Higgins et al. Rice (2021) 14:52
https://doi.org/10.1186/s12284-021-00481-0

Rice

ORIGINAL ARTICLE Open Access

Resequencing of 672 Native Rice Accessions to Explore Genetic Diversity and Trait Associations in Vietnam

Janet Higgins[1], Bruno Santos[2], Tran Dang Khanh[3,4], Khuat Huu Trung[3], Tran Duy Duong[3], Nguyen Thi Phuong Doai[3], Nguyen Truong Khoa[3], Dang Thi Thanh Ha[3], Nguyen Thuy Diep[3], Kieu Thi Dung[3], Cong Nguyen Phi[3], Tran Thi Thuy[3], Nguyen Thanh Tuan[4], Hoang Dung Tran[5], Nguyen Thanh Trung[6,7], Hoang Thi Giang[3], Ta Kim Nhung[3], Cuong Duy Tran[3], Son Vi Lang[3], La Tuan Nghia[8], Nguyen Van Giang[4], Tran Dang Xuan[9], Anthony Hall[1], Sarah Dyer[2], Le Huy Ham[3], Mario Caccamo[2] and Jose J. De Vega[1*]

Abstract

Background: Vietnam possesses a vast diversity of rice landraces due to its geographical situation, latitudinal range, and a variety of ecosystems. This genetic diversity constitutes a highly valuable resource at a time when the highest rice production areas in the low-lying Mekong and Red River Deltas are enduring increasing threats from climate changes, particularly in rainfall and temperature patterns.

Results: We analysed 672 Vietnamese rice genomes, 616 newly sequenced, that encompass the range of rice varieties grown in the diverse ecosystems found throughout Vietnam. We described four Japonica and five Indica subpopulations within Vietnam likely adapted to the region of origin. We compared the population structure and genetic diversity of these Vietnamese rice genomes to the 3000 genomes of Asian cultivated rice. The named Indica-5 (I5) subpopulation was expanded in Vietnam and contained lowland Indica accessions, which had very low shared ancestry with accessions from any other subpopulation and were previously overlooked as admixtures. We scored phenotypic measurements for nineteen traits and identified 453 unique genotype-phenotype significant associations comprising twenty-one QTLs (quantitative trait loci). The strongest associations were observed for grain size traits, while weaker associations were observed for a range of characteristics, including panicle length, heading date and leaf width.

Conclusions: We showed how the rice diversity within Vietnam relates to the wider Asian rice diversity by using a number of approaches to provide a clear picture of the novel diversity present within Vietnam, mainly around the Indica-5 subpopulation. Our results highlight differences in genome composition and trait associations among traditional Vietnamese rice accessions, which are likely the product of adaption to multiple environmental conditions and regional preferences in a very diverse country. Our results highlighted traits and their associated genomic regions that are a potential source of novel loci and alleles to breed a new generation of low input sustainable and climate resilient rice.

Keywords: Rice, Breeding, Adaptation, QTL, Genetic diversity, GWAS, Landraces

* Correspondence: jose.devega@earlham.ac.uk
[1]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK
Full list of author information is available at the end of the article

Background

Rice production in Vietnam is of great value for export and providing daily food for more than 96 million people. However, agricultural production, especially rice cultivation, is inherently vulnerable to climate variability across all regions in Vietnam. Based on the records of monthly precipitation and temperature from 1975 to 2014 (Nguyen et al. 2019), the areas of highest crop production in the low lying Mekong and Red River Deltas are particularly vulnerable to the increasing threat from climate change. In 2017, the total planted area of rice in Vietnam was 7.7 million hectares. This includes 4.2 million hectares in the Mekong River Delta and 1.1 million hectares in the Red River Delta (GSO-Database 2017). These are also the areas where most of the population of the county is concentrated. In the Mekong River Delta, the damaging effects of salinisation and drought to rice production have increasingly manifested themselves in recent years (Parker et al. 2019; Son et al. 2018; Tran et al. 2019; Yen et al. 2019).

Vietnam possesses a vast diversity of native and traditional rice varieties due to its geographical situation, latitudinal range and diversity of ecosystems (Fukuoka et al. 2003). This diversity constitutes a largely untapped and highly valuable genetic resource for local and international breeding programs. Vietnamese landraces are disappearing as farmers switch to modern elite varieties. To limit this erosion of genetic resources, several rounds of collection of landraces, particularly from the northern upland areas, have been undertaken since 1987. Thousands of rice accessions have been deposited in the Vietnamese National Genebank at the Plant Resources Center (PRC, Hanoi, Vietnam), together with passport information detailing their traditional name and province of origin. One hundred and eighty-two traditional Vietnamese accessions were selected for a genotype by sequencing (GBS) study in 2014 (Phung et al. 2014). This study yielded 25,971 single nucleotide polymorphisms (SNPs) that were used to describe four Japonica and six Indica subpopulations. These subpopulations were classified by region, ecosystem and grain-type using passport information (province and ecosystem) and phenotyping. This dataset had subsequently been used for genome-wide phenotype-genotype association studies (GWAS) relating to root development (Phung et al. 2016), panicle architecture (Ta et al. 2018), drought tolerance (Hoang et al. 2019b), leaf development (Hoang et al. 2019a) and jasmonate regulation (To et al. 2019) and phosphate efficiency (Mai et al. 2020; To et al. 2020).

An international effort to re-sequence Asian rice accessions known as the "3000 Rice Genomes Project" (3K RGP) has provided the rice community with a better understanding of Asian rice diversity and evolutionary history, as well as providing valuable knowledge to enable more efficient use of these accessions for rice improvement (Wang et al. 2018; Wing et al. 2018). However, only 56 of these accessions originated from Vietnam, suggesting that the rice diversity within this country may not be fully captured within the 3K RGP. While the original 3K RGP analysis described nine subpopulations (Wang et al. 2018), subsequent reanalysis had shown that the 3K RGP could be further subdivided into fifteen subpopulations (Zhou et al. 2020).

The primary objective of this study was to gain a clear understanding of the rice population and genome structure in Vietnam, leveraging the vast diversity of rice varieties due to the reasons previously highlighted, and then to contextualize this analysis in the 3K RGP. In addition, we aim to assess the phenotypic variability between subpopulations and expand on previously published QTLs using a larger dataset. For this, we newly sequenced 616 Vietnamese rice accessions using whole-genome sequencing (WGS), most of them being native landraces. One hundred sixty-four of these rice accessions were in common with a previous study [8] based on a genotyping-by-sequencing (GBS) approach. We supplemented this dataset with all 56 Vietnamese genotypes from the 3K RGP to form a native diversity panel with 672 accessions. We analysed this diversity panel of 672 accessions to explore how breeding and environmental pressures have shaped the rice genome in Vietnamese accessions. We also carried out a comprehensive analysis of the population structure of the 3635 rice genomes obtained from joining our diversity panel and the complete 3K RGP datasets. We completed a GWAS on the diversity panel with 672 accessions (and separately for the Japonica and Indica subtypes within it) on thirteen phenotypes, which are available for around two-thirds of the samples. Our results highlight genomic differences and trait associations in traditional Vietnamese landraces, which are likely the product of adaption to multiple environmental conditions and regional culinary preferences in a very diverse country.

Results

Sequencing Rice Diversity from Vietnam

Whole-genome sequencing was carried out on 616 rice accessions. Five hundred eleven of the accessions were obtained from the PRC (Plant Resource Centre, Hanoi, Vietnam, http://csdl.prc.org.vn), together with their passport data, which shows that they were collected from all eight administrative regions of Vietnam (Table S1). The remaining samples were obtained from AGI's collection (Agricultural Genomics Institute, Hanoi, Vietnam). Three reference accessions (Nipponbare, a temperate Japonica; Azucena, a tropical Japonica; and two accessions of IR64, an Indica) obtained from the PRC, were

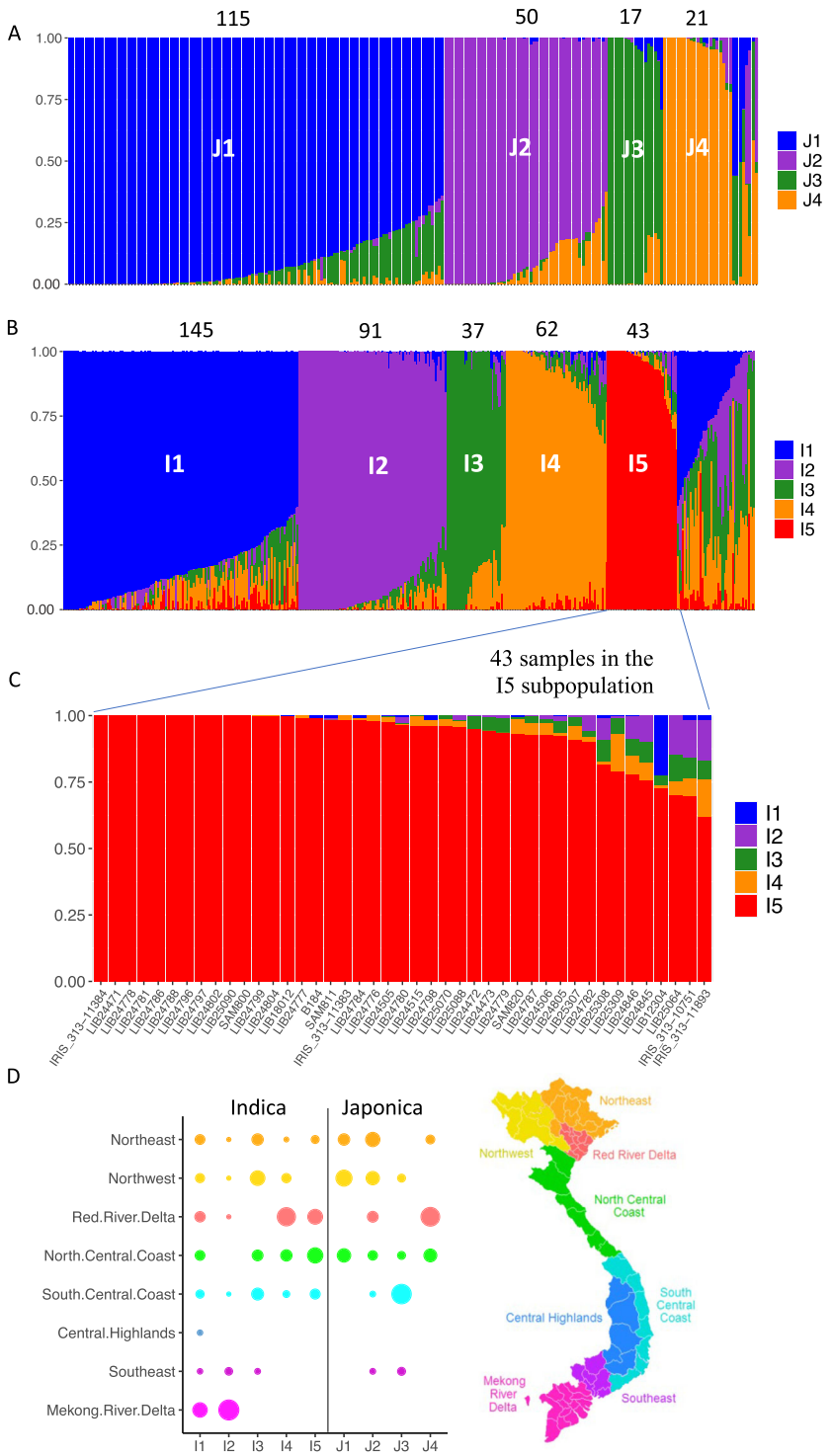

Fig. 1 Population structure and location of the Indica and Japonica subpopulations within Vietnam. **a** STRUCTURE results (mean of 10 replicates) at K = 4 for 211 Japonica subtypes. The cut off for inclusion in each subpopulation is 0.6. The number of samples in each subpopulation is shown above, a further 8 samples were classified as admixed. **b** STRUCTURE results (mean of 10 replicates) at K = 5 for 426 Indica subtypes. Each colour represents one subpopulation. Each accession is represented by a vertical bar and the length of each coloured segment in each bar represents the proportion contributed by each subpopulation. The cut off for inclusion in each subpopulation is 0.6. The number of samples in each subpopulation is shown above, a further 48 samples were classified as admixed. **c** STRUCTURE results for the I5 subpopulation expanded to show individual samples. **d** The proportion of each population originating from each of the 8 regions in Vietnam (based on a subset of 377 samples, 54% of Indica samples and 85% of Japonica samples)

included in the dataset. A total of 1174 Giga base-pairs (Gbps) of data was generated for the 616 samples representing an average sequencing depth of 30x for 36 "high coverage" samples and 3x for 580 "low coverage" samples (Table S1). These 616 newly-sequenced accessions were classified into 379 Indica and 202 Japonica subtypes, with the remaining 35 (including the Aus and Basmati varieties) being classified as admixed, based on the STRUCTURE (Pritchard et al. 2000) output for K = 2 using a subset of 163,393 SNPs.

Population Structure of Rice within Vietnam

The population structure of rice within Vietnam was analysed using the diversity panel of 672 samples, comprising 616 newly sequenced accessions and 56 Vietnamese genotypes from the 3K RGP. We assigned the 672 samples to four Japonica subpopulations and five Indica subpopulations (Table S1) using (i) the population structure information obtained from the STRUCTURE analysis (Fig. 1), (ii) the previous characterisation of a panel of Vietnamese native rice varieties using GBS (Phung et al. 2014), and (iii) the assessment of the optimal number of subpopulations (Fig. S1) using the method described in Evanno et al. (2005). Subpopulations were named as in Phung et al. (2014), except that we considered the I6 subpopulation to be part of the I3 subpopulation. Although the previous study used a limited number of GBS markers, 129 of the 164 common samples were assigned to the same subpopulations in both studies. Most differences were due to samples being classified as admixed in either one of the studies. We classified 48 (11%) of the Indica (Im), and eight (4%) of the Japonica samples (Jm) as admixed. The reference varieties Nipponbare (Temperate Japonica), Azucena (Tropical Japonica), and IR64 (Indica) were classified as J4, J1 and I1, respectively.

Each Indica subpopulation contained shared ancestry (admixed components) with other Indica subpopulations (Fig. 1a). The admixed components are shown in detail for the 43 samples in the I5 subpopulation (Fig. 1c) namely 38 samples from our dataset and the following five samples from the 3 K RGP; IRIS 313–11384 (IRGC 127275), B184 (IRGC 135862), IRIS 313–11383 (IRGC 127274), IRIS 313–10751 (IRGC 127577) and IRIS 313–11893 (IRGC 127519). The Japonica subtropical J1 subpopulation shared ancestry (between 0 and 25% of the genome) with the Japonica tropical J3 subpopulation, whereas the two temperate subpopulations, J2 and J4 shared ancestry dominantly with each other. The tropical J3 subpopulation contained four samples with around 20% of the haplotypes in common with the temperate J4 subpopulation. Using the passport information available from the PRC, the proportion of each subpopulation originating from each of the "administrative regions" of Vietnam is shown in Fig. 1d. Only the I1 and I2 Indica subpopulations were collected from the Mekong River Delta regions, I2 being almost exclusively grown there whereas I1 was more widespread than I2. The I4 and J4 subpopulations were mainly collected from the Red River Delta areas. The J1 and J3 subpopulations were closely related; the J1 subpopulation was predominantly from the North of Vietnam whereas the J3 subpopulation was concentrated around the South-Central Coast region. Small variations in the percentage of reads mapping were observed for each of the subpopulations (Fig. S2).

A Principal Component Analysis (Fig. 2a and b) showed the relationship between these nine Vietnamese subpopulations. Concerning the Vietnamese genotypes

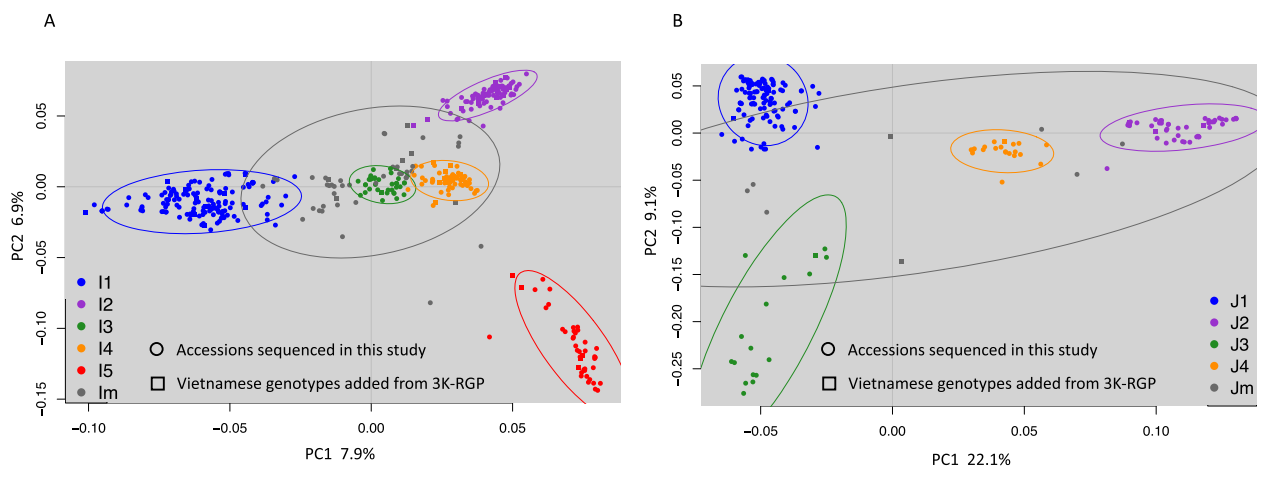

Fig. 2 PCA analysis of Indica and Japonica Vietnamese subpopulations. **a** PCA analysis of 426 accessions from Vietnam using the top two components to separate the five Indica subpopulations. The ellipses show the 95% confidence interval. **b** PCA analysis of 211 accessions from Vietnam using the top two components to separate the four Japonica subpopulations. The ellipses show the 95% confidence interval

from the 3K RGP dataset included in the diversity panel, the Indica I1 subpopulation included two XI-1B modern varieties and eight admixed (XI-adm) accessions. I2 included fourteen XI-3B1 genotypes, which comprises Southeast Asian accessions, and similarly, I3 and I4 included one and ten XI-3B2 genotypes, respectively. Finally, I5 included five XI-adm accessions and clustered distinctly away from all the other subpopulations (Fig. 2a). On the other hand, J1 included the two subtropical (GJ-sbtrp) accessions from the Vietnamese 3K RGP genotypes, and J3 included one tropical (GJ-trp1) accession from the Vietnamese 3K RGP genotypes (Fig. 2b). These results correlate well with the latitudinal distinction between these subpopulations. J2 and J4 included two and one temperate (GJ-tmp) accessions, respectively; and split into two clear subpopulations in Vietnam compared with the East Asian temperate subpopulation described by the 3K RGP.

Population Structure of the Combined 3635 Asian Cultivated Rice Genomes

Six hundred twelve of the 616 newly sequenced accessions from this study and the 3023 accessions from the 3K RGP were combined and classified into 9 and 15 subpopulations (Table S2), and compared with the subpopulations from the 3 K RGP analysis (Wang et al. 2018; Zhou et al. 2020). For clarity, we used the prefix Jap- and Ind- to label these subpopulations from our analysis.

When the combined dataset of 3635 samples was classified into nine subpopulations (Fig. S3a), we found that 95% of the 3 K RGP accessions (2882 out of 3023) were assigned into the same subpopulations. The remaining 5% lines were either (i) previously classified as admixture and our analysis placed into a subpopulation, or (ii) were previously classified in a subpopulation and were now classified as admixture. The 612 newly sequenced Vietnamese accessions were placed in three Indica clusters (187 accessions), three Japonica clusters (176 accessions), the Basmati and Sadri aromatic cB group (11 accessions), or the Aus cA subpopulation (one accession). In more detail, the three Indica clusters included three Im accessions in the East Asian cluster (Ind-1A), seventy-six I1 accessions in the cluster of modern varieties of diverse origins (Ind-1B), and 108 accessions (I2, I3 and Im) in the Southeast Asian cluster (Ind-3). Whereas, the three Japonica clusters included 54 accessions (J2, J4 and Jm) in the primarily East Asian temperate cluster (Jap-tmp), 119 accessions (J1, J3 and Jm) in the Southeast Asian subtropical cluster subpopulation (Jap-sbtrp) and three J3 accessions in the Southeast Asian Tropical subpopulation (Jap-trp). Any remaining accession with admixture components over 65% either Indica or Japonica were classified as Ind-adm (191 accessions) or Jap-adm (27 accessions), respectively. Finally, the remaining accessions were considered as Admix (19 accessions). Notably, all thirty-seven I5 accessions were placed in Ind-adm, and ten of the sixteen J3 accessions were placed in Jap-adm.

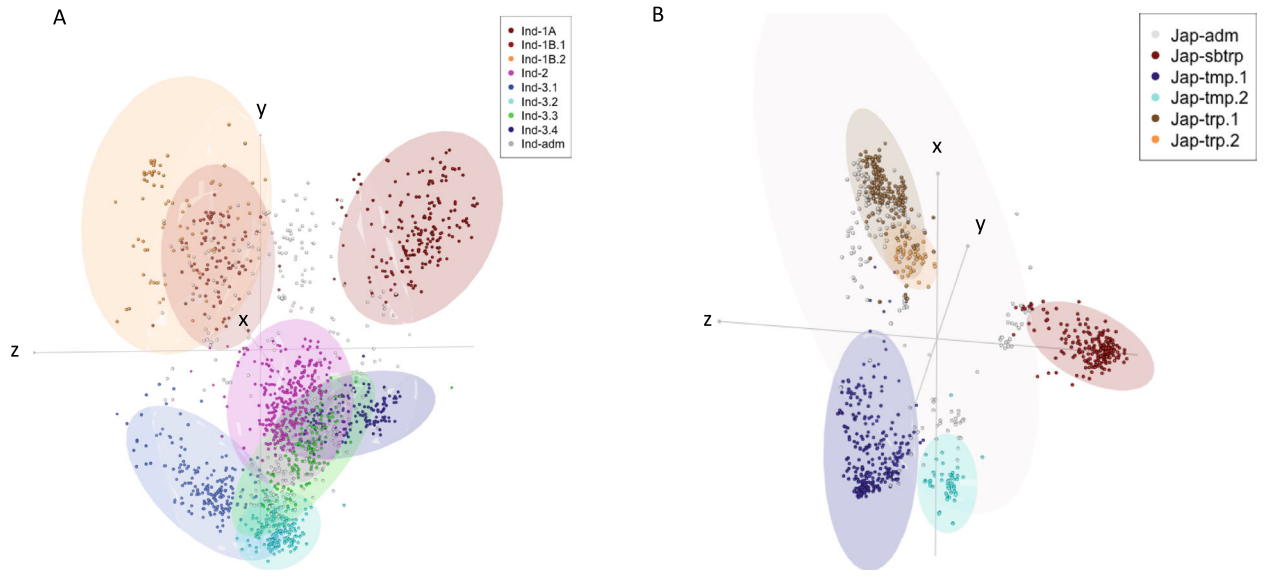

Fig. 3 PCO analysis of Indica and Japonica Vietnamese subpopulations. **a** PCO analysis of 1605 Indica samples (omitting the samples classified as XI-adm and Ind-adm outside Vietnam for clarity). The ellipses show the 95% confidence interval for the K15_new subpopulations (the K15_3KRGP and five Vietnamese Indica subpopulations are shown in Fig. S5). X = PC1, Y=PC4, Z = PC5. **b** PCO analysis of 982 Japonica samples (omitting the samples classified as GJ-adm and Jap-adm outside Vietnam for clarity) showing the K15_new subpopulations (the K15_3KRGP and four Vietnamese Japonica subpopulations are shown in Fig. S6) X = PC3, Y=PC4, Z = PC5

When the combined dataset of 3635 samples was reclassified into 15 subpopulations (K15_new, Fig. S3b), we noticed the following differences in the distribution of subpopulation compared to the 3K RGP analysis for the same number of 15 subpopulations (K15_3KRGP); we did not observe the division of the Aus samples into cA-1 and cA-2, and we subdivided the Indica subtypes and Japonica subtypes into eight and five subpopulations, respectively. A Principle Coordinate (PCO) analysis of the Indica and Japonica subpopulations is shown in Fig. 3, highlighting our new eight Indica and five Japonica subpopulations (In addition the Vietnamese and 3K RGP subpopulations are shown in Figs. S5 and S6).

The relation between the subpopulations in our comprehensive analysis (3635 accessions) and the 3 K RGP (3023 accessions) was as follows: (i) The Ind-1A, Ind-1B.1 and Ind-1B.2 were equivalent to XI-1A, XI-1B1 and XI-1B2, respectively. Forty-three of the Vietnamese I1 accessions were in the Ind-1B.1 subpopulation, and the remaining 102 I1 accessions were classified as admixed. (ii) The Ind-2 was equivalent to XI-2A and XI-2B, and as expected, this geographically distant South Asian subpopulation was not present in Vietnam. (iii) The previously observed split of the Indica-3 subpopulation into 3A and 3B was also observed in our analysis, where Ind-3.1 was equivalent to XI-3A and did not contain any Vietnamese accessions. (iv) The remaining Ind-3.2, Ind-3.3 and Ind-3.4 were a rearrangement of the XI-3B1 and XI-3B2 subpopulations. (v) The 89 Vietnamese I2 accessions belonged to Ind-3.2, which was a subset of XI-3B1. (vi) Ind-3.3 contained 16 of the 37 Vietnamese I3 accessions. (vii) 72% of the accessions in Ind-3.4 were from Vietnam, which contained 13 of the 37 I3 accessions, 61 of the 62 I4 accessions, and all I5 accessions. Within Ind-3.4, the admixture components of I3, I4 and I5 subpopulations (Fig. S7) showed that I3 accessions were highly admixed, some I4 and I5 accessions were completely within Ind-3.4, while other I4 and I5 accessions showed admixture with Ind-3.3 (I5) or Ind.2, Ind-3.2, and Ind-3.3 (I4). To clarify these relations, a principle component analysis (PCA) with a reduced number of accessions was carried out using the 723 sample dataset (672 Vietnamese accessions and 51 genotypes from neighbouring Southeast Asian Countries; Fig. S8), this supported the close relationships of I2 with XI-3B1, I4 with XI-3B2, I5 with XI-adm, J1 with GJ-sbtrp, and that both J2 and J4 were within GJ-tmp.

Phenotypic and Genetic Diversity Analysis of the Vietnamese Indica and Japonica Subpopulations

Phenotypic measurements for 19 traits were scored in field conditions in the Hanoi area by breeders from the Agricultural Genomics Centre (AGI) for approximately two-thirds of the samples in our study. For five of these traits, additional scores were also included from trials by the Vietnamese Plant Resource Centre. In addition, phenotypic data were available for eleven of the traits in 38 of the 56 samples sourced from the 3K-RGP dataset (Table S3, S4). Finally, the grain length to grain width ratio (GL/GW) was calculated to give a total of 20 traits (Table S5). Scores were available for between 328 and 503 of the 672 samples (Indica subpanel, 170–297 samples and Japonica subpanel, 134–178 samples).

Using a t-test, significant differences in measurements were found between the Indica and Japonica subtypes for ten of the traits; these are detailed in Table S5 and histograms are shown in Fig. 4 for selected phenotypes. The Indica subtypes had significantly (p-value < 0.0001) higher values for grain length to width ratio, leaf pubescence, culm number, culm length, and floret pubescence. In contrast, the Japonica subtypes had significantly higher values for grain width, leaf width, flag leaf angle, panicle length, and floret colour. The Indica I1 subpopulation (mostly elite varieties) was the most phenotypically distinct when compared to the rest of the Indica samples (mostly native landraces). I1 samples had longer grains (p-value = 2.2e-16), earlier heading date (p-value = 9.9e-12), higher culm strength (p-value = 2.2e-16), shorter leaf length (p-value = 2.7e-14) and shorter culm length (p-value < 2.2e-16). Similar values were obtained when comparing I1 to just the I5 subpopulation (Fig. 4). The I5 subpopulation was not phenotypically distinct (p-value < 0.001) from the other landrace subpopulations I2, I3 and I4, except for a significantly lower measurement of leaf pubescence (p-value = 0.0007). The Japonica J2 subpopulation had a significantly lower grain length to width ratio than J1 (p-value = 1.8e-13) and J3 (p-value = 5.7e-07). A correlation analysis carried out between the 20 phenotypes (Fig. S9) showed that the highest correlation (r = 0.6) was between leaf length and culm length (excluding the correlation between grain length to width ratio and grain length and grain width). Histogram and correlation plots are available for the 13 traits used for the GWAS analysis in Fig. S10 comparing the Indica and Japonica subtypes and in Fig. S11 comparing subpopulations I1 and I5. Further boxplots showing the phenotypic distribution according to subpopulation for culm length, grain length, grain width and heading date are available in Fig. S12.

The Japonica subtypes had a lower nucleotide diversity (π = 0.000912) than the Indica subtypes (π = 0.00167). Looking at the individual subpopulations (Table S6), the elite I1 subpopulation is the most diverse (π = 0.00144), and the I5 subpopulation is the least diverse (π = 0.00103). Regions of the genome with low diversity in all Indica subpopulations, and regions with low diversity in specific subpopulations, were observed when plotting diversity along each chromosome (Fig. S13). The J3

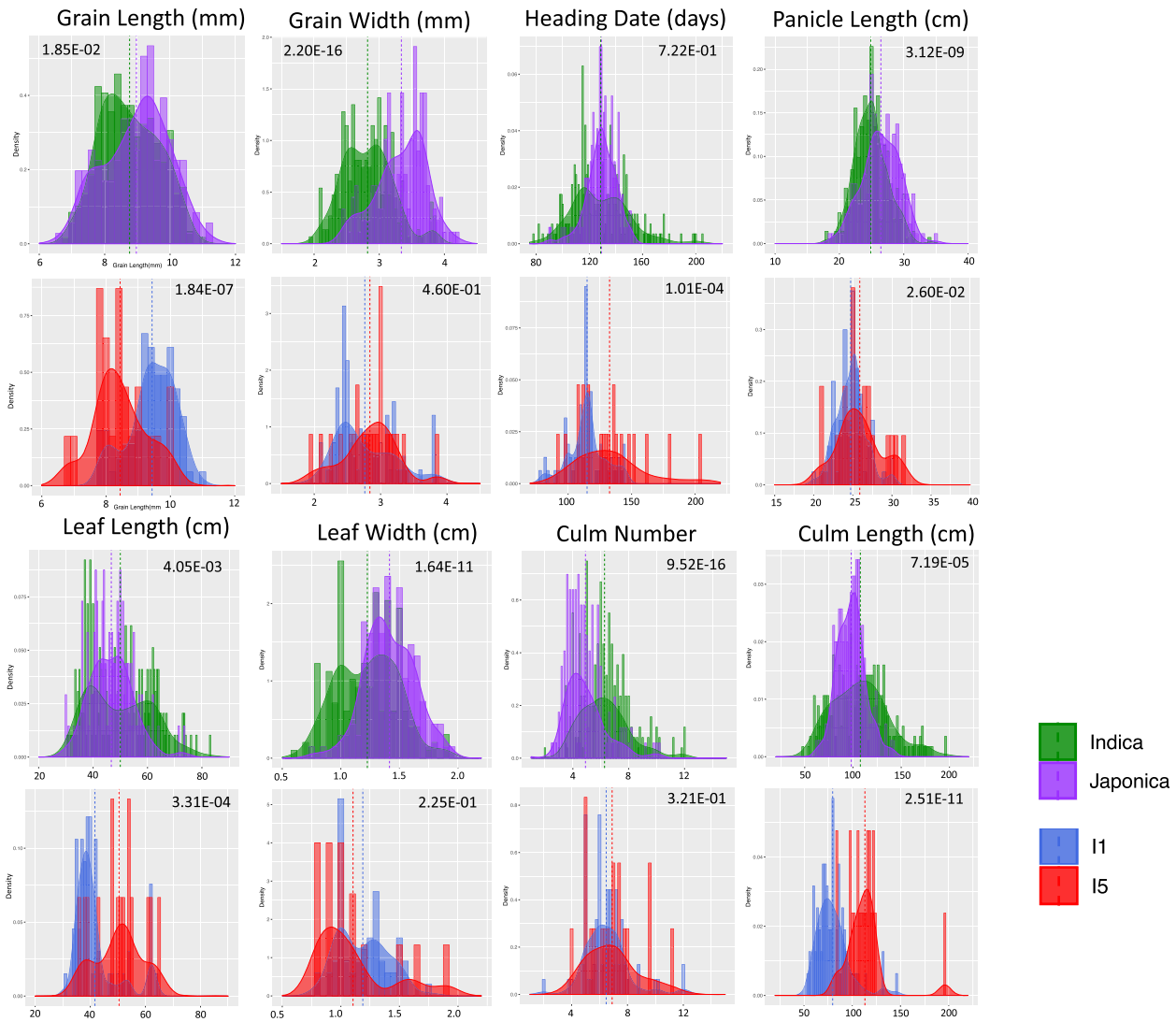

Fig. 4 Histograms comparing the Indica and Japonica subtypes and the I1 and I5 subpopulations. Histogram are shown for 8 of the 13 traits used in the GWAS analysis. The Japonica and Indica subtypes are shown in green and purple respectively and underneath a histogram is shown for a subset of the Indica values comparing subpopulations I1 and I5. The mean value is shown by a dotted line and the *p* value (T-test) is shown at the top of each plot. A ggpairs histogram and correlation plot is available for all 13 traits in Figs. S7 and S8

subpopulation is the most diverse of the four Japonica subpopulations. ($\pi = 0.000697$). Large genomic regions with very low diversity were observed in chromosomes 2, 3, 4 and 5 in all Japonica subpopulations (Fig. S14).

Genome-Wide Genotype-Phenotype Association Analysis

Three independent GWAS were conducted using the full panel (672 samples, 361,191 SNPs), the Indica subpanel (426 samples, 334,935 SNPs) and the Japonica subpanel (211 samples, 122,881 SNPs). Thirteen (13) of the 20 traits were suitable for GWAS based on the variance (Coefficient of Variation < 56% for the full panel). The full list of phenotypic measurements is available in Table S3. We found 643 significant phenotype-genotype associations. These associations were organised into 21 QTLs (Table 1, Table S7). The GWAS Manhattan and Quantile-Quantile plots are available in Fig. S17 and Fig. S18.

We used both the STRUCTURE and kinship matrix to control for population structure, however the QQ-plots suggest that *p*-values are still inflated, especially for the full panel, regarding several traits, such as GL GW ratio, heading date, diameter internode, and floret pubescence. We used a strict cut off of -log10(p) > 8.0 and required at least two significant SNPs per QTL, as detailed in Table 1.

The 21 QTLs contained 1730 genes and covered a total of 11 Mbp over ten chromosomes, and contained

Table 1 21 QTLs identified for plant description traits in the full panel, and Indica and Japonica subpanels. Detailing for the QTL analysis; significance threshold -$\log_{10}$ (*p* value) ≥ 8.0; panel in which significant associations were detected, highest level of significance for all panels and overlaps with publish QTLs for Vietnamese rice populations or for the 3 K RGP

QTL Name	Trait	Chrom	Panel	Segment position (bp)	Sig SNPs nb	min P.value	Number of genes	Overlap with QTLs
1_DI	Diameter_Internode	2	FP	6,805,273 - 6,923,410	3	3.12E-08	18	
2_GL	Grain_Length	2	FP & Jap	15,480,976 - 16,798,043	27	2.69E-12	197	panicle morphology [Ta 2018]
3_GL_jap	Grain_Length	2	Jap	35,638,527 - 35,927,940	4	3.16E-11	58	
4_GW_jap	Grain_Width	3	Jap	3,334,516–3,532,506	3	5.26E-09	34	Leaf Length [Phung 2016]
5_GS	Grain_Length	3	FP & Ind & Jap	16,520,656 - 16,908,475	30	9.26E-17	53	grain width and grain length [Mansueto 2017, Li 2018]
6_GS	Grain_Width	3	FP & Jap	17,686,248 - 20,833,777	355	2.02E-13	471	panicle morphology [Ta 2018]
7_GL	Grain_Length	4	FP	12,043,539 - 13,108,767	14	5.51E-11	167	
8_HD	Heading_Date	4	FP	16,165,354 - 16,384,087	4	1.72E-08	37	
9_PL	Panicle_Length	5	FP	667,557 - 767,557	2	6.17E-08	20	
10_GS	Grain_Width	5	FP & Ind	4,802,345 - 5,383,914	57	2.40E-11	75	grain width and grain length [Mansueto 2017, Li 2018]
11_GL	Grain_Length	6	FP & Ind	1,561,006 - 1,664,716	16	2.68E-10	17	
12_GL	Grain_Length	6	FP & Ind	6,680,831 - 7,190,137	51	1.81E-14	78	
13_GL	Grain_Length	6	FP	7,453,914 - 7,553,914	2	5.90E-08	13	
14_PL	Panicle_Length	6	FP	20,400,110 - 20,500,110	2	2.72E-08	13	
15_GL_jap	Grain_Length	7	Jap	11,519,294–12,296,525	3	5.76E-08	99	
16_FP	Floret_Pubescence	8	FP	18,004,654 - 18,104,654	2	1.64E-08	17	
17_FP	Floret_Pubescence	8	FP	26,175,268 - 26,275,268	2	6.06E-08	15	
18_FP	Floret_Pubescence	9	FP	6,656,837 - 7,940,621	51	7.23E-12	168	
19_HD	Heading_Date	9	FP	14,067,272 - 14,807,406	7	6.86E-09	115	
20_GW_jap	Grain_Width	10	Jap	1,098,998 - 1,404,807	6	3.61E-12	52	
21_LW	Leaf_width	12	FP	17,445,137 - 17,561,823	2	2.14E-09	13	

FP Full panel, *Ind* Indica subpanel, *Jap* Japonica subpanel, *Chrom* Chromosome, *Sig SNPs nb* Number of significant SNPs. References: Ta 2018 (Ta et al. 2018), Phung 2016 (Phung et al. 2016), Mansueto 2017 (Mansueto et al. 2017), Li 2018 (Li et al. 2018)

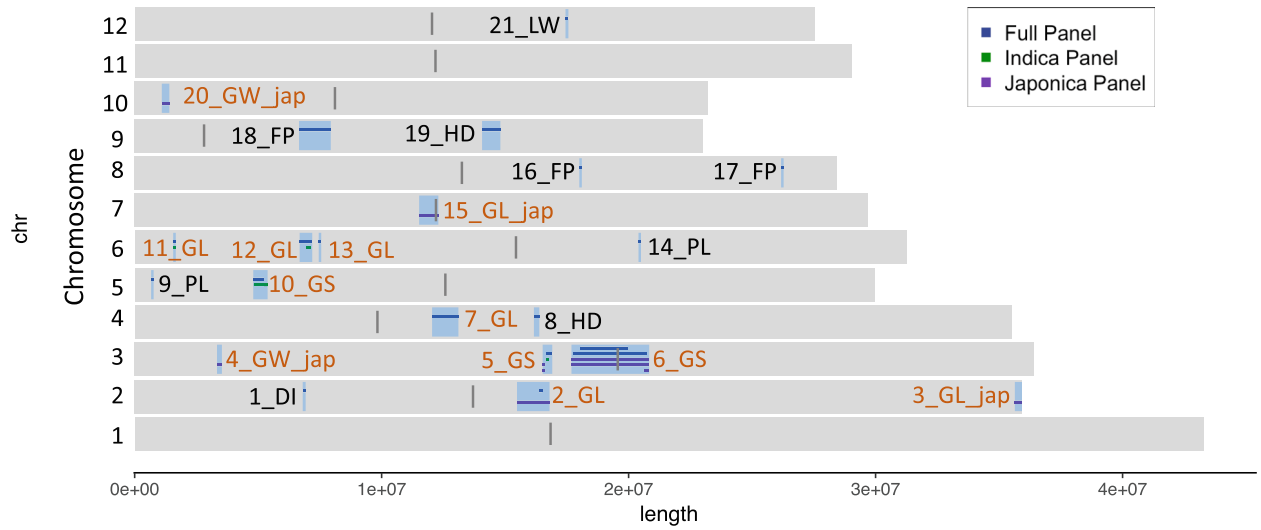

Fig. 5 The distribution of 21 QTL. 21 significant associations for 8 of the 13 traits ($-\log_{10}$ (p value) ≥ 8.0). The 33 individual associations for the full panel and the Japonica and Indica subpanels were merged to form the 21 final QTLs. The QTLs for grain length, grain width and grain length/width ratio were merged into QTLs for grain size, these are labelled in brown. The remaining QTLs are labelled in black; Leaf width (LW), Panicle Length (PL), Heading Date (HD), Floret Pubescence (FP), Diameter Internode (DI). Regions smaller than 100 kb are extended to 50 kb either side of SNP with maximum p value. Centromeric regions are shown as 100 kb regions in dark grey

453 SNPs with a significant association to a trait in at least one diversity panel (Fig. 5). The list of genes within each QTL is available in Table S8.

Seventeen QTLs were identified in the full diversity panel significantly associated with eight traits: grain length, grain width, grain length-to-width ratio, leaf width, panicle length, floret pubescence, heading date and internode diameter. A further 4 QTLs associated with grain length and grain width were observed only in the Japonica subpanel. Three of the QTLs, which were found in the full panel, were also observed in the Indica subpanel.

The set of 3.8 M SNPs (see methods), representing one SNP every 99 bases, was annotated based on the potential effect of each SNP on protein function using SnpEff (Table S11). 526,138 (4.79%) of the SNPs were in genes. There were 21,639 (0.197%) SNPs in 11,125 genes classified as having a putative "*High* impact" effect (E.g. Exon changes, frameshifts, gene fusions or rearrangements, protein structural changes, etc.). Following additional minimal allele frequency (MAF) filtering, in the Indica dataset (MAF 5%, 2,027,294 SNPs), there were 11,906 "*High* impact" SNPs in 7396 genes and in the Japonica dataset (MAF 5%, 1,125,716 SNPs), there were 6240 "*High* impact" SNPs in 4439 genes of which 2818 were present in both Indica and Japonica.

None of the 453 SNPs with a significant association was annotated as resulting in protein changes ("*High* impact" SNPs). However, "*High* impact" effects were identified in other SNPs within the QTL. Among the total 1730 genes in the 21 QTLs, we annotated 309 genes with "*High* impact" SNPs in the Indica subpanel, 248 genes with "*High* impact" SNPs in the Japonica subpanel, including 137 "*High* impact" SNPs common between the two sets. In addition, we looked for overlaps with the QTL in five published Vietnamese studies (Hoang et al. 2019a; Hoang et al. 2019b; Phung et al. 2016; Ta et al. 2018; To et al. 2019), which used 25,971 SNPs in 182 samples (164 in common). We found that 2_GL and 6_GS overlapped with QTL for panicle morphological traits (Ta et al. 2018); 2_GL overlapped with QTL9 for secondary branch number, and spikelet number (SBN and SpN), and 2_GS overlapped with QTL12 for secondary branch average length (SBL). 4_GW_jap overlapped with "q1" for longest leaf length (LLGHT) (Phung et al. 2016).

Discussion

Indica and Japonica Rice Subpopulations within Vietnam

Whole-genome sequencing of 616 Vietnamese rice accessions, predominantly landraces, plus 56 Vietnamese genotypes previously sequenced by the 3K RGP, provides us with a diversity panel to clarify the structure of rice subpopulations in Vietnam. Here, we describe five Indica subpopulations and four Japonica subpopulations using phenotypic measurements from this study, passport information available from the Vietnamese National Genebank (PRC), and the agronomic and geographical annotations from Phung et al. (Phung et al. 2014). In general terms, our population structure within Vietnam agreed with the previous study, which used a smaller

number of markers and 182 samples and is approximately a third of our diversity panel (Phung et al. 2014). Subpopulation I1 is the most phenotypically distinct of the Indica subpopulations and shows typical phenotypes of 'elite' varieties, such as short height, strong culm strength, long slender grains and a short growth-duration (less than 120 days from sowing to harvest). I1 accessions are grown throughout Vietnam in irrigated ecosystems but predominantly in the Mekong River Delta in the south of the country. Subpopulation I2 is mainly composed of long growth-duration (over 140 days), tall varieties grown in the rainfed lowland and irrigated ecosystems of the Mekong River Delta with a broad diversity of grain shapes. The remaining three Indica subpopulations are intermediate between I1 and I2 for growth-duration, height and culm strength, have a broad diversity of grain shapes, and are not grown in the Mekong River Delta. Subpopulation I3 has the highest proportion of upland varieties but also includes some lowland varieties from the "South Central Coast" region many of which were classified as an independent subpopulation (I6) by Phung et al. (Phung et al. 2014). Subpopulation I4 is mainly grown in the rainfed lowland and irrigated ecosystems of the Red River Delta. Subpopulation I5 is grown in a range of ecosystems but concentrated around the North Central Coast and Red River Delta regions, but excluding the Northwest region suggesting that it is the main lowland subpopulation. The J1 and J3 subpopulations are closely related upland varieties and the J2 and J4 subpopulations are closely related lowland varieties. Subpopulation J1 is mostly composed of medium growth-duration upland varieties from the mountainous regions in the North of Vietnam, with long large grains typical of upland varieties. Subpopulation J2 is grown throughout Vietnam in a range of ecosystems but has consistently short grains. Subpopulation J3 is mainly grown in the "South Central Coast" region and has long large grains. Subpopulation J4 is primarily grown in the Red River Delta region in lowland and mangrove ecosystems and has short grains.

Root traits measured by Phung et al. (2016) can be used to gain some information about the drought tolerance of the subpopulations. The J1 and J3 upland subpopulations have deeper and thicker roots than the thinner shallower roots in the J2 and J4 subpopulations, which are grown in irrigated and mangrove ecosystems (Phung et al. 2016). This suggests that the J1 and J3 subpopulations, which are grown mainly in rainfed upland regions, would be more drought tolerant than the others. Similarly, the I3 subpopulation has the deepest and thickest roots. It would, therefore, be more drought tolerant than the I1 and to a lesser extent the I5 subpopulation, which has the thinnest, shallowest root systems.

A Comprehensive Analysis of the Available 3635 Asian Cultivated Rice Genomes

The comprehensive analysis of the combined 3635 Asian cultivated rice genomes obtained by joining our diversity panel with the full 3K RGP dataset resulted in a similar assignation to the previous 3K RGP analysis in 84% of the cases. The largest differences were that the 3K RGP split the cA and XI-2 subpopulations, while our analysis split the GJ-tmp and rearranged the two XI-3B subpopulations into Ind-3.2, Ind-3.3 and Ind-3.4. The single temperate subpopulation (GJ-tmp) from the 3K RGP is further split in our study between the Jap-tmp.1 and Jap-tmp.2 subpopulations, with 88% of the samples in Jap-tmp.2 coming from Vietnam and forming the J2 subpopulation. These differences are likely due to changes in the distribution of genetic variants in subpopulations expanded within Vietnam.

Vietnamese Rice Subpopulations in the Context of the 3K RGP Asian Cultivated Rice Subpopulations

The Indica I1 subpopulation, which contains a high proportion of elite varieties, clustered with the X1-1B1 subpopulation of modern varieties. The Southeast Asian native subpopulations (XI-3B1 and XI-3B2) clustered with the I2 and I4 subpopulations, respectively. I3 appeared to include both XI-3B1 and XI-3B2 accessions. The subpopulations from East and South Asia (XI-1A, XI-2A, XI-2B, XI-3A) had no representatives from Vietnam and fell outside of the Vietnamese subpopulation clusters, as expected. Our four Vietnamese Japonica subpopulations relate to the tropical (J1), subtropical (J3) and temperate (J2 and J4) Japonica subpopulations from the 3K RGP according to their latitudinal origin from South to North Vietnam, respectively.

The most exciting subpopulation is I5. When all 3635 samples were considered, the subpopulation XI-3.4 included half of the I3, all but one of I4 and all I5 Vietnamese accessions, as well as half of the Southeast Asian native XI-3B2 genotypes from the 3K RGP. The remaining XI-3B2 were classified as Indica admix (Ind-adm). However, when only the Vietnamese samples were considered in the analysis, I5 clustered distinctly away from I3 and I4 subpopulations (Fig. 2a) and included five accessions from the 3K RGP, which had very low shared ancestry (admixture components) with other 3K RGP samples. Notably, Vietnamese landrace IRIS 313–11384 (IRGC 127275) had little shared ancestry with any other Vietnamese 3K RGP genotypes. Remarkably, a recent study on genomic signals of admixture and alien introgression in a core collection of 948 accessions representative of the earlier Asian Rice Landraces (Santos et al. 2019) included IRIS 313–10751 (IRGC 127577) and IRIS_313–11383 (IRGC 127274) from the I5 subpopulation.

Genome-Wide Association Analysis in Vietnamese Rice Landraces Highlighted 21 QTL

We have also extended upon seven published GWAS (Hoang et al. 2019a; Hoang et al. 2019b; Mai et al. 2020; Phung et al. 2016; Ta et al. 2018; To et al. 2019; To et al. 2020), which focussed on specific traits but used a smaller number of markers and a third of the samples from the Vietnamese dataset. We took a similar approach of carrying out the analysis on both the full panel and the Indica and Japonica subpanels. There are some interesting overlaps between the QTLs from the various studies, notably, the overlap of QTL for panicle morphology with our QTL for grain size (2_GL and 6_GS). These previous studies found QTL in the full panel and in the Indica subpanel, but not in the Japonica subpanel. However, we found QTL for grain size that were only present in the Japonica subpanel, and all the QTL found in the Indica subpanel were also found in the full panel. These differences probably reflect our larger dataset. Comparing our results with the GWAS results from the 3K RGP (https://snp-seek.irri.org/) (Mansueto et al. 2017; Mansueto et al. 2016), the QTL 5_GS on chromosome 3 is in the same region as a marker associated with grain length, and the QTL 10_GS on chromosome 5 is in the same region as a marker associated with both grain width and grain length. Underlying these two QTL, there are genes that have a putative role in the control of grain size in rice (Li et al. 2018), namely GS3 (Os03g0407400) in 5_GS and GSE5 (LOC_Os05g09520, Os05g0187500) in 10_GS. Functional nucleotide polymorphism (FNP) can be caused by either a SNP or Indel, as has been shown for a number of yield related traits such as grain size (Kim et al. 2016). However, we were only able to detect SNPs in our dataset. We also looked for genes with "*High* impact" SNPs in QTL, relevant candidates include bip130 (Zhou et al. 2019) (LOC_Os05g02260, Os05g0113500) with a stop gain mutation underlying the QTL 9_PL for panicle length and OsSPX-MFS3 (LOC_Os06g03860, Os06g0129400) (Wang et al. 2015) with a splice acceptor variant at the end of an intron underlying the QTL 11_GL for grain length.

Subpopulation I5 Constitutes an Untapped Resource of Cultivated Rice Diversity

The analysis restricted to Vietnamese accessions allowed us to observe differences among the accessions within the country. Although 38 accessions (including two genotypes from the same accession in our study) are deposited in the PRC in Hanoi, and the remaining five accessions are available from the 3 K RGP, there is limited information from the passport and phenotypic data to be able to understand the distinctiveness of this subpopulation fully. Further analysis of this subpopulation should encompass 'Indica specific genes' which may have been overlooked in our study as we used a Japonica reference.

Phung et al. (Phung et al. 2014) described subpopulation I5 as "medium growth-duration accessions from various ecosystems of the North and South Central Coast regions, with rather small and non-glutinous grains". Our I5 accessions are predominantly from the Red River Delta and contiguous coastal departments, the "North Central Coast" and "Northwest" administrative regions, but remarkably excluding the higher altitude Northwest region in the North, the more upper "Central Highlands", as well as the whole Mekong River Delta in the south. This suggests that I5 accessions are common traditional low yielding lowland varieties with specific environmental or culinary values.

Comparing the Vietnamese subpopulations to the fifteen Asian rice subpopulations identified from the 3K RGP highlighted the I5 subpopulation as a potential source of novel variation as it forms a well-separated cluster. Subpopulation I5 originates from lowland areas such as the Red River Delta and adjacent regions. For the range of phenotypes measured in this study, the I5 subpopulation did not differ phenotypically from the other landraces, which have undergone breeding selection within Vietnam. However, compared to the 'elite' I1 subpopulation, I5 accessions have shorter grains, take longer to flower, having lower culm strength, longer culms and leaves.

Conclusions

In this study, we generated a large genome-variation dataset for rice by sequencing 616 accessions from Vietnam and supplemented these with data obtained from the 3K RGP. Using this resource, we incorporated the Vietnamese rice diversity within the population structure of the Asian cultivated rice. Firstly, we incorporated ~ 50 representative samples into our dataset to define the five Indica and four Japonica subpopulations found in Vietnam using STRUCTURE population analysis. We then added a further ~ 50 samples from outside Vietnam and carried out a PCA analysis. Subsequently, we merged both our Vietnamese and the 3K RGP datasets and used admixture for the global population analysis. These approaches showed comparable population structure.

A GWAS analysis yielded associations for grain characteristics, panicle length, heading date and leaf width. These traits are likely under selection and associated with yield. Together with previously published QTLs using a subset of our Vietnamese rice dataset (Hoang et al. 2019a; Hoang et al. 2019b; Phung et al. 2016; Ta et al. 2018; To et al. 2020; To et al. 2019), we obtained a comprehensive set of QTLs that can be used to further

understand the breeding potential of varieties in Vietnam. The locally adapted varieties and subpopulations provide a source of novel alleles that can be exploited in rice breeding to develop a new generation of sustainable 'Green Super Rice' with lower input needs, enhanced nutritional content and suitability for growing on marginal lands (Wing et al. 2018). Also to develop climate-smart rice varieties, which are of particular relevance to Vietnam's high production areas in the low lying deltas, which are currently severely impacted by climate change.

Materials and Methods

Sequencing of 616 Accessions from Vietnam

We sequenced a total of 616 rice accessions, 612 accessions from Vietnam and three reference accessions, Nipponbare, a temperate Japonica; Azucena, a tropical Japonica; and IR64, an Indica (2 samples). Five hundred eleven accessions are available from the Vietnamese National Genebank (PRC) at http://csdl.prc.org.vn (Table S1). All Vietnamese native rice landraces were grown at Dai Dong Experimental Farm (Dai Dong commune, Thach That district, Hanoi, Vietnam) in 2015. The healthy seeds generated from one mature spikelet of the individual plant in each landrace were harvested and dried separately. After that, the selected seeds (35–40 seeds/landrace) were incubated and sown for 2 weeks to collect leaf samples (30 g/sample) for genomic DNA extraction. Total genomic DNA extraction of each rice landrace was made from young leaf tissue using the Qiagen DNeasy kit (Qiagen, Germany). DNA concentration and purity of the samples were measured by the UV-VIS NanoDrop ND-2000 spectrophotometer (Thermo Fisher Scientific) at OD 260/280 nm and OD 260/230 nm wavelengths.

Sequencing was performed by Genomic Services at the Earlham Institute (Norwich, UK). Around 1 μg of genomic DNA from each sample was used to construct a sequencing library. For the 36 high coverage samples (prefix: SAM) the Illumina TruSeq DNA protocol was followed, and the samples were sequenced on the HiSeq 2000 for 100 cycles. For the low coverage samples (prefix: LIB), genomic DNA was sheared to 500 bp using the Covaris S2 Sonicator (Covaris and Life technologies), and samples were processed using the KAPA high throughout Library Prep Kit (Kapa Biosystems, MA, USA). The ends of the DNA were repaired for the ligation of barcoded adapters. The resulting libraries were quality checked, pooled, and quantified by qPCR. The libraries were sequenced on a HiSeq 2500 instrument following the manufacturer's instructions.

Phenotyping

Phenotyping experiments were conducted at the Thach That Experimental Farm of AGI in 2014 and 2015 (Dai Dong commune, Thach That district, Hanoi, Vietnam). The seeds of each rice landrace were incubated in an oven at 45 °C for 5 days to break the seed dormancy. All rice seeds were soaked in tap water for 2 days and incubated at 35-40 °C for 4 days for germinating. The fully germinated seeds of each rice landrace were directly sown in the paddy field plot (1.5m^2 in the area). After 15 days of sowing, 24 seedlings of each landrace were carefully transplanted by hand in field plots (2x4m^2). The fertiliser and pesticide applications were performed following the conventional methods of rice cultivation in Vietnam. The phenotypic and agronomic characteristics were carried out following IRRI's standard evaluation system (IRRI 2002).

In addition, phenotypic data were available for eleven of the traits in 38 of the 56 genotypes sourced from the 3K RGP dataset. These eleven traits were included in our analysis because we did not observe a significant difference (p-value > 0.07) between our dataset and the 3K RGP dataset for the I2 subpopulation (Table S5).

Merging the SNP Called in the Sequenced Materials and the Complete 3 K RGP Dataset

Raw sequencing reads were mapped to the Nipponbare reference genome Os-Nipponbare-Reference-IRGSP-1.0 (IRGSP-1.0), using BWA-MEM with default parameters except for "-M -t 8". Alignments were compressed, sorted and merged using samtools. Picard tools were then used to mark optical and PCR duplicates and add read group information. We used freebayes v1.1.0 for variant calling using default parameters. A total of 21.2 M variants were identified of which 16.4 M were SNPs, and 4.8 M were indels. The resulting VCF file was then filtered for biallelic SNPs with a minimum SNP quality of 30, resulting in 16.0 M variants. PLINK v1.9 was used to convert the VCF into a PLINK BED format. These variants were then combined with the 3K RGP 29 M biallelic SNPs dataset v1.0 by downloading the PLINK BED files from the "SNP-seek" database (https://snp-seek.irri.org) excluding variants on scaffolds and 26,553 SNPs that were flagged as triallelic upon merging, resulting in 36.9 M SNPs. The SNPs present in both datasets were then extracted and filtered using an identical approach to Wang et al. (Wang et al. 2018), resulting in 5.9 M SNPs. For that, PLINK v1.9 "--hardy" (Purcell et al. 2007) was used to obtain observed and expected heterozygosity for 100,000 SNPs. We removed SNPs in which heterozygosity exceeds Hardy–Weinberg expectation for a partially inbred species, with inbreeding coefficient (F) estimated as the median value of "1 – Hobs/Hexp", in which Hobs and Hexp are the observed and expected heterozygosity for SNPs where "Hobs/Hexp < 1" and the minor allele frequency is > 5% and using the cut-off value of 0.479508 for the entire 3622 samples

dataset. A further filtered set of 3.4 M SNPs was obtained by removing SNPs with > 20% missing calls and MAF < 1%. Finally, a core set of 361,279 SNPs was obtained with PLINK by LD pruning SNPs with a window size of 10 SNPs, window step of one SNP and r2 threshold of 0.8, followed by another round of LD pruning with a window size of 50 SNPs, window step of one SNP and r2 threshold of 0.8. Samples with more than 50% missing data in this core set were then removed, resulting in dropping seven newly sequenced samples and one genotype from the 3K RGP dataset.

Population Structure of the Combined 3635 Samples

The population structure was analysed using the ADMIXTURE software (Alexander and Lange 2011) on the SNP set obtained in the previous section. First, ADMIXTURE was run from K = 5 to K = 15 in order to compare it with the analysis from IRRI (Wang et al. 2018; Zhou et al. 2020). For each K, ADMIXTURE was then run 50 times with varying random seeds. Each matrix was then annotated using the subpopulation assignment from the 3K RGP nine subpopulations. Then, up to 10 Q-matrices belonging to the largest cluster were aligned using CLUMPP software (Jakobsson and Rosenberg 2007), these were averaged to produce the final matrix of admixture proportions. Finally, the group membership for each sample was defined by applying a threshold of ≥0.65 to this matrix. Samples with admixture components < 0.65 were classified as follows. If the sum of components for subpopulations within the major groups (Ind and Jap) was ≥0.65, the samples were classified as Ind-adm or Jap-adm, respectively, and the remaining samples were deemed admixed (admix).

Multi-dimensional scaling analysis was performed using the 'cmdscale' function in R (R Core Team, 2020), using a distance matrix obtained in R using the 'Dist' function from the amap package (Lucas 2018). The resulting file was then passed to Curlywhirly (https://ics.hutton.ac.uk/curlywhirly/) and rgl v0.100.19 (https://r-forge.r-project.org/projects/rgl/) for visualisation.

Recalling the Diversity Panel with 723 Samples

The 616 rice samples were mapped to the Japonica Nipponbare (IRGSP-1.0) reference with BWA-MEM using default parameters, duplicate reads were removed with Picard tools (v1.128) and the bam files were merged using SAMtools v1.5 (Li et al. 2009). Variant calling was completed again on the merged bam file with FreeBayes v1.0.2 (Garrison and Marth, 2012) separately for each of the 12 chromosomes, but using the option "--min-coverage 10". Over 6.3 M bi-allelic SNPs with a minimum allele count of ≥3 and quality value above 30 and missing in < 50% of samples were obtained with VCFtools v0.1.13 (Danecek et al. 2011). BAM alignment files to the Nipponbare IRGSP 1.0 reference genome were downloaded from http://snp-seek.irri.org/ (Mansueto et al. 2017; Mansueto et al. 2016) for 107 selected samples. Alignment statistics are included in Table S11. These BAM files were merged and variant calling was similarly completed using FreeBayes v1.0.2 (Garrison and Marth, 2012) separately for each of the 12 chromosomes using the option --min-coverage 10, and filtered with VCFtools v0.1.13 as before to obtain 6.8 M bi-allelic SNPs with a minimum allele count of ≥3 and quality value above 30 and missing in < 50% of samples. The two sets of 6.3 M and 6.8 M SNPs were merged using BCFtools v1.3.1 isec to obtain 4.4 M SNPs which were present in both sets and in at least 70% of samples. These 4.4 M SNPs were then filtered to remove positions which fell outside the expected level of heterozygosity for this dataset, as previously indicated. The resulting estimate of F for the 723 samples was 0.882, so a SNP whose heterozygosity is >5x higher than the most likely value for a given frequency and the dataset's inbreeding rate will be deemed as having an excessive number of heterozygotes. The cut-off value was 0.591, which resulted in 3.8 M SNPs passing this filter, a scatter plot indicating the SNPs which were kept and removed is shown in Fig. S15. Missing data was imputed in this latest dataset using Beagle v4.1 with default parameters (Browning and Browning 2016). A comparison using PCA, between the imputed and non-imputed SNP sets showed that imputation did not change the clustering of these 723 samples (Fig. S16). The 3.8 M SNPs were subsequently filtered for minimum allele frequency (MAF), linkage disequilibrium (LD pruning or filtering), and distance between polymorphisms (thinning) in different subsets of samples to obtain fourteen sets of SNPs that ranged from 59 K to 3.8 M SNPs, which were appropriate for the various downstream analysis described below (Table S11).

Population Structure and Diversity Analysis for the Panel of 672 Vietnamese Samples

SNP sets were filtered for MAF 5%, followed by LD filtering using PLINK --indep-pairwise 50 10 0.2, with further thinning if required. We ran STRUCTURE (Pritchard et al. 2000) v2.3.5 using the default admixture model parameters; each run consisted of 10,000 burn-in iterations followed by 50,000 data collection iterations. STRUCTURE was run using K = 2 for the 616 samples using SNP set 1 (163,393 SNPs). Samples with admixture components < 0.75 were classified as admixed, and the remaining samples were classified as Indica or Japonica. STRUCTURE was run varying the assumed number of genetic groups (K) from 3 to 10 with three runs per K value for the 672 Vietnamese samples (SNP set 9–80,000 SNPs); from 1 to 8 with ten runs per K value for the 426

Indica subtypes from Vietnam (SNP set 10–108,420 SNPs) and the 211 Japonica subtypes from Vietnam (SNP set 11–59,815 SNPs). The output files were visualised using the R package POPHELPER v.2.2.7 (Francis 2017) including the calculation of the number of clusters (K) using the Evanno method (Evanno et al. 2005; Zheng et al. 2012). Using the combined-merged clumpp output from POPHELPER, Indica (K = 5) and Japonica (K = 4) samples were classified into Indica I1 to I5 and Japonica J1 to J4 subpopulations using a threshold of > = 0.6, with the remaining samples being classified as mixed (Im and Jm). The principal component analysis (PCA) was performed using the R package SNPRelate v1.16.0 (Zheng et al. 2012) using method = 'biallelic'. Nucleotide Diversity (π) was measured for each of the subpopulations with VCFtools v0.1.13 using 100-kbp windows and a step size of 10 kbp.

Determining the Effect of SNPs

The effects of all bi-allelic SNPs (low, medium and high effects) on the genome were determined based on the pre-built release 7.0 annotation from the Rice Genome Annotation Project (http://rice.plantbiology. msu.edu/) using SnpEff (Cingolani et al. 2012) release 4.3, with default parameters. The complete set of 3,750,621 SNPs (SNP set 2) which contained on average one variant every 99 bases was annotated. Using sequence ontology terms, the effect of each SNP was classified as described by SnpEff. A summary of the SNP effect analysis is available in Table S10.

Genome-Wide Association Analysis

Three independent analyses were conducted using the full panel (672 samples, 361,191 SNPs), the Indica subpanel (426 samples, 334,935 SNPs) and the Japonica subpanel (211 samples, 122,881 SNPs), SNP sets 12, 13 and 14 respectively (Table S11). The GWAS analysis was performed by employing the R package Genome Association and Prediction Integrated Tool (GAPIT) version 3.0 (Lipka et al. 2012; Tang et al. 2016). The covariate matrix was generated in STRUCTURE. We used the combined-merged output from POPHELPER for the full panel (K = 8), the Indica subpanel (K = 5) and the Japonica subpanel (K = 4). The covariate matrix and the kinship calculated in GAPIT were included in the GWAS model to control for false positives. The SUPER (Settlement of MLM Under Progressively Exclusive Relationship (Wang et al. 2014) method integrated into GAPIT, designed to increase the statistical power, was used to perform the association mapping analysis. The SUPER method was implemented in GAPIT by setting the parameter of "*sangwich.top*" and "*sangwich.bottom*" to *CMLM* and *SUPER*, respectively. A quantile-quantile (Q–Q) plot was used to check if the model was correctly accounting for both confounding variables. Associations held by peaks with $-\log_{10}$ (p-value) ≥ 8.0 (equivalent to a FDR < 0.01) were used to declare the significant associations.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12284-021-00481-0.

Additional file 1: Table S1. Name and details of 672 rice varieties. Detailing read number, mapping statistics, Vietnamese National Genebank number, local name, location, characteristic, subtype and subpopulation. **Table S2**. Name and details of 3635 rice varieties. Detailing the new subpopulation and PCO analysis. **Table S3**. Phenotypic measurements for 20 traits for 672 samples. Detailing individual measurements for each sample, description of phenotypes, statistics for all samples and individually for the Indica and Japonica subtypes. Phenotypes are available for around 75% of the samples. **Table S4**. Phenotype abbreviations and details. **Table S5**. Phenotype statistics (mean and coefficient of variation) and population comparisons (t-test). **Table S6**. Diversity (π) of each subpopulation. **Table S7**. GWAS results. List of the 21 QTL and the positions of the individual QTLs for each panel. **Table S8**. Gene lists for the 21 QTL. **Table S9** List of 107 IRRI rice samples. Detailing IRRI accession, country on origin, K9 and K15 group and Vietnamese subpopulation. **Table S10**. List of 14 SNP sets used for analysis. Detailing filtering parameters, sample and SNP numbers for each SNP set. **Table S11**. Summary count of SNPs with effects on the genome. Detailing SnpEff annotation of the full set of 3,750,621 SNPs using the *Oryza sativa* MSU release 7 rice annotation. Six tables detailing number of effects by impact, functional class, type, region, base changes and Ts/Tv ratio.

Additional file 2: Figure S1. Analysis of STRUCTURE output using the Evanno method. Evanno Plots output from Pophelper for 672 Vietnamese samples, 426 Indica samples and 211 Japonica samples. **Figure S2**. Mapping rate (%properly paired) for Japonica and Indica subpopulations. **Figure S3**. Principal coordinate analysis (PCO) of the 3635 Asian cultivated rice genomes. Plots are coloured by the subpopulations a K9_new, b K15_new. The first component represents the separation between the Indica and Japonica lines. The second components show the separation of cAus and to a lesser extent cBas while the third and fourth components represent the separation within Japonica and Indica respectively. Note for (a) we display the first 3 components and for (b) we display components 1, 2 and 4. **Figure S4**. Comparison between K15_3KRGP, K15_new and Vietnamese subpopulations. a Comparison between K15_3KRGP and K15_new using 3023 samples. b Comparison between K15_new and Vietnamese subpopulations using 668 samples (overlap of 56 samples from Vietnam with a). c Percentage of K15_new subpopulations from Vietnam. Arrow are shown for subpopulations which consist of > 50% of samples from Vietnam. Diagram generated using http://sankeymatic.com/. **Figure S5**. PCO analysis of 1605 Indica samples. Omitting the samples classified as XI-adm and Ind-adm outside Vietnam for clarity. Plot coloured by a K15_3KRGP, b K15_new including Vietnamese samples, c Five Vietnamese Indica subpopulations. The ellipses show the 95% confidence interval. X = PC1, Y=PC4, Z = PC5. Figure generated using rgl https://r-forge.r-project.org/projects/rgl/. **Figure S6**. PCO analysis of 982 Japonica samples. Omitting the samples classified as GJ-adm and Jap-adm outside Vietnam for clarity. Plot coloured by a K15_3KRGP, b K15_new including Vietnamese samples, c Four Vietnamese Japonica subpopulations. The ellipses show the 95% confidence interval. X = PC3, Y=PC4, Z = PC5. Figure generated using rgl https://r-forge.r-project.org/projects/rgl/. **Figure S7**. Admixture components of the Indica I3, I4 and I5 subpopulations. **Figure S8**. PCA analysis of Indica and Japonica Vietnamese subpopulations including 51 genotypes from outside Vietnam. a PCA analysis of 445 accessions using the top two components to separate the five Indica subpopulations. The ellipses show the 95% confidence interval. b PCA analysis of 233 accessions using the top two components to separate the four Japonica subpopulations. The ellipses show the 95% confidence interval. **Figure S9**. Correlation between the

20 phenotypes. **Figure S10**. Correlation between Indica and Japonica for the 13 phenotypes used for GWAS. The figure was created using "ggpairs" package in R. **Figure S11**. Correlation between Indica I1 and I5 subpopulations for the 13 phenotypes used for GWAS. The figure was created using "ggpairs" package in R. **Figure S12**. Boxplots showing the Phenotypic distribution per subpopulation for Culm Length, Grain Length, Grain Width and Heading Date. **Figure S13**. Indica subpopulation diversity. Diversity (π) plotted along the 12 rice chromosomes in sliding 100 kb windows. **Figure S14**. Japonica subpopulation diversity. Diversity (π) plotted along the 12 rice chromosomes in sliding 100 kb windows. **Figure S15**. SNP filtering for heterozygosity. Proportion of heterozygous calls versus allele frequency. Each dot represents a SNP from a random sample of 100,000 SNPs. The points have an opacity of 5% to highlight regions of higher point density. The bulk of the SNPs lie on the Hardy-Weinberg equilibrium curve scaled by a factor of around 0.118, which implies a Wright's inbreeding coefficient of F = 0.882. The SNPS have been filtered using cut off of 0.592 (5*(1-F)), the corresponding SNPs which are kept and removed are shown on the plot. **Figure S16**. PCA analysis of 723 samples before and after imputation. Comparing the 2,690,005 not imputed SNP set 3 to the 2,665,825 imputed SNP set 4 Both SNP set were filtered for 5% MAF. Using PC1 and PC2 to separate the Japonica subpopulations. Using PC3 and PC4 to separate the Indica subpopulations.

Additional file 3: Figure S17. GWAS Manhattan and qq plots for the full panel and Indica and Japonica subpanels for Grain Length, Grain Width, Grain length-to-width ratio, Heading Date, Culm Strength, Leaf Length and Leaf Width.

Additional file 4: Figure S18. GWAS Manhattan and qq plots for the full panel and Indica and Japonica subpanels for Leaf Pubescence, Culm Number, Diameter Internode, Culm Length, Panicle Length and Floret Pubescence.

Acknowledgements
We thank Professor Giles Oldroyd for his contributions to the conception of this project. We are grateful for the support from Dr. Nelzo Ereful, and Matt Heaton during outreach activities in Vietnam, and Dr. Luca Venturi, Dr. Ricardo Ramirez Gonzalez, Dr. Graham Etherington for their support during summer training activities in the UK, and Dr. Chris Watkins, Dr. Helen Chapman and the Genomics Pipelines team at the Earlham Institute for the sequencing support.

Authors' Contributions
TDK, KHT, AH, SD, LHH, MC and JDV designed and conceived the research. TDK, KHT, TDD, NTPD, NTK, DTTH, NTD, KTD, CNP, TTT, NTT, HDT, NTT, HTG, TKN, CDT, SVL, LTN, NVG and LHH performed the phenotyping and laboratory experiments. JH and BS performed the data analysis with assistance from TDD, NTPD, DTTH, NTD, KTD, NTT, LTN, TDX, MC and JDV. JH, BS and JDV wrote the paper. All authors read and approved the final manuscript.

Funding
The author(s) acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation; this research was funded by the BBSRC Core Strategic Programme Grant (Genomes to Food Security) BB/CSP1720/1 and its constituent work package BBS/E/T/000PR9818 (WP1 Signatures of Domestication and Adaptation); and BBSRC's grants BB/N013735/1 (Newton Fund), and the Newton Fund Institutional Links (Project 172732508), which is managed by the British Council.

Availability of Data and Materials
All sequence data used in this manuscript have been deposited as study PRJEB36631 in the European Nucleotide Archive.

Declarations

Ethics Approval and Consent to Participate
Not applicable.

Consent for Publication
Not applicable.

Competing Interests
The authors declare that there is no conflict of interest regarding the publication of this article.

Author details
[1]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK. [2]NIAB, 93 Lawrence Weaver Road, Cambridge CB3 0LE, UK. [3]Agriculture Genetics Institute (AGI), Hanoi, Vietnam. [4]Vietnam National University of Agriculture, Hanoi 131000, Vietnam. [5]Faculty of Biotechnology, Nguyen Tat Thanh University, Ho Chi Minh 72820, Vietnam. [6]Faculty of Pharmacy, Duy Tan University, Da Nang 550000, Vietnam. [7]Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam. [8]Plant Resource Center, An Khanh, Hoai Duc, Hanoi 152900, Vietnam. [9]Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima 739-8529, Japan.

Received: 1 September 2020 Accepted: 7 April 2021
Published online: 10 June 2021

References
Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinform 12(1):246. https://doi.org/10.1186/1471-2105-12-246

Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. Am J Hum Genet 98(1):116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Cingolani P, Platts A, Wangle L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6(2):80–92. https://doi.org/10.4161/fly.19695

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14(8):2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Francis RM (2017) Pophelper: an R package and web app to analyse and visualize population structure. Mol Ecol Resour 17(1):27–32. https://doi.org/10.1111/1755-0998.12509

Fukuoka S, Alpatyeva NV, Ebana K, Luu NT, Nagamine T (2003) Analysis of Vietnamese rice germplasm provides an insight into japonica rice differentiation. Plant Breed 122(6):497–502. https://doi.org/10.1111/j.1439-0523.2003.00908.x

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing arXiv, p 1207.3907

GSO-Database (2017) General statistic Office in Vietnam, database

Hoang GT, Gantet P, Nguyen KH, Phung NTP, Ha LT, Nguyen TT, Lebrun M, Courtois B, Pham XH (2019a) Genome-wide association mapping of leaf mass traits in a Vietnamese rice landrace panel. PLoS One 14(7):e0219274. https://doi.org/10.1371/journal.pone.0219274

Hoang GT, Van Dinh L, Nguyen TT, Ta NK, Gathignol F, Mai CD, Jouannic S, Tran KD, Khuat TH, Do VN, Lebrun M, Courtois B, Gantet P (2019b) Genome-wide association study of a panel of Vietnamese Rice landraces reveals new QTLs for tolerance to water deficit during the vegetative phase. Rice (N Y) 12(1):4. https://doi.org/10.1186/s12284-018-0258-6

IRRI (2002) Standard Evaluation System for Rice. http://www.knowledgebank.irri.org/images/docs/rice-standard-evaluation-system.pdf

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23(14):1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Kim S-R, Ramos J, Ashikari M, Virk PS, Torres EA, Nissila E, Hechanova SL, Mauleon R, Jena KK (2016) Development and validation of allele-specific SNP/indel markers for eight yield-enhancing genes using whole-genome sequencing strategy to increase yield potential of rice, Oryza sativa L. Rice 9(1)

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
Li N, Xu R, Duan P, Li Y (2018) Control of grain size in rice. Plant Reprod 31(3): 237–251. https://doi.org/10.1007/s00497-018-0333-6
Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28(18):2397–2399. https://doi.org/10.1093/bioinformatics/bts444
Lucas A (2018) Amap: another multidimensional analysis https://CRAN.R-project.org/package=amap
Mai NTP, Mai CD, Nguyen HV, Le KQ, Duong LV, Tran TA, To HTM (2020) Discovery of new genetic determinants of morphological plasticity in rice roots and shoots under phosphate starvation using GWAS. J Plant Physiol 257:153340
Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, Sanciangco M, Palis K, Copetti D, Poliakov A, Dubchak I, Solovyev V, Wing RA, Hamilton RS, Mauleon R, McNally KL, Alexandrov N (2017) Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res 45(D1): D1075–D1081. https://doi.org/10.1093/nar/gkw1135
Mansueto L, Fuentes RR, Chebotarov D, Borja FN, Detras J, Abriol-Santos JM, Palis K, Poliakov A, Dubchak I, Solovyev V, Hamilton RS, McNally KL, Alexandrov N, Mauleon R (2016) SNP-seek II: a resource for allele mining and analysis of big genomic data in Oryza sativa. Curr Plant Biol 7-8:16–25. https://doi.org/10.1016/j.cpb.2016.12.003
Nguyen DK, Ancev T, Randall A (2019) Evidence of climatic change in Vietnam: some implications for agricultural production. J Environ Manag 231:524–545. https://doi.org/10.1016/j.jenvman.2018.10.011
Parker L, Bourgoin C, Martinez-Valle A, Laderach P (2019) Vulnerability of the agricultural sector to climate change: the development of a pan-tropical climate risk vulnerability assessment to inform sub-national decision making. PLoS One 14(3):e0213641. https://doi.org/10.1371/journal.pone.0213641
Phung NT, Mai CD, Hoang GT, Truong HT, Lavarenne J, Gonin M, Nguyen KL, Ha TT, Do VN, Gantet P, Courtois B (2016) Genome-wide association mapping for root traits in a panel of rice accessions from Vietnam. BMC Plant Biol 16(1):64. https://doi.org/10.1186/s12870-016-0747-y
Phung NT, Mai CD, Mournet P, Frouin J, Droc G, Ta NK, Jouannic S, Le LT, Do VN, Gantet P, Courtois B (2014) Characterization of a panel of Vietnamese rice varieties using DArT and SNP markers for association mapping purposes. BMC Plant Biol 14(1):371. https://doi.org/10.1186/s12870-014-0371-7
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559–575. https://doi.org/10.1086/519795
R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna https://www.R-project.org/
Santos JD, Chebotarov D, McNally KL, Bartholome J, Droc G, Billot C, Glaszmann JC (2019) Fine scale genomic signals of admixture and alien introgression among Asian Rice landraces. Genome Biol Evol 11(5):1358–1373. https://doi.org/10.1093/gbe/evz084
Son NY, Yen BT, Sebastian LS. (2018) Development of climate-related risk maps and adaptation plans (climate smart MAP) for Rice production in Vietnam's Mekong River Delta CCAFS working paper 220. https://hdl.handle.net/10568/90253
Ta KN, Khong NG, Ha TL, Nguyen DT, Mai DC, Hoang TG, Phung TPN, Bourrie I, Courtois B, Tran TTH, Dinh BY, La TN, Do NV, Lebrun M, Gantet P, Jouannic S (2018) A genome-wide association study using a Vietnamese landrace panel of rice (Oryza sativa) reveals new QTLs controlling panicle morphological traits. BMC Plant Biol 18(1):282. https://doi.org/10.1186/s12870-018-1504-1
Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F, Su Z, Pan Y, Liu D, Lipka AE, Buckler ES, Zhang Z (2016) GAPIT version 2: an enhanced integrated tool for genomic association and prediction. Plant Genome 9:2
To HTM, Le KQ, Van Nguyen H, Duong LV, Kieu HT, Chu QAT, Tran TP, Mai NTP (2020) A genome-wide association study reveals the quantitative trait locus and candidate genes that regulate phosphate efficiency in a Vietnamese rice collection. Physiol Mol Biol Plants 26(11):2267–2281. https://doi.org/10.1007/s12298-020-00902-2
To HTM, Nguyen HT, Dang NTM, Nguyen NH, Bui TX, Lavarenne J, Phung NTP, Gantet P, Lebrun M, Bellafiore S, Champion A (2019) Unraveling the genetic elements involved in shoot and root growth regulation by Jasmonate in Rice using a genome-wide association study. Rice (N Y) 12:69
Tran TV, Tran DX, Myint SW, Huang CY, Pham HV, Luu TH, Vo TMT (2019) Examining spatiotemporal salinity dynamics in the Mekong River Delta using Landsat time series imagery and a spatial regression approach. Sci Total Environ 687:1087–1097. https://doi.org/10.1016/j.scitotenv.2019.06.056
Wang C, Yue W, Ying Y, Wang S, Secco D, Liu Y, Whelan J, Tyerman SD, Shou H (2015) Rice SPX-major facility Superfamily3, a vacuolar phosphate efflux transporter, is involved in maintaining phosphate homeostasis in Rice. Plant Physiol 169(4):2822–2831. https://doi.org/10.1104/pp.15.01005
Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z (2014) A SUPER powerful method for genome wide association study. PLoS One 9(9):e107684. https://doi.org/10.1371/journal.pone.0107684
Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann JC, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557(7703):43–49. https://doi.org/10.1038/s41586-018-0063-9
Wing RA, Purugganan MD, Zhang Q (2018) The rice genome revolution: from an ancient grain to green Super Rice. Nat Rev Genet 19(8):505–517. https://doi.org/10.1038/s41576-018-0024-z
Yen BT, Quyen NH, Duong TH, Van Kham D, Amjath-Babu TS, Sebastian L (2019) Modeling ENSO impact on rice production in the Mekong River Delta. PLoS One 14(10):e0223884. https://doi.org/10.1371/journal.pone.0223884
Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28(24):3326–3328. https://doi.org/10.1093/bioinformatics/bts606
Zhou X, Ni L, Liu Y, Jiang M (2019) Phosphorylation of bip130 by OsMPK1 regulates abscisic acid-induced antioxidant defense in rice. Biochem Biophys Res Commun 514(3):750–755. https://doi.org/10.1016/j.bbrc.2019.04.183
Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, Mohammed N, Al-Bader N, Sobel-Sorenson C, Parakkal P, Arbelaez LJ, Franco N, Alexandrov N, Hamilton NRS, Leung H, Mauleon R, Lorieux M, Zuccolo A, McNally K, Zhang J, Wing RA (2020) A platinum standard pan-genome resource that represents the population structure of Asian rice. Sci Data 7(1):113. https://doi.org/10.1038/s41597-020-0438-2

Publisher's Note