Table of Contents

Supplementary Figures

- Supplementary Figure 1. The schematic of experimental design of our study.
- Supplementary Figure 2. Neighbor-joining tree of 15 representative accessions in AA-genome complex group.
- Supplementary Figure 3. PCA plots of the first two components of 446 O. *rufipogon* accessions, with the proportions of the components indicated.
- Supplementary Figure 4. Full plotting of geographic origins of O. rufipogon accessions.
- Supplementary Figure 5. The plots of the population-differentiation statistic (Fst) in O. rufipogon population across 12 rice chromosomes.
- Supplementary Figure 6. Genome-wide average LD decay estimated from 446 O. rufipogon accessions.
- Supplementary Figure 7. Genome-wide association study of leaf sheath color in *O. rufipogon* population using compressed MLM and simple model.
- Supplementary Figure 8. Regional Manhattan plots of GWAS for leaf sheath color in *O. rufipogon* population identify a known gene *OsC1*.
- Supplementary Figure 9. Genome-wide association study of tiller angle in O. *rufipogon* population using compressed MLM and simple model.
- Supplementary Figure 10. Manhattan plots for a QTN simulating qualitative traits using compressed MLM.
- Supplementary Figure 11. Manhattan plots for a quantitative trait simulated from four QTNs using compressed MLM.
- Supplementary Figure 12. Comparison of the performances of GWAS in populations of cultivated and wild accessions.

- Supplementary Figure 13. PCA plots of the first two components of a full population (446 *O. rufipogon* accessions plus 1,083 *O. sativa* varieties).
- Supplementary Figure 14. Neighbor-joining tree of accessions in *indica* (O. sativa) and Or-I group (O. rufipogon).
- Supplementary Figure 15. Neighbor-joining tree of accessions in *japonica* (O. *sativa*) and Or-III group (O. *rufipogon*).
- Supplementary Figure 16. The distributions of population-differentiation statistic (Fst) across the rice genomes.
- Supplementary Figure 17. The alteration of allele frequencies of 212,818 SNPs that were nearly fixed between *indica* and *japonica*.
- Supplementary Figure 18. Whole-genome screens of selective sweep regions in the population of *O. rufipogon Or-I* and *O. sativa indica*.
- Supplementary Figure 19. Whole-genome screens of selective sweep regions in the population of *O. rufipogon Or*-III and *O. sativa japonica*.
- Supplementary Figure 20. Whole-genome screening of selection signals in the *indica* population and the *japonica* population using the XP-EHH method.
- Supplementary Figure 21. Phylogenetic trees of 446 O. rufipogon accessions and 1,083 O. sativa varieties calculated from SNPs for the known domestication loci.
- Supplementary Figure 22. The genetic distance of *indica* and *japonica* to two clades of *O. rufipogon* across the whole genome and across 55 domestication loci.
- Supplementary Figure 23. The genetic distance of *indica* and *japonica* to O. *rufipogon* in different geographic regions.
- Supplementary Figure 24. Five models for the domestication of *indica* and *japonica* rice.

- Supplementary Figure 25. Computational simulations of *in silico* data sets under five demographic scenarios using the forward-simulation program SFS CODE.
- Supplementary Figure 26. Frequency distribution of variation of fifteen domestication-related traits.
- Supplementary Figure 27. Ten QTLs identified with known causal genes.
- Supplementary Figure 28. The distribution of selective sweeps and domestication-related QTLs in 12 rice chromosomes.
- Supplementary Figure 29. The numbers of all QTLs identified and the QTLs within selective sweep regions for fifteen domestication-related traits.
- Supplementary Figure 30. Frequency distribution of QTL effects.
- Supplementary Figure 31. Illustration of the integrated genomics approach for functional annotations of domestication loci and precise detections of causal variants around a known gene *Bh4*.
- Supplementary Figure 32. Illustration of the integrated genomics approach for functional annotations of domestication loci and precise detections of causal variants around a known gene *sh4*.
- Supplementary Figure 33. Illustration of the integrated genomics approach for functional annotations of domestication loci and precise detections of causal variants around a known gene *Rc*.
- Supplementary Figure 34. Causal variants detection and haplotype analysis on a domestication sweep controlling hull color.



Supplementary Figure 1 | The schematic of experimental design of our study. The plants were sequenced by the second-generation sequencing platforms. Genome sequences were aligned with rice reference sequence for variation identification. Whole-genome assembly was performed for the deep sequencing data of W1943. The whole-genome variations were used for the construction of a comprehensive rice haplotype map and the in-depth analyses of rice domestication. The parents of the *O. rufipogon× O. sativa* population were W1943 and Guangluai-4 (*indica*).



Supplementary Figure 2 | Neighbor-joining tree of 15 representative accessions in AA-genome complex group. Details on sampling for the phylogenetic relationship reconstruction and sequence differences of the rice accessions compared with reference genome are provided in Supplementary Table 1 and Supplementary Sections.



Supplementary Figure 3 | PCA plots of the first two components of 446 *O. rufipogon* accessions, with the proportions of the components indicated. The three groups, *Or*-I, *Or*-II and *Or*-III, were identified from the neighbor-joining tree of the 446 accessions and color-coded as Fig. 1a.



Supplementary Figure 4 | Full plotting of geographic origins of *O*. *rufipogon* accessions throughout Asia and Oceania spanning all native geographic range of the species, with accessions in different groups colored as in Fig. 1a.



Supplementary Figure 5 | The plots of the population-differentiation statistic (Fst) in O. rufipogon population across 12 rice chromosomes. The plots of the population genetic differentiation along chromosomes in O. rufipogon. Fst (in red) was calculated in 100-kb windows, among the three clades of O. rufipogon. Orange and gray blocks on chromosomes indicate centromeres and physical gaps of the reference genome, respectively



Supplementary Figure 6 | Genome-wide average LD decay estimated from 446 *O. rufipogon* accessions. The maximum value LD (expressed as r^2) and half of the maximum are indicated with corresponding distances.



Supplementary Figure 7 | Genome-wide association study of leaf sheath color in *O. rufipogon* population using compressed MLM and simple model. (a) Manhattan plots for leaf sheath color using simple model. Negative $\log_{10} p$ -values from a genome-wide scan are plotted against position on each of 12 chromosomes. (b) Quantile-quantile plot for leaf sheath color using simple mode. The horizontal axis shows -log10 transformed expected *P* values, while the vertical axis indicates - log10 transformed observed *P* values (c) Manhattan plots for leaf sheath color using compressed MLM. Negative $\log_{10} p$ -values from a genome-wide scan are plotted against position on each of 12 chromosomes. The horizontal dash-dot lines indicate the genome-wide significance threshold (1×10^{-6}) . (d) Quantile-quantile plot for leaf sheath color using compressed MLM. The horizontal axis shows -log10 transformed expected *P* values, while the vertical axis indicates - log10 transformed observed *P* values from a genome-wide scan are plotted against position on each of 12 chromosomes. The horizontal dash-dot lines indicate the genome-wide significance threshold (1×10^{-6}) . (d) Quantile-quantile plot for leaf sheath color using compressed MLM. The horizontal axis shows -log10 transformed expected *P* values, while the vertical axis indicates -log10 transformed expected *P* values.



Supplementary Figure 8 | Regional Manhattan plots of GWAS for leaf sheath color in *O. rufipogon* population identify a known gene *OsC1*. Negative $\log_{10} p$ -values from a genome-wide scan are plotted against position on a 2-Mb region. The horizontal dash-dot lines indicate the genome-wide significance threshold (1×10⁻⁶).



Supplementary Figure 9 | Genome-wide association study of tiller angle in *O. rufipogon* population using compressed MLM and simple model. (a) Manhattan plots for tiller angle using simple model. Negative $\log_{10} p$ -values from a genome-wide scan are plotted against position on each of 12 chromosomes. (b) Quantile-quantile plot for tiller angle using simple mode. The horizontal axis shows -log10 transformed expected P values, while the vertical axis indicates - log10 transformed observed P values. (c) Manhattan plots for tiller angle using compressed MLM. Negative $\log_{10} p$ -values from a genome-wide scan are plotted against position on each of 12 chromosomes. The horizontal dash-dot lines indicate the genome-wide significance threshold (1×10^{-6}) . (d) Quantile-quantile plot for tiller angle using compressed MLM. The horizontal axis shows -log10 transformed expected P values, while the vertical axis indicates -log10 transformed expected P values.



Supplementary Figure 10 | Manhattan plots for a QTN simulating qualitative traits using compressed MLM. The simulation test was performed in the two populations of wild rice (**a**) and cultivated rice (**b**), respectively. Negative $\log_{10} p$ -values from a genome-wide scan are plotted against position on each of 12 chromosomes. QTL interval is estimated from the size of the genomic region with strong association signals.



Supplementary Figure 11 | Manhattan plots for a quantitative trait simulated from four QTNs using compressed MLM. The simulation test was performed in the two populations of wild rice (**a**) and cultivated rice (**b**), respectively. Negative $\log_{10} p$ -values from a genome-wide scan are plotted against position on each of 12 chromosomes. QTL interval is estimated from the size of the genomic region with strong association signals.



Supplementary Figure 12 | Comparison of the performances of GWAS in populations of cultivated and wild accessions. The simulation tests were performed using a total of twenty simulated traits from a single QTN (simulating qualitative traits) or multiple QTNs (simulating quantitative traits) with various MAF. **a.** Comparison of the interval of associated loci in populations of cultivated and wild accessions. **b.** Comparison of sensitivity (recall rate) of GWAS in populations of cultivated and wild accessions.



Supplementary Figure 13 | PCA plots of the first two components of a full population (446 *O. rufipogon* accessions plus 1,083 *O. sativa* varieties). The two subspecies of *O. sativa* (*indica* and *japonica*) and the three types of *O. rufipogon* (*Or*-I, *Or*-II and *Or*-III) were color-coded as Fig. 2a. Among them, *O. rufipogon* accessions and *O. sativa* varieties were represented by triangle and circle dot, respectively.



Supplementary Figure 14 | Neighbor-joining tree of accessions in *indica* (*O. sativa*) and *Or*-I group (*O. rufipogon*). Accessions of *indica* and *Or*-I group are colored in green and pink, respectively. Different ecotypes are circled with dash-dot lines and labeled correspondingly.



Supplementary Figure 15 | Neighbor-joining tree of accessions in *japonica* (*O. sativa*) and *Or*-III group (*O. rufipogon*). Accessions of *japonica* and *Or*-III group are colored in orange and blue, respectively. Different ecotypes are circled with dash-dot lines and labeled correspondingly.



Supplementary Figure 16 | The distributions of population-differentiation statistic (Fst) across the rice genomes between *indica* and *japonica* (in black), between *Or*-I and *Or*-III (in grey), between *japonica* and *Or*-III (in blue) and between *indica* and *Or*-I (in red).



Supplementary Figure 17 | The alteration of allele frequencies of 212,818 SNPs that were nearly fixed between *indica* and *japonica*. In cultivated rice, these SNPs have an allele frequency >0.95 in one subspecies and <0.05 in the other. The average frequencies of Nipponbare allele and non-Nipponbare allele are represented to be blue and red, respectively. The values of frequencies and alterations are indicated.



SUPPLEMENTARY INFORMATION

RESEARCH

Supplementary Figure 18 | Whole-genome screens of selective sweep regions in the population of *O. rufipogon Or-I* and *O. sativa indica*. The value of π_{or-I}/π_{Ind} are plotted against the position on each of the 12 chromosomes. The horizontal dashed line indicates the genome-wide threshold of selection signals. The loci with known domestication genes are indicated.



Supplementary Figure 19 | Whole-genome screens of selective sweep regions in the population of *O. rufipogon Or*-III and *O. sativa japonica*. The value of $\pi_{or-III\,a}/\pi_{Jap}$ are plotted against the position on each of the 12 chromosomes. The horizontal dashed line indicates the genome-wide threshold of selection signals. The loci with known domestication genes are indicated.



Supplementary Figure 20 | Whole-genome screening of selection signals in the *indica* population and the *japonica* population using the XP-EHH method. The XP-EHH scores from a genome-scan are plotted against the position on each of the 12 chromosomes. The horizontal dashed lines indicate the genome-wide threshold (XP-EHH score > 5). Eight well-characterized domestication loci are shown in red. Of them, only two loci, *sh4* and *qSW5*, show selection signals in the *indica* population. The XP-EHH results were not used in the followed analyses.



Supplementary Figure 21 | Phylogenetic trees of 446 *O. rufipogon* accessions and 1,083 *O. sativa* varieties calculated from SNPs for the known domestication loci. The SNPs within 40 kb genomic regions of each gene were used in phylogenetic relationship construction. The two subspecies of *O. sativa* (*indica* and *japonica*) and the three groups of *O. rufipogon* (*Or*-I, *Or*-II and *Or*-III) were color-coded as Fig. 2a, in pink, sky blue, red, grey and blue, respectively. *Or*-III population from Southern China, probably the immediate progenitors of cultivated rice, is indicated in the trees.



Supplementary Figure 22 | The genetic distance of *indica* and *japonica* to two clades of *O. rufipogon* across the whole genome (a,b) and across 55 domestication loci (c,d). The genetic distance between any two groups A and B, which was computed based on the simple matching coefficient matrix of the SNP data, was expressed as d(A,B).



Supplementary Figure 23 | The genetic distance of *indica* and *japonica* to *O. rufipogon* in different geographic regions. (a) The average distance of *O. rufipogon* accessions from different countries in Southeast Asia, South Asia and Southern China to *japonica* around 55 domestication sweeps. The distance was estimated by simple matching distance of all SNPs within the 55 domestication sweeps. (b) The average distance of *O. rufipogon* accessions from different countries in Southeast Asia, South Asia and Southern China to *indica* around 55 domestication sweeps.(c) The average distance of *O. rufipogon* accessions from different countries in Southeast Asia, South Asia and Southern China to *indica* around 55 domestication sweeps.(c) The average distance of *O. rufipogon* accessions from different countries in Southeast Asia, South Asia and Southers Asia, South Asia and Southern China to *japonica* across the whole genome. The distance was estimated by simple matching distance of all SNPs in the rice genome. (d) The average distance of *O. rufipogon* accessions from different countries in Southeast Asia, South Asia and Southern China to *indica* across the whole genome. The distance was estimated by simple matching distance of all SNPs in the rice genome. (d) The average distance of *O. rufipogon* accessions from different countries in Southeast Asia, South Asia and Southern China to *indica* across the whole genome.



Supplementary Figure 24 | Five models for the domestication of *indica* and *japonica* rice. (a) The hypothesis of a simple single origin of cultivated rice. (b) The hypothesis of independent domestication origins of cultivated rice. (c) This hypothesis considers two origins of cultivated rice followed with bidirectional shares of domestication alleles. (d) This hypothesis considers that *indica* rice was first domesticated and *japonica* rice was subsequently developed from crosses between *indica* rice was subsequently developed from crosses between *japonica* rice and another clade of *O. rufipogon*. (e) This hypothesis considers that *japonica* rice was first domesticated and *indica* rice was subsequently developed from crosses between *japonica* rice and another clade of *O. rufipogon*.



Supplementary Figure 25 | Computational simulations of *in silico* data sets under five demographic scenarios using the forwardsimulation program SFS_CODE. The five demographic scenarios of the rice domestication history, H1-H5, were described in the Supplementary Section S5. The *in silico* domestication loci of *indica* and *japonica* were created by either the domestication model or the admixture model in SFS_CODE. The D-values of the genetic distances between the subspecies of *O. sativa* and each clade of *O. rufipogon* were calculated for both the sImulated datasets and the real data. D-values were computed as followed: D-value(*indica*) = d(indica,Or-I) - d(indica,Or-IIIa) and D-value(*japonica*) = d(japonica,Or-I) - d(japonica,Or-IIIa). The genetic distance between any two groups A and B, which was expressed as d(A,B), was computed based on the simple matching coefficient matrix of the SNP data, .



Supplementary Figure 26 | Frequency distribution of variation of fifteen domestication-related traits in a permanent recombination population from a cross between *O. sativa* Guangluai-4 and *O. rufipogon* W1943. Phenotypic values of the two mapping parents, *O. sativa* Guangluai-4 and *O. rufipogon* W1943, are indicated by triangles of dark red and blue, respectively.



Supplementary Figure 27 | Ten QTLs identified with known causal genes. Curves indicate chromosomal locations and LOD values of detected QTLs. Corresponding traits and phenotypic effect (R^2) of the QTL are indicated. Recombination bins around the LOD peaks of the QTL are illustrated as horizontal bars with their precise physical positions labeled. The physical position of the known causal gene in the bin is indicated by a black triangle.



Supplementary Figure 28 | The distribution of selective sweeps and domestication-related QTLs in 12 rice chromosomes. (a) The distribution of the numbers of selective sweeps in *indica*, *japonica* and the full population in 12 rice chromosomes, indicated by black, blue and red, respectively. (b) The distribution of the numbers of QTLs identified for domestication-related traits (purple) and the QTLs within selective sweep regions (yellow) in 12 rice chromosomes.



Supplementary Figure 29 | The numbers of all QTLs identified (purple) and the QTLs within selective sweep regions (yellow) for fifteen domestication-related traits. For the traits of stigma exsertion, stigma color, shattering, pericarp color and grain weight, all the QTLs identified were located within domestication sweeps. In contrast, no or few QTLs were located in the domestication sweeps for plant height and heading date.



Supplementary Figure 30 | Frequency distribution of QTL effects. The effect size was evaluated by the value of r^2 from QTL analysis using the software WinQTLcart.



Supplementary Figure 31 | Illustration of the integrated genomics approach for functional annotations of domestication loci and precise detections of causal variants around a known gene Bh4. (a) QTL detected in chromosome 4 for the domestication-related trait of hull color. The local region with peak LOD value was within a selective sweep identified, showing a strong selection signal in rice domestication. The value of π_w/π_c are plotted against the position on the local genomic region. The precise physical location of the known domestication gene Bh4 is indicated by a black triangle. (b) Causal variant detection around the locus of Bh4. The top of the panel shows the genomic location of the region with gene structure represented. Coding regions and introns are depicted as rectangles and lines, respectively. The points indicate the orientations of the genes. The black lines indicate the locations of sequence variants identified between *de novo* assembly of wild rice W1943 and reference genome of cultivated rice Nipponbare. Among them, the variants with effects on protein coding are colored with dark red. Bottom of the panel shows the allele frequency of each sequence variant in different populations. Each column represents each sequence variant, and each line represents each group (from top to bottom: outgroup, *Or*-I, *Or*-II, *Or*-III, *indica* and *japonica*). Color index indicates the frequency of derived allele, from 100% ancestral allele in light green to 100% derived allele in dark allele. No data available (NA) is indicated by white. The variants showing a high differentiation in allele frequencies between *O. sativa* and *O. rufipogon*, thus candidate causal variant site, are indicated in red triangle.



Supplementary Figure 32 | Illustration of the integrated genomics approach for functional annotations of domestication loci and precise detections of causal variants around a known gene sh4. (a) QTL detected in chromosome 4 for the domestication-related trait of hull color. The local region with peak LOD value was within a selective sweep identified, showing a strong selection signal in rice domestication. The value of $\pi_{Or.V}/\pi_{Ind}$ are plotted against the position on the local genomic region. The precise physical location of the known domestication gene sh4 is indicated by a black triangle. (b) Causal variant detection around the locus of sh4. The top of the panel shows the genomic location of the region with gene structure represented. Coding regions and introns are depicted as rectangles and lines, respectively. The points indicate the orientations of the genes. The black lines indicate the locations of sequence variants with effects on protein coding are colored with dark red. Bottom of the panel shows the allele frequency of each sequence variant in different populations. Each column represents each sequence variant, and each line represents each group (from top to bottom: outgroup, Or-I, Or-II, Or-III, *indica* and *japonica*). Color index indicates the frequency of derived allele, from 100% ancestral allele in light green to 100% derived allele in dark allele. No data available (NA) is indicated by white. The variants showing a high differentiation in allele frequencies between *O. sativa* and *O. rufipogon*, thus candidate causal variant site, are indicated in red triangle.



Supplementary Figure 33 | Illustration of the integrated genomics approach for functional annotations of domestication loci and precise detections of causal variants around a known gene Rc. (a) QTL detected in chromosome 4 for the domestication-related trait of hull color. The local region with peak LOD value was within a selective sweep identified, showing a strong selection signal in rice domestication. The value of $\pi_{\text{Or-IIIA}}/\pi_{Jap}$ are plotted against the position on the local genomic region. The precise physical location of the known domestication gene Rc is indicated by a black triangle. (b) Causal variant detection around the locus of Rc. The top of the panel shows the genomic location of the region with gene structure represented. Coding regions and introns are depicted as rectangles and lines, respectively. The points indicate the orientations of the genes. The black lines indicate the locations of sequence variants identified between *de novo* assembly of wild rice W1943 and reference genome of cultivated rice Nipponbare. Among them, the variants with effects on protein coding are colored with dark red. Bottom of the panel shows the allele frequency of each sequence variant in different populations. Each column represents each sequence variant, and each line represents each group (from top to bottom: outgroup, *Or-I, Or-II, Or-III, indica* and *japonica*). Color index indicates the frequency of derived allele, from 100% ancestral allele in light green to 100% derived allele in dark allele. No data available (NA) is indicated by white. The variants showing a high differentiation in allele frequencies between *O. sativa* and *O. rufipogon*, thus candidate causal variant site, are indicated in red triangle.

RESEARCH SUPPLEMENTARY INFORMATION



Supplementary Figure 34 | Causal variants detection and haplotype analysis on a domestication sweep controlling hull color. **a**. Candidate causal variants were detected from the sequence variants between cultivated and wild rice, by investigating the spectrum of allele frequencies and predicting the effect of the variants on gene coding. Two candidate causal variants were detected in a ~100-kb genomic region of chromosome 4. **b**. The top of the panel shows the structure of the known causal gene *Bh4*. Bottom of the panel shows the frequencies of the derived allele (with the domestication mutation) of the variants in the groups of cultivated and wild rice. **c**. Common haplotypes around *Bh4* are represented with the phenotypic and geographic information of the representative accessions. **d**. Extended haplotypes around *Bh4* for the representative accessions.