論文の内容の要旨

論文題目　Molecular evolution of domesticated Asian rice revealed by genome-wide comparative analyses（アジア栽培イネにおける分子進化のゲノム比較解析による解明）

氏　　名　楊　靜佳

Introduction

Domestication is the process in which a population of wild animals or plants has evolved to match human requirements through a series of artificial selection. Because domestication has had strong influence on human culture and history, it is of particular interest to know when and from where the domesticated species originated, for what progenitors have been utilized, and how cultivars were created. In case of the Asian cultivated rice *Oryza sativa*, these basic questions remained unsolved, because the relationship of wild and cultivated rice varieties is quite complex. *O. sativa* consists of diverse varieties, many of which can be classified into two major groups, *japonica* and *indica*. There are two alternative hypotheses about the process of domestication from its progenitor *O. rufipogon*. One suggests that *japonica* and *indica* were derived from a single ancestor and then diverged, while the other suggests that these two groups were independently domesticated from different varieties of *O. rufipogon* that diverged much earlier than domestication. Since either of the hypotheses was apparently supported by several lines of evidence, the origin(s) of the cultivated rice and divergence time of *japonica* and *indica* remained a matter of debate. In addition, it was suspected that some genes might have been transferred between *japonica* and *indica* through hybridization, but the amount and directions of genes transferred were unknown. Here, using large-scale data generated by the next-generation sequencers, I conduct genome and transcriptome analyses of wild and cultivated rice varieties. I attempt to reveal the origin(s) and divergence time of the Asian cultivated rice, and to elucidate contribution of gene flow between *japonica* and *indica* to the process of their domestication.

Results

1. Independent domestication revealed by large-scale data analysis

Gene tree discordance, which is widely observed in closely related species, is a serious problem that gene trees are inconsistent with an expected species relationship because of ancestral polymorphisms. To examine gene tree discordance in rice, the genome sequences of a *japonica* cultivar Nipponbare (NB) and an *indica* cultivar 93-11 were compared with 2,026 full-length cDNA sequences obtained from *O. rufipogon* W1943, which was sampled in Jiangxi, China. As a result, a total of 167 gene trees could be reconstructed with statistical

significance, and as theory suggests, all the three possible tree topologies were observed. However, one topology that supported closeness of *japonica* and W1943 were clearly predominant, whereas the other two were minor (Figure 1A). Therefore, it is suggested that the major topology showing a sister group of *japonica* and W1943 represents the species relationship of these groups and the others were caused by ancestral polymorphisms. This result shows that, despite the gene tree discordance, a careful examination of large-scale data should be useful to solve phylogeny of closely related species.

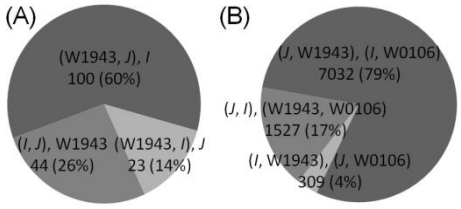

Figure 1. Ratios of gene trees. (A) NB, 93-11, and W1943 were used. (B) NB, GLA4, W1943, and W0106 were used. J: *japonica*; I: *indica*.

To answer the question of the origin(s) of the cultivated rice, RNA-seq data of two diverse varieties (W1943 and W0106) of *O. rufipogon* were used. Because W0106 was sampled in Orissa, India, it was expected to show a phylogenetic pattern different from that of W1943. In addition, the genomic sequences of another *indica* cultivar Guangluai-4 (GLA4), for which ~20X coverage Illumina reads were available, was employed. Aligning all the short reads to the *japonica* genome, I obtained 6,142,852 sites that were shared among all the four groups and 8,868 parsimony-informative sites were found. Under the maximum parsimony principle, a tree topology can be inferred at each of these 8,868 sites. It was expected that one of the three possible unrooted tree topologies, which is consistent with the evolutionary relationship of these groups, should be predominant, whereas the other two minor topologies would be observed because of ancestral polymorphisms. The data clearly indicates that at 7032 sites (79%) *japonica* and *indica* are close to W1943 and W0106, respectively, and *japonica* and *indica* do not form a sister group (Figure 1B). Hence, I conclude that *japonica* and *indica* were derived independently from distinct groups of *O. rufipogon*.

2. Gene flow from *japonica* to *indica*

If ancestral polymorphisms are the only cause of gene tree discordance, similar numbers of minor tree topologies should be observed. However, the minor tree showing that *japonica* and *indica* consist a sister group was supported greater than the other one (Figure 1). In a genome alignment of NB and 93-11 (Figure 2A) and that of NB and GLA4 (Figure 2B), several nearly identical regions, which were possibly due to hybridization between *japonica* and *indica*, were found. Thus, a larger number of trees showing *japonica-indica* closeness can be accounted for by hybridization events after their split.

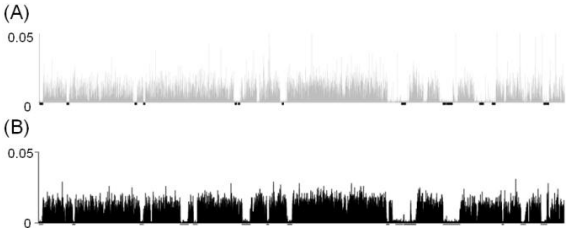

Figure 2. Alignment in chromosome 3 (A) between NB and 93-11 and (B) between NB and GLA4. Thick bars below x-axis represent "nearly identical regions" that showed <0.2 nt / kb differences and >75 kb.

In many cases, the positions of the nearly identical regions were similar between two alignments (Figure 2). This is possibly because hybridization occurred in the common ancestor of the two *indica* cultivars. To investigate this possibility, a window analysis of 10 kb was performed to compare nucleotide differences between NB and 93-11 ($d_{jp\text{-}9311}$) and those between NB and GLA4 ($d_{jp\text{-}GLA4}$). Since 93-11 and GLA4 are equally distant to NB, $d_{jp\text{-}9311}$ and $d_{jp\text{-}GLA4}$ should be scattered around the diagonal line in an x-y plot, although they may slightly fluctuate by chance (Figure 3A, region I). In addition, because recent hybridization before the split of 93-11 and GLA4 was expected, there should be windows with small $d_{jp\text{-}9311}$ and $d_{jp\text{-}GLA4}$ (Figure 3A, region II). In fact, the

x-y plot of $d_{jp\text{-}9311}$ and $d_{jp\text{-}GLA4}$ showed that a significant number of dots were located in the regions II (Figure 3B), suggesting hybridization before divergence of the *indica* cultivars.

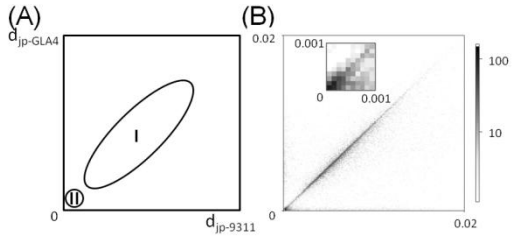

Figure 3. (A) Expected and (B) observed differences between *japonica* and *indica*.

Since recent gene flow was found between *japonica* and *indica*, next question was its amount and direction. If a genomic region was introgressed from *japonica* to *indica*, the region should be more closely related to W1943 than to W0106. This can be examined by comparing $d_{jp/in\text{-}W1943}$ and $d_{jp/in\text{-}W0106}$, where jp/in means the nearly identical regions between NB and GLA4. It was found that at least 13 Mb of the nearly identical regions, which contain ~2,000 genes, were introgressed from *japonica* to *indica*, while only 0.77 Mb were from *indica* to *japonica*. These observations suggest that there might have been much more genes transferred from *japonica* to *indica*. Since relatively strict criteria were used to define nearly identical regions, these should be regarded as minimum estimates of the amount of introgressed genes.

3. *Japonica* and *indica* groups diverged earlier than domestication

The estimated divergence time of *japonica* and *indica* ranges from 8,200 years to 0.44 million years in previous studies. This considerable range of the time might be due to inappropriate assumption such as single origin hypothesis or disregard of ancestral polymorphisms and hybridization. It is believed that molecular data such as gene sequences obtained from different species or groups are useful to accurately estimate divergence time. However, because genes (alleles) are diverged before the divergence of two groups (Figure 4), observed divergence time of genes (t_{obs}) is always larger than the real divergence time of groups (t). The difference between t_{obs} and t is due to ancestral polymorphisms and depends on the ancestral effective population size. Here I employed a model that includes ancestral polymorphisms and estimated both divergence time and the effective population size using the maximum likelihood method. The number of nucleotide differences was counted in alignments of 2 kb between NB and GLA4 and a distribution of the differences was created. The genome alignment of these two cultivars consists of two different regions that diverged at ancient time and were introgressed recently. Thus, the distribution of the nucleotide differences can be a combination of two distributions (Figure 5A), each of which has its own divergence time and effective population size. Moreover, the proportions of the two distributions, which correspond to the amounts of introgressed and non-introgressed regions, should be estimated. As a result, the distribution based on estimated parameters (Figure 5B, black curve) fit well with the observed distribution (Figure 5B, gray bars). The estimated divergence time of *japonica* and *indica* was 120,000 years, which is much earlier than the widely accepted domestication time, 9,000 years. Thus, this result provides a strong support that *japonica* and *indica* were independently derived from different varieties of wild rice. The ancient effective population size was 125,000, while the effective population size of the *japonica* group

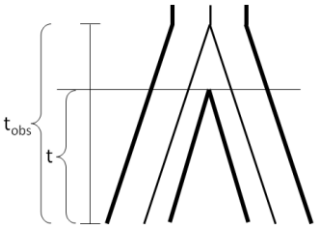

Figure 4. Observed and real divergence time.

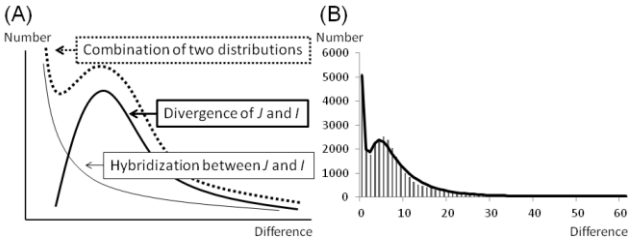

Figure 5. (A) Expected and (B) observed distributions.

before hybridization was 6,700. This reduced effective population size of *japonica* might be related to a severe bottleneck in *japonica*. It is reasonable to think that only a small number of *japonica* ancestors were selected during domestication. It is noteworthy that the estimated proportion of the introgressed regions was 21%, which is larger than my previous estimate. This inconsistence may be due to the strict criteria used for selecting nearly identical regions in previous analysis. These results show that *japonica* genes may have played important roles in shaping current *indica* cultivars.

Conclusion

Several significant aspects of molecular evolution of rice were revealed. First, this study provides clear evidence that *japonica* and *indica* were originated from different varieties of wild rice. Intriguingly, W1943 was sampled near the Middle Yangtze, which is well-known as the center of early rice cultivation, whereas W0106 was sampled in East India, which is close to the Middle Ganges valley where archeological evidence indicates rice cultivation of ~7,000 years ago. This geological evidence and the phylogeny disclosed in this study suggest that rice was domesticated multiple times in distant places. Second, a significant portion (21%) of the *indica* genome was derived from *japonica*. This finding suggests that even though *japonica* and *indica* diverged much earlier than their domestication and were separately domesticated, they recently crossed, so that some *japonica* genes contributed to create the modern *indica* rice. The introgressed genes detected here are possibly related to important traits, and can be useful resources for breeding of rice in the future.

47-087358：楊靜佳