Supplementary Material

CandiHap: a haplotype analysis toolkit for natural variation study

Xukai Li[1,*], Zhiyong Shi[1], Jianhua Gao[1], Xingchun Wang[1], Kai Guo[2,*]

[1]Shanxi Key Laboratory of Minor Crop Germplasm Innovation and Molecular Breeding, College of Life Sciences, Shanxi Agricultural University, Taigu 030801, China, [2]Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA

There has been recently great interest on gene haplotypes (Aung, et al., 2017; Basu, et al., 2019; Basu, et al., 2019; Basu, et al., 2019; Chao, et al., 2017; Hou, et al., 2017; Jäger, et al., 2013; Li, et al., 2019; Li, et al., 2014; Malik, et al., 2016; Mao, et al., 2019; Miao, et al., 2017; Mopidevi, et al., 2019; Sato, et al., 2009; Tu, et al., 2018; Wang, et al., 2018; Wang, et al., 2018; Webster, et al., 2015; Yang, et al., 2018; Zeng, et al., 2020; Zhang, et al., 2017; Zhang, et al., 2019), but to the best of our knowledge, those analyses were carried out manually. However, currently available tools are web-based or command-lines implemented for studies on human and rice traits, severely limiting their wide applications (Adzhubei, et al., 2010; Johnson, et al., 2008; Kumar, et al., 2009; Lee and Shatkay, 2008; Mi, et al., 2010; Saccone, et al., 2010; Schmitt, et al., 2010; Wang, et al., 2020; Xu and Taylor, 2009; Yuan, et al., 2006; Yue, et al., 2006). We developed a user-friendly local software, CandiHap (https://github.com/xukaili/CandiHap), which can preselect candidate causal SNPs from Sanger or next-generation sequencing data, and applied to any species of plant, animal and microbial. It could be operated on Windows, UNIX or Mac computer platforms. Users can use CandiHap to specify a gene or linkage sites based on GWAS results and explore favourable haplotypes of candidate genes for target traits.

There are mainly three steps included in the CandiHap analytical through command lines, and the test data files can be freely downloaded at https://github.com/xukaili/CandiHap.

1. To annotate the vcf by ANNOVAR:
   1.1 gffread test.gff -T -o test.gtf
   1.2 gtfToGenePred -genePredExt test.gtf si_refGene.txt
   1.3 retrieve_seq_from_fasta.pl --format refGene --seqfile genome.fa si_refGene.txt --outfile si_refGeneMrna.fa
   1.4 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
3. To run CandiHap:
   perl GWAS_LD2haplotypes.pl -f genome.gff -m ann.hmp -p Phenotype.txt -l LDkb -c Chr:position
   e.g. perl GWAS_LD2haplotypes.pl -f test.gff -m haplotypes.hmp -p Phenotype.txt -l 50kb -c 9:54583294

   Or to run CandiHap by one gene:
   perl CandiHap.pl -m Your.hmp -f Genome.gff -p Phenotype.txt -g Your_gene_ID
   e.g. 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 -s 0.5 -u 2000 -d 500 -l 1 -n Structure.txt

The primary step in the 'sanger_CandiHap.sh' is made through only one simple command (The PHYC.txt is reference gene sequence):
   sh sanger_CandiHap.sh PHYC.txt

References

Adzhubei, I.A.*, et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* 2010;7(4):248-249.

Aung, T.*, et al.* Genetic association study of exfoliation syndrome identifies a protective rare variant at LOXL1 and five new susceptibility loci. *Nat. Genet.* 2017;49(7):993-1004.

Basu, U.*, et al.* Genetic dissection of photosynthetic efficiency traits for enhancing seed yield in chickpea. *Plant Cell Environ.* 2019;42(1):158-173.

Basu, U.*, et al.* CLAVATA signaling pathway genes modulating flowering time and flower number in chickpea. *Theor. Appl. Genet.* 2019;132(7):2017-2038.

Basu, U.*, et al.* ABC transporter-mediated transport of glutathione conjugates enhances seed yield and quality in chickpea. *Plant Physiol.* 2019;180(1):253-275.

Chao, M.J.*, et al.* Haplotype-based stratification of huntington's disease. *Eur. J. Hum. Genet.* 2017;25(11):1202-1209.

Hou, J.*, et al.* ADP-glucose pyrophosphorylase genes, associated with kernel weight, underwent selection during wheat domestication and breeding. *Plant Biotechnol. J.* 2017;15(12):1533-1543.

Jäger, J.*, et al.* IL7RA haplotype-associated alterations in cellular immune function and gene expression patterns in multiple sclerosis. *Genes Immun.* 2013;14(7):453-461.

Johnson, A.D.*, et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008;24(24):2938-2939.

Kumar, P., Henikoff, S. and Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* 2009;4(7):1073-1081.

Lee, P.H. and Shatkay, H. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.* 2008;36(Database issue):D820-D824.

Li, R.*, et al.* Combined linkage mapping and bsa to identify QTL and candidate genes for plant height and the number of nodes on the main stem in soybean. *Int. J. Mol. Sci.* 2019;21(1):42.

Li, Y.*, et al.* Chalk5 encodes a vacuolar H+-translocating pyrophosphatase influencing grain chalkiness in rice. *Nat. Genet.* 2014;46(4):398-404.

Malik, N.*, et al.* An integrated genomic strategy delineates candidate mediator genes regulating grain size and weight in rice. *Sci. Rep.* 2016;6(1):23253.

Mao, D.*, et al.* Natural variation in the HAN1 gene confers chilling tolerance in rice and allowed adaptation to a temperate climate. *Proc. Natl. Acad. Sci. USA.* 2019;116(9):3494-3501.

Mi, H.*, et al.* PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* 2010;38(Database issue):D204-D210.

Miao, L.*, et al.* Elite haplotypes of a protein kinase gene TaSnRK2.3 associated with important agronomic traits in common wheat. *Front. Plant Sci.* 2017;8(368).

Mopidevi, B.*, et al.* A polymorphism in intron I of the human angiotensinogen gene (hAGT) affects binding by HNF3 and hAGT expression and increases blood pressure in mice. *J. Biol. Chem.* 2019;294(31):11829-11839.

Saccone, S.F.*, et al.* SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res.* 2010;38(Web Server issue):W201-W209.

Sato, K.*, et al.* Strong evidence of a combination polymorphism of the tyrosine kinase 2 gene and the signal transducer and activator of transcription 3 gene as a DNA-based biomarker for susceptibility to crohn's disease in the japanese population. *J. Clin. Immunol.* 2009;29(6):815-825.

Schmitt, A.O.*, et al.* CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. *Bioinformatics* 2010;26(7):969-970.

Tu, Y.*, et al.* Monitoring conservation effects on a Chinese indigenous chicken breed using major histocompatibility complex B-G gene and DNA Barcodes. *Asian-Australas J. Anim. Sci.* 2018;31(10):1558-1564.

Wang, C.-C.*, et al.* Towards a deeper haplotype mining of complex traits in rice with RFGB v2.0. *Plant Biotechnol. J.* 2020;18(1):14-16.

Wang, M.*, et al.* Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.* 2018;50(10):1435-1441.

Wang, Q.*, et al.* Genetic variations in ARE1 mediate grain yield by modulating nitrogen utilization in rice. *Nat. Commun.* 2018;9(1):735.

Webster, M.T.*, et al.* Linked genetic variants on chromosome 10 control ear morphology and body mass among dog breeds. *BMC Genomics* 2015;16(1):474.

Xu, Z. and Taylor, J.A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* 2009;37(Web Server issue):W600-W605.

Yang, M.*, et al.* Genome-wide association studies reveal the genetic basis of ionomic variation in rice. *Plant Cell* 2018;30(11):2720-2740.

Yuan, H.-Y.*, et al.* FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.* 2006;34(Web Server issue):W635-W641.

Yue, P., Melamud, E. and Moult, J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 2006;7:166-166.

Zeng, X.*, et al.* Genome-wide Dissection of Co-selected UV-B Responsive Pathways in the UV-B Adaptation of Qingke. *Mol. Plant* 2020;13(1):112-127.

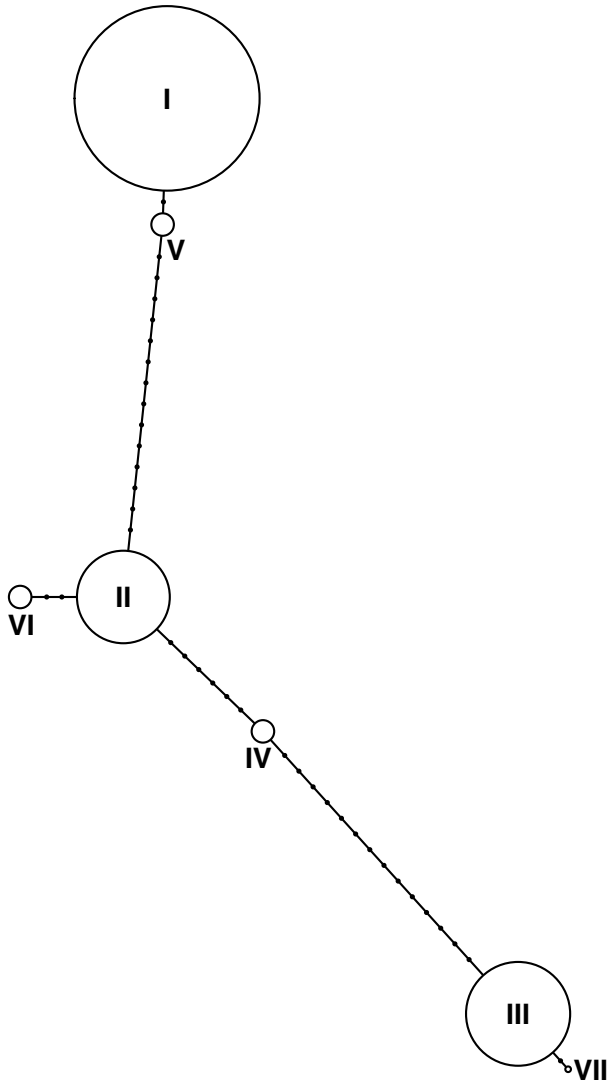
Zhang, F.*, et al.* Genome-wide association analysis identifies resistance loci for bacterial blight in a diverse collection of indica rice germplasm. *PLoS One* 2017;12(3):e0174598.

Zhang, H.-J.*, et al.* Transcription factor gene TaNAC67 involved in regulation spike length and spikelet number per spike in common wheat. *Acta Agronomica Sinica* 2019;45(11):1615-1627.

a

Pos	I	II	III	IV	V	VI	VII
NO.	284	17	24	2	2	2	2
1	G	G	T	G	G	G	T
2	G	G	A	G	G	G	A
3	T	T	C	T	T	C	C
4	A	A	G	A	A	A	G
5	T	C	C	C	T	C	C
6	T	T	C	T	T	T	C
7	T	T	G	G	T	T	G
8	T	T	C	T	T	T	C
9	G	G	A	G	G	G	A
10	G	T	T	T	G	T	T
11	C	T	T	T	C	T	T
12	G	G	C	G	G	G	C
13	T	A	A	A	A	A	A
14	C	C	A	A	C	C	A
15	C	C	T	T	C	C	T
16	G	G	A	A	G	G	A
17	T	T	G	G	T	T	G
18	C	C	T	C	C	C	T
19	T	G	G	G	T	G	G
20	A	C	C	A	A	C	C
21	T	A	A	A	T	A	A
22	C	T	T	T	C	T	C
23	C	C	T	C	C	C	T
24	G	G	A	G	G	G	A
25	T	G	G	G	T	T	G
26	C	G	G	G	C	G	G
27	C	T	T	T	C	T	T
28	A	A	G	A	A	A	G
29	C	G	G	G	C	G	G
30	T	C	C	C	T	C	C
31	A	C	C	C	A	C	C
32	C	C	A	C	C	C	A
33	A	G	G	G	A	G	G

b



Supplementary Fig. 1 | Haplotype network analysis of the *Si9g49990* gene in foxtail millet. a, The difference of haplotypes. **b**, Haplotype network. Note: only the SNPs and haplotypes found in ≥2 accessions were used to construct the haplotype network. The value of circle size had been transformed into $\log_2$.

The gene haplotype showed in some articles

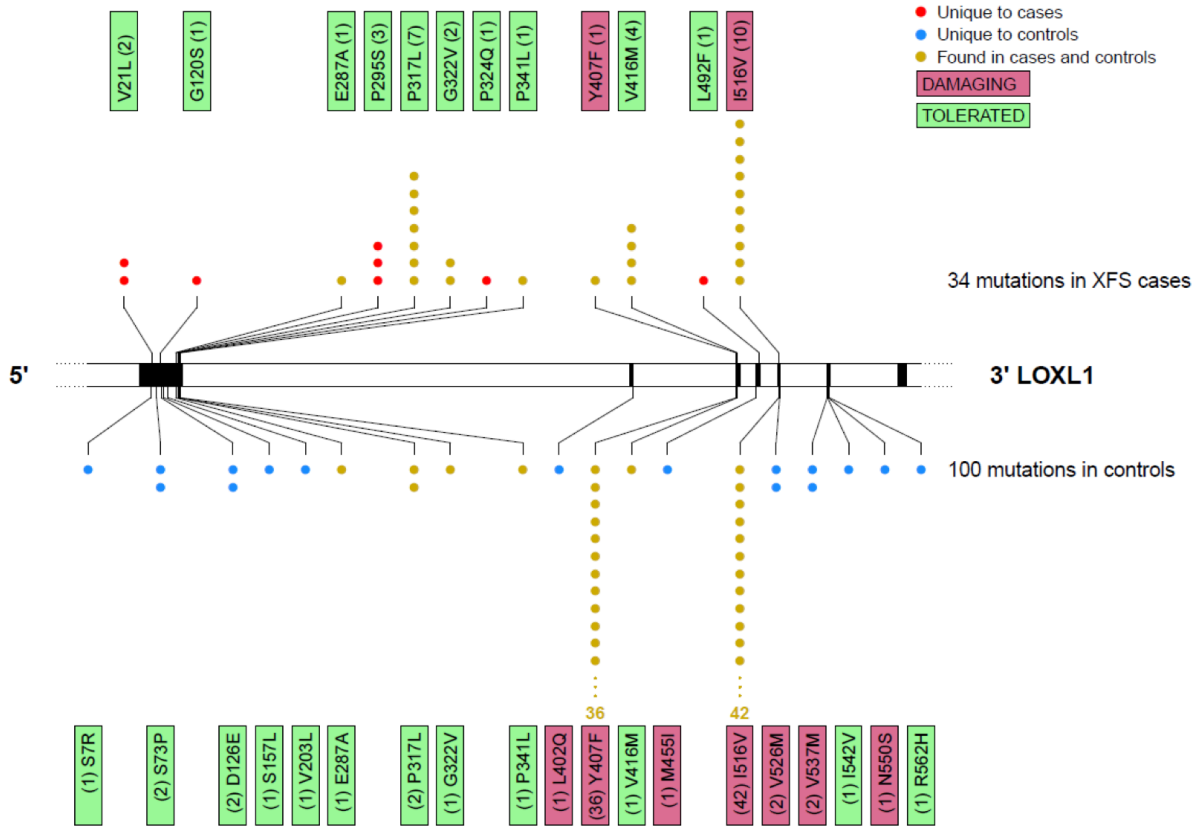
Aung, T. *et al.* Genetic association study of exfoliation syndrome identifies a protective rare variant at LOXL1 and five new susceptibility loci. *Nat. Genet.* **49**, 993-1004, doi:10.1038/ng.3875 (2017).

Supplementary Figure 4

Analysis of non-synonymous variant burden from the *LOXL1* resequencing exercise.

a) *LOXL1* non-synonymous mutations detected after sequencing 2,827 XFS and glaucoma cases and 3,014 controls from Japan. Mutations are labelled as 'DAMAGING' if they were predicted to be damaging or deleterious by all five protein prediction soft wares (SIFT, Polyphen2-HumDiv, LRT score, MutationTaster, and Condel). Otherwise, they are labelled as 'TOLERATED'.

The number of individuals carrying a particular mutation is given in parenthesis next to the annotation.



Basu, U. *et al.* ABC transporter-mediated transport of glutathione conjugates enhances seed yield and quality in chickpea. *Plant Physiol.* **180**, 253-275, doi:10.1104/pp.18.00934 (2019).

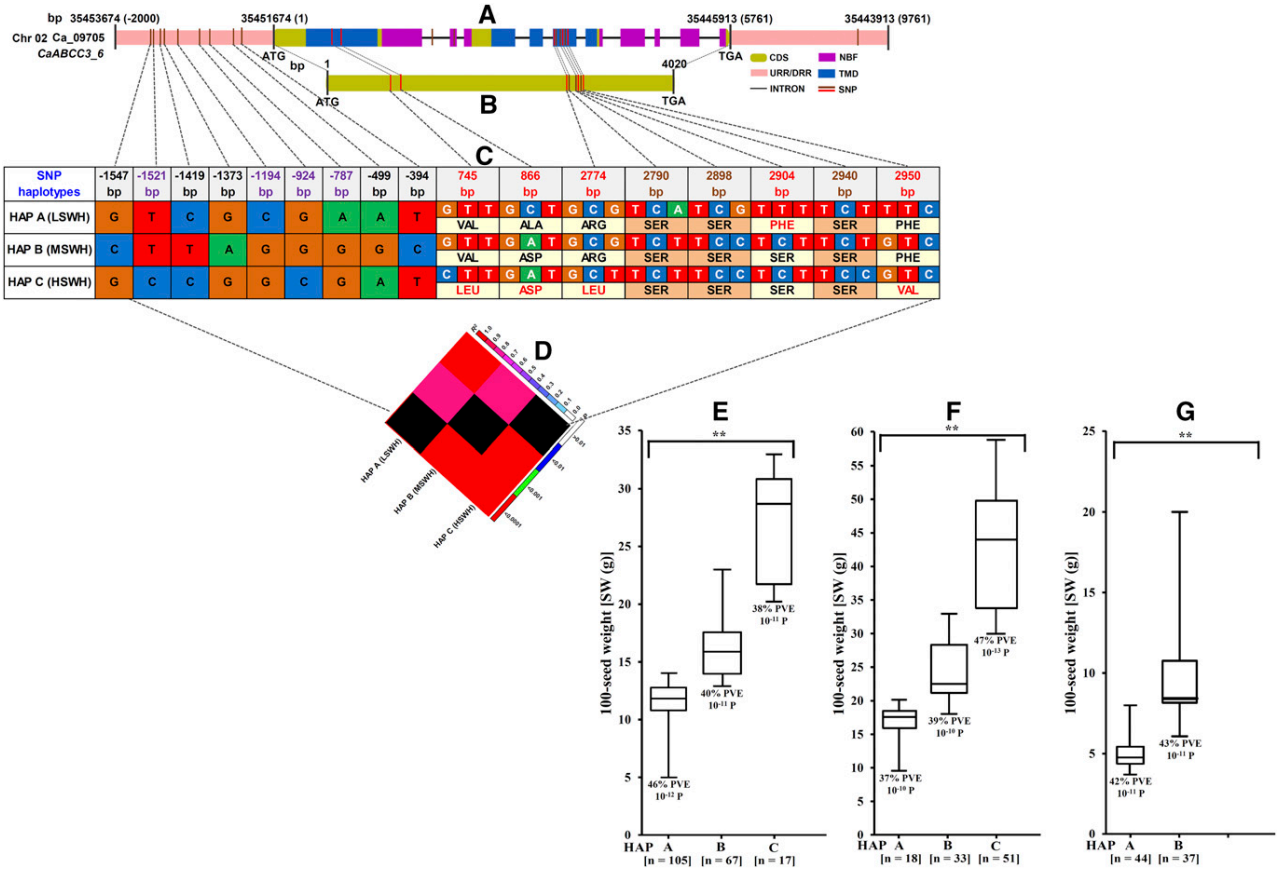


Figure 7. Haplotype-specific LD and association mapping in a strongly SW-associated gene, *CaABCC3*(*6*), delineated by GWAS, gene-by-gene regional association analysis and map-based cloning. Genomic organization/constitution of the *CaABCC3*(*6*) gene (A) including its (B) CDS, exhibiting the distribution of SNPs in different sequence components of this gene. C, The genotyping of 20 SNPs (A and B) in different coding and noncoding sequence components of *CaABCC3*(*6*) in all 291 cultivated (*desi* and *kabuli*) and 81 wild chickpea accessions constituted three haplotypes (D). Three haplotypes, HAP A, HAP B, and HAP C, exhibited strong association with low, medium, and high SW, respectively. The nonsynonymous and regulatory SNPs exhibiting differentiation, especially between LSWH (HAP A) and HSWH (HAP C), are highlighted in red and violet, respectively. The value r^2 indicates the frequency correlation between pairs of alleles across a pair of SNP loci. Boxplots for 100-SW based on three haplotypes, HAP A, HAP B, and HAP C, constituted in (E) *desi* (189 accessions), (F) *kabuli* (102), and (G) wild (81) chickpea, demonstrating their strong associations with low, medium, and high SW, respectively. Box edges represent the upper and lower quantiles, with the median value in the middle of the box. The digits within the square brackets denote the number of accessions representing each class of haplotype associated with SW. $*P < 0.0001$, two-sided Wilcoxon test. HAP, haplotype; LSWH/MSWH/HSWH, low-/medium-/high-SW haplotype.

Basu, U. *et al.* CLAVATA signaling pathway genes modulating flowering time and flower number in chickpea. *Theor. Appl. Genet.* **132**, 2017-2038, doi:10.1007/s00122-019-03335-y (2019).

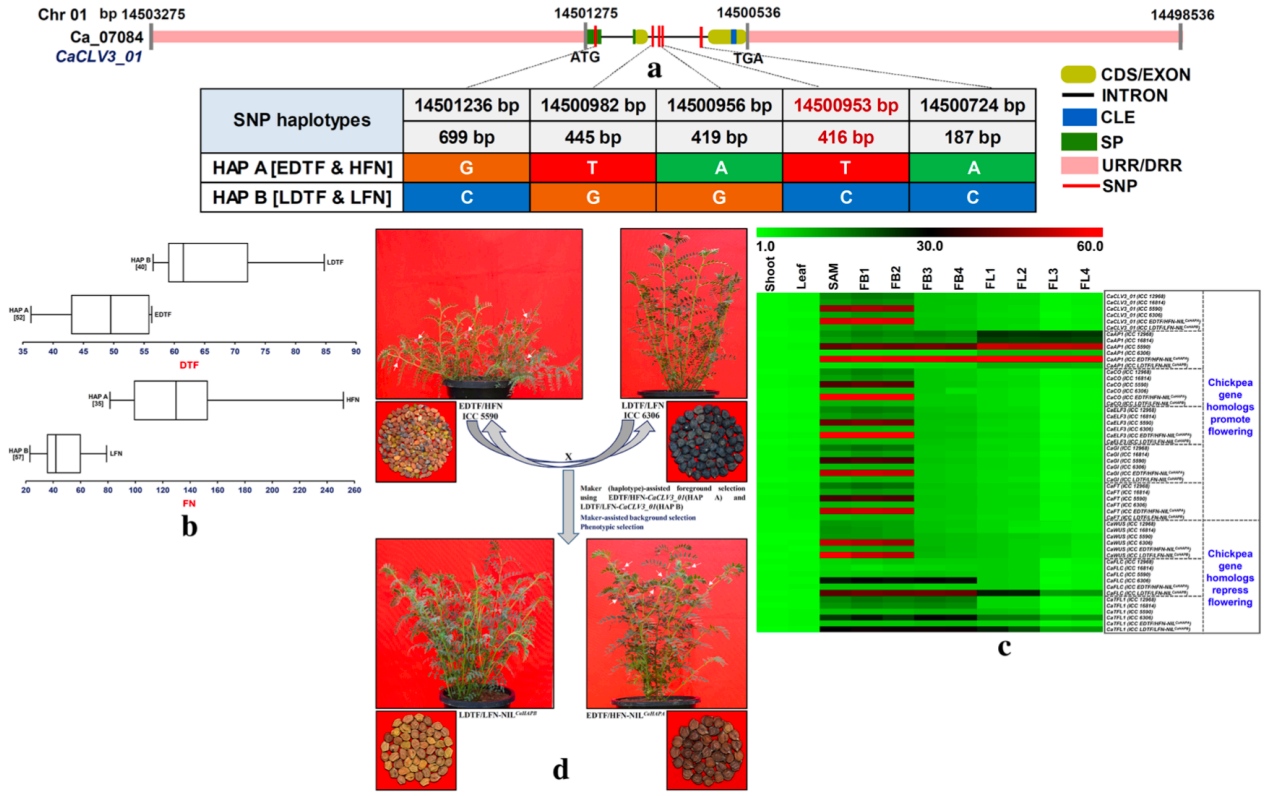


Fig. 7 **a** The genotyping of five SNPs of *CaCLV3_01* among association panel (92 chickpea accessions) constituted two major haplotypes, HAP A (EDTF and HFN) and HAP B (LDTF and LFN). **b** Boxplots ICC 6306) contrasting with DTF and FN traits. The average log signal expression value of genes is represented at the top with a color scale, in which green, black and red colors denote low, medium and

Basu, U. *et al.* Genetic dissection of photosynthetic efficiency traits for enhancing seed yield in chickpea. *Plant Cell Environ.* **42**, 158-173, doi:10.1111/pce.13319 (2019).

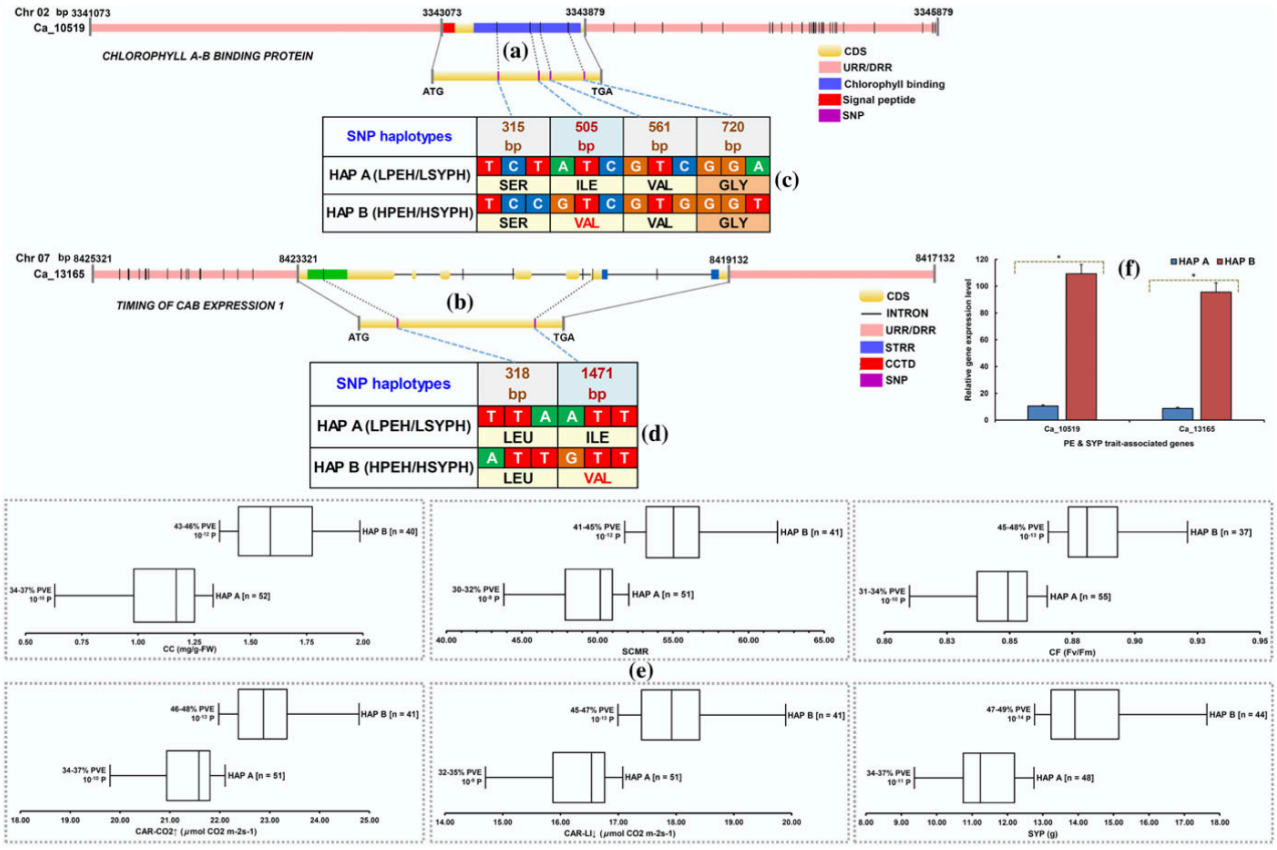


FIGURE 5 Constitution of haplotypes and their association mapping and expression profiling of a strong photosynthetic efficiency (PE) and seed yield per plant (SYP)-associated chlorophyll A-B binding protein-coding gene and its interacting gene, Timing of CAB Expression 1 (delineated by association analysis, QTL mapping, and expression profiling), validating potential of the gene haplotypes in regulating PE and SYP traits in chickpea. Genomic organization and constitution of a (a) chlorophyll A-B binding protein-coding gene and its interacting gene, (b) Timing of CAB Expression 1 exhibiting distribution of SNPs in different sequence components of these genes. (c,d) The genotyping of SNPs in different coding and noncoding sequence components of these two genes among 92 *desi* and *kabuli* cultivated chickpea accessions constituted two major haplotypes from each gene. (e) Two haplotypes, HAP A and HAP B, represented by the *desi* and *kabuli* accessions (*n*) demonstrating strong association with low and high PE and SYP trait differentiation, respectively, are illustrated by the Box-Whisker Plots. (f) Haplotype-specific transcript profiling of two haplotypes constituted from chlorophyll A-B binding protein-coding gene (*Ca_10519*) and Timing of CAB Expression 1 (*Ca_13165*) gene using the young/mature leaf tissues of the two selected chickpea accessions representing low (HAP A) and high (HAP B) PE and SYP haplotypes. Error bars represent standard error ($n = 3$). (*$p < .0001$, two-tailed *t* test). URR = upstream regulatory region; DRR = downstream regulatory region; SNP = single nucleotide polymorphism; CC = chlorophyll content; SCMR = SPAD chlorophyll meter reading; CAR↓LI = CO_2 assimilation rate at decreasing light intensity; CAR↑CO_2 = assimilation rate at increasing CO_2 concentration; STRR = signal transduction response regulator; CCTD = CCT (CONSTANS, CONSTANS-like and TOC1) domain

Chao, M. J. *et al.* Haplotype-based stratification of huntington's disease. *Eur. J. Hum. Genet.* **25**, 1202-1209, doi:10.1038/ejhg.2017.125 (2017).

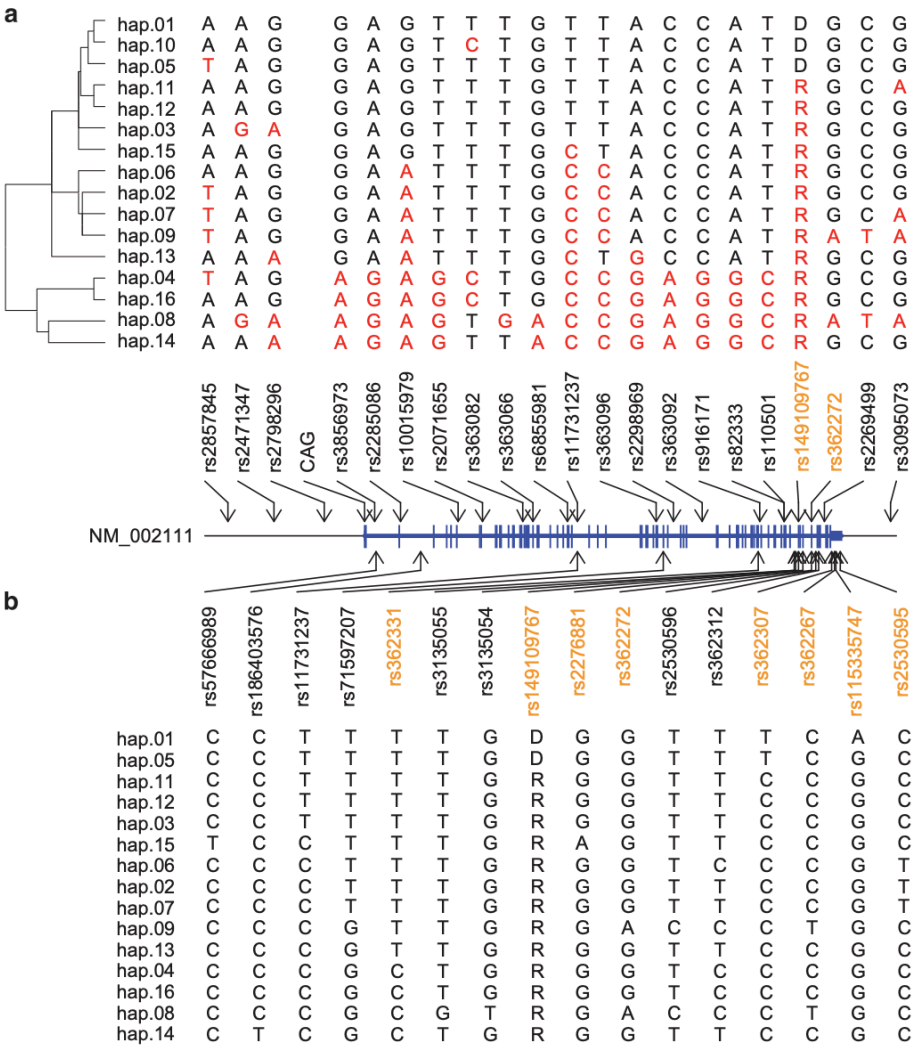


Figure 1 Definitions and sequence relationships of *HTT* haplotypes. (**a**) Twenty SNPs, one 3 bp indel (rs149109767, alleles R-reference and D-deletion) and the CAG repeat polymorphism are shown at their genomic locations relative to that of the *HTT* RefSeq transcript (NM_002111). Genotype at each marker on each of 16 *HTT* haplotypes, defined in the text, is shown above the marker. Haplotypes are ordered based upon a neighbor-joining method (p-distance model) in a dendrogram with two main branches, each with different sizes of sub-clusters. Alleles in red represent differences from hap.01, the most frequent haplotype on CAG-expanded HD chromosomes. (**b**) Consensus alleles of 10 exon SNPs and 10 intron SNPs that showed the biggest cumulative heterozygosity were determined for each haplotype based on 1000 Genomes Project data. A consensus allele for a given SNP site represents the most frequent allele among a collection of chromosomes with same haplotype. Since hap.10 is not present in 1000 Genomes data (Phase 1), hap.10 was excluded in this analysis. Subsequently, alleles of SNPs that show variable alleles in 15 haplotypes and alleles of two exon SNPs that were used to define the haplotypes are indicated. SNPs in orange and black font colors represent SNPs on exons and introns of RefSeq NM_002111, respectively.

Hou, J. *et al*. ADP-glucose pyrophosphorylase genes, associated with kernel weight, underwent selection during wheat domestication and breeding. *Plant Biotechnol. J.* **15**, 1533-1543, doi:10.1111/pbi.12735 (2017).

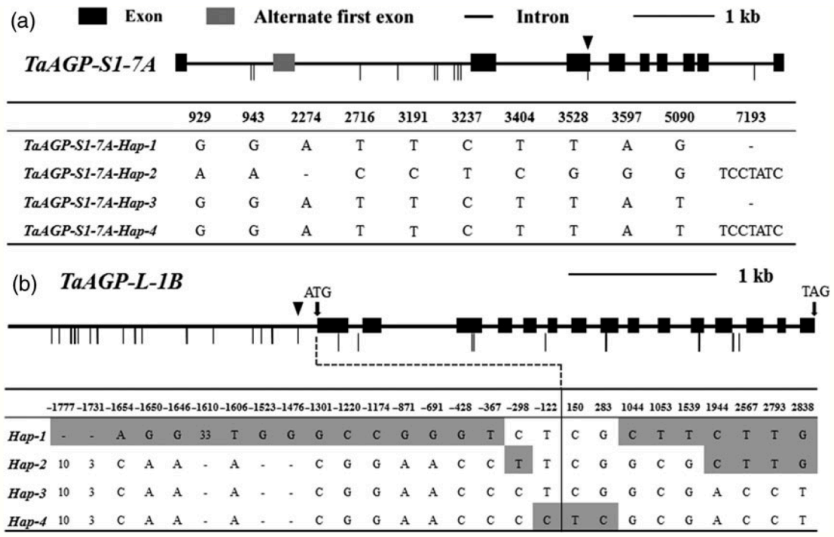
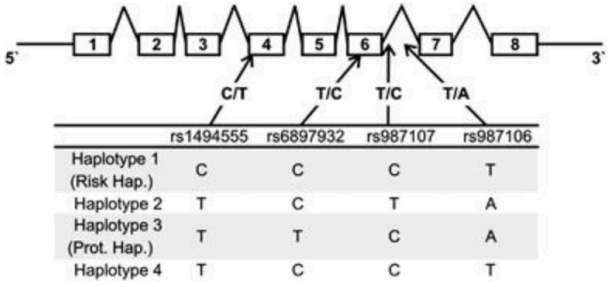


Figure 1 Haplotypes of *TaAGP-S1-7A* and *TaAGP-L-1B*. (a) Coding regions of *TaAGP-S1-7A*. ▼SNP at position 5090. (b) Coding and 2-kb upstream regions, and polymorphic sites of *TaAGP-L-1B*. ▼SNP at position -122. Numbers indicate deletion size (bp). Vertical thin lines indicate polymorphic site differences between haplotypes.

Jäger, J., Schulze, C., Rösner, S. & Martin, R. IL7RA haplotype-associated alterations in cellular immune function and gene expression patterns in multiple sclerosis. *Genes Immun.* **14**, 453-461, doi:10.1038/gene.2013.40 (2013).

Figure 1



	rs1494555	rs6897932	rs987107	rs987106
Haplotype 1 (Risk Hap.)	C	C	C	T
Haplotype 2	T	C	T	A
Haplotype 3 (Prot. Hap.)	T	T	C	A
Haplotype 4	T	C	C	T

Schematic diagram of the *IL7RA* sequence showing the four SNPs used to stratify the Hamburg cohort into four common haplotypes.

Jain, S. *et al.* A haplotype of human angiotensinogen gene containing -217A increases blood pressure in transgenic mice compared with -217G. *Am. J. Physiol. Regul. Integr. Comp. Physiol*. **295**:R1849-R1857, doi: 10.1152/ajpregu.90637 (2008).

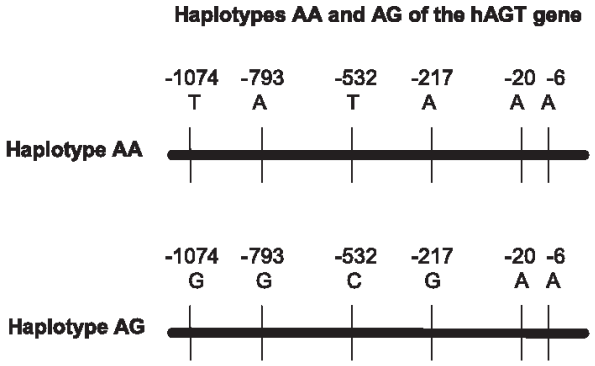
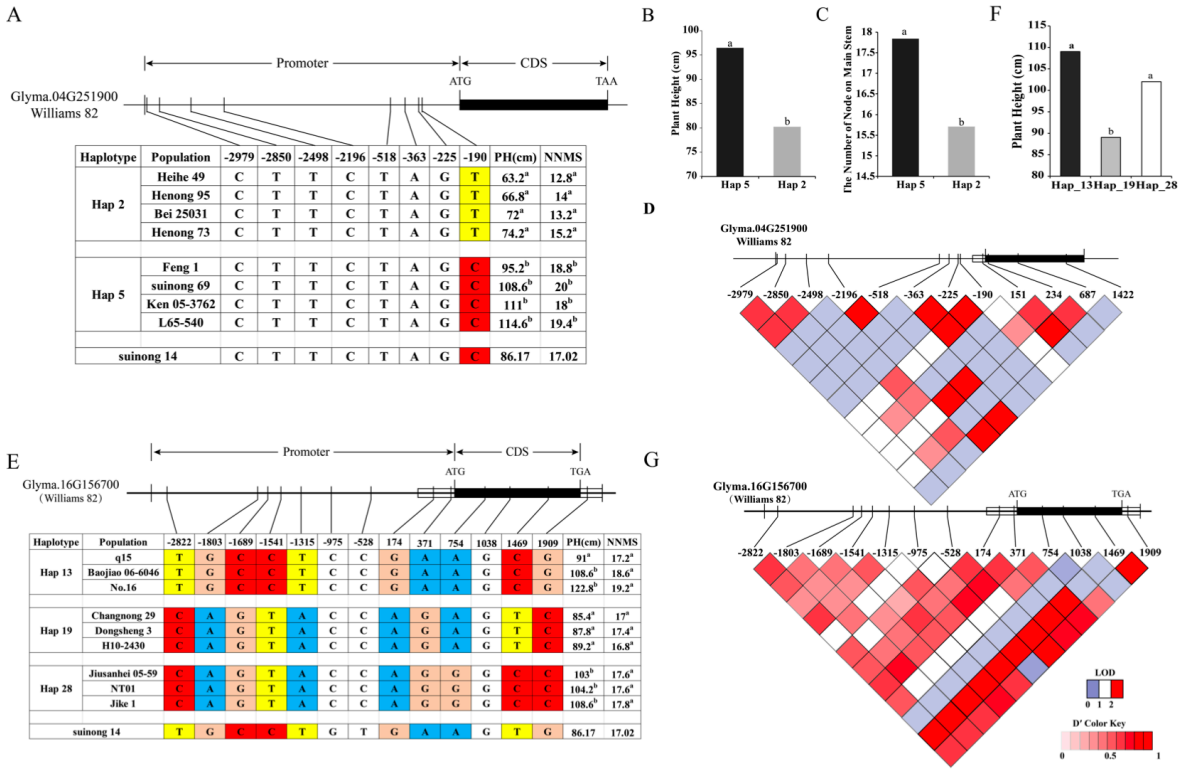


Fig. 1. Different haplotypes of the human angiotensinogen (hAGT) gene. The

Li, R. *et al.* Combined linkage mapping and bsa to identify QTL and candidate genes for plant height and the number of nodes on the main stem in soybean. *Int. J. Mol. Sci.* **21**, 42 (2019).



Haplotype	Population	-2979	-2850	-2498	-2196	-518	-363	-225	-190	PH(cm)	NNMS
Hap 2	Heihe 49	C	T	T	C	T	A	G	T	63.2[a]	12.8[a]
	Henong 95	C	T	T	C	T	A	G	T	66.8[a]	14[a]
	Bei 25031	C	T	T	C	T	A	G	T	72[a]	13.2[a]
	Henong 73	C	T	T	C	T	A	G	T	74.2[a]	15.2[a]
Hap 5	Feng 1	C	T	T	C	T	A	G	C	95.2[b]	18.8[b]
	suinong 69	C	T	T	C	T	A	G	C	108.6[b]	20[b]
	Ken 05-3762	C	T	T	C	T	A	G	C	111[b]	18[b]
	L65-540	C	T	T	C	T	A	G	C	114.6[b]	19.4[b]
suinong 14		C	T	T	C	T	A	G	C	86.17	17.02

Haplotype	Population	-2822	-1803	-1689	-1541	-1315	-975	-528	174	371	754	1038	1469	1909	PH(cm)	NNMS
Hap 13	q15	T	G	C	C	T	C	C	G	A	A	G	C	G	91[a]	17.2[a]
	Baojiao 06-6046	T	G	C	C	T	C	C	G	A	A	G	C	G	108.6[b]	18.6[a]
	No.16	T	G	C	C	T	C	C	G	A	A	G	C	G	122.8[b]	19.2[a]
Hap 19	Changnong 29	C	A	G	T	A	C	C	A	G	A	G	T	C	85.4[a]	17[a]
	Dongsheng 3	C	A	G	T	A	C	C	A	G	A	G	T	C	87.8[a]	17.4[a]
	H10-2430	C	A	G	T	A	C	C	A	G	A	G	T	C	89.2[a]	16.8[a]
Hap 28	Jiusanhei 05-59	C	A	G	T	A	C	C	A	G	G	G	C	C	103[b]	17.6[a]
	NT01	C	A	G	T	A	C	C	A	G	G	G	C	C	104.2[b]	17.6[a]
	Jike 1	C	A	G	T	A	C	C	A	G	G	G	C	C	108.6[b]	17.8[a]
suinong 14		T	G	C	C	T	G	T	G	A	A	G	T	G	86.17	17.02

Figure 5. Haplotype analysis of the candidate gene. (**A**) Haplotype analysis of *Glyma.04G251900* from 92 soybean resource. (**B**) PH of Hap-2 and Hap-5. (**C**) NNMS of Hap-2 and Hap-5. (**D**) linkage disequilibrium (LD) analysis of SNPs located on *Glyma.04G251900*. Red from light to dark represents the degree of linkage between SNPs. (**E**) Haplotype analysis of *Glyma.16G156700* from 92 soybean resource. (**F**) PH of Hap-13, Hap-19, and Hap-28. (**G**) LD analysis of SNPs located on *Glyma.16G156700*. [a], [b]: Different letters represent significant differences between each other at the 0.05 level.

Li, Y. *et al.* Chalk5 encodes a vacuolar H+-translocating pyrophosphatase influencing grain chalkiness in rice. *Nat. Genet.* **46**, 398-404, doi:10.1038/ng.2923 (2014).

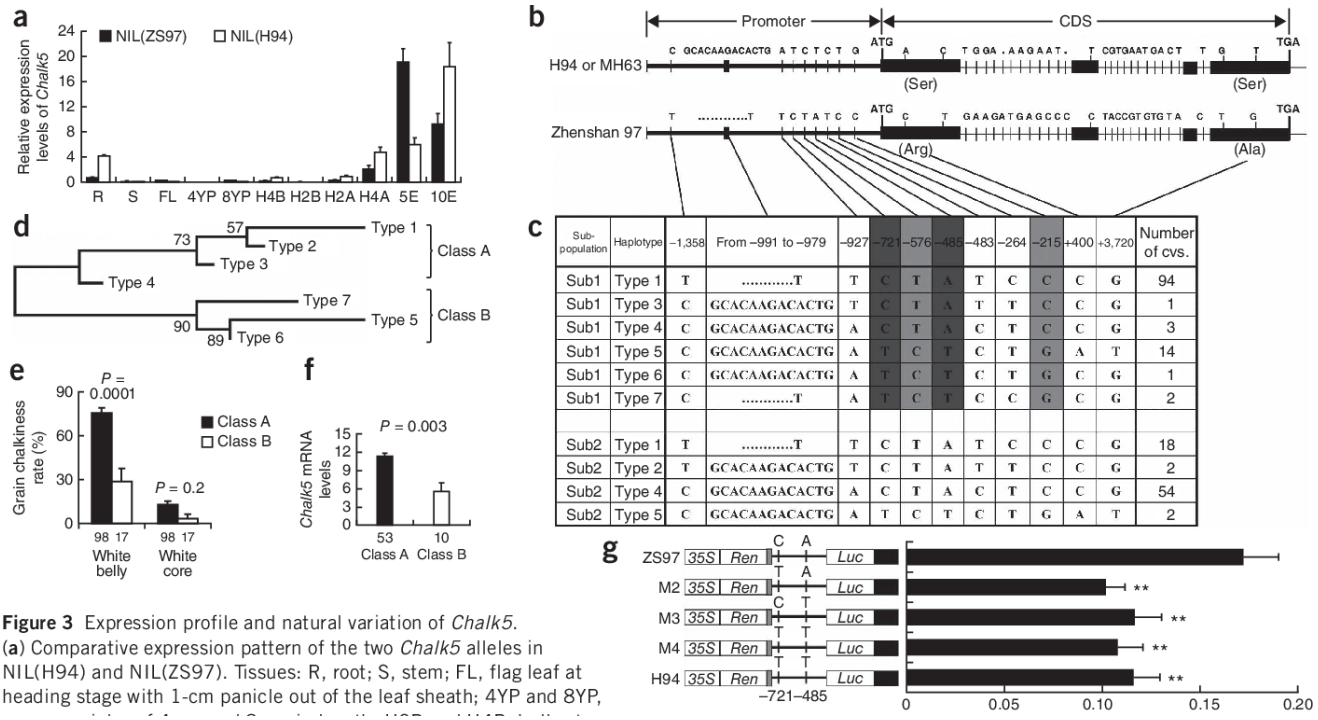


Figure 3 Expression profile and natural variation of *Chalk5*. (**a**) Comparative expression pattern of the two *Chalk5* alleles in NIL(H94) and NIL(ZS97). Tissues: R, root; S, stem; FL, flag leaf at heading stage with 1-cm panicle out of the leaf sheath; 4YP and 8YP,

Malik, N. *et al.* An integrated genomic strategy delineates candidate mediator genes regulating grain size and weight in rice. *Sci. Rep.* **6**, 23253, doi:10.1038/srep23253 (2016).

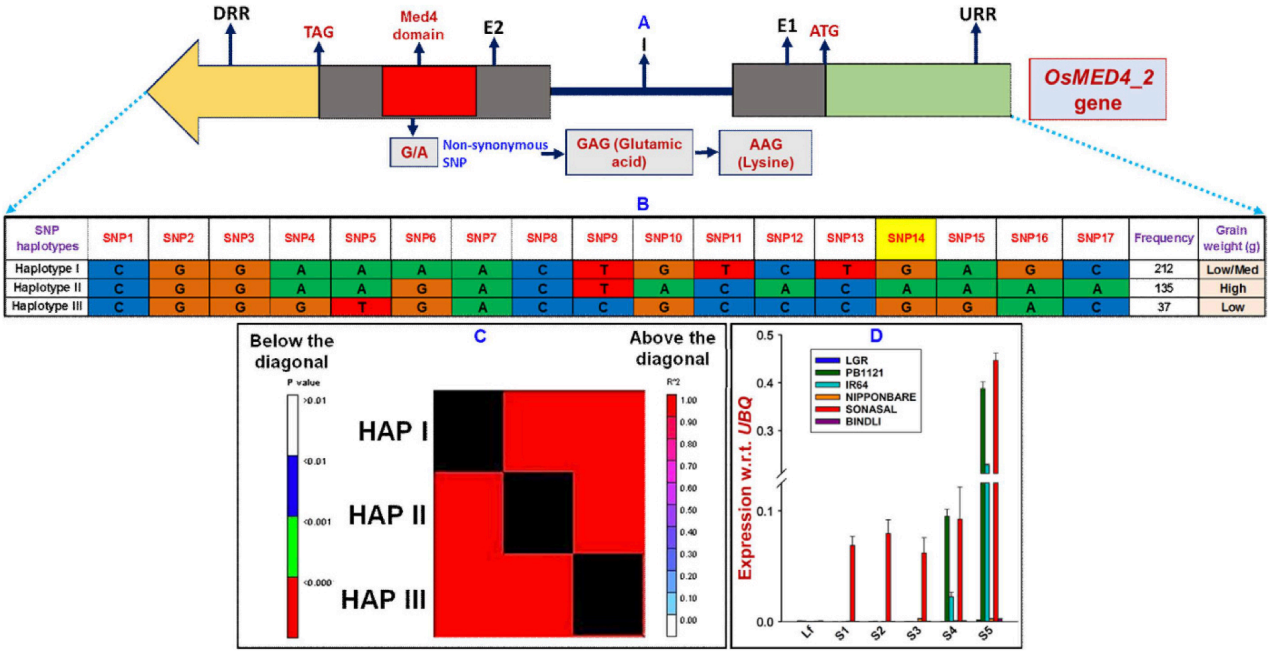


Figure 5. The molecular haplotyping and SNP haplotype-specific association analysis/LD mapping in an *OsMED4_2* gene (A) validating its strong association potential for grain weight/size differentiation in rice. The genotyping of 17 SNPs, including one missense non-synonymous SNP (G/A, shaded with yellow colour) [encoding for Glutamic acid (GAG) to Lysine (AAG)] among 384 rice accessions (association panel) constituted three haplotypes (B). (C) Three SNP haplotype marker-based genotyping information produced a higher LD estimate and resolution covering the entire gene. (D) The differential expression profiling of *OsMED4_2* gene in five seed developmental stages (S1–S5) and flag leaves (Lf) of six contrasting low (Sonasal and Bindli) and high (LGR, PusaBasmati 1121, Nipponbare and IR 64) grain weight rice accessions. E1: Exon1, E2: Exon2, I: Intron, URR: upstream regulatory region and DRR: downstream regulatory region.

Mao, D. *et al.* Natural variation in the HAN1 gene confers chilling tolerance in rice and allowed adaptation to a temperate climate. *Proc. Natl. Acad. Sci. USA.* **116**, 3494-3501, doi:10.1073/pnas.1819769116 (2019).

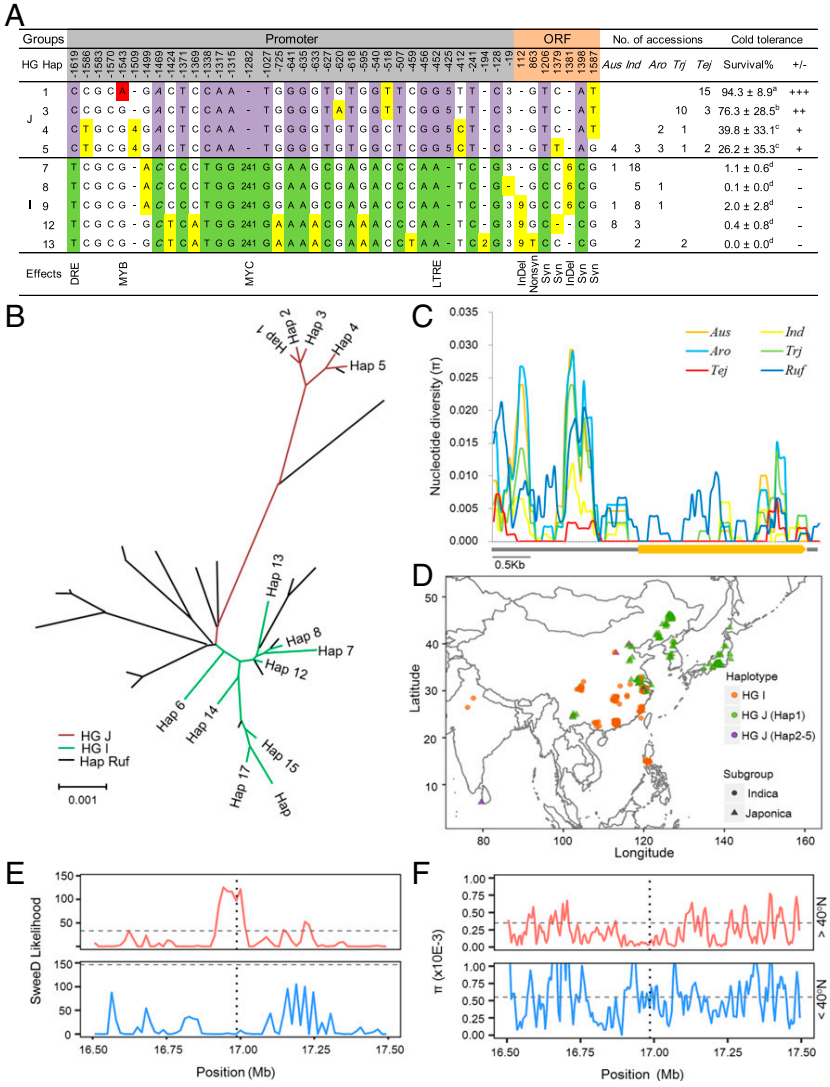
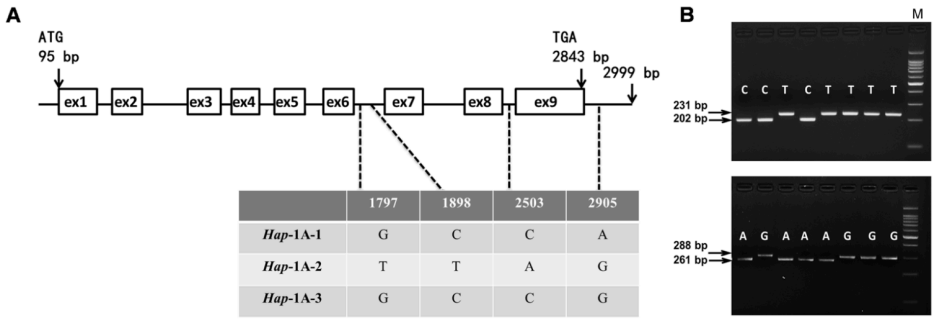


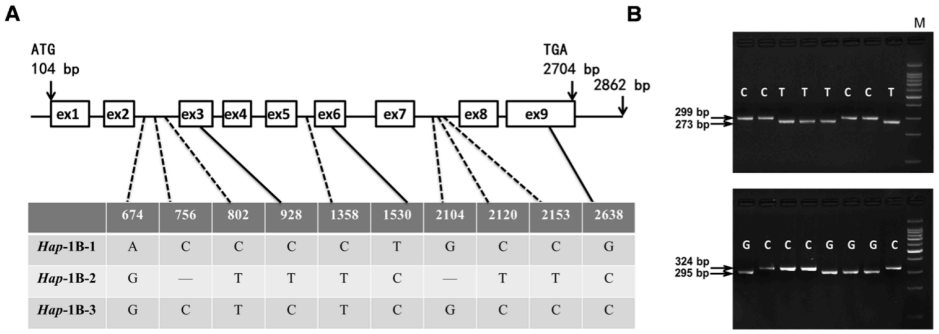
Fig. 5. The haplotypes and evolutionary analysis of *HAN1*. (*A*) Haplotype analysis of *HAN1* in 101 rice germplasm accessions. The purple and green shaded SNPs or InDels represent polymorphisms in the *japonica* and *indica* types, respectively. The highlighted red A is found only in Hap1 in temperate *japonica* rice, while yellow polymorphisms are not unique among the rice cultivars (There were significant differences among different letters by Student's *t* test with $P < 0.01$ and $n \geq 3$). (*B*) A phylogenetic tree showing the relationships between *HAN1* haplotypes from cultivated rice and *O. rufipogon* accessions. (*C*) Nucleotide diversity of *HAN1* in the five cultivated rice subgroups of *O. sativa* and *O. rufipogon* accessions. The thick bar represents CDS region of *HAN1*, and the thin bars before or after the CDS are the 1.8-kb promoter or the 0.5-kb downstream region of *HAN1*, respectively. Aus, Ind, Aro, Trj, Tej, and Ruf represent the five ecotypes of cultivated rice, such as *aus*, *indica*, *aromatic* rice, tropical *japonica* and temperate *japonica* rice, and wild rice (*O. rufipogon*). (*D*) The geographical distribution of cultivated rice with different *HAN1* haplotypes. There are two groups, the *indica* group HG I and the *japonica* group HG J. HG J is further divided into two subgroups HG J (Hap 1) and HG J (Haps 2–5) based on presence or absence of the MYB cis element. (*E*) The likelihood of a selective sweep of the 1.0-Mb region surrounding *HAN1* in the rice population cultivated in typical temperate climate areas in high latitude regions (>40° N), or elsewhere in low latitude regions (<40° N). (*F*) Nucleotide diversity in the 1.0-Mb region flanking *HAN1* in the two populations in *E*. The analyses performed in *A*, *B*, and *C* were based on DNA sequence information generated by Sanger sequencing in a population of 101 rice germplasm accessions, while the analyses in *D*, *E*, and *F* were based on NGS data and geographic information in a collection of 572 rice accessions.

Miao, L. *et al.* Elite haplotypes of a protein kinase gene TaSnRK2.3 associated with important agronomic traits in common wheat. *Front. Plant Sci.* **8**, doi:10.3389/fpls.2017.00368 (2017).



	1797	1898	2503	2905
Hap-1A-1	G	C	C	A
Hap-1A-2	T	T	A	G
Hap-1A-3	G	C	C	G

FIGURE 2 | Nucleotide polymorphisms and functional marker development of *TaSnRK2.3-1A*. (A) Four single nucleotide polymorphisms (SNPs) were detected in the non-exon regions of *TaSnRK2.3-1A*. **(B)** A dCAPS marker 2.3AM1 was developed based on SNP-1898 (C/T) in the upper segment. Digestion of the amplified 231 bp fragment with *Hha*I produced fragments of 202 bp/29 bp for accessions with SNP-1898C, yet unacted for accessions with SNP-1898T. Similarly, a CAPS marker 2.3AM2 from SNP-2905 (A/G) was also developed in the lower segment. M, 100 bp DNA Ladder.



	674	756	802	928	1358	1530	2104	2120	2153	2638
Hap-1B-1	A	C	C	C	C	T	G	C	C	G
Hap-1B-2	G	—	T	T	T	C	—	T	T	C
Hap-1B-3	G	C	T	C	T	C	G	C	C	C

FIGURE 3 | Nucleotide polymorphisms and functional marker development of *TaSnRK2.3-1B*. (A) Eight SNPs and two InDels were detected in the coding region of *TaSnRK2.3-1B*. "—" indicates deletion. **(B)** CAPS marker 2.3BM1 was developed based on SNP-2153(C/T). Digestion of the amplified 299 bp fragment with *N1a*III produced fragments of 273 bp/26 bp for accessions with SNP-2153T, and a single 299-bp band for accessions with SNP-2153C. Similarly, a dCAPS marker 2.3BM2 for SNP-2638(C/G) was developed. M, 100 bp DNA Ladder.

Mopidevi, B. *et al.* A polymorphism in intron I of the human angiotensinogen gene (hAGT) affects binding by HNF3 and hAGT expression and increases blood pressure in mice. *J. Biol. Chem.* **294**, 11829-11839, doi:10.1074/jbc.RA119.007715 (2019).

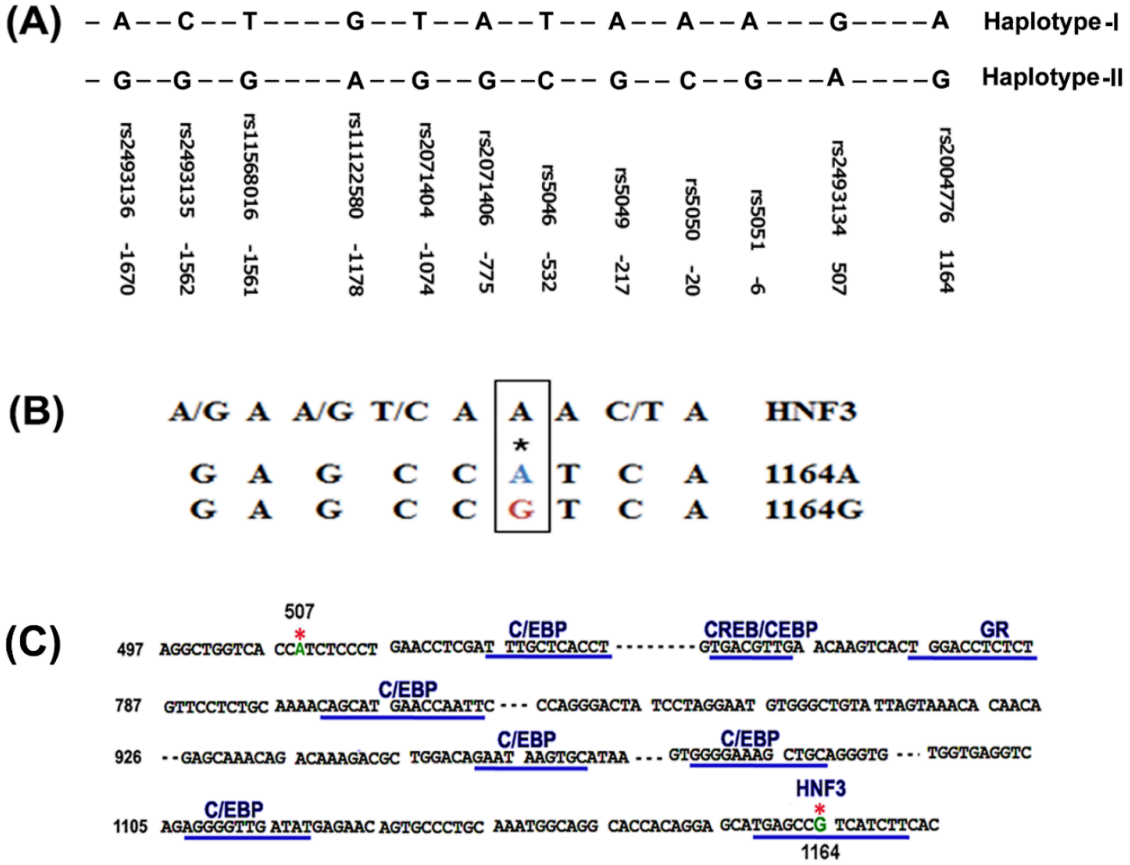


Figure-2: **(A)** The Nucleotide sequence of SNPs present in Hap-I and Hap-II of hAGT gene. **(B)** Homology between consensus HNF-3 binding site and nucleotide sequence containing +1164A and +1164G in intron-I of the hAGT gene. **(C)** In silico analysis of transcription factor binding sites in intron I of the hAGT gene; polymorphic sites at +507 and +1164 are marked by asterisk.

Sato, K. *et al.* Strong evidence of a combination polymorphism of the tyrosine kinase 2 gene and the signal transducer and activator of transcription 3 gene as a DNA-based biomarker for susceptibility to crohn's disease in the japanese population. *J. Clin. Immunol.* **29**, 815-825, doi:10.1007/s10875-009-9320-x (2009).

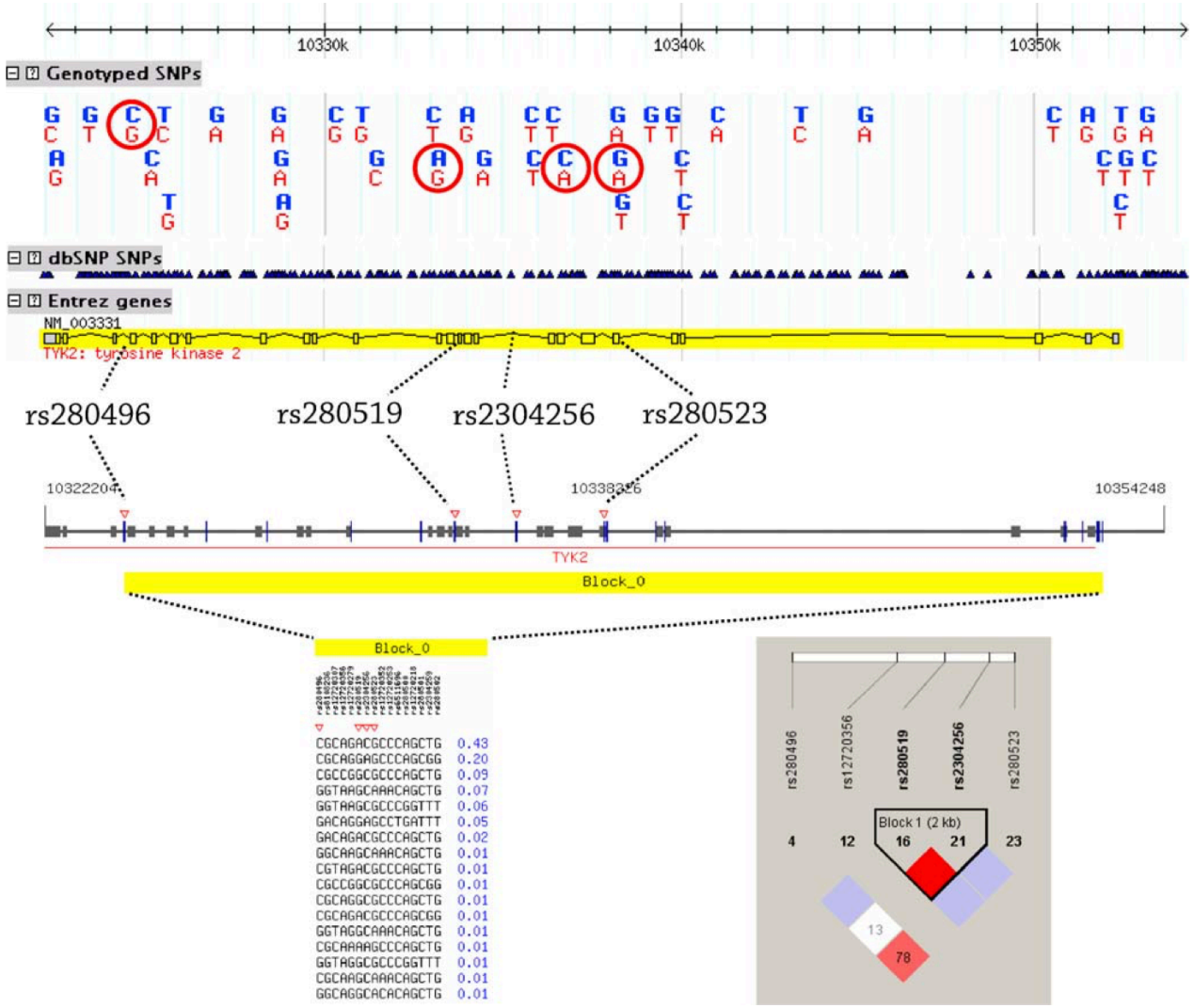


Fig. 2 Locations of the genotyped tag SNP sites in *TYK2* in the International HapMap (*upper*) and iHap (*lower*) data. The *horizontal bars in the middle* indicate the genomic sequence of *TYK2*. *Blue vertical bars* indicate the positions of all SNP sites. A *yellow rectangle* represents the positions of a linkage disequilibrium block. A list and the locations of candidate tag SNPs are shown in the *lower right position* using Haploview 4.0 software. *Red inverted triangles* indicate the genotyped tag SNPs sites in this study, and their names are presented *above each inverted triangle*

Tu, Y., Shu, J., Ji, G., Zhang, M. & Zou, J. Monitoring conservation effects on a Chinese indigenous chicken breed using major histocompatibility complex B-G gene and DNA Barcodes. *Asian-Australas J. Anim. Sci.* **31**, 1558-1564, doi:10.5713/ajas.17.0627 (2018).

```
MATRIX
                11
           112524
          9155632
          3465469
Hap_1     CTTGCGC
Hap_2     ...A..T
Hap_3     .CC....
Hap_4     ...ATAT
Hap_5     G..A..T
```

Figure 1. Variable sites in cytochrome oxidase I gene of haplotype in generations 0, 5, 10, 15, 16, 17 of Langshan conservation population. -, indicate the same base.

```
Hap#  Freq. Sequences
Hap_1:  3    1 21 31
Hap_2:  164   2 4-20 22-24 26-30 35-52 54 56-86 90-97 99-102 104-116 118-122,123-150,158-187
Hap_3:  4    3 33-34 53
Hap_4:  9    25 32 87 98 151-153 188 189
Hap_5:  10    55 88-89 103 117 154-157 190
```

Figure 2. Haplotype frequency sequences of cytochrome oxidase I gene in generations 0, 5, 10, 15, 16, 17 of Langshan conservation population.

Wang, M. *et al.* Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.* **50**, 1435-1441, doi:10.1038/s41588-018-0229-2 (2018).

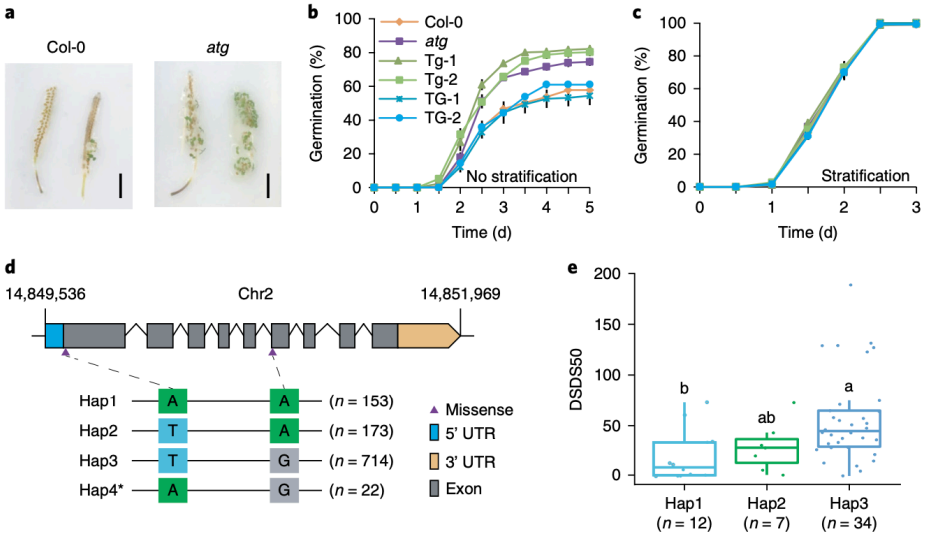


Fig. 4 | *AtG* controls seed dormancy in *Arabidopsis*. **a**, Seed germination of Col-0 (*AtG*) and *atg*. Photographs were taken two weeks after imbibition. Experiment repeated three times with similar results. Scale bar, 3.0 mm. **b**, Germination phenotype of freshly harvested seeds of Col-0, *atg*, and transgenic lines without stratification. **c**, Germination phenotype of freshly harvested seeds of Col-0, *atg*, and transgenic lines after 3 d of stratification at 4 °C. The transgenic lines are from soybean *G* CDSs driven by the promoter of *AtG*. For TG-1 and TG-2, the CDS is from Kuaiqingpi (*GmG*), and for Tg-1 and Tg-2, the CDS is from DN50 (*Gmg*). For *Arabidopsis* seed germination assays, means ± s.e.m. are shown for $n = 5$ independent experiments. Each experiment consists of about 50 seeds. **d**, Haplotype analysis of *AtG* in the published sequence of 1,062 *Arabidopsis* accessions[35]. Asterisk indicates haplotype without dormancy data. Sample number *n* for each haplotype is shown in the figure. **e**, Dormancy behavior values (DSDS50) of different *AtG* haplotypes according to published dormancy data in *Arabidopsis*[36]. Box edges depict interquartile range, whiskers 1.5 × the interquartile range, and center lines the median. Sample number *n* for each haplotype is shown under the boxes. The significance was calculated by one-way analysis of variance with Tukey's multiple comparisons test, $\alpha < 0.05$. Different letters indicate distinct groups.

Wang, Q. *et al.* Genetic variations in ARE1 mediate grain yield by modulating nitrogen utilization in rice. *Nat. Commun.* **9**, 735, doi:10.1038/s41467-017-02781-w (2018).

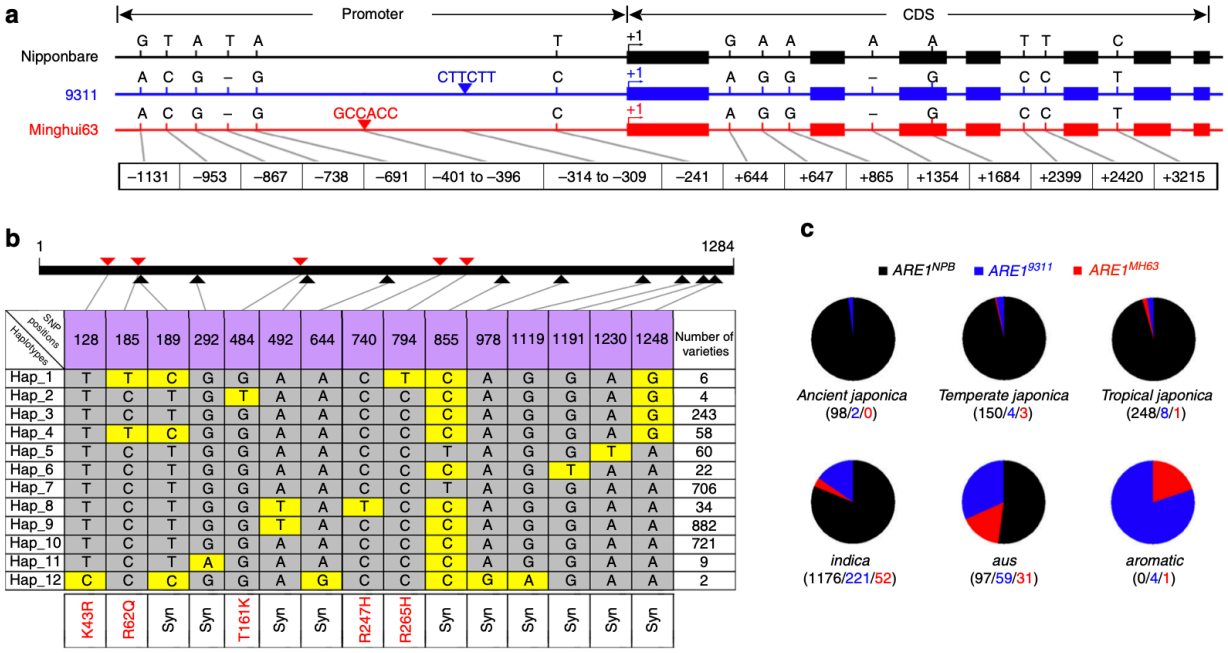


Fig. 5 Analysis of genetic variations in *ARE1*. **a** Schematic representation of genetic variations in *ARE1* in a *japonica* variety Nipponbare (NPB) and two *indica* varieties 9311 and Minghui63 (MH63). Exons are shown by filled boxes and other sequences are shown by lines. Numbers at the bottom indicate positions of variations (the putative transcription start is referred to as +1). CDS, coding sequences. **b** Major haplotypes of single nucleotide polymorphisms (SNPs) in the *ARE1* coding region. Major SNP haplotypes and casual variations in the encoded amino acid residues are shown. The *ARE1* coding sequences of 2747 rice varieties were compared with that of NPB (Hap_7). Twelve haplotypes were identified from these accessions and polymorphic nucleotides of each haplotype are highlighted by yellow boxes. The numbers of the identified varieties of each haplotype are shown at right. Syn, synonymous variations. **c** Distribution of three haplotypes of insertion-deletion polymorphisms (InDels) in the *ARE1* promoter in various accessions. The numbers of the detected haplotypes (specified by different colors) are given below each group

Webster, M. T. *et al.* Linked genetic variants on chromosome 10 control ear morphology and body mass among dog breeds. *BMC Genomics* **16**, 474, doi:10.1186/s12864-015-1702-2 (2015).

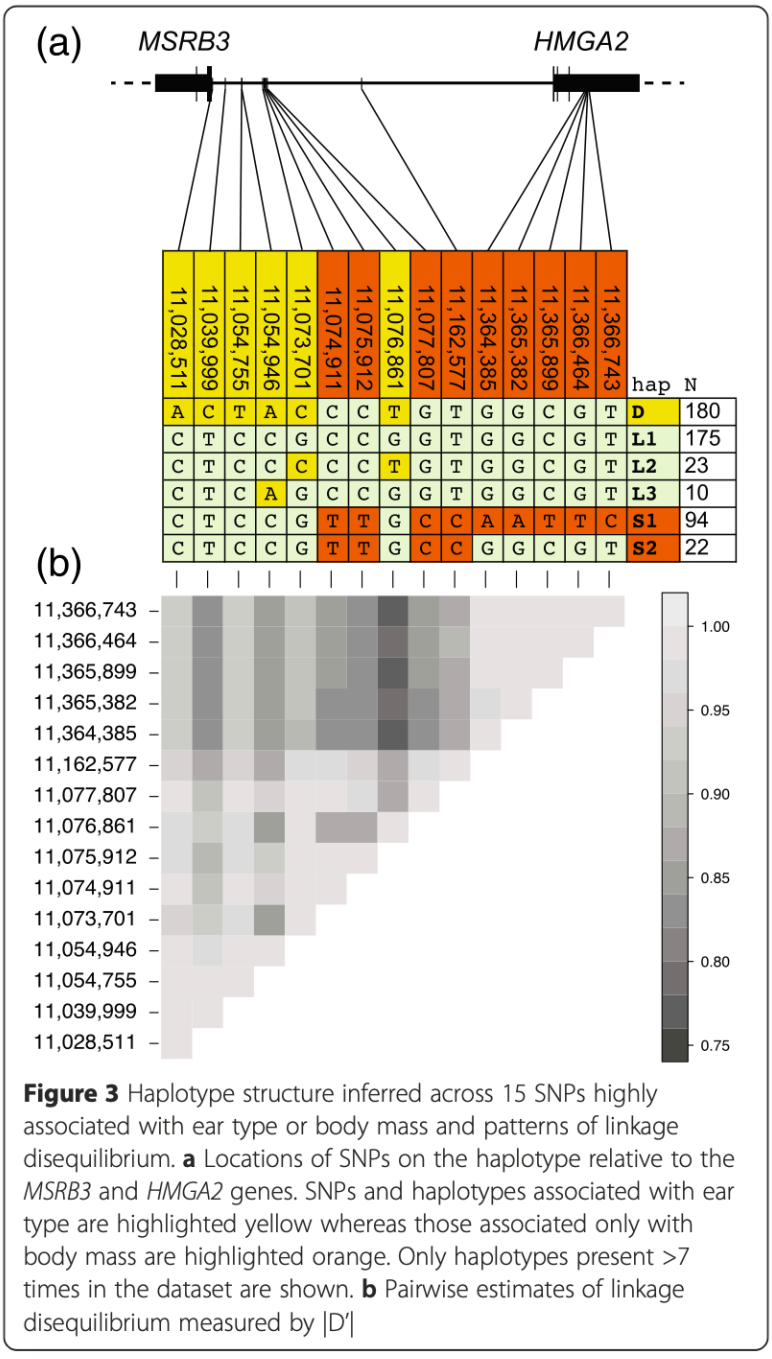


Figure 3 Haplotype structure inferred across 15 SNPs highly associated with ear type or body mass and patterns of linkage disequilibrium. **a** Locations of SNPs on the haplotype relative to the *MSRB3* and *HMGA2* genes. SNPs and haplotypes associated with ear type are highlighted yellow whereas those associated only with body mass are highlighted orange. Only haplotypes present >7 times in the dataset are shown. **b** Pairwise estimates of linkage disequilibrium measured by |D'|

Yang, M. *et al.* Genome-wide association studies reveal the genetic basis of ionomic variation in rice. *Plant Cell* **30**, 2720-2740, doi:10.1105/tpc.18.00375 (2018).

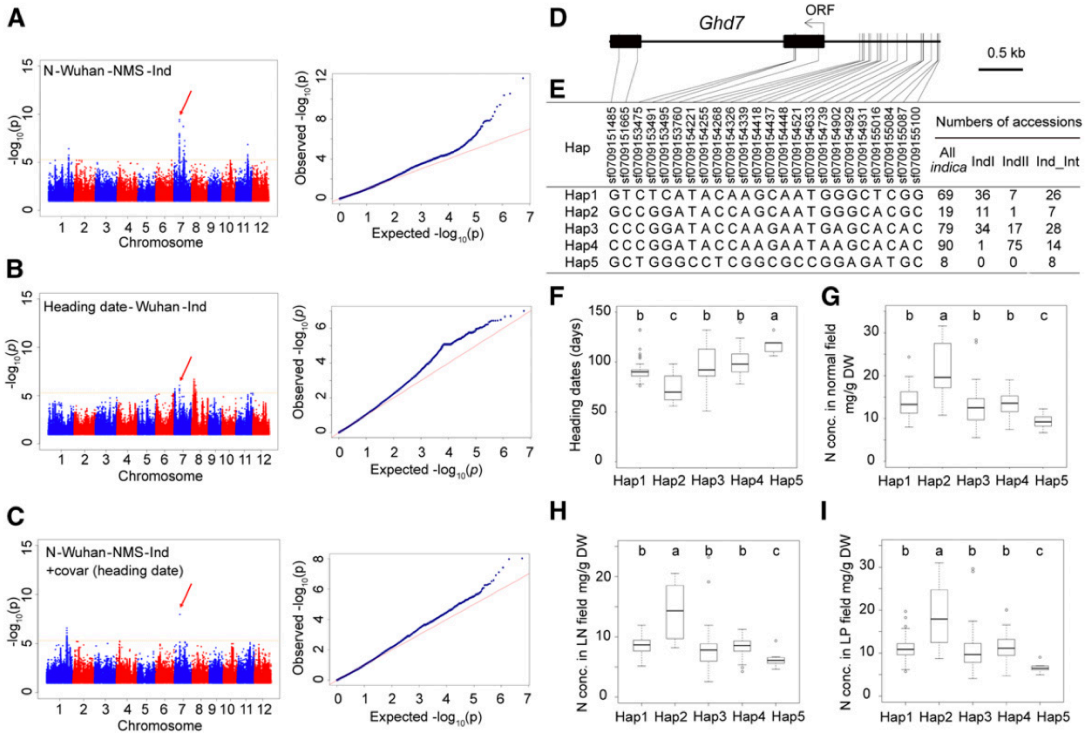


Figure 8. Characterization of the Role of *Ghd7* in N Accumulation in Rice by GWAS.

(A) Manhattan (left) and Q-Q (right) plots displaying the GWAS results for N concentration in shoots of the *indica* subpopulation at the heading stage in the LP field in Wuhan.

(B) Manhattan and Q-Q plots displaying the GWAS results of heading date in the *indica* subpopulation.

(C) Manhattan and Q-Q plots displaying the GWAS results of N concentration in shoots of the *indica* subpopulation at the heading stage in the LP field in Wuhan using heading date as a covariate.

Red arrows in **(A)**, **(B)**, and **(C)** point to a same lead SNP, which is located close to *Ghd7*.

(D) Gene model of *Ghd7*. The black filled boxes represent the coding sequence. The gray vertical lines mark the polymorphic sites identified by high-throughput sequencing in the *indica* subspecies. ORF, open reading frame.

(E) Haplotype analysis of the *Ghd7* gene region in the *indica* subspecies based on the polymorphic sites shown in **(D)**. Only haplotypes with total number of accessions ≥ 5 were analyzed.

(F) Box plot for heading dates of different Ghd7 haplotypes.

(G) to **(I)** Box plots for shoot N concentrations of different Ghd7 haplotypes at the heading stage in the NF **(G)**, LN field **(H)**, and LP field **(I)** in Wuhan. Significant differences at $P < 0.05$ within each group are indicated by different letters (one way ANOVA test). DW, dry weight.

Box plots represent the interquartile range, the thick line in the middle of the box represents the median, the whiskers represent 1.5 times the interquartile range, and the dots represent outlier points. The data are based on two biological replicates.

Zeng, X. *et al.* Genome-wide Dissection of Co-selected UV-B Responsive Pathways in the UV-B Adaptation of Qingke. *Mol. Plant* **13**, 112-127, doi:https://doi.org/10.1016/j.molp.2019.10.009 (2020).

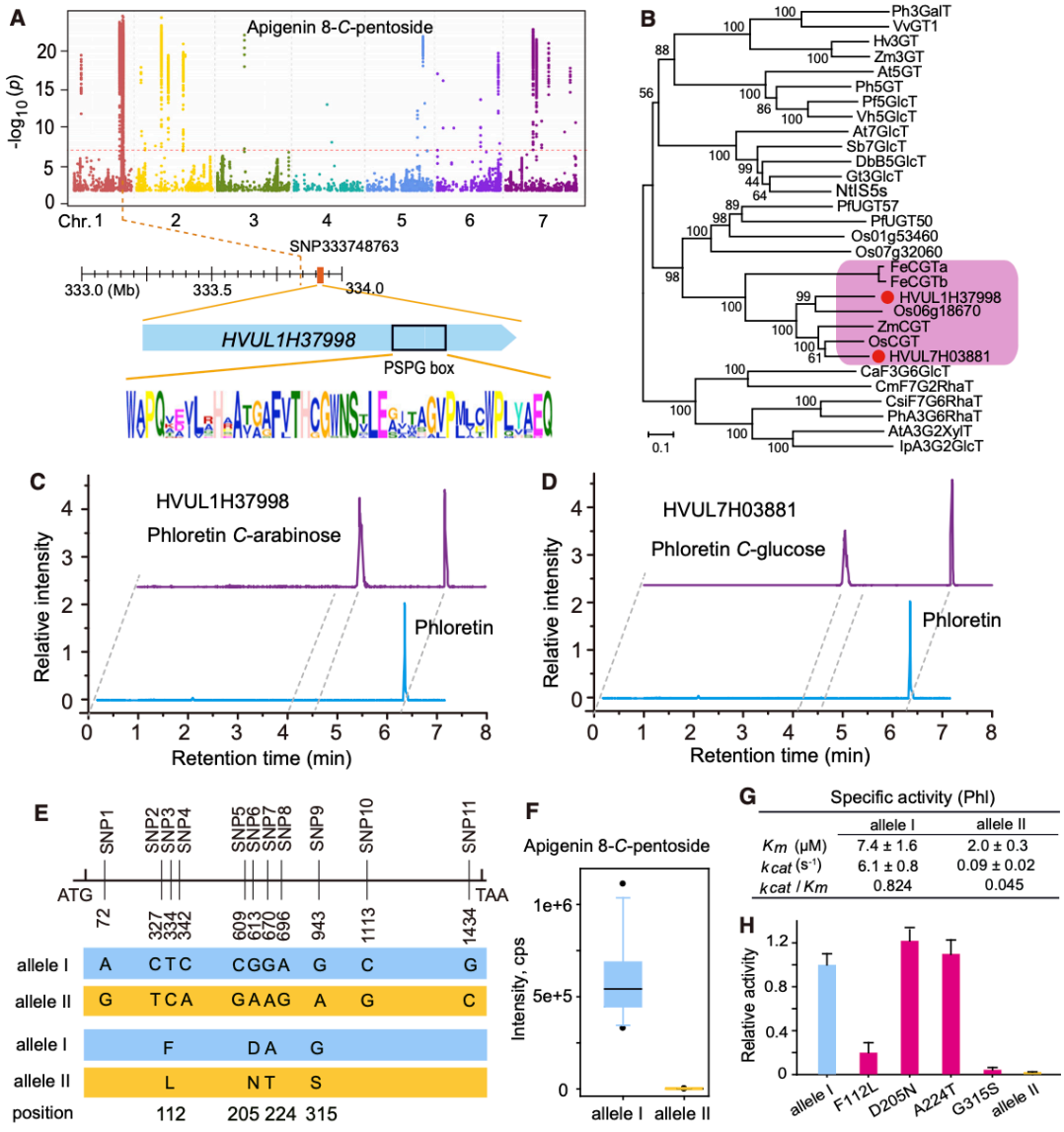


Figure 4. Functional Validation and Natural Variation of HVUL1H37998.

(A) Manhattan plot displaying the GWAS result for the content of apigenin 8-*C*-pentoside. Gene model of *HVUL1H37998*, which is located15 kb from the lead SNP (SNP 1:333748763), is shown. Conserved sequence of the plant secondary product glycosyltransferase box was obtained by collection of reported UGT.

(B) An unrooted phylogenetic tree was constructed as described in Methods. Bootstrap values > 70% (based on 1000 replications) are indicated at each node (bar: 0.1 amino acid substitutions per site).

(C and D) HPLC chromatograms of the products of the reactions of HVUL1H37998 **(C)** and HVUL7H03881 **(D)** with UDP-arabinose and UDP-glucose, respectively. Phloretin was used as a sugar acceptor.

(E) Nucleotide polymorphisms identified in the coding sequence of *HVUL1H37998*.

(F) Boxplot showing the content of apigenin 8-*C*-pentoside; plotted as an associated site at 11th SNP in the *HVUL1H37998* coding sequence.

(G) Enzymatic activity of HVUL1H37998 from two alleles. Allele I indicates the high production genotype, whereas allele II indicates the low production genotype. Assay was repeated three times and bars indicate mean ± SD, n = 3. $^{**}P < 0.01$.

(H) CGT activity assays of mutants compared with wild-type enzyme (average ± SD, n = 3).

Zhang, F. *et al.* Genome-wide association analysis identifies resistance loci for bacterial blight in a diverse collection of indica rice germplasm. *PLoS One* **12**, e0174598, doi:10.1371/journal.pone.0174598 (2017).

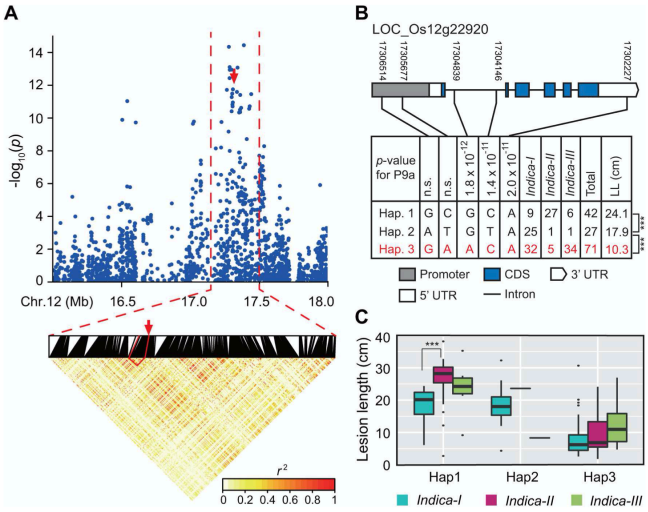


Fig 3. Hotspot region for the resistance to *Xanthomonas oryzae* pv. *oryzae* race P9a and haplotype analysis of the peak associated with the gene on chromosome 12. (A) Local Manhattan plot (top) (16.5–17.5 Mb) and linkage disequilibrium heatmap (bottom) (17.2–17.5 Mb) surrounding the hotspot region on chromosome 12. The arrow indicates the position of the peak single nucleotide polymorphism (SNP) located in *xa25* (*LOC_Os12g29220*). Dashed lines indicate the *xa25* region. (B) Gene structure and haplotype analysis of *xa25* in 140 accessions based on five significant SNPs in *xa25*. Haplotypes with fewer than five accessions are not shown. (C) Lesion lengths caused by P9a infections of accessions in three haplotypes of *xa25* in different *indica* subgroups. Box edges represent the 0.25 and 0.75 quantiles with median values indicated by bold lines. Whiskers extend to data no more than 1.5-times the interquartile range, and the remaining data are represented by dots. '***' refers to a significant difference based on Duncan's multiple comparison tests ($p < 0.001$).

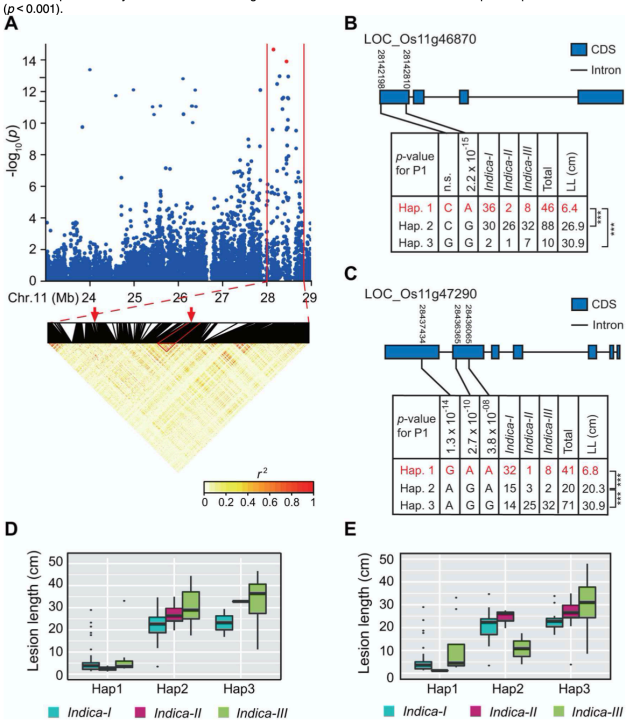


Fig 4. Hotspot region for the resistance to *Xanthomonas oryzae* pv. *oryzae* race P1 and haplotype analysis of the peak associated with the gene on chromosome 11. (A) Local Manhattan plot (top) (23–29 Mb) and linkage disequilibrium heatmap (bottom) (28.0–28.8 Mb) surrounding the hotspot region on chromosome 11. Red arrows and points indicate the positions of the peak single nucleotide polymorphisms located in the *Xa4* candidate gene (i.e., *LOC_Os11g46870*) and *Xa26* paralog (i.e., *LOC_Os11g47290*), respectively. Dashed lines indicate the *xa25* region. (B) Gene structure and haplotype analysis of the *Xa4* candidate gene (i.e., *LOC_Os11g46870*). (C) Gene structure and haplotype analysis of the *Xa26* paralog (i.e., *LOC_Os11g47290*). Lesion lengths caused by P1 infections of accessions in three haplotypes of *LOC_Os11g46870* (D) and *LOC_Os11g47290* (E) in different *indica* subgroups. '***' refers to a significant difference based on Duncan's multiple comparison tests ($p < 0.001$).

Zhang, H.-J. *et al.* Transcription factor gene TaNAC67 involved in regulation spike length and spikelet number per spike in common wheat. *Acta Agronomica Sinica* **45**, 1615-1627, doi:10.3724/sp.J.1006.2019.91009 (2019).

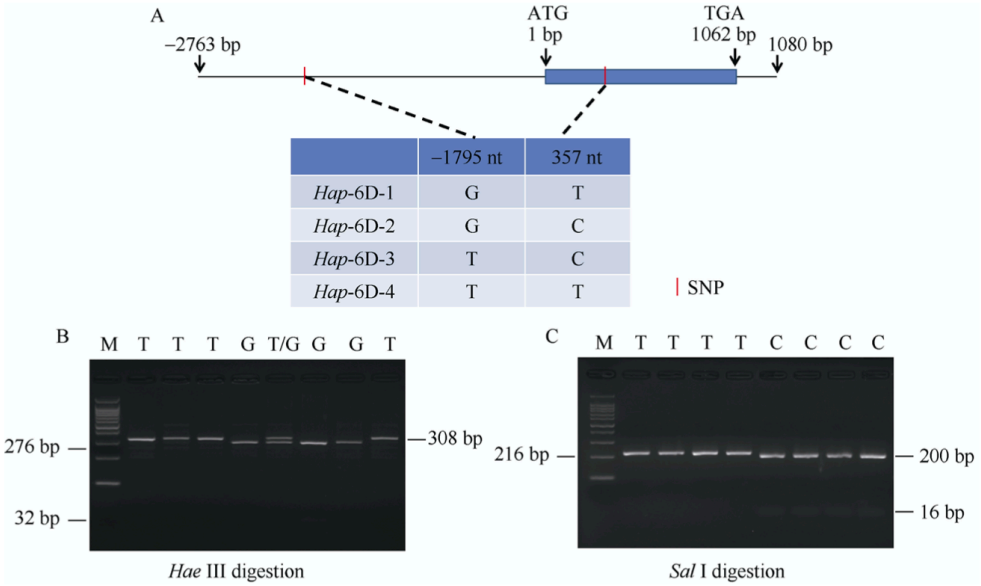


图 3　*TaNAC67-6D* 序列多态性(A)以及 CAPS 标记 SNP-D-1 (B)和 dCAPS 标记 SNP-D-2 (C)

Fig. 3　*TaNAC67-6D* sequence polymorphisms (A) and CAPS marker SNP-D-1 (B) and dCAPS marker SNP-B-2 (C)

图 B: 当–1795 nt 基因型是 G 时能被酶切，当基因型为 T 时不能被酶切；图 C: 当 357 nt 基因型是 T 时不能被酶切，当基因型是 C 时能被酶切。M: 100 bp DNA ladder。

Fig. B: if the genotype at –1795 nt is G, the PCR products can be digested; if it is T the PCR products cannot be digested. Fig. C: if the genotype at 357 nt is T the PCR products cannot be digested, if it is C the PCR products can be digested. M: 100 bp DNA ladder.