RESEARCH

Population Structure of Nation-wide Rice in Thailand

Phanchita Vejchasarn[1], Jeremy R. Shearman[2], Usawadee Chaiprom[3], Yotwarit Phansenee[1], Tatpong Tulyananda[4], Jirapong Jairin[1] and Chainarong Amornbunchornvej[5*]

Abstract

Background:: Thailand is a country with large diversity in rice varieties due to its rich and diverse ecology. In this paper, 300 rice varieties from all across Thailand were sequenced to identify SNP variants allowing for the population-structure to be explored.

Results:: The result of inferred population structure from admixture and clustering analysis illustrated strong evidence of substructure in each geographical region. The results of phylogenetic tree, PCA analysis, and machine learning on SNPs selected by QTL analysis also supported the inferred population structure.

Conclusion:: The population structure, which was inferred in this study, contains five populations s.t. each population has a unique ecological system, genetic pattern, as well as agronomic traits. This study can serve as a reference point of the nation-wide population structure for supporting breeders and researchers who are interested in Thai rice.

Keywords: Admixture; Oryza sativa; SNPs

*Correspondence:
chainarong.amo@nectec.or.th
[5]National Electronics and Computer Technology Center (NECTEC), 112 Phahonyothin Road, Khlong Nueng, Khlong Luang District, 12120 Pathum Thani, Thailand
Full list of author information is available at the end of the article

Background

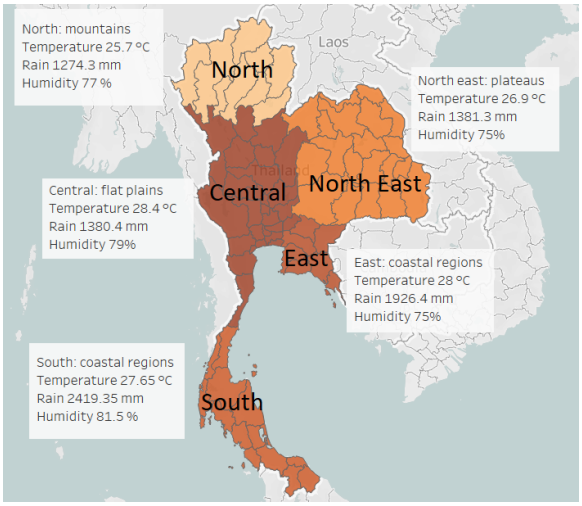

Figure 1 The climate of Thailand in the aspect of average temperature, amount of rain, and humidity in 2018 separated by regions [1].

Rice (*Oryza sativa*) has been the main carbohydrate source in Thailand for more than 4,000 years [2], and Thailand has been a major rice exporter since 1851 [3]. Accelerated cultivar selection for specific environments is important for rice breeding

programs. The long time period of rice domestication has yielded many rice cultivars with wide variation in size, flowering time, grain quality, and yield to name a few.

Thailand has large diversity in ecological systems [4]. In the north, most of the area is covered by mountains and tropical rain forests. In central Thailand, the region consists of plains and fields that are prone to flood. In the north-eastern part, plateaus are the main type of area. In the south are tropical coastal regions and tropical islands. See Figure 1 for more details.

Due to the diverse ecology in Thailand, rice varieties need to be adapted to their intended growth region and there is some degree of association between genetic variation and geographical origin of Thai rice [5]. Moreover, there is a higher level of diversity in Thai rice accessions compared to International Rice Research Institute (IRRI) germplasm [4]. Upland Thai rice forms a cluster of tropical japonica [6, 4, 7], while lowland rice forms Indica clusters.

Understanding population structure and genetic diversity is an important step before Genome-wide association studies (GWAS) [8], which paves the way for studies of traits and functional gene investigation. Studies in population structure and genetic diversity of Thai rice has been conducted using different sets of rice accessions and molecular markers. Comparison of genetic diversity among 43 Thai rice and 57 IRRI rice accessions was investigated in [4], using single-stranded conformation polymorphism (SSCP) indels markers. Additionally, [9] used 12 simple sequence repeat (SSR) markers to examine ongoing gene flow among three types of rice samples in Thailand, including 42 wild rice populations, 12 weedy rice populations, and 37 cultivated rice varieties. Recently, with a greater number of rice germplasm accessibility, 144 Thai and 23 exotic rice accessions were included to evaluate genetic diversity using SSR markers in [6]. Another study assessed the population gene pool of 15 Thai elite rice cultivars using InDel markers ([10]). It is worth to note that there are some limitations in these previous works regarding the access to a high number of varieties for each region of Thailand. Additionally, genome-wide SNP markers were not used in these previous works.

To fill the gap in the literature, in this study, we mainly focused on the population structure of 300 rice varieties from all over Thailand, which are grown in rich and diverse ecological systems. We use both InDel and SNP markers to infer subpopulations. These 300 varieties are a good representation of the nation-wide rice population structure.

Results

Population Structure

After clustering the 300 samples, five populations were found in the dataset. These five inferred populations generally group according to geological areas of rice sample cultivation. POP1 represents Indica samples from Central Thailand. POP3 represents Indica samples from Northeastern Thailand. POP2 represents rice samples from both Northeastern and Central Thailand. POP4 represents samples from Southern Thailand. And lastly, POP5 represents Japonica samples from Northern Thailand.

A principal component analysis showed that PC1 separated the Japonica population varieties (POP5) from the rest of the varieties, while PC2 separated the

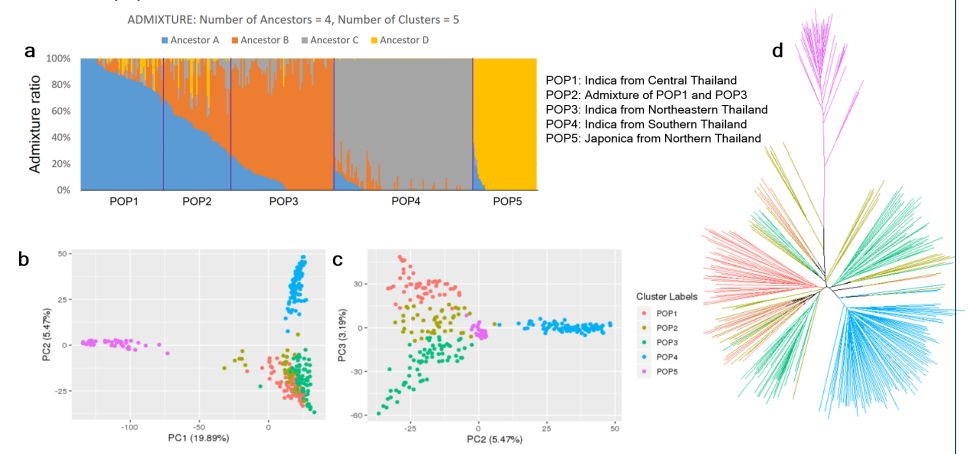

Figure 2 Population structure of 300 Thai rice varieties inferred from 69,777 SNPs and 47,277 Indels. **a** Admixture plot of Thai rice varieties. The vertical axis represents an ancestry ratio of each variety. The horizontal axis represents individual varieties grouped by clustering analysis. Groups were assigned by clustering analysis on individual-admixture ratios. There are four ancestors with five populations inferred by clustering analysis. **b** The first and second principal components (PCs) from a principal component analysis. **c** The second and third PCs. Cluster colors were assigned according to ADMIXTURE clustering analysis results. The PC1 separates the japonica varieties (POP5) from the indica varieties. PC2 separates southern indica varieties (POP4) from central and northern varieties (POP1, POP2, and POP3). Lastly, PC3 separates central indica (POP1), from northern indica (POP3) with their admixture varieties appearing in between the two (POP2). **d** Phylogenetic tree of the 300 varieties, created by NJ tree, color coded according to the ADMIXTURE result.

southern population varieties (POP4) from the other three populations central and northern varieties of Indica samples (Figure 2). Lastly, PC3 separated the central Indica varieties (POP1) from the northern Indica varieties (POP3) with the varieties identified as admixed (POP2) joining the two, showing that the geographical separation is reflected in the genotypes of each variety. A phylogenetic tree was constructed and showed that the Japonica population (POP5) was separated from the Indica populations (Figure 2 (d)). Admixed varieties (POP2) were distributed among central (POP1) and northern (POP3) branches, suggesting that POP2 is an admixed group of POP1 and POP3, while POP1, POP3, and POP4 were clearly separated from each other. Admixture analysis showed that POP1, POP3, POP4, and POP5 were grouped into different ancestors (different colors). POP2, however, had mixed ratios of ancestor A and B, which were the ancestors of POP1 and POP3. This indicates that POP2 is an admixed population of POP1 and POP3. POP1, POP3, POP4, and POP5 have high bootstrap support around 0.9, while POP2 has average support at 0.69 (Table 1). This is consistent with POP2 representing an admixed population of POP1 and POP3.

In the aspect of population genetic distance, the F_{ST} between admixture ancestor populations, which is a widely-used measure of genetic variation among populations [11], were reported in Table 2. The table shows that Ancestor D, which was the ancestor of the Japonica population (POP5) has a higher distance than was observed among other populations. Ancestors A and B were closer compared to C. While it is unclear whether POP4 was Indica or Japonica population, the

Table 1 Number of samples and support of clustering assignment from bootstrapping for each population. The support number represents the likelihood that each cluster has the same set of members. Higher support implies a higher chance that cluster members are in the same population.

	Number of samples	Average support
POP1	54	0.98
POP2	45	0.69
POP3	67	0.92
POP4	92	0.89
POP5	42	0.99

Table 2 F_{ST} divergences between ancestry populations inferred by ADMIXTURE. A is an ancestor of Indica (elite line), B is an ancestor of Indica (modern variety), and D is the ancestor of Japonica. By using a threshold of $F_{ST} \leq 0.3$ to consider populations to have a similar type: either Japonica or Indica, C was assigned to be an ancestor of Indica (landrace in southern part of Thailand).

F_{ST}	Ancestor A	Ancestor B	Ancestor C
Ancestor B	0.178	-	-
Ancestor C	0.208	0.209	-
Ancestor D	0.480	0.497	0.507

F_{ST} values suggest that the ancestor of POP4 (C) was closer to ancestors A and B (Indica) than the ancestor D (Japonica). This implies that ancestor C should be an Indica ancestor and that POP4 is an Indica population.

Agronomic traits of subpopulations

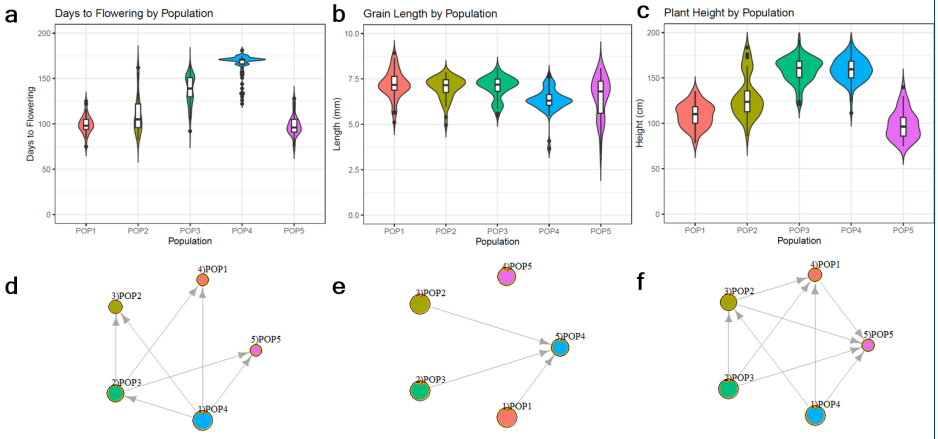

Figure 3 Subpopulation distributions of three phenotypes: days to flowering (a), grain length (b), and plant height (c). Domination graphs represent relationships between pairs of populations for days to flowering (d), grain length (e), and plant height (f). Arrow directions point from the population with a significantly higher phenotype value to the population with a lower phenotype value (with Mann Whitney test at $\alpha = 0.001$).

There are three agronomic traits that have been compared among subpopulations: days to flowering, grain length, and plant height. Figure 3 shows the details of these agronomic traits for each subpopulation. The distributions for traits are in the above figures (a-c), while the significance tests results are in the below figures (d-f). A significance test shows that whether one population has a trait significantly different from another.

For the days to flowering trait, central Indica varieties (POP1) flower earlier than north-eastern Indica varieties (POP3). The admixed population (POP2) has a flowering time roughly between that of POP1 and POP3, as expected. Southern Indica

Table 3 The result of 10-fold cross validation based on 268 SNPs for population classification using Random Forest algorithm

	Precision	Recall	F1
POP1	0.83	0.93	0.88
POP2	0.76	0.62	0.68
POP3	0.90	0.91	0.90
POP4	0.97	0.98	0.97
POP5	1.00	1.00	1.00

varieties (POP4) have the latest flowering time out of the 300 varieties investigated. Lastly, the Japonica varieties (POP5) had a similar flowering time as POP1 (Figure 3 a,d).

For the grain-length trait, POP1, POP2, and POP3 have similar grain length, while POP4 has a significantly shorter grain length compared to POP1, POP2, and POP3. POP5 has high variation of grain length. This indicates that Japonica (POP5) cannot be distinguished from Indica (POP1 - POP4) by using the grain-length trait (Figure 3 b,e).

For the plant-height trait, ordering by the ascending heights, the order is POP5, POP1, POP2, and POP3. POP3 and POP4 have no significantly different in the height trait (Figure 3 c,f).

Unique SNPs of subpopulations

A QTL analysis was used to identify SNPs with large variation in allele frequency between populations and 50-100 of the SNPs with the greatest allele frequency difference between populations were selected to train a random forest model to identify which population any given sample is from based on genotype. A total of 268 SNPs were selected (Supplementary Table 1).

Only POP5 had population specific SNPs that allowed for accurate population identification, this was not surprising as this population is Japonica and the other populations are all Indica (Table 3). The Indica populations had too much allele sharing to allow for each variety to be accurately assigned to their population. The admixed population had the lowest rate of correct population assignment, while the other populations were all in the 80-90% range (Table 3.)

While a QTL analysis to identify population specific SNPs might be unconventional, it is well known that population stratification can result in false positives. In this particular case the populations in question are not discrete populations, but rather groupings of varieties that tend to correlate with location and have genetic mixing between varieties.

The majority of SNPs most predictive for POP1 occurred on chromosome 1 in an interval between 21.6 and 22.5 Mb and an interval on chromosome 3 between 8.4 and 8.8 Mb. The majority of SNPs most predictive for POP2 occurred on chromosome 3 between 31 and 31.5 Mb with some small intervals on chromosomes 5, 6 and 7. There were 5 intervals of predictive SNPs for POP3 and several small intervals. Chromosome 3 had a interval from 27.59 to 27.65 Mb, chromosome 5 had an interval from 18.71 to 18.78 Mb, chromosome 6 had two intervals from 7.61 to 7.68 Mb and 11.02 to 11.06 Mb, chromosome 10 had an interval from 14.74 to 14.8 Mb. POP4 had the most distinctive allele frequencies with SNP intervals on chromosome 1 at 21.07 to 21.11 Mb, chromosome 2 at 5.32 to 5.35 Mb and 16.41

to 16.45 Mb, chromosome 5 at 23.71 to 23.84 Mb, and chromosome 11 at 2.7 to 2.8 Mb and 23.36 to 23.42. Of the 268 SNPs, there were 110 within 75 genes, although the majority of these are predicted genes with no known function (Supplementary Table 2). There were 259 genes within the intervals of these predictive SNPs and most were predicted genes of unknown function (Supplementary Table 3).

Discussion

According to the work in [4], upland Thai rice were grouped in Japonica cluster: other were clustered in Indica cluster, which are consistent with the population structure found in this work. In the aspect of agronomic traits, all inferred subpopulations posses unique traits that might suit to their growing environment since they were grown in the different ecological conditions; northern areas are upland, central areas are flat plain, north-eastern areas are plateaus, and south areas are coastal regions and tropical islands.

The inferred subpopulation in the north is a Japonica cluster (POP5). Other four inferred subpopulations are Indica clusters in the central area (POP1), north-east (POP3), south (POP4), and the admixture of POP1 and POP3 (POP2). All inferred subpopulations were different and separated well using 268 selected SNPs from QTL analysis on Random forest classifier except the admixture cluster (POP2). This implies that inferred subpopulations were unique.

An interesting finding was that the most predictive SNPs for each population occurred within a few small intervals, rather than randomly spread throughout the genome, which suggests a selection pressure, perhaps selecting for a trait that makes the variety better in the area it is grown. However, the population groupings are broad, each covering a quite diverse range of environments, and the allele frequencies between populations have a large amount of overlap, so many of these regions could be due to chance rather than function. Some interesting genes around 268 SNPs, for example, were Os03g0262000, Os06g0677800, and Os05g0203800. Os03g0262000, which is a homolog of AtPIP5K1 that is induced by water stress and abscisic acid in *A. thaliana* [12]. Os06g0677800 (OsARF17) is a target for viral infection [13]. Os05g0203800 (OSMADS58) plays a crucial role for flower development [14].

Conclusion

Thailand is a country with large diversity in rice varieties due to its rich and diverse ecology. In this paper, 300 rice varieties from all across Thailand were sequenced to identify SNP variants allowing for the population-structure to be explored.

The result of inferred population structure from admixture and clustering analysis illustrated strong evidence of substructure in each geographical region. The results of phylogenetic tree, PCA analysis, and machine learning on SNPs selected by QTL analysis also supported the inferred population structure. Moreover, by using only 268 SNPs, Random forest classifier was able to classify four out of five subpopulations except the admixture well. This indicates these subpopulations are unique enough to be distinguished by a small number of SNPs. A unique ecological system where rice is grown might play a key role in this uniqueness. The 268 SNPs can be served as a markers of these subpopulations for future study.

This study can serve as a reference point of the nation-wide population structure for supporting breeders and researchers who are interested in Thai rice.

Methods

Plant material

The list of 300 representative Thai rice varieties is at Supplementary Table 4. The Thai rice accessions were collected from all regions of Thailand: northern, north-eastern, southern, and central region. All plants were grown in the wet season of 2018 at Ubon Ratchathani Rice Research Center (URRC) of Ubonratchatani province, Thailand (15°19'55.2"N, 104°41'27.9"E).

Genotyping by sequencing and variance calling

The genotypic sequences were generated from Ion S5™ XL Sequencer (Thermo Fisher Scientific). The data were obtained as BAM files. The ApeKI enzyme was used for genomic DNA digestion to prepare the DNA library. In the sequencing step, E-Gel™ SizeSelect™ agarose gels (Invitrogen) were used to select DNA fragments for 250–300 bp. The Nipponbare reference genome by Ion Torrent™ Suite Software Alignment Plugin v5.2.2. was used for analyzing all sequencing data. The fastq files were created from BAM files using Samtools v1.9 [15]. Then, fastq files were re-aligned with the Japonica reference genome using Burrow–wheeler aligner (BWA) v0.7.17 [16] and SAMtools. Variants were called using using GATK v4.1.4.1 [17].

Population structure analysis

Numerical genotype function

Genotype was converted into a numerical value, such that homozygous reference allele was 1.0, homozygous alternate allele was 0.0, and heterozygous was 0.5 using TASSEL [18]. The SNPs were filtered to have a minimum allele frequency of 0.05 and a minimum call rate of 70% per SNP. The SNP number was reduced from 3,366,491 to 117,054 sites after filtering.

Admixture analysis

Numerical genotypes were used to create .ped, .map and .bed files for ADMIXTURE [19] analysis to estimate ancestry ratios of all individual samples. The optimal number of ancestors was found to be four by the Elbow method.

Clustering analysis

Their ancestry-ratio vectors of each SNP were used for data clustering. The individual assignments of clustering were inferred by applying a k-means clustering approach [20] in the R software package [21]. The Elbow method was applied to infer the optimal number of clusters based on Between-cluster and Total Sum-of-Square (BCTSS) Ratio. The BCTSS ratio represents a ratio of difference of distance from individuals to their cluster centroid between having current clustering assignment compared to having only one cluster. The optimal number k^* of clustering assignment should reduce BCTSS ratio significantly compared against $k^* - 1$ and $k^* + 1$ cases.

A 10,000 iteration bootstrap approach [22] was deployed to estimate the support of clustering assignment of each population. The clustering assignment that maximized BCTSS ratio with the optimal k along with the support of assignment from bootstrap were used to represent the subgroups of the population.

Principal components analysis

PCs were generated from numeric genotype data using TASSEL [18].

Phylogenetic tree construction

A phylogenetic tree was generated by Neighbor-Joining method [23] using the numerical genotype data in TASSEL [18].

Domination graphs inference

Domination graphs represents relationships between pairs of populations for three phenotypes were inferred using EDOIF package [24]. For each phenotype, nodes of domination graph are subpopulations while there is an edge from a population with a significantly higher phenotype value to a population with a lower phenotype value. The Mann Whitney test was deployed to infer edges of a domination graph with $\alpha = 0.001$.

Population specific SNPs

We investigated the potential of identifying SNPs that were specific to each population identified by the admixture analysis. These groupings can include a large number of varieties and the varieties have varying levels of relatedness, which means varying levels of SNP sharing occur within and between populations, so a large number of SNPs would be required to discriminate between populations. The variants were filtered to select for bi-allelic SNPs where all samples were homozygous and a series of Quantitative trait locus (QTL) analyses were performed to identify the most discriminatory SNPs. The phenotype for each QTL analysis was set as a binary trait of 'same population' or 'other populations' using the population groupings identified by the admixture analysis. A separate QTL analysis was performed for each population and the SNPs with the highest LOD score and largest allele frequency difference were taken as being the most predictive for that population. These SNPs were then used to train a random forest model [25] using the R randomForest package [26] and the R caret package [27]. Gene information from the GFF was overlaid on the SNP data to identify any population discriminatory SNP that was within a gene. In addition, genes within intervals of closely spaced predictive SNPs were also investigated.

Population classification

We deployed machine learning data classification to investigate whether the set of population specific SNPs we selected can be used to discriminate between the five populations. We used 10-fold cross validation [28], which is a technique in machine learning to measure the performance of prediction from a set of classifiers. We used random forest model [25] as a main classifier in the analysis training on the 268 selected SNPs to classify the five populations of 300 rice varieties. A true positive (TP) is when the predicted class was the same as the ADMIXTURE derived class. A false positive (FP) is the case when the classifier predicts that a sample belongs to some specific class but it is not the member of that class. A false negative (FN) is when a sample that belongs to a specific class is not predicted to be a member of that class. The precision is the ratio of the number of TP cases to the number of

TP and FP cases. The recall is the ratio of the number of TP cases to the number of TP and FN cases. The F1 score is calculated from precision and recall as follows.

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (1)$$

Competing interests

The authors declare that they have no competing interests.

Author's contributions

P. Vejchasarn: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper. J.R. Shearman: Contributed reagents, materials, analysis tools or data; Analyzed and interpreted the data; Wrote the paper. U. Chaiprom: Contributed reagents, materials, analysis tools or data; Analyzed and interpreted the data; Wrote the paper. Y. Phansenee: Performed the experiments; Analyzed and interpreted the data; T. Tulyananda: Analyzed and interpreted the data; Wrote the paper. J. Jairin: Conceived and designed the experiments; Analyzed and interpreted the data; C. Amornbunchornvej: Contributed reagents, materials, analysis tools or data; Analyzed and interpreted the data; Wrote the paper.

Author details

[1]Ubonratchathani Rice Research Institute, 34000 Ubonratchathani, Thailand. [2]National Omics Center, National Science and Technology Development Agency, 111 Thailand Science Park, Paholyothin Road, Khlong Nueng, Khlong Luang, 12120 Pathum Thani, Thailand. [3]National Biobank of Thailand (NBT), 144 Thailand Science Park, Phahonyothin Road, Khlong Nueng, Khlong Luang, 12120 Pathum Thani, Thailand. [4]School of Bioinnovation & Bio-based Product Intelligence, Faculty of Science, Mahidol University, 10400 Bangkok, Thailand. [5]National Electronics and Computer Technology Center (NECTEC), 112 Phahonyothin Road, Khlong Nueng, Khlong Luang District, 12120 Pathum Thani, Thailand.

References

1. (NSO), T.N.S.O.: Thailand Environment Statistics 2020. International series of monographs on physics. Thailand's National Statistical Office (NSO), Bangkok (2020). http://service.nso.go.th/nso/nsopublish/pubs/e-book/Thailand_Environment_2020/files/assets/common/downloads/publication.pdf
2. Weber, S., Lehman, H., Barela, T., Hawks, S., Harriman, D.: Rice or millets: early farming strategies in prehistoric central thailand. Archaeological and Anthropological Sciences **2**(2), 79–88 (2010). doi:10.1007/s12520-010-0030-3
3. Siamwalla, A.: A history of rice policies in thailand. Food Research Institute Studies **14**(1387-2016-115909), 233–249 (1975)
4. Chakhonkaen, S., Pitnjam, K., Saisuk, W., Ukoskit, K., Muangprom, A.: Genetic structure of thai rice and rice accessions obtained from the international rice research institute. Rice **5**(1), 19 (2012)
5. Pusadee, T., Wongtamee, A., Rerkasem, B., Olsen, K.M., Jamjod, S.: Farmers drive genetic diversity of thai purple rice (oryza sativa l.) landraces. Economic Botany **73**(1), 76–85 (2019)
6. Pathaichindachote, W., Panyawut, N., Sikaewtung, K., Patarapuwadol, S., Muangprom, A.: Genetic diversity and allelic frequency of selected thai and exotic rice germplasm using ssr markers. Rice Science **26**(6), 393–403 (2019)
7. Kladmook, M., Kumchoo, T., Hongtrakul, V.: Genetic diversity analysis and subspecies classification of thailand rice landraces using dna markers. African Journal of Biotechnology **11**(76), 14044–14053 (2012)
8. Reig-Valiente, J.L., Viruel, J., Sales, E., Marqués, L., Terol, J., Gut, M., Derdak, S., Talón, M., Domingo, C.: Genetic diversity and population structure of rice varieties cultivated in temperate regions. Rice **9**(1), 58 (2016)
9. Pusadee, T., Schaal, B.A., Rerkasem, B., Jamjod, S.: Population structure of the primary gene pool of oryza sativa in thailand. Genetic Resources and Crop Evolution **60**(1), 335–353 (2013)
10. Moonsap, P., Laksanavilat, N., Tasanasuwan, P., Kate-Ngam, S., Jantasuriyarat, C.: Assessment of genetic variation of 15 thai elite rice cultivars using indel markers. Crop Breeding and Applied Biotechnology **19**(1), 15–21 (2019)
11. Holsinger, K.E., Weir, B.S.: Genetics in geographically structured populations: defining, estimating and interpreting f st. Nature Reviews Genetics **10**(9), 639 (2009)
12. Mikami, K., Katagiri, T., Iuchi, S., Yamaguchi-Shinozaki, K., Shinozaki, K.: A gene encoding phosphatidylinositol-4-phosphate 5-kinase is induced by water stress and abscisic acid in arabidopsis thaliana. The Plant Journal **15**(4), 563–568 (1998). doi:10.1046/j.1365-313X.1998.00227.x. https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-313X.1998.00227.x
13. Zhang, H., Li, L., He, Y., Qin, Q., Chen, C., Wei, Z., Tan, X., Xie, K., Zhang, R., Hong, G., Li, J., Li, J., Yan, C., Yan, F., Li, Y., Chen, J., Sun, Z.: Distinct modes of manipulation of rice auxin response factor osarf17 by different plant rna viruses for infection. Proceedings of the National Academy of Sciences **117**(16), 9112–9121 (2020). doi:10.1073/pnas.1918254117. https://www.pnas.org/content/117/16/9112.full.pdf
14. Yamaguchi, T., Lee, D.Y., Miyao, A., Hirochika, H., An, G., Hirano, H.-Y.: Functional diversification of the two c-class mads box genes osmads3 and osmads58 in oryza sativa. The Plant Cell **18**(1), 15–28 (2006)
15. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The sequence alignment/map format and samtools. Bioinformatics **25**(16), 2078–2079 (2009)

16. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprint arXiv:1303.3997 (2013)
17. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.*: The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome research **20**(9), 1297–1303 (2010)
18. Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S.: TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23**(19), 2633–2635 (2007). doi:10.1093/bioinformatics/btm308. http://oup.prod.sis.lan/bioinformatics/article-pdf/23/19/2633/451862/btm308.pdf
19. Alexander, D.H., Novembre, J., Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. Genome research **19**(9), 1655–1664 (2009)
20. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics **21**, 768–769 (1965)
21. R Development Core Team, R., et al.: R: A language and environment for statistical computing. R foundation for statistical computing Vienna, Austria (2011)
22. Efron, B.: Bootstrap Methods: Another Look at the Jackknife, pp. 569–593. Springer, New York, NY (1992). doi:10.1007/978-1-4612-4380-9_41
23. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4), 406–425 (1987). doi:10.1093/oxfordjournals.molbev.a040454. http://oup.prod.sis.lan/mbe/article-pdf/4/4/406/11167444/7sait.pdf
24. Amornbunchornvej, C., Surasvadi, N., Plangprasopchok, A., Thajchayapong, S.: A nonparametric framework for inferring orders of categorical data from category-real pairs. Heliyon **6**(11), 05435 (2020). doi:10.1016/j.heliyon.2020.e05435
25. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
26. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)
27. Kuhn, M.: Caret: Classification and Regression Training. (2020). R package version 6.0-86. https://CRAN.R-project.org/package=caret
28. Allen, D.M.: The relationship between variable selection and data agumentation and a method for prediction. Technometrics **16**(1), 125–127 (1974). doi:10.1080/00401706.1974.10489157. https://www.tandfonline.com/doi/pdf/10.1080/00401706.1974.10489157

Supplementary

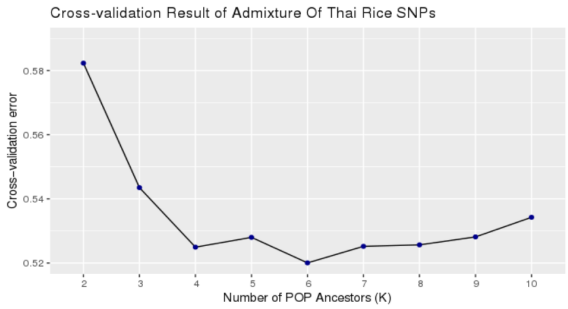

Figure 4 The Elbow method result that was used to find the optimal number of ancestors from ADMIXTURE. The optimal number is 4 ancestors.

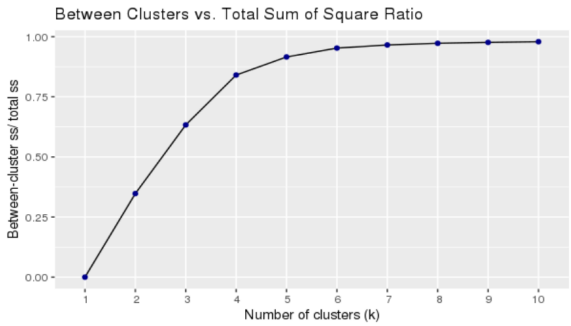

Figure 5 The Elbow method result that was used to find the optimal number of clusters for k-means clustering. The optimal number is 5 clusters.