Population Structure of 93 Varieties of Rice (Oryza Sativa L. subsp. Hsien Ting) in Qinba, China

Yu Zhang (yuzhang20160315@outlook.com)
Shaanxi University of Technology https://orcid.org/0000-0002-5379-6773
Ye wen Wang
Hanzhong Agricaltural Sciences institute
Yue xing Wang
Xinjiang Agricultural University
Shi mao Zheng
Shaanxi University of Technology
Wan ying Zhou
Shaanxi University of Technology
Hong Liu
Hanzhong Vocational and Technical College

Research article

Keywords: Rice, SSR, SNPs, Population structure, Gene flow

Posted Date: December 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-113864/v1

Abstract

Background

The Qinba region is the transition region between indica and japonica varieties with a long history of indica varieties planting. 72,824 SNPs data based on GBS method, 48 pairs core primers of SSRs, and 15 agronomic traits were employed to explore the population structure of 93 rice varieties. The Mantel test was used to analyze the distance matrix generated using *NlaIII*-GBS only, *MseI*-GBS only, by combining *NlaIII*-GBS and *MseI*-GBS data and SSR.

Result

In this study, a total of 379 alleles were obtained using 48 pairs core primer of SSR, encompassing an average of 8.0 alleles per primer. The PPB and PIC was 88.65% and 0.77, respectively. Among these, RM278 possess the highest TNB and NPB, and the PPB in 29 pairs of SSR markers was 100%. RM176 showed the highest PIC. MAF was set to 0.05, and 39,872, 35,547 and 67,621 SNPs were obtained via NlaIII-GBS only, *MseI*-GBS only, and merged *NlaIII*-GBS and *MseI*-GBS data, respectively. The IBS genetic similarity coefficient average was 0.74. The results showed that the correlation between the genetic distance matrix based on *NlaIII*-GBS and *MseI*-GBS was the largest (R^2=0.88), followed by *NlaIII*-GBS and SSR (R^2=0.35), then by merged *NlaIII*-GBS and *MseI*-GBS data and SSR (R^2=0.33), and the smallest by *MseI*-GBS and SSR (R^2=0.27). The results showed that the 93 rice varieties could be clustered into two subgroups. Molecular variance analysis revealed that the genetic variation was 2% among populations and 98% within populations. Tajima's D value was 1.66, and the F_{ST} between the two populations was 0.61, and the Nm was 0.16.

Conclusion

The population genetic variation explained by SNP was larger than that explained by SSR. Through cluster analysis, the 93 samples were divided into 2 subgroups, with more than 97% of the samples clustered into one subgroup. The gene flow of 93 samples used in this study is larger than that of naturally self-pollinated crops, which may be caused by long-term breeding selection of indica varieties in the Qinba region. However, the genetic structure of the rice population is simple and lacked rare alleles.

Background

The Qinba region is a climate transition area between North and South China, and it is also the transition area from *indica* varieties to *japonica* varieties and the best suitable planting area for *indica* varieties. To date, no genetic research studies on the rice germplasm resources in this region have been conducted. The population genetic structure is the non-random distribution of genes or genotypes in space and time, including genetic variations within populations and genetic differentiation among populations. Population structure analysis is essential to explore the biological adaptability, the population formation process, evolutionary mechanism, protection, and development of biological resources[17]. At the same time, the population with identical or similar genetic backgrounds is most suitable for genome-wide association studies (GWAS), therefore, the study of population genetic structure plays an important role in the field of biology. Genetic markers have played an important role in studying population structure, ranging from earlier morphological markers to more recent DNA molecular markers. Eukaryotic genomes have simple sequence repeats (SSRs) that span approximately 10-50 kb. In the last few decades, SSR molecular markers have become important tools in the field of biology, particularly in terms of population structure, genetic mapping, and other related fields. SSR markers have also become the designated markers of the International Fingerprint Mapping Center. These are employed in judicial identification, identification of new varieties of plants, such as rice, rape, and corn[5, 14, 23, 26]. SSR markers are used in DNA fingerprinting for breed protection[24]. However, SSR markers are few in number, show unbalanced distribution in the genome, have weak electrophoretic resolution, and are relatively time-consuming and labor-intensive, and thus it is difficult to construct high-density genetic maps. With the recent development of next-generation sequencing technology, most biological studies have rapidly improved. In particular, the use of single nucleotide polymorphisms(SNPs) based on genome-wide scans[1], and with the release of extensive rice genome sequencing data, one SNP in every hundreds of base pairs or even every dozens of base pair has been identified, indicating that there are numerous SNPs in the rice genome[4, 9, 18] . A small number of SNPs can be used to resolve many problems, so the sequencing technology was born based on simplified genome by restriction site-associated DNA (RAD) tags. The frontrunner among these technologies is genotyping-by-sequencing (GBS), which has recently gained attention because it utilizes methylation-sensitive restriction endonucleases (type II enzyme), thereby avoiding repetitive regions of the genome (methylated regions) , GBS technology can rapidly identify high-density mutations, especially SNPs mutations. Currently, there are fewer commercially available restriction enzymes for selecting RAD tags for sequencing. In this study, two type II enzymes (*NlaIII* and *MseI*) were selected for digestion of the genome, which generated RAD tags for sequencing to obtain SNPs. Simultaneously, 48 core primer pairs of SSRs from NY/T1433-2014 that originated the Agricultural Standards of the People's Republic of China and 15 agronomic traits, namely, sowing date, plant height, leaf length, leaf width, effective number of panicles per plant, panicle length, total number of grains per panicle, number of filled grains per panicle, 1,000-grain weight, browning rate, milled rice rate, head milled rice rate, chalky grain rate, chalkiness degree, and length/width ratio, were employed to explore gene flow and population genetic structure of 93 rice varieties and to provide reference for future research studies using different genetic markers employed in related fields.

Results

Genotyping

SSR genotyping

A total of 378 bands was detected using 48 core SSR primer pairs (Table 1). Among these, 336 polymorphic bands were detected. The average number of polymorphic fragments was 7 ranging from 1 to 14. The maximum number of 14 polymorphic bands was detected by RM278 while RM311 is the least. The

average PPB was 88.87% ranging from 50% to 100%. The average PIC value was 0.77 ranging from 0.19 to 0.88.Those data showed that core SSR in rice can produces rich bands and high polymorphic rate.

Table 1 Information and amplification results of SSR primers

Primer name	Chr.	Sequence（5′-3′）	Annealing temperature(°C)	TNB	NPB	PPB(%)	PIC
RM583	1	F:agatccatccctgtggagag; R:gcgaactcgcgttgtaatc	55	10	10	100	0.86
RM71	2	F:ctagaggcgaaaacgagatg; R:gggtgggcgaggtaataatg	55	8	8	100	0.84
RM85	3	F:ccaaagatgaaacctggattg; R:gcacaaggtgagcagtcc	55	9	9	100	0.85
RM471	4	F:acgcacaagcagatgatgag; R:gggagaagacgaatgtttgc	55	8	6	75	0.86
RM274	5	F:cctcgcttatgagagcttcg; R:cttctccatcactcccatgg	55	12	12	100	0.84
RM190	6	F:ctttgtctatctcaagacac; R:ttgcagatgttcttcctgatg	55	5	5	100	0.74
RM336	7	F:cttacagagaaacggcatcg; R:gctggtttgtttcaggttcg	55	7	7	100	0.79
RM72	8	F:ccggcgataaaacaatgag; R:gcatcggtcctaactaaggg	55	12	9	75	0.86
RM219	9	F:cgtcggatgatgtaaagcct; R:catatcggcattcgcctg	55	2	2	100	0.36
RM311	10	F:tggtagtataggtactaaacat; R:tcctatacacatacaaacatac	55	2	1	50	0.37
RM209	11	F:atatgagttgctgtcgtgcg; R:caacttgcatcctcccctcc	55	4	3	75	0.67
RM19	12	F:caaaaacagagcagatgac; R:ctcaagatggacgccaaga	55	12	9	75	0.86
RM1195	1	F:atggaccacaaacgaccttc; R:cgactcccttgttcttctgg	55	8	8	100	0.84
RM208	2	F:tctgcaagccttgtctgatg; R:taagtcgatcattgtgtggacc	55	5	4	80	0.75
RM232	3	F:ccggtatccttcgatattgc; R:ccgacttttcctcctgacg	55	10	10	100	0.87
RM119	4	F:catccccctgctgctgctgctg; :cgccggatgtgtgggactagcg	67	7	4	57.14	0.79
RM267	5	F:tgcagacatagagaaggaagtg; R:agcaacagcacaacttgatg	55	9	5	56.56	0.85
RM253	6	F:tccttcaagagtgcaaaacc; R:gcattgtcatgtcgaagcc	55	6	6	100	0.75
RM481	7	F:tagctagccgattgaatggc; R:ctccacctcctatgttgttg	55	7	7	100	0.80
RM339	8	F:gtaatcgatgctgtgggaag; R:gagtcatgtgatagccgatatg	55	8	8	100	0.79
RM278	9	F:gtagtgagcctaacaataatc; R:tcaactcagcatctctgtcc	55	14	14	100	0.85
RM258	10	F:tgctgtatgtagctcgcacc; R:tggcctttaaagctgtcgc	55	7	6	85.71	0.80
RM224	11	F:atcgatcgatcttcacgagg; R:tgctataaaaggcattcggg	55	8	8	100	0.84
RM17	12	F:tgccctgttattttcttctctc; R:ggtgatcctttcccatttca	55	9	9	100	0.78
RM493	1	F:tagctccaacaggatcgacc; R:gtacgtaaacgcggaaggtg	55	7	7	100	0.83
RM561	2	F:gagctgttttggactacggc; R:gagtagctttctcccacccc	55	8	5	62.50	0.85
RM8277	3	F:agcacaagtaggtgcatttc; R:atttgcctgtgatgtaatagc	55	7	7	100	0.75
RM551	4	F:agcccagactagcatgattg; R:gaaggcgagaaggatcacag	55	6	6	100	0.68
RM598	5	F:gaatcgcacacgtgatgaac; R:atgcgactgatcggtactcc	55	9	5	55.56	0.75
RM176	6	F:cggctcccgctacgacgtctcc; :agcgatgcgctggaagaggtgc	67	10	7	70	0.88
RM432	7	F:ttctgtctcacgctggattg; R:agctgcgtacgtgatgaatg	55	5	5	100	0.71
RM331	8	F:gaaccagaggacaaaaatgc; R:catcatacatttgcagccag	55	8	7	87.50	0.82
OSR28	9	F:agcagctatagcttagctgg; R:actgcacatgagcagagaca	55	10	9	90	0.80
RM590	10	F:catctccgctctccatgc; R:ggagttggggtcttgttcg	55	9	6	66.67	0.87
RM21	11	F:acagtattccgtaggcacgg; R:gctccatgagggtggtagag	55	11	11	100	0.87
RM3331	12	F:cctcctccatgagctaatgc; R:aggaggagcggatttctctc	50	6	4	66.67	0.80
RM443	1	F:gatggttttcatcggctacg; R:agtcccagaatgtcgtttcg	55	10	7	70	0.75
RM490	1	F:atctgcacactgcaaacacc; R:agcaagcagtgctttcagag	55	9	9	100	0.82
RM424	2	F4:tttgtggctcaccagttgag; R:tggcgcattcatgtcatc	55	5	5	100	0.72
RM423	2	F:agcacccatgccttatgttg; R:ccttttttcagtagccctccc	55	7	7	100	0.82
RM571	3	F:ggaggtgaaagcgaatcatg; R:cctgctgctctttcatcagc	55	7	7	100	0.67
RM231	3	F:ccagattatttcctgaggtc; R:cacttgcatagttctgcattg	55	12	12	100	0.84
RM567	4	F:atcagggaaatcctgaaggg; R:ggaaggagcaatcaccactg	55	10	10	100	0.78

RM289	5	F:ttccatggcacacaagcc; R:ctgtgcacgaacttccaaag	55	10	10	100	0.88
RM542	7	F:tgaatcaagcccctcactac; R:ctgcaacgagtaaggcagag	55	8	7	87.50	0.84
RM316	9	F:ctagttgggcatacgatggc; R:acgcttatatgttacgtcaac	55	2	2	100	0.19
RM332	11	F:gcgaaggcgaaggtgaag; R:catgagtgatctcactcaccc	55	10	8	80	0.88
RM7102	12	F:taggagtgtttagagtgcca; R:tcggtttgcttatacatcag	55	3	3	100	0.43

SNPs genotyping

A total of 39,872 SNPs and 35,547 SNPs passed the minor allele frequency (MAF) lower limit of 0.05 by *NlaIII*-GBS only and *MseI*-GBS only, respectively. Then mergered *NlaIII*-GBS and *MseI*-GBS data, a total of 72,824 SNPs including 67,621 SNPs that aligned specific chromosomes and 5,023 SNPs that do not aligned specific chromosomes. The average MAF was 0.21 of 93 samples.

Linkage disequilibrium (LD) decay and Haplotype construction

In a total of 6,288,753 loci, among which 326,873 (5.198%) were heterozygous based on 67,621 SNPs. The 67,621 SNPs were unevenly distributed among the 12 chromosomes (Fig.1a); chromosome 1 contained the largest amount of makers (8,425), while chromosome 8 included the least (3,953). Among the 84,255 SNP pairs, R^2 value had a minimum of 0.2 and an average of 0.73. 46,322 SNP pairs (54.98%) had R^2values higher than 0.8, while 7,841 pairs (9.31%) were in complete LD (R^2=1). The inter-marker genetic distance between all pairs, between pairs of SNPs with R^2 inferior to 0.8, and between pairs with R^2 equal or superior to 0.8 had average values of 154,130 bp, 171,768 bp, and 139,686 bp, respectively. LD, as represented by inter-loci R^2 values, decreases as the physical distance between loci increases (Fig. 1b). The 12 chromosomes yielded a total of 6568 predicted haplotypes (Fig.1c), with chromosome 1 possessing the most haplotypes (776) and chromosome 10 possessing the least (349). The largest haplotype was composed of 95 SNPs. The longest haplotype spanned 200.0 kb, the average length of the haplotype is 33.71kb.

Population genetic structure analysis

We employed two types of genetic markers including three kinds of DNA markers and 15 agronomic traits to perform cluster analysis.

PC analysis

Principal component analysis was performed to select the first three PC(eigenvalue) and their cumulative contribution of variance accounted for 40.69%,39.76% 40.10% and 15.76% by *NlaIII*-GBS only, by *MseI*-GBS only, by mergered *NlaIII*-GBS and *MseI*-GBS data and by SSR, respectively. PCA separated the 93 genotypes into two subgroups (Fig.4) which were consistent with the UPGMA and STRUCTURE results. W366 and W367 has been separately clustering (Fig. 2).

UPGMA

The unweighted pair-group method with arithmetic means (UPGMA) algorithm was performed on the 93 genotypes, which demonstrated that the 93 genotypes could be divided into 2 subgroups (Fig. 3). Group I included 1 to 3 samples, respectively. and group II contained 92 to 90 samples. The average genetic distance was 0.29 ranging from 0.02 to 0.55 based on mergered *NlaIII*-GBS and *MseI*-GBS data,which the two most closely related materials were W710 and W711, the two most greatest materials were W366 and W740.

Bayesian clustering

MAF <5% , 70824 SNPs were used to assess the population structure of the entire pool of rice germplasms. Delta K reached a maximum value at K=2, suggesting that the 93 rice samples were divided into two subgroups (consisting of 70 and 23 samples) (Fig. 4).In the population structure analysis, the results from K = 2 to K = 5 revealed the occurrence of gene introgression between groupI and groupII, accounting for approximately 76.34% of the observed variations (calculated by K = 2).

The analysis performed using PCA, UPGMA and Bayesian clustering got similar results, The population structure is relatively simple,the matrix delamination is not distinctive.

Agronomic traits clustering

Clustering result based on 15 agronomic traits show in fig. 5a. the 92 materials are gathered together in addition to W669, showed a single genetic basis for the population.

Clustering of different category materials

We analyze population genetic information of the different category materials including 57 restoring lines, 19 maintainer lines and 17 special rices, respectively (Table 2), and UPGMA clustering (Fig. 5b, 5c, 5d) .The results showed the genetic basis of the restorer line was more abundant than that of the maintainer line,and the genetic basis of the special rice was wider than that of the conventional rice.

Table 2 Population genetic analysis of different category materials

Samples	Tajima'D	Range of IBS genetic distance	the average genetic distance	the two most closely related materials	the two most greatest related materials
whole materials(93)	1.66	0.0229-0.5452	0.3007	W710/W711	W366/W740
Restoring lines(57)	1.36672	0.0229-0.3927	0.2666	W710/W711	W685/W697
Maintainer lines(19)	0.43533	0.0242-0.3745	0.2293	W740/W741	W725/W738
Special rice(17)	0.62542	0.0285-0.5315	0.3280	W375/W380	W300/W366

Correlation analysis among genetic distance matrices by three-types of marker dataset

In the present study, the coefficients of correlation (R^2) between the genetic distance matrices of *NlaIII*-GBS only and *MseI*-GBS only, *NlaIII*-GBS only and SSR, *MseI*-GBS only and SSR, mergered *NlaIII*-GBS and *MseI*-GBS and SSR were 0.88, 0.35, 0.27,0.33, respectively (Fig. 6), this may be due to the different number of markers used.

AMOVA and gene flow

Tajima's D value was 1.66, which signifies low levels of both low- and high-frequency polymorphisms, indicating a decrease in population size and/or balancing selection. This resulted in more haplotypes and lacked rare alleles in the population. Analysis of molecular variance showed that the genetic variation within the population was 98% and between populations was 2%, which indicated the existence of slight genetic variation among 93 samples. The genetic differentiation coefficient (F_{ST}) between the two populations was 0.61, and gene flow (Nm) was 0.16. Further investigation showed that the gene flow of selfing crops was the smallest, and that of annual herbaceous plants was the lowest. If Nm > 1, which indicates that the level of gene flow between populations is high, then genetic differentiation among populations is small; if Nm > 4, then gene communication between populations is more adequate and genetic differentiation is smaller; and Nm < 1 indicates population differentiation may have occurred due to genetic drift. The gene flow was 0.16, which indicates that the gene flow among rice populations in the Qinba region is lower, but nearly 2.5-fold larger than that of conventional inbred plants, which may result in long-term artificial selection, leading to reduced genetic differentiation.

Discussion

Germplasm resources are the basis of rice breeding. The analysis of genetic diversity and genetic structure of rice germplasm resources is beneficial to mining excellent breeding materials and improving breeding efficiency. The results showed that 48 pairs core primers of SSRs had rich bands and high polymorphism. It has been widely used in the study of genetic diversity and identification of varieties [27, 17, 12, 3]. A total of 72824 SNPs were obtained by genome sequencing based on two enzymes (*NlaIII* and *MseI*), indicating that there are abundant SNPs in rice. Previous sequencing results showed that there is one SNP in every hundreds or even tens of base pair in rice genome [8, 16, 13, 14, 24, 19, 21], indicating that there were a large number of SNPs in the rice genome, which had been applied in gene mapping, functional gene detection, assisted breeding selection and other aspects of rice [21, 29]. The results of this study are consistent with previous sequencing results.

Different DNA markers reflect different ranges of polymorphisms in the genome. In recent decades, SSR markers, which represent second-generation DNA molecular markers, have been widely used for plant population genetic analysis, phylogenetic reconstruction, and quantitative trait mapping. Theoretically, the more markers used in a study, the more accurate the results are. SSR markers are mostly distributed in centromeres, telomeres, introns, and 3'UTR regions, and most of these are non-functional markers of genes. If combined with functional SNP markers, the genetic structure of the population revealed would be more accurate. With the rapid development of the next-generation genome sequencing technology, it is now possible to obtain a large number of sequence data at a low cost as well as increase the reliability of the results. The PCA, UPGMA and Bayesian clustering are three commonly used methods for population genetic structure, the analysis of the population genetic structure is based on the estimation of genetic distance or genetic similarity coefficient matrix between samples, For example, in PC analysis, the accumulative contribution rate of the first three major factors analyzed by SSR was only 15.76%, far less than that using SNP data, it indicates that the more DNA polymorphism, the more accurate the population variation can be explained.The results showed that the genetic structure of 93 materials was relatively simple and the matrix stratification was not obvious. The cluster based on 15 agronomic traits also showed that the genetic basis of the other 92 materials was single except W669. The above studies showed that the parents of rice breeding in Qinba area had higher genetic similarity, single genetic background and low genetic polymorphism. The research results are consistent with the previous conclusions of our research group [28]. This may be due to many reasons, such as the wide exchange of variety resources among breeding units in the process of breeding, and the similar breeding goals,which makes the genetic basis of breeding varieties similar. The narrow genetic basis limits the cultivation of rice varieties, the breeding of excellent characters and increasing rice yield.

Through the analysis of different types of materials (57 restorer lines, 19 sterile lines and 17 special rice), the results showed that the genetic basis of restorer lines was richer than that of maintainer lines, which was consistent with the conclusion of Ying Jiezheng et al. [27]. The main reason may be that most cytoplasmic male sterile lines currently used in production are related to Zhenshan 97B, □ - 32B, zhong9a and gang46a, or they are derived from Chinese Dwarf early rice varieties aizazhan and aijiaonante. At present, the restorer lines used in combination production are from the Yangtze River Basin of China, Sichuan, Southeast Asia, South Korea, etc. and the restorer lines created by crossing indica and japonica rice. Compared with other rice germplasm resources, special rice has rich genetic basis and high breeding potential.

By analyzing different types of materials (57 restorer lines, 19 sterile lines and 17 special rice), the results showed that the genetic basis of restorer lines was richer than that of maintainer lines, which was consistent with the conclusion of Ying Jiezheng et al. [27]. Perhaps the main cause is the production on the application of cytoplasmic male sterile lines most with Jane shanyou 97 b, □- 32 b, 9 a and 46 a post in the blood, or China's variety of dwarf rice dwarf is

accounted for and short feet nantes derivative. The restorer line is rich in sources. Currently, the restorer line used in combination production is consanguineous from Yangtze River Basin, Sichuan, Southeast Asia, Korea, etc., as well as restorer line created by crossing indica and japonica. Special rice has an abundant genetic basis compared to other rice germplasm resources and has high breeding potential.

Conclusions

Among the three commonly used methods of population structure analysis, Bayesian algorithm is more practical than UPGMA and PC analysis, considering the known pedigree knowledge. At the same time, the size of gene flow of each samples can be seen from the population genetic structure graph based on Bayesian algorithm. Genetic effects in populations depend on the opportunity distribution of MAFs across the genome-wide, and different populations have different MAF values. Although gene flow in the population consisting of 93 rice varieties is large, the average MAF of the population was only 0.21. It is an important measure to improve the genetic diversity of rice varieties and the main way to improve the yield of rice varieties in Qinba area. Any one of the two sets of GBS data used in this study can be utilized in the analysis of population genetic diversity. However, particularly for subsequent QTL mapping, the greater the number of polymorphic sites, the higher the mapping resolution, especially for *NlaIII*-GBS data, which possess CATG recognition sequences, so the selected RAD tags digested by *NlaIII* enzyme may be the functional regions of the gene and will be studied in subsequent QTL mapping efforts.

Methods

Plant Materials

A total of 93 accessions, representing most of the Oryza sativa L. subsp. Hsien Ting varieties (lines) germplasm resource of the Qinba area in China, were collected from the rice experimental farm during the 2018 growing season in the Shaanxi Rice Research Institute and comprised 57 restoring lines, 19 maintainer lines, and 17 special rice.

Genomic DNA extraction, primer synthesis, PCR amplification,and GBS

The genomic DNA of 93 rices was extracted from fresh leaves using the SDS technique and detected with 0.8 % agarose gel electrophoresis. The 48 SSR primers were synthesized by Beijing Aoke Biotechnology Co., Ltd. (Beijing, China). PCR were carried out in a 10 μL volume containing 1 μL DNA template, 2μL(10μM)of forward and reverse primers(1μL each), 5μL 2×Taq Master Mix, 2 μL RNase-free water. the reactions were programed as follows: initial denaturation at 94.0 °C for 5 minutes, denaturation at 94.0 °C for 1 minute, annealing at 50-60.0 °C for 1 minute, and extension at 72.0 °C for 1 minute, for a total of 35 cycles. Electrophoresis was performed using 8% non-denaturing polyacrylamide gel under 95V voltage; the bands were visualized via silver staining.

The genomic DNA of 93 rices was digested using NlaIII and MseI enzyme, respectively. GBS performed by the Illumina Hiseq 2000 platform of Novo Gene Bioinformatics Technology Co.,Ltd (Beijing, China). The GBS data obtained with NlaIII digestion, MseI digestion were recorded as *NlaIII*-GBS, *MseI*-GBS, respectively. Then after filtering the polymorphisms with dp. Miss and MAF of 2,0.3 and 0.05, respectively. SNPs filtered were annotated based on reference genome (ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/oryza_ indica/dna/).

LD decay and Haplotype construction

Genotype data were then used to calculate LD between SNPs and to construct haplotypes using the EM algorithm implemented by PLINK1.07 (https://www.cog-genomics.org/plink2). The command line "–r2" and "–blocks" were used to calculate LD and assign SNPs to their respective haplotypes by calculating inter-maker LD within a 200kb window, respectively. The figures were constructed using the Origin8 platform(http://www.originlab.com/).

SSR marker efficiency analysis

Following electrophoresis, each amplification band corresponded to a primer hybridization locus and was considered as an effective molecular marker. Each polymorphic band detected by a same given primer represented an allelic mutation. In order to generate molecular data matrices, clear bands for each fragment were scored in every accession for each primer pair and recorded as 1 (presence of a fragment), 0 (absence of a fragment), and 9 (complete absence of band). The value of the polymorphism information content (PIC) was calculated by PIC_Calc 0.6 program, where PIC represents the PIC value of the ith locus and Pij represents the frequency that allele j appears in the ith locus. The value of PIC varies from 0 to 1, with 0 indicating an absence of polymorphism at a given locus and 1 reflecting multiple alleles at a given locus. The level of polymorphism of each marker was assessed by the polymorphism information content [2], which measures the extent of genetic variation: PIC values smaller than 0.25 indicates low levels of polymorphism associated to a locus, PIC values between 0.25 and 0.5 imply moderate levels of polymorphism, while PIC values greater than 0.5 indicate high levels of polymorphism

PC analysis

PC analysis was performed under Eigen module by NTSYS-pc2.10e [20].

UPGMA

Identity-by-state (IBS) distance matrix generated by TASSEL5.0 (http://www.maizegenetics.net/tassel), building a UPGMA tree, MEGA5.0 (http://http://www.megasoftware.net/) editing and visualizing.

Bayesian clustering

The Bayesian clustering algorithm implemented in STRUCTURE 2.3.4 (http://taylor0.biology.ucla.edu/structureHarvesteroybase.org/tools.php) was used to simulate population genetic structure with the assumption that all of the genetic markers were independent. To obtain an estimate of the most probable number of population (K), K values from 1 to 5 were simulated with 5 iterations for each K, using 100000 burn-in periods followed by 100000 Markov Chain Monte Carlo iterations. Delta K was plotted against K values and the best number of clusters was determined following the method proposed by Evanno et al [7].

In this study, STRUCTURE 2.3.4, which applies a Bayesian clustering algorithm, was used to simulate population genetic structure based on the assumption that the 72824 loci were independent. Using a membership probability threshold of 0.60, population K values from 1 to 5 were simulated with 5 iterations for each K using 10,000 burn-in periods followed by 10,0000 Markov Chain Monte Carlo iterations in order to obtain an estimate of the most probable number of population. Delta K was plotted against K values; the best number of clusters was determined following the method proposed by Evanno et al and obtained via the Structure Harvester platform (http://taylor0.biology.ucla.edu/structureHarvester/)[6].

Agronomic traits clustering

SPSS20 was employed to perform agronomic traits clustering by nearest neighbor analysis.

Correlation analysis among genetic distance matrices by diffrent DNA marker dataset

Mantel tests were used to measure the correlation between the genetic distance matrices generated by *NlaIII*-GBS only, *MseI*-GBS only, merged *NlaIII*-GBS and *MseI*-GBS data and SSR. It was carried out by the GenAlEx software with permutation 9999 times[21]. $r \geq 0.9$, $0.8 \leq r < 0.9$, $0.7 \leq r < 0.8$, and $r < 0.7$ represented significant correlation, moderate correlation, weak correlation, and no correlation, respectively.

AMOVA and Gene flow

For the analysis of molecular variance (AMOVA), 93 accessions were classified into two groups by 72824 SNPs. The components of variance attributable to different varieties and breeding lines were estimated from the genetic distance matrix using Tajima & Nei method, as specified in the AMOVA procedure in ARLEQUIN 3.1[8]. A nonparametric permutation procedure with 9999 permutations was used to test the significance of variance components associated with the different possible levels of genetic structure in this study. The pairwise Fst values, a value of F statistic analogs computed from AMOVA, were used to compare genetic distances between any two groups.

Abbreviations

AMOVA: Analysis of molecular variance; DNA: Deoxyribonucleic acid; GWAS: genome-wide association studies; GBS: genotyping by sequencing; IBS: Identity by state; LD: Linkage disequilibrium; MAF: Minor allele frequency; NPB: number of polymorphic bands; PPB: Percentage of polymorphic bands; PIC: Polymorphism information content; PCA: Principal component analyses; RAD: restriction site-associated DNA; SNP: Single nucleotide polymorphism; SSR: Simple sequence repeats; TNB: total number of bands; UPGMA: Unweighted pair group method with arithmetic mean

Declarations

Acknowledgments

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Authors' contributions

Yu Zhang designed the study and analyzed the data, Yewen Wang wrote the manuscript, Yuexing Wang and Shimao Zheng performed the experiments, Wanying Zhou and Hong Liu edited and revised the manuscript. and all authors approved the manuscript.

Funding

This study was supported by the Sci-technological Project of Shaanxi Province (NYKJ-2016-35), the Sci-technological Project of Shaanxi Province (2013K02-10-01), the Sci-technological Project of Shaanxi Province (2020NY-050), the Sci-technological Project of Shaanxi Province (2019NY-041)

Ethics approval and consent to participate

All authors read and approved the manuscript.

Consent for publication

All authors agreed to publish this manuscript.

Competing interests

The authors declare that they have no competing interests.

Author Details

[1]School of Biological Sciences and Engineering, Shaanxi University of Technology, 72300 Hanzhong Shaanxi, China.

[2]Hanzhong Agricultural Sciences Institute, 723000 Hanzhong Shaanxi,China .

[3]College of Agronomy, Xinjiang Agricultural University, 830052 Urumqi, China.

[4]Hanzhong Vocational and Technical College, 723000 Hanzhong Shaanxi, China.

References

1. Aslam ML, Bastiaansen JWM, Elferink MG. Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo). BMC Genomics. 2012;13:391.
2. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32(3):314-331.
3. Ceng CS, Peng D, Shi YY, Xie w, Liu AM. Fingerprinting Construction of Rice Core Parental Lines with SSR Markers. Crop Research. 2016;30(05):481-486+511.
4. Chen HD, Xie WB, He H. A High-Density SNP Genotyping Array for Rice Biology and Molecular Breeding. Molecular Plant. 2014;7(3):541-553.
5. Delphine VI, Albrecht EM, Claude L, Benjamin S. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theoretical and Applied Genetics. 2010;120(7): 1289 -1299.
6. Earl, Dent A. and vonHoldt, Bridgett M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. 2012;4(2):359-361 .
7. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol, 2005;14(8):2611-2620.
8. Excoffier L, Laval G, Schneider S.Arlequin: an integrated software package for population genetics data analysis. Version 3.0. Computational and Molecular Population Genetics Laboratory (CMPG). Berne (Switzerland): Institute of Zoology, University of Berne.
9. Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH. An SNP resource for rice genetics and breeding based on subspecies incica and japonica genome alignments.Genome Res. 2004;14(9): 1812-1819.
10. Chen HD, Xie WB, He H. A High-Density SNP Genotyping Array for Rice Biology and Molecular Breeding. Molecular Plant. 2014;7(3):541-553.
11. He GL, Fu GP, Peng CS, Deng W, Zhu S, Yang Y, et al. DNA Fingerprint Map and Analysis of Genetic Diversity of the Japonica Rice Varieties in the Regional Test in Jiangxi Province in 2018. Acta Agriculturae Universitatis Jiangxiensis. 2019;41(05):843-852.
12. Hong G, Cue Y, Han B. Sequence and analysis of rice chromosome 4. Nature. 2002;420: 316-320.
13. Li C, Zhang Y, Ying K, Liang XL, Han B. Sequence variations of simple sequence repeats on chromosome 4 in two subspecies of the Asian cultivated rice. Thear Appl Genet. 2004;108(3): 392-400.
14. Li FB , Yang J, Lv ZW, Yi B, Wen J, Fu TD, et al. Screening of Brassica napus core SSR primers. Chinese Journal of Oil Crop Sciences. 2010;32(3):329-336.
15. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, et al. A draft sequence of the rice genome . Science. 2002;296: 79-92.
16. Lin YX, Wang AX, Liu H, Wang Z, Liang MZ, Dai XJ, et al. Research on DNA Molecular Digital Fingerprint Database Based on 48 Pairs of SSR Primers for 94 Hybrid Rice Parents in NYT 1433-2014. Chinese Journal of Rice Science. 2016;30(06):593-602.
17. Loveless MD, Hamrick JL. Ecological determinants of genetic structure in plant populations. Annu.Rev.Ecol.Syst. 1984; 15:65-95.
18. Nasu S, Suzuki J, Ohta R, Hasegawa K, Yui R, Kitazawa N, et al. Search for and analysis of single nucleotide polymorphisms (SNPS) in rice (Oryza sativa, Oryza rufipogon) and establishment of SNP markers. DNA Res. 2002;9(5): 163-171.
19. Pan CH, Wang ZB, Ma YY, Yin YJ, Zhang YF, Zuo SM, et al. InDel and SNP Markers and Their Application in Map-Based Cloning of Rice Genes. Chinese Journal of Rice Science. 2007;21(5):447-453.
20. Rohlf F J. NTSYS-pc. Numerical Taxonomy and Multivariate Analysis System, Version 2.02. Exeter Software, New York. 1998.
21. Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. Bioinformatics. 2012;28(19):2537-2539.
22. Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, et al. Development of genome-wide DNA Polymorphism database for map-based cloning of rice genes. Plant Physial. 2004;125(3): 1198-1205.
23. Sui GL, Yu SC, Yang JX, Wang WH, Su TB, Zhang FL, et al. Validation of a Core Set of Microsatellite Markers and Its Application for Varieties Identification in Chinese Cabbage. Acta Horticulturae Sinica. 2014;41(10):2021-2034.
24. Teng HT, Lv B, Zhao JR, Xu Y, Wang FG, Du YY, et al. DNA Fingerprint Profile Involved in Plant Variety Protection Practice. Biotechnology Bulletin. 2009; (1):1-6.
25. Wang MH, Zhang XT, Wu GL, Jiang Q, Shi YH. DNA Fingerprints Construction and Purity Identification Based on SSR Markers for Rice Varieties in Ningbo City. China Rice. 2019;25(06):50-54.
26. Yin GY, Yang XY, He QB, Qu CM, Zhang JH, Dai XM. Screen and Identification of SSR Core Primers for Tobacco Germplasm. Journal of Plant Genetic Resources. 2013;14(5): 960-965.
27. Ying JZ, Shi YF, Zhuang JY, Xue QZ. Microsatellite Marker Evaluation on Genetic Diversity of the Major Commercial Rice Varieties in China. Scientia Agricultura Sinica. 2007;(04):649-654.

28. Zhang Y, Chen XY, Wang SB, Mo D. Analysis of SSR Marker-based Polymorphism of Rice Mainly Popularized in Hanzhong. Chinese Agricultural Science Bulletin. 2011;27(07):34-37.
29. Zhang Y, Zhang XJ, Yang FJ, Feng ZF. SNP-based Detecting Method of Rice Blast Resistance Gene Pi-ta. Chinese Journal of Rice Science. 2013;27(3):325-328.

Figures

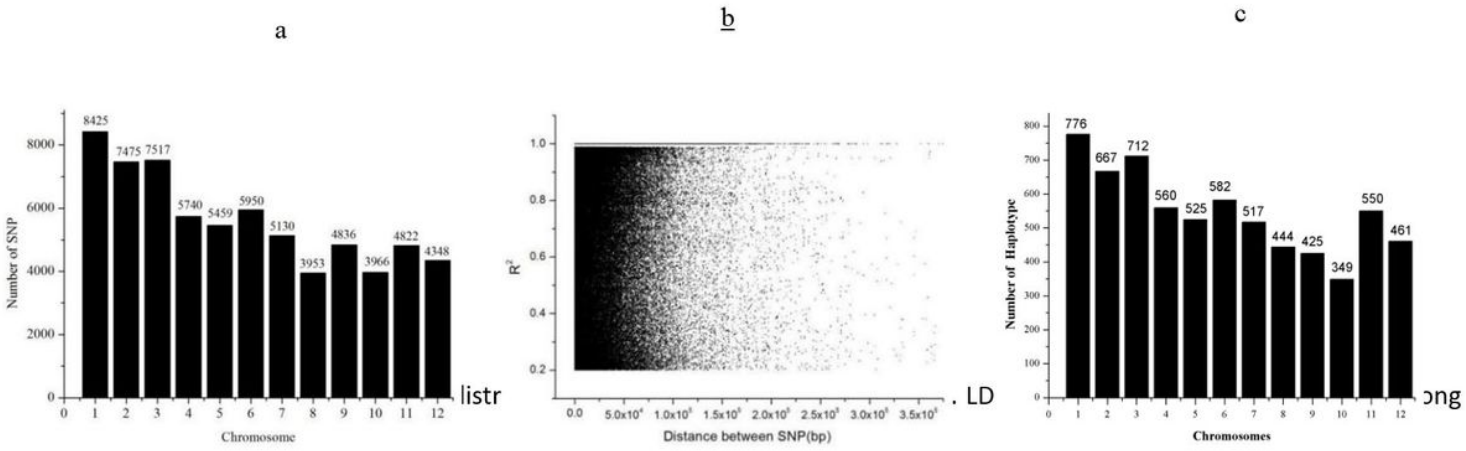

Figure 1

In a total of 6,288,753 loci, among which 326,873 (5.198%) were heterozygous based on 67,621 SNPs. The 67,621 SNPs were unevenly distributed among the 12 chromosomes (Fig.1a); chromosome 1 contained the largest amount of makers (8,425), while chromosome 8 included the least (3,953). Among the 84,255 SNP pairs, R2 value had a minimum of 0.2 and an average of 0.73. 46,322 SNP pairs (54.98%) had R2 values higher than 0.8, while 7,841 pairs (9.31%) were in complete LD (R2=1). The inter-marker genetic distance between all pairs, between pairs of SNPs with R2 inferior to 0.8, and between pairs with R2 equal or superior to 0.8 had average values of 154,130 bp, 171,768 bp, and 139,686 bp, respectively. LD, as represented by inter-loci R2 values, decreases as the physical distance between loci increases (Fig. 1b). The 12 chromosomes yielded a total of 6568 predicted haplotypes (Fig.1c), with chromosome 1 possessing the most haplotypes (776) and chromosome 10 possessing the least (349). The largest haplotype was composed of 95 SNPs. The longest haplotype spanned 200.0 kb, the average length of the haplotype is 33.71kb.

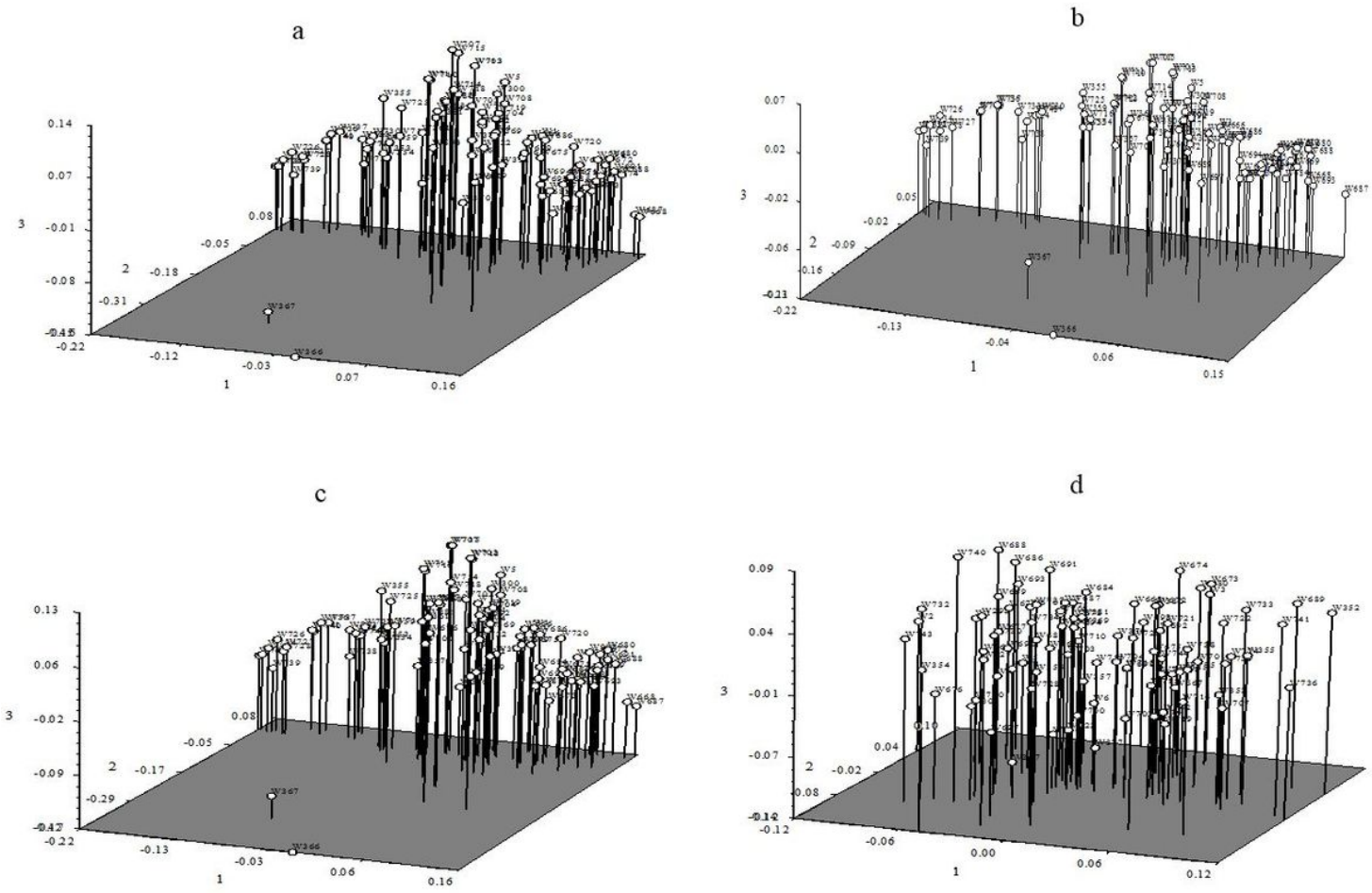

Figure 2

PCA plots by the different DNA markers. a. by NlaIII-GBS only; b. by MseI-GBS only; c. by merged NlaII-GBS and MseI-GBS data; d. by SSR

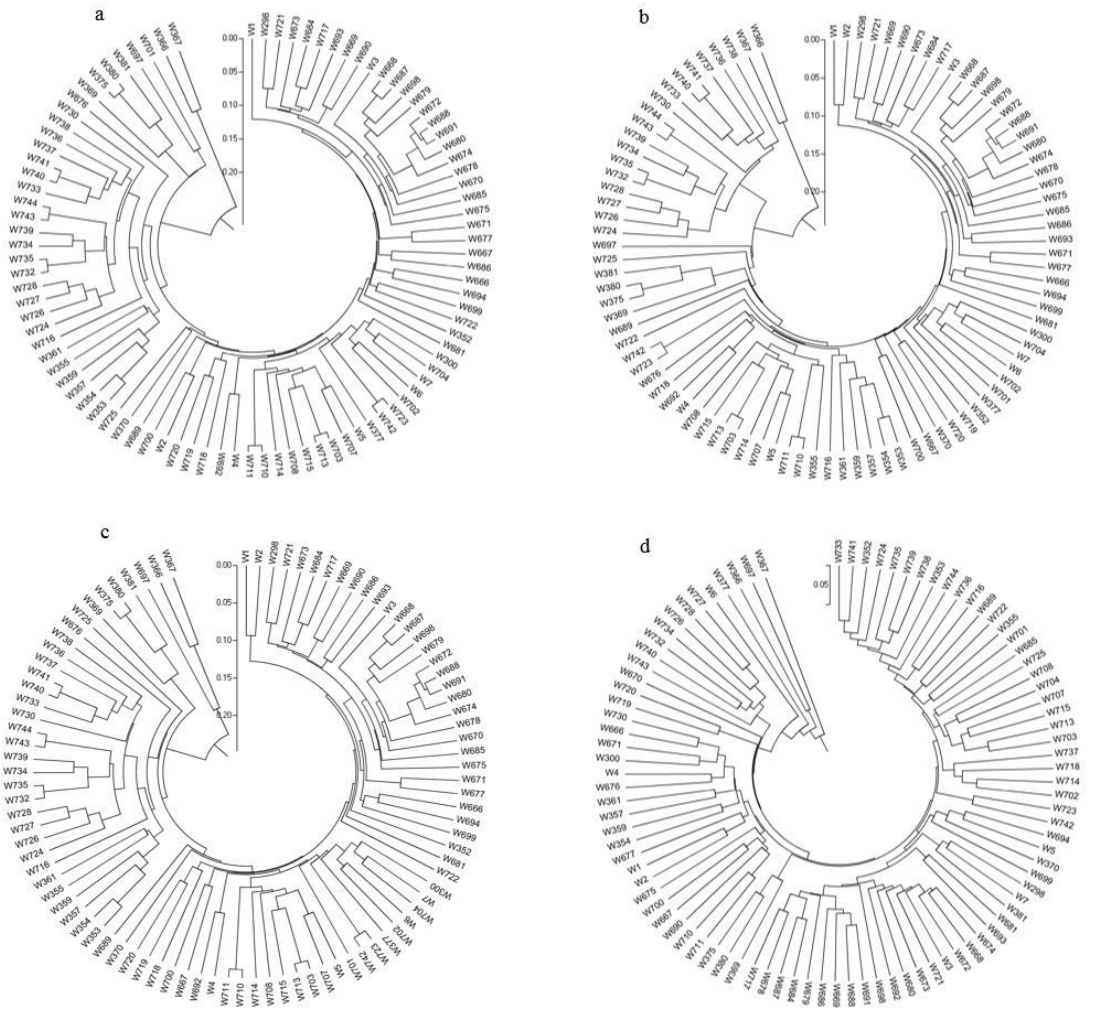

Figure 3

UPGMA Clustering by different DNA markers. a. by NlaIII-GBS only; b. by MseI-GBS only; c. by merged NlaIII-GBS and MseI-GBS; d.by SSR

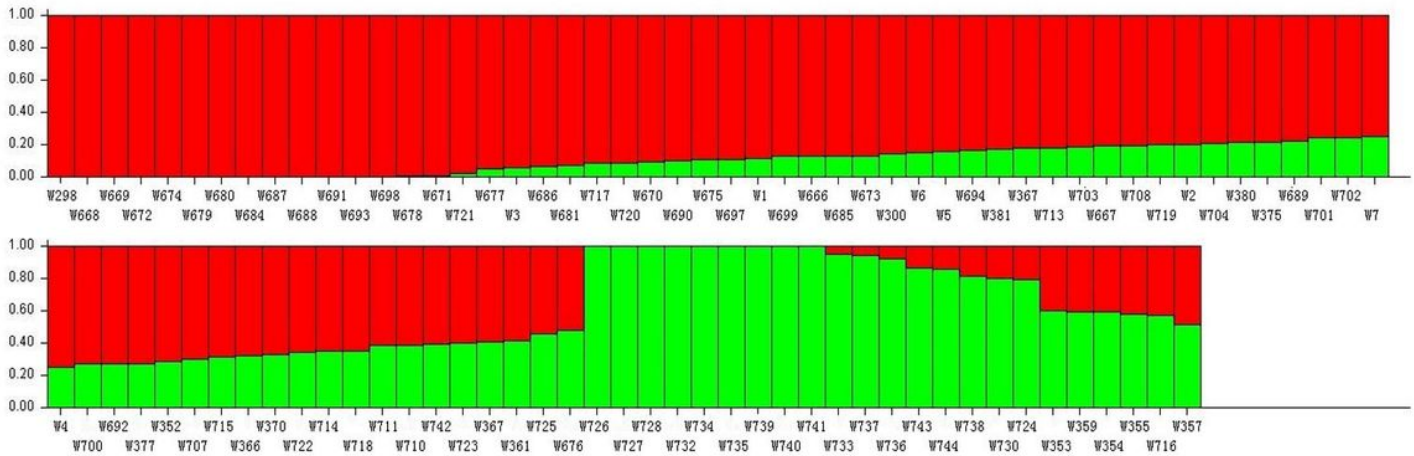

Figure 4

Bayesian clustering based on 72824 SNPs for 93 rice; Red: group I; Green: group II. Vertical lines on the X-axis refer to each sample. The proportion of each color represents probability rate with which a given genotype belongs to each group.

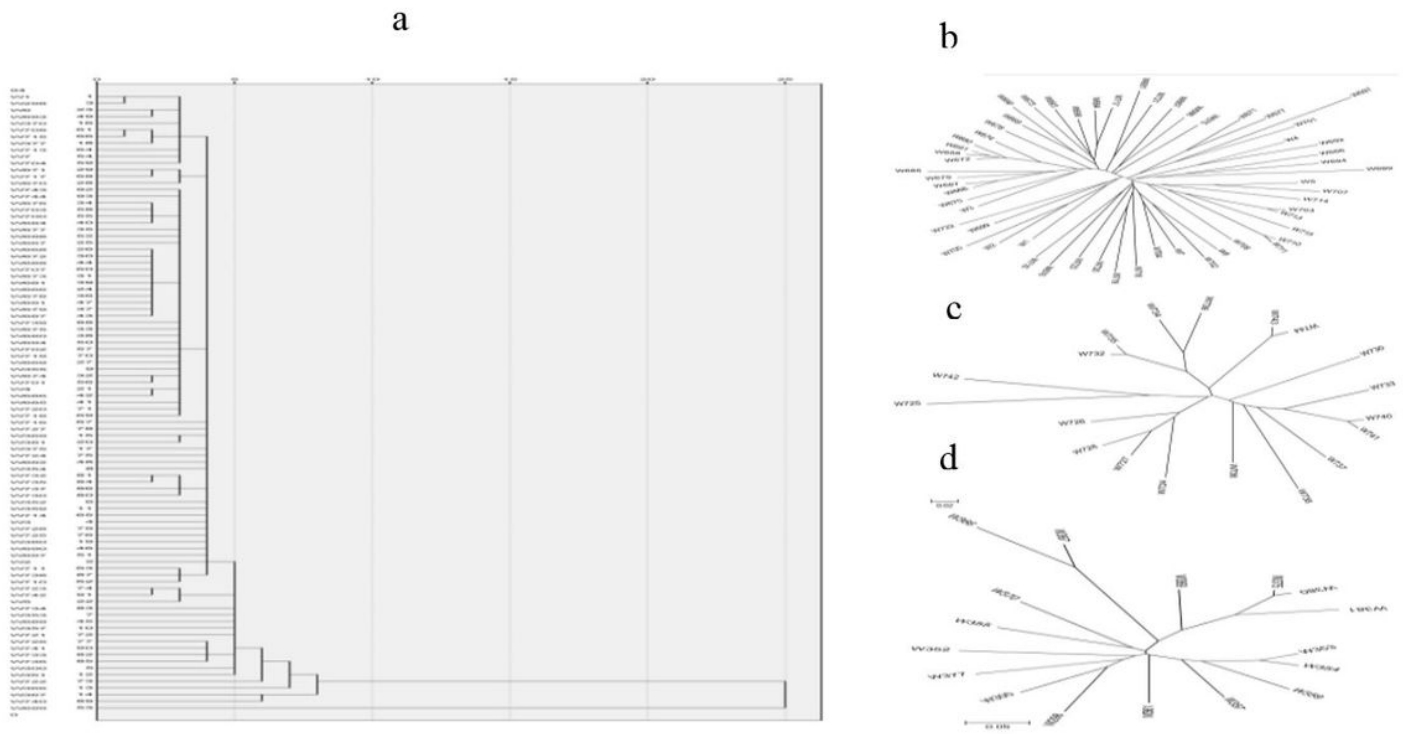

Figure 5

Clustering. a. 15 agronomic traits clustering. b. 57 restoring lines clustering. c. 19 maintainer lines clustering. d. 17 special rices clustering.

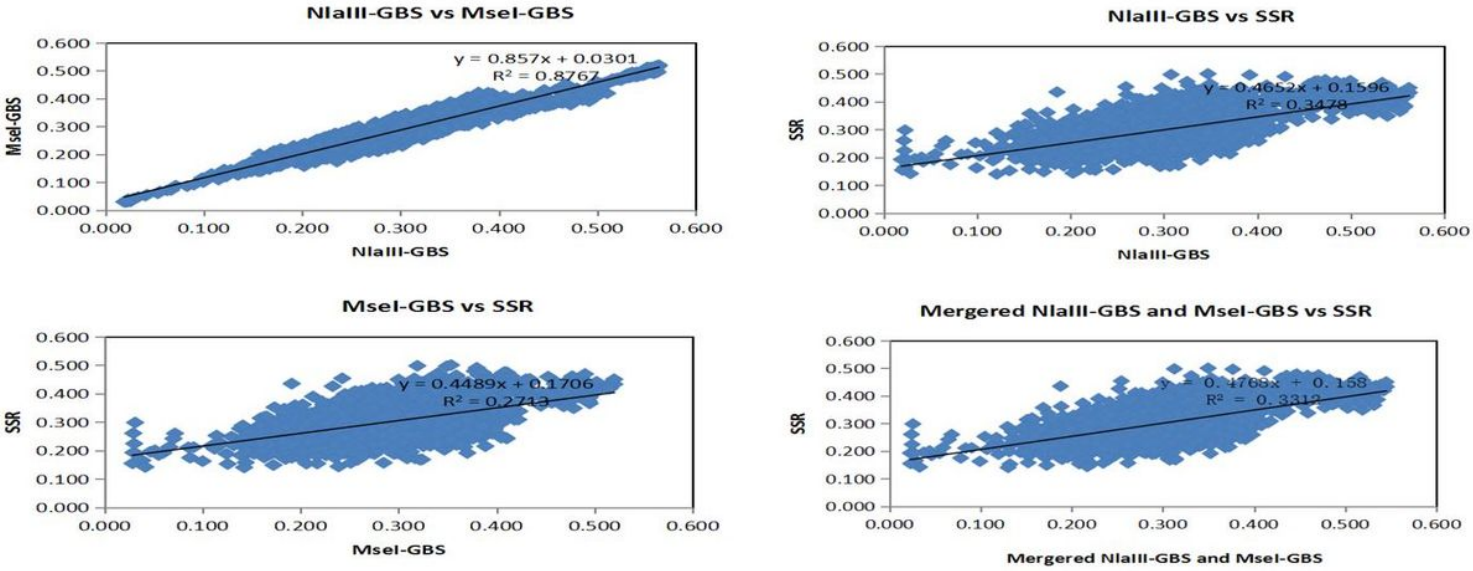

Figure 6

The correlation between the genetic distance matrices by different genetics markers