Mol Breeding (2023) 43:21
https://doi.org/10.1007/s11032-023-01366-4

CandiHap: a haplotype analysis toolkit for natural variation study

Xukai Li · Zhiyong Shi · Jianhua Gao · Xingchun Wang · Kai Guo

Received: 9 June 2022 / Accepted: 22 February 2023 / Published online: 15 March 2023

Abstract Haplotype blocks greatly assist association-based mapping of casual candidate genes by significantly reducing genotyping effort. The gene haplotype could be used to evaluate variants of affected traits captured from the gene region. While there is a rising interest in gene haplotypes, much of the corresponding analysis was carried out manually. CandiHap allows rapid and robust haplotype analysis and candidate identification preselection of candidate causal single-nucleotide polymorphisms and InDels from Sanger or next-generation sequencing data. Investigators can use CandiHap to specify a gene or linkage sites based on genome-wide association studies and explore favorable haplotypes of candidate genes for target traits. CandiHap can be run on computers with Windows, Mac, or UNIX platforms in a graphical user interface or command line, and applied to any species, such as plant, animal, and microbial. The CandiHap software, user manual, and example datasets are freely available at BioCode (https://ngdc.cncb.ac.cn/biocode/tools/BT007080) or GitHub (https://github.com/xukaili/CandiHap).

Keywords CandiHap · Haplotype · GWAS · SNPs · InDels

Introduction

With the rapid development of next-generation sequencing (NGS) technologies, genome sequencing is becoming inexpensive, routine, and convenient to obtain large numbers of single-nucleotide polymorphisms (SNPs) (Goodwin et al. 2016). Whole-genome re-sequencing (WGRS), genotyping-by-sequencing (GBS), and restriction site-associated DNA (RAD) are essential strategies in medical, biological, and agricultural research to elucidate the genetic basis of phenotypic traits, such as disease or economically important features (Visscher et al. 2017, 2012; Patil et al. 2019; Thudi et al. 2016; Tinker et al. 2016; Miller et al. 2007). These strategies are based on sequencing

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s11032-023-01366-4.

Xukai Li and Zhiyong Shi contributed equally to this work.

X. Li (✉) · J. Gao · X. Wang
Hou Ji Laboratory in Shanxi Province, Shanxi Agricultural University, Taigu 030031, China
e-mail: xukai_li@sxau.edu.cn

X. Li · Z. Shi · J. Gao · X. Wang
College of Life Sciences, Shanxi Agricultural University, Taigu 030801, China

K. Guo (✉)
Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA
e-mail: kaiguo@umich.edu

of whole genomes, or representative genome fractions, across many individuals to determine loci with sequence variations. SNPs can alter the amino acid sequence of a protein directly (e.g., via non-synonymous SNPs, alterations of stop-codons, frameshift SNPs, or SNPs in splice sites) or can change gene expression patterns by affecting gene regulatory regions. Using many genome-wide variants, genome-wide association studies (GWAS) generally identify SNPs that are statistically associated with certain traits. Such SNPs provide the basis to understand mechanisms that drive a trait; however, a key challenge is to rapidly and robustly identify causal SNPs (McCarthy and Hirschhorn 2008).

In GWAS, millions of genetic variations are tested across numerous genomes to identify those statistically associated with a specific trait or disease. Interpreting these associations in a biological and genomic context is very difficult (Uffelmann et al. 2021). Due to linkage disequilibrium, previous GWAS have demonstrated that most traits are significantly correlated with both causal and non-causal variants that are physically close (Slatkin 2008). This limits the identification of causative variants without additional research (Uffelmann and Posthuma 2021). Based on observed patterns of linkage disequilibrium and association statistics, fine-mapping is an in silico procedure created to prioritize the set of variants within each genetic locus found by GWAS that are most likely to be causal to the target phenotype (Raychaudhuri 2011; Schaid et al. 2018). The fine-mapping approaches are based on stepwise conditional, such as GCTA-COJO (Yang et al. 2012) and Bayesian models, including CAVIAR (Hormozdiari et al. 2014), FINEMAP (Benner et al. 2016), PAINTOR (Kichaev et al. 2014), and SuSIE (Wang et al. 2018a). Prioritizing the likely affected gene is perhaps the most crucial part of the functional interpretation of GWAS. GWAS fine-mapped to coding variants, tools such as ANNOVAR (Wang et al. 2010) or SnpEff (Cingolani et al. 2012) can be used to infer their potential effect on genes and to determine the affected gene (Visscher et al. 2017). However, the vast majority of the tools used for these analyses are web-based or command lines implemented and mainly focused on human and rice traits, which severely limit wider applications. In fact, researchers would benefit from identifying candidate causal variants of the most significant SNPs from the species on which GWAS was conducted. The corresponding manual tasks are laborious, time consuming, and prone to errors and omission. To resolve these problems, we aimed to develop a software for fast identification of candidate causal variants or gene(s) from GWAS data.

Haplotype blocks greatly assist association-based mapping of casual candidate genes by significantly reducing the genotyping effort (Zhang et al. 2002a). Different definitions are used to define the haplotype block structure (Patil et al. 2001; Gabriel et al. 2002; Wang et al. 2002; Zhang et al. 2002b). Here, we refer to the haplotype not as the strong inter-marker LD, but rather the SNPs and indels within a gene region, including upstream, downstream, exonic, and intronic regions. We term this as the gene haplotype, which could be adopted to evaluate the variants of the affected traits captured from the gene region. While there is a rising interest in gene haplotypes, much of the corresponding analyses were carried out manually (Supplementary material). In addition, DnaSP is a software package for comprehensive analysis of DNA polymorphism data and also allow analysis on haplotype phasing (Rozas et al. 2017). HaplotypeMiner is an R package developed for exploring allelic diversity at genes of interest in a plant breeding context. The program minimally takes as input a dataset of SNP markers and the genomic position of a gene of interest, and outputs a set of possible haplotypes defined by the genotypes of a reduced number of neighboring SNPs (Tardivel et al. 2019). The RFGB (Rice Functional and Genomic Breeding) contains a set of core collection of about 3000 rice accessions (3 K rice genome). This dataset provides a base for the large-scale novel allele mining for important traits in rice with various bioinformatics and genetic approaches. It also contains Haplotype module, which was designed to fulfill the increasing demands of mining the associations between sequence variations within certain gene/region/SNP and target traits (Wang et al. 2020). However, a graphical haplotype analysis tool (no basic programming skill is required) that can handle any species is in urgent need.

Methods

Materials, genotyping, and annotation

A set of 398 accessions, including 360 foxtail millets and 38 *S. viridis*, was used to examine and calibrate our approach. These accessions have recently been analyzed via whole-genome sequencing (WGS) (Li et al. 2022). The clean reads were mapped onto the *xiaomi* reference genome (Yang et al. 2020). SNPs and InDels were detected using GATK (McKenna et al. 2010) (ver. 3.8.0). The SNPs and InDels were considered valid if they met the following requirements: (1) two alleles only; (2) excluding sites on the basis of the proportion of missing data > 0.9; (3) minor allele frequency ≥ 0.05; (4) mean depth values ≥ 5. SNPs that did not meet these four criteria were excluded from the study. We generated a final set of 4,158,075 filtered single-nucleotide polymorphisms (SNPs) and insertion-deletion polymorphisms (InDels) for further analysis. All identified SNPs that passed quality screening were further annotated with ANNOVAR (Wang et al. 2010) or SnpEff (Cingolani et al. 2012) based on the gene annotation of the reference genome (Wang et al. 2010). In practical application, users can adjust the above parameters. When a VCF file was submitted, ANNOVAR or SnpEff was computed to rapidly categorize the effects of variants in the reference genome sequence. ANNOVAR or SnpEff annotates variants based on their genomic locations (annotated genomic locations can be intronic, exonic, or intergenic) and predicts coding effects (mainly synonymous or non-synonymous amino-acid replacement). The process can be applied to any plant, animal, and bacteria species, by providing the genome file and its GFF (generic feature format) annotation file. The annotated vcf file was converted to HapMap using a Perl script vcf2hmp.pl, and this step would normally take several hours (~ 3 h for 4 million SNPs). Finally, a haplotypes.hmp file was generated for further haplotype analysis.

Put vcf2hmp.pl test.gff, test.vcf, and genome.fa files in the same directory, then run:

```
# 1. To annotate the vcf by ANNOVAR:
  gffread  test.gff   -T -o test.gtf
  gtfToGenePred -genePredExt test.gtf  si_refGene.txt
  retrieve_seq_from_fasta.pl --format refGene --seqfile  genome.fa  si_refGene.txt --outfile si_refGeneMrna.fa
  table_annovar.pl  test.vcf ./ --vcfinput --outfile  test --buildver  si --protocol refGene --operation g -remove

# 2. To convert the txt result of annovar to hapmap format:
  perl  vcf2hmp.pl  test.vcf  test.si_multianno.txt
```

Platforms and install

CandiHap was written in Perl 5 and R, which supported Windows, Mac, or UNIX platform computers in the graphical user interface (GUI) or command lines. Graphics were created by R. In addition to the GUI, users can also run CandiHap through command lines and install the R software environment (https://www.r-project.org). The code was compiled for the UNIX platforms and Windows 64-bit environment, and tested with CentOS 7, Windows 10/11 as well as Mac OS 10/13. For a given SNP that was found significantly in GWAS, the run time was about 1 min for a set of 398 samples with 4,158,075 marks. The CandiHap tool is open source, available on multiple platforms, and freely available online (https://github.com/xukaili/CandiHap or https://bigd.big.ac.cn/biocode/tools/BT007080).

A user-friendly graphical user interface software package of CandiHap, installable on Windows platforms, is implemented using electron development toolkit, which is freely available and not required for registration. For Windows users' convenience, the installation package integrates Perl and R

modules needed to run independently, meaning no more software installation is required. However, for the Mac Os or UNIX platform users, installation of the R software environment is required, followed by three packages by command install.packages (c("ggplot2","agricolae", "ggbeeswarm")) in R.

Parameter for CandiHap

Intergenic SNPs are SNPs that are located at least 5 kb up- or downstream of a gene. In general, they are not associated with a gene and not located in a known regulatory region. We set a default parameter in CandiHap, which limits the mapping SNPs to 2000 bp upstream and 500 bp downstream of a gene. The default settings ensure that the result is based on the association signals in gene(s) with statistical significance. Users may also adjust the parameter in "CandiHap.pl." Using Perl and R, the analysis provided and displayed various statistical results for haplotypes such as annotation statistics, types of variation, number of varieties, variety ID and phenotypes, mean, SD (standard deviation) of phenotypes, and significant phenotype differences. A boxplot of the gene showed a significant difference in haplotype–phenotype association analysis. The least significant difference (LSD) test is used to determine whether or not the difference between or among group means is significant.

Put CandiHap.pl and Phenotype.txt, Your.hmp, and genome.gff files in the same directory, then run:

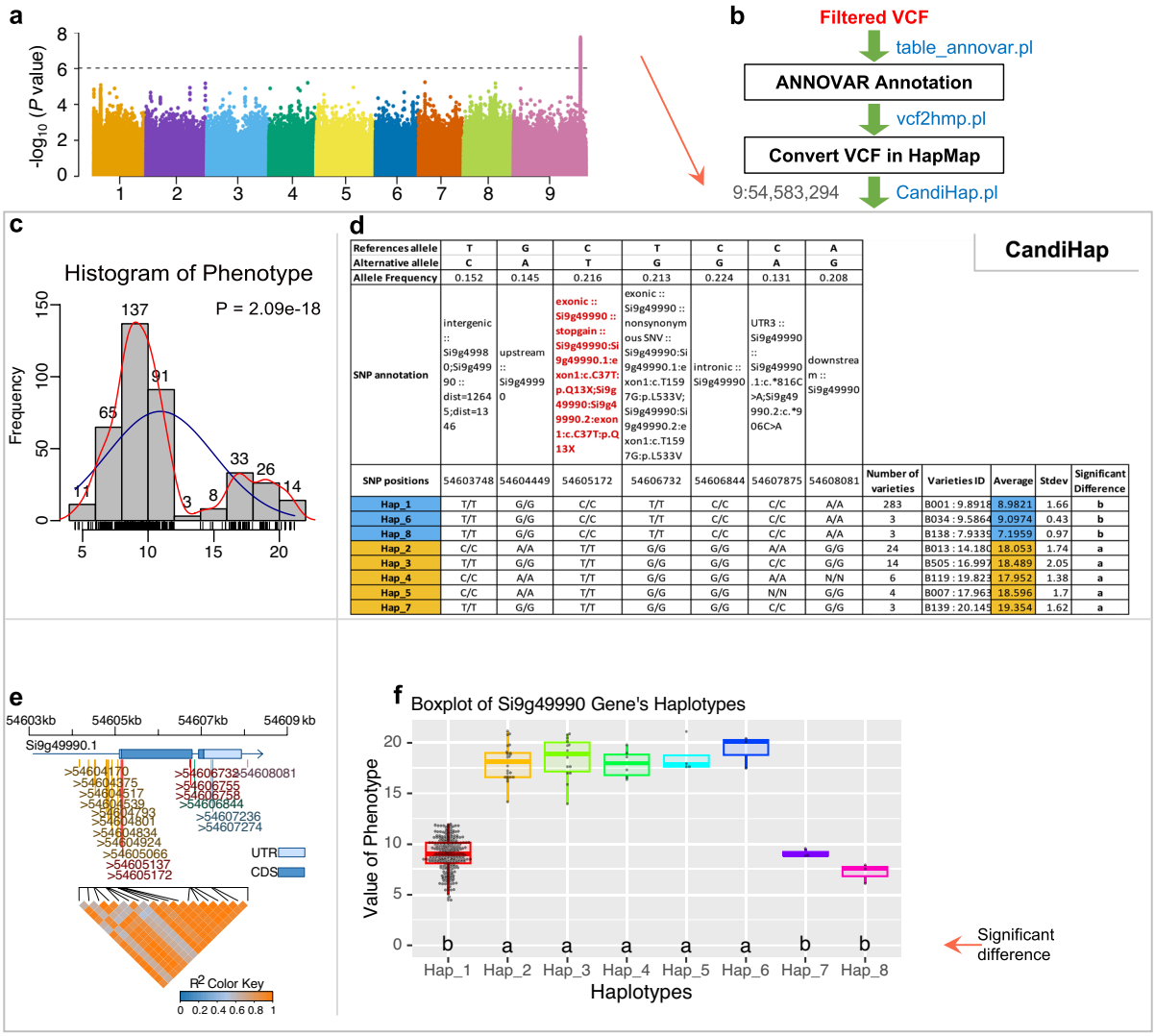
```
# 3. To run CandiHaplotypes
    perl  CandiHap.pl  -m haplotypes.hmp  -f test.gff  -p Phenotype.txt  -g Si9g49990
    perl  CandiHap.pl  -m haplotypes.hmp  -f test.gff -p Phenotype.txt  -g Si9g49990 -d 1000 -l 1 -n Structure.txt
    The command parameters are:
       -m    input hmp file name (Must).
       -p    input phenotype file name (Must).
       -f    input gff file name (Must).
       -g    Your gene ID (Must).
       -s    p value of wilcox test. default is 1.
       -u    gene upstream. default is 2000 bp.
       -d    gene downstream. default is 500 bp.
       -l    Plot LDheatmap (1) or not (0). default is not 0. require R package "LDheatmap" and "genetics".
       -n    input pop file name and plot haploNet figure. default is NULL. require R package "pegas" and "sf".
       -k    keek all tmp files.
       -h    this (help) message.
```

Graphical user interface

In addition to the command-line environment, we also provide user-friendly graphical interface software package of CandiHap on Windows platforms. For the convenience of the Windows platform, the installation package integrates necessary Perl and R modules for running independently by PyInstaller (5.7.0, https://pyinstaller.org/en/stable/), meaning no more software installation required. The installation package also contains the test dataset as the default value of the GUI version.

Sanger data for haplotype analysis

The "sanger_CandiHap.sh" was written in Shell, Perl 5, and R (with sangerseqR), which only supported the UNIX platforms in command lines. We developed a Perl script "ab1-fastq.pl" for reading ABI Sanger sequencing trace file and simulating the primarySeq and secondarySeq to fastq reads by extracting 90-bp blocks from Sanger sequence and shifting 1 bp in turns. As an example, a 200-bp Sanger sequence would obtain 110 fastq reads within the length of 90 bp. Then, mapping the new fastq reads to reference gene sequence is transferred into



References allele	T	G	C	T	C	C	A
Alternative allele	C	A	T	G	G	A	G
Allele Frequency	0.152	0.145	0.216	0.213	0.224	0.131	0.208
SNP annotation	intergenic :: Si9g49980;Si9g49990 :: dist=12645;dist=1346	upstream :: Si9g49990	exonic :: Si9g49990 :: stopgain :: Si9g49990:Si9g49990.1:exon1:c.C37T:p.Q13X;Si9g49990:Si9g49990.2:exon1:c.C37T:p.Q13X	exonic :: Si9g49990 :: nonsynonymous SNV :: Si9g49990:Si9g49990.1:exon1:c.T1597G:p.L533V; Si9g49990:Si9g49990.2:exon1:c.T1597G:p.L533V	intronic :: Si9g49990	UTR3 :: Si9g49990 :: Si9g49990.1:c.*816C>A;Si9g49990.2:c.*906C>A	downstream :: Si9g49990

SNP positions	54603748	54604449	54605172	54606732	54606844	54607875	54608081	Number of varieties	Varieties ID	Average	Stdev	Significant Difference
Hap_1	T/T	G/G	C/C	T/T	C/C	C/C	A/A	283	B001 : 9.8918	8.9821	1.66	b
Hap_6	T/T	G/G	C/C	T/T	C/C	C/C	A/A	3	B034 : 9.5864	9.0974	0.43	b
Hap_8	T/T	G/G	C/C	T/T	C/C	C/C	A/A	3	B138 : 7.9339	7.1959	0.97	b
Hap_2	C/C	A/A	T/T	G/G	G/G	A/A	G/G	24	B013 : 14.180	18.053	1.74	a
Hap_3	T/T	G/G	T/T	G/G	G/G	C/C	G/G	14	B505 : 16.997	18.489	2.05	a
Hap_4	C/C	A/A	T/T	G/G	G/G	A/A	N/N	6	B119 : 19.823	17.952	1.38	a
Hap_5	C/C	A/A	T/T	G/G	G/G	N/N	G/G	4	B007 : 17.963	18.596	1.7	a
Hap_7	T/T	G/G	T/T	G/G	G/G	C/C	G/G	3	B139 : 20.145	19.354	1.62	a

Fig. 1 Overview of the CandiHap process. **a** A GWAS result. **b** General scheme of the process. **c** The histogram of phenotype and Shapiro test to check whether the data is normally distributed data or not. **d** The statistics of haplotypes and significant differences haplotypes are highlighted by color boxes. **e** Gene structure and SNPs of a critical gene. **f** Boxplot of a critical gene's haplotypes and the LSD test is used to determine whether or not the difference between or among group means is significant

a call SNP process for next-generation sequencing. Burrows–Wheeler aligner software (Li and Durbin 2010) (BWA mem, ver. 0.7.17) was used to map the fastq reads with default parameters onto the gene reference sequence of all samples. Mapped reads were converted into BAM files using SAMtools (ver. 1.7). The variants including SNPs and indels were detected using GATK (McKenna et al. 2010) (ver. 3.8.0). Hard filtering was applied to the raw variant set using GATK. The results are filtered to retain the homozygous mutation sites. It is an open source and freely available online (https://github.com/xukaili/CandiHap/tree/master/Sanger_ab1_Linux or https://ngdc.cncb.ac.cn/biocode/tools/7080/releases/v1.0.1).

Put sanger_CandiHap.sh, Gene_VCF2haplotypes.pl, ab1-fastq.pl, and all.ab1 files in the same directory (require R package "sangerseqR"), then run:

```
sh  sanger_CandiHap.sh  PHYC.txt
```

Results

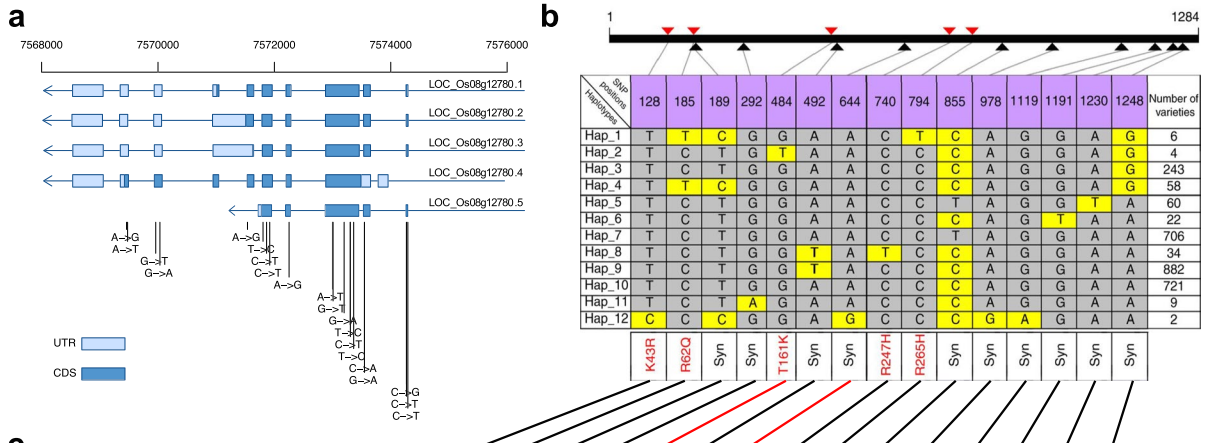
We developed a user-friendly software, CandiHap, that may be operated on a range of computer platforms. In CandiHap, users can identify polymorphisms based on the models of gene haplotypes within vcf file and to report results in a variety of formats, including tables and figures. CandiHap allows researchers to explore favorable haplotypes of candidate genes for target traits, providing a guide to study underlying genetic mechanisms. In addition, some researchers use Sanger sequences to detect the mutations that underlie a number of traits, yet it is challenging to determine heterozygotes from Sanger ab1 files and conduct haplotype analysis. The "Sanger_CandiHap.sh" in CandiHap allows fast identification of the haplotype from Sanger ab1 files.

An overview of the process is presented in Fig. 1. Starting from a VCF file as an entry point, CandiHap first annotates the variants using an annotated reference genome to produce a new VCF file. This new VCF file is then used to mine variants and genotyping data, and sent into a series of modules in charge of various processes. Users can then analyze variants ranging from genome to single gene levels. The GWAS results of genomic regions (Fig. 1a) and LD can be defined by entering the limits, and the application would loop and process all genes in the LD regions. The CandiHap implements a three-stage analysis (Fig. 1b): the first annotates the VCF file for GWAS by ANNOVAR (table_annovar.pl); the second converts the txt result of ANNOVAR to hapmap format (vcf2hmp.pl); and the third stage requires input data of hapmap file, GFF file of your reference genome, the phenotype data, the LD, and the most significant SNP position of GWAS result. If users need only to run one gene, the vcf, phenotype, gff, and gene ID need to be input. Besides the graphical user interface (GUI) software, users can run CandiHap through command lines on UNIX, Mac, or DOS platforms. The output includes a txt file of haplotypes with detailed information and three pdf files of figures (Fig. 1c–f). The results of haplotypes include references allele, alternative allele, allele frequency, SNP annotation, SNP positions, and haplotypes (Fig. 1d). The information for each haplotype also includes number of varieties, varieties ID and its phenotype, average, SD of phenotype, and significant difference (Fig. 1d). For the graphical user interface (GUI), CandiHap analytical pipeline is divided into three functional modules, vcf2hmp, CandiHap, and GWAS_LD2haplotypes, which correspond to the command line steps. First, the "txt" and "vcf" files of ANNOVAR results with genotype information are required as input for module vcf2hmp to convert the txt result of ANNOVAR to hapmap format. Then, CandiHap module can detect a single specific gene or GWAS_LD2haplotypes module for an LD region.

To exemplify CandiHap, we performed a GWAS analysis of foxtail millet (Li et al. 2022). Approximately 3679 K SNPs were tested; of them, 531 SNPs passed the threshold of $P\text{-value} < 9.42 \times 10^{-7}$. The most significant SNP was located at chr9 at position of 54,583,294 with $P\text{-value} = 1.23 \times 10^{-8}$ (Fig. 1a), and CandiHap identified a candidate causal gene (*Si9g49990*) within 50 kb LD of this SNP. We have identified a signal at position 54,605,172 ($P\text{-value} = 1.03 \times 10^{-7}$), leading to stop gain of *Si9g49990* (Fig. 1d). The boxplot of *Si9g49990*, for haplotype-phenotype association analysis, showed significant differences in the phenotype of each haplotype between Hap 1, 2, 6 and Hap 3, 4, 5, 7, 8, 9, with intuitive supporting evidence (Fig. 1f). Only the SNPs and haplotypes found in ≥ 2 accessions were used to construct the haplotype network for *Si9g49990* (Supplementary Fig. 1).

To further test the universality of CandiHap in haplotype analysis, we analyzed the haplotypes of the *ARE1* gene in rice (Wang et al. 2018b), and the same result was obtained except that five more SNPs and two errors were identified in our study (highlighted by blue and red boxes). The discrepancy is due to the fact that there are 276 more rice varieties used in our study, and authors analyze the haplotype of *ARE1* gene manually (Fig. 2).

In addition to NGS data, Sanger sequencing technology is also widely used in natural variation analysis. To meet this demand, "sanger_CandiHap.sh" was developed for process of Sanger sequencing data (Fig. 3). Starting from ab1 files as the entry point, the process first simulates ABI Sanger sequencing trace data to fastq reads and then maps the fastq reads to reference gene sequence as for re-sequencing program (Fig. 3a). The output is a txt file of haplotypes with detailed information, including references allele, alternative allele, SNP positions, and haplotypes. The information for each haplotype consists of the number of samples and sample ID (Fig. 3c). As an example, the



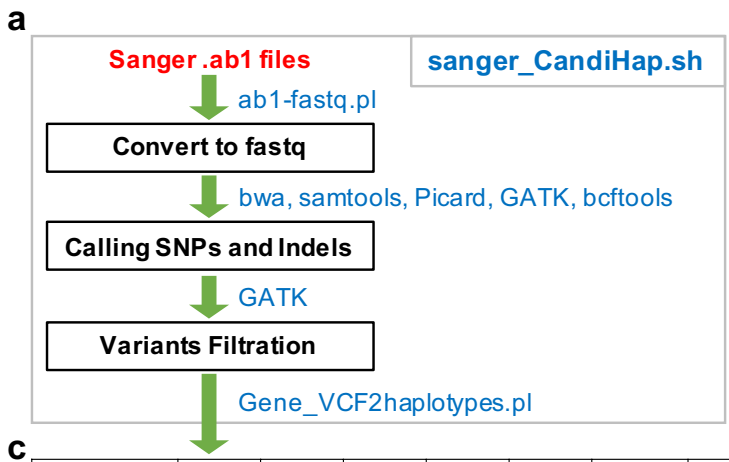
References allele	C	C	C	G	C	T	C	T	G	G	A	A	C	C	T	A	G	G	A	A	
Alternative allele	T	T	G	A	A	C	T	C	A	T	T	G	T	T	C	G	A	T	T	G	Total varieties: 3023
LOC_Os08g12780.4:CDS						128	185	189	292	482	492	644	740	794	855	978	1119	1191	1230	1248	
Allele Frequency	0.00401	0.0015	0.001	0.00415	0.00365	0.00149	0.02671	0.02738	0.00315	0.00133	0.33444	0.00133	0.01228	0.00332	0.27473	0.00166	0.00116	0.00813	0.02261	0.12155	
SNP annotation	Syn : T->T	Syn : T->T	splicing: 33+1G>C	Syn : C->C	nonSyn : A->S	nonSyn : K->R	nonSyn : R->Q	Syn : P->P	Syn : L->L	nonSyn : T->K	Syn : P->P	nonSyn : V->A	nonSyn : R->H	nonSyn : R->H	Syn : L->L	Syn : S->S	Syn : V->V	Syn : A->A	Syn : L->L	Syn : S->S	
SNP positions	7574287	7574263	7574262	7573542	7573541	7573357	7573300	7573296	7573193	7573003	7572993	7572248	7571916	7571862	7571801	7571534	7570031	7569959	7569476	7569458	Number of varieties
Hap_16	C/C	C/C	C/C	G/G	C/C	T/T	T/T	C/C	G/G	G/G	A/A	A/A	C/C	T/T	C/C	A/A	G/G	G/G	A/A	G/G	8
Hap_20	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	T/T	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	G/G	4
Hap_4	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	G/G	259
Hap_5	C/C	C/C	C/C	G/G	C/C	T/T	T/T	C/C	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	G/G	59
Hap_23	C/C	T/T	C/C	G/G	C/C	T/T	T/T	C/C	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	G/G	3
Hap_40	C/C	C/C	C/C	G/A	C/C	T/T	T/T	C/C	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	G/G	1
Hap_6	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	T/T	A/A	G/G	G/G	T/T	A/A	59
Hap_41	C/C	C/C	C/C	G/A	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	T/T	A/A	G/G	G/G	T/T	A/A	1
Hap_10	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	T/T	A/A	A/A	22
Hap_2	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	T/T	A/A	G/G	G/G	A/A	A/A	727
Hap_42	C/C	C/T	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	T/T	A/A	G/G	G/G	A/A	A/A	1
Hap_8	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	T/T	A/A	T/T	C/C	C/C	A/A	G/G	G/G	A/A	A/A	35
Hap_1	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	T/T	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	A/A	914
Hap_43	C/T	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	T/T	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	A/A	1
Hap_3	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	A/A	719
Hap_13	T/T	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	A/A	11
Hap_17	C/C	C/C	C/C	A/A	A/A	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	A/A	8
Hap_24	C/C	C/C	G/G	A/A	A/A	T/T	C/C	T/T	G/G	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	A/A	3
Hap_15	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	A/A	G/G	A/A	A/A	C/C	C/C	C/C	A/A	G/G	G/G	A/A	A/A	9
Hap_29	C/C	C/C	C/C	G/G	C/C	C/C	C/C	C/C	G/G	G/G	A/A	G/G	C/C	C/C	C/C	G/G	A/A	G/G	A/A	A/A	2
Hap_18	C/C	C/C	C/C	G/G	C/C	T/T	C/C	T/T	G/G	G/G	T/T	A/A	C/C	C/C	T/T	A/A	G/G	G/G	A/A	A/A	6
Hap_44	C/C	C/C	C/C	G/G	C/C	T/T	T/T	C/C	G/G	G/G	A/A	A/A	C/C	C/C	T/T	A/A	G/G	G/G	A/A	G/G	1

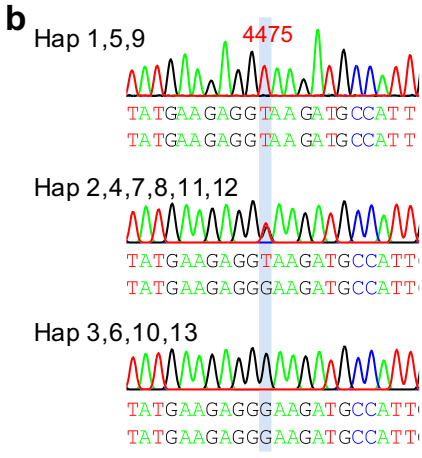
Fig. 2 Haplotype analysis of the *ARE1* gene in rice compared with the results by Wang et al. (2018b). **a** Gene structure and SNPs of *ARE1*. **b** Major haplotypes of SNPs in the *ARE1* coding region of 2747 rice varieties (from Wang's article Fig. 5b, https://www.nature.com/articles/s41467-017-02781-w/figures/5). **c** The haplotype results of *ARE1* coding region of 3023 rice varieties using CandiHap (SNPs data were downloaded from RFGB, http://www.rmbreeding.cn). Major SNP haplotypes and casual variations in the encoded amino acid residues are shown. The five more SNPs is due to the fact that there are 276 more rice varieties used in our study (highlighted by blue boxes), and two errors are highlighted by red boxes

PHYC gene has a mutation at 4475 and has homozygous samples (Fig. 3b). Runtime is ~20 min for a set of 100 samples of Sanger sequences. The results are filtered to retain the homozygous mutation sites, allowing users to find the mutation easily (Fig. 3b, c). Moreover, the heterozygous mutation was also shown in the results (Fig. 3c). There are 57 wild types and 43 variations, and 40 homozygous samples showed wild-type phenotype (Fig. 3c). The test data files can be freely downloaded (https://github.com/xukaili/CandiHap/tree/master/Sanger_ab1 or https://ngdc.cncb.ac.cn/biocode/tools/7080/releases/v1.0.1/file/download?filename=sanger_CandiHap-1.0.1.zip).

Discussion

We refer to the haplotype not as the strong inter-marker LD, but rather the SNPs and indels within a gene region, including upstream, downstream, exonic, and intronic regions. We term this as the gene haplotype, which specifically aims to select a collection of SNPs that are useful for defining SNP haplotypes





c

References allele Alternative allele	CA C	G C	G C	G T	A AG	Number of samples	Sample ID
SNP positions	4243	4271	4273	4475	4887		
Hap_1	CA/CA	G/G	G/G	T/T**	A/A	36	B1, B10, B11, B14, B15, B16, B17, B18, B19, B20, B21, B22, B23, B24, B25, B26, B28, B29, B3, B30, B31, B32, B33, B34, B35, B36, B37, B38, B4, B41, B42, B5, B6, B7, B8, B9
Hap_5	CA/CA	G/C	G/G	T/T**	A/A	6	B12, B13, B27, B39, B40, B44
Hap_9	CA/CA	C/C**	G/G	T/T**	A/A	1	B43
Hap_2	CA/CA	G/G	G/G	G/T	A/A	24	W-1, W-11, W-13, W-18, W-20, W-21, W-22, W-31, W-32, W-34, W-35, W-37, W-41, W-43, W-45, W-46, W-48, W-8, W-9, W101, W54, W72, W86, W97
Hap_4	CA/CA	G/C	G/C	G/T	A/A	11	W-14, W-15, W-17, W-23, W-24, W-25, W-26, W-47, W-6, W85, W91
Hap_7	CA/CA	G/G	G/C	G/T	A/A	2	W-3, W98
Hap_8	CA/CA	G/G	G/C	G/T	A/AG	1	W-7
Hap_11	CA/CA	G/G	C/C**	G/T	AG/AG**	1	W71
Hap_12	CA/CA	G/G	G/G	G/T	AG/AG**	1	W-40
Hap_3	CA/CA	G/G	G/G	G/G	A/A	11	GBS, W-10, W-16, W-28, W-33, W-36, W-38, W-39, W-44, W87, W99
Hap_6	CA/CA	G/C	G/C	G/G	A/A	4	W-30, W-4, W-5, W93
Hap_10	C/C**	G/C	G/C	G/G	A/A	1	W79
Hap_13	CA/CA	G/G	G/C	G/G	A/A	1	W-12

Fig. 3 Overview of the sanger_CandiHap process. **a** General scheme of the process from Sanger ab1 files. **b** PeakTrace of ab1 images of three main genotypes. **c** The statistics of haplotypes

around a target gene and could be adopted to evaluate the variants of the affected traits captured from the gene region. Also, all the individuals that share a given gene are systematically pooled into haplotypes. The CandiHap can be widely used for the investigations of natural variations. It should be noted that CandiHap is not intended to be used to predict true causal SNPs and gene(s) for complex traits. Therefore, CandiHap outputs are candidate causal SNPs and gene(s), which allows users to screen for useful ones. An essential application of the CandiHap results is to allow investigators to test "a priori" hypothesis by using candidate causal SNPs as a practical starting point.

In terms of other software, the DnaSP input data are required in FASTA format that is not suitable for analyzing large-scale population data, such as vcf (Rozas et al. 2017); the HaplotypeMiner is a command-line tool using parameters (Tardivel et al. 2019), and RFGB can only analyze the 3 K rice genome data (Wang et al. 2020). Due to the above limitations, the CandiHap is a graphical haplotype analysis tool (no basic programming skill is required) that can handle any species, which is in urgent need.

Several factors can affect the best settings for haplotype definition of a given gene; we have therefore allowed users to select the desired settings for some parameters: (1) the *P*-value of Wilcox test, (2) the LD threshold, (3) the length of gene upstream and downstream, and (4) the phenotype data. Collapsing SNPs to retain a single tag SNP by using a higher threshold (*P*-value of Wilcox test = 1) leads to no loss of information from the dataset but has the disadvantage of being sensitive to genotyping errors. On the contrary, a lower threshold, as was used here (*P*-value of Wilcox test = 0.01), will mask small genotyping errors but may also lead to the loss of haplotypes specific to rare alleles. The ability to capture informative markers in the vicinity of the gene is dependent on the density of genotyping. For this approach to work best, the genotyping should be sufficiently dense to deliver at least one marker at a gene. We successfully identified gene haplotypes offering a very good match with the allelic status at *Si9g49990* and *LOC_Os08g12780* (*ARE1*) genes under study. Because of manual analysis on *ARE1* gene in rice, these five alleles were misidentified, and two errors were identified in Wang et al. (2018a, b) (Fig. 2).

Breeding aims to produce new lines that carry an array of alleles that jointly produce a superior phenotype. A key part of this work is to assess the allelic diversity present in germplasm collections and to identify individuals carrying favorable alleles at these genes. In this work, we demonstrate how our approach can provide accurate and essential information for breeding by delivering a quick and clear picture of the allelic diversity for a gene within a given germplasm collection. This approach was also found to be reproducible on two distinct genotypic datasets (next-generation sequencing and Sanger sequencing) with similar results. Ultimately, our approach for identification of gene haplotypes from large SNP datasets represents a promising approach to routinely assess allelic variation in a gene region. In the future, CandiHap will be regularly updated and expanded to perform more functions with more user-friendly options.

Acknowledgements The authors thank the CandiHap users for helpful comments and discussions. We also thank Dr. Staffan Persson (University of Melbourne) and Dr. Yiwei Jiang (Purdue University) for their critical reading of the manuscript.

Author contribution XL and KG supervised the study. XL, XW, and KG conceived of the study idea. JG carried out the sampling process and participated in the material preparation. XL and ZS performed most of the experiments. XL, ZS, and KG analyzed the data. XL drafted the manuscript. All of the authors discussed the results and commented on the manuscript.

Funding This work has been supported by the National Key R&D Program of China (2022YFC3400300); the National Natural Science Foundation of China (32001608); the Major Special Science and Technology Projects in Shanxi Province (202101140601027); the National Laboratory of Minor Crops Germplasm Innovation and Molecular Breeding (in preparation) (202204010910001–02).

Data availability The raw sequencing data reported in this paper have been deposited in the NCBI under BioProject accession no. PRJNA633413. This data is also available in the BIG Data Center under the accession number CRA002636. SNP data of *ARE1* coding region of 3023 rice varieties were downloaded from RFGB (http://www.rmbreeding.cn). CandiHap code is available online (https://github.com/xukaili/CandiHap or https://ngdc.cncb.ac.cn/biocode/tools/BT007080).

Declarations

Conflict of interest The authors declare no competing interests.

References

Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32(10):1493–1501. https://doi.org/10.1093/bioinformatics/btw018

Cingolani P, Platts A, le Wang 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 (austin) 6(2):80–92. https://doi.org/10.4161/fly.19695

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296(5576):2225–2229. https://doi.org/10.1126/science.1069424

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333–351. https://doi.org/10.1038/nrg.2016.49

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E (2014) Identifying causal variants at loci with multiple signals of association. Genetics 198(2):497–508. https://doi.org/10.1534/genetics.114.167908

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet 10(10):e1004722. https://doi.org/10.1371/journal.pgen.1004722

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595. https://doi.org/10.1093/bioinformatics/btp698

Li X, Gao J, Song J, Guo K, Hou S, Wang X, He Q, Zhang Y, Zhang Y, Yang Y, Tang J, Wang H, Persson S, Huang M, Xu L, Zhong L, Li D, Liu Y, Wu H, Diao X, Chen P, Wang X, Han Y (2022) Multi-omics analyses of 398 foxtail millet accessions reveal genomic regions associated with domestication, metabolite traits, and anti-inflammatory effects. Mol Plant 15(8):1367–1383. https://doi.org/10.1016/j.molp.2022.07.003

McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17(R2):R156–R165. https://doi.org/10.1093/hmg/ddn289

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky 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(9):1297–1303. https://doi.org/10.1101/gr.107524.110

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17(2):240–248. https://doi.org/10.1101/gr.5681207

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294(5547):1719–1723. https://doi.org/10.1126/science.1065573

Patil GB, Lakhssassi N, Wan J, Song L, Zhou Z, Klepadlo M, Vuong TD, Stec AO, Kahil SS, Colantonio V, Valliyodan B, Rice JH, Piya S, Hewezi T, Stupar RM, Meksem K, Nguyen HT (2019) Whole-genome re-sequencing reveals the impact of the interaction of copy number variants of the rhg1 and Rhg4 genes on broad-based resistance to soybean cyst nematode. Plant Biotechnol J 17(8):1595–1611. https://doi.org/10.1111/pbi.13086

Raychaudhuri S (2011) Mapping rare and common causal alleles for complex human diseases. Cell 147(1):57–69. https://doi.org/10.1016/j.cell.2011.09.011

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34(12):3299–3302. https://doi.org/10.1093/molbev/msx248

Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet 19(8):491–504. https://doi.org/10.1038/s41576-018-0016-z

Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9(6):477–485. https://doi.org/10.1038/nrg2361

Tardivel A, Torkamaneh D, Lemay MA, Belzile F, O'Donoughue LS (2019) A systematic gene-centric approach to define haplotypes and identify alleles on the basis of dense single nucleotide polymorphism datasets. Plant Genome 12(3):1–11. https://doi.org/10.3835/plantgenome2018.08.0061

Thudi M, Khan AW, Kumar V, Gaur PM, Katta K, Garg V, Roorkiwal M, Samineni S, Varshney RK (2016) Whole genome re-sequencing reveals genome-wide variations among parental lines of 16 mapping populations in chickpea (Cicer arietinum L.). BMC Plant Biol 16 Suppl 1 (Suppl 1):10–10 https://doi.org/10.1186/s12870-015-0690-3

Tinker NA, Bekele WA, Hattori J (2016) Haplotag: software for haplotype-based genotyping-by-sequencing analysis. G3 6 (4):857–863 https://doi.org/10.1534/g3.115.024596

Uffelmann E, Posthuma D (2021) Emerging methods and resources for biological interrogation of neuropsychiatric polygenic signal. Biol Psychiatry 89(1):41–53. https://doi.org/10.1016/j.biopsych.2020.05.022

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D (2021) Genome-wide association studies. Nature Reviews Methods Primers 1(1):59. https://doi.org/10.1038/s43586-021-00056-9

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90(1):7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101(1):5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71(5):1227–1234. https://doi.org/10.1086/344398

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164–e164. https://doi.org/10.1093/nar/gkq603

Wang Q, Nian J, Xie X, Yu H, Zhang J, Bai J, Dong G, Hu J, Bai B, Chen L, Xie Q, Feng J, Yang X, Peng J, Chen F, Qian Q, Li J, Zuo J (2018b) Genetic variations in ARE1 mediate grain yield by modulating nitrogen utilization in rice. Nat Commun 9(1):735. https://doi.org/10.1038/s41467-017-02781-w

Wang G, Sarkar A, Carbonetto P, Stephens M (2018a) A simple new approach to variable selection in regression, with

application to genetic fine-mapping. bioRxiv https://doi.org/10.1101/501114

Wang C-C, Yu H, Huang J, Wang W-S, Faruquee M, Zhang F, Zhao X-Q, Fu B-Y, Chen K, Zhang H-L, Tai S-S, Wei C, McNally KL, Alexandrov N, Gao X-Y, Li J, Li Z-K, Xu J-L, Zheng T-Q (2020) Towards a deeper haplotype mining of complex traits in rice with RFGB v2.0. Plant Biotechnol J 18 (1):14–16 https://doi.org/10.1111/pbi.13215

Yang Z, Zhang H, Li X, Shen H, Gao J, Hou S, Zhang B, Mayes S, Bennett M, Ma J, Wu C, Sui Y, Han Y, Wang X (2020) A mini foxtail millet with an *Arabidopsis*-like life cycle as a C_4 model system. Nat Plants 6(9):1167–1178. https://doi.org/10.1038/s41477-020-0747-7

Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, Frayling TM, McCarthy MI, Hirschhorn JN, Goddard ME, Visscher PM (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44 (4):369–375, s361–363 https://doi.org/10.1038/ng.2213

Zhang K, Calabrese P, Nordborg M, Sun F (2002a) Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71(6):1386–1394. https://doi.org/10.1086/344780

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002b) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99(11):7335–7339. https://doi.org/10.1073/pnas.102186799