




ORIGINAL ARTICLE

Open Access



Explore the genetics of weedy traits using rice 3K database

Yu-Lan Lin¹, Dong-Hong Wu² , Cheng-Chieh Wu^{3,4}  and Yung-Fen Huang^{1*} 

Abstract

Background: Weedy rice, a conspecific weedy counterpart of the cultivated rice (*Oryza sativa* L.), has been problematic in rice-production area worldwide. Although we started to know about the origin of some weedy traits for some rice-growing regions, an overall assessment of weedy trait-related loci was not yet available. On the other hand, the advances in sequencing technologies, together with community efforts, have made publicly available a large amount of genomic data. Given the availability of public data and the need of “weedy” allele mining for a better management of weedy rice, the objective of the present study was to explore the genetic architecture of weedy traits based on publicly available data, mainly from the 3000 Rice Genome Project (3K-RGP).

Results: Based on the results of population structure analysis, we have selected 1378 individuals from four sub-populations (*aus*, *indica*, *temperate japonica*, *tropical japonica*) without admixed genomic composition for genome-wide association analysis (GWAS). Five traits were investigated: awn color, seed shattering, seed threshability, seed coat color, and seedling height. GWAS was conducted for each sub-population × trait combination and we have identified 66 population-specific trait-associated SNPs. Eleven significant SNPs fell into an annotated gene and four other SNPs were close to a putative candidate gene (± 25 kb). SNPs located in or close to *Rc* were particularly predictive of the occurrence of seed coat color and our results showed that different sub-populations required different SNPs for a better seed coat color prediction. We compared the data of 3K-RGP to a publicly available weedy rice dataset. The profile of allele frequency, phenotype-genotype segregation of target SNP, as well as GWAS results for the presence and absence of awns diverged between the two sets of data.

Conclusions: The genotype of trait-associated SNPs identified in this study, especially those located in or close to *Rc*, can be developed to diagnostic SNPs to trace the origin of weedy trait occurred in the field. The difference of results from the two publicly available datasets used in this study emphasized the importance of laboratory experiments to confirm the allele mining results based on publicly available data.

Keywords: *Oryza sativa* L., Weedy rice (WR), 3K Rice Genome Project (3K-RGP), Genome-wide association study (GWAS), SNP-Seek

Background

Weedy rice (WR), a conspecific weedy counterpart of the cultivated rice (*Oryza sativa* L.), is probably the most intractable problem to deal with in rice-growing countries. Herbicide inapplicability, as well as the

morphological resemblance, have made WR management an uphill struggle for rice farmers around the world (Olofsdotter et al. 2000). Each year, WR causes important economic loss in rice production due to the reduction in yield and in grain quality (Ziska et al. 2015). Although sharing a high level of genetic similarities with cultivated rice, WR is often characterized by some wild-like traits, such as rapid seedling growth, the presence of awns, high degree of shattering, red seed coat color, and strong dormancy, which together contribute to increasing WR's

*Correspondence: huangy@ntu.edu.tw

¹ Department of Agronomy, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd, Da'an Dist., Taipei 10617, Taiwan
Full list of author information is available at the end of the article

ability to survive in the field (Ziska et al. 2015). Indeed, such “weedy” traits and the domestication syndromes are like the two sides of a coin. While cultivated rice has undergone intense selection against shattering to reduce grain loss during harvest, shattering habit of WR facilitates seed dispersal in the field (Delouche et al. 2007a). WR possess awns to avoid grain predation by animals (Delouche et al. 2007b) while cultivated rice is mainly characterized by awn-less grains. Red seed coat, or pericarp, is common to WR after which “weedy red rice” is named although colored seed coat is also observed in some cultivars. A number of identified genes know to contribute to the pigmentation of rice pericarp, among which a major factor is *Rc*, a bHLH protein involving in proanthocyanidin synthesis (Furukawa et al. 2007; Sweeney et al. 2006). Potentially descended from cultivated ancestors, the weedy traits that WR possesses are likely to be derived from standing variation in wild or cultivated rice (Huang et al. 2018; Vigueira et al. 2019).

The occurrence of WR is generally acknowledged as polyphylogenetic. For example, the U.S. has suffered serious WR infestation which mainly comprised of two morphologically different ecotypes, the black-hull awned (BHA) and the straw awn-less (SH) populations (Londo and Schaal 2007). Studies based on phylogenetic analysis through nucleic and cytoplasmic DNA evidence suggested that BHA and SH were possibly derived from *aus* and *indica*, respectively (Reagon et al. 2010). None of the cultivars grown in the U.S. belongs to these two groups, indicating that BHA and SH may be originated from stock seeds contamination or from escaped breeding materials (Olsen et al. 2007). On the other hand, WR in northern China is most likely originated from hybridization between local *japonica* landraces whereas WR in southern China was genetically similar to *indica*, the most cultivated sub-species in the south of China (Sun et al. 2019). Previous studies have also implied that wild rice may have participated in the formation of WR through inter-specific hybridization (Song et al. 2014; Vigueira et al. 2019). To date, most of the WR genetic studies focused on its occurrence using whole genome sequencing (Huang et al. 2018; Li et al. 2017; Qiu et al. 2017; Vigueira et al. 2019). Loci related to weedy traits were identified through genome-wide selection signature analysis with a focus on known domestication-related genes such as *Bh4*, *PROG1*, *Rc*, *sh4* (Huang et al. 2018; Li et al. 2017; Qiu et al. 2017). The actual phenotype-genotype studies on weedy traits were relatively few (Nguyen et al. 2019; Ye et al. 2013). However, the knowledge on weedy trait genetics would help in a better WR management.

Rapid development of next-generation sequencing (NGS) technologies has brought plant genetics into a new

era (Bräutigam and Gowik 2010; Varshney et al. 2014). Large amounts of data have been generated and shared openly with the scientific community. In rice, in addition to the reference genome (cv. Nipponbare, IRGSP-1.0), many genomic tools are available. SNP-Seek database (<https://snp-seek.irri.org/>) is a major outcome of the 3000 rice genome project (3K-RGP). SNP-Seek harbors genotypic and phenotypic data for more than 3000 accessions of cultivated rice (Mansueto et al. 2016a, b), has provided valuable resources for genomic discovery of *Oryza sativa*, and has opened the doors for large-scale association studies and for an efficient molecular breeding (Angira et al. 2019; Kumar et al. 2020; Leung et al. 2015; Mansueto et al. 2016b; Tang et al. 2019; Tatarinova et al. 2016). Indeed, whole genome association results based on the full 3K-RGP data are available on SNP-Seek. Meanwhile, if one is interested in sub-population specific trait-marker association, a re-analysis may be necessary.

Given the availability of public data and the need of allele mining for weedy traits in view of a better WR management, the objective of the present study was to explore the genetic architecture of rice weedy traits based on publicly available data. We first investigated the abundant data from 3K-RGP and then cross-checked the results with another WR panel, also obtained from public resources.

Methods

Source data

Both phenotypic data and genotypic data of the 3000 Rice Genome Project (3K-RGP) were downloaded from the Rice SNP-Seek Database (<http://snp-seek.irri.org>). The full set of phenotypes consisted of 47 agronomic traits recorded in ordinal scale between 1994 and 2010, maintained in the International Rice Genebank Collection Information System (IRGCIS). For our purpose, we have selected five traits which are often viewed as notable features distinguishing WR from cultivated rice (Meyer and Purugganan 2013): awn color (AWCO_REV), degree of panicle shattering (PSH), panicle threshability (PTH), seed coat color (SCCO_REV), and seedling height (SDHT_CODE). We chose to follow the trait abbreviation according to IRGCIS for an easier cross comparison between studies. For genotypic data, we have chosen the 404K core SNP set. The 404K core SNPs consisted of 404,388 bi-allelic SNPs across 3024 rice accessions, generated through a two-step linkage disequilibrium (LD) pruning procedure to reduce the number of markers at a minor allele frequency (MAF) > 0.01 and a call rate \geq 0.8. Detailed data filtering process is described at http://snp-seek.irri.org/_download.zul. As different datasets were generated for different analyses, details for different

datasets will be provided accordingly. Meanwhile, a summary genotypic data workflow can be found in Fig. 1.

To compare the trait-associated markers identified in rice 3K dataset to those identified in WR background, we have obtained phenotypes and sequence reads of 205 WR accessions from two studies (Li et al. 2017; Qiu et al. 2017), among which 198 were unique. These accessions mainly encompassed four different geographical regions within China (162 accessions), the U.S. (42 accessions), and South Korea (1 accession). Raw sequence reads were retrieved from NCBI (SRR4334499 and PRJNA295802). We performed variant calling in accordance with the SNP discovery pipeline described in Mansueto et al. (2016a, b): raw sequence reads were first mapped to *japonica* reference genome (cv. Nipponbare, IRGSP-1.0) via Burrows-Wheeler Alignment tool (BWA) version 0.7.17-r1188 (Li and Durbin 2009) with default settings after removing adapter sequences. The resulting alignments were processed using MarkDuplicates of Picard (<http://broadinstitute.github.io/picard/>) and IndelRealigner implemented in The Genome Analysis Toolkit (GATK) version 3.8-0-ge9d806836 (McKenna et al. 2010). Final variant calling from the 198 WR samples were performed using BCFTools version 1.10.2 (<http://www.sanger.ac.uk/science/tools/samtools-bcftools-htslib>). We used BEAGLE (Browning et al. 2018) to impute missing genotype with default parameters.

Population structure and linkage disequilibrium

Population structure was inferred using both principle component analysis (PCA) and the model-based maximum likelihood approach implemented in software ADMIXTURE v1.23 (Alexander et al. 2009). Data for population structure inference was prepared as follows.

We first excluded accessions with no phenotypic data available for all five target traits. We then used PLINK v1.9 (Purcell et al. 2007) for the subsequent SNP filtering and pruning. We applied the 1.5× interquartile range to remove outliers in terms of missing rate at per accession level (>0.08817) and at per SNP level (>0.1408). SNPs with MAF ≥ 0.05 were retained. In order to generate independent SNP for population structure analysis, we pruned SNPs at a window of 50 SNPs. SNPs showing pair-wise r^2 higher than 0.2 were removed and the pruning was advanced along the genome at a step of five SNPs. This resulted in a dataset of 2883 individuals × 14,462 SNPs (Fig. 1).

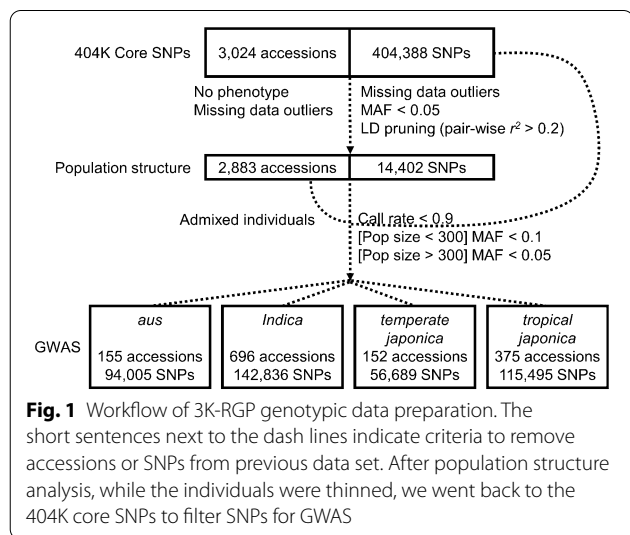
PCA was performed in GCTA (Yang et al. 2011). For ADMIXTURE, K=1 to 15 were tested. Since the cross-validation error barely differed between K values, we defined K=5 based on the prior knowledge on sub-populations within the rice 3K samples (Wang et al. 2018). To avoid the interference of population structure in marker-trait association, only accessions belonging clearly to one sub-population (defined as proportion of genome from a single source ≥ 0.8) were retained for further analyses. This resulted in a final of 1378 accessions divided into four sub-populations (Fig. 1): *aus* (155), *indica* (696), *temperate japonica* (152), and *tropical japonica* (375).

LD was estimated as the squared correlation coefficient of allele state (r^2) within each sub-population using PLINK. Pairwise r^2 values were computed between all SNPs within 300 kb along the same chromosome according to the study of Wang et al. (2018). Average r^2 was calculated for each 1 kb-bin across the whole genome for graphical visualization.

Genome-wide Association Study (GWAS)

Genotypic data for each sub-population were prepared from 404K core SNPs using different marker filtering criteria: for sub-populations contained fewer than 300 accessions, i.e., *aus* and *temperate japonica*, SNPs were excluded from GWAS if call rate < 0.9 and MAF < 0.1; for sub-population contained more than 300 accessions, i.e., *indica* and *tropical japonica*, SNPs were excluded from GWAS if call rate < 0.9 and MAF < 0.05. This gave a total number of 94,005 SNPs for *aus*, 142,836 SNPs for *indica*, 115,495 SNPs for *tropical japonica*, and 56,689 SNPs for *temperate japonica* (Fig. 1).

We used the Fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al. 2016) for GWAS in our study. This method divides the multiple loci linear mixed model into two parts: a fixed effect model (FEM) and a random effect model (REM) which were used iteratively. FEM includes one-by-one marker testing; multiple associated markers, or pseudo quantitative trait nucleotide (pseudo QTN), were further



treated as covariates to control false positives. To avoid the model over-fitting problem in FEM, the pseudo QTN were used to define the kinship in order to model individuals' total genetic effect in REM. Through this two-stage iterative process, FarmCPU has not only increased statistical power but also reduced computation time while controlling both false positives and false negatives. The inclusion of covariate to control for population structure can increase the detection power and eliminate false positives than without the inclusion of covariates (Liu et al. 2016). Therefore, principle components (PC) generated from PCA were used as covariates. For each given trait within a sub-population, GWAS models including zero to five PC were compared with each other based on quantile–quantile (Q–Q) plots (expected p-values vs. observed p-values) to identify the best-fit model. Significant marker-trait association were identified using Bonferroni threshold at an error rate of 0.05.

Cross-validation of associated SNPs

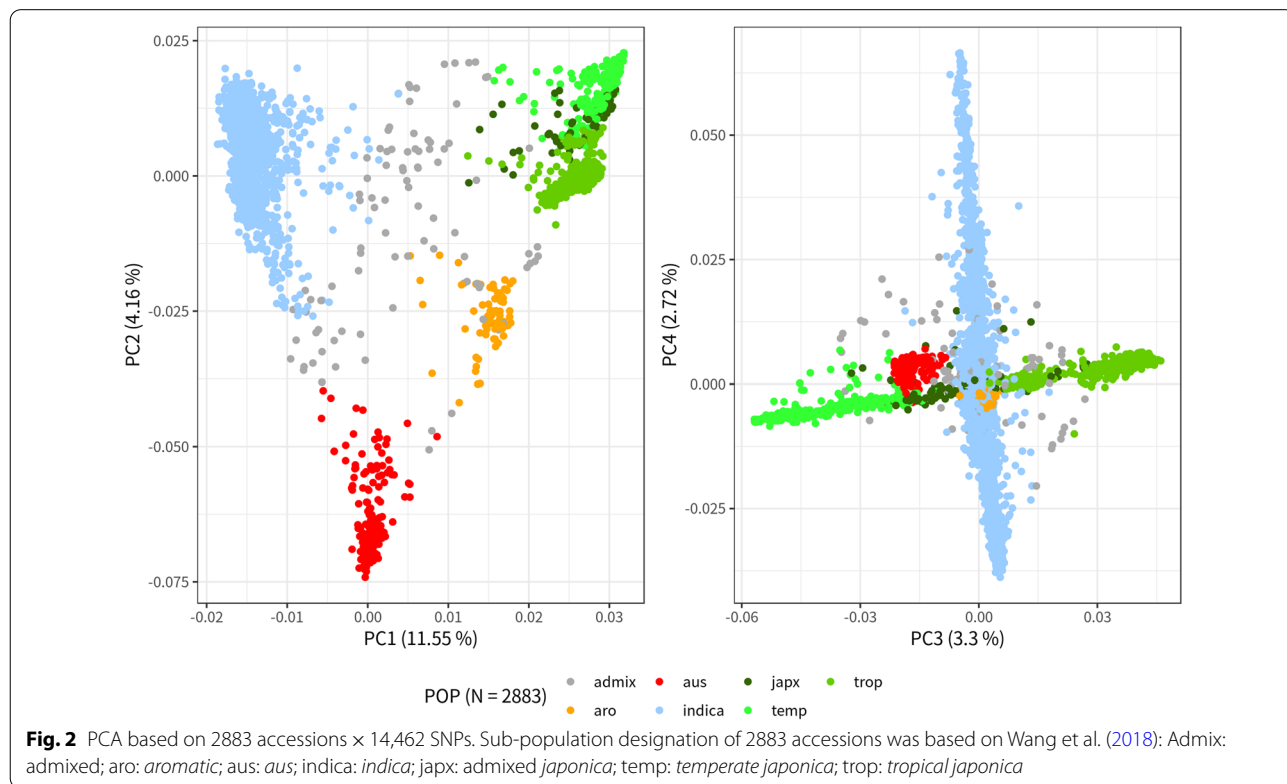
For each significant SNP, we retrieved all genes located within ±100-kb flanking region and defined candidate genes using the gene annotation from the Rice Annotation Project Database (RAP-DB, <http://rapdb.dna.affrc.go.jp/>) (Sakai et al. 2013). For trait-associated SNPs, contingency tables between SNP alleles and phenotype were made and visually inspected to examine the association

between genotype and phenotype. Based on the GWAS results of 3K RGP, we verified the associated variants using WR data through the inspection of genotype–phenotype distribution.

Results

Population structure and LD analysis

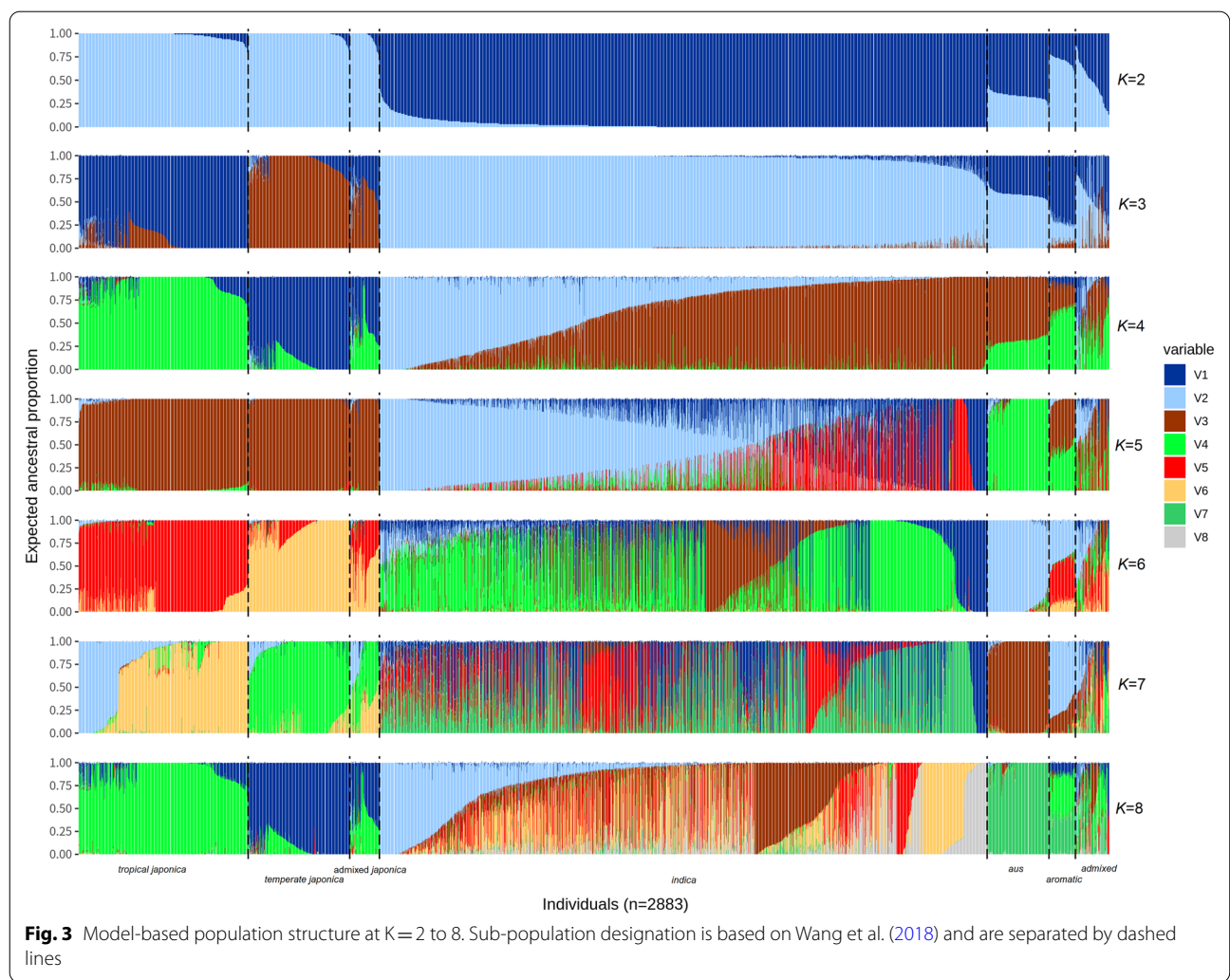
PCA was first conducted to assess population structure of the 2883 accessions after removing some individuals (cf. Methods/Population structure and linkage disequilibrium). The amount of genetic variation explained by PC1 to PC4 was 11.5%, 4.76%, 3.3%, and 2.72%, respectively (Fig. 2). Despite the fact that the first PC explained relatively less variance compared to previous studies (Wang et al. 2014; Zhao et al. 2011), the five main sub-populations, *aromatic*, *aus*, *indica*, *temperate japonica* and *tropical japonica* of Asian rice were clearly distinguished within the sample (Fig. 2). Population structure inferred through the model-based approach was tested for number of sub-populations (K) ranging from 1 to 15. The decision of an appropriate K within a dataset usually relies on a sudden drop in the cross-validation error between K values. Meanwhile, the error estimated in five-fold cross validation decreased smoothly with the increase of K and was minimized at the largest value at K=15 (Additional file 1: Figure S1). At K=2, *indica* and *japonica*

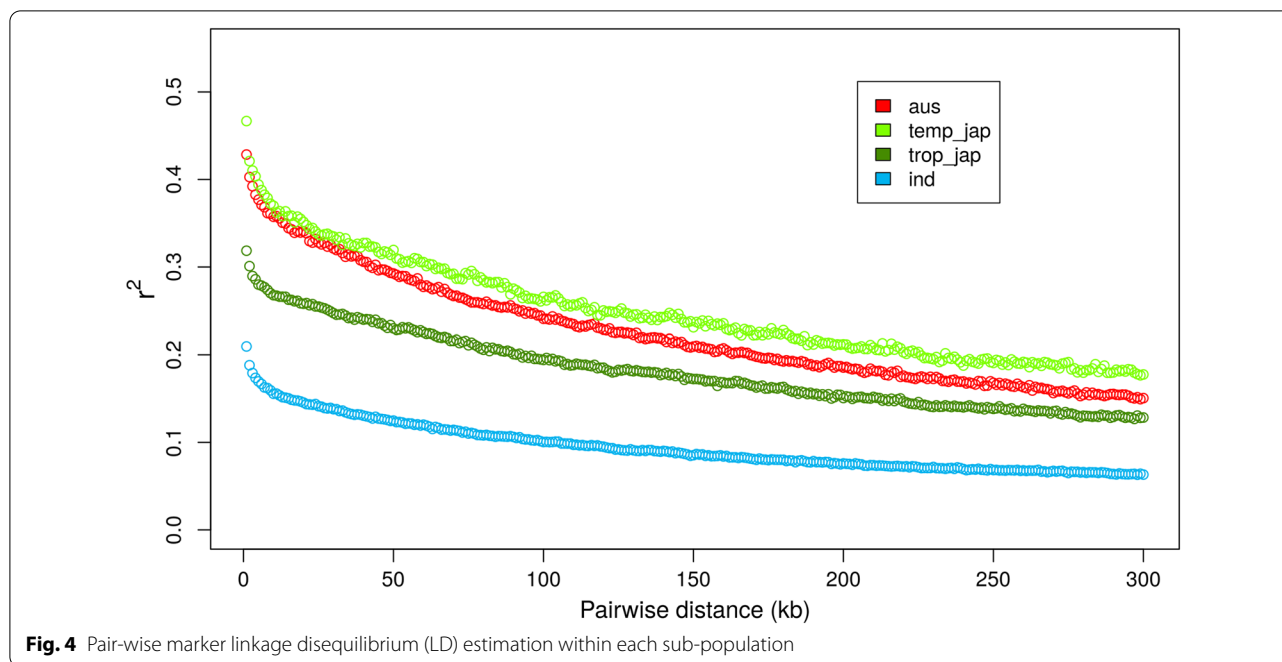


were clearly distinguished, followed by the separation within the *japonica* group at K=3 (Fig. 3). At K=5, *aus* could be distinguished from the other sub-populations, although the separation between *japonica* accessions disappeared, which may be due to the bias of algorithm under the circumstances where strong genetic drift specific to a sub-population was present in the sample, or a sub-population accounted for a large proportion (Marees et al. 2018). *aromatic* could be distinguished from K=7, meanwhile the *temperate japonica* and a set of *indica* would belong to the same sub-population. Considering the population structure revealed by the PCA and the model-based analysis, as well as information from a previous study using the rice 3K panel (Wang et al. 2018), and the need of a certain sample size to achieve reasonable statistical power for GWAS, we have selected K=5 for subsequent analysis. Potential admixed accessions were removed for a better control over false positives resulting from hindered

stratification or genetic heterogeneity (Korte and Farlow 2013; Tian et al. 2008). A final of 1378 accessions were retained for further analyses: 155 accessions were assigned to *aus*, 696 to *indica*, 152 to *temperate japonica*, and 375 to *tropical japonica*. No *aromatic* accessions passed our criteria of selection for non-admixed individuals (Fig. 3) therefore the *aromatic* sub-population was discarded from subsequent analyses.

LD observed in each sub-population was generally lower than that estimated in previous studies (McNally et al. 2009; Xu et al. 2012) (Fig. 4) while the pattern became more congruent with the original study when calculating LD based on another dataset of 4.8 million SNPs that were not LD-pruned (Wang et al. 2018) (Additional file 1: Figure S2). In the core SNP dataset that we used, all sub-populations exhibited an LD decay rate of 100–200 kb at which r^2 dropped to half of its maximum value. The sub-population of *temperate japonica* exhibited the shortest LD decay (91 kb), followed by





aus (146 kb) and *tropical japonica* (157 kb). The mean r^2 value observed in *indica* tended to be much lower than other populations, implying almost no LD between markers within the *indica* sub-population. Nevertheless, the genome-wide marker density with such low LD was already high (between 94,005 SNPs for *aus* and 142,836 SNPs for *indica*; cf. Methods/Genome-wide association study) and markers were well distributed along the genome (Additional file 1: Figure S3).

Trait distribution

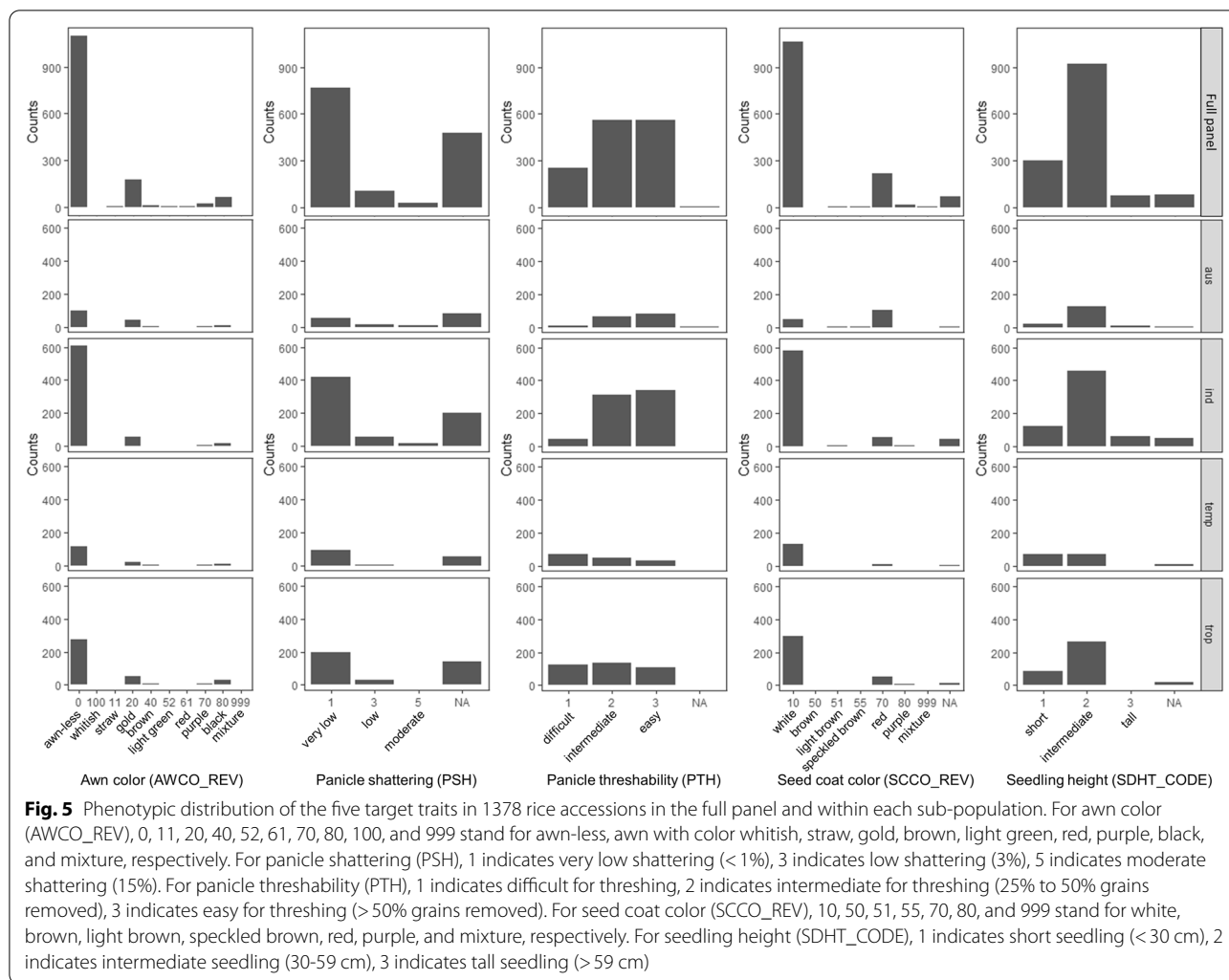
Across the full panel of 1378 accessions, target traits usually had a missing rate smaller than 0.1, except for shattering degree (PSH) which had a missing rate of 0.35 (Fig. 5). Within each sub-population, the distribution of awn color (AWCO_REV), shattering degree (PSH) and seed coat color (SCCO_REV) were highly skewed toward cultivated phenotype since most accessions were improved cultivars, characterized by awn-less panicle, loss of shattering, and white pericarp (Fig. 5). However, there were individuals showing awned panicle and colored pericarp in all sub-populations, especially in *aus* where awn color and seed coat color showed relatively balanced distribution. Panicle threshability (PTH) described here is an agronomic trait distinct from shattering. While shattering is defined as the proportion of grains detached at reproductive stage, threshability is defined as the proportion of grains removed when firmly grasped by hand (IRRI, 2002). Threshability showed relatively balanced distribution within *aus*, *temperate japonica* and *tropical*

japonica, whereas most *indica* accessions showed an easy grain removal during threshing process. Many accessions, particularly *indica*, exhibited a low to intermediate (30–59 cm) seedling height at 5-leaf stage.

GWAS and investigation of trait-associated SNPs

GWAS were conducted separately in each sub-population given that the existence of population structure may result in spurious associations and that sub-population specific variants may be ignored in a combined population (Wang et al. 2018). There is no single model appropriate for a given trait neither for a given population since the genetic architecture underlying a trait varies between populations. Therefore, model selection is needed for each population × trait combination. Different numbers of PC were included as fixed effects in the final GWAS model for each trait within each sub-population (Fig. 6, Additional file 1: Table S1, Figure S4). As we performed GWAS within sub-populations, most sub-population × trait combinations did not require the inclusion of PC (12 out of 20 sub-population × trait combinations included zero PC). The 8 sub-population × trait combinations which required the inclusion of PC in GWAS model were PSH in *aus* (1 PC), AWCO_REV, PTH, and SDHT_CODE in *indica* (1 PC each), AWCO_REV in *temperate japonica* (2 PC) and PSH, PTH, and SDHT_CODE in *tropical japonica* (1, 2, and 2 PC, respectively).

Overall, 66 trait-associated SNPs were identified over five target traits (Table 1): 23 for awn color (AWCO_REV), 8 for panicle shattering degree (PSH), 3 for panicle



threshability (PTH), 19 for seed coat color (SCCO_REV), and 13 for seedling height (SDHT_CODE). All trait-associated SNPs were population-specific. Meanwhile, one region within 26 kb on Chromosome 4 harbored three awn-associated SNPs identified in *aus* and one other awn-associated SNP identified in *temperate japonica*. Three SNPs within 5.3 kb which were close to or inside *Rc* gene on Chromosome 7 were identified from *aus*, *indica*, and *tropical japonica* (Table 1). Twelve out of 66 SNPs, associated with seed coat color, awn color, or panicle shattering degree, fell inside of 11 annotated genes (Table 1). Based on our LD estimation, we further assessed the ± 100 kb region of every trait-associated SNPs in order to identify other potential candidate genes. The number of annotated genes within the ± 100 kb region varied from 8 to 42, most of them were hypothetical protein, of unknown function, or of functions without clear association with the trait of interest. Nevertheless, we have identified four candidate genes close to the

trait-associated SNP, including *Rc* (143 bp downstream of S7_6062746 identified within *tropical japonica*), *RAE2* (23.4 kb upstream of S8_24022229 associated to awn), *WOX10* (2.4 kb upstream of S8_25508381 associated to seedling height), and *GRF8* (7.3 kb downstream of S11_20514671 associate to seedling height).

As our objective was not only to identify trait-associated loci, but also to assess the probability a weedy trait would appear when an individual carries a certain genotype at loci of interests, we inspected the segregation ratio between target traits and associated markers for each sub-population \times trait combination. We have identified nine SNPs whose allele could indicate with at least 60% of accuracy the target phenotype (Table 2). Here, accuracy is defined as the probability that a phenotype occurs at a given allele. For example, among 145 individuals of *aus*, the frequency of genotype AA at S6_7624195 was 0.74. Among all the individuals carrying AA genotype of S6_7624195, 73.8% were awn-less individuals.

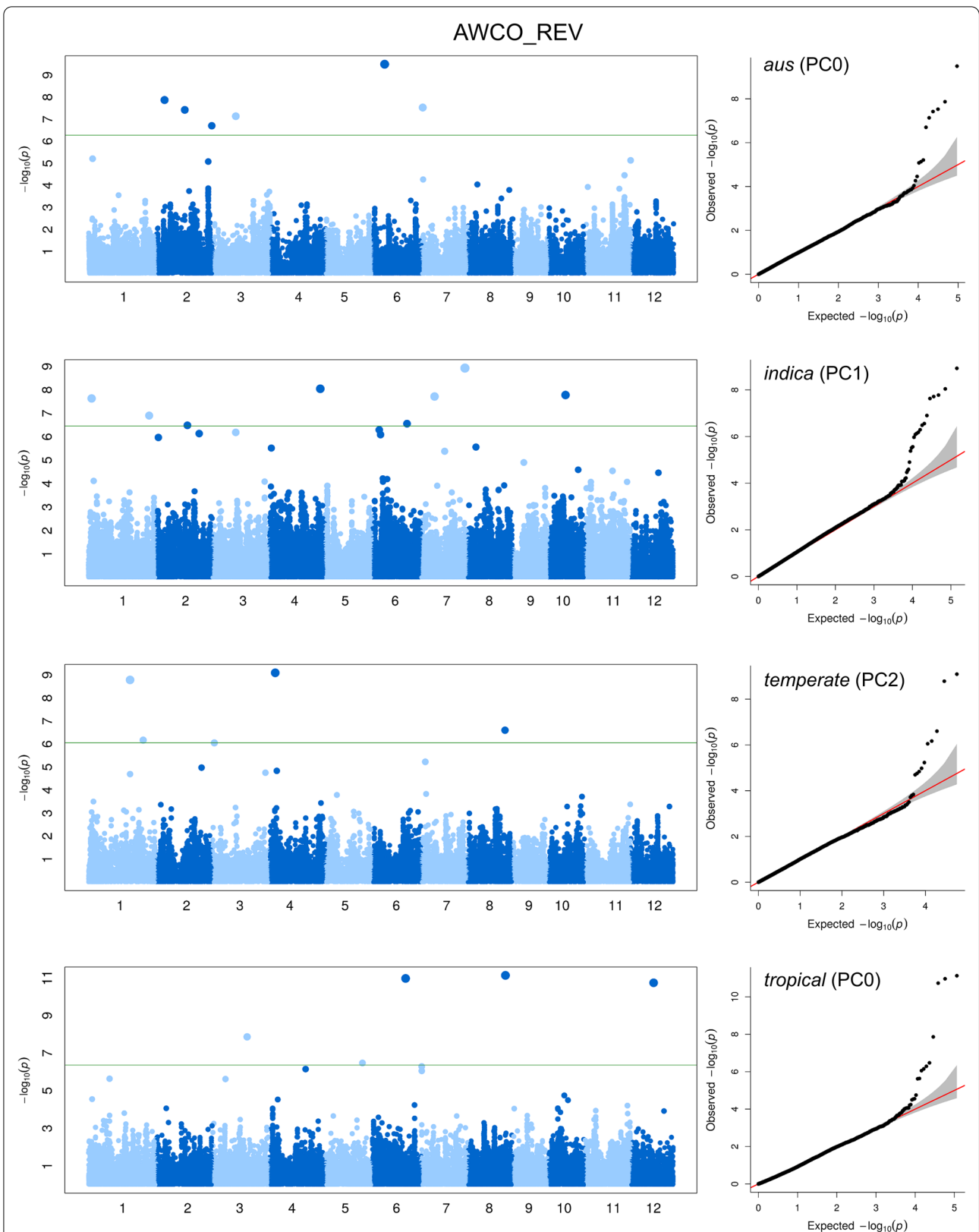


Fig. 6 GWAS results for awn color (AWCO_REV) in each sub-population. Manhattan plot was at the left side, quantile–quantile plot was at the right side. From top to bottom: GWAS results for *aus*, *indica*, *temperate japonica*, and *tropical japonica*, the number of PC included in the final model was indicated between parentheses

Table 1 Trait-associated SNP identified within each sub-population of 3K-RGP

Trait	Pop	SNP	P-value	No. gene \pm 100 kb	Candidate genes	Annotation
AWCO_REV	ind	S1_2222938	2.34E-08	19		
	temp	S1_26642467	1.64E-09	18		
	temp	S1_34999638	6.77E-07	25	<i>Os01g0820700</i>	<i>OsWRKY116</i>
	ind	S1_38893251	1.26E-07	32	<i>Os01g0894500</i>	Sep15/SeIM redox domain containing protein
	aus	S2_5126277	1.35E-08	36		
	aus	S2_18016425	3.79E-08	15		
	ind	S2_20021623	3.27E-07	19		
	aus	S2_35221348	1.99E-07	42		
	aus	S3_14604626	7.34E-08	26		
	trop	S3_22066093	1.37E-08	22		
	temp	S4_3569126	8.05E-10	13		
	ind	S4_32368836	9.04E-09	40	<i>Os04g0636500</i>	BTB domain containing protein
	trop	S5_23650885	3.36E-07	30		
	aus	S6_7624195	3.23E-10	25		
	trop	S6_21261276	1.08E-11	19		
	ind	S6_22232682	2.78E-07	25		
	aus	S7_671056	2.97E-08	27		
	ind	S7_8538306	1.93E-08	8	<i>Os07g0253100</i>	Conserved hypothetical protein
	ind	S7_27892037	1.19E-09	36		
	temp	S8_23747966	2.51E-07	22		
trop	S8_24022229	7.43E-12	37	<i>Os08g0485800</i>	Barwin-related endoglucanase domain containing protein	
				<i>Os08g0485500</i> (-23.4 kb)	<i>RAE2</i> ; Regulation of awn development and elongation	
	ind	S10_10954998	1.67E-08	16		
	trop	S12_14872482	1.84E-11	11		
PSH	temp	S2_18492267	1.99E-07	19		
	temp	S3_3398933	1.35E-08	34		
	aus	S4_4462336	7.23E-08	21		
	aus	S4_4462447	2.33E-07	21		
	temp	S4_4464494	1.28E-07	21		
	aus	S4_4488660	9.51E-08	21	<i>Os04g0166000</i>	Conserved hypothetical protein
	temp	S4_5387818	4.08E-07	10		
	temp	S12_1917244	5.88E-08	35	<i>Os12g0139400</i>	<i>OsRR10</i>
PTH	trop	S2_24505237	1.22E-07	23		
	trop	S7_13096799	5.14E-12	12		
	trop	S8_20373414	1.06E-10	20		
SCCO_REV	ind	S1_8264505	2.21E-07	30		
	ind	S1_14709154	2.38E-14	18		
	ind	S1_19343008	2.26E-07	18		
	trop	S2_20541410	5.66E-08	10		
	ind	S2_35121783	1.35E-07	34	<i>Os02g0818700</i>	Similar to tumor-related protein
	temp	S3_4766628	8.81E-07	36		
	temp	S3_12419022	3.69E-09	32		
	trop	S3_16978020	4.30E-07	18		
	aus	S3_29290186	4.03E-08	20		
	trop	S7_6062746	3.98E-54	15	<i>Os07g0211500</i> (-0.1 kb)	<i>Rc</i>
ind	S7_6067855	1.15E-114	15	<i>Os07g0211500</i>	<i>Rc</i>	
aus	S7_6068017	1.81E-40	15	<i>Os07g0211500</i>	<i>Rc</i>	

Table 1 (continued)

Trait	Pop	SNP	P-value	No. gene \pm 100 kb	Candidate genes	Annotation
	<i>temp</i>	S7_23073822	7.43E-15	29	<i>Os07g0571500</i>	Similar to transmembrane protein 49
	<i>ind</i>	S8_8581885	1.39E-08	15		
	<i>aus</i>	S8_9105408	4.44E-11	22	<i>Os08g0249100</i>	<i>OsRLCK249</i>
	<i>ind</i>	S9_1298020	2.85E-08	14		
	<i>temp</i>	S10_10394027	1.15E-08	13	<i>Os10g0346300</i>	<i>OsLAC17</i>
	<i>trop</i>	S11_17682735	8.45E-09	10		
	<i>trop</i>	S11_19853800	6.99E-08	12		
SDHT_CODE	<i>aus</i>	S1_15921134	1.38E-07	11		
	<i>aus</i>	S1_30508051	1.69E-08	37		
	<i>temp</i>	S1_31283533	4.33E-07	27		
	<i>ind</i>	S2_19154582	8.69E-08	19		
	<i>aus</i>	S2_19678534	3.74E-08	24		
	<i>aus</i>	S4_21716120	1.96E-07	35		
	<i>ind</i>	S8_8643208	4.52E-13	11		
	<i>ind</i>	S8_25508381	2.02E-09	25	<i>Os08g0242400</i> (-2.4 kb)	<i>OsWOX10</i>
	<i>ind</i>	S11_20514671	4.20E-11	20	<i>Os11g0551900</i> (+7.3 kb)	<i>OsGRF8</i>
	<i>ind</i>	S11_24872440	2.62E-11	12		
	<i>aus</i>	S12_154948	4.55E-08	26		
	<i>ind</i>	S12_365399	6.12E-08	34		
	<i>ind</i>	S12_25154542	2.89E-11	26		

The meaning for each trait abbreviation is as follows: AWCO_REV: awn color; PSH: panicle shattering degree; PTH: panicle threshability, SCCO_REV: seed coat color; SDHT_CODE: seedling height. Pop: sub-populations, including *aus*, *indica* (*ind*), *temperate japonica* (*temp*), and *tropical japonica* (*trop*). Markers were named based on their genomic position on the reference sequence (cv Nipponbarre, IRGSP v1) for S[chromosome]_[bp position on the chromosome]. When the trait-associated SNP fell into an annotated gene, the gene name was indicated in the column of "Candidate gene". If a gene, whose function was related to the trait of interest, fell within \pm 100 kb of the trait-associated SNP, the distance between the SNP and the start of the gene is indicated in the parentheses: minus sign means before the SNP, plus sign means after the SNP. Annotation was from RAP-DB. Closely positioned SNPs for which the same genes were identified within the \pm 100 kb region were underlined in bold

The homozygous G at S7_13096799 in *tropical japonica* indicated the tendency of difficult threshing, while the homozygous A, although of lower frequency in the population (0.23), indicated the tendency of easy threshing. Seed coat color could be predicted with an average of 88.7% accuracy by the genotype at target loci: two SNPs identified for *aus*, one SNP for *indica*, and one SNP for *tropical japonica*, located in or in the vicinity of *Rc*, a bHLH protein known to involved in the pigment synthesis at seed coat (Singh et al. 2017; Sweeney et al. 2006) for which we will provide a detail description further. SNPs whose genotypes could predict with a certain accuracy seedling height were identified for *indica* (S11_24872440) and *temperate japonica* (S1_31283533).

The high prediction accuracy of SNP genotypes for seed coat color drove us to investigate in detail these seed coat color-associated SNPs located within or in the vicinity of *Rc* on chromosome 7 (Fig. 7) and we have included the functional 14-bp InDel for a comparison. S7_6062746 was identified in the sub-population of *tropical japonica* and was located in the 5'- untranslated region (5'-UTR). The two alleles T and G of this

SNP had no functional evidence yet. S7_6067855 was identified within *indica* and the two alleles were synonymous mutation (Fig. 7a). S7_6068017 was identified within *aus* whose color-less allele, A, would cause an early stop codon and lead to an incomplete protein product. This allele has been previously assigned as *rc-s* allele (Singh et al. 2017; Sweeney et al. 2006). Further, we verified the frequency of accessions carrying the "colored" allele and the frequency of individuals carrying "colored" allele showing indeed "colored phenotype" (Freq (Col|G_c) in Fig. 7b) in the full panel (1378 accessions) and within each sub-panel (parallel information for "color-less" allele is provided in Table S2). We observed that, although the genotype of the 14-bp InDel generally explained well the occurrence of red seed coat, the "colored" allele of SNPs identified for each sub-population explained even better the occurrence of red seed coat for sub-populations. The most striking case was related to *aus*: the frequency of individuals carrying the 14-bp insertion showing effective red seed coat in *aus* was 0.66, while 96% of individuals carrying the "colored" allele of S7_6068017 possessed

Table 2 Phenotype prediction accuracy based on genotype of selected trait-associated SNPs

Trait	Pop	No. Ind	SNP	Genotype	Freq	Prediction accuracy
AWN	<i>aus</i>	145	S6_7624195	AA	0.74	73.8% awn-less
				GG	0.26	68.4% awned
AWN	<i>trop</i>	375	S8_24022229	GG	0.9	80.1% awn-less
				AA	0.1	89.1% awned
PTH	<i>trop</i>	378	S7_13096799	GG	0.76	73.4% difficult (41.6%) to intermediate (31.8) threshing
				AA	0.23	90.9% easy (38.6%) to intermediate (52.2%) threshing
SCCO	<i>aus</i>	153	S7_6068017	CC	0.68	96.1% colored seed coat
				AA	0.31	93.8% white seed coat
SCCO	<i>aus</i>	153	S8_9105408	GG	0.65	86% colored seed coat
				AA	0.33	66.7% white seed coat
SCCO	<i>ind</i>	645	S7_6067855	GG	0.92	96.1% white seed coat
				AA	0.07	91.5% colored seed coat
SCCO	<i>trop</i>	362	S7_6062746	TT	0.9	90.5% white seed coat
				GG	0.1	88.9% colored seed coat
SDHT	<i>ind</i>	641	S11_24872440	CC	0.85	92.1% short (21%) to intermediate (70%) seedling
				TT	0.13	95.2% intermediate (71%) to tall (23%) seedling
SDHT	<i>temp</i>	137	S1_31283533	AA	0.75	62.1% short seedling
				GG	0.25	94% intermediate seedling

The meaning for each trait abbreviation is as follows: AWN: awn color; PTH: panicle threshability, SCCO: seed coat color; SDHT: seedling height. Sub-population \times trait combinations are shown in the table if the genotype of trait-associated SNP can predict the phenotype with at least 60% of chance to accuracy. Otherwise we do not present the Sub-population \times trait combinations for a clear and concise presentation

colored seed coat in *aus*. No seed coat color-associated SNP was identified for the sub-population of *temperate japonica* which should be a result of too few individuals exhibiting colored seed coat.

Comparison between 3K-RGP and weedy rice

In addition to the analysis using the 3K-RGP data, we attempted to compare our results with publicly available WR data. For this purpose, we have obtained sequence reads of 205 WR accessions from two studies (Li et al. 2017; Qiu et al. 2017) which provided both phenotypic and genotypic data. We identified SNPs of WR panel by aligning WR sequence reads to the reference genome (cv. Nipponbare, IRGSP-1.0). After removing monomorphic and low MAF SNPs ($MAF < 0.05$), we obtained 71,343 polymorphic SNPs for WR panel, among which 41,044 were also found in the 3K-RGP dataset. Detailed observation on the 41,044 common markers revealed different allele profile between the two panels: the minor allele between panels were different, and the allele frequency of the same allele between the two panels was not positively correlated (Pearson's $r = -0.48$, p -value < 0.001 ; Fig. 8a). In terms of phenotypic data, only the presence and absence of awn was available for the WR panel. Therefore, we examined first whether the phenotype–genotype segregation of the two awn-associated explanatory SNPs identified in 3K-RGP (Table 2) could explain the phenotype–genotype relationship in the WR panel.

S8_24022229 was absent from the WR panel, while for S6_7624195, no individual carried the “awn-less” genotype (AA); only three individuals carried allele A in heterozygous state (Fig. 8b). We have also conducted GWAS on WR panel (Fig. 8c, Table 3) whose profile and significant SNPs were different from that of 3K-RGP (Fig. 6, Table 1).

Discussion

The aim of this study was to make use of publicly available data to explore the genetics of “weedy” traits, or, in other words, traits related to domestication. The main dataset that we used were subsets from the 404K core SNPs of the 3K-RGP. We have chosen the 404 core SNP dataset considering both marker density and computational efficiency. Population structure analyses clearly showed the presence of sub-populations within the germplasm in spite of the fact that the model-based approach did not provide a clear-cut K value for the number of sub-populations and prior knowledge was needed for such determination. Knowledge on population structure on a collection of germplasm is particularly important for the subsequent genetic analyses, especially GWAS. It has been demonstrated in plants that population structure could interfere in the identification of trait-associated markers (Alonso-Blanco et al. 2016; Morris et al. 2013; Zhao et al. 2011). Therefore, we have used a relatively stringent criterion

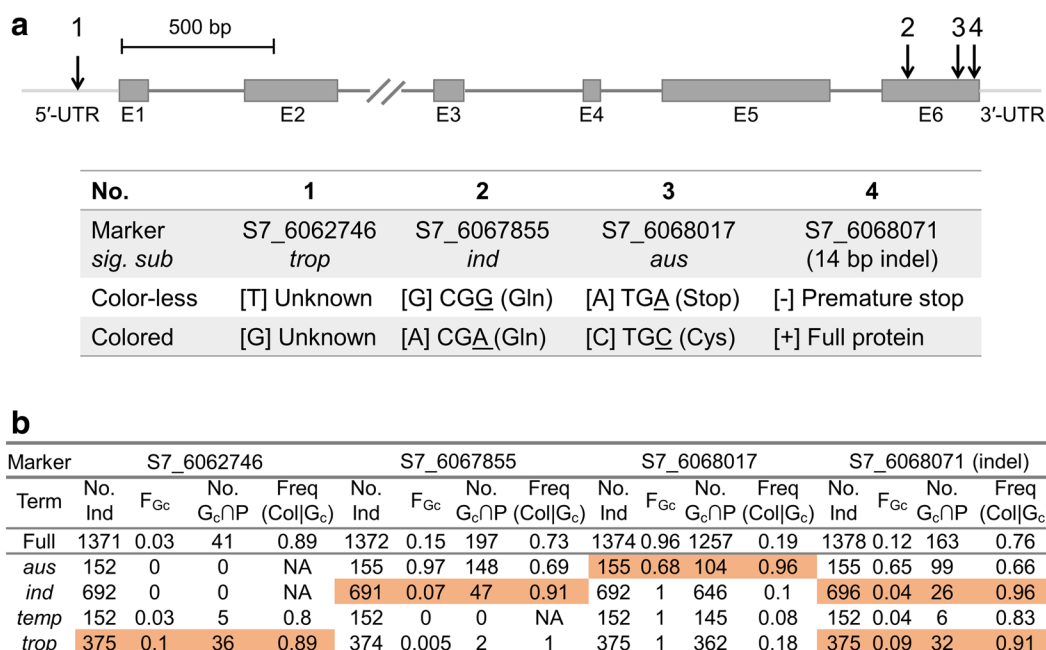


Fig. 7 Relationship between genotype and phenotype for seed coat/pericarp color-associated SNP inside or close to the known candidate gene *Rc*. **a** Position of the SNPs were indicated by arrows on the diagram of the structure of *Rc*. E1 to E6 stand for exon 1 to exon 6. The gene structure of *Rc* is proportional to the actual base-pair length, except for intron 2 which is ca. 2.7 kb. *sig. sub* indicates the sub-population from which the corresponding seed coat color-associated SNP was identified. The color-less or colored seed coat-associated allele, as well as the associated functional alteration were indicated. **b** For each trait-associated marker, its number of individuals used for the calculation of frequency of individuals carrying the “colored” genotype (F_{Gc}), the number of individuals carrying the “colored” genotype which had seed coat color information (No. $G_c \cap P$), and the frequency that the individuals carrying colored genotype had indeed colored seed coat ($Freq(Col | G_c)$). Within a sub-population, SNPs whose alleles correspond well to phenotypes are highlighted in orange. Meanwhile, only SNP \times sub-population combinations with No. $G_c \cap P$ exceeding 20 are highlighted

(sub-population genome belonging ≥ 0.8) to define sub-populations. As such, no individuals from *aromatic* sub-population were retained for our analysis since they appeared as the results of hybridization between *aus* and *japonica* (Civán et al. 2015). Our approach seemed to be effective since the inclusion of PC to correct for population structure in GWAS model was mostly unnecessary (Additional file 1: Table S1) and the inclusion of PC for some sub-population \times trait combinations could be explained by the fact that weak structure did exist within the sub-populations that we defined (Wang et al. 2018).

Self-pollinated feature of cultivated rice often favors the maintenance of long-range LD (McNally et al. 2009; Zhao et al. 2011). Genome-wide LD decay in cultivated rice has been estimated at a 100–300 kb range across different sub-populations (McNally et al. 2009; Zhao et al. 2011). The relatively low LD value observed in our study should be due to the fact that the 404K core SNP was previously LD-pruned, yet the trend of the LD decay remained similar when we estimated the LD using the 4.8 million un-pruned dataset (Additional file 1: Figure S2).

We used GWAS to detect trait-associated SNPs within each sub-population. Relatively few trait-associated SNPs were identified compared to the number of SNPs used for GWAS. One possible reason was that the available phenotypic data were categorical data which reduced the variation and information of quantitative traits. The other possible reason was that the phenotypic distributions were skewed towards cultivated types, such as absence of awn, low degree of shattering, color-less seed coat, and short stature (Fig. 5). A more balanced phenotypic distribution would help in the estimation of allelic contrast within a marker. Eleven trait-associated SNPs fell into an annotated gene, while four other putative candidate genes were identified using an average LD decay length (± 100 kb). They were actually closer to the putative candidate (within ± 25 kb) compared to the average LD decay. We have observed the same phenomenon in another study where most promising candidate genes were identified within ± 10 kb of the significant SNP (Ma et al. unpublished data). Among the candidate genes identified in this study, the most promising one is *Rc*, a gene encoding a bHLH protein controlling

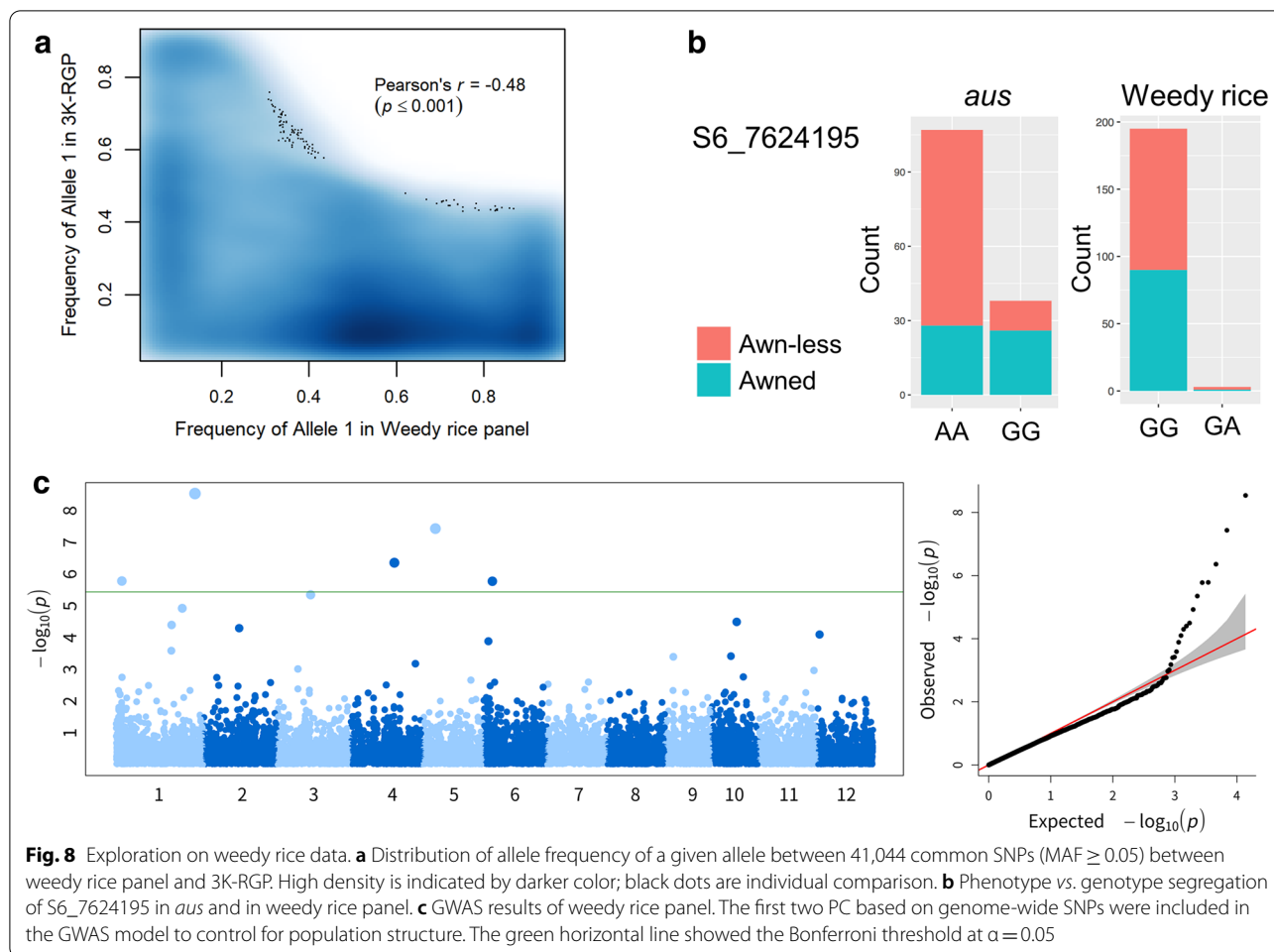


Fig. 8 Exploration on weedy rice data. **a** Distribution of allele frequency of a given allele between 41,044 common SNPs (MAF ≥ 0.05) between weedy rice panel and 3K-RGP. High density is indicated by darker color; black dots are individual comparison. **b** Phenotype vs. genotype segregation of S6_7624195 in *aus* and in weedy rice panel. **c** GWAS results of weedy rice panel. The first two PC based on genome-wide SNPs were included in the GWAS model to control for population structure. The green horizontal line showed the Bonferroni threshold at $\alpha = 0.05$

Table 3 Awn-associated SNPs identified within WR panel

SNP	P.value	No. Gene ± 100 kb	Candidate genes	Annotation
S1_2958149	1.65E-06	42	<i>Os01g0155000</i>	Alpha/beta hydrolase fold-3 domain containing protein
S1_38916854	2.89E-09	34	<i>Os01g0895200</i>	DOMON related domain containing protein
S4_21610862	4.36E-07	27	<i>Os04g0434600</i>	Similar to OSIGBa0102D10.3 protein
S5_6339023	3.68E-08	20		
S6_4425573	1.67E-06	31	<i>Os06g0187700</i>	Conserved hypothetical protein

Markers were named based on their genomic position on the reference sequence (cv Nipponbarre, IRGSP v1) for S[chromosome]_[bp position on the chromosome]. When the trait-associated SNP fell into an annotated gene, the gene name was indicated in the column of "Candidate gene". Annotation was from RAP-DB

the proanthocyanidin synthesis in rice, a major player in seed coat color formation (Furukawa et al. 2007; Sweeney et al. 2006). A 14-bp deletion within exon 6 of *Rc*, and a shorter *Rc* transcript was found to correlate to the color-less seed coat (Sweeney et al. 2006), which was further confirmed through functional validation that *Rc* did code for a bHLH protein and its 14-bp InDel contribute to the color-less seed coat phenotype (Furukawa et al. 2007). Among the 1378 3K-RGP individuals

selected for our study, 163 carried the 14-bp "red" allele of which 76% of individual showed effective red seed coat (Fig. 7). Our detailed analysis showed that the 14-bp "red" allele correlated relatively well to the occurrence of red seed coat within *indica*, *temperate japonica*, and *tropical japonica*, although the number of individuals carrying the red allele was small for the three populations, therefore less representative to a larger scale (Fig. 7). There are other genes known to be involved in

seed coat color formation in rice, such as *Rd*, coding for an enzyme of the proanthocyanidin biosynthetic pathway (Furukawa et al. 2007). This could explain that although polymorphisms in *Rc* correlate well with seed coat color but cannot predict 100% the phenotype. Meanwhile, for each sub-population, the trait-associated SNPs identified in our study were more representative to explain the occurrence of red seed coat both in terms of number of individuals for the analysis (No. $G_c \cap P$ in Fig. 7) and the frequency of “red” allele carrying individuals with effective red phenotype ($\text{Freq}(\text{Col} | G_c)$ in Fig. 7). The most striking case was S7_6068017 identified in *aus* whose genotype could explained close to the total occurrence of red seed coat. The A allele of S7_6068017 was previous identified as *rc-s* allele (Sweeney et al. 2006) and has been noticed to be present at moderate frequency in *aus* (five out of nine white seed coat individuals) and in *aromatic* (two out of 17 white seed coat individuals) but not at all in *indica*, *temperate* or *tropical japonica* (Sweeney et al. 2007). Using a larger sample, we found that both alleles of S7_6068017 were presented in a relative balanced frequency in *aus*. As red seed coat is widely observed for WR (Huang et al. 2018; Li et al. 2017; Qiu et al. 2017), our results on *Rc* could be used to diagnose in part the origin of WR occurred in the field. Some significant trait-associated SNPs were in the vicinity of other candidate genes: the most significant awn color-associated SNP identified in *tropical japonica*, S8_24022229 (Table 1), was 23.4 kb downstream *RAE2*, which was involved in awn development and grain size in Asian rice (Bessho-Uehara et al. 2016). For seedling height, two associated SNPs were found to be close to two candidate genes: S8_25508381 was 2.4 kb downstream of *WOX10*, while S11_20514671 was upstream of *GRF8*. *WOX10* belongs to *WUSCHEL* (*WUS*)-related Homeobox (*WOX*) gene family which have been shown to coordinate gene expression both in shoot and root meristem in Arabidopsis (Haecker et al. 2004). *OsWOX11* was shown to involve in the activation of rice crown root emergence and growth, related to seedling growth. On the other hand, *GRF8* belongs to growth-regulating factor (GRF) family, a plant specific transcription factor. GRF are involved in various plant developmental processes and take part in the coordination of growth under adverse environmental conditions (Omidbakhshfard et al. 2015). Therefore, it can be a good candidate of seedling vigor, an important characteristic for weeds.

Since our work using the data from 3K-RGP was to pave the way for our understanding of the WR occurrence, we have also obtained WR data from public resources to get a first glance of the usefulness of our results from cultivated pool. Although our most promising diagnostic markers were for seed coat color, the common trait

from WR panel was the presence/absence of awn, which restricted our analysis. A substantial number of markers were in common between the 3K-RGP panel and the WR panel, but when we investigated quality-filtered markers, allele frequency behaved differently between the two panels, as well as the genotype vs. phenotype segregation and, the genomic position of awn-associated markers (Fig. 8). Indeed, the genotypic data showed that the WR accessions did not grouped with the cultivated rice from the 3K-RGP (data not shown) which implied a different evolution history of the two panels. Such differences did not discredit the result from each panel. Instead, they remind us that although publicly available data is a great tool to explore plant genetics and genomics, certain limits remain and data from complementary laboratory experiment will always be necessary to validate the dry-lab results.

Conclusion

We have analyzed publicly available data to explore the genetic architecture of weediness-related traits on cultivated pool from 3K-RGP. After assessing population structure and LD, we have performed GWAS within each sub-population and have identified trait-associated markers. Candidate genes were identified for weediness-related traits and the most significant were SNPs from *Rc* on chromosome 7. We have shown that *rc-s* was particularly abundant in *aus* and that *Rc* alleles could therefore serve as diagnostic markers to assess the origin of weedy rice in the field, especially to understand the occurrence of colored seed coat. Further, we have compared the data from 3K-RGP to another publicly available data of WR. The constitution of the two panels was so different that their results did not converge toward the same direction. This work showed the potential of publicly available data but also remind the indispensable validation by lab experiments.

Supplementary information

The online version contains supplementary material available at <https://doi.org/10.1186/s40529-020-00309-y>.

Additional file 1: Table S1. Number of PC included in the final GWAS model. **Table S2.** Relationship between genotype and phenotype for seed coat/pericarp “color-less” allele of SNPs inside or close to the known candidate gene *Rc*. **Figure S1.** cross-validation error for K = 1 to 15. **Figure S2.** LD decay based on 4.8 M SNPs. **Figure S3.** Genome-wide marker distribution for each sup-population. **Figure S4.** Manhattan plots and Q-Q plots for final GWAS model for different sub-population × trait combination.

Abbreviations

3K-RGP: 3000 Rice Genome Project; BHA: Black-hull awned; GWAS: Genome-wide association analysis; LD: Linkage disequilibrium; PCA: Principle

component analysis; SH: Straw awn-less; SNP: Single nucleotide polymorphism; WR: Weedy rice.

Acknowledgements

We thank Dr. Yue-le C. Hsing for her comments and for providing computing power for sequence alignment. We thank Dr. Chih-Wei Tung for her comments on this work. We thank Dr. Fu-Jin Wei for his assistance in server maintenance and software uses. We thank the editor and the two anonymous reviewers for their critical reading and helpful suggestions for manuscript improvement.

Authors' contributions

YFH conceived and supervised the project. YLL and YFH analyzed the data. CCW performed sequence alignment and variant calling. DHW provided the functional marker genotypic data. YLL, YFH and DHW interpreted the results. YLL and YFH drafted the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by National Taiwan University and the Council of Agriculture, Executive Yuan (Grant No. 109AS-1.1.5-ST-a1).

Availability of data and materials

The data used and analyzed for the current study can be obtained from the corresponding author.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Department of Agronomy, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd, Da'an Dist., Taipei 10617, Taiwan. ² Taiwan Agricultural Research Institute, Council of Agriculture, Executive Yuan, No. 189, Zhongzheng Rd, Wufeng Dist, Taichung City 41362, Taiwan. ³ Institute of Plant and Microbial Biology, Academia Sinica, Taipei 11529, Taiwan. ⁴ Institute of Plant Science, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd, Da'an Dist., Taipei 10617, Taiwan.

Received: 12 August 2020 Accepted: 29 December 2020

Published online: 12 January 2021

References

- Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–1664
- Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezaan TM, Ding W, Ecker JR, Exposito-Alonso M, Farlow A, Fitz J, Gan X, Grimm DG, Hancock AM, Henz SR, Holm S, Horton M, Jarsulic M, Kerstetter RA, Korte A, Korte P, Lanz C, Lee C-R, Meng D, Michael TP, Mott R, Mulyati NW, Nägele T, Nagler M, Nizhynska V, Nordborg M, Novikova PY, Picó FX, Platzer A, Rabanal FA, Rodríguez A, Rowan BA, Salomé PA, Schmid KJ, Schmitz RJ, Seren Ü, Sperone FG, Sudkamp M, Svardal H, Tanzer MM, Todd D, Volchenboum SL, Wang C, Wang G, Wang X, Weckwerth W, Weigel D, Zhou X (2016) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166:481–491
- Angira B, Addison CK, Cerioli T, Rebong DB, Wang DR, Pumphlin N, Ham JH, Oard JH, Linscombe SD, Famoso AN (2019) Haplotype characterization of the sd1 Semidwarf gene in United States Rice. *Plant Genome* 12:190010
- Bessho-Uehara K, Wang DR, Furuta T, Minami A, Nagai K, Gamuyao R, Asano K, Angeles-Shim RB, Shimizu Y, Ayano M, Komeda N, Doi K, Miura K, Toda Y, Kinoshita T, Okuda S, Higashiyama T, Nomoto M, Tada Y, Shinohara H, Matsubayashi Y, Greenberg A, Wu J, Yasui H, Yoshimura A, Mori H, McCouch SR, Ashikari M (2016) Loss of function at *RAE2*, a previously unidentified EPFL, is required for awnlessness in cultivated Asian rice. *Proc Natl Acad Sci* 113:8969–8974
- Bräutigam A, Gowik U (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biol* 12:831–841
- Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from Next-Generation reference panels. *Am J Hum Genet* 103:338–348
- Civáň P, Craig H, Cox CJ, Brown TA (2015) Three geographically separate domestications of Asian rice. *Nat Plants* 1:15164
- Delouche JC, Burgos NR, Gealy DR, de San Martín GZ, Labrada R (2007b) Diversity of weedy rice populations Weedy rices-origin, biology, ecology and control, p17-44. Food and Agriculture Organization of the United Nations, Rome. pp. 144
- Delouche JC, Burgos NR, Gealy DR, de San Martín GZ, Labrada R (2007a) Seed shattering and dormancy in weedy rices Weedy rices-origin, biology, ecology and control, p45-6. Food and Agriculture Organization of the United Nations, Rome. pp. 144
- Furukawa T, Maekawa M, Oki T, Suda I, Iida S, Shimada H, Takamura I, Kadowaki K-I (2007) The *Rc* and *Rd* genes are involved in proanthocyanidin synthesis in rice pericarp. *Plant J* 49:91–102
- Haecker A, Groß-Hardt R, Geiges B, Sarkar A, Breuninger H, Herrmann M, Laux T (2004) Expression dynamics of *WOX* genes mark cell fate decisions during early embryonic patterning in *Arabidopsis thaliana*. *Development* 131:657–668
- Huang Z, Kelly S, Matsuo R, Li L-F, Li Y, Olsen KM, Jia Y, Caicedo AL (2018) The role of standing variation in the evolution of weedy traits in South Asian weedy rice (*Oryza* spp.). *Genes Genomes Genet* 8:3679–3690
- IRRI. 2002. Standard evaluation systems for rice (SES), 56. IRRI
- Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29
- Kumar A, Daware A, Kumar A, Kumar V, Gopala Krishnan S, Mondal S, Patra BC, Singh AK, Tyagi AK, Parida SK, Thakur JK (2020) Genome-wide analysis of polymorphisms identified domestication-associated long low-diversity region carrying important rice grain size/weight quantitative trait loci. *Plant J*. <https://doi.org/10.1111/tpj.14845>
- Leung H, Raghavan C, Zhou B, Oliva R, Choi IR, Lacorte V, Jubay ML, Cruz CV, Gregorio G, Singh RK, Ulat VJ, Borja FN, Mauleon R, Alexandrov NN, McNally KL, Sackville Hamilton R (2015) Allele mining and enhanced genetic recombination for rice breeding. *Rice* 8:34
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760
- Li L-F, Li Y-L, Jia Y, Caicedo AL, Olsen KM (2017) Signatures of adaptation in the weedy rice genome. *Nat Genet* 49:811–814
- Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet* 12:e1005767
- Londo JP, Schaal BA (2007) Origins and population genetics of weedy red rice in the USA. *Mol Ecol* 16:4523–4535
- Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, Sanciangco M, Palis K, Copetti D, Poliakov A, Dubchak I, Solovoyev V, Wing RA, Hamilton RS, Mauleon R, McNally KL, Alexandrov N (2016a) Rice SNP-seek database update: new SNPs, indels, and queries. *Nucleic Acids Res* 45:D1075–D1081
- Mansueto L, Fuentes RR, Chebotarov D, Borja FN, Detras J, Abriol-Santos JM, Palis K, Poliakov A, Dubchak I, Solovoyev V, Hamilton RS, McNally KL, Alexandrov N, Mauleon R (2016b) SNP-Seek II: a resource for allele mining and analysis of big genomic data in *Oryza sativa*. *Curr Plant Biol* 7–8:16–25
- Marees AT, de Kluiver H, Stringer S, Vorspan J, Curis E, Marie-Claire C, Derks EM (2018) A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatric Res* 27:e1608
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303
- McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Rötter G, Buell CR, Leung H, Leach JE (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Genetics* 182:12273–12278
- Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet* 14:840–852

- Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci* 110:453–458
- Nguyen T, Zhou C, Zhang T, Yu J, Miao R, Huang Y, Zhu X, Song W, Liu X, Mou C, Lan J, Liu S, Tian Y, Zhao Z, Jiang L, Wan J (2019) Identification of QTL for seed dormancy from weedy rice and its application to elite rice cultivar ‘Ninggeng 4’. *Mol Breeding* 39:123
- Olofsson M, Valverde BE, Madsen KH (2000) Herbicide resistant rice (*Oryza sativa* L.): global implications for weedy rice and weed management. *Ann Appl Biol* 137:279–295
- Olsen KM, Caicedo AL, Jia Y (2007) Evolutionary genomics of weedy rice in the USA. *J Integr Plant Biol* 49:811–816
- Omidbakhshfar Mohammad A, Proost S, Fujikura U, Mueller-Roeber B (2015) Growth-regulating factors (GRFs): a small transcription factor family with important functions in plant biology. *Mol Plant* 8:998–1010
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
- Qiu J, Zhou Y, Mao L, Ye C, Wang W, Zhang J, Yu Y, Fu F, Wang Y, Qian F, Qi T, Wu S, Sultana MH, Cao Y-N, Wang Y, Timko MP, Ge S, Fan L, Lu Y (2017) Genomic variation associated with local adaptation of weedy rice during de-domestication. *Nat Commun* 8:15323
- Reagon M, Thurber CS, Gross BL, Olsen KM, Jia Y, Caicedo AL (2010) Genomic patterns of nucleotide diversity in divergent populations of U.S. weedy rice. *BMC Evol Biol* 10:180
- Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, Wakimoto H, Yang C-C, Iwamoto M, Abe T, Yamada Y, Muto A, Inokuchi H, Ikemura T, Matsumoto T, Sasaki T, Itoh T (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol* 54:e6–e6
- Singh N, Singh B, Rai V, Sidhu S, Singh AK, Singh NK (2017) Evolutionary insights based on SNP haplotypes of red pericarp, grain size and starch synthase genes in wild and cultivated rice. *Front Plant Sci* 8:972
- Song B-K, Chuah T-S, Tam SM, Olsen KM (2014) Malaysian weedy rice shows its true stripes: wild *Oryza* and elite rice cultivars shape agricultural weed evolution in Southeast Asia. *Mol Ecol* 23:5003–5017
- Sun J, Ma D, Tang L, Zhao M, Zhang G, Wang W, Song J, Li X, Liu Z, Zhang W, Xu Q, Zhou Y, Wu J, Yamamoto T, Dai F, Lei Y, Li S, Zhou G, Zheng H, Xu Z, Chen W (2019) Population genomic analysis and *de novo* assembly reveal the origin of weedy rice as an evolutionary game. *Mol Plant* 12:632–647
- Sweeney MT, Thomson MJ, Pfeil BR, McCouch S (2006) Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* 18:283–294
- Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, Bustamante CD, McCouch SR (2007) Global dissemination of a single mutation conferring white pericarp in rice. *PLoS Genet* 3:1418–1424
- Tang W, Ye J, Yao X, Zhao P, Xuan W, Tian Y, Zhang Y, Xu S, An H, Chen G, Yu J, Wu W, Ge Y, Liu X, Li J, Zhang H, Zhao Y, Yang B, Jiang X, Peng C, Zhou C, Terzaghi W, Wang C, Wan J (2019) Genome-wide associated study identifies NAC42-activated nitrate transporter conferring high nitrogen use efficiency in rice. *Nat Commun* 10:5279
- Tatarinova TV, Chekalin E, Nikolsky Y, Bruskin S, Chebotarov D, McNally KL, Alexandrov N (2016) Nucleotide diversity analysis highlights functionally important genomic regions. *Sci Rep* 6:35730
- Tian C, Gregersen PK, Seldin MF (2008) Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet* 17:R143–R150
- Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLoS Biol* 12:e1001883
- Vigueira CC, Qi X, Song B-K, Li L-F, Caicedo AL, Jia Y, Olsen KM (2019) Call of the wild rice: *Oryza rufipogon* shapes weedy rice evolution in Southeast Asia. *Evol Appl* 12:93–104
- Wang CH, Zheng XM, Xu Q, Yuan XP, Huang L, Zhou HF, Wei XH, Ge S (2014) Genetic diversity and classification of *Oryza sativa* with emphasis on Chinese rice germplasm. *Heredity* 112:489–496
- 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 J-C, 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 3010 diverse accessions of Asian cultivated rice. *Nature* 557:43–49
- Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 30:105–111
- Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82
- Ye H, Beighley DH, Feng J, Gu X-Y (2013) Genetic and physiological characterization of two clusters of quantitative trait loci associated with seed dormancy and plant height in rice. *Genes Genomes Genetics* 3:323–331
- Zhao K, Tung C-W, Eizenga G, Wright M, Ali M, Price A, Norton G, Islam M, Reynolds A, Mezey J, McClung A, Bustamante C, McCouch S (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* 2:467
- Ziska LH, Gealy DR, Burgos N, Caicedo AL, Gressel J, Lawton-Rauh AL, Avila LA, Theisen G, Norsworthy J, Ferrero A, Vidotto F, Johnson DE, Ferreira FG, Marchesan E, Menezes V, Cohn MA, Linscombe S, Carmona L, Tang R, Merotto A (2015) Chapter Three-Weedy (Red) Rice: an emerging constraint to global rice production. In: Sparks DL (ed) *Advances in Agronomy*, vol 129. Academic Press, Cambridge, pp 181–228

Publisher's Note

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

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)