Selective sweep with significant positive selection serves as the driving force for the differentiation of japonica and indica rice cultivars

Yang Yuan1,2†, Qijun Zhang3†, Shuiyun Zeng2†, Longjiang Gu2, Weina Si2, Xiaohui Zhang2, Dacheng Tian2, Sihai Yang2* and Long Wang2*

Abstract

Background: Asian cultivated rice (Oryza sativa L.), including japonica and indica, is unarguably the most important crop in Asia as well as worldwide. However, a decisive conclusion of its origination and domestication processes are still lacking. Nowadays, the ever-increasing high-throughput sequencing data of numerous rice samples have provided us new opportunities to get close to the answer of these questions.

Results: By compiling 296 whole-genome sequenced rice cultivars and 39 diverse wild rice, two types of domesticated regions (DR-I and DR-II) with strong selective sweep signals between different groups were detected. DR-I regions included 28 blocks which significantly differentiated between japonica and indica subspecies, while DR-II regions were consisted of another 28 blocks which significantly differentiated between wild and cultivated rice, each covered 890 kb and 640 kb, respectively. In-depth analysis suggested that both DR-IIs and DR-IIs could have originated from Indo-China Peninsula to southern China, and DR-IIs might be introgressed from indica to japonica. Functional bias with significant positive selection has also been detected in the genes of DR-I, suggesting important role of the selective sweep in differentiation of japonica and indica.

Conclusions: This research promoted a new possible model of the origin of the cultivated rice that DR-IIs in japonica and indica maybe independently originated from the divergent wild rice in the Indo-China Peninsula to southern China, and then followed by frequent introgression. Genes with significant positive selection and biased functions were also detected which could play important roles in rice domestication and differentiation processes.

Keywords: japonica, indica, Wild rice, Domestication, Resequencing, Functional genes

Background

As one of the most important cereal grains, rice has been used as a major food source for more than half of the world’s population [1]. Asian cultivated rice is divided into two subspecies, Oryza sativa ssp. japonica and Oryza sativa ssp. indica. The japonica varieties are adapted to more temperate climates, whereas most indica varieties are generally adapted to tropical lowland cultivation [2]. The distinction between the two rice subspecies has been recognized in China since at least the Han dynasty [3]. However, the origin(s) of the two subspecies has long been under scrutiny and debate, particularly on whether the two subspecies were derived from a single or multiple domestications [4]. Numerous studies have suggested that the two subspecies of indica and japonica were probably independently domesticated from different isolates of the wild rice, Oryza rufipogon [5, 6], which was then following by further differentiation [7]. Londo et al. believed that indica might have originated from a region south of the Himalaya mountain range, whereas japonica originated from wild rice in southern
China [8]. On the other hand, Molina et al. reported that Asian rice might have been domesticated from a single origin [9]. A more recent study supports the viewpoint of a single origin, indicating that Asian rice cultivars might have been domesticated from southern China [10]. The comprehensive and deeper survey of the genetic basis of domestication may contribute to improved domestication strategies of organisms and provide novel approaches in deciphering the process of domestication [11–13]. To date, the debate on the origin of Asian rice cultivars remains elusive, thus requiring additional evidence to resolve this issue.

Ongoing works have already characterized a large number of genes, such as Sub1A [14], S-5 [15], and NRT1.1B [16], which were demonstrated to be involved in differentiation between the two subspecies. Similarly, some other genes are related to domestication from wild to cultivated rice, including Blh4 [17], PROG1 [18] and sh4 [19]. It is essential to utilize these genes in distinguishing the two subspecies or wild rice from rice cultivars. However, information on the mechanism underlying the differences in morphological, physiological, and biochemical features, as well as some other aspects between the two subspecies or wild rice and cultivated rice is limited. Moreover, the agronomic traits may be controlled by multiple genes, which were difficult to be identified [20]. Therefore, additional studies that aim to identify genes that are involved in the domestication of rice at the whole-genome scale using new methods such as whole genome sequencing technology are warranted [10, 20].

Recently, a batch of 3,000 rice accessions have been fully sequenced with high coverage [21], which provided unprecedented opportunities to explore the differentiation between indica and japonica via genomic approaches. By characterizing the two types of domesticated regions according to selection signatures and subsequent phylogenetic analysis, we delimited the potential zone for origins of the two subspecies, that is, the two subspecies may have a common origin in the Indo-China Peninsula. In addition, significant positive selection and particular genes that possibly control important morphological traits between cultivars and wild rice, as well as differentiation between indica and japonica were identified in these two types of domesticated regions, which suggests that selective sweep with significant positive selection may serve as the driving force for the differentiation of the two rice subspecies.

Results
Selection of the rice accessions
Although the 3,000 rice genomes dataset [21] provides an unprecedented resource for detecting the selective sweep regions in rice cultivars, most of which have low-coverage or inadequate sequencing depths that are difficult to employ in the high-resolution detection analysis due to the frequent coverage gaps. Therefore, 330 cultivars with ≥15× sequencing depths were downloaded from the 3,000 samples (Additional file 1: Table S1). To detect the regions with significant differentiation between japonica and indica, only those cultivars with significant differentiation that was detected by diversity calculation and PCA were retained (Additional file 1: Figure S1 and Table S2 see Methods). Finally, 296 out of the 330 rice cultivars, including 154 indica and 142 japonica cultivars, from 45 different countries (or regions), were used for further analysis, which have 23.8× average sequencing depth (ranging from 15.0× to 51.1×; Additional file 1: Table S1). The other 34 cultivars were removed due to the potential genetic admixtures of these two subspecies (Additional file 1: Table S1). And the retained cultivars exhibited scattered geographic distribution. About 1/3 of these cultivars were collected from China and India, both of which are geographically larger. Another 1/3 were from countries in Southeast Asia, including Thailand, Burma, Cambodia and so on. The last 1/3 were from other countries, such as Japan, South Korea and so on.

In addition, the sequences of 39 diverse wild rice O. rufipogon or O. nivara (also referred to annual of O. rufipogon) accessions, which were believed to be the immediate progenitor of the Asian cultivated rice of O. sativa, were collected from previous reports [10] (Additional file 1: Table S3). These sequences have 11× sequencing depth on average (ranging from 4.0× to 68.6×). On the other hand, the whole-genome sequences of 20 African rice O. glaberrima accessions, which are closely related to O. sativa and O. rufipogon, were downloaded at a 42.0× average sequencing depth (ranging from 5.9× to 120.3×; Additional file 1: Table S3) as outgroups for further analysis.

Nucleotide diversity/divergence within or between these two subspecies
After trimming and removing low-quality bases, the clean reads of 296 Asian cultivated rice, 39 wild rice, and 20 Africa cultivated rice accessions were mapped to the Nipponbare reference genome. Then, joint SNP calling and genotyping of the sequenced samples were performed (see Methods). After stringent filtering, a total of 23,147,437 SNPs across all the 355 various rice samples were called and used to estimate diversity among samples or divergence among groups.

Based on the SNP data, π was estimated at 0.00196 within japonica and 0.00265 within indica groups (Additional file 1: Figure S2), which were higher than that within japonica (0.0006) and indica (0.0016) that were estimated in 517 Chinese indica and japonica
landraces [22]. These discrepancies may be due to (i) higher genetic diversity in the worldwide cultivars than that only in Chinese landraces; (ii) an underestimation of the genetic diversity in these 517 Chinese landraces caused by the low-coverage sequencing depth (average 1x for each sample). $D_{xy}$ between *indica* and *japonica* was about 0.00617 in our selected samples (Additional file 1: Figure S2), suggesting a significant differentiation between these two subspecies, which also had been confirmed by the PCA analysis. As expected, the diversity within these wild rice accessions was expected about 0.0067 (Additional file 1: Figure S2), which was significantly higher than that in *japonica* or *indica*, suggesting the strong bottleneck during the domestication that has also been reported by other studies [10, 23, 24]. As is known, the short-read sequencing technology could generate massive erroneous SNP calls if not properly handled. However, it was less likely for these false signals to mimic the truly differentiated SNPs. Such erroneous calls would distribute more randomly, both in *japonica* and in *indica*, which rarely resulted in low diversities within each subspecies. Furthermore, we did not observe any inflation in the estimated diversities, or any significant departure in the PCA analysis. This implies a conservative SNP call set, given the high coverage dataset and the stringent criteria used in this study.

**Detection of regions that underwent selective sweep**

The large SNP data set detected in our samples provides an opportunity to identify artificial selected regions by comparing polymorphism levels in these cultivated and wild rice accessions. The selective signature from domestication with selective sweep includes a reduction in nucleotide diversity/divergence and altered allele frequency in these domestication loci. As mentioned above, the average diversity within *japonica* or *indica* is about 0.002. Therefore, the value of 0.002 was employed as one cutoff to detect the divergent regions between *japonica* and *indica*. Among these divergent regions, only those which have at least 10-fold lower diversity (0.0002) between the two subspecies can be defined as selective sweep regions both in *japonica* and *indica*. Compared with previous studies, two advantages can be found in our definition: i) the ten-fold-lower diversity is a more stringent criterion than previous study [10], ii) by using 0.002, we can easily exclude these regions having very low divergence between these two subspecies and very low diversity within each of the two subspecies, which may not be the result of divergence and selection, instead, caused by sequencing or analytical errors.

Therefore, two types of domesticated regions were detected (Tables 1, 2 and 3): domesticated region type I (defined as DR-I), which have undergone selective sweep within each subspecies ($\pi \leq 0.0002$) but have high divergence between these two subspecies ($D_{xy} \geq 0.002$); domesticated region type II (defined as DR-II), which have low diversity regions shared by both subspecies (not only $\pi \leq 0.0002$ within each subspecies, but also $D_{xy} \leq 0.0002$ between the two subspecies), but have high diversity within the wild rice ($\pi \geq 0.001$) (See Methods for details).

DR-I showed that differentiation has already occurred in these wild rice, and selective sweep was independently imposed on the domestication of *indica* or *japonica*, respectively. Therefore, the *indica*-*japonica* differentiation was driven not only by genetic drifts or possible natural selection of wild rice, but also by artificial selection of cultivated rice. The artificial selection should be much more powerful in accelerating the differentiation progress. For this type of domesticated regions, a total of 28 blocks, including about 890 kb of DNA (average: 31.8 kb; ranging from 10 kb to 198 kb) and 163 genes, were detected (Tables 1 and 2). Meanwhile, since *japonica* had been subdivided into temperate *japonica* and tropical *japonica* [25, 26], $\pi$ within temperate *japonica* and tropical *japonica*, and $D_{xy}$ between temperate *japonica* and *indica* and tropical *japonica* and *indica* were calculated (Table 2). All the data showed a similar result, suggesting that the DR-I regions were conserved in temperate *japonica* and tropical *japonica*, and divergent between temperate *japonica* and *indica* or tropical *japonica* and *indica*.

For DR-II, normal diversity was detected within these wild rice accessions ($\pi \geq 0.001$), whereas very low diversity was detected not only within each subspecies ($\pi \leq 0.0002$), but also between the two subspecies ($D_{xy} \leq 0.0002$), suggesting that selective sweep was simultaneously imposed on both *indica* and *japonica*, and these overlapping DR-IIs may have originated only once or as a result of selection and subsequent introgression from one subspecies to another. For this type of domesticated region, a total of 28 blocks, including about 640 kb genomic length (average: 23.0 kb, ranging from 10 kb to 80 kb), was detected (Tables 1 and 3).

**Table 1** Blocks with selective sweep

|        | DR-I | DR-II |
|--------|------|-------|
| Block numbers | 28   | 28    |
| Average length (kb) | 31.8 | 23    |
| Total length (Mb)  | 0.89 | 0.64  |
| Including genes    | 163  | 110   |

DR-I: domesticated regions of type I that have undergone selective sweep within each subspecies ($\pi \leq 0.0002$) but have high divergence between these two subspecies ($D_{xy} \geq 0.002$); DR-II: domesticated regions of type II that have low diversity shared by both subspecies (not only $\pi \leq 0.0002$ within each subspecies, but also $D_{xy} \leq 0.0002$ between the two subspecies), but have high diversity within the wild rice ($\pi \geq 0.001$).
Origin of DR-I
When using the SNPs of DR-I to reconstruct a phylogenetic tree, as expected, all japonica or indica samples clustered together within its own clade, respectively, whereas the japonica clade was distantly located from the indica clade (Fig. 1), which was consistent with the results of PCA analysis that the japonica varieties clearly segregate from the indica individuals (Additional file 1: Figure S1). Furthermore, Tajima’s D and FST statistics also indicated significant differences of the genetic structure between japonica and indica populations in the DR-I region (Table 2). On the other hand, some wild rice accessions were clustered within the japonica or indica clade, respectively, and some wild rice accessions were scattered outside the two cultivated rice clades, which was also consistent with the findings of previous studies [10].

In these 28 DR-I blocks, the average diversity was 0.00013 (ranging from 0.00003 to 0.00020) within japonica and 0.00017 (ranging from 0.00011 to 0.00025) within indica (Table 2). However, the average divergence between these two groups was 0.0034, which was 11- to 101-fold higher (~27-fold on average) than the diversity within each group (Table 2), suggesting significant differentiation between indica and japonica in these regions. The average diversity is ~15-fold lower within japonica and ~16-fold lower within indica than their corresponding genome-wide diversity, suggesting strong selection with selective sweep on these regions. On the other
hand, in these blocks, significantly positive correlations were displayed either in the diversity within indica vs. japonica, or Djap-jap vs. $\pi_{jap}$ or $\pi_{ind}$ or $D_{jap-wild}$ and $D_{ind-wild}$ (Additional file 1: Figure S3). This was consistent with the possibility that japonica and indica may have undergone strong artificial selection from different isolates of the wild rice O. rufipogon during domestication at least in these domesticated blocks.

For this reason, a phylogenetic tree was reconstructed using SNPs of the 28 DR-Is to determine which wild rice was more similar to the 28 domesticated blocks that were fixed or near fixed in japonica or indica, respectively. Similar to the findings of previous studies [10], four clear clades were detected, including the African rice clade (also as an outgroup clade), japonica clade with some O. rufipogon accessions, indica clade with some O. rufipogon accessions, and an independent O. rufipogon clade (Fig. 1). All japonica or indica samples were clustered together, whereas all japonica samples were far from all of these indica samples, which matched the PCA result that the japonica varieties clearly segregated from the other groups, suggesting that, at least in these 28 DR-I regions, these two subspecies were probably independently domesticated from different wild rice isolates. Interestingly, the wild rice accessions with the highest similarity to japonica varieties were W1725, W2198, P46 and W3046 (Fig. 1), which were collected from Thailand to Southern China, suggesting that these japonica domesticated regions may have independently originated

| Chr | Range of blocks | Length (bp) | $\pi_{jap}$ *10^5 | $\pi_{ind}$ *10^5 | $D_{jap-ind}$ *10^5 | $\pi_{cultivar}$ *10^5 | $D_{jap-wild}$ *10^5 | $D_{ind-wild}$ *10^5 | $D_{cultivar-wild}$ *10^5 |
|-----|----------------|-------------|-------------------|-------------------|----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| 1   | chr01          | 8645929–8660126 | 14198            | 1.0               | 18.0                | 16.0                  | 13.0                  | 118.0                 | 104.0                   |
| 2   | chr03          | 26199939–26210003 | 10065            | 10.0              | 15.0                | 14.0                  | 13.0                  | 141.0                 | 144.0                   |
| 3   | chr04          | 25959995–26040007 | 80013            | 10.0              | 15.0                | 14.0                  | 13.0                  | 181.0                 | 228.0                   |
| 4   | chr04          | 26089892–26100078 | 10187            | 6.0               | 8.0                 | 7.0                   | 7.0                   | 155.0                 | 250.0                   |
| 5   | chr04          | 26229998–26260103 | 30106            | 10.0              | 15.0                | 15.0                  | 14.0                  | 230.0                 | 327.0                   |
| 6   | chr04          | 26359991–26370003 | 10013            | 5.0               | 6.0                 | 6.0                   | 6.0                   | 125.0                 | 168.0                   |
| 7   | chr04          | 26789966–26810053 | 20088            | 10.0              | 16.0                | 14.0                  | 14.0                  | 155.0                 | 251.0                   |
| 8   | chr04          | 26949945–26970083 | 20139            | 9.0               | 14.0                | 13.0                  | 12.0                  | 141.0                 | 190.0                   |
| 9   | chr04          | 27009995–27060009 | 50015            | 7.0               | 20.0                | 17.0                  | 16.0                  | 286.0                 | 360.0                   |
| 10  | chr04          | 32520159–32540000 | 10842            | 18.0              | 13.0                | 16.0                  | 16.0                  | 109.0                 | 104.0                   |
| 11  | chr04          | 3389971–33900005 | 10035            | 15.0              | 10.0                | 20.0                  | 16.0                  | 173.0                 | 228.0                   |
| 12  | chr04          | 33979898–3399038  | 10050            | 15.0              | 10.0                | 16.0                  | 14.0                  | 100.0                 | 135.0                   |
| 13  | chr04          | 34219891–34249999 | 30109            | 12.0              | 7.0                 | 12.0                  | 11.0                  | 277.0                 | 316.0                   |
| 14  | chr05          | 29729953–29740204 | 10252            | 9.0               | 5.0                 | 13.0                  | 12.0                  | 141.0                 | 190.0                   |
| 15  | chr07          | 2739991–2770008  | 30018            | 15.0              | 13.0                | 18.0                  | 16.0                  | 406.0                 | 456.0                   |
| 16  | chr07          | 2799992–2810002  | 10011            | 11.0              | 6.0                 | 9.0                   | 9.0                   | 262.0                 | 311.0                   |
| 17  | chr07          | 287923–2910069  | 30147            | 14.0              | 10.0                | 13.0                  | 13.0                  | 280.0                 | 437.0                   |
| 18  | chr07          | 2979996–2991842  | 11847            | 1.0               | 4.0                 | 3.0                   | 3.0                   | 140.0                 | 104.0                   |
| 19  | chr07          | 3096102–3109999  | 13898            | 10.0              | 9.0                 | 11.0                  | 10.0                  | 110.0                 | 116.0                   |
| 20  | chr07          | 3699995–3710053  | 10059            | 12.0              | 7.0                 | 10.0                  | 10.0                  | 364.0                 | 394.0                   |
| 21  | chr07          | 3759873–3790006  | 30134            | 9.0               | 7.0                 | 10.0                  | 9.0                   | 264.0                 | 273.0                   |
| 22  | chr07          | 4019981–4030093  | 10113            | 12.0              | 6.0                 | 12.0                  | 10.0                  | 138.0                 | 131.0                   |
| 23  | chr07          | 4070000–4090006  | 20007            | 9.0               | 17.0                | 15.0                  | 14.0                  | 236.0                 | 256.0                   |
| 24  | chr08          | 23749999–23770091| 20093            | 12.0              | 14.0                | 14.0                  | 14.0                  | 246.0                 | 269.0                   |
| 25  | chr08          | 2380970–23860336 | 50067            | 9.0               | 16.0                | 15.0                  | 14.0                  | 549.0                 | 500.0                   |
| 26  | chr08          | 23919999–23940003 | 20005            | 5.0               | 10.0                | 9.0                   | 8.0                   | 206.0                 | 283.0                   |
| 27  | chr08          | 23989975–24040080 | 50106            | 7.0               | 14.0                | 11.0                  | 11.0                  | 273.0                 | 315.0                   |
| 28  | chr12          | 24930000–24950016 | 20017            | 13.0              | 1.0                 | 11.0                  | 9.0                   | 168.0                 | 124.0                   |
| Average |                  | 22951          | 10.0              | 11.0              | 12.0              | 11.0                  | 213.0                 | 247.0                 | 246.0                   |
from the regions of Southern China [10, 27] to the Indo-China Peninsula. On the other hand, the wild rice accessions with the highest similarity to *indica* were W1559, CA97, L89_12 and IRGC88812, all of which were collected from Indo-China Peninsula, suggesting that these *indica* domesticated regions may have independently originated from the regions of Indo-China Peninsula. Therefore, the Indo-China Peninsula might be the overlapping place with the closest wild relatives both for *japonica* and *indica*.

Due to limited sampling size of wild rice, one possible explanation for the topology of the phylogenetic tree was that the DR-Is are retained by frequent introgression from cultivars to some wild rice. To test this hypothesis, more wild rice samples are necessary for further investigation of the origin of these domesticated regions. We expect if most other wild rice accessions have a similar pattern with above regions, then we can exclude the introgression from cultivars to wild rice in these two regions. Conversely if there were a variety of wild rice accessions exhibiting dissimilar pattern in these two regions, then introgression from cultivars to wild rice may have an impact in our selected wild rice samples. Then, more than 400 whole-genome sequenced wild rice samples were collected to enlarge the sample panel, mostly having low sequence depth (~1×) (Additional file 1: Table S4). Additional file 1: Table S5 repeatedly shows that the wild rice accessions from the

---

**Fig. 1** Phylogenetic tree of the 28 DR-I blocks. The low depth wild rice (whole genome depth <4) were not used in this tree.
Indo-China Peninsula (e.g., W1725 and W1506 from Thailand) or Southern China (e.g., W3093, P46 and W3040) have the highest identity to the genotype of *japonica*, whereas most of the wild rice accessions from the Indo-China Peninsula (e.g., W1559, W1086 and W1930) have the highest identity to the genotype of *indica*. This is consistent with the above result that, at least in these 28 DR-I regions, the closest wild relatives for *japonica* and *indica* have the overlapping place in Indo-China Peninsula, suggesting that both of the two domesticated subspecies may have a common place of the origin.

### Origin of DR-II
In the 28 DR-II blocks, the average diversity was 0.0001 either within *japonica* or *indica*, even in all of these cultivars, which was about 21-fold (from 6.1 to 168-fold) lower than that of wild rice accessions (Table 3), suggesting strong selective sweep both in *indica* and *japonica* accessions. This result also suggested that these 28 DR-II blocks should have a common origin and subsequently introgressed from one subspecies to another [10]. The diversity in these regions within *japonica*, *indica*, or both was significantly lower than the diversity in the 28 DR-I blocks within *japonica* (t-test, \( P = 0.011 \)) or *indica* (t-test, \( P < 0.0001 \)), suggesting that the DR-I blocks may be independently fixed in *indica* or *japonica* prior to the fixation of the DR-II blocks.

A phylogenetic tree was also reconstructed using the SNPs from the 28 DR-II blocks. As expected, all *japonica* and *indica* samples were clustered together as a
cultivar clade (Fig. 2), which was different from the topology of the DR-Is, in that tree *japonica* samples were distantly located from these *indica* samples and in different clades. The wild rice with the highest similarity to *japonica* and *indica* in this tree was also the wild rice of W1559, which had the highest similarity only to *indica* but not *japonica* samples in DR-Is, suggesting that these 28 DR-II blocks might have introgressed from *indica* to *japonica*.

To further investigate the direction of introgression in these DR-Is, more whole-genome sequenced wild rice samples, including those with low depth, were used as earlier described. The wild rice accessions of W1086 (India), W0178 (Thailand), W1090 (India), CA97_053 (Cambodia), and W0639 (Burma) (Additional file 1: Table S6) have the highest identity to DR-Is in both *indica* and *japonica*. On the other hand, all these wild rice accessions were clustered within the
indica-type wild rice clade in the phylogenetic tree. This result further suggested that the 28 DR-Is were introgressed from indica to japonica, which was contradictory to the reported 55 major domestication sweeps introgressed from japonica to indica [10], suggesting that not only introgression from japonica to indica, but also from indica to japonica could be detected in our rice cultivars.

Functional classification and selection pressure of specific genes

In the DR-I and DR-II regions, 163 and 110 genes were detected, respectively (Additional file 1: Table S7 and S8). To determine the functional categories of these genes, we summarized the possible functions of these genes using rice Gene Ontology (GO) annotation (http://geneontology.org/) (Figure S4).

Most of the genes in the DR-I regions have a significant functional enrichment in cell growth, anatomical structure morphogenesis, cellular component organization, DNA metabolic process, reproduction, embryo development, and photosynthesis (Additional file 1: Figure S4); for example, some functions responding to seed or coleoptile development (Os01g61380, Os03g05820, Os03g06010-Os03g06060, Os03g06120, Os03g06360, Os03g06890, and Os05g41030), flower, pollen, or anther development (Os03g05140, Os05g38990, Os07g08170, and Os10g39880), root development (Os03g43400 and Os03g43410) (Additional file 1: Table S7). This functional enrichment analysis was consistent with the findings of previous reports that the differentiation of morphological and physiological traits was an adaptation to distinct climatic, ecogeographic, and cultural conditions between indica and japonica [28–30].

Because the selective sweep was respectively detected in indica and japonica, and significant differentiation between these two subspecies was observed, we expected that more genes under positive selection would be detected in the genes of DR-I during the differentiation of indica and japonica. Therefore, the Ka/Ks between indica and japonica samples was employed to evaluate positive selection of these genes. Interestingly, 43 out of the 163 genes (25.9%) with Ka/Ks > 1 (or Ks = 0, Ka > 0) were detected (Additional file 1: Table S7), which showed a significantly higher proportion ($\chi^2$ with Yates correction = 32.4, d.f. = 1, $P < 0.0001$) than that (2,678 out of 27,384, ~8.7%) of the genome-wide reports [7], suggesting that a large number of the DR-I genes were under positive selection during the differentiation of indica and japonica, and the differentiation of the two subspecies was driven by both artificial and natural selection with strong positive selection, which directly acted on many characteristics.

Genes in the DR-II regions were enriched in cell differentiation, anatomical structure morphogenesis, cell growth, photosynthesis, awns development, flower development, cell cycle, embryo development, and reproduction (Additional file 1: Figure S5 and Table S8). This functional enrichment was consistent with the reports on the differentiation of morphological and physiological traits for cultivars and wild rice. For example, sh4 [19], a gene related to shattering, which was included in the selective sweep regions in the research of Huang et al. [10], was also found in our DR-II. Remarkably, a more recent research reported a highly conserved region in chromosome 4, which was also included in our DR-II regions [26], LABA1 (LOC_Os04g43840) and GAD1 (LOC_Os08g37890) were also detected in our results, which were associated with long, barbed awns in wild rice (Table S8) [31, 32].

Discussion

There is currently an ongoing debate both in the genetic and archaeological arenas on the origins of Asian cultivated rice [27]. Numerous studies have suggested that indica and japonica may have multiple origins and that the two subspecies might have been independently domesticated from different isolates of the wild rice of O. rufipogon [8, 27], and indica might have originated in eastern India and japonica from southern China [8]. However, a recent report has shown that a single origin for domesticated Asia rice was more likely to have occurred than multiple origins, as indicated by the results of Bayesian phylogenetic analyses [9]. Interestingly, a more recent study based on genome-wide variation has reported that japonica might have been first domesticated from wild rice of Or-III in southern China, and was subsequently followed by dispersal to Southeast Asia, and then crossed with local wild rice of Or-I to generate indica after several cross-differentiation-selection cycles [10]. Another model was recently proposed by Choi et al., in support of separate origins for different rice subspecies, but only a single de novo domestication of rice which occurred in japonica. [33]. Other studies have shown that the domestication process of rice might have been initiated multiple times, followed by extensive introgression of strongly selected alleles, e.g., some important domesticated genes of sh4, rc, and waxy, which originated in japonica and spread to indica [19, 34, 35]. Therefore, frequent introgression between indica and japonica is a widely accepted theory for the domestication of rice [27]. However, the geographical discordance relating to the concept that japonica was domesticated in Southern China, and indica was generated in Southeast Asia or South Asia by japonica crossing with local wild rice after its dispersal remains elusive.

Interestingly, our data have shown that japonica and indica may have a common place of the origin in Indo-China Peninsula in the DR-I and DR-II, which have the
highest identity to that of the wild rice from Indo-China Peninsula and South Asia, suggesting that all these regions with strongly selective sweep may have a common place with the highest identity for some accessions of the wild rice. This result was consistent with those of a recently published study [26]. On the other hand, our data also showed that the average diversity in the DR-II regions either within japonica, indica, or both was significantly lower than that in the DR-I blocks within japonica (t-test, $P = 0.011$) or indica (t-test, $P < 0.0001$) (Tables 2 and 3), suggesting that the DR-I blocks were independently fixed in indica or japonica prior to the fixation of these DR-II blocks. Third, our data also showed that all of the 28 DR-IIs were introgressed from indica to japonica, which was inconsistent with the findings of previous reports in that 55 major domestication sweeps might have introgressed from japonica to indica [10], suggesting that not only introgression from japonica to indica but also from indica to japonica can be detected in the rice cultivars. Therefore, a possible model of the origin of the cultivated rice (Fig. 3) may be that (i) the proto-indica and proto-japonica might have independently originated from the divergent wild rice in the Indo-China Peninsula, in which the DR-IIs with significant divergence between the two subspecies may have been domesticated during this period; (ii) then, followed by frequent introgression from indica into proto-japonica also in the Indo-China Peninsula (detected in this study), or from japonica into proto-indica in southern China [10, 27], modern japonica and indica formed and spread to different regions. During this period, DR-IIs may have been domesticated in all these cultivated rice (Fig. 4).

Rice have undergone a series of similar phenotypic domestication for effective harvest and planting from their wild progenitors, including a reduction in seed shattering and dormancy, synchronization of seed maturation, increase in seed sizes, and decrease in culm number and branches [11, 19, 36–40]. On the other hand, the two major rice subspecies are differentiated by a number of morphological and physiological characters, along with a substantial sterility barrier [41]. Therefore, we expected that the genes with strongly selective sweep should have functional bias. Interestingly, for the genes found in DR-I regions, on one hand, a significant functional bias was detected in seed or coleoptile development, flower, pollen or anther development, and root development (Additional file 1: Table S7), which was consistent with the differentiation of the morphological and physiological traits, e.g., seed size, between the two subspecies. On the other hand, a large number of these domesticated genes were detected to have undergone positive selection, suggesting that the differentiation of the two subspecies was driven by both natural and artificial selection with strong positive selection in these domesticated genes.

Furthermore, some genes under strong positive selection had a tendency of being clustered within small
regions. For instance, a 0.6-Mb block containing 2 DR-Is showed a significantly lower diversity in *japonica* than *indica*, which strongly implied selection sweep and genetic bottleneck (Additional file 1: Figure S6). Within this block, 8 genes (*Os03g05820*, *Os03g06010–Os03g06060*, and *Os03g06120*), which may be related to root development and coleoptile development, were clustered [29–31]. In addition, not only these 8 genes, but also the nearby genes were under a significantly positive selection (Additional file 1: Table S7), although the function of the other genes was unclear or apparently showed no relationship with the differentiation of *japonica* and *indica*. Hitchhiking effect may explain such a high positive selection of this region, or that the genes with unknown functions may have played major roles in the differentiation of the two rice subspecies. Similar results were observed in DR-II (Additional file 1: Figure S7 and Table S8), suggesting that these regions may contain a large number of important genes that were related to the domestication of the cultivated rice. Coincidentally, two clustered regions in chr04: 26.0 Mb–27.1 Mb and chr08:23.7 Mb–24.9 Mb were also detected by Hua et al. [31] and Jin et al. [32]. Thus, the present research not only determined the origin of the Asian rice cultivars in higher detail, but also facilitated in the discovery of more genes relative to domestication and breeding. Our method could be applied to molecular breeding and agricultural schemes of various rice cultivars.

We need to notice that, although the samples were only selected according to the coverage threshold, we could not totally rule out sampling bias. In this study, we only focused on those most highly-differentiated *indica* and *japonica* accessions, this could be failed to reflect the truly diversity between *indica* and *japonica*. The somewhat controversial conclusions from different researches [10, 33] might in turn reflect the importance of sampling strategy. However, as we are not going to draw a final conclusion on the debate of rice domestica-

tion, we called for more extensive samplings to further the generalizability of our conclusions.

**Conclusion**

By using 296 whole-genome sequenced rice cultivars in the 3000 rice project and 39 diverse wild rice of *O. rufipogon*, our research had detected 28 DR-Is and 28 DR-II$s$, which covered 890 kb and 640 kb regions, and harbored 163 and 110 genes, respectively. The results suggested that both DR-Is and DR-II$s$ might have originated from the Indo-China Peninsula to southern China. In addition, functional bias with significant positive selection has been detected in the genes of DR-I, suggesting that the selective sweep with significant positive selection might drive for the differentiation of the *japonica* and *indica* rice subspecies.

**Methods**

**Data source**

The resequencing data of *japonica* and *indica* individuals was obtained from the 3,000 rice project [21] (Additional file 1: Table S1). For wild rice (*O. rufipogon*), 371 rice samples were obtained from Huang et al. [10], 10 were obtained from Xu et al. [25], one was obtained from Ohyanagi et al. [42] and one was obtained from Zhang et al. [43] (Additional file 1: Table S3). The resequencing data of 20 African domesticated rice (*O. glaberrima*) were collected from Wang et al. [44] (Additional file 1: Table S3).

**Analysis of sequencing data**

All available *indica* and *japonica* individuals were selected for further analysis based on the sequencing depth (*japonica* ≥15× and *indica* ≥20×). All reads were mapped against the reference genome Nipponbare (IRGSP1.0, http://rice.plantbiology.msu.edu/) using BWA-mem (version 0.7.9a-r786) [45] with option “-M”. Picard-MarkDuplicates (version 1.114, http://broadinstitute.-

github.io.picard/) and GATK-IndelRealigner (version 3.2) [46] were applied to correct mapping results. GATK-UnifiedGenotyper was applied to identify SNPs (single nucleotide polymorphism) of each rice individual. To reduce possible sequencing and mapping errors, SNPs with poor quality (quality <50), low depth (depth <3) or low frequency (MAF <5%) were excluded. These retained SNPs were then used for further analysis.

**Selection of *japonica* and *indica* samples**

Nucleotide diversity (π) or divergence (Dxy), which was defined as the ratio between SNP numbers and the corresponding region length [47], served as criterion for the selection of *indica* and *japonica* individuals. For example, as to one *japonica* individual, we defined π as the average π value between this individual and all the other *japonica* samples. Similarly, Dxy was the average Dxy value between this *japonica* individual and all the *indica* samples. When π ≥ Dxy, this *japonica* individual was considered to be a non-typical *japonica* rice and excluded from the analysis. Finally, principal components analysis (PCA) was also performed with VCFtools (http://vcftools.sourceforge.net/) and GCTA (http://
cnsgenomics.com/software/gcta/), using all the SNPs in the whole genome. The first two components of the PCA analysis were adapted to further remove the *japonica*/*indica* individuals that were not clustered.

**Detection of the differentiated/domesticated regions**

All the genomes were divided into 37,332 10-kb-windows and π/Dxy was applied as the genetic parameter to identify domesticated regions. For each window, when the π values within *japonica* or *indica* individuals were both

---

Yuan et al. BMC Genomics (2017) 18:307

Page 11 of 13
were merged into one block, even when the addition, adjacent uniform-type domesticated regions proportion of repeat sequence, were discarded. In mapping quality, or obvious translocation, or high IGV [48], regions with low mapping depth, or low domesticated region type II (DR-II). Furthermore, using these blocks were used for further analysis as completed domesticated regions (DR-I or DR-II).

**Calculation of the genetic parameters**

For all the SNPs within the entire genome of selected japonica and indica individuals, when most (≥80%) japonica individuals shared one genotype whereas most (≥80%) indica individuals shared another genotype, we defined this SNP site as a near-fixed marker between japonica and indica. Then, using these near-fixed markers, non-synonymous (Ka) or synonymous (Ks) substitutions of the coding sequences between japonica and indica were calculated using the Nei-Gojobori method with Jukes-Cantor correction. Whole-genome Fst values and Tajima’s D indices were calculated by VCFtools. Calculation of the Fst and its significance in domesticated regions were performed by Arlequin3.1 [49]. A 5% confidence interval within the whole rice genome was used to identify significance of Tajima’s D.

SNPs in DR-I or DR-II were used to build corresponding neighbor-joining (NJ) trees with Jukes-Cantor model using MEGA v5.0 [50]. Bootstrap analysis with 1,000 replicates was used to estimate the stability of internal nodes and gaps/missing data treatment was performed to identify pairwise deletions.

**Additional file**

**Additional file 1:** Figure S1. PCA plots of the first two components before (a) and after (b) sample selection. Figure S2. The proportions of the genome-wide diversity within the groups of japonica, indica and wild rice and divergence between japonica and indica group. Figure S3. Diversity/divergence relationship between rice groups. Figure S4. GO statistic of the DR-I regions. Figure S5. GO statistic of the DR-II regions. Figure S6. Clustered regions of DR-I. Figure S7. Clustered regions of DR-II. Table S1. List of 330 rice cultivars downloaded from the 3 K-rice project. Table S2. PCA value for each sample. Table S3. List of African cultivated rice (O. glaberrima) and wild rice (O. rufipogon and O. nivara) used to generate phylogenetic tree. Table S4. List of wild rice obtained from Huang et al. used in this project. Table S5. SNP genotype of DR-I. Table S6. SNP genotype of DR-II. Table S7. List of 163 genes in the 28 DR-I regions. Table S8. List of 110 genes in the 28 DR-I regions. (PDF 2465 kb)

**Abbreviations**

DR-I: Domesticated regions of type I; DR-II: Domesticated regions of type II; GO: Gene Ontology; PCA: Principle component analysis; SNP: Single nucleotide polymorphism

**Acknowledgements**

Not applicable.

**Funding**

The present work was supported by the National Natural Science Foundation of China (31601041, 31570368), Program for Changjiang Scholars and Innovative Research Team in University (IRT_14R27), the National Major Special Project on New Varieties Cultivation for Transgenic Organisms (No. 2016ZX08002-003), Fundamental Research Funds for the Central Universities and Jiangsu Collaborative Innovation Center for Modern Crop Production, China Postdoctoral Science Foundation (2016 M601772).

**Availability of data and materials**

All the sequencing data were downloaded from NCBI Sequence Read Archive (SRA) under accession PRJEB6180, ERP001143, ERP000729, ERP000106, PRJNA13765, DRA000438, SRA023116 and SRP051089 (see Additional file 1: Table S1, Table S3 and Table S4 for reference detail). The phylogenetic data has been deposited in the website (http://gattaca.nju.edu.cn/pub_data.html).

**Authors’ contributions**

D.T., S.Y. and L.W. designed the study. Y.Y., Q.Z., S.Z., L.G., W.S. and X.Z. performed research; Y.Y., Q. Z., S.Z., L.G. and L.W. analyzed data; S.Y. and L.W. composed the paper. All authors read and approved the final manuscript.

**Competing interests**

The authors declare no competing interests.

**Consent for publication**

Not applicable.

**Ethics approval and consent to participate**

Not applicable.

**Publisher’s Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Author details**

1. The Applied Plant Genomics Laboratory, College of Agricultural Sciences, Nanjing Agricultural University, Nanjing 210095, China. 2. State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China. 3. Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China.

Received: 21 June 2016 Accepted: 8 April 2017

**References**

1. Gao Z-Y, Zhao S-C, He W-M, Guo L-B, Peng Y-L, Wang J-J, et al. Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. Proc Natl Acad Sci. 2013;110:14492–7.

2. Yang Y, Zhu K, Xia H, Chen L, Chen K. Comparative proteomic analysis of indica and japonica rice varieties. Genet Mol Biol. 2014;37:652–61.

3. Oka H. Origin of Cultivated Rice. Elsevier. 1988.

4. Oka H. Experimental studies on the origin of cultivated rice. Genetics. 1974;78:475–86.

5. Huang P, Molina J, Flowers JM, Rubinstein S, Jackson SA, Purugganan MD, et al. Phylogeography of Asian wild rice, Oryza rufipogon: a genome-wide view. Mol Ecol. 2012;21:4593–604.

6. Chang T-T. The origin, evolution, cultivation, dissemination, and diversification of Asian and African rices. Euphytica. 1976;25:425–41.

7. Sun X, Jia Q, Guo Y, Zheng X, Liang K. Whole-genome analysis revealed the positively selected genes during the differentiation of indica and temperate japonica rice. PLoS One. 2015;10:e0119239.
8. Londo JP, Chiang Y-C, Hung K-H, Chiang T-Y, Schaal BA. Phylogeography of Asian wild rice, Oryza rufipogon, reveals multiple independent domestications of cultivated rice. Oryza sativa. Proc Natl Acad Sci U S A. 2006;103:9578–83.

9. Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, et al. Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci U S A. 2011;108:8351–6.

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

11. Li C, Zhou A, Sang T. Genetic analysis of rice domestication syndrome with the wild annual species. Oryza nivara. New Phytol. 2006;170:185–93.

12. Si W, Yuan Y, Huang J, Zhang X, Zhang Y, Zhang Y, et al. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. New Phytol. 2015;206(4):1491–502.

13. Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen J-Q, et al. Parent-pregnancy sequencing indicates higher mutation rates in heterozygotates. Nature. 2015;523:463–7.

14. Xu K, Xu X, Fukuo T, Canlas P, Maghirang-Rodriguez R, Heuer S, et al. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature. 2006;442:705–8.

15. Yanagihara S, McCouch SR, Ishikawa K, Ogi Y, Maruyama K, Ikehashi H. Molecular analysis of the inheritance of the S-S locus, conferring wide compatibility in Indica/Japonica hybrids of rice (O. sativa L.). TAG Theor. Appl Genet Theor Angew Genet. 1995;90:182–8.

16. Hu B, Wang W, Ou S, Tang J, Li H, Chen R, et al. Variation in SNRT1.B contributes to nitrate-use divergence between rice subspecies. Nat Genet. 2015;47:834–8.

17. Zhu B-F, Si L, Wang Z, Zhou Y, Zhu J, Shuangguan Y, et al. Genetic control of a key transition from black to straw-white seed hull in rice domestication. Nat Genet. 2010;42:961–6.

18. Tan L, Li X, Liu F, Sun X, Li C, Zhu Z, et al. Control of a key transition from prostrate to erect growth in rice domestication. Nat Genet. 2008;40:1360–4.

19. Li C, Zhou A, Sang T. Genetic analysis of rice domestication syndrome with the wild annual species. Oryza nivara. New Phytol. 2006;170:185–93.

20. Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, et al. Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci U S A. 2011;108:8351–6.

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

22. Li C, Zhou A, Sang T. Genetic analysis of rice domestication syndrome with the wild annual species. Oryza nivara. New Phytol. 2006;170:185–93.

23. Lu J, Tang T, Tang H, Huang J, Shi S, Wu C-I. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication of Asian rice followed by gene flow from the wild annual species. Oryza nivara. New Phytol. 2006;170:185–93.

24. Yang C, Kawahara Y, Mizuno H, Wu J, Matsumoto T, Itoh T. Independent domestications of Asian rice. Nat Plants. 2015;1:15164.

25. Xu K, Xu X, Fukuo T, Canlas P, Maghirang-Rodriguez R, Heuer S, et al. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature. 2006;442:705–8.

26. Civá N. Development in Rice Domestication. Plant Cell. 2016;28:2453–63.

27. Choi JY, Platts AE, Fuller DQ, Hsing Y, Wing RA, Purugganan MD. The rice paradox: Multiple origins but single domestication in Asian rice. Mol Biol Evol. 2017;34(4):969–79.

28. Konishi S, Iwaiz T, Lin SY, Ebana K, Fukuta Y, Sasaki T, et al. An SNP caused loss of seed shattering during rice domestication. Science. 2006;312:1392–6.

29. Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, Bustamante CD, et al. Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet. 2007;3:e133.

30. Nunes AL, Delatore CA, Merotto A. Gene expression related to seed shattering and the cell wall in cultivated and weedy rice. Plant Biol Studt. Gen. 2014:16888–96.

31. Oba S, Sumo N, Fujimoto F, Yasue T. Association between Grain Shattering Habit and Formation of Abscission Layer Controlled by Grain Shattering gene sh-2 in Rice (Oryza sativa L.). Jpn J Crop Sci. 1995;64:607–15.

32. Yan S, Zou G, Li S, Wang H, Liu H, Zhai G, et al. Seed size is determined by the combinations of the genes controlling different seed characteristics in rice. TAG Theor. Appl Genet Theor Angew Genet. 2011;123:1173–81.

33. Jiang H, et al. Genetic Analysis and Gene-Mapping of Two Reduced-Culm-Number Mutants in Rice. J Integr Plant Biol. 2006;48:341–7.

34. Fujita D, Ebron LA, Arai E, Kato H, Khush GS, Sheehy JE, et al. Fine mapping of a gene for low-tiller number, Ltn, in japonica rice (Oryza sativa L.) variety Aikawa 1. TAG Theor. Appl Genet Theor Angew Genet. 2010;120:233–40.

35. Yang J, Zhao X, Cheng K, Du H, Ouyang Y, Chen J, et al. A killer-protector system regulates both hybrid sterility and segregation distortion in rice. Science. 2012;337:1336–40.

36. Ohyanagi H, Ebata T, Huang X, Gong H, Fujita M, Mochizuki T, et al. OryzaGenome: Genome Diversity Database of Wild Oryza Species. Plant Cell Physiol. 2015;57:e1.

37. Zhang Q-J, Zhu T, Xie E-H, Shi C, Liu Y-L, Zhang Y, et al. Rapid diversification of five Oryza AA genomes associated with rice adaptation. Proc Natl Acad Sci. 2014;111:E4954–62.

38. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicochea JL, et al. The genome sequence of African rice (Oryza glaterrima) and evidence for independent domestication. Nat Genet. 2014;46:982–8.

39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxlf. Engl. 2009;25:1754–60.

40. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curt Proct Bioinform Bioedit Andreas Bae finishing Al. 2013;43:11.0.1–33.

41. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76:5269–74.

42. Thovaleddottir H, Robinson JT, Mesarov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.

43. Cecillon L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010;10:564–7.

44. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–9.

45. Becker RA, Wilks AR. Maps in S. AT&T Bell Laboratories Statistics Research Report, vol. 93. 1993. p. 2.

Submit your next manuscript to BioMed Central and we will help you at every step:
• We accept pre-submission inquiries
• Our selector tool helps you to find the most relevant journal
• We provide round the clock customer support
• Convenient online submission
• Thorough peer review
• Inclusion in PubMed and all major indexing services
• Maximum visibility for your research

Submit your manuscript at www.biomedcentral.com/submit