Multiple origin but single domestication led to domesticated Asian rice

by

Jae Young Choi[1] and Michael D. Purugganan[1,2]

[1]Center for Genomics and Systems Biology, Department of Biology, 12 Waverly Place, New York University, New York, NY 10003 USA

[2]Center for Genomics and Systems Biology, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

Corresponding Author: Jae Young Choi. Email: jyc387@nyu.edu

Abstract

The domestication scenario that led to Asian rice (*Oryza sativa*) is a contentious topic. Here, we have reanalyzed a previously published large-scale wild and domesticated rice dataset, which were also analyzed by two studies but resulted in two contrasting domestication scenario. Our result indicates Asian rice originated from multiple wild progenitor subpopulations, however, domestication occurred only once and the domestication alleles were transferred between rice subpopulation through introgression.

Elucidating the origins of Asian rice (*Oryza sativa*) domestication has been a contentious field (Gross and Zhao 2014). With whole genome data, it is becoming apparent that each Asian rice variety group/subspecies (aus, indica, and japonica) had distinct subpopulations of wild rice (*O. nivara* or *O. rufipogon*) as its progenitor (Huang *et al.* 2012). However, whether rice was domesticated once and subsequent varieties were formed by introgression with different wild progenitors, or whether each variety was domesticated independently in different parts of Asia is debatable.

The debate mainly arose from two studies analyzing the same data but surprisingly arriving at two different domestication scenarios: Huang et al. (2012) supporting the single domestication with introgression model whereas Civán et al. (2015) supporting the multiple domestication model. Both studies used a reduction in polymorphism levels as a metric to detect local genomic regions associated with domestication, and the evolutionary history of those regions were interpreted as the domestication history for Asian rice. However, even population genetic model-based methods of detecting selective sweeps are prone to false positives and with the right condition any evolutionary scenario can be interpreted with a false positive selective sweep region (Pavlidis *et al.* 2012). Given that each Asian rice had separate wild progenitor population of origin, any false positive selective sweep region will likely to be concordant with the underlying species phylogeny, and spuriously support the multiple domestication model. In addition, both studies used genotype calls made from a low coverage (1~2X) resequencing data (Huang *et al.* 2012). However, uncertainty associated with genotype calls made from low coverage data (Nielsen *et al.* 2011) could be another source that led to the difference in results for the two studies. Thus, we revisited the

domestication scenarios proposed by the two studies and reanalyzed the Huang et al. data using a complete probabilistic framework that takes the uncertainty in SNP and genotype likelihoods into consideration (Fumagalli *et al.* 2014; Korneliussen *et al.* 2014). We then carefully compared our results against the two domestication models and contrasted it against results from both Huang et al. (2012) and Civán et al. (2015) studies.

In both Huang et al. (2012) and Civán et al. (2015) studies, the phylogeny based on genome-wide data versus putative domestication region sequences were compared to determine which domestication scenario is best supported by the data. We reconstructed the genome-wide phylogeny by estimating genetic distances between domesticated and wild rice using genotype probabilities (Vieira *et al.* 2016). Three different parameters were used to estimate genotype probabilities, which were subsequently used to estimate genetic distances and build neighbor-joining trees for each chromosome (Supplemental Fig 1). Comparing trees built from the three different parameters, each chromosomal phylogeny was largely concordant with each other. Further, the trees corroborated the results of Huang et al. where the japonicas were most closely related to the Or-III wild rice subpopulation, while indica and aus were most closely related to the Or-I wild rice subpopulation.

We then scanned for local genomic regions associated with domestication related selective sweeps to infer the domestication history of Asian rice. Sweeps were identified using sliding windows that were estimating the ratio of wild to domesticate polymorphism (π_w/π_d). To identify putative selective sweep regions, we chose the approach of Civán et al. (2015) and identified sweep regions separately for each rice subpopulation. If rice had a single domestication origin, all three rice subpopulations

would have identical sweep regions with shared haplotypes; otherwise, the single domestication model cannot be supported. These regions with co-located low-diversity genomic regions (CLDGRs; (Civáň *et al.* 2015)) were identified using a 20 kbp sliding window. To identify significant CLDGRs we chose a stringent cutoff to conservatively identify candidate regions (see Material and Method for detail) and identified a total of 39 CLDGRs (Supplemental Table 1).

Neighbor-joining trees were then reconstructed for each 39 CLDGRs (Supplemental Figure 2). The majority of CLDGRs showed monophyletic relationships among the domesticated rice subpopulation, where japonica, indica, and aus were clustering between and not within subpopulation types. A few windows (e.g. 2:11,660,000-11,680,000) showed phylogenetic relationships where each domesticated sample were clustering within the same subpopulation type. This initially suggested the evolutionary history of CLDGRs were most consistent with the single domestication origin model. We then examined larger window sizes of 100 kbp, 500 kbp, and 1000 kbp for candidate CLDGRs (Supplemental Table 1) and reconstructed phylogenies for those regions (Supplemental Fig 3,4, and 5). Larger window sizes have less number of windows for analysis, hence leading to lesser number of CLDGRs being identified (Supplemental Table 2). Nonetheless, with increasing window sizes CLDGR phylogenies were becoming more congruent with the genome-wide phylogenies, consistent with the multiple domestication origins model.

CLDGRs, however, are candidate regions for domestication and false-positive CLDGRs may represent regions affected by domestication-related bottlenecks. As population bottlenecking can decrease effective population sizes, false positive CLDGRs

may represent regions of the genome with increased lineage sorting and becoming more concordant with the underlying species phylogeny (Pamilo and Nei 1988). Hence, it is crucial that a CLDGR have additional evidence that can associate it with selection and differentiate its evolutionary history from the underlying species phylogeny. To do so we searched CLDGRs that overlapped genes with functional genetic evidence related to domestication. We found three known domestication genes: long and barbed awn gene *LABA1* (chr4:25,959,399-25,963,504), the prostrate growth gene *PROG1* (chr7:2,839,194-2,840,089), and shattering locus *sh4* (chr4:34,231,186-34,233,221) (Li *et al.* 2006; Tan *et al.* 2008; Hua *et al.* 2015). Interestingly, the gene *sh4* was the only gene detected across multiple sliding window sizes excluding the largest 1000 kbp window (Supplemental Table 1).

Phylogenetic trees were then reconstructed for the three domestication loci that included 20 kbp upstream and downstream of their coding sequence. We note for all three genes the casual variant resulting in the domestication phenotype were located in the protein coding sequences (Li *et al.* 2006; Jin *et al.* 2008; Hua *et al.* 2015). For all three genomic regions, the phylogenies were clustering different subpopulation types of domesticated rice together (Figure 1), consistent with the single domestication scenario. Further, in all three regions the most closely related wild rice corresponded to the Or-III subpopulation, supporting the hypothesis that the domestication alleles were introgressed from japonica into indica and aus (Huang *et al.* 2012; Choi *et al.* 2017).

Interestingly, *sh4* was identified as a candidate gene with evidence of selective sweep in this study and both Huang et al. (2012) and Civán et al. (2015). However, only Civán et al. (2015) did not find evidence of single origin in a phylogenetic tree

reconstructed from a 240 kbp region surrounding *sh4*. When we reconstructed phylogenies for 40 kbp windows surrounding the *sh4* region, the downstream region of *sh4* had phylogenies in which the domesticated rice were clustering with the same subpopulation types (Supplemental Fig 6). We then reconstructed the phylogeny for large genetic regions surrounding each three domestication loci and discovered with each increased window size, the phylogeny of the region increasingly corroborated the genome-wide phylogeny by clustering with the same subpopulation type (Supplemental Fig 7). Thus, the domestication-related evolutionary history for sh4 is limited to the gene and its upstream region. Thus, including large flanking regions can lead to phylogenies that are concordant with the genome-wide species phylogeny, spuriously concluding it as evidence for the multiple domestication origin model.

In this study we have used the same approach as Huang et al. (2012) and Civán et al. (2015) to search for regions of domestication related selective sweeps and investigated those regions' evolutionary history. With stringent thresholds and conservative assumptions to exclude false positive CLDGRs we were able to narrow down to three genes (*LABA1*, *PROG1*, and *sh4*), which were likely to be the key genes involved in the domestication of Asian rice (Meyer and Purugganan 2013). Civán et al. (2015) had criticized the role of *PROG1* and *sh4* in domestication due to several wild rice alleles clustering with the domesticated alleles (Figure 1). However, evidence from de-domesticated weedy rice shows feralized rice can carry causative domestication allele but not retain any of the domestication phenotypes (Li *et al.* 2017), suggesting some of the wild rice in the Huang et al. (2012) dataset may actually represent different stages of feralized domesticated rice (Wang *et al.* 2017). Thus, clustering of wild rice with

domesticated rice in candidate domesticated genes does not preclude those genes from having an important role in domestication.

In the end, our evolutionary analysis for the domestication loci *LABA1*, *PROG1*, and *sh4* are consistent with both Sanger and next-generation sequencing results (Li *et al.* 2006; Tan *et al.* 2008; Xu *et al.* 2011; Huang *et al.* 2012; Hua *et al.* 2015). Our results are also consistent with the archaeological and genomic evidences (Fuller *et al.* 2010; Choi *et al.* 2017). Here then, we propose the Asian rice has evolved from multiple origins but de novo domestication had only occurred once (Figure 2). Specifically, our model hypothesizes each domesticated rice subpopulation had distinct wild rice subpopulation as its immediate progenitor, but domestication only occurred once in japonica involving the *genes LABA1, PROG1,* and *sh4*. The domestication alleles for these genes were then subsequently introgressed into the wild progenitors of aus and indica by gene flow and ultimately led to their domestication.

Materials and Method

Raw paired-end FASTQ data from the Huang et al. study was download from the National Center for Biotechnology Information website under bioproject ID numbers PRJEB2052, PRJEB2578, PRJEB2829. We excluded the aromatic rice group from the analysis, as their sample sizes were too small and we excluded the few samples that had too high coverage. In the end a total of 1477 samples were selected for analysis (Supplemental Table 3).

Raw reads were then trimmed for adapter contamination and low quality bases using trimmomatic ver. 0.36 (Bolger *et al.* 2014) with the command:

```
java -jar trimmomatic-0.36.jar PE \
    $FASTQ1 $FASTQ2 \
    $FASTQ1_paired $FASTQ1_unpaired $FASTQ2_paired $FASTQ2_unpaired \
     ILLUMINACLIP:TruSeq2-PE.fa:2:30:10:4 \
     LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:30
```

Quality controlled FASTQ reads were then realigned to the reference japonica genome downloaded from EnsemblPlants release 30 (ftp://ftp.ensemblgenomes.org/pub/plants/). Reads were then mapped to the reference genome using the program BWA-MEM ver. 0.7.15 (Li 2013) with default parameters. Alignment files were then processed with PICARD ver. 2.9.0 (http://broadinstitute.github.io/picard/) and GATK ver. 3.7 (McKenna *et al.* 2010) toolkits to remove PCR duplicates and realign around INDEL regions (DePristo *et al.* 2011).

Using the processed alignment files genotype probabilities were calculated with the program ANGSD ver. 0.913 (Korneliussen *et al.* 2014). The genotype probabilities were then used by the program ngsTools (Fumagalli *et al.* 2014) to conduct population

genetic analysis. To estimate theta (θ) ngsTools uses the site frequency spectrum as a prior to calculate allele frequency probabilities. Usually site frequency spectrum requires an appropriate outgroup sequence to infer the ancestral state of each site. However, for calculating Watterson and Tajima's θ it is not necessary to know whether each polymorphic site is a high or low frequency variant (Korneliussen *et al.* 2013). Hence, we used the same reference japonica genome as the outgroup but strictly for purposes of calculating θ. Per site allele frequency likelihood was calculated using ANGSD with the commands:

```
angsd -b $BAMLIST -ref $REF -anc $REF -out $SFS -r $CHR \
    -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0 \
    -C 50 -baq 1 -minMapQ 20 -minQ 30 \
    -minInd $minInd \
    -setMinDepth $setMinDepth \
    -setMaxDepth $setMaxDepth \
    -doCounts 1 -GL 1 -doSaf 1
```

Per site allele frequency for each domesticated and wild subpopulation was calculated separately with different filtering parameters using the options –minInd, -setMinDepth, -setMaxDepth. Specifically, -minInd and –setMinDepth were set as a third of the number individuals in the subpopulation while –setMaxDepth was set as five times the number individuals in the subpopulation. Overall site frequency spectrum was then calculated with the realSFS program from the ANGSD package. Using each subpopulation's site frequency spectrum as prior, we then calculated θ for each subpopulation using ANGSD with the command:

```
angsd -b $BAMLIST -ref $REF -anc $REF -out $THETA -r $CHR \
    -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0 \
    -C 50 -baq 1 -minMapQ 20 -minQ 30
    -minInd $minInd \
    -setMinDepth $setMinDepth \
    -setMaxDepth $setMaxDepth
    -doCounts 1 -GL 1 -doSaf 1 \
    -doThetas 1 -pest $SFS
```

Sliding window analysis was then conducted with the thetaStat program from the ANGSD package using window length and step sizes of 20 kbp, 100 kbp, 500 kbp, and 1000 kbp.

For each window θ per site was estimated by dividing Tajima's theta (θ_π) against the total number of sites with data in the window. Windows with less then 25% of sites with data were discarded from downstream analysis. This resulted in a minimum of 90% of the windows being analyzed (Supplemental Table 2). To calculate π_w/π_d values we chose the Or-II subpopulation to calculate π_w since Or-II subpopulation was most distantly related to all three domesticated rice subpopulation (Supplemental Fig 1). π_w/π_d values were calculated separately for each domesticated rice subpopulations. Windows with large π_w/π_d values were designated as candidate domestication selective sweep region, and significance was determined using an empirical distribution of π_w/π_d values.

Japonica has demographic history that is consistent with more intense domestication related bottlenecks then aus and indica (Xu *et al.* 2011). Thus, many π_w/π_d values for japonica are expected to be similar between true domestication sweep and neutral regions, causing difficulties in identifying true positive selective sweeps. Hence, we chose the approach of Civán et al. (2015) by using a single threshold π_w/π_d value to determine significance for all three subpopulation. In contrast to Civán et al. (2015) we

chose our threshold based on the empirical distribution of each subpopulation. The 97.5 percentile π_w/π_d values were determined for each domesticated rice subpopulation, and the subpopulation with the lowest 97.5 percentile π_w/π_d values was decided as the significance threshold. The threshold percentile that is represented by each subpopulation and window size is listed in Supplemental Table 4. This threshold assumes at least for one subpopulation, to represent the true π_w/π_d value seen in a window after domestication related selective sweeps, while in the other two subpopulations the threshold maybe seen after a selective sweep or a population bottleneck. These CLDGRs then, represent candidate domestication related selective sweep regions for all three subpopulations, and it is necessary for each CLDGR to have additional information to differentiate itself from the background domestication related bottleneck scenarios. We assumed CLDGRs overlapping genes with functional genetic evidence related to domestication phenotypes (Meyer and Purugganan 2013) as true candidate domestication genes.

To account for the uncertainty in the underlying data, phylogenetic analysis were conducted by estimating pairwise genetic distances from genotype probabilities (Vieira *et al.* 2016). We ran the program ANGSD to calculate genotype probabilities for all 1477 domesticated and wild rice samples using the command:

```
angsd -b $BAMLIST -ref $REF -out $GENOPP -r $CHR \
    -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0 \
    -C 50 -baq 1 -minMapQ 20 -minQ 30 \
    -minInd $minInd
    -setMinDepth $setMinDepth
    -setMaxDepth $setMaxDepth
    -doCounts 1 -GL 1 -doMajorMinor 1 -doMaf 1 \
    -skipTriallelic 1 -SNP_pval 1e-3 -doGeno 8 -doPost 1
```

Initially, the effects of different filtering parameters on the downstream phylogenetic analysis were examined by using three different parameter values for the options –minInd, -setMinDepth, -setMaxDepth: 1) minInd=492, setMinDepth=492, setMaxDepth=4920; 2) minInd=738, setMinDepth=738, setMaxDepth=8862; 3) minInd=492, setMinDepth=369, setMaxDepth=8862. Afterwards all subsequent phylogenetic analysis were conducted with genotype posterior probabilities calculated using the minInd=492, setMinDepth=492, setMaxDepth=4920 parameter set. Genotype posterior probabilities were then used by the program ngsDist from the ngsTools package to estimate all pairwise genetic distances. Neighbor-joining trees were reconstructed with the genetic distances using the program FastME ver. 2.1.5 (Lefort *et al.* 2015).

References

Bolger A. M., Lohse M., Usadel B., 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Choi J. Y., Platts A. E., Fuller D. Q., Hsing Y.-I., Wing R. A., *et al.*, 2017 The rice paradox: Multiple origins but single domestication in Asian rice. Mol. Biol. Evol. 34: msx049.

Civáň P., Craig H., Cox C. J., Brown T. A., Fuller D. Q., *et al.*, 2015 Three geographically separate domestications of Asian rice. Nat. Plants 1: 15164.

DePristo M. A., Banks E., Poplin R., Garimella K. V, Maguire J. R., *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Fuller D. Q., Sato Y.-I., Castillo C., Qin L., Weisskopf A. R., *et al.*, 2010 Consilience of genetics and archaeobotany in the entangled history of rice. Archaeol. Anthropol. Sci. 2: 115–131.

Fumagalli M., Vieira F. G., Linderoth T., Nielsen R., 2014 ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics 30: 1486–7.

Gross B. L., Zhao Z., 2014 Archaeological and genetic insights into the origins of domesticated rice. Proc. Natl. Acad. Sci. 111: 6190–6197.

Hua L., Wang D. R., Tan L., Fu Y., Liu F., *et al.*, 2015 LABA1, a Domestication Gene Associated with Long, Barbed Awns in Wild Rice. Plant Cell 27: 1875–1888.

Huang X., Kurata N., Wei X., Wang Z.-X., Wang A., *et al.*, 2012 A map of rice genome variation reveals the origin of cultivated rice. Nature 490: 497–501.

Jin J., Huang W., Gao J.-P., Yang J., Shi M., *et al.*, 2008 Genetic control of rice plant architecture under domestication. Nat. Genet. 40: 1365–9.

Korneliussen T. S., Moltke I., Albrechtsen A., Nielsen R., 2013 Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics 14: 289.

Korneliussen T. S., Albrechtsen A., Nielsen R., 2014 ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15: 356.

Lefort V., Desper R., Gascuel O., 2015 FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol. Biol. Evol. 32: 2798–2800.

Li C., Zhou A., Sang T., 2006 Rice domestication by reducing shattering. Science 311: 1936–9.

Li H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li L.-F., Li Y.-L., Jia Y., Caicedo A. L., Olsen K. M., 2017 Signatures of adaptation in the weedy rice genome. Nat. Genet.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–303.

Meyer R. S., Purugganan M. D., 2013 Evolution of crop species: genetics of domestication and diversification. Nat. Rev. Genet. 14: 840–852.

Nielsen R., Paul J. S., Albrechtsen A., Song Y. S., 2011 Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12: 443–51.

Pamilo P., Nei M., 1988 Relationships between gene trees and species trees. Mol. Biol.

Evol. 5: 568–83.

Pavlidis P., Jensen J. D., Stephan W., Stamatakis A., 2012 A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. Mol. Biol. Evol. 29: 3237–3248.

Tan L., Li X., Liu F., Sun X., Li C., *et al.*, 2008 Control of a key transition from prostrate to erect growth in rice domestication. Nat. Genet. 40: 1360–1364.

Vieira F. G., Lassalle F., Korneliussen T. S., Fumagalli M., 2016 Improving the estimation of genetic distances from Next-Generation Sequencing data. Biol. J. Linn. Soc. 117: 139–149.

Wang H., Vieira F. G., Crawford J. E., Chu C., Nielsen R., 2017 Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. Genome Res.: gr.204800.116.

Xu X., Liu X., Ge S., Jensen J. D., Hu F., *et al.*, 2011 Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. 30: 105–111.

Supplemental Fig1. Neighbor-joining trees for 12 chromosomes. Each row represent trees built using genotype posterior probabilities calculated from 3 different parameters. Top, middle, and bottom row represents tree built from genotype posterior probabilities calculated from parameter 1, 2, and 3 listed in materials and method.

Supplemental Fig2. Neighbor-joining trees for the 39 CLDGRs identified after 20 kbp sliding window.

Supplemental Fig3. Neighbor-joining trees for the 10 CLDGRs identified after 100 kbp sliding window.

Supplemental Fig4. Neighbor-joining trees for the 4 CLDGRs identified after 500 kbp sliding window.

Supplemental Fig5. Neighbor-joining trees for the 2 CLDGRs identified after 1000 kbp sliding window.

Supplemental Fig6. Neighbor-joining trees for 40 kbp windows surrounding the gene sh4 (chr4:34231186..34233221).

Supplemental Fig7. Neighbor-joining trees for three different window sizes flanking the domestication genes LABA1, PROG1, and sh4. First row, 50 kbp upstream and

downstream of gene; Second row, 250 kbp upstream and downstream of gene; Thrid row, 500 kbp upstream and downstream of gene.

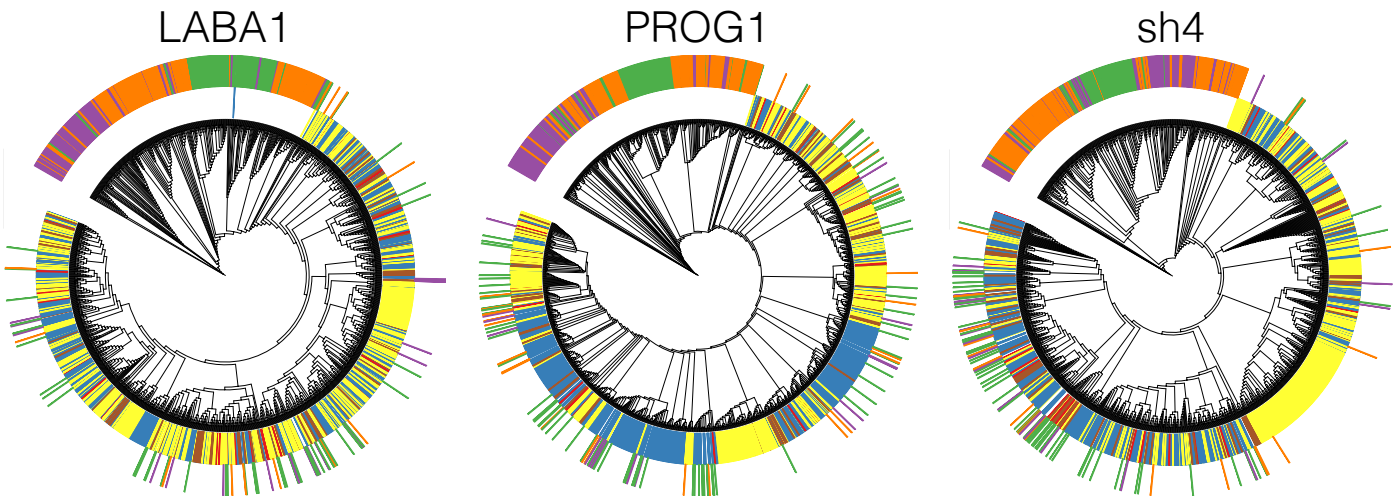

Fig1. Neighbor-joining tree for 20 kbp upstream and downstream of domestication genes LABA1, PROG1, and sh4. Inner circle of colors represent domesticated rice: red, aus; blue, indica; yellow, temperate japonica; brown, tropical japonica. Outer circle of colors represent wild rice that were designated by Huang et al.: green, Or-I; purple, Or-II; orange, Or-III.

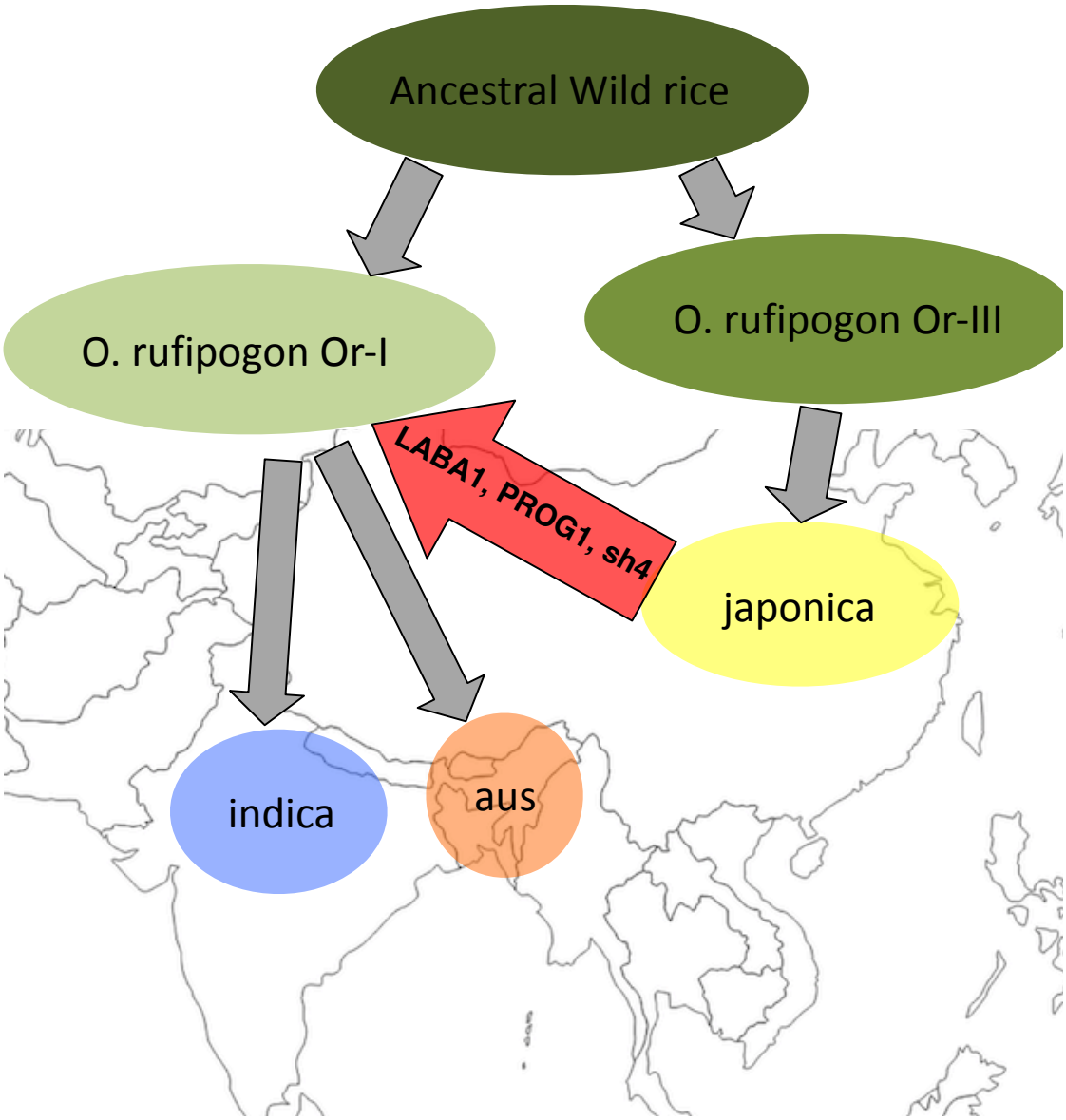

Fig2. Domestication scenario that led to Asian rice.