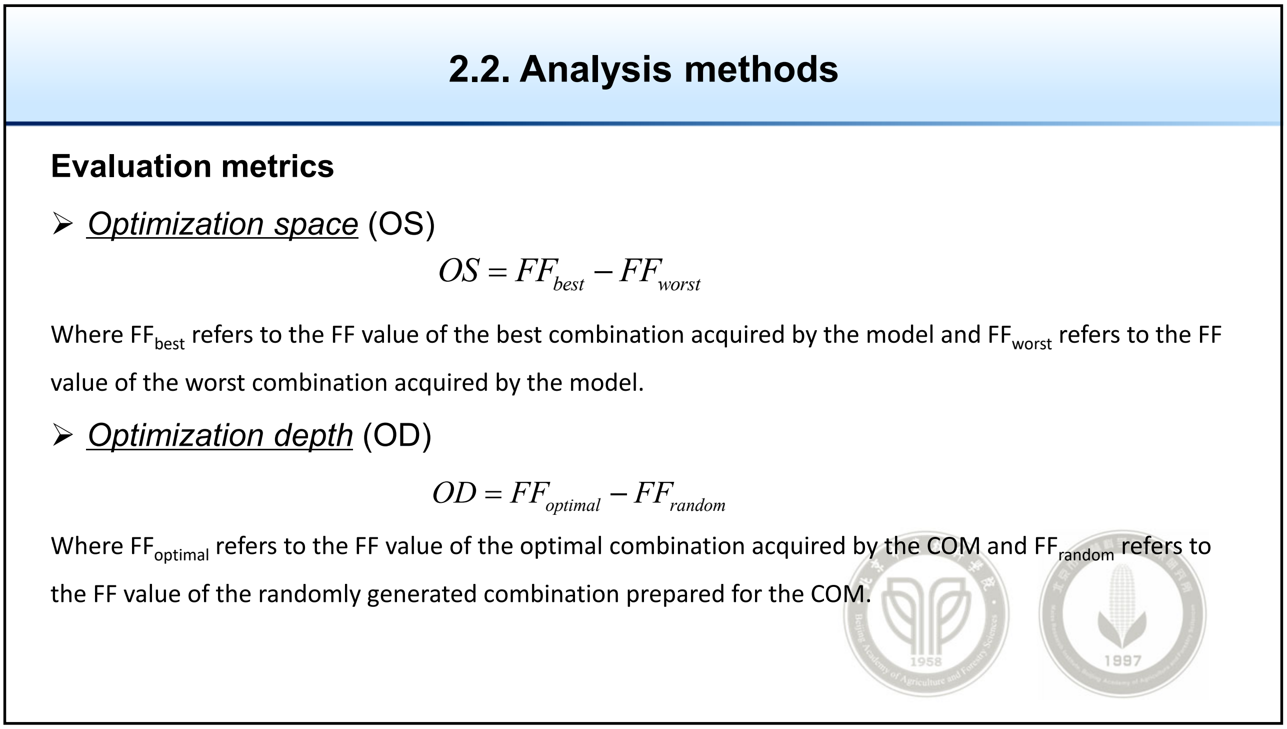
$$R\text{-}\mathrm{V}DP = \frac{D + \sum_{k=1}^{p} G_k}{\mathrm{m}}$$

- origin-based variety discrimination power (O-VDP), (Yang, et al. 2024)

$$O\text{-}\mathrm{V}DP = \frac{D}{\mathrm{m}}$$

Where D refers to the number of samples that are not identical to others and defined as different varieties, p is the total number of categories in which a group contains two or more samples, G_k is the number of groups in category k, and *m* refers to the total number of training samples.

11



2.2. Analysis methods

Evaluation metrics

- *Optimization space* (OS)

$$OS = FF_{best} - FF_{worst}$$

Where FF_{best} refers to the FF value of the best combination acquired by the model and FF_{worst} refers to the FF value of the worst combination acquired by the model.

- *Optimization depth* (OD)

$$OD = FF_{optimal} - FF_{random}$$

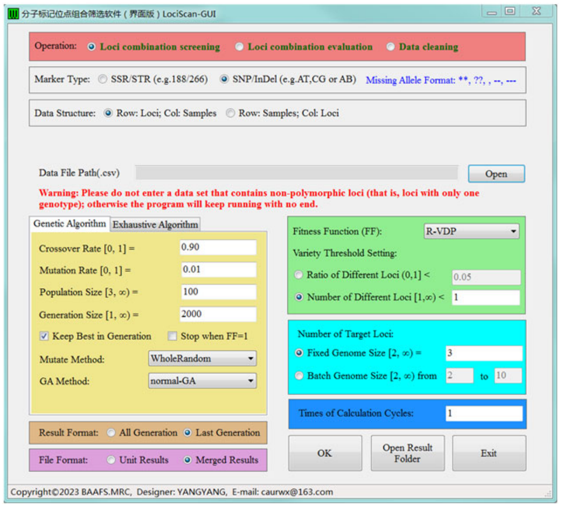
Where $FF_{optimal}$ refers to the FF value of the optimal combination acquired by the COM and FF_{random} refers to the FF value of the randomly generated combination prepared for the COM.

12

3. Result

3.1. Develop a software tool named LociScan contains 2 versions

- ➢ LociScan-GUI version main interface



分子标记位点组合筛选软件（界面版）LociScan-GUI

Operation: ● Loci combination screening ○ Loci combination evaluation ○ Data cleaning

Marker Type: ○ SSR/STR (e.g.188/266) ● SNP/InDel (e.g.AT,CG or AB) Missing Allele Format: **, ??, , --, ---

Data Structure: ● Row: Loci; Col: Samples ○ Row: Samples; Col: Loci

Data File Path(.csv)

Warning: Please do not enter a data set that contains non-polymorphic loci (that is, loci with only one genotype); otherwise the program will keep running with no end.

Genetic Algorithm | Exhaustive Algorithm

Crossover Rate [0, 1] = 0.90
Mutation Rate [0, 1] = 0.01
Population Size [3, ∞) = 100
Generation Size [1, ∞) = 2000
☑ Keep Best in Generation □ Stop when FF=1
Mutate Method: WholeRandom
GA Method: normal-GA

Fitness Function (FF): R-VDP
Variety Threshold Setting:
○ Ratio of Different Loci (0,1] < 0.05
● Number of Different Loci [1,∞) < 1

Number of Target Loci:
● Fixed Genome Size [2, ∞) = 3
○ Batch Genome Size [2, ∞) from 2 to 10

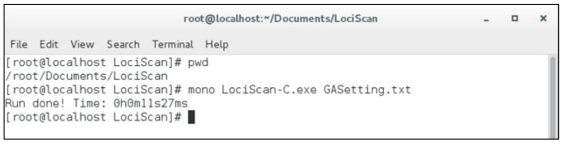
Times of Calculation Cycles: 1

Result Format: ○ All Generation ● Last Generation

File Format: ○ Unit Results ● Merged Results

Copyright©2023 BAAFS.MRC, Designer: YANGYANG, E-mail: caurwx@163.com

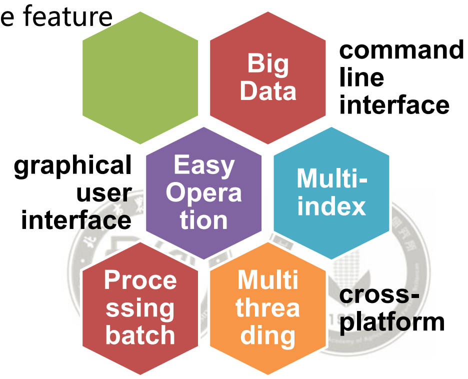
- ➢ LociScan-CLI version running interface



root@localhost:~/Documents/LociScan
File Edit View Search Terminal Help

```
[root@localhost LociScan]# pwd
/root/Documents/LociScan
[root@localhost LociScan]# mono LociScan-C.exe GASetting.txt
Run done! Time: 0h0m11s27ms
[root@localhost LociScan]# █
```

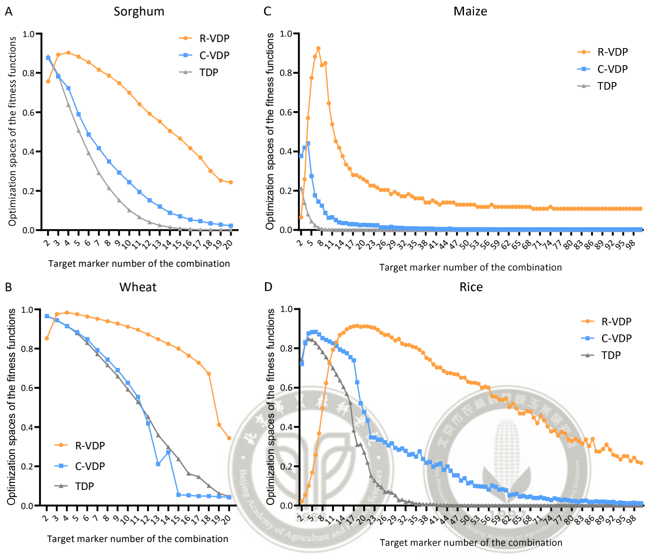
- ➢ Software feature



3.2. Optimization space evaluation of fitness functions

Result 1

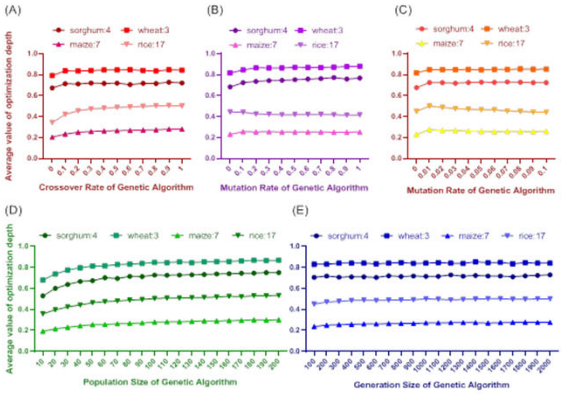
- Combinations with the same marker number but different marker composition result in differences in the capacity of PVD, smaller marker number leads to greater difference.
- R-VDP had the largest OS and scope of target marker numbers where OS existed, followed by C-VDP and TDP.
- Recommend R-VDP as the default FF in the following studies.

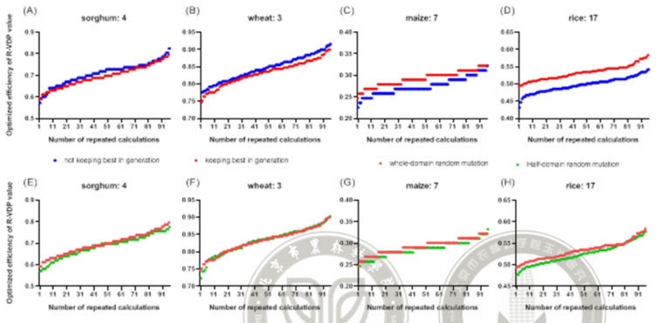


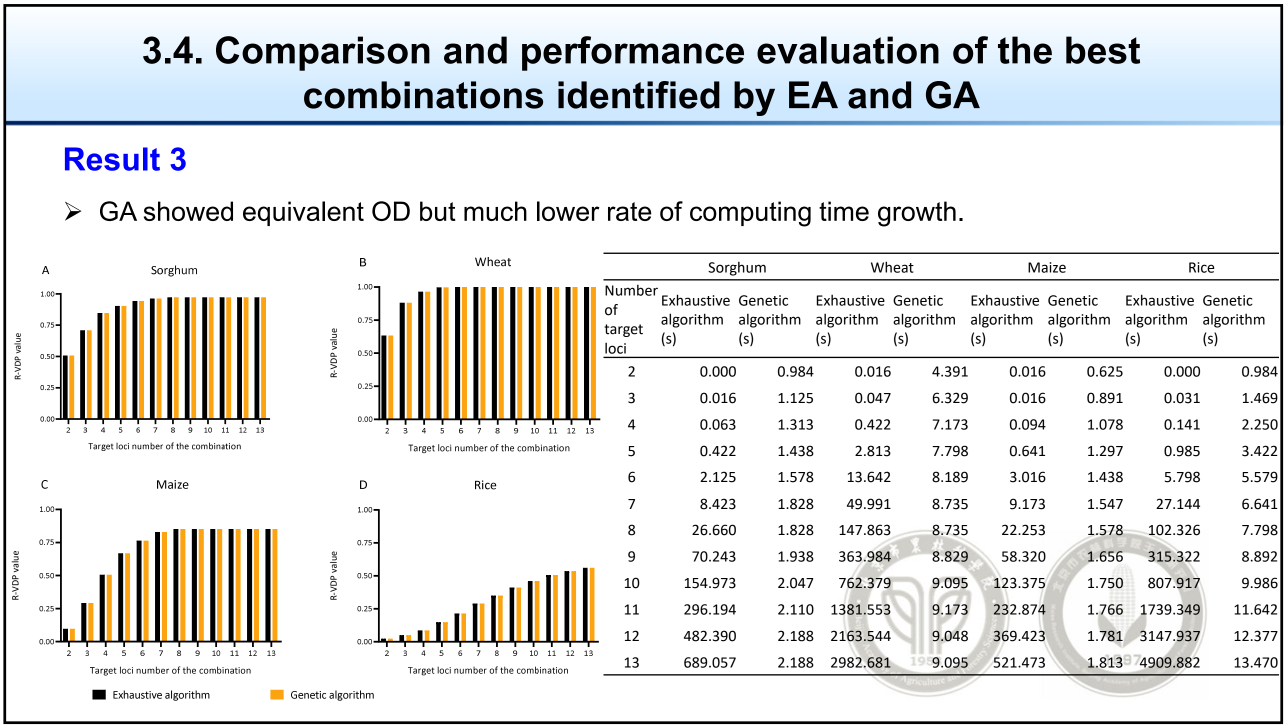
3.3. Influences of model parameters on genetic algorithm

Result 2

- The OD became higher when the population size or generation size was bigger.
- The GA's parameters have universal but varying influence on the OD of COM.





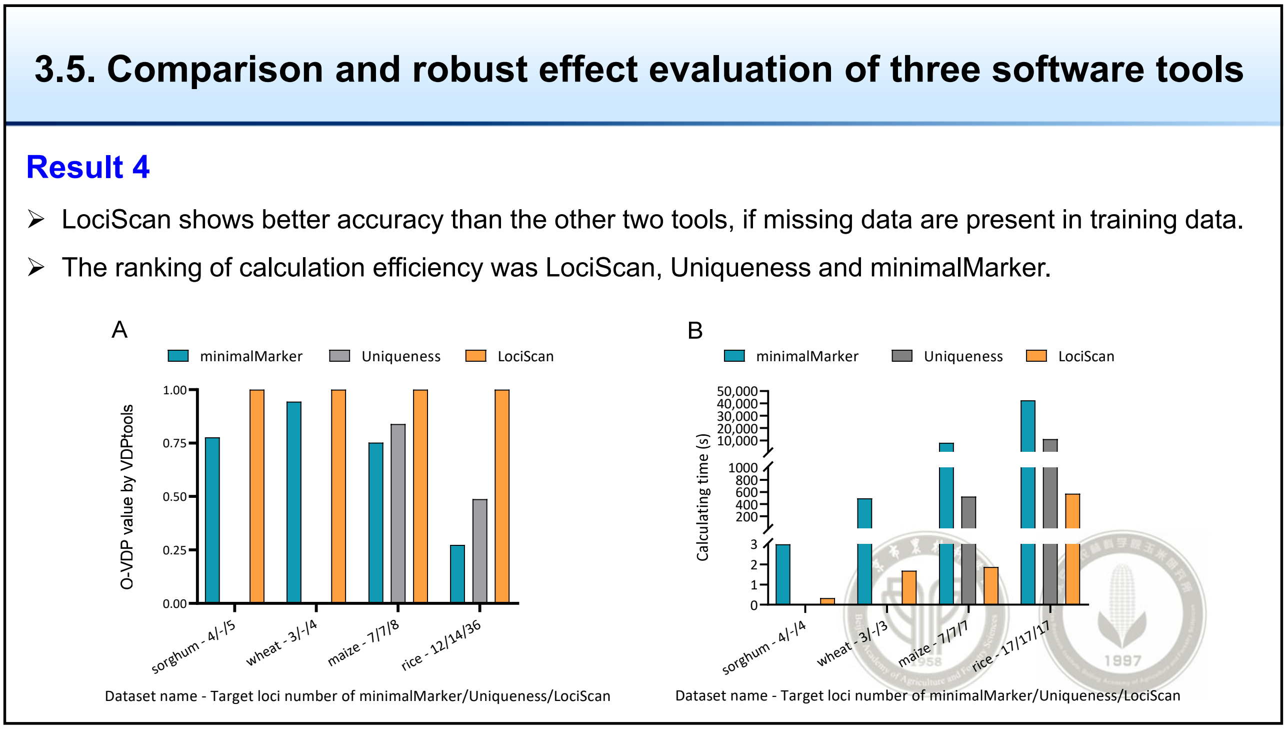


3.4. Comparison and performance evaluation of the best combinations identified by EA and GA

Result 3

- GA showed equivalent OD but much lower rate of computing time growth.

	Sorghum		Wheat		Maize		Rice	
Number of target loci	Exhaustive algorithm (s)	Genetic algorithm (s)	Exhaustive algorithm (s)	Genetic algorithm (s)	Exhaustive algorithm (s)	Genetic algorithm (s)	Exhaustive algorithm (s)	Genetic algorithm (s)
2	0.000	0.984	0.016	4.391	0.016	0.625	0.000	0.984
3	0.016	1.125	0.047	6.329	0.016	0.891	0.031	1.469
4	0.063	1.313	0.422	7.173	0.094	1.078	0.141	2.250
5	0.422	1.438	2.813	7.798	0.641	1.297	0.985	3.422
6	2.125	1.578	13.642	8.189	3.016	1.438	5.798	5.579
7	8.423	1.828	49.991	8.735	9.173	1.547	27.144	6.641
8	26.660	1.828	147.863	8.735	22.253	1.578	102.326	7.798
9	70.243	1.938	363.984	8.829	58.320	1.656	315.322	8.892
10	154.973	2.047	762.379	9.095	123.375	1.750	807.917	9.986
11	296.194	2.110	1381.553	9.173	232.874	1.766	1739.349	11.642
12	482.390	2.188	2163.544	9.048	369.423	1.781	3147.937	12.377
13	689.057	2.188	2982.681	9.095	521.473	1.813	4909.882	13.470



3.5. Comparison and robust effect evaluation of three software tools

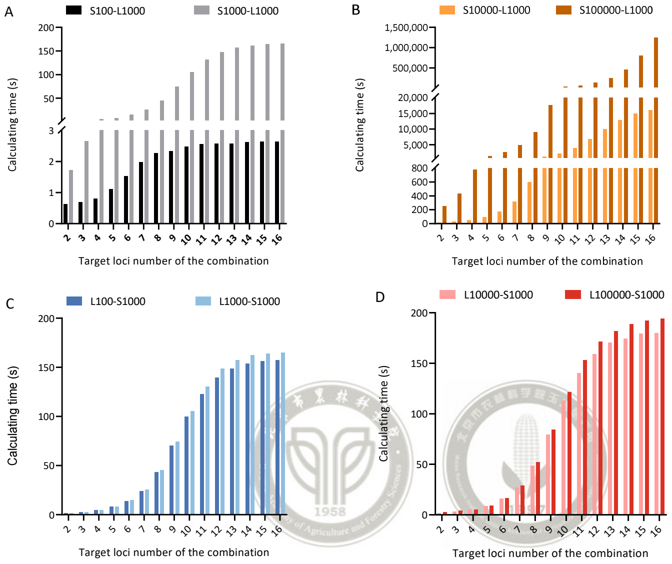
Result 4

- LociScan shows better accuracy than the other two tools, if missing data are present in training data.
- The ranking of calculation efficiency was LociScan, Uniqueness and minimalMarker.

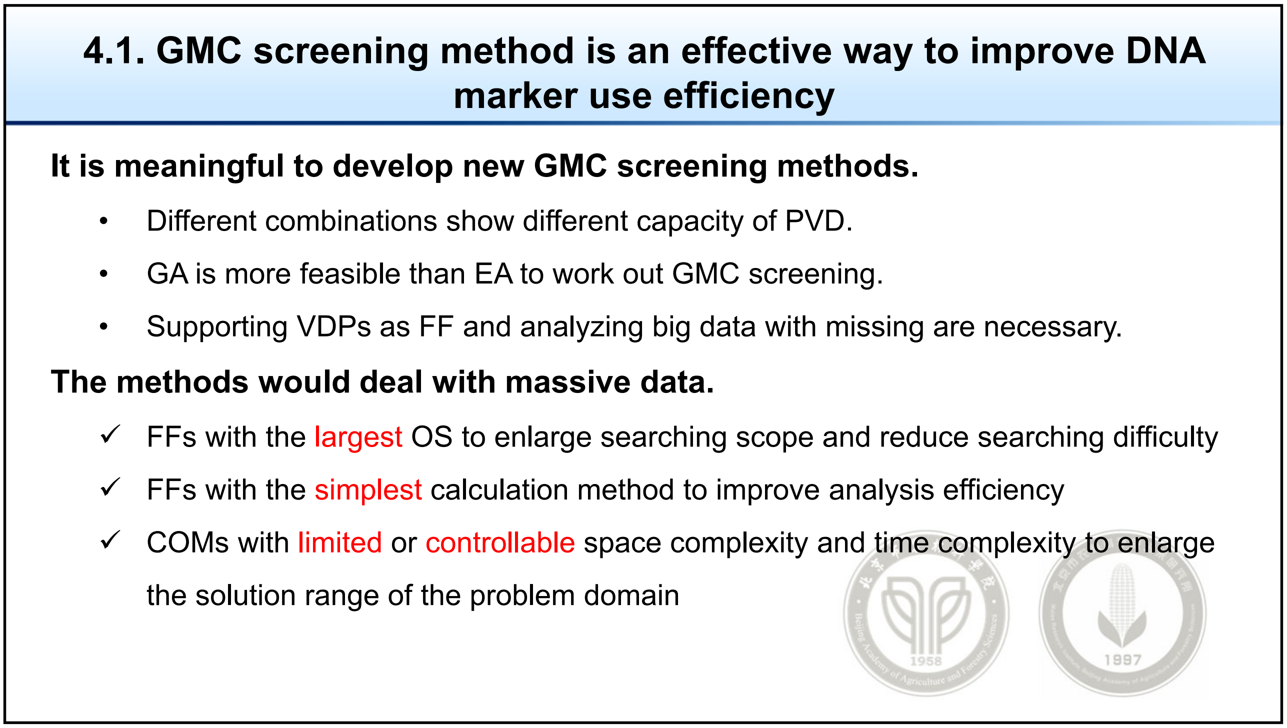
3.6. Performance evaluation of LociScan by data quantity scale

Result 5

- The sample number was the main influence factor for the performance of LociScan, while the marker number had little influence.
- The effect of target marker number on the analysis performance was limited, such that the calculation time of LociScan reached a plateau when the target marker number increased to a certain threshold.



4. Discussion



4.1. GMC screening method is an effective way to improve DNA marker use efficiency

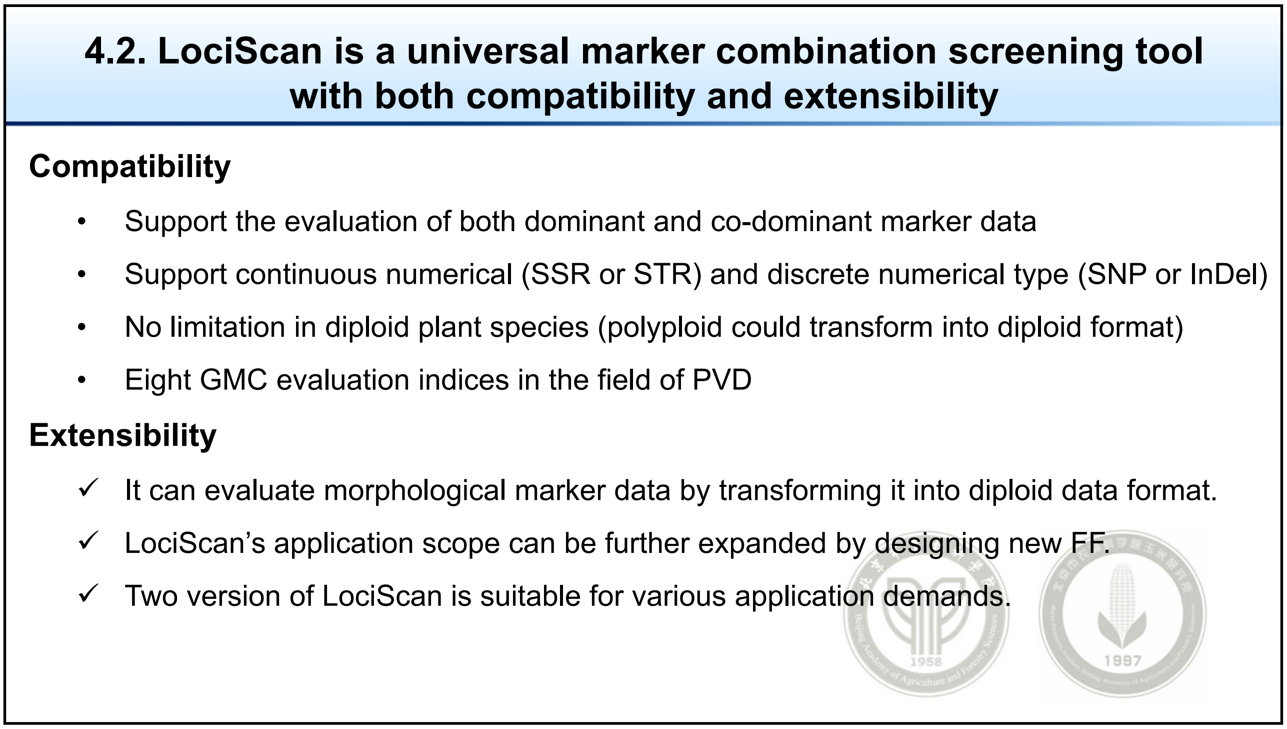
It is meaningful to develop new GMC screening methods.

- Different combinations show different capacity of PVD.
- GA is more feasible than EA to work out GMC screening.
- Supporting VDPs as FF and analyzing big data with missing are necessary.

The methods would deal with massive data.

- ✓ FFs with the largest OS to enlarge searching scope and reduce searching difficulty
- ✓ FFs with the simplest calculation method to improve analysis efficiency
- ✓ COMs with limited or controllable space complexity and time complexity to enlarge the solution range of the problem domain

21



4.2. LociScan is a universal marker combination screening tool with both compatibility and extensibility

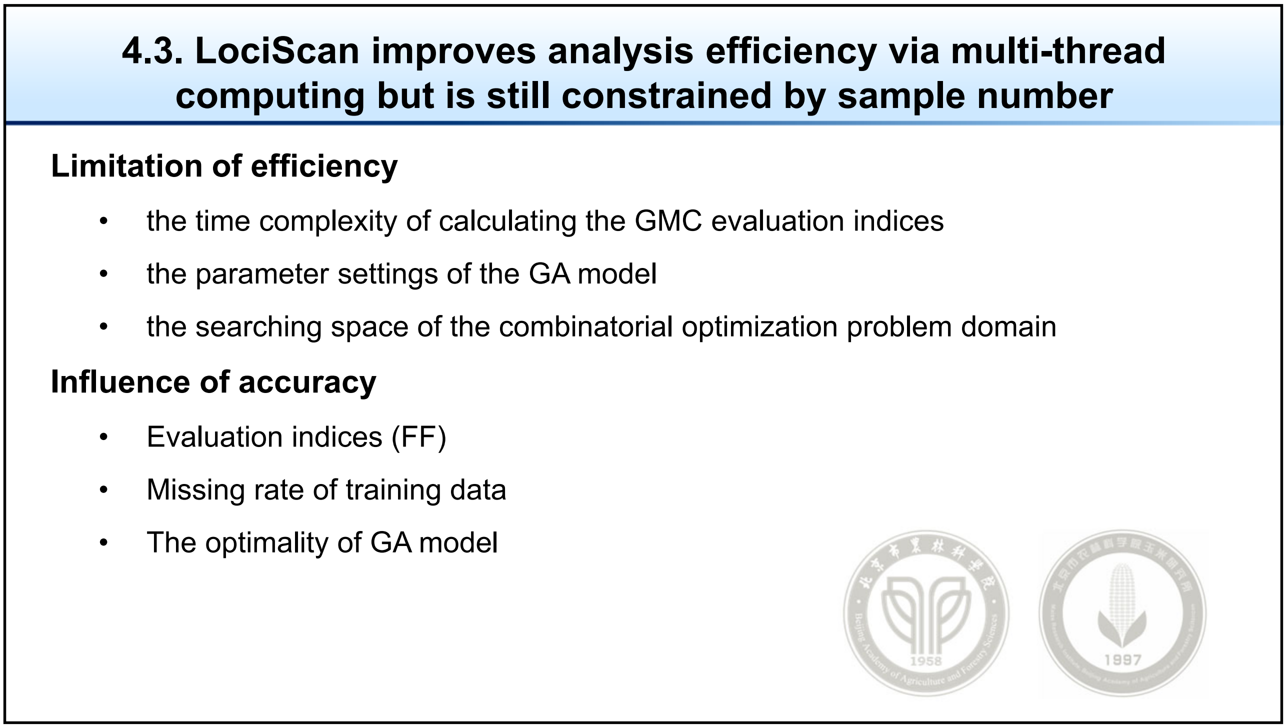
Compatibility

- Support the evaluation of both dominant and co-dominant marker data
- Support continuous numerical (SSR or STR) and discrete numerical type (SNP or InDel)
- No limitation in diploid plant species (polyploid could transform into diploid format)
- Eight GMC evaluation indices in the field of PVD

Extensibility

- ✓ It can evaluate morphological marker data by transforming it into diploid data format.
- ✓ LociScan’s application scope can be further expanded by designing new FF.
- ✓ Two version of LociScan is suitable for various application demands.

22



4.3. LociScan improves analysis efficiency via multi-thread computing but is still constrained by sample number

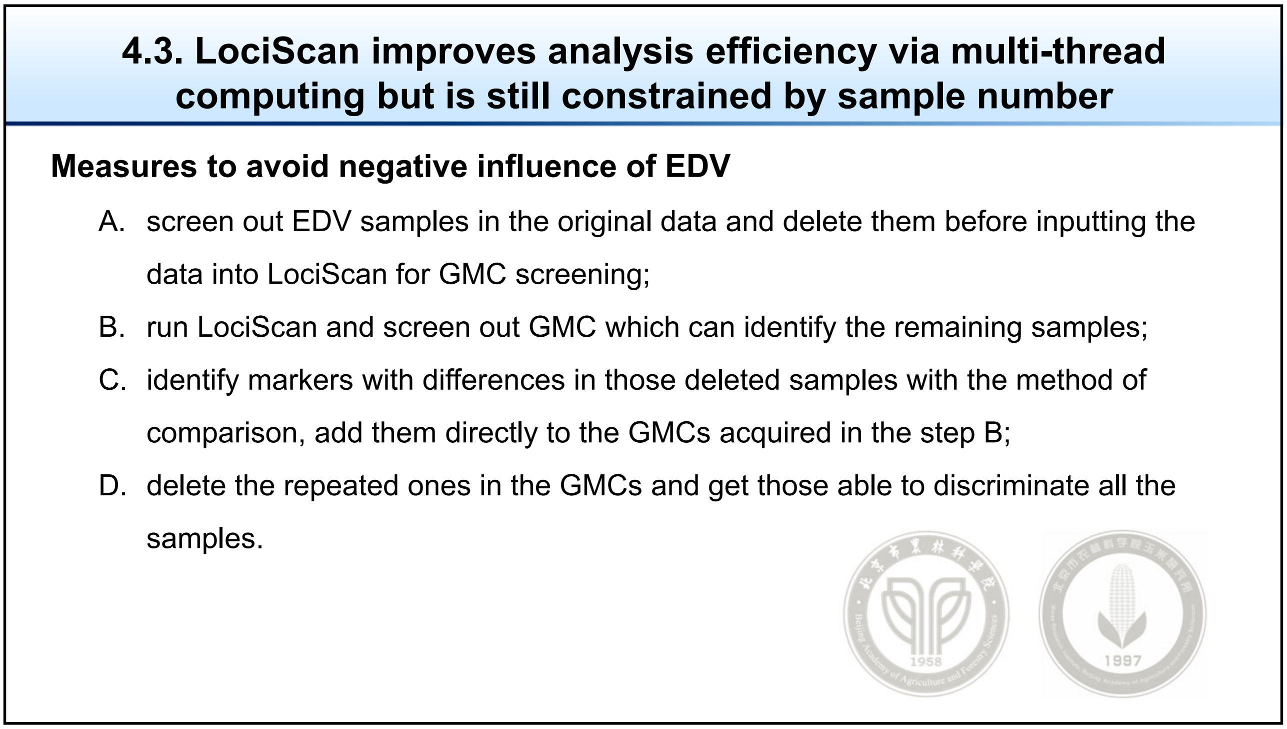
Limitation of efficiency

- the time complexity of calculating the GMC evaluation indices
- the parameter settings of the GA model
- the searching space of the combinatorial optimization problem domain

Influence of accuracy

- Evaluation indices (FF)
- Missing rate of training data
- The optimality of GA model

23



4.3. LociScan improves analysis efficiency via multi-thread computing but is still constrained by sample number

Measures to avoid negative influence of EDV

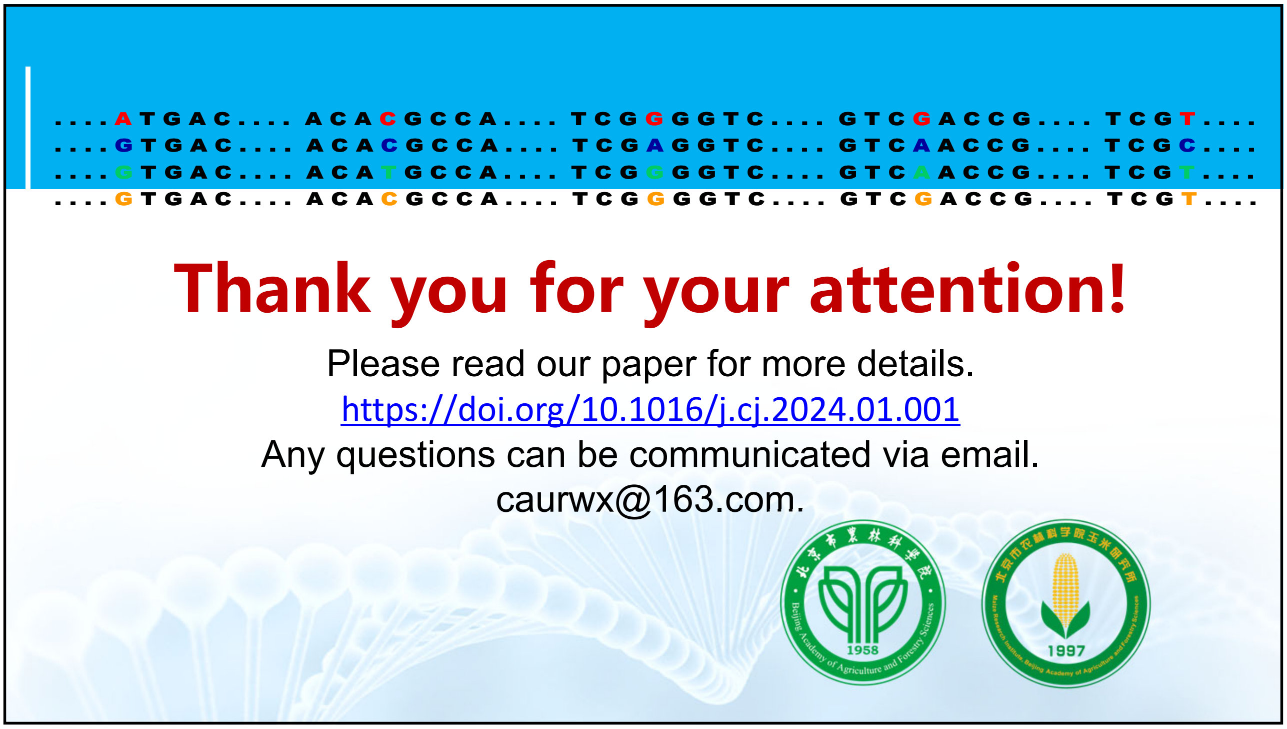
A. screen out EDV samples in the original data and delete them before inputting the data into LociScan for GMC screening;

B. run LociScan and screen out GMC which can identify the remaining samples;

C. identify markers with differences in those deleted samples with the method of comparison, add them directly to the GMCs acquired in the step B;

D. delete the repeated ones in the GMCs and get those able to discriminate all the samples.

24



....ATGAC.... ACACGCCA.... TCGGGGTC.... GTCGACCG.... TCGT....
....GTGAC.... ACACGCCA.... TCGAGGTC.... GTCAACCG.... TCGC....
....GTGAC.... ACATGCCA.... TCGGGGTC.... GTCAACCG.... TCGT....
....GTGAC.... ACACGCCA.... TCGGGGTC.... GTCGACCG.... TCGT....

Thank you for your attention!

Please read our paper for more details.
https://doi.org/10.1016/j.cj.2024.01.001
Any questions can be communicated via email.
caurwx@163.com.

25

[End of Annex and of document]