Determination of minimum number of random SNP for accurate population classification in rice (*Oryza sativa* L.)

Shan Wang
China National Rice Research Institute
Junhua Ye
China National Rice Research Institute
Qun Xu
China National Rice Research Institute
Xin Xu
China National Rice Research Institute
Yingying Yang
China National Rice Research Institute
Mengchen Zhang
China National Rice Research Institute
Yue Feng
China National Rice Research Institute
Xiaoping Yuan
China National Rice Research Institute
Hanyong Yu
China National Rice Research Institute
Yiping Wang
China National Rice Research Institute
Xinghua Wei
China National Rice Research Institute
Yaolong Yang (yangxiao182@126.com)
China National Rice Research Institute https://orcid.org/0000-0003-4409-120X

Original article

Keywords: SNP number, rice classification, population structure, genetic diversity

Posted Date: June 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-35906/v1

Abstract

Background: Classification of germplasm collections is of great importance for both the conservation and utilization of genetic resources. Thus, it is necessary to estimate and classify rice varieties in order to utilize these germplasms more efficiently for rice breeding. However, molecular classification of large germplasm collections can be costly and labor-intensive. Development of an informative panel of a few markers would allow for rapid and cost-effective assignment of crops to genetic sub-populations.

Results: Here, the minimum number of random SNP for rice classification (MNRSRC) was studied using a panel of 51 rice varieties belonging to different sub-groups. Through the genetic structure analysis, the rice panel can be obviously divided into five subgroups. The estimation of the MNRSRC was performed using SNP random sampling method based on genetic diversity and population structure analysis. In the genetic diversity analysis, statistical analysis of the coefficient of variation (CV) was performed for MNRSRC estimation, and we found that CV variation tended to plateau when the number of SNP was around 200, which was verified by the both cross-validation error of K value and correlation analysis of genetic distance. When the number of SNPs was greater than 200, the distribution of cross-validation error value tended to be similar, and correlation coefficients, almost greater than 0.95, exhibited small range of variation. In addition, we found that MNRSRC might not be affected by the number of varieties and the type of varieties.

Conclusion: The estimation of the MNRSRC was performed using SNP random sampling method based on genetic diversity and population structure analysis. The results demonstrated that at least about 200 random filtered SNP loci were required for classification in a rice panel. In addition, we also found that MNRSRC might not be affected by the number of varieties and the type of varieties. The study on MNRSRC in this study can provide a reference and theoretical basis for classification of different types of rice panels.

Introduction

Rice is a major crop in the world and feeds more than half of the world's population (Khush, 1997). Large germplasm collections provide abundant genetic resources for modern breeding program. Since the 1950s, yield and total production have increased significantly due to the release and utilization of new varieties, such as dwarf germplasm and male-sterile resources. And a few breeding lines have played a particularly key role in the improvement of rice varieties and are designed as cornerstone breeding parents. So far, under both natural and artificial selection, more than 120,000 distinct rice varieties have been recognized worldwide, which exhibits high levels of morphological and genetic diversity (Sang and Ge, 2007; Vaughan et al., 2008). Faced with such a great number of rice varieties, taxonomic classification of germplasm collections is of great importance for both the conservation and potential utilization of genetic resources collected in genebanks. (Wang et al. 2014). And clustering analysis of crop varieties or inbred lines is an essential method for studying genetic relationships of crop germplasms.

There is no doubt that *indica* and *japonica* are the two sub-species in Asian cultivated rice (Izawa 2008; Vaughan et al. 2008). However, the sub-group division of the sub-species is different according to many distinct investigations. Phenotypic markers, biochemical markers and heterosis indices were previously applied to germplasm cluster analysis. In a landmark study using 15 polymorphic enzyme loci on 1688 landraces, Glaszmann (1987) identified six varietal groups (I to VI), with two of the largest groups (groups I and IV) corresponding to typical Indica and Japonica varieties (subspecies) (Khush, 1997). By screening a sample of 234 rice accessions using 169 nuclear simple sequence repeats (SSRs) and two chloroplast loci, Garris et al. (2005) detected five distinct groups and referred to them as *indica*, *aus*, *aromatic*, *temperate japonica* and *tropical japonica*, which was subsequently supported by analyses of genome-wide single-nucleotide polymorphism (SNP) data (Caicedo et al., 2007; Zhao et al., 2010; Huang et al., 2012). Therefore, it is indicated that different number and types of markers and diverse rice accession panels contribute to these variable classification results.

However, molecular classification of large germplasm collections can be costly and labor-intensive. Development of an informative panel of a few markers would allow for rapid and cost-effective assignment of crops to genetic sub-populations and would facilitate breeding efforts to utilize allelic diversity within a sub-species without concern for cross incompatibility. Yuan et al. (2015) analyzed genetic variation of 69 rice varieties by using 120 SSR makers and found that the minimum of SSR markers for analyzing genetic diversity and population structure of rice was 72 and 60, respectively. Correlation analysis and t test were performed as major method to determine the minimum number of SSR markers. Agrama et al. (2012) selected 14 SSR markers with high discriminatory which was effective in assigning germplasm accessions to any five sub-populations. The study of minimum of markers always applied to the determination of essential derived variety (EDV) in maize. Kahler et al. (2010) reported the selection and evaluation of a panel of 285 SSR loci to help determine essential derivation in maize in the United States. A similar study conducted by the French Maize Breeders Association resulted in the publication of a set of 163 SSRs that were recommended for use to help determine essential derived variety status in maize germplasm in France (Andreau et al., 2003; Heckenberger et al., 2003; Kahler et al., 2010). In wheat, You et al. (2003) suggested that 73 loci with good polymorphism were needed to reflect genetic relationships among common wheat varieties from the 10 wheat growing regions of China with more than 90% certainty.

With the development of high-throughput sequencing technologies and bioinformatics, single nucleotide polymorphism (SNP) marker was rapidly used in population structure and genetic diversity instead of SSR (Rousselle et al., 2015; Xu et al., 2016). In addition, Although SNP exhibits lower polymorphism than SSR, due to its huge number, low cost and widely distribution on the whole genome, more and more researchers prefer to use SNP as the molecular marker recently (Lin et al. 2020; Yang et al. 2020). In this study, 51 rice varieties belonging to different sub-groups were sequenced. Based on genetic diversity and population structure analysis, minimum number of random SNP marker for rice classification (MNRSRC) was determined in order to make the molecular classification for large number of rice germplasm more efficient.

Results

Genetic structure and subdivision of rice panel

A rice panel consisted of 51 varieties were used in order to analyze the population structure (Additional file 1: Table S1). A model-based manner was used to estimate the genetic component of each variety. The K value was increased from 1 to 9 for cross-validation, and the standard error of the cross-validation estimate for each K was obtained. And we found that when K=5, the cross-validation error displayed the lowest value, suggesting K = 5 is a sensible modeling choice (Figure 1a). Thus, these varieties can be divided into five subgroups. According to the Q matrix, two major subgroups corresponding to two rice subspecies (*indica* and *japonica*) are apparent at K = 2, whereas five well-clustered subgroups was displayed at K = 5, which was consistent with the original information of every variety (Additional file 1: Table S1). It was found that 3 and 2 subgroups were under *japonica* and *indica*, respectively, that is, *temperate japonica*, *tropical japonica*, *aromatic*, *aus*, and *indica* (Figure 1b).

Similarly, the principle components analysis (PCA) and neighbor-join (NJ) tree also presented the same results. In the PCA, PC1 and PC2 totally explained more than 43% genetic variation, suggesting a strong genetic structure of our rice panel. The first two eigenvectors clearly separate the population into five subgroups. The PC2 and PC3 also provide the same result (Figure 1c). Additionally, a phylogenetic tree was constructed based on genetic distance with clearly five clusters. Totally, all these results indicated that our rice panel can be obviously divided into five subgroups (Figure 1d).

Estimation of the minimum SNP number by genetic diversity

We divided the population into five subgroups by using huge number of SNPs. However, what is the minimum number of SNP makers required for rice classification? Here, we used the different number of random SNP markers to study. After filtering, a SNP marker set containing about 119,568 SNPs was obtained to classification, and it was named as “original SNP-set”. We calculated π value of every SNP loci, which can reflect the genetic diversity of the rice panel. The average of π value is 0.3 in original SNP-set. A series of different numbers of SNP sets from 20 to 1000 were selected from original SNP-set to estimate the appropriate number of SNPs with 1000-time repetition, and the average π value of all SNP subsets were calculated. The π values of each SNP number fluctuated around 0.3 (Figure 2a). Moreover, with the increase of SNP number, the variation of π value became smaller. Coefficient of variation (CV) of each SNP number analysis revealed that CV value was sharply decreased when the SNP number was less than 200, and slowly decreased when SNP number was greater than 200 (Figure 2b). Furthermore, we used the statistical analysis to verify the inflection point of the CV value. We compared every five continuous SNP numbers with next second five continuous SNP numbers, and found that before about 200 SNPs, there was a significant difference between them ($p<0.05$) (Figure 2c). Therefore, we considered that 200 SNPs was the MNRSRC.

Estimation of the minimum SNP number by population structure

We can divide the rice panel into several subgroups referring to the lowest cross-validation error value. We estimated the standard error of the cross-validation of increased K value from 1 to 9 based on 50, 100, 150, 200, 250, 300, 350, 400, and 450 random SNP subsets selected from the original SNP-set with 100 repeats (Figure 3). When the SNP number was 50, it showed that the values were completely scattered. But with the increase of SNP number, the cross-validation error of every K became more and more concentrated. When the SNP number reached 200, the values' distribution became similar (Figure 3). Therefore, to get the accurate classification, the best MNRSRC should be greater than 200.

Estimation of minimum SNP number by correlation analysis

Genetic distance reflected the relationship between two varieties, which played the key role in rice classification. The correlation coefficients between genetic distance matrices based on original SNP-set and different number of random SNP subsets were calculated (Figure 4). The results suggested that the correlation coefficients between SNP original SNP-set and low number of random SNP subset showed large range of variation. When the number was larger than 200, the correlation coefficients exhibited small range of variation, and almost all values were larger than 0.95 (Figure 4). Thus, using the subset with more than 200 SNPs can acquire the similar genetic relationship to that obtained by using original SNP-set among the rice panel.

Minimum SNP number required for different types of rice panel

First, to study the effect of variety number on the minimum SNP number, a series of variety sets which consisted of 10, 20, 30, and 40 varieties were used (Table 1). In order to eliminate the influence of variety type, each variety set contained an equal number of the five variety types. The results showed that with the increase of the variety number, the minimum number of SNP didn't change much and remained about 200. However, the number of filtered SNP was positive correlated with the number of varieties. Therefore, variety number had no effect on the minimum number evaluation of SNP. Then, we also wondered whether the variety type affected the MNSRC. A total of 15 different rice panels including two, three, and four types of variety combinations were used for calculating the MNSRC (Table 2). Interestingly, all sets presented a similar result that the MNSRC ranged from 180 to 210. Therefore, MNRSRC might not be affected by the number of varieties and the types of varieties.

Discussion

Rice classification lays the foundation for future utilization of rice germplasm. Many previous studies have shown that Asian rice can be divided into five types (Garris et al. 2005; Zhao et al. 2018). Although it was also said that Asian rice can be divided into six types including *temperate japonica*, *tropical japonica*, *aromatic*, *rayada*, *aus*, and *indica* (Wang et al. 2014). In our study, we have not enough number of *rayada* rice, and thus only other five type of rice were selected. Due to the small number of *rayada*, the five types of rice varieties in our study also represented most Asian rice variety types.

Coefficient of variation (CV) can reflect the degree of data dispersion. In general, the greater CV is, the greater data dispersion degree will be. Otherwise, the data dispersion degree will be smaller. However, CV is not only influenced by the degree of data dispersion, but also by the average level of data values. Therefore, it is difficult to judge whether the data has reached a plateau according to the value of CV. In our study, as the number of SNPs increases, CV value will not remain stable, but become smaller and smaller. Many previous studies have demonstrated that reaching a plateau just depends on their own judgement through the curve graphs (Wang et al. 2003; Yuan et al. 2015; Zhang et al. 2015). In this study, there are two aspects different from the previous studies in obtaining the minimum number. Firstly, the huge volume of data was used here. Random SNPs were selected 1,000 repetitions for every number of SNPs from 20 to 1,000, which can be better to reflect the true results. Statistical analysis was used to identify whether the SNP number had reached a plateau. It was judged that the plateau period was reached when the CV values of any 5 SNP numbers were not significantly different from the CV values of the next second 5 SNP numbers. So, it made our results more convincing. Secondly, we used three methods to obtain the minimum SNP number, including CV analysis of genetic diversity parameter, population structure and correlation analysis of genetic distance, all of which were also used to verify the result of minimum SNP number. Similarly, different SNP number with lots of repetitions was selected in our study. According to the previous research, when the correlation coefficient of genetic distance matrix is greater than 0.9, it always implied that the genetic distance matrix is very consistent with the original one (Mantel 1967). In our study, when the SNP number reached 200, most of the repetitions' correlation coefficient was greater than 0.95. Furthermore, rice classification can be directly affected by population structure which is an important index for MNRSRC study.

The minimum SNP number used here is the random SNP. If the special SNPs were selected, fewer SNPs might be required. However, the fewer special SNP set has a deficiency that it can only be applied to specific or similar populations, but not to other different populations. In our study, we chose different number varieties to study and found that MNRSRC could not be influenced by variety number. Similarly, MNRSRC might not be affected by the type of varieties, either. According to the previous study, the genetic variation between *temperate japonica* and *indica* was larger than that between *temperate japonica* and *tropical japonica*. But, regardless of the number of subgroups and what subgroups were included in the population, the MNRSRC was about 200. It suggested that MNRSRC might not be affected by different genetic variation in the population. Maybe, it required a greater number of varieties for validation. The study on MNRSRC can provide reference and theoretical basis for the classification of different rice panels.

MNRSRC in our study was selected from the filtered SNP set, therefore, different population had different SNP original SNP-set. Due to the study of random markers, the minimum number is always larger than the number of special markers. Agrama et al. (2012) used a computer program WHICHLOCI to select the best combination SNP loci for population classification in rice, and suggested that only four SSR markers, including RM551, RM11, RM224, and RM44, could classify rice varieties to five subgroups with 99.4% accuracy. Nevertheless, because of the high workload and high cost, the classification with only 72 SSR markers were used as the accurate results. In fact, the 72 SSR markers may not represent the true

genetic relationship among the rice collections. In our study, about 100,000 filtered SNP loci were used as the genetic information of rice varieties, which ensured the accurateness in rice classification. It is common that SNP number is greater than SSR number, because SNPs are widely distributed in the genome, and SSR has higher polymorphism. An SSR marker may have more than ten alleles, while a SNP marker only have two alleles, which makes computer language-assisted research very convenient.

Additionally, in other crops, there are also some studies for minimum number of markers. In wheat, it was said that 73 loci with good polymorphism were needed to reflect genetic relationships among accessions with more than 90% certainty (You et al. 2004). Furthermore, correlation coefficients among random samples of alleles suggested that 350 to 400 alleles were needed to detect genetic relationships among common wheat varieties (Zhang et al. 2002).In soybean, at least 570 alleles (about 50 SSR loci) were required to reflect the genetic relationships of Chinese soybean cultivars. In maize, Wu et al. (2010) use 112 SSR primers and 97 maize inbred lines to determine the number of SSR alleles, and the regression equation to determine the number of SSR alleles was obtained. Nelson et al. (2011) evaluate the number of SNP required to measure genetic distance in maize and found that SNP markers were only two to three times as many as SSR markers, where the special SNP markers were selected. Among these previous studies of minimum number, the original numbers of markers were relatively small, and the markers were not random, so their applicability is limited. Moreover, correlation analysis was the main method to determine the minimum number of markers, which might affect the practicability and accuracy of selected marker. Therefore, the method in this study is more reasonable and the results is more accurate and applicable.

Conclusion

The estimation of the MNRSRC was performed using SNP random sampling method based on genetic diversity and population structure analysis. The results demonstrated that at least about 200 random filtered SNP loci were required for classification in a rice panel. In addition, we also found that MNRSRC might not be affected by the number of varieties and the type of varieties. The study on MNRSRC in this study can provide a reference and theoretical basis for classification of different types of rice panels.

Materials And Methods

Plant materials

A total of 51 Asian cultivated rice varieties were obtained from the National Mid-term Rice Genebank at the China National Rice Research Institute. Most of them were previously reported by Wang et al. (2014). Of them, 21 were labelled as *Indica* and 30 as *Japonica*. According to the previous report, these varieties included 11 *indica*, 10 *aus*, 11 *aromatic*, 8 *tropical japonica* and 11 *temperate japonica*.

Sequencing and genotyping

Total genomic DNA was extracted for next-generation sequencing. Dried rice seeds were soaked in water for about 2 days and allowed to germinate. Rice shoot were sampled and grinded into powder in liquid nitrogen. About 0.2 g powder was used for DNA extract using DNeasy Plant Mini Kit (Qiagen). DNA extracts were quantified by fluorescence using Qubit 4.0 (Invitrogen, USA) and integrity of DNA was checked by 1.5% agarose gel electrophoresis. The DNA was finally dissolved in sterile distilled water and stored at -20ºC for further DNA sequencing library construction. The sequencing library was qualified by Agilent 2100 high sensitivity DNA reagents. The genomic DNA was sequenced using next-generation sequencing technology on the Illumina platform for about 20 × coverage of rice genome and generating 150-bp paired-end reads. All paired-end reads were aligned and mapped using BWA software against the reference the reference genome sequence (Os-Nipponbare-Reference-IRGSP-1.0) (Li and Durbin 2009). SNP calling were conducted using the GATK pipeline as previously described (McKenna et al. 2010).

Filtering process for the SNP calling and analysis

Before the analysis of sequencing data, filtering process should be accomplished to guarantee accuracy. There are some aspects of filtering criteria are as follow: filtering out of low-quality variants; Select the common variants from the results between samtools and GATK pipeline to improve the accuracy; filtering out of the SNPs for which all varieties present one allele or more than two alleles; selecting variants when the sequencing depth > 10; filtering out of the SNPs for which the missing data ratio > 10%; selecting the SNPs for which the frequency of minor allele (MAF) >0.05; filtering out of the redundant SNPs by the threshold of linkage disequilibrium coefficient r^2>0.1; SNP number analysis was then applied to reveal the MNRSRC.

Genetic diversity and population structure

Genetic diversity indices including nucleotide diversity (π) were calculated by using vcftools using the parameter of “site-pi” (http://vcftools.sourceforge.net/). EIGENSOFT was used to conduct a principal component analysis (PCA) to estimate the number of subpopulations (Patterson et al. 2006). ADMIXTURE software was used to calculate the genetic component for each variety (Alexander et al. 2009) A phylogenetic tree was constructed using FastTree based on the SNP information of 51 rice varieties (Price et al. 2009), and visualized using the online tool iTOL (https://itol.embl.de/) (Letunic et al. 2016).

Determination of minimum number of random SNP

The filtered SNP panel was used to analyze the genetic diversity, population structure, and the results were defined as the actual value. Then different number of SNPs was selected from the filtered SNP panel to form diverse subsets, and each number of SNP subset was randomly selected 100-1000 times. Through the analysis of genetic diversity and population structure by using SNP subsets, the random value can be obtained. The MNRSRC can be determined by analyzing the difference between the actual values and the random values of different numbers of SNP subsets. Statistical analysis including t-test

and one-way analysis of variance was performed on Microsoft Office Excel 2010 and SAS 9.2 (SAS, Inc., Cary, North Carolina, USA), respectively.

Abbreviations

SNP: single nucleotide polymorphism; MNRSRC: minimum number of random SNP for rice classification; SSR: simple sequence repeats; PCA: principle components analysis; CV: coefficient of variation.

Declarations

Ethical Approval and Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request

Competing interest

The authors have declared that no competing interests exist.

Funding

This research was funded by Zhejiang Province Natural Science Foundation of China (LQ19C130006), Special Program for Breeding of Zhejiang Province (2016C02050-6-1) and Chinese Academy of Agricultural Sciences (Grant No. CAAS-ASTIP-201X-CNRRI).

Authors' contributions

YY (Yaolong Yang) AND XW conceived and designed research. YY (Yaolong Yang), SW, JY, YY (Yingying Yang), and QX conducted sequencing experiments. SW, QX, YF, HY and XY analyzed the data. SW and JY wrote the manuscript. WY, XW and YY (Yaolong Yang) helped to revise the manuscript. All the authors read and approved the final manuscript.

Acknowledgements

The authors would thank the International Rice Research Institute (Los Banos, Philippines) for providing most of the seed samples.

References

Agrama HA, McClung AM, Yan WG (2012) Using minimum DNA marker loci for accurate population classification in rice (*Oryza sativa* L.). Mol Breeding 29:413-425 doi:10.1007/s11032-011-9558-x

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome research 19:1655-1664 doi:10.1101/gr.094052.109

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, Bustamante CD, Purugganan MD (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. Plos Genet 3:1745-1756 doi:10.1371/journal.pgen.0030163

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631-1638 doi:10.1534/genetics.104.035642

Glaszmann JC (1987) Isozymes and classification of Asian rice varieties. TAG Theoretical and applied genetics Theoretische und angewandte Genetik 74:21-30 doi:10.1007/BF00290078

Huang XH, Zhao Y, Wei XH et al. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32-U53 doi:10.1038/ng.1018

Izawa T (2008) The Process of Rice Domestication: A New Model Based on Recent Data. Rice 1:127-134 doi:10.1007/s12284-008-9014-7

Kahler AL, Kahler JL, Thompson SA, Ferriss RS, Jones ES, Nelson BK, Mikel MA, Smith S (2010) North American Study on Essential Derivation in Maize: II. Selection and Evaluation of a Panel of Simple Sequence Repeat Loci. Crop Sci 50:486-503 doi:10.2135/cropsci2009.03.0121

Khush GS (1997) Origin, dispersal, cultivation and variation of rice. Plant Mol Biol 35:25-34 doi:10.1023/A:1005810616885

Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242-W245 doi:10.1093/nar/gkw290

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760 doi:10.1093/bioinformatics/btp324

Lin ZC, Qin P, Zhang XW et al. (2020) Divergent selection and genetic introgression shape the genome landscape of heterosis in hybrid rice. P Natl Acad Sci USA 117:4623-4631 doi:10.1073/pnas.1919086117

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-

generation DNA sequencing data. Genome research 20:1297-1303 doi:10.1101/gr.107524.110

Nelson BK, Kahler AL, Kahler JL, Mikel MA, Thompson SA, Ferriss RS, Smith S, Jones ES (2011) Evaluation of the Numbers of Single Nucleotide Polymorphisms Required to Measure Genetic Distance in Maize (Zea mays L.). Crop Sci 51:1470-1480 doi:10.2135/cropsci2010.07.0401

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. Plos Genet 2:e190 doi:10.1371/journal.pgen.0020190

Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Mol Biol Evol 26:1641-1650 doi:10.1093/molbev/msp077

Rousselle Y, Jones E, Charcosset A et al. (2015) Study on Essential Derivation in Maize: III. Selection and Evaluation of a Panel of Single Nucleotide Polymorphism Loci for Use in European and North American Germplasm. Crop Sci 55:1170-1180 doi:10.2135/cropsci2014.09.0627

Sang T, Ge S (2007) Genetics and phylogenetics of rice domestication. Current opinion in genetics & development 17:533-538 doi:10.1016/j.gde.2007.09.005

Vaughan DA, Lu BR, Tomooka N (2008) Was Asian Rice (Oryza sativa) Domesticated More Than Once? Rice 1:16-24 doi:10.1007/s12284-008-9000-0

Wang B, Chang RZ, Tao L, Guang RX, Yan L, Zhang MH, Feng ZF, Qiu LJ (2003) Identification of SSR Primer Numbers for Analyzing Genetic Diversity of Chinese Soybean Cultivated Soybean. Molecular Plant Breeding 1:82-88

Wang CH, Zheng XM, Xu Q, Yuan XP, Huang L, Zhou HF, Wei XH, Ge S (2014) Genetic diversity and classification of Oryza sativa with emphasis on Chinese rice germplasm. Heredity 112:489-496 doi:10.1038/hdy.2013.130

Wu CL, Li SF, Dong BX, Zhang QQ, Zhang CQ (2010) Determination of the Number of SSR Alleles Necessary for the Analysis of Genetic Relationships Between Maize Inbred Lines. Agr Sci China 9:1713-1725 doi:10.1016/S1671-2927(09)60270-4

Xu Q, Yuan XP, Wang S, Feng Y, Yu HY, Wang YP, Yang YL, Wei XH, Li XM (2016) The genetic diversity and structure of indica rice in China as detected by single nucleotide polymorphism analysis. BMC Genet 17 doi:10.1186/s12863-016-0361-x

Yang YL, Xu X, Zhang MC, Xu Q, Feng Y, Yuan XP, Yu HY, Wang YP, Wei XH (2020) Genetic Basis Dissection for Eating and Cooking Qualities of *Japonica* Rice in Northeast China. Agronomy-Basel 10 doi:10.3390/Agronomy10030423

You GX, Zhang XY, Wang LF (2004) An estimation of the minimum number of SSR loci needed to reveal genetic relationships in wheat varieties: Information from 96 random accessions with maximized genetic

diversity. Mol Breeding 14:397-406 doi:10.1007/s11032-004-0285-4

Yuan XP, Wang CH, Deng HZ, Xu Q, Feng Y, Yu HY, Wang YP, Wei XH (2015) Minimum of SSR Markers for Analyzing Genetic Variation of *Oryza sativa* L. Chin J Rice Sci 29:578-586

Zhang QP, Liu WS, Liu S, Wer X, Liu N (2015) An estimation of SSR alleles numbers for analyzing genetic diversity of apricot. Journal of Fruit Science 32:186-191

Zhang XY, Li CW, Wang LF, Wang HM, You GX, Dong YS (2002) An estimation of the minimum number of SSR alleles needed to reveal genetic relationships in wheat varieties. I. Information from large-scale planted varieties and cornerstone breeding parents in Chinese wheat improvement and production. Theoretical and Applied Genetics 106:112-117 doi:10.1007/s00122-002-1016-z

Zhao KY, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung CW, Reynolds A, Bustamante CD, McCouch SR (2010) Genomic Diversity and Introgression in O. sativa Reveal the Impact of Domestication and Breeding on the Rice Genome. Plos One 5 doi:10.1371/journal.pone.0010780

Zhao Q, Feng Q, Lu HY et al. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice (vol 50, pg 278, 2018) Nat Genet 50:1196-1196 doi:10.1038/s41588-018-0136-6

Tables

Table 1 MNRSRC required for different number of varieties in rice panel

NO. of varieties	NO. of filtered SNP	MNRSRC			F-test	P-value
		mean±SD	min	max		
10	96173	190.00±1.95	170	210	1.1157	0.3647
20	110806	186.25±10.60	170	200		
30	116547	196.25±10.60	180	210		
40	118821	190.00±7.56	180	200		
51	119568	187.50±10.35	170	200		

Note: F-test was performed using one-way analysis of variance statistical method; the MNRSRC were obtained from different number of varieties with 8 repetitions.

Table 2 MNRSRC required for different variety types in rice panel

rice panel	NO. of filtered SNP	MNRSRC			F-test	P-value
		mean±SD	min	max		
tej-trj	77641	190.00±13.09	170	210	0.8080	0.6588
tej-aro	83896	193.75±10.60	180	210		
tej-aus	95069	183.75±11.88	170	200		
tej-ind	102794	186.25±16.85	170	210		
aus-ind	88320	181.25±12.46	170	200		
tej-trj-aro	100188	183.75±9.16	170	190		
tej-aro-aus	101455	187.50±13.89	170	210		
tej-aus-ind	109405	188.75±16.42	170	210		
trj-aro-aus	104757	188.75±14.58	170	210		
trj-aro-ind	110540	183.75±14.08	180	210		
tej-aro-aus-ind	111630	192.50±14.88	170	210		
tej-trj-aro-aus	111672	181.25±11.26	170	200		
trj-aro-aus-ind	113866	192.50±10.35	180	210		
tej-trj-aus-ind	118632	182.50±8.86	170	190		
tej-trj-aro-ind	116922	187.50±13.89	170	210		

Note: F-test was performed using one-way analysis of variance statistical method; tej: *temperate japonica*; trj: *tropical japonica*; aro: *aromatic*; aus: *aus*; ind: *indica*; the MNRSRC were obtained from different type of rice panels with 8 repetitions. The name of rice panel represents the classification of the rice panel. “tej-trj” represents the rice panel can be divided into *temperate japonica* and *tropical japonica* subgroups.

Figures

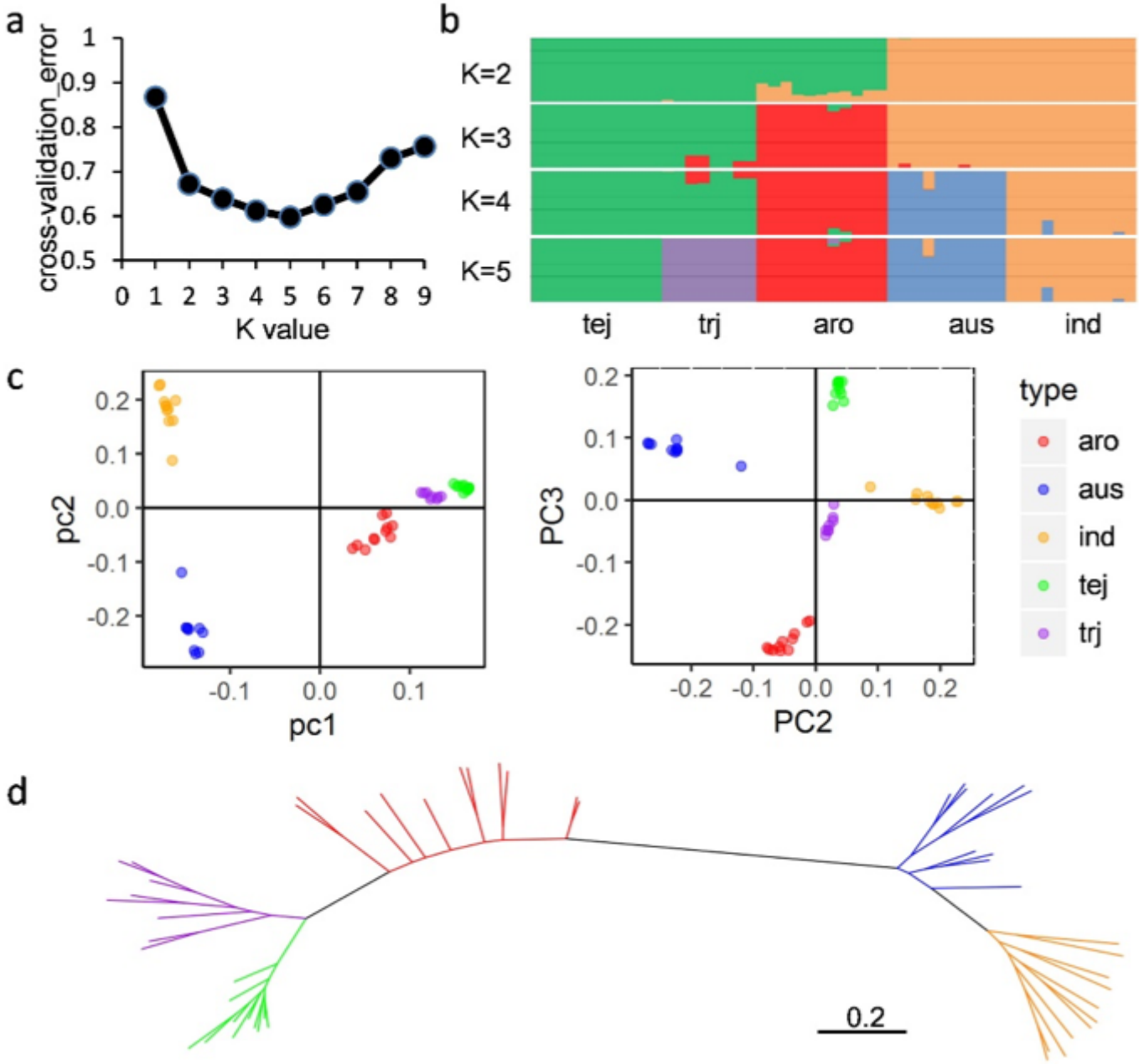

Figure 1

Population structure of 51 rice varieties. (a) the cross-validation error estimation of every K value; (b) Model-based population assignment of the rice panel; (c) Principal component analysis; (d) phylogenetic tree based on genetic distance.

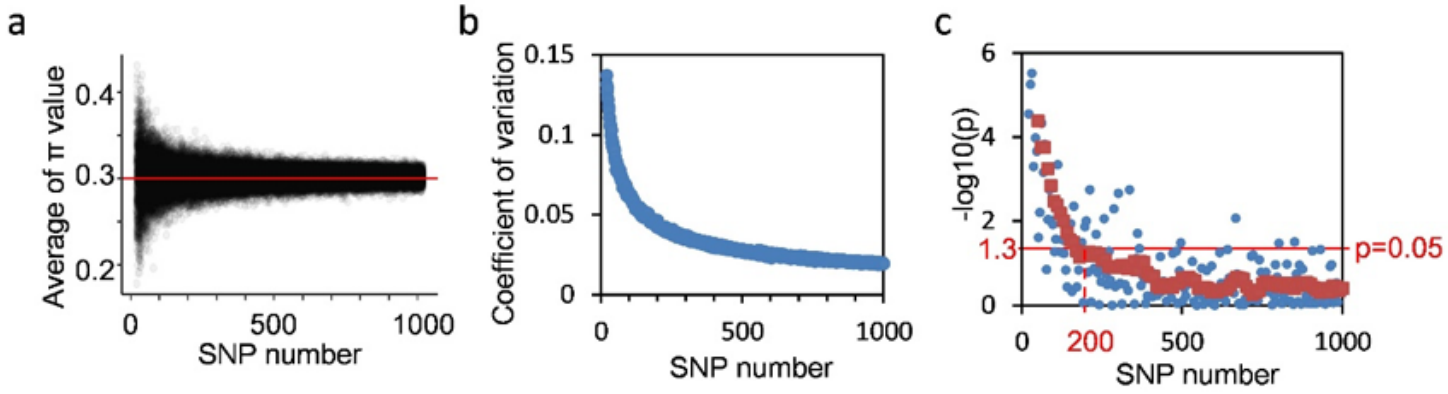

Figure 2

Estimation of MNRSRC by genetic diversity. (a) scatter diagram of average of π value of different number of SNP with 1000 repetitions. (b) coefficient of variation estimation of different SNP number; (c) statistical analysis for determining the MNRSRC. P value was calculated from the t-test of every five continuous SNP numbers with next second five continuous SNP numbers. Red line indicates the threshold of p value (p = 0.05), blue point indicates -log10(p) value, and red square indicates the average -log10(p) value of each ten SNP numbers.

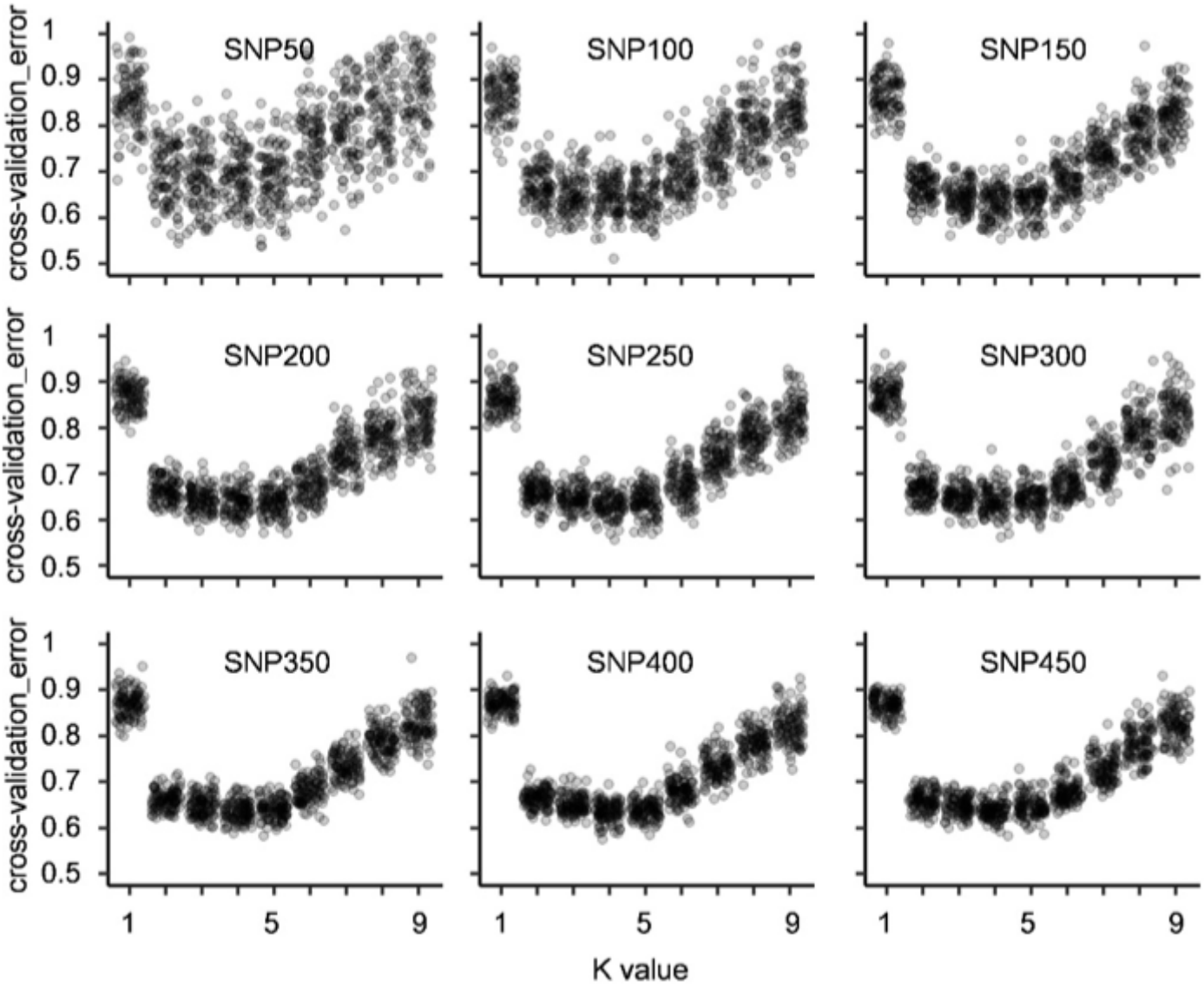

Figure 3

Cross-validation error analysis for different number of SNP. Standard error of the cross-validation of increased K value from 1 to 9 were estimated based on 50, 100, 150, 200, 250, 300, 350, 400, and 450 random SNP subsets selected from the original SNP-set with 100 repeats.

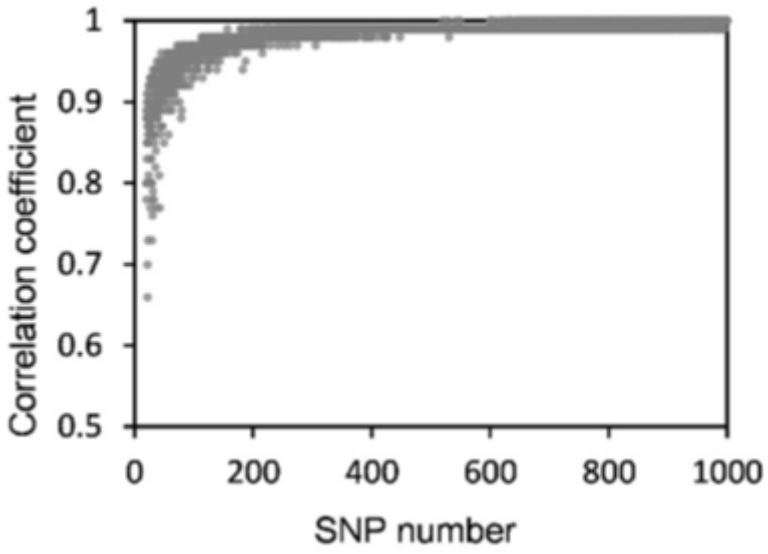

Figure 4

Correlation coefficients between genetic distance matrices based on original SNP-set and different number of random SNP subset. The SNP subset was selected from 20 to 1000 SNP number with the interval of 10 SNP number, and every SNP subset selected 100 repetitions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.xlsx