

1 **Asian wild rice is a hybrid swarm with extensive gene flow**
2 **and feralization from domesticated rice**

3 **Authors**

4 Hongru Wang^{1,2,5}, Filipe G. Vieira^{3,5}, Jacob E. Crawford⁴, Chengcai Chu^{1,6}, Rasmus
5 Nielsen^{4,6}.

6 ¹ State Key Laboratory of Plant Genomics, National Center for Plant Gene Research
7 (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of
8 Sciences, Beijing 100101, China.

9 ² College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100101,
10 China.

11 ³ Centre for GeoGenetics, University of Copenhagen, 1350 Copenhagen, Denmark.

12 ⁴ Department of Integrative Biology, University of California, Berkeley, California
13 94720 USA.

14 ⁵ These authors contribute equally to this work.

15 ⁶ Corresponding authors: R.N. (rasmus_nielsen@berkeley.edu) or C.C.
16 (ccchu@genetics.ac.cn).

1 **Abstract**

2 The domestication history of rice remains controversial with multiple studies reaching
3 different conclusions regarding its origin(s). These studies have generally assumed that
4 populations of living wild rice, *O. rufipogon*, are descendants of the ancestral
5 population that gave rise to domesticated rice, but relatively little attention has been
6 paid to the origins and history of wild rice itself. Here, we investigate the genetic
7 ancestry of wild rice by analyzing a diverse panel of rice genomes consisting of 203
8 domesticated and 435 wild rice accessions. We show that most modern wild rice is
9 heavily admixed with domesticated rice through both pollen and seed mediated gene
10 flow. In fact, much presumed wild rice may simply represent different stages of
11 feralized domesticated rice. In line with this hypothesis, many presumed wild rice
12 varieties show remnants of the effects of selective sweeps in previously identified
13 domestication genes, as well as evidence of recent selection in flowering genes possibly
14 associated with the feralization process. Furthermore, there is a distinct geographical
15 pattern of gene flow from *aus*, *indica* and *japonica* varieties into co-located wild rice.
16 We also show that admixture from *aus* and *indica* is more recent than gene flow from
17 *japonica*, possibly consistent with an earlier spread of *japonica* varieties. We argue that
18 wild rice populations should be considered a hybrid swarm, connected to domesticated
19 rice by continuous and extensive gene flow.

1 **Introduction**

2 Asian cultivated rice is one of the most ancient and widely consumed staple food crops.
3 Its domestication and cultivation contributed to the rise of agricultural civilization in
4 Asia. Rice is believed to have been domesticated ~9000 years ago from one of its
5 sympatric wild species, *O. rufipogon* (Oka 1988; Fuller et al. 2010). Molecular studies
6 have identified multiple varietal groups in cultivated rice, including two major ones:
7 *japonica* (*keng*) and *indica* (*hsien*) (Glaszmann 1987; Garris et al. 2005; Sweeney et al.
8 2007). *Indica* and *japonica* are highly differentiated and partially reproductively isolated
9 by a postzygotic barrier (Chang 2003). Despite numerous archaeological and genetic
10 studies on the history of rice domestication, no consensus has been reached on the
11 number of origins of different rice subgroups (Sang and Ge 2007; Huang et al. 2012b;
12 Civián et al. 2015). Some researchers argue for a single-origin model, which hypothesizes
13 that rice domestication was a single event followed by a post-domestication
14 diversification that created divergent subgroups. This model is supported by molecular
15 research on the domestication genes, *sh4* (Li et al. 2006; Lin et al. 2007) and *PROG1* (Jin
16 et al. 2008; Tan et al. 2008), which are responsible for two of the most critical
17 domestication traits in rice, non-shattering grains and erect growth, respectively. It has
18 been shown that different varietal groups of cultivated rice share identical sequences at
19 these two domestication genes (Lin et al. 2007; Tan et al. 2008). Additionally, multiple
20 studies that inferred the demographic histories of domesticated rice using independent
21 datasets favor the single-origin model (Gao and Innan 2008; Molina et al. 2011).
22 However, phylogenetic analyses using both nuclear and cytoplasmic DNA markers

1 consistently show that *indica* and *japonica* are each associated with different subgroups
2 of *O. rufipogon* (Cheng et al. 2003; Zhu and Ge 2005; Londo et al. 2006; Rakshit et al.
3 2007; Huang et al. 2012b; Civián et al. 2015). Some have used these results to argue that
4 rice domestication occurred more than once, and they attribute the sharing of key
5 domestication loci to gene flow after domestication (Londo et al. 2006; Rakshit et al.
6 2007; Sang and Ge 2007) or independent selection from standing ancestral variation
7 (Civián et al. 2015).

8 Despite these seemingly conflicting viewpoints, it is well accepted that
9 understanding the genetic variation of the primary gene pool, from which rice was
10 domesticated, is critical in studying rice domestication (Vaughan et al. 2008). The
11 primary gene pool is a concept used among plant breeders to define a set of
12 species/subspecies comprised of three components: the cultivated species, its wild
13 ancestor, and in many cases, its weedy counterparts (Harlan and DeWet 1971). Within
14 this gene pool, hybridization occurs easily and hybrid swarms are occasionally formed as
15 a result of crossing between the constituent components (Harlan 1992).

16 There has been extensive work on the population structure and genetic relatedness of
17 different subgroups within the rice primary gene pool, but incongruent phylogenetic
18 patterns have been observed. Wild rice has an annual ecotype, *O. nivara*, and its
19 phylogenetic position with the perennial type is inconclusive (Lu et al. 2002; Londo et
20 al. 2006), thus in this study we will not separate it from *O. rufipogon*. Early molecular
21 phylogenetic studies using isozymes identified two genetic groups of *O. rufipogon*, with
22 closer genetic affinity to *indica* and *japonica*, respectively (Second 1982). Multiple
23 other DNA studies also identified different genetic subgroups in wild rice populations

1 associated with different domesticated rice subgroups. In addition, they also identified
2 more ancestral genetic groups in wild rice population (Sun et al. 1996; Cheng et al. 2003;
3 Zhu and Ge 2005; Londo et al. 2006). Using genome-wide markers from 48 sequence
4 tagged sites, Huang (2012a) concluded that there were two distinct groups of wild rice,
5 one genetically related to *indica*, and one without particular relatedness to any
6 domesticated group. A whole genome sequencing study (Huang et al. 2012b)
7 categorized wild rice into three groups, two of which cluster with *japonica* and *indica*,
8 respectively, in the phylogeny constructed with genome-wide SNP markers. Recently, a
9 genotype-by-sequencing study on 286 diverse *O. rufipogon* species complex accessions
10 identified 6 subpopulations and suggested that there were gene flow between *O. rufipogon*
11 species complex and *O. sativa* (Kim et al. 2016). To account for the range of genetic
12 variation within the rice primary gene pool, various modeling analyses were also
13 performed (Caicedo et al. 2007; Zhu et al. 2007; Gao and Innan 2008; Molina et al.
14 2011). They consistently found that domesticated species had suffered severe
15 bottlenecks and that models of non-independent rice domestication provided better
16 explanation for the pattern of genetic variation within the gene pool (Gao and Innan
17 2008; Molina et al. 2011). Also, field sampling studies in different regions observed
18 ongoing gene flow among different components of the gene pool (Pusadee et al. 2013;
19 Pusadee et al. 2016) (also summarized in Oka 1988). Numerous concerns were raised
20 regarding the conservation of genetic diversity in wild rice populations, because
21 frequent gene flow from domesticated rice into wild rice populations could cause
22 genetic erosion and diversity loss in wild rice (Oka 1988). It is also well recognized that
23 gene flow between domesticated and wild rice populations is an important factor that

1 might confound phylogenetic studies and demographic history inferences on the rice
2 primary gene pool (Vaughan et al. 2008; Huang et al. 2012a). However, there is no study
3 to date that estimate the amount of gene flow from domesticated rice into natural wild
4 rice populations and/or determine the extent gene flow has shaped the genetic landscape
5 of wild rice.

6 **Results**

7 **Admixture analysis in the primary gene pool of Asian rice**

8 To investigate population structure and admixture patterns in the primary gene pool of
9 Asian rice, we combined whole genome sequencing data from 203 cultivated rice
10 varieties (Wang et al. 2016) and 435 accessions of *O. rufipogon* (Huang et al. 2012b).
11 The cultivated rice accessions were collected from 71 countries and were systematically
12 selected to be representative of rice diversity from more than 18,000 accessions in the
13 USDA rice germplasm seed bank (Agrama et al. 2009). The wild rice samples were
14 collected *in situ* in wild rice natural habitats (Supplemental Text S1) by scientists from
15 National Institute of Genetics in Japan (Morishima 2002).

16 We first estimated ancestry proportions for individuals using NGSadmix (Skotte et
17 al. 2013), which implements a clustering method similar to the one in the popular
18 program ADMIXTURE (Alexander et al. 2009), while incorporating uncertainty in the
19 genotype calls inherent in Next Generation Sequencing (NGS) data. We fit admixture
20 models by varying the number of presumed ancestral populations (K) from 2 to 15
21 (Supplemental Fig. S1-S3). Generally, the results fit those found in previous studies and
22 expected from prior knowledge of rice population genetics (see Supplemental Text S3).

1 However, accessions of domesticated rice are identified to have a small amount (<5%) of
2 wild rice ancestry, possibly reflecting introgression from wild rice, which was not
3 observed in previous studies (Wang et al. 2016). In the most remarkable case, one
4 domesticated rice accession (GSOR311586) was identified to be of 99% wild ancestry.
5 We conducted field observations which showed that this accession has shattering grains
6 and black-hull seeds with long awns that are hallmark phenotypes of wild rice
7 (Supplemental Fig. S6). PCR also confirmed that this accession contained a wild allele
8 of *sh4*. It is very likely that this is, in fact, a wild rice accession that was misidentified as
9 domesticated during germplasm collection.

10 In the wild rice population, however, we identified six subgroups (Fig. 1A), which
11 we denote as *Or-A*, *Or-B*, *Or-C*, *Or-D*, *Or-E* and *Or-F*, respectively, according to the
12 order of emergence when increasing *K* in the admixture analyses (Fig. 1A, Supplemental
13 Fig. S1). We also find good correspondence between subpopulations assigned here and
14 previously described genetic subgroups (Huang et al. 2012b) based on phylogenetic
15 analyses (Supplemental Table S1, Supplemental Fig. S7). Notably, a large proportion
16 (42%) of wild rice individuals seems to be substantially admixed and thus could not be
17 assigned to a single ancestry group, suggesting a complicated history of hybridization
18 and differentiation among wild rice. Among the identified clusters, four components
19 (*Or-A*, *Or-B*, *Or-C* and *Or-D*) are unique to wild rice. The *Or-A* component is the first to
20 emerge in wild rice when we increase *K* from 2 to 3. This component has a broad
21 geographic distribution, with highest ancestry proportions concentrated in the Oceanic
22 regions and lower ancestry proportions in West India and Sri Lanka (Fig. 1B). *Or-B*
23 emerged when five ancestral populations were included in the model. Geographically,

1 *Or-B* is found almost exclusively in China and it has been hypothesized that *Or-B* may
2 represent a wild ancestor of both *indica* and *japonica* since this population harbors
3 ancestral alleles at domestication-related loci shared by *indica* and *japonica* (Huang et al.
4 2012b). Adding one additional ancestral population to the model ($K = 6$) results in the
5 emergence of *Or-C*, which is found mostly in South and Southeast Asia and comprises
6 the majority of the wild rice genomes in West India and Sri Lanka population. *Or-D* is
7 found almost exclusively in the Indochina Peninsula, Bangladesh and East India (Fig.
8 1B). Intriguingly, for the last two subgroups (*Or-E* and *Or-F*), the major genetic
9 components are shared with *aus* and *indica*, respectively. To further characterize the
10 genetic relationships among subgroups in this gene pool, we carried out a principal
11 component analysis (PCA) (Fig. 1C, Supplemental Fig. S8). In the PCA space
12 constructed with the first two PCs, *japonica* forms an isolated cluster, while *indica* and
13 wild rice form a separate, more diffuse cluster. *Or-E* and *Or-F* co-localize with *aus* and
14 *indica* in the PCA plot. PC3 separates *indica* and *aus*, each forming a cluster. However,
15 *Or-E* and *Or-F* still cluster with *aus* and *indica*, respectively and the clustering pattern
16 still persists even at higher dimensions of the PCA space (Supplemental Fig. S8). This
17 suggests a very high degree of genetic relatedness between wild rice subgroups
18 *Or-E/Or-F* and the domesticated rice subgroups *aus/indica*, respectively.

19 **Gene flow between *O. rufipogon* and *O. sativa***

20 The exceptional genetic similarity between *Or-E/Or-F* and the corresponding
21 domesticated subgroups revealed by PCA and admixture analyses can be explained by
22 two possible hypotheses. First, *Or-E* and *Or-F* could be extant representatives of the
23 ancestral source population used in the domestication process and the genetic affinity

1 with *aus* and *indica* could result from standing ancestral polymorphism segregating in
2 these domesticated subgroups. Second, it could be caused by gene flow between
3 domesticated rice and the corresponding wild subgroups. To test these hypotheses, we
4 first conducted a correlation analysis between geographic distance and genetic distance
5 in all *sativa-rufipogon* pairs. We find a highly significant correlation ($\rho = 0.15$, $P <$
6 2.2×10^{-16} ; Supplemental Fig. S9), indicating that geographically close
7 *sativa-rufipogon* sample pairs tend to be more genetically related than expected. One
8 possible explanation for the correlation could be that the correlation is driven by shared
9 ancestral polymorphism between two species, but this is only tenable when there are
10 multiple geographic sites where rice was domesticated independently. Moreover, the
11 correlation is also present within smaller regions, such as India ($\rho = 0.18$, $P =$
12 1.1×10^{-12}) and Bangladesh ($\rho = 0.27$, $P = 3.1 \times 10^{-3}$) (Supplemental Fig. S10). An
13 explanation of the correlation based solely on multiple independent domestications
14 would further require multiple such domestication events within each country, with
15 local variability and structure preserved since the time of domestication — a very
16 unlikely scenario. A more tenable hypothesis is substantial local gene flow between
17 domesticated and wild rice in these regions.

18 To further examine the hypothesis of gene flow between domesticated and wild rice
19 populations, we analyzed the local ancestry at two known domestication-related genes,
20 *sh4* and *PROG1*, and asked whether the domesticated alleles are found in the wild
21 population or *vice versa*. These two genes have previously been shown to be
22 responsible for key morphological transitions from wild to domesticated rice: a
23 mutation (G→T) in the coding sequence of *sh4* causes reduced shattering of rice grains

1 (Li et al. 2006; Lin et al. 2007) and genetic variants in *PROG1* contribute to the
2 transition from prostrate to erect growth in domesticated rice (Jin et al. 2008; Tan et al.
3 2008). To our knowledge, these are the only two genes in the rice genome that control
4 critical traits distinguishing wild and domesticated rice, meanwhile all domesticated
5 rice share identical domesticated alleles at these loci (Lin et al. 2007; Tan et al. 2008),
6 despite enormous allelic diversity commonly observed at other genomic loci among
7 subgroups of domesticated rice. The domestication alleles confer traits strongly
8 preferred by humans, but they are presumably highly deleterious in the wild: the
9 non-shattering phenotype will increase the probability of herbivory of rice seeds and
10 erect growth will make rice plant more easily spotted and grazed by herbivores (Tan et
11 al. 2008). We first examined the haplotype content at the *sh4* locus using a clustering
12 approach (see Methods; Supplemental Fig. S11). Despite varying *K* from 2 to 5, all
13 domesticated rice accessions except the ‘misidentified’ GSOR311586 remain assigned
14 to a single component (Fig. 2A), suggesting they harbor closely related haplotypes, as
15 previously argued (Tan et al. 2008). Surprisingly, 94 samples (21.6% of all wild
16 samples) from the wild rice population are also consistently assigned to the
17 domesticated cluster, suggesting that they have the domesticated allele at *sh4*. Using a
18 PCR assay, we confirmed that all the assayed samples contained the derived allele (T)
19 at the functional SNP position, supporting the local ancestry assignment method as an
20 effective approach in discerning alleles (Supplemental Text S6, Supplemental Table
21 S1). Since we adopted the 95% ancestry cutoff for identifying domesticated allele (see
22 Methods), the result suggests that 94 may represent a conservative estimate of the
23 number of wild samples harboring the domesticated allele at *sh4*. This estimate is

1 consistent with a previous study which determined that ~27% of wild rice contain the
2 non-shattering allele at *sh4* (Zhu et al. 2012).

3 The observation that these ‘wild’ accessions contain the domestication allele at this
4 key domestication gene can be explained by two hypotheses: introgression from
5 domesticated rice or shared ancestral variation. In the first scenario, we would expect
6 that these individuals might share the signal of the domestication-related selective
7 sweep at the *sh4* locus and show a reduction in genetic distance to domesticated rice
8 relative to the distance between other wild rice and domesticated rice at this locus.
9 However, if these varieties harbor the domestication allele simply due to shared
10 ancestry from before domestication, they should not show the signal of a domestication
11 related selective sweep. To test this hypothesis, we examined local diversity at this
12 locus on wild rice carrying the domesticated allele of *sh4* (hereafter WRDS) and found
13 a four-fold reduction in relative nucleotide diversity across the 200kb region that
14 perfectly coincides with a similar diversity reduction in domesticated rice (Fig. 2B;
15 Supplemental Fig. S12). Also, Tajima’s *D* (Tajima 1989) is -2.63 in this region
16 (Supplemental Fig. S15), indicating an excess of rare alleles relative to equilibrium
17 expectations, which is also consistent with the scenario of a recent selective sweep. At
18 the sweep region, the genetic divergence between WRDS and domesticated rice drops
19 to 0 (Fig. 2C), indicating they share nearly identical haplotypes. However, the
20 divergence between WRDS-wild and cultivated-wild population is consistently high
21 and resemble background genomic levels (Fig. 2C; Supplemental Fig. S13). Taken
22 together, these results show that the genetic similarity between WRDS and
23 domesticated rice at *sh4* is caused by sharing of the same domestication allele

1 transferred by gene flow from domesticated into wild rice populations. The fact that
2 nominal wild rice has the shattering phenotype (Zhu et al. 2012), even when carrying
3 the domesticated *sh4* haplotype, suggests that one or more compensatory mechanisms
4 have evolved in wild rice populations in order to compensate for the extremely high
5 influx of the domesticated *sh4* allele through continuous gene flow from domesticated
6 rice.

7 When applying the same analysis to the *PROG1* locus, we identified 113 wild rice
8 accessions (26.0% of all wild rice samples) carrying the domestication allele *prog1*
9 (see Methods; Supplemental Fig. S14), and in these, the nucleotide diversity is reduced
10 and Tajima's *D* is -2.42 (Supplemental Fig. S15-S18), similar to the pattern observed
11 for *sh4*. A significant excess of these accessions ($n = 66$; $P < 0.01$, χ^2 test) also carry the
12 domestication allele at *sh4*. In total, 23 out of 25 accessions in subgroup *Or-E* carry
13 *prog1*, and 20 accessions carry the domesticated *sh4* allele. In the *Or-F* subgroup, 11
14 out of the 12 accessions carry the *prog1* allele, and all of them harbor the domestication
15 allele of *sh4*. When combined with the genome-wide admixture inferences, these
16 results strongly argue that the *Or-E* and *Or-F* subgroups either emerged as a result of
17 feralization of domesticated rice, or have received very high levels of gene flow, most
18 likely from the *aus* and *indica* varieties, respectively. Therefore, the shared ancestry of
19 *Or-E/Or-F* with domesticated subgroups observed under the $K = 9$ should be
20 interpreted as a consequence of extensive gene flow from domesticated rice. Moreover,
21 it is noteworthy that 104 accessions of other subgroups of wild rice harbor the
22 domesticated allele at either *PROG1* or *sh4*, resulting in a total 32% of annotated wild
23 rice accessions carrying domestication alleles, suggesting that gene flow/feralization is

1 substantial and not limited to only a subset of the wild rice subgroups (Supplemental
2 Fig. S19).

3 Morphologically, domesticated rice has closed floret, making crossed pollination
4 difficult and keeping them largely self-fertilized. Wild rice, however, typically has
5 open floret with exerted stigma, resulting in a higher rate of outcrossing, and this is
6 mirrored by a lower inbreeding coefficient estimates when compared with
7 domesticated rice (*t*-test, $P \ll 0.01$; Supplemental Fig. S20). Thus, morphological
8 differences predict an asymmetric pattern of gene flow, with its dominant direction
9 from domesticated into wild populations. Moreover, the census sizes of domesticated
10 rice populations are much larger relative to wild rice populations, this also suggest that
11 gene flow will predominantly be from domesticated to wild rice. Consistent with these
12 expectations, we find 207 domestication alleles at *sh4/PROG1* in wild rice population,
13 while the wild alleles in domesticated accessions rarely are observed ($n = 3$).
14 Genome-wide admixture analyses are also consistent with this hypothesis: varying K
15 from 2 to 9, we consistently observe domestication components in wild rice population,
16 but very little wild ancestry in domesticated rice (Supplemental Fig. S1). For the $K = 9$
17 model, 50% of wild rice have >10% domesticated ancestry (Supplemental Fig. S21).
18 Interestingly, wild rice populations are enriched with accessions containing 50-60% or
19 90-100% domesticated ancestry (Supplemental Fig. S21), possibly due to very recent
20 gene flow.

21 **Geographic pattern of gene-flow**

22 Monitoring the geographic pattern of the gene flow is important and may help guide the
23 protection of wild rice germplasm. Using introgression of *sh4* and *PROG1* as an

1 indicator, we found a significantly biased geographic distribution, and could reject the
2 hypothesis of an uniform amount of gene flow in all regions ($P < 0.01$, χ^2 test; Fig. 3A;
3 Supplemental Table S2). In Bangladesh, 75% of wild accessions have domesticated
4 alleles at one of the loci and 45% have domestication alleles at both loci; in East India,
5 we find 60.4% have one domesticated allele and 41.5% have both. These numbers are
6 much higher than the average level of 32.4% and 15.2%. By contrast, the northeast
7 ranges of wild rice habitat show little or no introgression at either locus, e.g., only
8 17.5% of wild rice in China and 6.7% in Laos harbored domesticated alleles. In
9 Indonesia, none of the rice accessions show evidence of introgression. Estimates of
10 domesticated ancestry proportions ($K = 9$) in the genome of wild rice show a pattern of
11 gene flow similar to that inferred using the two domestication loci (Fig. 3B,
12 Supplemental Fig. S22). Varieties from Bangladesh have the highest proportion of
13 domesticated rice ancestry among wild rice populations, with an estimate of 60%
14 (Supplemental Fig. S22). The high level of gene flow in this region is consistent with
15 field observations arguing that wild rice collected in this region may be heavily
16 admixed (Morishima 2002). The neighboring regions of East India and Malaysia also
17 have high estimates of 50% and 43% domesticated ancestry, respectively. By contrast,
18 wild accessions from regions such as China, Laos and Indonesia are relatively
19 unadmixed with domesticated admixture proportions as low as 8%, 10%, and 15%,
20 respectively (Supplemental Fig. S22).

21 When examining gene flow from the perspective of the donors, we find a great
22 difference in contribution from different domesticated rice subgroups, with 50%
23 *indica*, 46% *aus*, and only 4% *japonica*. There are several factors likely contributing to

1 this pattern. First, *indica* and *aus* varieties more readily shatter than *japonica* varieties
2 (Konishi et al. 2006; Vaughan et al. 2008), so they are more likely to contribute to
3 feralization. Second, wild rice is more likely to be sampled from areas in which
4 varieties from the *indica* and *aus* subgroups are cultivated. *Japonica* varieties are
5 mainly cultivated in the north, including North China, Korea and Japan (Gurdev
6 Khush, personal communication), where wild rice is rare and hence has not been
7 included in wild rice sampling efforts. The extensive overlap of *indica/aus* planting
8 area with wild rice habitat provides more opportunity for gene flow. In countries such
9 as Laos, Vietnam and Thailand, where gene flow mainly comes from the *indica*
10 subgroup (100%, 94%, 93%, respectively). However, in Bangladesh and India, gene
11 flow is mostly contributed by the *aus* subgroup (75% and 55%, respectively).
12 Interestingly, consistent with the broad distribution of *indica* cultivation, gene flow
13 from the *indica* subgroup is present in wild populations from most geographic regions
14 with an average of 50% of admixed samples carrying >5% *indica* ancestry (Fig. 3B). In
15 contrast, *aus* and *japonica* are planted in more restricted geographic regions and the
16 distribution of gene flow into wild populations reflects these geographic biases (Fig.
17 3B). The proportion of wild accessions with >5% *aus* ancestry is high in Bangladesh
18 and India (86% and 61%, respectively), which coincides well with the traditional
19 planting area of *aus* varieties (Glaszmann 1987; Khush 1997). A considerable
20 proportion of wild accessions from Malaysia and Sri Lanka (38% and 36%,
21 respectively) also carry substantial *aus* ancestry. Wild accessions with >5% *japonica*
22 ancestry are found in high proportions specifically in regions such as China, Burma and
23 Vietnam, representing the Northeast range of wild populations where the planting

1 region of *japonica* varieties and wild rice populations overlap.

2 To determine whether gene flow from each domesticated subgroup occurred during
3 the same or different time periods, we used local ancestry inference in admixed wild
4 rice to identify introgressed domesticated chromosomal segments. Since the
5 introgressed segments are broken into smaller segments by recombination over time,
6 the distribution of introgressed tract lengths is informative about the age of admixture
7 (Pool and Nielsen 2009; Moreno Estrada et al. 2013). The results of the local ancestry
8 inference are consistent with our global ancestry inferences (Supplemental Fig. S23)
9 and further support geographic biases in domesticated sources of gene flow (Fig. 4A).
10 We summarized the length distribution of introgressed tracts from each domesticated
11 subgroup and found that the length distribution of *japonica* haplotypes is enriched for
12 smaller segments with an average of 8cM (Fig. 4B). The distribution of *japonica*
13 haplotypes is significantly shorter than that of both *indica* (*t*-test, $P < 1 \times 10^{-8}$) and *aus*
14 (*t*-test, $P < 1 \times 10^{-8}$), which have average haplotype length of 27cM and 18cM,
15 respectively. This result indicates that the gene flow from *japonica* to wild rice is older
16 than that of *aus* and *indica*.

17 **Feralization plays an important role in gene flow**

18 The gene flow from *O. sativa* to *O. rufipogon* may follow two different evolutionary
19 pathways: by pollen dissemination or seed dispersal. If seed spillage were involved, we
20 would expect to find cytoplasmic genomes with domesticated rice haplotype in wild
21 populations. In this study, we took advantage of the high copy number of the
22 chloroplast genome, providing an average of 200× sequencing coverage for each
23 accession in the sequencing data (See Methods), to obtain highly accurate haplotype

1 information. We first estimated a Maximum Likelihood (ML) phylogenetic tree of the
2 chloroplast haplotypes (Supplemental Fig. S24). The domesticated rice samples were
3 found in two clusters, corresponding to the *indica* and *japonica* subgroups.
4 Interestingly, many ‘wild’ rice chloroplast genomes were nested within the
5 domesticated rice clusters. To further quantify the number of ‘wild’ rice accessions that
6 are closely related to domesticated rice chloroplast haplotype, we constructed a
7 haplotype network using common polymorphic sites across rice chloroplast genomes
8 (Supplemental Text S4), which summarizes all major chloroplast haplotypes in the
9 primary gene pool of rice and the phylogeny among them (Fig. 5). Surprisingly, we
10 found 98 accessions (28.8% of 340) of wild rice with identical chloroplast haplotypes
11 to those of domesticated rice. For both *Or-E* and *Or-F*, which we have shown to carry
12 domesticated nuclear ancestry, an excess of accessions harbor domesticated chloroplast
13 haplotypes as well (17 out of 24 for *Or-E*, $P = 0.01$; 8 out of 12 for *Or-F*, $P = 0.06$, χ^2
14 test). This further supports that these accessions in fact are established by seed
15 dispersal, i.e. feral rice. These results suggest an evolutionary scenario that includes
16 ancient feralization events followed by subsequent backcrossing with wild rice
17 populations. In line with the analysis at domestication loci, gene flow from
18 domesticated rice is not limited to just *Or-E* and *Or-F* subgroups, because domesticated
19 chloroplast genomes are carried by other groups of wild rice as well (Supplemental Fig.
20 S25).

21 **Selection and adaptation in feral rice**

22 The exceptional relatedness of both nuclear and chloroplast genomes between *Or-E*
23 and *aus* indicates that *Or-E* might have arisen from *aus* varieties in the very recent past,

1 and then diverged during adaptation to the local wild environments. Thus a comparison
2 of *Or-E* and *aus* genomes provides a unique opportunity to investigate the genetic basis
3 of plant feralization. In order to identify loci that might have been differentially
4 selected between domesticated and feral rice, we first scanned the genome using F_{ST} to
5 identify highly differentiated genes between *Or-E* and *aus*. We performed gene
6 ontology (GO) enrichment analysis on genes with F_{ST} values ranking in the top 5% of
7 the empirical distribution. The top enriched GO terms are mostly high-hierarchy terms
8 which are too general to provide any specific biological hints (Supplemental Table S3).
9 But among the top enriched GO terms that refer to explicit biological functions, abiotic
10 and biotic resistance terms including response to fungus ($P = 8.1 \times 10^{-7}$), bacterium ($P =$
11 1.5×10^{-9}), salt ($P = 8.4 \times 10^{-11}$), cold ($P = 4.8 \times 10^{-6}$) and wounding ($P = 3.6 \times 10^{-8}$), are
12 prominently enriched. This suggests that rice might have faced different biotic and
13 abiotic selection pressures under domestic and wild conditions. Interestingly, the GO
14 term ‘long-day photoperiodism’ is also enriched, an enrichment which persists even if
15 the GO analysis is limited to genes with top 1% F_{ST} values, indicating that genes
16 underlying flowering time in long-day condition are among the most differentiated
17 genes between *Or-E* and *aus*. We subsequently identified genes under selection in *Or-E*
18 that may have been targeted by natural selection during the feralization process.
19 Interestingly, *HDI*, a gene underlying major quantitative trait locus (QTL) for
20 photoperiod-dependent flowering (Yano et al. 2000), is among those with the most
21 dramatic diversity reduction across *Or-E* rice genomes, ranking in the top 0.3% of
22 diversity-reduction genes across *Or-E* rice genomes, suggesting strong selection on
23 this locus in the *Or-E* population. A comparison of the haplotypes of *Or-E* and *aus* at

1 this gene identified the most differentiated SNP as a non-synonymous polymorphism
2 (G/A, G387S) that is fixed for G in *Or-E* but has low allele frequency in *aus* (13.3%), a
3 potential candidate causal mutation. It is likely that *HDI* is a target of selection for rice
4 feralization and that the non-synonymous mutation has contributed to the flowering
5 time adaptation of rice in the wild habitat.

6 **Implications for rice domestication**

7 The high level of gene flow between wild and domesticated rice has consequences for
8 our understanding of the process of rice domestication. To illustrate this, we estimated
9 admixture graphs of geographically defined wild rice and major groups of
10 domesticated rice using TreeMix (Pickrell and Pritchard 2012), which uses a Maximum
11 Likelihood (ML) method based on a Gaussian model of allele frequency change. We
12 divided wild rice into five regional populations based on geographic characteristics of
13 the wild rice area and potential boundaries between subgroups (Methods, Figure 1B).
14 Four major subgroups of domesticated rice were also included. Though the topology of
15 the ML trees changes depending on the number of migration events (m) allowed in the
16 model (Supplemental Fig. S26), certain patterns persist and are robust towards
17 assumptions regarding m . First, the domesticated rice subgroups consistently show
18 evidence of more genetic drift, likely because they underwent strong bottlenecks
19 caused by the domestication process and by artificial selection. The *japonica*
20 subgroups have exceptionally long branches consistent with the previously reported
21 much stronger bottleneck in their domestication history (Caicedo et al. 2007; Zhu et al.
22 2007; Gao and Innan 2008). Wild rice populations in the Ganges Basin (GBW)
23 consistently form a clade with *indica* and *aus* (Supplemental Fig. S26). Two

1 hypotheses could explain this pattern: (1) *indica* and *aus* were domesticated from the
2 GBW very recently, or (2), as suggested by the previous analyses in this manuscript,
3 the GBW populations are a product of feralization from domesticated *aus* and *indica*
4 rice. Similarly, *temperate* and *tropical japonica* forms a clade with Chinese rice when
5 assuming no migration.

6 Allowing just one migration event (Supplemental Fig. S26; $m = 1$), we observe an
7 admixture event from *indica* into the Indochina wild rice population (ICW)
8 contributing (46%) of the DNA in Indochina. This is consistent with the results that a
9 substantial amount of *indica* ancestry is observed in ICW (Figure 1B, Figure 4A).
10 Allowing two admixture events (Supplemental Fig. S26; $m = 2$), a substantial amount
11 of gene-flow from Indochina to China is observed. This is possibly a consequence of
12 Chinese wild rice being admixed between original wild rice and domesticated rice.
13 This is supported by the fact that when $m = 3$, wild Chinese rice groups with wild rice in
14 Indochina and the Archipelago, but with substantial gene-flow (49%) from the ancestor
15 of *japonica* domesticated rice (Fig. 6). Likely, the true wild ancestor of *japonica* rice is
16 not represented in the sample by any current wild descendant population. The *Or-B*
17 component found in the China may not be an ‘authentic’ wild component, but rather it
18 is a product of admixture between wild rice and ancient *japonica*. The wild rice
19 ancestral to the domesticated *japonica* may be, in fact, already extinct. For models with
20 $m = 3$, we observed an admixture event, with a proportion of 19%, from *aus* to *tropical*
21 *japonica* (Fig. 6), indicating substantial genetic ancestry shared between these two
22 subgroups. We consistently observe *japonica* sharing high residuals with *aus/indica*
23 (Supplemental Fig. S26), which likely reflects that they share many genomic

1 components caused by hybridization in their domestication and breeding history.

2 **Discussion**

3 Elucidating the pattern of gene flow among wild and domesticated rice is important for
4 understanding the history of rice domestication. Multiple studies argued for the
5 independent domestication of rice based on reciprocal monophyly of *indica* and
6 *japonica* when using different nuclear DNA markers in independent rice collections
7 (Cheng et al. 2003; Zhu and Ge 2005; Rakshit et al. 2007; Civián et al. 2015). In
8 contrast, treating *O. rufipogon* as a single homogenous group in an analyses of
9 divergence times and population trees, Molina *et al.* (2011) argued for a single
10 domestication event. Based on comparisons to wild rice samples, Huang *et al.* (2012b)
11 similarly argued that rice domestication originated in South China. Recently, Civián *et*
12 *al.* (2015) argued that the *aus* group had been independently domesticated in the
13 Ganges Basin area. In this study, we showed that there is extensive, continuous gene
14 flow from domesticated rice into wild rice populations. It suggests that the patterns
15 described in previous studies are likely caused by gene flow from domesticated rice
16 into wild rice populations after rice domestication. We show that wild rice in the
17 Ganges Basin is likely feral rice, recently diverged from domesticated rice, and
18 Chinese wild rice has received extensive gene flow from an ancient *japonica*
19 population. Furthermore, the *indica* and *aus* groups are always sister-groups,
20 suggesting a single domestication event for these two groups. TreeMix results are
21 largely compatible with a dual origin of domestication given the deep divergence
22 observed between *indica* and *japonica* subgroups. The divergence even spans the

1 diversity of present-day Asian wild rice, but we caution that current wild rice samples
2 may be biased due to incomplete sampling or loss of ‘authentic’ wild rice sample in
3 germplasm centers during preservation. We cannot exclude the possibility of a single
4 domestication hypothesis because the deep divergence could also be caused by
5 substantial independent gene flow from other wild rice species into different
6 domesticated rice subgroups, which is practiced in rice breeding (Brar and Khush
7 1997). The single domestication hypothesis would require either (1) extensive
8 gene-flow from wild rice into the *indica/aus* subgroups so that their genomes now are
9 dominated by gene-flow from wild rice, combined with a subsequent loss (or lack of
10 representation) of true “ancestral” wild rice. Alternatively, (2) a single domestication
11 hypothesis could also be compatible with the data if all wild rice populations
12 represented in the panel are dominated by gene-flow from local domesticated rice
13 occurring continuously over the past ~10,000 years. However, the dual domestication
14 model is arguably a simpler scenario.

15 Rice was introduced into the United States less than 400 years ago, and rice
16 cultivation was not widely expanded until the 1750s (Dethloff 2003). However, weedy
17 rice is now common in rice growing regions in the United States and is one of the major
18 weeds limiting rice production (Ziska et al. 2015). Genetic analysis has shown that
19 American weedy rice population arose from *indica* and *aus* varieties independently
20 (Londo and Schaal 2007). These observations indicate that rice could revert to the wild
21 state in domestication traits frequently. Weedy rice is a conspecific form of cultivated
22 rice, while displaying distinguishing features including shattering grains and strong
23 seed dormancy typical of wild rice (Ferrero 2003; Song et al. 2014). The shattering and

1 seed dormancy phenotypes acquired in weedy rice are presumably adaptive in wild
2 conditions, potentially further facilitating feralization. A crop-weed-wild complex is
3 found throughout regions where wild and cultivated rice overlap, and gene flow
4 between components within the species complex is frequently observed (Ellstrand et al.
5 2013; Pusadee et al. 2013; Song et al. 2014; Pusadee et al. 2016). In Asia, rice
6 cultivation has been performed for thousands of years, and rice feralization has likely
7 happened throughout this period as well (Vaughan et al. 2005). In fact, much presumed
8 wild rice in many parts of Asia may possibly be descendants of ancient
9 feralization/hybridization events. Wild and weedy rice found all over the world might
10 simply represent different stages of the feralization process. It is even possible that
11 what we today characterize as wild rice in Asia, may largely be feral rice that has
12 undergone thousands of generations of natural selection in the wild, and that the
13 original species from which *O. sativa* was domesticated is either extinct or has been
14 almost entirely overwhelmed by the massive amounts of gene flow from domesticated
15 rice. *O. rufipogon* may then represent a nominal species created by human
16 domestication and subsequent feralization/hybridization.

1 **Materials and Methods**

2 **Genomic data acquisition**

3 The genomic data of wild rice was downloaded from the European Nucleotide Archive
4 under the accession number ERP001143 (Supplemental Text S2). Domesticated rice
5 data was downloaded from the NCBI BioProject Repository (project number:
6 PRJNA301661).

7 **Short reads mapping**

8 Reads were mapped to the rice genome (IRGSP-1.0) (Kawahara et al. 2013) with BWA
9 (version 0.7.0) (Li and Durbin 2009) and the mapping was further improved with
10 Stampy (version 1.0.20) (Lunter and Goodson 2011). PCR duplicates were removed by
11 “rmdup” in SAMtools (version 0.17) (Li et al. 2009). We realigned reads at gapped
12 regions with GATK (version 2.6) (DePristo et al. 2011).

13 **Population structure, phylogeny, network, and TreeMix analyses**

14 We estimated genotype likelihoods of populations with the “-GL” option in ANGSD
15 (version 0.542) (Korneliussen et al. 2014). Inbreeding coefficients for each individual
16 were calculated using a probabilistic framework implemented in ngsF (Vieira et al.
17 2013). The variability and allele frequency of each genomic site was estimated by
18 ANGSD using the “-doMaf” command. Variable sites were extracted and used for
19 further analyses. A genotype likelihoods-based method, implemented in NGSadmix
20 (Skotte et al. 2013), was used for global ancestry inference. The analysis was
21 conducted on the combined population including 203 domesticated and 435 wild rice

1 accessions. We randomly picked one variable site for every 5kb genomic region from
2 variable sites to reduce effects of linkage disequilibrium. In total, 60,722 evenly
3 distributed markers were used. With these markers, we successively tested 14
4 clustering models in the population with K (presumed cluster number) ranging from 2
5 to 15. For each K , we run 200 independent replicate optimizations, picked the
6 clustering model with highest log likelihood value and the corresponding log
7 likelihoods are shown in Supplemental Fig. S3. PCA was performed with the same
8 genotype likelihoods dataset using ngsCovar from the ngsTools package (Fumagalli et
9 al. 2014). All plots were generated with R (version 3.0.2) (R Core Team 2016). We
10 estimated admixture trees, phylogenies, and haplotype networks using standard
11 methods explained in Supplemental Text S4 and S5.

12 **Introgression analyses at two domestication loci**

13 To identify domestication haplotypes at the *sh4* locus in wild rice, we inferred local
14 ancestry in a 10kb region centered on *sh4*. Using genotype likelihoods, we ran
15 NGSadmix for varying values of K , and domesticated rice accessions were consistently
16 assigned to one component from $K = 2$ to $K = 5$ except for the mis-identified sample,
17 GSOR311586. At $K = 6$, the domesticated rice population splits into two major
18 components, which conflicted with prior knowledge that there is one haplotype at this
19 locus in the domesticated rice population (Li et al. 2006; Lin et al. 2007), suggesting
20 that $K = 6$ model is over-fitting. To further investigate this issue, we randomly sampled
21 3 samples from each domesticated rice population assigned to different ancestries
22 under the $K = 6$ model and PCR amplified the *sh4* locus in these accessions. They
23 turned out in all cases to harbor the domesticated allele at the causal variant site.

1 Consequently, we proceeded to use the $K = 5$ model for allele identification of the
2 domestication haplotype. Wild rice samples with at least 95% domesticated ancestry at
3 the locus were inferred to carry the domesticated allele. For the *PROG1* locus, there is
4 also a strong selective sweep (He et al. 2011) and all domesticated rice share identical
5 haplotype in this region (Tan et al. 2008). We thus applied the same procedures to
6 identify introgression at this locus. For both genes, we confirmed that the domesticated
7 haplotypes identified from wild rice population contained the domesticated allele at the
8 functional SNP site through PCR amplification (Supplemental Text S6, Supplemental
9 Table S1).

10 Tajima's D and θ_π statistics were calculated under a probabilistic framework
11 designed for low-coverage data (Korneliussen et al. 2013). The methods are
12 implemented in ANGSD and can be invoked by parameter “-doThetas”. d_{XY} between
13 populations was calculated in 10kb non-overlapping windows. For each window, d_{XY}
14 values were calculated for all paired polymorphic sites and then averaged over sites.
15 For each polymorphic site, the allele frequencies in population X and Y are denoted as
16 p_X and q_X , and p_Y and q_Y , respectively and d_{XY} for a site is calculated as $d_{XY} = p_X q_Y +$
17 $p_Y q_X$.

18 **Genotype calling and local ancestry inference**

19 To infer the local ancestry in admixed wild rice genomes, we first set up reference wild,
20 *temperate japonica*, *aus* and *indica* panels. Under the $K = 9$ admixture model, wild rice
21 individuals whose ancestry were inferred to be $\geq 80\%$ from one of four wild rice
22 specific components, and which contained neither *prog1* nor domesticated allele of

1 *sh4*, were used in the reference wild rice panel. Domesticated rice with $\geq 80\%$ inferred
2 ancestry from one of *indica*, *aus* or *temperate japonica* was used as reference *indica*,
3 *aus* or *temperate japonica* panel, respectively. Wild rice samples with combined
4 ancestry of *indica*, *aus* and *temperate japonica* $\geq 20\%$ were included as admixed
5 accessions. Local ancestry assignment was performed on admixed rice genomes with
6 RFMix (Maples et al. 2013). Since this algorithm uses haplotypes as input, we called
7 genotypes in both admixed and reference panel samples with ANGSD (Korneliussen et
8 al. 2014). Imputation and phasing was further performed on the datasets with BEAGLE
9 (version 3.3.2) (Browning and Browning 2007).

10 **Selection detection in feral rice**

11 We calculated F_{ST} between the *Or-E* and *aus* populations for all rice genes using
12 ngsTools (Fumagalli et al. 2014) which calculates F_{ST} using genotype likelihoods,
13 taking genotyping uncertainty into account. We performed GO analysis on the ranked
14 gene list based on the F_{ST} values: GO annotation of all rice genes was downloaded from
15 Gramene (<http://www.gramene.org/>, release 49), the enrichment of each GO was tested
16 using Fisher's exact test, corrected for multiple tests using a Bonferroni correction. The
17 significantly enriched GO terms for the top 5% F_{ST} genes can be found on
18 Supplemental Table S3. Nucleotide diversity reduction in both *aus* and *Or-E* genomes
19 was estimated by comparing with diversity in wild rice populations. The diversity for
20 each population was estimated using the “-doThetas” command in ANGSD
21 (Korneliussen et al. 2014).

1 **Acknowledgements**

2 This work was supported by grants from Chinese Academy of Sciences (XDA08010401)
3 and National Natural Science Foundation of China (31430063). J.C. was supported by a
4 National Institutes of Health Ruth L. Kirschstein National Research Service Award. We
5 would like to thank Qi Feng and Bin Han from National Center for Gene Research,
6 Chinese Academy of Sciences for providing wild rice DNA for PCR assay.

7 **Authors' contributions**

8 R.N. and C.C. supervised the project. H.W. F.G.V. and J. C. performed the analysis. H.W.
9 and R.N. wrote the manuscript with critical input from all the authors. All authors read
10 and approved the final manuscript.

11 **Competing interests**

12 The authors declare that they have no competing interest.

13 **Supplemental files**

14 All supplemental figures, tables and data can be found on the attached files.

15 **Data access**

16 PCR amplified sequences are deposited in NCBI under accession number KY701787 -
17 KY701861 (for *sh4*) and KY701862 - KY701970 (for *PROG1*), respectively.

1

1 **Figure Legends**

2 **Figure 1.**

3 **Population structure of the rice primary gene pool including *O. sativa* and *O.***
4 ***rufipogon*.** (A) Clustering using NGSadmix assuming $K = 2$ and $K = 9$. At $K = 2$, the
5 samples are divided into *indica* and *japonica* components. At $K = 9$, five subgroups of
6 domesticated rice are recovered, while four unique components of wild rice are
7 identified. The color bars beneath the clusters denote the subgroup assignments. The
8 red arrow points to the mis-identified domesticated accession (GSOR311586) which
9 was confirmed to have wild rice ancestry. The abbreviations of subgroups in cultivated
10 rice are as follows: ADM for *admixture*; IND for *indica*; AUS for *aus*; ARO for
11 *aromatic*; TRJ for *tropical japonica* and TEJ for *temperate japonica*. (B) Geographic
12 distribution of rice samples. South and Southeast Asia, which are the major habitats for
13 wild rice and also major rice cultivation areas, are shown on the map. The area was
14 divided into five regions: (1) South Asia; (2) Ganges Basin; (3) Indochina Peninsula;
15 (4) China; (5) Archipelago countries. The color code of the bar beneath the clustering
16 plot indicates cultivated (green) and wild (black) rice. The abbreviations at the bottom
17 left are as follows: AS for Asia, all rice samples from Asia but not shown on the map are
18 included in this category; AM for America; AF for Africa; EU for Europe. (C) PCA of
19 the combined population with wild (triangle) and cultivated (dot) rice samples. The
20 abbreviation codes are the same as those in (A).

21

1 **Figure 2.**

2 ***sh4* haplotypes in wild and domesticated rice populations.** (A) Local ancestry
3 inference at *sh4* locus for $K = 5$. The bar at the bottom denotes *O. sativa* (green) and *O.*
4 *rufipogon* (black) accessions, respectively. (B) Diversity reduction at the selective
5 sweep region of *sh4*. The y-axis shows the ratio of pairwise differences estimator (π) of
6 nucleotide diversity between populations. (C) d_{XY} values between populations at the
7 selective sweep region of *sh4*. Grey line indicates the *sh4* gene region in (B) and (C).

8

9 **Figure 3.**

10 **Geographic and sub-specific pattern of gene flow.** (A) Geographic distribution of
11 domesticated alleles introgression at *PROG1* and *sh4* loci. Each pie chart represents the
12 wild rice population of a region, and the area is proportional to its sample size. Each
13 chart was divided into four categories according to the haplotype information at the two
14 domestication loci: the domestication *sh4* allele only (yellow), the wild *prog1* allele
15 only (blue), domestication alleles for both *sh4* and *PROG1* (purple), and wild *sh4* and
16 *PROG1* alleles (black). Regions with less than 10 samples are not shown. (B)
17 Geographic distribution of gene flow from *indica*, *japonica* and *aus*. The proportions of
18 admixed wild accessions with $>5\%$ ancestry of a certain subspecies in different regions
19 are plotted. Wild accessions with *indica* have a pandemic distribution across rice
20 cultivation regions, while accessions with *japonica* ancestry are endemic to regions
21 including China, Burma and Vietnam. Accessions with *aus* ancestry are mainly found in
22 Ganges Basin region, Sri Lanka and Malaysia.

1 **Figure 4.**

2 **Distribution of chromosomal segments with domesticated ancestry in ‘wild’ rice.**

3 (A) Ancestry assignment for ‘wild’ rice from different regions. Each row represents a
4 chromosome of one individual. Data for chromosome 7 is presented. Bar colors indicate
5 ancestry as follows: wild rice is grey, *aus* is coral, *indica* is blue, and *japonica* is green.
6 Introgression in Bangladesh is dominated by *aus*, Thailand is dominated by *indica* .
7 Chinese ‘wild’ rice harbors tracts of both *indica* and *japonica* ancestry. (B) Cumulative
8 distribution of ancestry tract lengths from different subgroups of domesticated rice.

9

10 **Figure 5.**

11 **Chloroplast haplotype network among the 28 common haplotypes in rice primary**

12 **gene pool.** The haplotypes were defined using 74 common SNPs from the rice
13 chloroplast genome. Each pie chart in the figure represents one haplotype, and it was
14 further divided according to the subgroup information of the samples. All domesticated
15 rice samples were colored in green, and wild rice samples were divided into seven
16 subgroups. The root of the haplotype network was inferred using the chloroplast
17 genome of *O. meridionalis*, and is indicated by a red arrow in the figure. The length of
18 lines connecting pie charts is proportional to the pairwise distance between haplotypes.
19 The areas of the pie charts are proportional to the number of samples with the
20 haplotype.

21 **Figure 6.**

22 **Maximum-likelihood admixture graph on the primary gene pool of Asian**

1 **domesticated rice.** The wild rice (*O. rufipogon*) population was divided into five
2 geographic populations (see Methods). The abbreviations for the major domesticated
3 rice subgroups are the same as in Fig. 1. African wild rice, *O. barthii* (BAR), was used
4 to root the tree. The bootstrap values on the tree are based on 1,000 replicates. Arrows
5 on the graph represents admixture events between different rice populations.

6

1 **References**

- 2 Agrama H, Yan W, Lee F, Fjellstrom R, Chen MH, Jia M, McClung A. 2009. Genetic
3 assessment of a mini-core subset developed from the USDA rice genebank.
4 *Crop Sci* **49**(4): 1336-1346.
- 5 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry
6 in unrelated individuals. *Genome Res* **19**(9): 1655-1664.
- 7 Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and
8 missing-data inference for whole-genome association studies by use of localized
9 haplotype clustering. *Am J Hum Genet* **81**(5): 1084-1097.
- 10 Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL,
11 Polato NR, Olsen KM, Nielsen R, McCouch SR et al. 2007. Genome-wide
12 patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* **3**(9):
13 1745-1756.
- 14 Chang TT. 2003. Origin, domestication, and diversification. In *Rice: Origin, History,*
15 *Technology, and Production*, (ed. CW Smith, RH Dilday), pp. 3-25.
- 16 Cheng C, Motohashi R, Tsuchimoto S, Fukuta Y, Ohtsubo H, Ohtsubo E. 2003.
17 Polyphyletic origin of cultivated rice: based on the interspersed pattern of
18 SINEs. *Mol Biol Evol* **20**(1): 67-75.
- 19 Civáň P, Craig H, Cox CJ, Brown TA. 2015. Three geographically separate
20 domestications of Asian rice. *Nat Plants* **1**: 15164.
- 21 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA,
22 del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation
23 discovery and genotyping using next-generation DNA sequencing data. *Nat*
24 *Genet* **43**(5): 491-498.
- 25 Dethloff HC. 2003. American rice industry: historical overview of production and
26 marketing. In *Rice: Origin, History, Technology, and Production*, (ed. CW
27 Smith, RH Dilday), pp. 67-86.
- 28 Ellstrand NC, Meirmans P, Rong J, Bartsch D, Ghosh A, de Jong TJ, Haccou P, Lu BR,
29 Snow AA, Neal Stewart Jr C et al. 2013. Introgression of crop alleles into wild
30 or weedy populations. *Annu Rev Ecol Evol Syst* **44**: 325-345.
- 31 Ferrero A. 2003. Weedy rice, biological features and control. *FAO Plant Production*
32 *and Protection Paper*: 89-107.
- 33 Fuller DQ, Sato YI, Castillo C, Qin L, Weisskopf AR, Kingwell Banham EJ, Song J,
34 Ahn SM, Van Etten J. 2010. Consilience of genetics and archaeobotany in the
35 entangled history of rice. *Archaeol Anthropol Sci* **2**(2): 115-131.
- 36 Fumagalli M, Vieira FG, Linderth T, Nielsen R. 2014. ngsTools: methods for
37 population genetics analyses from next-generation sequencing data.
38 *Bioinformatics* **30**(10): 1486-1487.
- 39 Gao LZ, Innan H. 2008. Nonindependent domestication of the two rice subspecies,
40 *Oryza sativa* ssp. *indica* and ssp. *japonica*, demonstrated by multilocus
41 microsatellites. *Genetics* **179**(2): 965-976.
- 42 Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S. 2005. Genetic structure and
43 diversity in *Oryza sativa* L. *Genetics* **169**(3): 1631-1638.
- 44 Glaszmann JC. 1987. Isozymes and classification of Asian rice varieties. *Theor Appl*

- 1 *Genet* **74**(1): 21-30.
- 2 Harlan JR. 1992. *Crop and Man*. American Society of Agronomy, Crop Science Society
3 of America, Madison, Wisconsin.
- 4 Harlan JR, DeWet MJ. 1971. Toward a rational classification of cultivated plants.
5 *Taxon* **20**(4): 509-517.
- 6 He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu CI, Shi S.
7 2011. Two evolutionary histories in the genome of rice: the roles of
8 domestication genes. *PLoS Genet* **7**(6): e1002100.
- 9 Huang P, Molina J, Flowers JM, Rubinstein S, Jackson SA, Purugganan MD, Schaal
10 BA. 2012a. Phylogeography of Asian wild rice, *Oryza rufipogon*: a
11 genome-wide view. *Mol Ecol* **21**(18): 4593-4604.
- 12 Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W et
13 al. 2012b. A map of rice genome variation reveals the origin of cultivated rice.
14 *Nature* **490**(7421): 497-501.
- 15 Jin J, Huang W, Gao JP, Yang J, Shi M, Zhu MZ, Luo D, Lin HX. 2008. Genetic control
16 of rice plant architecture under domestication. *Nat Genet* **40**(11): 1365-1369.
- 17 Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S,
18 Schwartz DC, Tanaka T, Wu J, Zhou S et al. 2013. Improvement of the *Oryza*
19 *sativa* Nipponbare reference genome using next generation sequence and optical
20 map data. *Rice* **6**(1): 4.
- 21 Khush GS. 1997. Origin, dispersal, cultivation and variation of rice. In *Oryza: From*
22 *Molecule to Plant*, pp. 25-34. Springer.
- 23 Kim H, Jung J, Singh N, Greenberg A, Doyle JJ, Tyagi W, Chung J-W, Kimball J,
24 Hamilton RS, McCouch SR. 2016. Population dynamics among six major
25 groups of the *Oryza rufipogon* species complex, wild relative of cultivated
26 Asian rice. *Rice* **9**(1): 56.
- 27 Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. 2006. An SNP
28 caused loss of seed shattering during rice domestication. *Science* **312**(5778):
29 1392-1396.
- 30 Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation
31 sequencing data. *BMC Bioinformatics* **15**(1): 356.
- 32 Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. 2013. Calculation of Tajima's *D*
33 and other neutrality test statistics from low depth next-generation sequencing
34 data. *BMC Bioinformatics* **14**: 289.
- 35 Li C, Zhou A, Sang T. 2006. Rice domestication by reducing shattering. *Science*
36 **311**(5769): 1936-1939.
- 37 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
38 transform. *Bioinformatics* **25**(14): 1754-1760.
- 39 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
40 Durbin R, Genome Project Data Processing S. 2009. The Sequence
41 Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
- 42 Lin Z, Griffith ME, Li X, Zhu Z, Tan L, Fu Y, Zhang W, Wang X, Xie D, Sun C. 2007.
43 Origin of seed shattering in rice (*Oryza sativa* L.). *Planta* **226**(1): 11-20.
- 44 Londo J, Schaal B. 2007. Origins and population genetics of weedy red rice in the USA.
45 *Mol Ecol* **16**(21): 4523-4535.
- 46 Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA. 2006. Phylogeography of

- 1 Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications
2 of cultivated rice, *Oryza sativa*. *Proc Natl Acad Sci USA* **103**(25): 9578-9583.
- 3 Lu BR, Zheng K, Qian H, Zhuang J. 2002. Genetic differentiation of wild relatives of
4 rice as assessed by RFLP analysis. *Theor Appl Genet* **106**(1): 101-106.
- 5 Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast
6 mapping of Illumina sequence reads. *Genome Res* **21**(6): 936-939.
- 7 Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative
8 modeling approach for rapid and robust local-ancestry inference. *Am J Hum*
9 *Genet* **93**(2): 278-288.
- 10 Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, Huang P, Jackson
11 S, Schaal BA, Bustamante CD et al. 2011. Molecular evidence for a single
12 evolutionary origin of domesticated rice. *Proc Natl Acad Sci USA* **108**(20):
13 8351-8356.
- 14 Moreno Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz
15 Tello PA, Martínez RJ, Hedges DJ, Morris RW et al. 2013. Reconstructing the
16 population genetic history of the Caribbean. *PLoS Genet* **9**(11): e1003925.
- 17 Morishima H. 2002. Reports of the study-tours for investigation of wild and cultivated
18 rice species. Part II., pp. 196-199.
- 19 Oka HI. 1988. *Origin of Cultivated Rice*. Elsevier, New York.
- 20 Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from
21 genome-wide allele frequency data. *PLoS Genet* **8**(11): e1002967.
- 22 Pool JE, Nielsen R. 2009. Inference of historical changes in migration rate from the
23 lengths of migrant tracts. *Genetics* **181**(2): 711-719.
- 24 Pusadee T, Jamjod S, Rerkasem B, Schaal B. 2016. Life-history traits and geographical
25 divergence in wild rice (*Oryza rufipogon*) gene pool in Indochina Peninsula
26 region. *Ann Appl Biol* **168**(1): 52-65.
- 27 Pusadee T, Schaal BA, Rerkasem B, Jamjod S. 2013. Population structure of the
28 primary gene pool of *Oryza sativa* in Thailand. *Genet Resour Crop Ev* **60**(1):
29 335-353.
- 30 R Core Team. 2016. R: A Language and Environment for Statistical Computing. R
31 Foundation for Statistical Computing, <https://http://www.R-project.org/>.
- 32 Rakshit S, Rakshit A, Matsumura H, Takahashi Y, Hasegawa Y, Ito A, Ishii T,
33 Miyashita NT, Terauchi R. 2007. Large-scale DNA polymorphism study of
34 *Oryza sativa* and *O. rufipogon* reveals the origin and divergence of Asian rice.
35 *Theor Appl Genet* **114**(4): 731-743.
- 36 Sang T, Ge S. 2007. The puzzle of rice domestication. *J Integr Plant Biol* **49**(6):
37 760-768.
- 38 Second G. 1982. Origin of the genic diversity of cultivated rice (*Oryza* spp.): study of
39 the polymorphism scored at 40 isozyme loci). *Jpn J Genet* **57**: 25257.
- 40 Skotte L, Korneliussen TS, Albrechtsen A. 2013. Estimating individual admixture
41 proportions from next generation sequencing data. *Genetics* **195**(3): 693-702.
- 42 Song BK, Chuah TS, Tam SM, Olsen KM. 2014. Malaysian weedy rice shows its true
43 stripes: wild *Oryza* and elite rice cultivars shape agricultural weed evolution in
44 Southeast Asia. *Mol Ecol* **23**(20): 5003-5017.
- 45 Sun C, Wang X, Atsushi Y, Kazuyuki D, Nobuo I. 1996. RFLP analysis of nuclear DNA
46 in common wild rice (*O. rufipogon* Griff.) and cultivated rice (*O. sativa* L.).

- 1 *Scientia Agricultura Sinica* **30**(4): 37-44.
- 2 Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, Bustamante CD,
3 McCouch SR. 2007. Global dissemination of a single mutation conferring white
4 pericarp in rice. *PLoS Genet* **3**(8): e133.
- 5 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
6 polymorphism. *Genetics* **123**(3): 585-595.
- 7 Tan L, Li X, Liu F, Sun X, Li C, Zhu Z, Fu Y, Cai H, Wang X, Xie D et al. 2008. Control
8 of a key transition from prostrate to erect growth in rice domestication. *Nat*
9 *Genet* **40**(11): 1360-1364.
- 10 Vaughan DA, Lu BR, Tomooka N. 2008. The evolving story of rice evolution. *Plant Sci*
11 **174**(4): 394-408.
- 12 Vaughan DA, Sanchez PL, Ushiki J, Kaga A, Tomooka N. 2005. Asian rice and weedy
13 rice evolutionary perspectives. In *Crop Fertility and Volunteerism*, (ed. J
14 Gressel), pp. 257-277. CRC Press.
- 15 Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R. 2013. Estimating inbreeding
16 coefficients from NGS data: Impact on genotype calling and allele frequency
17 estimation. *Genome Res* **23**(11): 1852-1861.
- 18 Wang H, Xu X, Vieira FG, Xiao Y, Li Z, Wang J, Nielsen R, Chu C. 2016. The power of
19 inbreeding: NGS based GWAS of rice reveals convergent evolution during rice
20 domestication. *Mol Plant* **9**(7): 975-985.
- 21 Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, Fuse T, Baba T, Yamamoto
22 K, Umehara Y, Nagamura Y et al. 2000. *Hd1*, a major photoperiod sensitivity
23 quantitative trait locus in rice, is closely related to the Arabidopsis flowering
24 time gene *CONSTANS*. *Plant Cell* **12**(12): 2473-2484.
- 25 Zhu Q, Ge S. 2005. Phylogenetic relationships among A-genome species of the genus
26 *Oryza* revealed by intron sequences of four nuclear genes. *New Phytol* **167**(1):
27 249-265.
- 28 Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. 2007. Multilocus analysis of nucleotide
29 variation of *Oryza sativa* and its wild relatives: severe bottleneck during
30 domestication of rice. *Mol Biol Evol* **24**(3): 875-888.
- 31 Zhu Y, Ellstrand NC, Lu BR. 2012. Sequence polymorphisms in wild, weedy, and
32 cultivated rice suggest seed-shattering locus *sh4* played a minor role in Asian
33 rice domestication. *Ecol Evol* **2**(9): 2106-2113.
- 34 Ziska LH, Gealy DR, Burgos N, Caicedo AL, Gressel J, Lawton Rauh AL, Avila LA,
35 Theisen G, Norsworthy J, Ferrero A et al. 2015. weedy (red) rice: an emerging
36 constraint to global rice production. In *Adv Agron*, Vol 129 (ed. DL Sparks), pp.
37 181-228. Elsevier, New York.
- 38
- 39

Figure 1

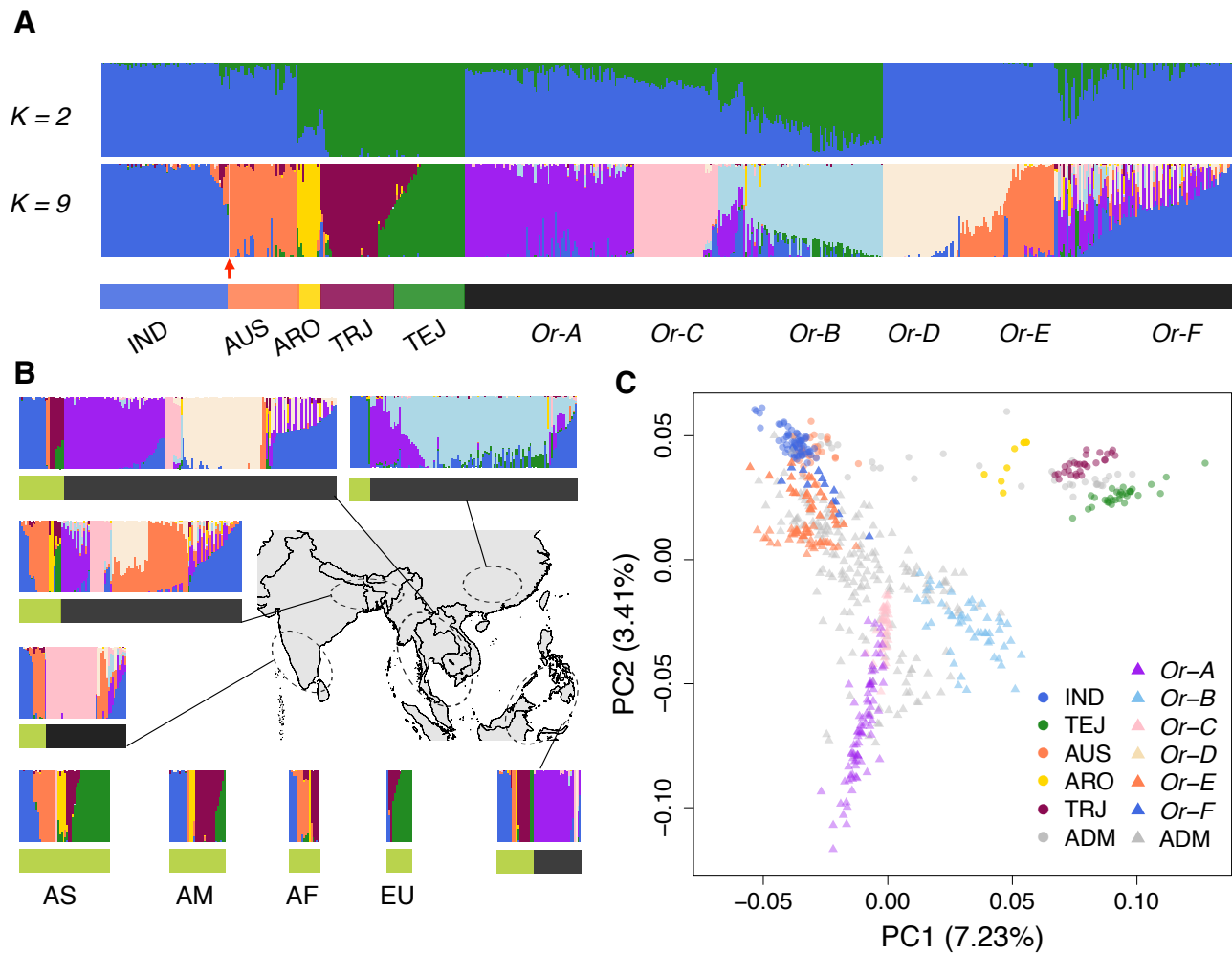


Figure 2

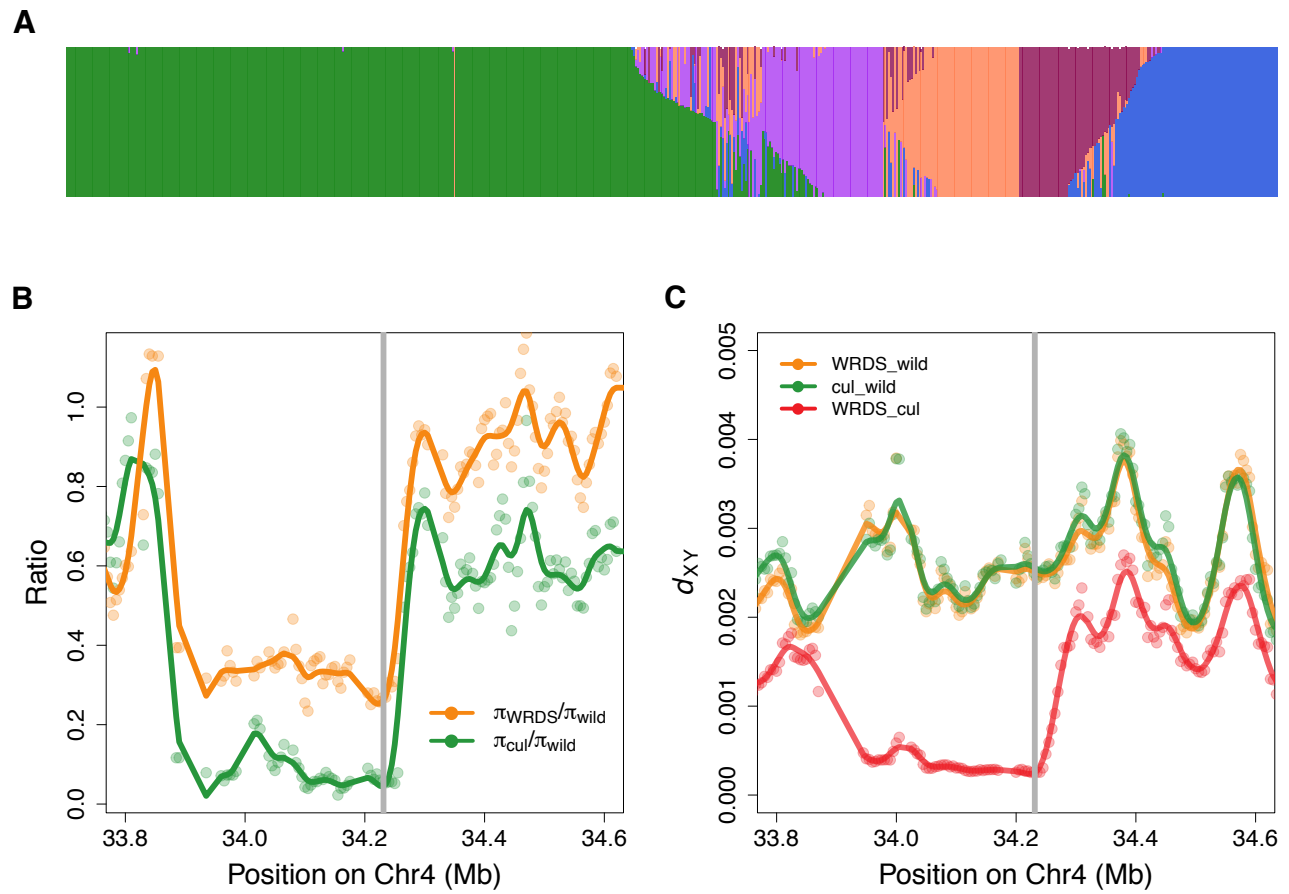
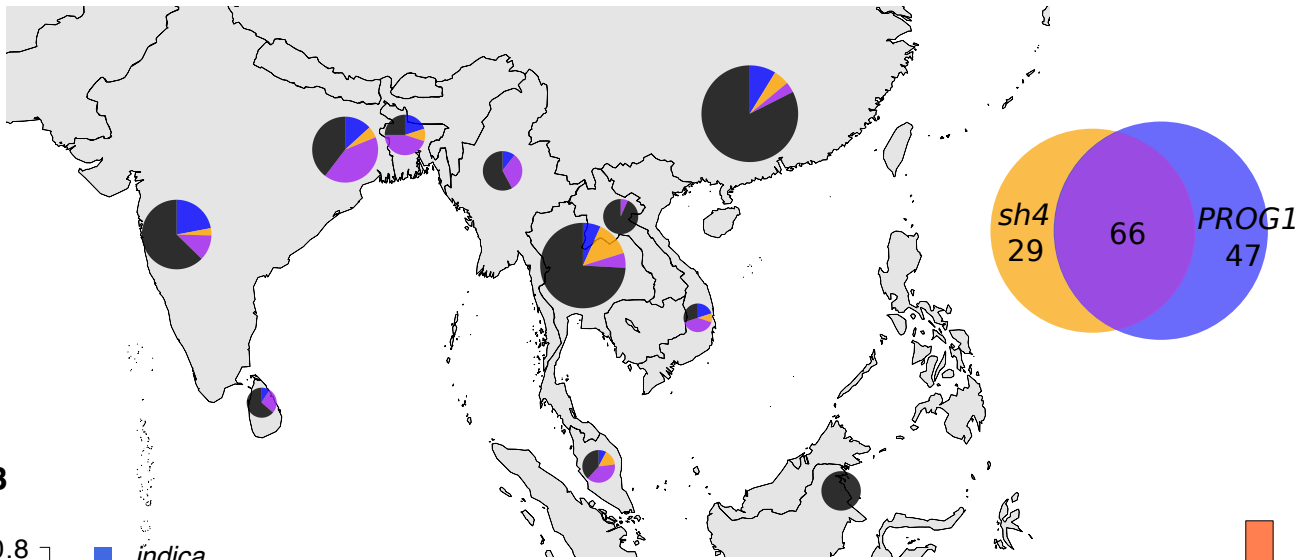


Figure 3

A



B

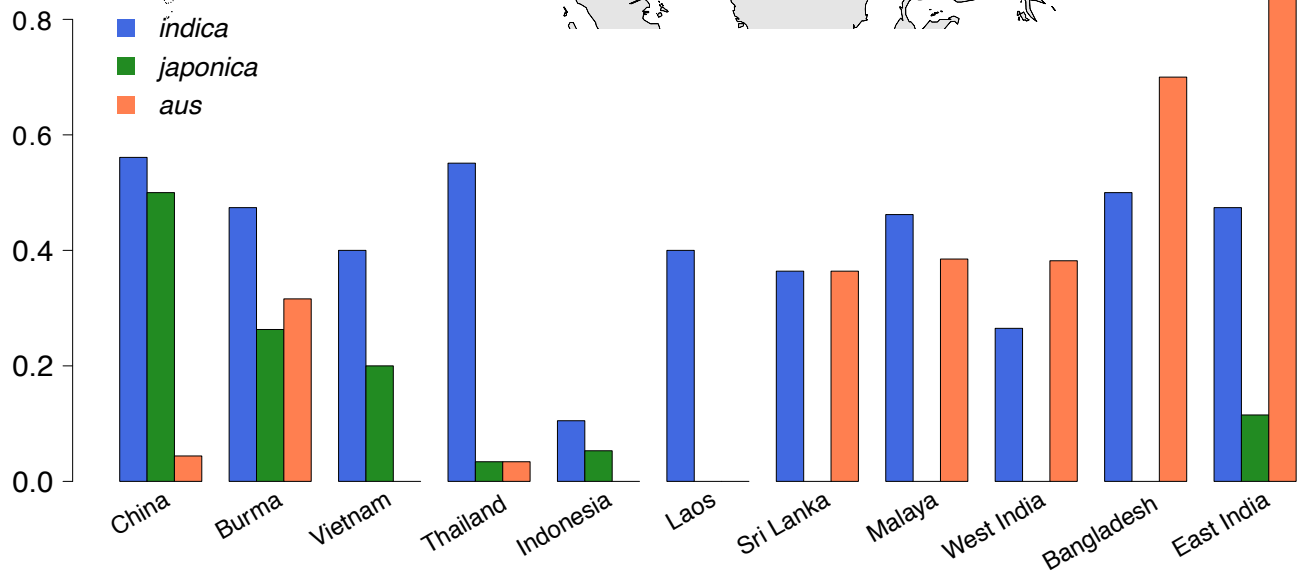


Figure 4

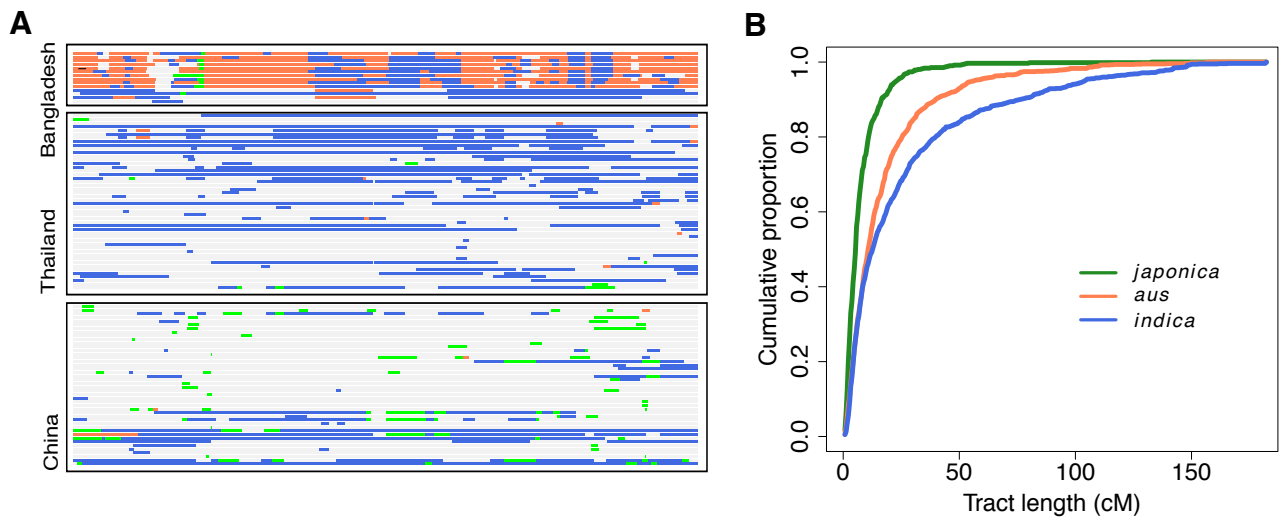


Figure 5

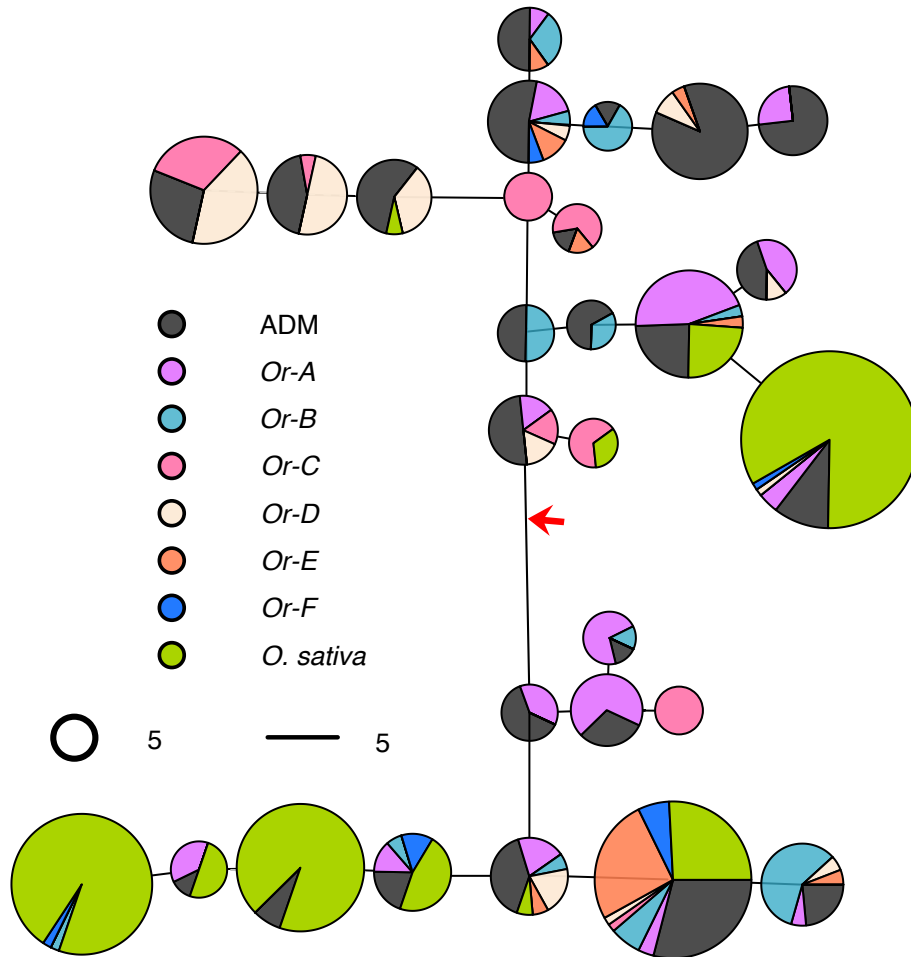
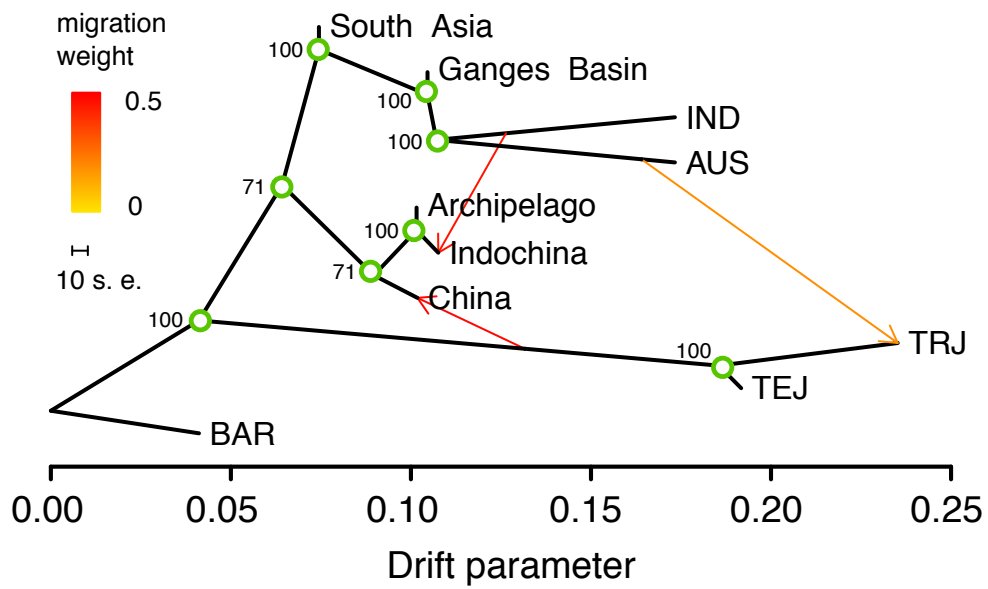


Figure 6





Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice

Hongru Wang, Filipe G. Vieira, Jacob E. Crawford, et al.

Genome Res. published online April 6, 2017

Access the most recent version at doi:[10.1101/gr.204800.116](https://doi.org/10.1101/gr.204800.116)

Supplemental Material

<http://genome.cshlp.org/content/suppl/2017/04/18/gr.204800.116.DC1>

P<P

Published online April 6, 2017 in advance of the print journal.

Accepted Manuscript

Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access

Freely available online through the *Genome Research* Open Access option.

Creative Commons License

This manuscript is Open Access. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International license), as described at <http://creativecommons.org/licenses/by/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>
