Safder I *et al.* (2021)
Notulae Botanicae Horti Agrobotanici Cluj-Napoca
Volume 49, Issue 3, Article number 12346
DOI:10.15835/nbha49312346
Research Article

Identification of SNPs in rice *GPAT* genes and in silico analysis of their functional impact on *GPAT* proteins

Imran SAFDER[a], Gaoneng SHAO[b], Zhonghua SHENG, Peisong HU, Shaoqing TANG*

*State Key Laboratory of Rice Biology and China National Center for Rice Improvement, China National Rice Research Institute, Hangzhou 310006, China; imransafdar82@gmail.com; shaogaoneng@caas.cn; shengzhonghua@caas.cnhupeisong@caas.cn; shaoqingtangcnrri@126.com (*corresponding author)*

[a,b] These authors contributed equally to the work

Abstract

SNPs are the most common nucleotide variations in the genome. Functional SNPs in the coding region, known as nonsynonymous SNPs (nsSNPs), change amino acid residues and affect protein function. Identifying functional SNPs is an uphill task as it is difficult to correlate between variation and phenotypes in association studies. Computational *in silico* analysis provides an opportunity to understand the SNPs functional impact to proteins and facilitate experimental approaches in understanding the relationship between the phenotype and genotype. Advancement in sequencing technologies contributed to sequencing thousands of genomes. As a result, many public databases have been designed incorporating this sequenced data to explore nucleotide variations. In this study, we explored functional SNPs in the rice *GPAT* family (as a model plant gene family), using 3000 Rice Genome Sequencing Project data. We identified 1056 SNPs, among hundred rice varieties in 26 *GPAT* genes, and filtered 98 nsSNPs. We further investigated the structural and functional impact of these nsSNPs using various computational tools and shortlisted 13 SNPs having high damaging effects on protein structure. We found that rice *GPAT* genes can be influenced by nsSNPs and they might have a major effect on regulation and function of *GPAT* genes. This information will be useful to understand the possible relationships between genetic mutation and phenotypic variation, and their functional implication on rice *GPAT* proteins. The study will also provide a computational pathway to identify SNPs in other rice gene families.

Keywords: 3000 Rice Genome project; functional SNPs; *in silico* analysis; nucleotide variation

Introduction

Rice genome was initially sequenced among cereal crops. It paved the path to sequence other complex plant genomes. The impact of sequencing rice genome was immediate: elicited high citations, DNA marker usage and research groups curated public databases (Jackson, 2016). An easy access to genome data stored in public databases provides opportunity to identify and explore nucleotide variations.

Received: 18 Apr 2021. Received in revised form: 22 Aug 2021. Accepted: 13 Sep 2021. Published online: 27 Sep 2021.
From Volume 49, Issue 1, 2021, Notulae Botanicae Horti Agrobotanici Cluj-Napoca journal uses article numbers in place of the traditional method of continuous pagination through the volume. The journal will continue to appear quarterly, as before, with four annual numbers.

SNPs are the key nucleotide variations found predominantly in genome, and are simplest in form. For instance, in the human genome, 90% of the variations belong to SNPs, located after every 100-300bp, and 170 bp in rice, but their densities may vary in different regions (Goff *et al.*, 2002; Nelson *et al.*, 2004; Kharabian, 2010). They have a wide presence in the genome and can be found in any part such as intronic region, mRNA, or in the intergenic regions.

SNPs are involved in many processes like the occurrence of disease in humans where a single nucleotide variation can be responsible for life-threatening disease (Lek *et al.*, 2016). In plants they can induce physiological, biochemical and phenotypic changes and alter function (Gailing *et al.*, 2009; Majeed *et al.*, 2019; Sandhu *et al.*, 2020). SNPs are named and categorized according to the region they are present and functions they perform. Among them, SNPs in the coding regions, termed as non-synonymous SNPs (nsSNPs) or missense SNPs are particularly important as they are responsible to change amino acid residues. These nsSNPs, therefore influence protein function, provide harmful or neutral effects to the protein structure, or reduce protein solubility (Zhang *et al.*, 2018) . They can alter gene regulation, protein charge, or either change inter and intra protein interactions. Therefore, they are crucial and it is beneficial to catalogue functional SNPs in different species. Although nsSNPs are important but SNPs including the coding synonymous SNPs (SNPs present in coding regions, but don't change the amino acid residues) or SNPs positioned outside the coding regions still have functions. They are attributed to impact gene expression due to changes in regulatory elements, exonic splice enhancers, binding of a transcription factor or in splicing processes (De Alencar and Lopes, 2010).

With the passage of time SNPs are gradually taking the place of SSRs, to use as markers in breeding purposes, as they are stable, efficient, have high presence, and bear less cost (McCouch *et al.*, 2010; Mammadov *et al.*, 2012). Marker-assisted selection is one of the methods used to attain molecular markers, discovered in large scale SNP-genotyping. Similarly, GWAS studies are also used to establish relationships with the SNP and the candidate genes (Zhang *et al.*, 2018).

Identification of SNPs is an initial yet tedious task to explore functional effects of SNPs and their correlation with the phenotype(Tibbs Cortes *et al.*, 2021), as it needs multi testing of hundreds of SNPs in the gene of interest. Still, the question would be which SNP set required for selection is important for association studies success, providing the strong reason for SNP variation. Many SNP assays have been employed, but because of not having accurate phenotypic and genotypic data, it's not easy to conduct these experiments. A particular breeding segregation population or near-isogenic line would be required to identify SNPS functional effects (Pea *et al.*, 2013).

Similarly, another task after the explosion of genome wide studies is to understand functional significance of the identified SNP, and apply it to application studies. Regarding experimental assays, many SNPs from the GWAS are not as impactful to select for certain traits, and breeders found it difficult to use these molecular markers. As a successful breaded crop depend largely on the accuracy of these functional SNPs (Tibbs Cortes *et al.*, 2021).

SNPs detected by experimental studies provide the strongest evidence to demonstrate functional significance. However, because of the lack of exactness of the phenotypic and genotypic data, these experiments are not easy to identify. An alternative, to explore the possible significant pattern of SNPs is to prioritize SNPs on their functional significance using *in silico* computational tools. These computational tools can identify and differentiate the neutral SNPs from the functionally significant SNPs, and infer nature of mutation and changes in protein structure caused by the particular SNP (Kharabian, 2010; Arshad and Attya Bhatti, 2018). In addition, similar to past *in silico* studies, they could also provide an independent evidence source alongside experimental studies to explore functionally important SNPs (Yang *et al.*, 2004; Jiang *et al.*, 2010; Chaisan *et al.*, 2012; Bhardwaj *et al.*, 2016; Withana *et al.*, 2020).

Advances in big data analytics and artificial intelligence systems provide opportunities to build machine learning models that can predict accurate DNA variations and protein models identified from sequence data. For instance, Alphafold recently gained appreciation among the scientific community. Alphafold is an artificial

intelligence and deep learning based prediction model for protein structures that achieved high accuracy even for sequences having lower homologous sequences (Senior *et al.*, 2020; Yang *et al.*, 2020). Such breakthroughs reveal the potential of AI on genome data and it is presumed that computational approaches will be frequently used in plants to recruit SNPs in the future (Korani *et al.*, 2019).

There are several public databases to curate SNPs in rice genome. 3000 Rice Genome Project (3KRGP) consortium has sequenced 3000 rice varieties from 89 different countries to explore genomic diversity in rice crop, and approximately 18.9 million SNPs were identified from this project (Li *et al.*, 2014). The consortium provides rice breeders and scientists a massive resource. In the past years, some comprehensive databases have been created using the 3KRGP data including, SNP-Seek database (Alexandrov *et al.*, 2015), Rice Functional and Genomic Breeding (RFGB) (Wang *et al.*, 2020a), RiceVarMap (Zhao *et al.*, 2015). These public datasets provide opportunities to identify large-scale discovery of genomic variants associated with various traits, and can be tapped to increase yield potential. SNP-Seek database platform has incorporated millions of variants from 3000 rice accessions and provide easy access to mine alleles (Mansueto *et al.*, 2016).

In this study, we aim to analyze all the functional SNPs in the rice Glycerol-3-phosphate acyltransferase (*GPAT)* gene family, using 3 K RGP data among 100 Chinese rice varieties from different geographical regions. We used rice *GPAT* genes as a model family, we previously identified in a genome wide study (Safder *et al.*, 2021) to demonstrate *in silico* analysis in a gene family. We used different computational tools to explore SNPs, and their impact to the structure and function on rice *GPAT* proteins and prioritized SNPs having significant impact on *GPAT* proteins. The study will provide useful information about important SNPs that can affect protein functions and would be useful in future investigations.

Materials and Methods

SNP Dataset and missense SNP identification in GPAT genes

We retrieved the SNP data set of 100 rice accessions from the SNP-Seek database (*http://snp-seek.irri.org*) (Alexandrov *et al.*, 2015) at each rice *GPAT* locus. We used 26 *GPAT* locus positions identified in a genome wide study (Table S6). Then, we searched each locus position among 100 rice varieties in the SNP-Seek database and find all the SNPs in 26 *GPAT* genes. Afterwards, we filtered all missense or nonsynonymous SNPs (nsSNPs) among initially identified SNPs. We find and documented nsSNPs details (SNP position, protein remnant change) to investigate further.

SNPs identification with damaging effects

We used three computational databases to explore functional impact of nsSNPs on protein structures including, Protein Variation Effect Analyzer (PROVEAN) [*http://provean.jcvi.org/index.php*] (Choi *et al.*, *2012*; Choi and Chan, 2015), Sorting Intolerant from Tolerant (SIFT) (Ng and Henikoff, 2003), and PolyPhen-2 (Adzhubei *et al.*, 2010). These computational tools predicts whether an amino acide substitution by a DNA variant affects protein function based on sequence homology and physical properties of amino acids. The nsSNPs having a damaging or deleterious effect identified by these three tools were regarded as high risk nsSNPs and taken further for more investigation to analyze their putative effect.

SNPs influence on structural and functional properties of GPAT proteins

GPAT protein sequences carrying the high risk nsSNPs identified in the previous step were further examined along with the mutated amino acid residues, and submitted to the MutPred v1.2 (Li *et al.*, 2009a) database. The MutPred database investigates the outcome of mutations at proteins and predicts the molecular mechanism associated with the mutation and also provides different gain and loss structural properties of a protein.

Effects on protein stability

We used I-Mutant to examine protein stability affected by nonsynonymous SNPs by submitting the normal and mutated amino acid sequences. I-Mutant evaluates protein stability changes or any structural change after variation. The tool provides a relativity RI index (RI) of results, ranges 0 to 10 score, showing the reliability of the score.

Evolutionary conservation of amino acid positions

We used ConSurf tool to determine amino acids evolutionary conservation, affected by nsSNPs; to analyze if this amino acid position is highly conserved in a protein sequence. Consurf analyzes the degree of evolutionary conservation using 50 homologous sequences from different species based on phylogenetic relation. We considered those positions important, if they were highly conserved and located on sites affected by nsSNPs. (Ashkenazy *et al.*, 2010; Celniker *et al.*, 2013; Ashkenazy *et al.*, 2016).

3D modeling of protein structures and RMSD calculation

Native and mutated protein structure models were generated using protein homology tools to evaluate effects to protein structure caused by high risk nsSNPs. Phyre2 (Kelley *et al.*, 2015) was used to generate the protein models, and these structures were further viewed by Chimera(Pettersen *et al.*, 2004). In addition, we used two more tools including Root Mean Square Deviation (RMSD), and Template Modeling Score (TM-align). RMSD indicates variation between two protein structures; a high score shows more variation between native and mutant structure (Carugo and Pongor, 2001; Zhang and Skolnick, 2005).

Post translational modification (PTM) sites

Post translation modification sites (PTM) in the rice *GPAT* Proteins were predicted including the phosphorylation, ubiquitylation and methylation sites. We used two computational tools to predict every PTM event. Phosphorylation PTM sites at Serine (S), Threonine (T), and Tyrosine (Y) amino acid residues were predicted using GPS 5.0 (Xue *et al.*, 2005) and NetPhos 3.1 (Blom *et al.*, 1999). In NetPhos 3.1, residues having at least 0.5 score were considered as phosphorylated. Ubiquitylation PTM sites were predicted using BDM-PUB (Li *et al.*, 2009b) and UbPred (Radivojac *et al.*, 2010). In UbPred only those residues were considered having a 0.62 or more score. Methlyation sites were predicted using PSSMe (Wen *et al.*, 2016) and iMethyl-PseAAC (Qiu *et al.*, 2014) tools. PSSMe predicted those lysine or arginine residues that had higher probability ratios.

Results

Functional SNPs extracted from the rice SNP-seek database

We used the rice SNP-seek database to explore all the SNPs in 26 *GPAT* genes. We used rice GPAT gene family as a model plant gene family, we reported in a genome wide study (Safder *et al.*, 2021). SNP-seek database comprised 18 million SNPs identified in 3KRGP. To recruit SNPs, we shortlisted 100 Chinese rice accessions from different geographical regions (including indica, japonica) (Table S1). Initially we find 1056 SNPs, among 100 varieties at each *GPAT* locus; 99 were found in the UTR regions, 683 in the intronic regions and 140 in the codon synonymous region (Figure 1). We find 98 SNPs (non-synonymous SNPs or nsSNPs) in the coding regions, which can influence protein structure and function (Table 1). We selected these 98 nsSNPs for further investigation.

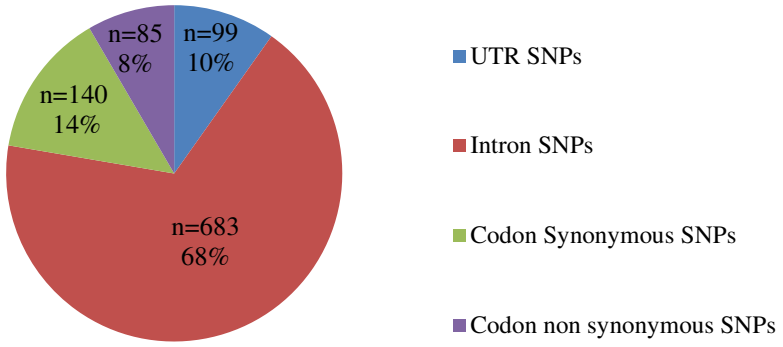

Figure 1. SNPs percentage in 26 rice *GPAT* genes
SNPs in different regions, including the UTR region, intronic region, and coding synonymous and nonsynonymous region

Table 1. Detailed information of 98 nsSNPs in 26 *GPAT* genes

Gene ID	SNP position	Allele change	Amino acid position and residue change
Os*GPAT*1	8348616	G>T	Ala242Ser
	8350199	C>G	Ala362Gly
	8350403	G>A	Arg430His
	8350709	C>T	Ala532Val
Os*GPAT*2	10970025	G>C	Thr223Ser
	10972698	C>T	Glu62Lys
	10972857	C>T	Gly9Ser
Os*GPAT*4	25257427	T>C	Phe59Ser
Os*GPAT*5	33154556	C>A	Gly348Cys
	33156296	C>A	Met88Ile
	33156464	C>A	Lys32Asn
Os*GPAT*6	3686407	G>A	Ala255Val
Os*GPAT*7	40865354	A>G	Val452Ala
	40865364	T>C	Asn449Asp
	40865933	T>C	Gly292Gly
	40867989	A>G	Val81Ala
	40868136	G>A	Ala32Val
Os*GPAT*8	776045	C>T	Val406Met
	776197	C>G	Arg355Pro
	776230	G>A	Ala344Val
	778246	C>T	Gly8Glu
Os*GPAT*9	30152069	C>A	Gln361His
Os*GPAT*10	34985769	G>A	Met160Ile
	34985895	G>C	Gln202His
	34985957	G>A	Gly223Asp
Os*GPAT*11	31786785	T> C	Val45Ala
	31787035	G>C	Lys99Asn
	31787058	A>G	Asn107Ser
	31787320	C>G	His160Asp
	31787530	A>G	Asp191Gly
	31787808	C>G	Asp252Glu
	31787824	A>G	Arg258Gly
Os*GPAT*12	34058002	G>A	Ala34Val
Os*GPAT*13	11754042	G>A	Ala357Val
	11755949	C>A	Ala127Ser
	11756221	G>T	Thr36Lys
Os*GPAT*16	22481879	G>A	Ala51Thr
	22484156	C>A	Gln207Lys
	22483271	G>C	Ala502Pro
Os*GPAT*17	24725803	C>T	Pro157Leu

Os*GPAT*18	30123014	G>C	Asp36His
	30127549	G>C	Gly545Arg
Os*GPAT*19	20809659	C>T	Ser231Asn
	20810625	T>G	Met122Leu
	20811166	T>C	Ile90Val
	20812394	G>C	Pro15Ala
Os*GPAT*20	1757860	C>T	Gly447Ser
	1757860	G>T	Phe412Leu
	1758474	G>A	His399Tyr
	1758474	A>G	Val242Ala
	1758624	G>A	Ser161Ile
	1758717	C>A	Ala155Val
	1758774	G>A	Ala142Val
	1758815	G>A	Phe128Leu
	1758876	C>A	Arg108Leu
Os*GPAT*22	22060049	G>C	Ala258Pro
Os*GPAT*24	25200311	T>C	Gln327Arg
	25200336	C>T	Ala319Thr
	25201064	G>A	.Pro297Leu
	25201089	C>A	Gly289Cys
	25201303	C>T	Ala256Thr
Os*GPAT*25	27486970	C>T	Pro2Leu
	27486991	G>A	Arg9Lys
	27487044	T>C	Ser27Pro
	27487108	T>C	Phe48Ser
	27487132	C>T	Ala56Val
	27487269	G>A	Val102Ile
	27487390	A>T	Glu142Val
	27487401	G>A	Asp146Asn
	27487501	C>T	Ser179Leu
	27487577	G>C	Glu204Asp
	27487578	G>A	Val205Met
	27487604	G>T	Leu213Phe
	27487627	C>T	Thr221Ile
	27489105	G>A	Arg242His
	27489194	G>A	Ala272Thr
	27489200	G>A	Gly274Ser
	27489701	G>C	Asp441His
	27489766	C>C	Ile462Met
	27489853	C>C	Asp491Glu
	27489854	G>A	Val492Met
	27489880	G>T	Met500Ile
	27489956	G>C	Val526Leu
	27489979	C>G	Asn533Lys
	27489986	G> G/T	Ala536Ser
	23091766	C>G	Ala44Gly
	23091780	A>G	Thr49Ala
	23091784	T>G	Val50Gly
	23091790	T>C	Met52Thr
	23091918	G>C	Val95Leu
	23091937	G>C	Gly101Ala
	23091943	G>C	Gly103Ala
	23092023	A>G	Thr130Ala
	23092134	G>A	Ala167Thr
	23092180	A>T	Tyr182Phe
	23092335	G>A	Val234Ile
	23093266	C>T	Pro268Ser
	23094109	G>A	Val549Ile

Deleterious nsSNPs explored from Computational tools

By employing *in silico* tools, SIFT, PROVEAN and PolyPhen, we explored 98 nsSNPs effect to variant amino acid residues. These tools predict functional effects of amino acid substitutions to the protein structure (Table 2).

Each tool has a cut-off value; a nsSNP is considered functional below this value (Table 2.). For instance, PROVEAN score is -2.5, below this cutoff, the substitutions are considered "deleterious", having high risk or deleterious effects on protein, and above this cutoff variant are considered Neutral. We considered nsSNPs as high risk or having damaging effects, if predicted by 2 of these tools. We find at least 13 nsSNPs (Table 2) in different genes having high risk or damaging effects on the protein structure.

Table 2. Effect of nsSNPs on protein structures analyzed by different computational tools

Aminoacid change	SIFT		PROVEAN		Polyphen-2	
	SIFT prediction	SIFT tolerance	Score	Cutoff (-2.5)	Effect	Score
Ala242Ser	Tolerated	0.88	0.708	Neutral	Benign	
Ala362Gly	Tolerated	0.37	-0.054	Neutral	Benign	
Arg430His	**Affected**	0.05	-1.533	Neutral	**Probably damaging**	0.999
Ala532Val	Tolerated	1	-3.007	**Deleterious**	Possibly damaging	0.915
Thr223Ser	Tolerated	0.44	-0.186	Neutral	Benign	
Glu62Lys	Tolerated	0.06	-3.674	**Deleterious**	**Probably damaging**	0.997
Gly9Ser	Tolerated	0.62	-0.835	Neutral	Possibly damaging	0.762
Phe59Ser	Tolerated	0.42	0.462	Neutral	Benign	
Gly348Cys	Tolerated	0.12	-1.308	Neutral	Benign	
Met88Ile	Tolerated	0.39	-0.138	Neutral	Benign	
Lys32Asn	**Affected**	0	-1.235	Neutral	Possibly damaging	0.625
Ala255Val	Tolerated	0.18	1.378	Neutral	Benign	
Val452Ala	**Affected**	0	-0.456	Neutral	Possibly damaging	0.889
Asn449Asp	**Affected**	0	0.192	Neutral	Benign	
Gly292Gly	**Affected**	0	0	Neutral	Benign	
Val81Ala	**Affected**	0	0.18	Neutral	Benign	
Ala32Val	**Affected**	0	-0.578	Neutral	**Probably damaging**	0.998
Val406Met	Tolerated	0.51	0.981	Neutral	Benign	
Arg355Pro	**Affected**	0	-6.543	**Deleterious**	**Probably damaging**	1
Ala344Val	Tolerated	0.23	2.209	Neutral	Benign	
Gly8Glu	Tolerated	1	0.107	Neutral	**Probably damaging**	1
Gln361His	Tolerated	0.46	-1.184	Neutral	Possibly damaging	0.513
Met160Ile	Tolerated	0.41	-0.922	Neutral	Benign	
Gln202His	Tolerated	0.2	-0.086	Neutral	Benign	
Gly223Asp	Tolerated	0.26	-0.506	Neutral	**Probably damaging**	0.997
Val45Ala	Tolerated	0.31	-0.73	Neutral	Benign	
Lys99Asn	Tolerated	0.07	-0.922	Neutral	Benign	
Asn107Ser	Tolerated	0.72	-1.117	Neutral	Benign	
His160Asp	Tolerated	1	5.701	Neutral	Benign	
Asp191Gly	Tolerated	1	6.58	Neutral	Benign	
Asp252Glu	Tolerated	0.8	-0.411	Neutral	Benign	
Arg258Gly	Tolerated	0.28	-1.074	Neutral	Benign	
Ala34Val	Tolerated	0.09	-0.907	Neutral	Possibly damaging	0.62

Ala357Val	Tolerated	1	-0.489	Neutral	Benign	
Ala127Ser	Tolerated	0.7	-1.866	Neutral	**Probably damaging**	0.997
Thr36Lys	Tolerated	1	-0.455	Neutral	Benign	
Ala51Thr	Tolerated	0.07	-0.77	Neutral	Possibly damaging	0.863
Gln207Lys	Tolerated	0.97	-0.497	Neutral	Possibly damaging	0.78
Ala502Pro	**Affected**	0	-0.911	Neutral	Probably damaging	0.997
Pro157Leu	**Affected**	0	-8.92	**Deleterious**	**Probably damaging**	1
Asp36His	**Affected**	0	-0.25	Neutral	**Probably damaging**	0.988
Gly545Arg	Tolerated	0.08	0.216	Neutral	Possibly damaging	0.641
Ser231Asn	Tolerated	0.11	1.647	Neutral	Benign	
Met122Leu	Tolerated	1	-0.159	Neutral	Benign	
Ile90Val	Tolerated	0.63	-0.503	Neutral	Benign	
Pro15Ala	Tolerated	1	0.036	Neutral	Benign	
Gly447Ser	**Affected**	0	-4.122	**Deleterious**	**Probably damaging**	1
Phe412Leu	Tolerated	0.73	-2.128	Neutral	Benign	
His399Tyr	Tolerated	1	-1.996	Neutral	Possibly damaging	0.748
Val242Ala	Tolerated	0.15	-1.585	Neutral	Benign	
Ser161Ile	Tolerated	0.27	-0.482	Neutral	**Probably damaging**	0.984
Ala155Val	Tolerated	1	1.86	Neutral	Benign	
Ala142Val	Tolerated	0.34	-0.772	Neutral	Benign	
Phe128Leu	**Affected**	0	-5.526	**Deleterious**	**Probably damaging**	0.984
Arg108Leu	Tolerated	0.08	-1.221	Neutral	Possibly damaging	0.458
Ala258Pro	Tolerated	0.2	2.652	Neutral	Benign	
Gln327Arg	Tolerated	0.16	-0.172	Neutral	Benign	
Ala319Thr	Tolerated	0.39	-0.591	Neutral	Benign	
.Pro297Leu	Tolerated	0.27	-2.273	Neutral	Possibly damaging	0.548
Gly289Cys	**Affected**	0.01	-4.878	**Deleterious**	**Probably damaging**	0.999
Ala256Thr	Tolerated	0.06	0.161	Neutral	Benign	
Pro2Leu	**Affected**	0	0.883	Neutral	Benign	
Arg9Lys	**Affected**	0	-0.021	Neutral	Benign	
Ser27Pro	**Affected**	0	-0.847	Neutral	Possibly damaging	0.777
Phe48Ser	**Affected**	0	-0.213	Neutral	Benign	
Ala56Val	Tolerated	0.61	-1.302	Neutral	Possibly damaging	0.938
Val102Ile	Tolerated	0.81	0.474	Neutral	Benign	
Glu142Val	Tolerated	0.41	1.225	Neutral	Benign	
Asp146Asn	Tolerated	0.2	-2.508	**Deleterious**	Possibly damaging	0.752
Ser179Leu	Tolerated	0.58	2.299	Neutral	Benign	
Glu204Asp	Tolerated	0.63	-0.936	Neutral	Possibly damaging	0.593
Val205Met	**Affected**	0.04	-0.176	Neutral	Possibly damaging	0.503
Leu213Phe	Tolerated	0.7	-0.862	Neutral	Benign	
Thr221Ile	Tolerated	0.2	-2.353	Neutral	Benign	
Arg242His	Tolerated	0.14	-0.896	Neutral	Benign	
Ala272Thr	Tolerated	1	4.744	Neutral	Benign	
Gly274Ser	Tolerated	0.35	1.037	Neutral	Benign	
Asp441His	Tolerated	0.28	2.559	Neutral	Benign	
Ile462Met	Tolerated	0.16	0.233	Neutral	Benign	

Asp491Glu	Tolerated	1	0.511	Neutral	Benign	
Val492Met	**Affected**	0	-2.467	Neutral	**Probably damaging**	0.973
Met500Ile	Tolerated	0.73	1.228	Neutral	Benign	
Val526Leu	Tolerated	0.46	-1.888	Neutral	Benign	
Asn533Lys	Tolerated	1	0.533	Neutral	Benign	
Ala536Ser	**Affected**	0	-0.279	Neutral	Benign	
Ala44Gly	Tolerated	0.38	-0.537	Neutral	Benign	
Thr49Ala	Tolerated	0.1	-0.571	Neutral	Benign	
Val50Gly	Tolerated	0.07	-0.242	Neutral	Benign	
Met52Thr	Tolerated	0.05	-2.242	Neutral	Benign	
Val95Leu	Tolerated	0.64	0.973	Neutral	Benign	
Gly101Ala	Tolerated	0.51	0.101	Neutral	Benign	
Gly103Ala	Tolerated	0.59	-0.942	Neutral	Benign	
Thr130Ala	Tolerated	1	2.286	Neutral	Benign	
Ala167Thr	Tolerated	0.32	0.039	Neutral	Possibly damaging	**0.691**
Tyr182Phe	Tolerated	0.25	-3.812	**Deleterious**	**Probably damaging**	1
Val234Ile	Tolerated	0.59	-0.048	Neutral	Benign	
Pro268Ser	Tolerated	0.52	-0.526	Neutral	Possibly damaging	0.859
Val549Ile	**Affected**	0	-0.659	Neutral	**Probably damaging**	1

Structural and functional modifications

We submitted all the high risk nsSNPs to MuPred, to explore functional changes due to amino acid variations (Table 3). Deleterious nsSNPs influenced several mechanisms; some with high probability score in Arg355Pro (altered trans membrane protein), Pro157Leu (gain of helix), E62 (gain of Acetylation, loss of loop). Variations in V492M and V549I affected most mechanisms. Some amino acid variations had low probability score; therefore, no mechanism was detected. These functional changes revealed nsSNPs could influence functional variation in *GPAT* proteins.

Table 3. Effect of high-risk nsSNPs to rice *GPATs* functional and structural mechanisms

Mutation	Probability of deleterious mutation	Molecular mechanism altered by mutations
R430H	0.148	
D36H	0.185	
E62K	0.627	Gain of acetylation at E62 (0.0013)
A32V	0.084	
R355P	0.925	Altered transmembrane protein (0.0086)
A502P	0.224	
P157L	0.941	Gain of Helix (0.02)
D36H	0.185	
G447S	0.456	
F128L	0.761	Altered Metal binding (0.04)
G289C	0.655	Loss of Ubiquitylation at K284(0.03) Altered Transmembrane protein (0.03) Gain of Disulfide linkage at G289 (0.05)
V492M	0.66	Altered DNA binding (0.05)

V549I	0.542	Altered DNA binding (0.05) Gain of Relative solvent accessibility (0.01) Altered Metal binding (0.00026) Altered Disordered interface (0.04) Altered Ordered interface (0.05) Altered DNA binding (0.0008) Gain of Allosteric site at N545(0.01)

Effects on protein stability

We used I-Mutant to analyze stability changes to *GPAT* proteins by the nsSNPs. This tool predicts the reliability index (RI) and energy change values. The results revealed that all the variations decreased the stability of *GPAT* proteins (Table 4). Protein stability is vital to maintain its three dimensional structure; a reduction in stability causes protein denaturation, unfolding, and protein aggregation.(Ortbauer *et al.*, 2013; Deller *et al.*, 2016).

Table 4. nsSNPs impact to protein stability, TM-Score and, RMSD values

Gene ID	nsSNP position	Amino acid change	Stability	RI	DDG	TM-score	RMSD
OsGPAT 1	8350403	R430H	Decrease	9	-1.12	0.51851	5.67
OsGPAT 2	10972698	E62K	Decrease			0.49351	2.35
OsGPAT 7	40868136	A32V	Decrease	6	-0.67	0.52881	4.28
OsGPAT 8	776197	R355P	Decrease	0	-0.32	0.42009	3.55
OsGPAT 16	22484156	A502P	Decrease	7	-1.07	0.49196	4.48
OsGPAT17	24725803	P157L	Decrease	4	-0.42	0.68016	3.8
OsGPAT 18	30123014	D36H	Decrease	5	-1.09	0.46905	3.02
OsGPAT20	1757860	G447S	Decrease	4	-0.96	0.48605	5.13
OsGPAT 20	1758815	F128L	Decrease	6	-0.85	0.58509	4.77
OsGPAT 24	25201089	G289C	Decrease	7	-0.92	0.783	2.88
OsGPAT 25	27489854	V492M	Decrease	7	-0.86	0.4237	2.99
OsGPAT 26	23092180	Y182F	Decrease	0	-0.17	0.45534	3.85
OsGPAT 26	23094109	V549I	Decrease	2	0.62	0.45009	3.31

Free Energy change value (DDG); Reliability index (RI); Tm-Align score -similarity between native and mutant protein 3D models; Root mean square difference (RMSD)-variation between wild type and mutant protein; 3D model score

Evolutionary conservation of amino acid residue positions in rice GPAT genes

We used Consurf web tool to infer the evolutionary conservation of amino acid residues positions effected by nsSNPs. Consurf analyzes amino acid residues and allocate a conservation scale ranging between highly conserved, average, and variable to each amino acid residues, where variable reflect the lowest score. The score is calculated by merging the solvent accessibility predictions and evolutionary conservation data.

Table 5 demonstrates the conservation score of some high-risk nsSNPs in each *GPAT* gene. We found that amino acid positions R355P, P157L, F128L, V549I were highly conserved, whereas R430H, E62K, A502P, D36H were conserved. These results suggested amino acid residues affected by nsSNPs are located at evolutionary conserved positions. We have shown two highly conserved and exposed nsSNPs structures in Figures 2A and 2B respectively.

Amino acid sites that are largely conserved, participates in important biological functions (Berezin *et al.*, *2004*; Arshad and Attya Bhatti, 2018).In this context, nsSNP positions which are also found in conserved positions are significant for protein functions.

Table 5. Evolutionary conservation profile of amino acid positions affected by high-risk nsSNPs

Gene ID	nsSNP position	Amino acid change	Conservation score	Prediction
OsGPAT1	8350709	R430H	6	Conserved (e)
OsGPAT2	10972698	E62K	7	Conserved (e)
OsGPAT7	40868136	A32V	8	buried residue (b)
OsGPAT8	776197	R355P	9	Highly conserved and exposed (f)
OsGPAT16	22484156	A502P	2	Conserved (e)
OsGPAT17	24725803	P157L	9	highly conserved and buried (s)
OsGPAT18	30123014	D36H	2	Conserved (e)
OsGPAT20	1757860	G447S	5	exposed residue (e)
OsGPAT20	1758815	F128L	9	Highly conserved and exposed (f)
OsGPAT24	25201089	G289C	5	exposed residue (e)
OsGPAT25	27487401	D146N	6	exposed residue (e)
OsGPAT26	23092180	Y182F	6	buried residue (b)
OsGPAT26	23094109	V549I	8	Highly conserved and exposed (f)

The Conservation scale:
Variable: 1-3, Average: 4-6, Conserved amino acid: 7-9; (f): highly conserved and exposed; (s): highly conserved and buried (e): exposed amino acid residues, (b): buried amino acid residue

Structural analysis and modeling of wild type and mutant protein 3D structures

To explore whether the high risk nsSNPs impact protein structure, we generated 3D protein models of native and variant structures in genes carrying high impact nsSNPs mentioned in Table 4. We initially used Phyre2 that generated structures in pdb format, and then visualized by using Chimera. Besides, we analyzed the TM-align score and RMSD scores for the protein models (Table 4). A higher RMSD value demonstrates the deviation between mutant and the native model. Based on RMSD values, most mutant models showed high variation in structures with 2 or more RMSD value. Similarly, no model revealed low or zero value, demonstrating each nsSNPs have affected protein structures in some capacity. Finally, to demonstrate 3D structures, we selected four proteins having at least 2 RMSD values and low TM values. Figure 3 display the location of amino acid substitutions in each protein native and variant model. In these models, there were residual changes, along with variation in parameters including total energy, decrease in protein stability, suggesting they can influence protein folding. We further superimposed these variant models in figure 3 over wild types to explore the structural difference between them. Every superimposed model showed a difference in 3D structure, suggesting these nsSNPs affected the protein structure (Figure 4).

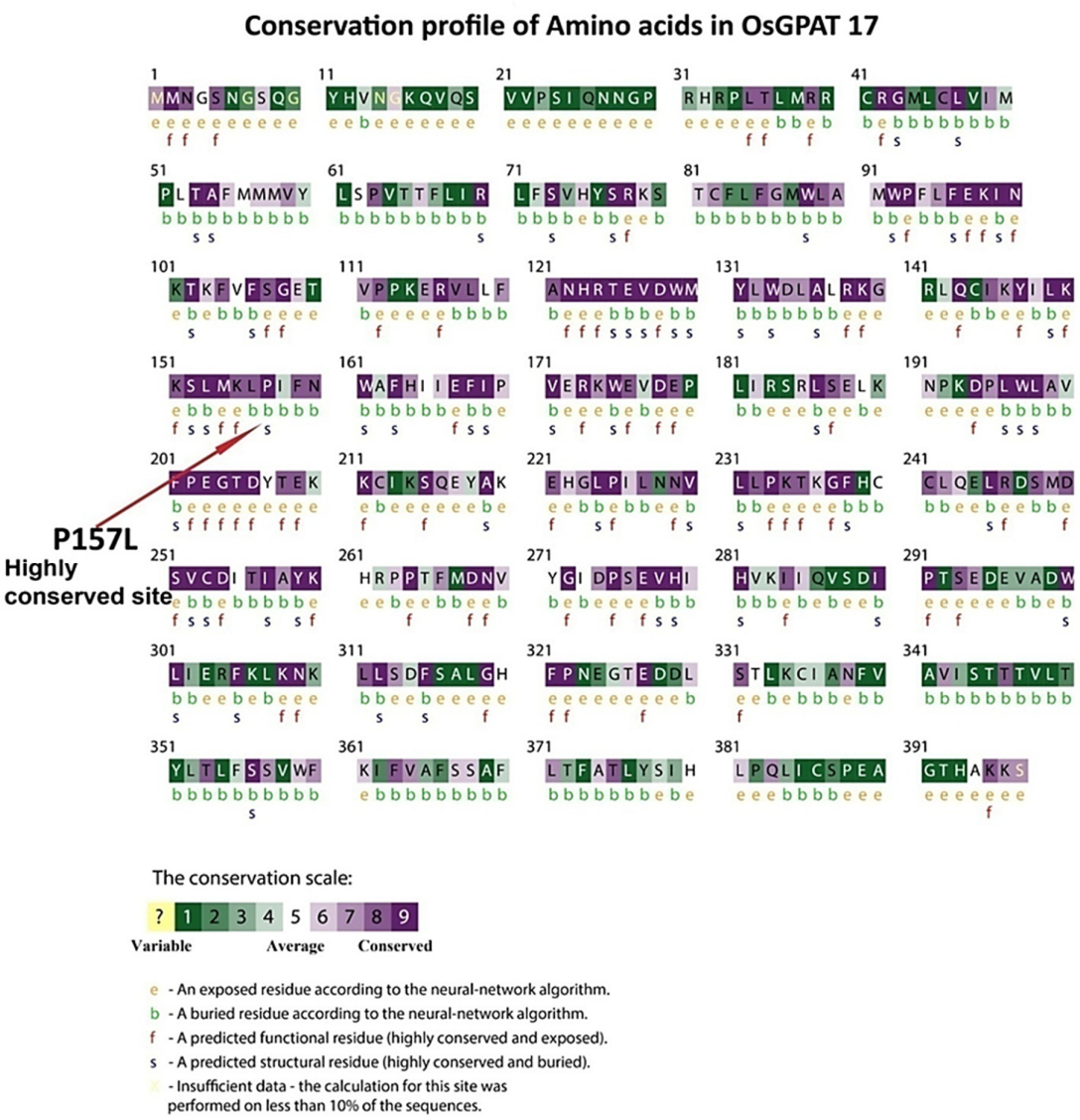

Figure 2A. Evolutionary conservation profile in *OsGPAT 17* demonstrating P157 amino acid position
Amino acid residues conservation profile demonstrating the extent of evolutionary conservation. The conservation scale ranged from 1-10, where 10 is the highly conserved and 1 the lowest. Protein regions associated with function tend to be highly conserved over the course of evolution

Figure 2B. Evolutionary conservation profile in *OsGPAT 8* demonstrating R355P amino acid position
Amino acid residues conservation profile demonstrating the extent of evolutionary conservation. The conservation scale ranged from 1-10, where 10 is the highly conserved and 1 the lowest. Protein regions associated with function tend to be highly conserved over the course of evolution

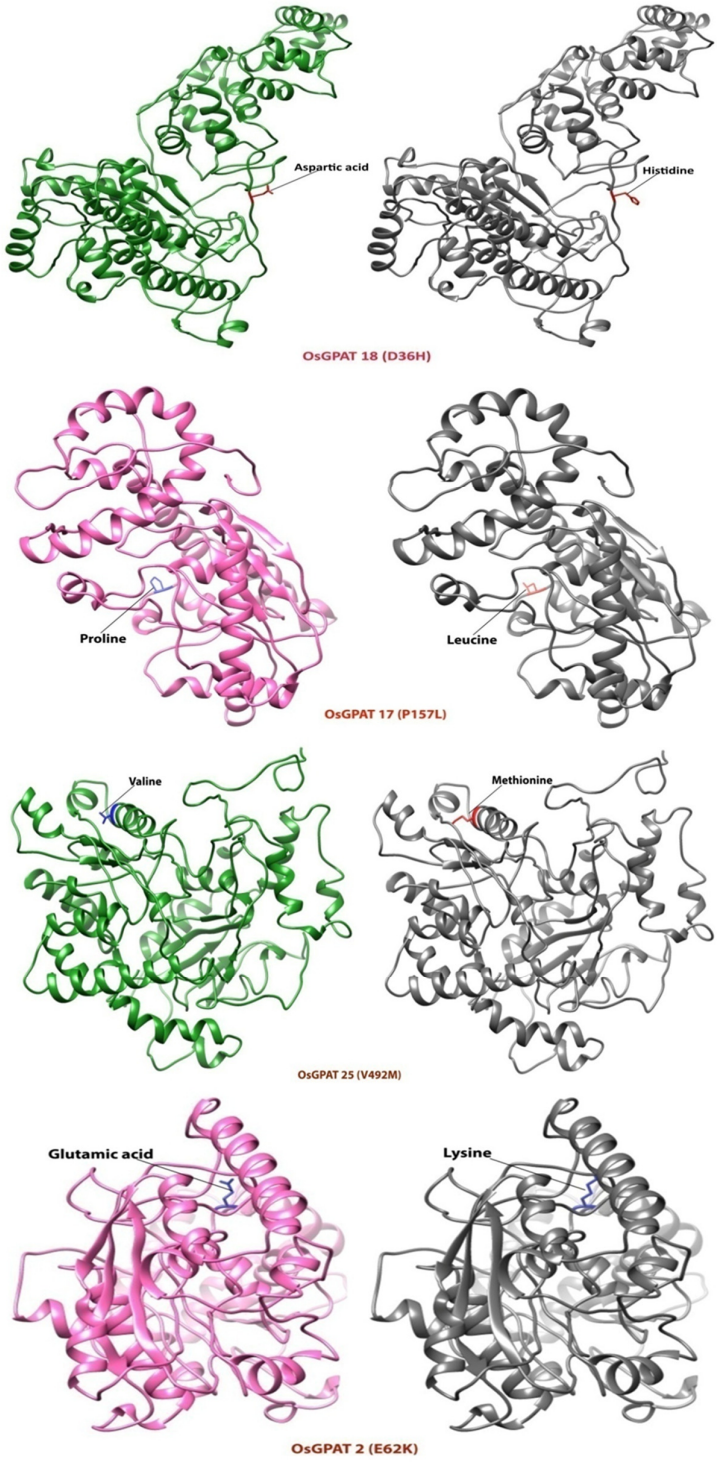

Figure 3. Protein 3D models of the rice *GPAT* proteins at native and variant positions generated by Phyre
Full-length models of *GPAT* proteins monomer having native residues on left sides and variant residues on the right side. Arrows indicate the specific location of substitution at amino acid residues derived from important SNPs. Every picture represents the substituted amino acid and the residual change in the protein structure. Each amino acid residue is demonstrated with color

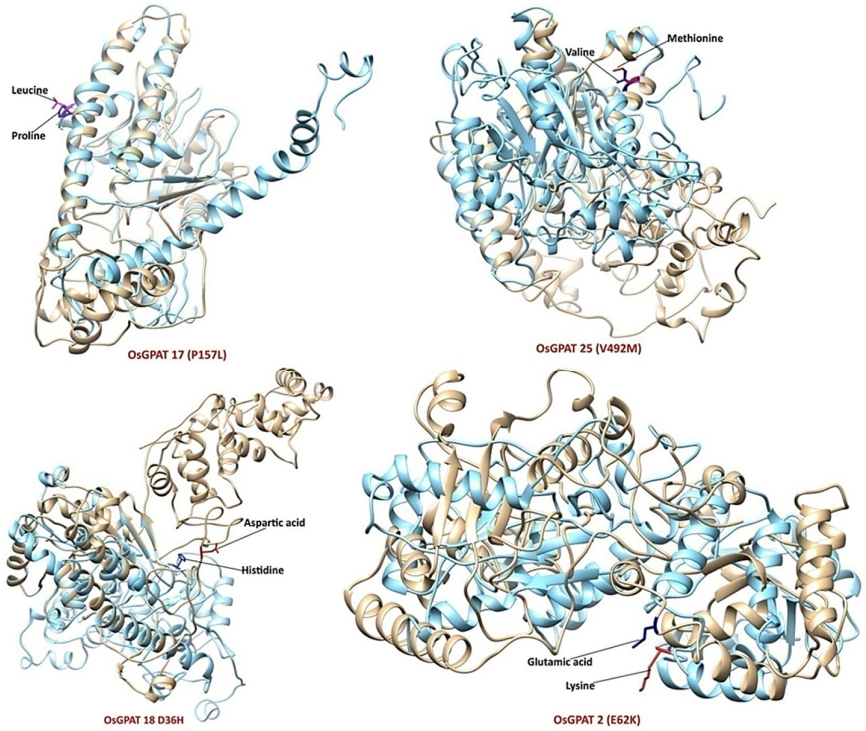

Figure 4. Superimposed 3D structures of selected native and variant rice *GPAT* proteins caused by the high-risk nsSNPS

Superimposed models were created by Chimera software. The arrows demonstrate the variant positions in superimposed protein models. Native and variant models are shown by blue and brown colors respectively

Post Translational modification prediction

Post translational modifications (PTM) influence protein conformation, stability, activity, localization and their interactions. They provide proteome diversity, impact signaling pathways, influence gene expression and enzyme kinetics (Piquerez *et al.*, 2014; Friso and van Wijk, 2015). Therefore, we estimated if nsSNPS resided on the PTM sites and they had any influence on PTM sites. We utilized many in silico tools to predict PTM sites in *GPAT* proteins. We used at least two *in silico* tools for each PTM event including the phosphorylation, methylation and ubiquitylation and selected positions predicted by both tools. We also explored whether any PTM site was conserved as well.

We investigated Phosphorylation sites by GPS 5.0 and NetPhos 3.1; both tools predicted 100 sites in various rice *GPAT* proteins (Table S2). 79 amino acid residues were Serine, 16 residues were Threonine and only 4 residues were Tyrosine. For methylation PTMs of lysine residues PSSMe and iMethyl-PseAAC tools were used. We considered those sites predicted by both these tools at 0.5 SVM Probability thresholds. There were 127 lysine residues that were common in both tools (Table S3). We used BDM-PUB and Ub-Pred to predict Ubiquitylation sites. 36 common amino acid sites were predicted for having a potential ubiquitylation site (Table S4).

We used Consurf results to explore the highly conserved PTM sites (Table S5). But we could not find any nsSNP coincided with the conserved PTM sites. Past studies have demonstrated a variation in the conserved PTM site affect protein function.

Discussion

The current study demonstrates various nonsynonymous SNPs could disturb rice *GPAT* protein structure and function. We filtered the nsSNPs (Table 1), shortlisted the high risk or deleterious SNPs (Table 2), and analysed their impact on protein structure and functions.

Initially, we find 1056 SNPs among 100 Chinese rice varieties in 26 *GPAT* genes (Table S6) as a model gene family, we identified in a genome-wide study (Safder *et al.*, 2021). We shortlisted 98 coding nsSNPs from these 1056 SNPs, that can change amino acid residues and alter protein function. We further used various computational tools including Provean, SIFT, Polyphen-2 to find significant non-synonymous SNPs having deleterious or potentially damaging effects to proteins. Since every computational tool uses different algorithms, their output could slightly differ; still, we find similarity in the predictions. For instance, many nsSNPs predicted as deleterious by an individual tool (Table 2), were often predicted by a second tool. Though these nsSNPs (predicted by a single tool in Table 2) can't be neglected and should be considered important, we only selected nsSNPs predicted by all the tools to strengthen our results. Hence, we found 13 SNPs (predicted by all three tools), considered as high-risk nsSNPs among the 98 nsSNPs (Table 3). We also explored functional PTM amino acid residues at phosphorylation, ubiquitylation and methylation sites, and investigated if the nsSNPs coincide with the PTM site; as modification in PTM sites can alter protein function.

SNPs can impact phenotypic diversity among different plant traits, influence gene expression, and are associated with various functions (Kharabian-Masouleh *et al.*, 2012; Wang *et al.*, 2018; Wang *et al.*, 2020b). Therefore, SNP identification and their functional elucidation provide a better understanding of their effects on gene functions and their phenotypes. Identifying functional SNPs is a complicated task in large population studies. Hence, it is beneficial to prior explore putative functional SNPs.

Past studies suggest nsSNPs influence different mechanisms. Biosynthesis of secondary metabolite, and signal transduction pathways are involved in fruit ripening and defense responses. A change in the cell wall structure and starch conversion to sugar, play roles in fruit ripening. Deleterious nsSNPs are involved in these signal transduction and metabolic pathways. An SNP in tomato invertase gene changed amino acid residue near proteins catalytic site and affected enzyme activity. In rice, 66 functional SNPs were discovered from 18 genes involved in starch biosynthesis. An important SNP was reported at the 1188 nucleotide position in Glucose-6-Phosphate Translocator 1 (*GPT1*), changed amino acid residues associated with amylase content. In another study, SNPs in several genes were identified involving in maize root development and associated with seedling root traits. Four functional SNPs in the *HSP17.8* gene in barley varieties were associated with agronomic traits. Similarly, in tomato, functional SNPs affect gene expression and are associated with phenotypic differences among the tomato lines. These examples demonstrate functional SNPs influence gene functions, reflecting the significance of the current study, as the identified nsSNPs could have a major effect on *GPAT* genes (Kharabian-Masouleh *et al.*, 2012; Hirakawa *et al.*, 2013; Seymour *et al.*, 2013; Xia *et al.*, 2013; Kumar *et al.*, 2014; Schreiber *et al.*, 2014; Bhardwaj *et al.*, 2016; Huq *et al.*, 2016; Zaynab *et al.*, 2018).

In past studies we find, in-silico approaches used in humans, plants, and identified functionally significant SNPs (De Alencar and Lopes, 2010; Kharabian, 2010; Kamaraj and Purohit, 2013; Arshad and Attya Bhatti, 2018; Zhang *et al.*, 2020). In rice, the Granule Bound Starch Synthase I *GBSSI* was used as a model plant gene and explored functional SNPs. The study identified a candidate SNP imparting a major impact on *GBSSI* and its phenotype. This nsSNP at exon 6, showed the highest effect on amylose content according to the SIFT prediction results (Kharabian, 2010), also reported in a previous study (Chen *et al.*, 2008), suggesting coherence between *in silico* analysis results and experimental approaches. Although experimental approaches provide strong evidence for estimating SNPs functional significance, they are tedious, lack exactness in phenotypic and genotypic data, so these experiments are not easy to detect SNPs (Cobb *et al.*, 2013). Besides, high false positive results crop up due to population structures and multiple testing. These results, therefore, mislead when a casual SNP is considered more significant in contrast to a truly casual variant (Tibbs Cortes *et al.*).

Computational methods are used to explore functional variants to reduce the burden from statistical association studies. Therefore, *in silico* tools provide an opportunity to facilitate experimental approaches and they can be used to recruit, identify and characterize functionally important SNPs having a major impact on phenotype (Tam *et al.*, 2019). In this regard, an *in-silico* study identified SNP diversity in cultivated and wild tomato genomes, investigated 1838 nsSNPs among 988 genes. There were 28 deleterious SNPs distributed among 27 genes predicted in hormonal and plant pathogen pathways similar to the current study. Further, they selected nsSNPs deleterious effect on the protein structure (Bhardwaj *et al.*, 2016). Likewise, a study explored important SNPs in the *TGF-β* receptor type 3 gene in chicken by using tools Sift, PANTHER, and I-mutant, and found a nsSNP in the coding region with deleterious impact, decreasing the protein stability(Rasal *et al.*, 2015). Another study predicted putative effects of SNPs on 58 *Prunus* rootstocks genes using SnpEff. The SNPs were categorized as a modifier, low, moderate, and high impact; high impact SNPs were further explored using *in silico* tools (Guajardo *et al.*, 2020).

Our results identified many high impact nsSNPs found in *GPAT*s conserved protein regions. For instance, nsSNPs including P157L, R355P, F128L, V549I, and A32V located at highly conserved positions (Table 5). Past studies have suggested mutations in conserved regions can affect protein functions. Mutations in the ZP domain in the *TGFBR3* gene affected its protein function; a conserved amino acid is essential for product and substrate specificity in triterpene synthases. Regarding variation between native and variant protein structures, we found most protein models had high RMSD value (two or more), reflecting differences between native and variant models affecting the protein stability (Jovine *et al.*, 2002; Han *et al.*, 2006; Salmon *et al.*, 2016; Islam *et al.*, 2019; Liao *et al.*, 2019; Bhardwaj and Purohit, 2020).

We could not find any PTM site coincided with the nsSNP, but many PTM sites were highly conserved. Past studies demonstrated PTM sites influence protein function, if there is a variation in their conserved positions, and affect protein stability or inter protein structures (Arif *et al.*, 2017; Gulzar *et al.*, 2017).

Conclusions

Briefly, in the study we identified all candidate SNPs most likely to affect rice *GPAT* proteins and influence their function. We demonstrated *in silico* tools could help us to characterize functional SNPs which possibly have potential impact on *GPAT* genes and related phenotypes. These functional SNPs could provide value in developing functional markers by associating their link to phenotypic traits. The study will also provide a computational pathway to find candidate SNPs in other rice gene families.

Authors' Contributions

Conceptualization, I.S., G.S., P.H. and S.T.; Data curation, I.S., G.S., P.H. and S.T.; Formal analysis, I.S., G.S., Z.S.; Funding acquisition, P.H. and S.T.; Methodology, I.S., G.S., Z.S.; Project administration, P.H. and S.T.; Resources, P.H. and S.T.; Supervision, P.H. and S.T.; Validation, Z.S.; Writing – original draft, I.S. and G.S.; Writing – review & editing, P.H. and S.T. All authors read and approved the final manuscript.

Acknowledgements

We appreciate the support of the Central Public-interest Scientific Institution Basal Research Fund of China (No. Y2020PT07) and the Chinese Academy of Agricultural Sciences (CAAS) Agricultural Sciences and Technologies Innovation Program for the financial assistance. We are also thankful to the China Scholarship Council (CSC) for providing the fellowship to Mr. Imran Safder (CSC No: 2017SLJ022683) during this research project.

Conflict of Interests

The authors declare that there are no conflicts of interest related to this article.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, ... Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nature Methods 7:248-249. *https://doi.org/10.1038/nmeth0410-248*

Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, ... Li Z (2015). SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Research 43:D1023-D1027. *https://doi.org/10.1093/nar/gku1039*

Arif R, Akram F, Jamil T, Mukhtar H, Lee SF, Saleem M (2017). Genetic variation and its reflection on posttranslational modifications in frequency clock and mating Type a-1 proteins in *Sordaria fimicola*. BioMed Research International 2017:1268623. *https://doi.org/10.1155/2017/1268623*

Arshad M, Attya Bhatti PJ (2018). Identification and in silico analysis of functional SNPs of human TAGAP protein: A comprehensive study. PloS One 13(1):e0188143. *https://doi.org/10.1371/journal.pone.0188143*

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Research 44:W344-W350. *https://doi.org/10.1093/nar/gkw408*

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Research 38:W529-W533. *https://doi.org/10.1093/nar/gkq399*

Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N (2004). ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 20:1322-1324. *https://doi.org/10.1093/bioinformatics/bth070*

Bhardwaj A, Dhar YV, Asif MH, Bag SK (2016). *In silico* identification of SNP diversity in cultivated and wild tomato species: insight from molecular simulations. Scientific Reports 6:38715-38715. *https://doi.org/10.1038/srep38715*

Bhardwaj VK, Purohit R (2020). Structural changes induced by substitution of amino acid 129 in the coat protein of Cucumber mosaic virus. Genomics 112:3729-3738. *https://doi.org/10.1016/j.ygeno.2020.04.023*

Blom N, Gammeltoft S, Brunak S (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology 294:1351-1362. *https://doi.org/10.1006/jmbi.1999.3310*

Carugo O, Pongor S (2001). A normalized root-mean-spuare distance for comparing protein three-dimensional structures. Protein Science 10:1470-1473. *https://doi.org/10.1110/ps.690101*

Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, Mayrose I, Pupko T, Ben-Tal N (2013). ConSurf: using evolutionary data to raise testable hypotheses about protein function. Israel Journal of Chemistry 53:199-206. *https://doi.org/10.1002/ijch.201200096*

Chaisan T, Van K, Kim MY, Kim KD, Choi B-S, Lee S-H (2012). *In silico* single nucleotide polymorphism discovery and application to marker-assisted selection in soybean. Molecular Breeding 29:221-233. *https://doi.org/10.1007/s11032-010-9541-y*

Chen M-H, Bergman CJ, Pinson SRM, Fjellstrom RG (2008). Waxy gene haplotypes: Associations with pasting properties in an international rice germplasm collection. Journal of Cereal Science 48:781-788. *https://doi.org/10.1016/j.jcs.2008.05.004*

Choi Y, Chan AP (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics (Oxford, England) 31:2745-2747. *https://doi.org/10.1093/bioinformatics/btv195*

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012). Predicting the functional effect of amino acid substitutions and indels. PloS One 7. *https://doi.org/10.1371/journal.pone.0046688*

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013). Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theoretical and Applied Genetics 126:867-887. *https://doi.org/10.1007/s00122-013-2066-0*

De Alencar S, Lopes JC (2010). A comprehensive in silico analysis of the functional and structural impact of SNPs in the IGF1R gene. BioMed Research International 715139. *https://doi.org/10.1155/2010/715139*

Deller MC, Kong L, Rupp B (2016). Protein stability: a crystallographer's perspective. Acta Crystallographica Section F: Structural Biology Communications 72:72-95. *https://doi.org/10.1107/S2053230X15024619*

Friso G, van Wijk KJ (2015). Posttranslational protein modifications in plant metabolism. Plant Physiology 169:1469-1487. *https://doi.org/10.1107/S2053230X15024619*

Gailing O, Vornam B, Leinemann L, Finkeldey R (2009). Genetic and genomic approaches to assess adaptive genetic variation in plants: forest trees as a model. Physiologia Plantarum 137:509-519. *https://doi.org/10.1111/j.1399-3054.2009.01263.x*

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92-100. *https://doi.org/10.1126/science.1068275*

Guajardo V, Solís S, Almada R, Saski C, Gasic K, Moreno MÁ (2020). Genome-wide SNP identification in Prunus rootstocks germplasm collections using genotyping-by-sequencing: phylogenetic analysis, distribution of SNPs and prediction of their effect on gene function. Scientific Reports 10:1467-1467. *https://doi.org/10.1038/s41598-020-58271-5*

Gulzar N, Dingerdissen H, Yan C, Mazumder R (2017). Impact of nonsynonymous single-nucleotide variations on post-translational modification sites in human proteins. Protein Bioinformatics, Springer, pp 159-190. *https://doi.org/10.1007/978-1-4939-6783-4_8*

Han JH, Kerrison N, Chothia C, Teichmann SA (2006). Divergence of interdomain geometry in two-domain proteins. Structure (London, England, 1993) 14:935-945. *https://doi.org/10.1016/j.str.2006.01.016*

Hirakawa H, Shirasawa K, Ohyama A, Fukuoka H, Aoki K, Rothan C, Sato S, Isobe S, Tabata S (2013). Genome-wide SNP genotyping to infer the effects on gene functions in tomato. DNA Research 20:221-233. *https://doi.org/10.1093/dnares/dst005*

Huq MA, Akter S, Nou IS, Kim HT, Jung YJ, Kang KK (2016). Identification of functional SNPs in genes and their effects on plant phenotypes. Journal of Plant Biotechnology 43:1-11. *https://doi.org/10.5010/JPB.2016.43.1.1*

Islam MJ, Khan AM, Parves MR, Hossain MN, Halim MA (2019). Prediction of deleterious non-synonymous SNPs of human STK11 gene by combining algorithms, molecular docking, and molecular dynamics simulation. Scientific Reports 9:16426. *https://doi.org/10.1038/s41598-019-52308-0*

Jackson SA (2016). Rice: the first crop genome. Rice 9:1-3. *https://doi.org/10.1186/s12284-016-0087-4*

Jiang D, Ye Q-l, Wang F-S, Cao L (2010). The mining of citrus EST-SNP and its application in cultivar discrimination. Agricultural Sciences in China 9:179-190. *https://doi.org/10.1016/S1671-2927(09)60082-1*

Jovine L, Qi H, Williams Z, Litscher E, Wassarman PM (2002). The ZP domain is a conserved module for polymerization of extracellular proteins. Nature Cell Biology 4:457-461. *https://doi.org/10.1038/ncb802*

Kamaraj B, Purohit R (2013). *In silico* screening and molecular dynamics simulation of disease-associated nsSNP in TYRP1 gene and its structural consequences in OCA3. BioMed Research International 2013:697051-697051. *https://doi.org/10.1155/2013/697051*

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10:845. *https://doi.org/10.1107/S2053230X15024619*

Kharabian-Masouleh A, Waters DLE, Reinke RF, Ward R, Henry RJ (2012). SNP in starch biosynthesis genes associated with nutritional and functional properties of rice. Scientific Reports 2:557. *https://doi.org/10.1038/srep00557*

Kharabian A (2010). An efficient computational method for screening functional SNPs in plants. Journal of Theoretical Biology 265:55-62. *https://doi.org/10.1016/j.jtbi.2010.04.017*

Korani W, Clevenger JP, Chu Y, Ozias-Akins P (2019). Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. Plant Genome 12. *https://doi.org/10.3835/plantgenome2018.05.0023*

Kumar B, Abdel-Ghani AH, Pace J, Reyes-Matamoros J, Hochholdinger F, Lübberstedt T (2014). Association analysis of single nucleotide polymorphisms in candidate genes with root traits in maize (*Zea mays* L.) seedlings. Plant Science 224:9-19. *https://doi.org/10.1007/s11103-015-0314-1*

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, ... MacArthur DG (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285-291. *https://doi.org/10.1038/nature19057*

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P (2009a). Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744-2750. *https://doi.org/10.1093/bioinformatics/btp528*

Li J-Y, Wang J, Zeigler RS (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience 3:2047-217X. *https://doi.org/10.1186/2047-217X-3-8*

Li X, Gao X, Ren J, Jin C, Xue Y (2009b). BDM-PUB: Computational prediction of protein ubiquitination sites with a Bayesian discriminant method. *https://doi.org/10.2174/138920291966619101409125O*

Liao M-l, Somero GN, Dong Y-W (2019). Comparing mutagenesis and simulations as tools for identifying functionally important sequence changes for protein thermal adaptation. Proceedings of the National Academy of Sciences 116:679-688. *https://doi.org/10.1073/pnas.1817455116*

Majeed S, Rana IA, Atif RM, Ali Z, Hinze L, Azhar MT (2019). Role of SNPs in determining QTLs for major traits in cotton. Journal of Cotton Research 2:5. *https://doi.org/10.1186/s42397-019-0022-5*

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012). SNP markers and their impact on plant breeding. International Journal of Plant Genomics 2012:728398. *https://doi.org/10.1155/2012/728398*

Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, ... Alexandrov N (2016). Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Research 45:D1075-D1081. *https://doi.org/10.1093/nar/gkw1135*

McCouch SR, Zhao K, Wright M, Tung C-W, Ebana K, Thomson M, Reynolds A, Wang D, DeClerck G, Ali ML (2010). Development of genome-wide SNP assays for rice. Breeding Science 60:524-535. *https://doi.org/10.1270/jsbbs.60.524*

Nelson MR, Marnellos G, Kammerer S, Hoyal CR, Shi MM, Cantor CR, Braun A (2004). Large-scale validation of single nucleotide polymorphisms in gene regions. Genome Research 14:1664-1668. *https://doi.org/10.1101/gr.2421604*

Ng PC, Henikoff S (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research 31:3812-3814. *https://doi.org/10.1093/nar/gkg509*

Ortbauer M, Vahdati K, Leslie C (2013). Abiotic stress adaptation: protein folding stability and dynamics. Abiotic Stress-Plant Responses and Applications in Agriculture 1:3-25. *https://doi.org/10.5772/53129*

Pea G, Aung HH, Frascaroli E, Landi P, Pè ME (2013). Extensive genomic characterization of a set of near-isogenic lines for heterotic QTL in maize (*Zea mays* L.). BMC Genomics 14:61. *https://doi.org/10.1186/1471-2164-14-61*

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004). UCSF Chimera–A visualization system for exploratory research and analysis. Journal of Computational Chemistry 25:1605-1612. *https://doi.org/10.1002/jcc.20084*

Piquerez SJ, Balmuth AL, Sklenář J, Jones AM, Rathjen JP, Ntoukakis V (2014). Identification of post-translational modifications of plant protein complexes. JoVE Journal of Visualized Experiments e51095. *https://doi.org/10.3791/51095*

Qiu W-R, Xiao X, Lin W-Z, Chou K-C (2014). iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. BioMed Research International 947416. *https://doi.org/10.1155/2014/947416*

Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM (2010). Identification, analysis, and prediction of protein ubiquitination sites. Proteins: Structure, Function, and Bioinformatics 78:365-380. *https://doi.org/10.1002/prot.22555*

Rasal KD, Shah TM, Vaidya M, Jakhesara SJ, Joshi CG (2015). Analysis of consequences of non-synonymous SNP in feed conversion ratio associated TGF-β receptor type 3 gene in chicken. Meta Gene 4:107-117. *https://doi.org/10.1016/j.mgene.2015.03.006*

Safder I, Shao G, Sheng Z, Hu P, Tang S (2021). Identification and analysis of the structure, expression and nucleotide polymorphism of the GPAT gene family in rice. Plant Gene 100290. *https://doi.org/10.1016/j.plgene.2021.100290*

Salmon M, Thimmappa RB, Minto RE, Melton RE, Hughes RK, O'Maille PE, Hemmings AM, Osbourn A (2016). A conserved amino acid residue critical for product and substrate specificity in plant triterpene synthases. Proceedings of the National Academy of Sciences 113:E4407-E4414. *https://doi.org/10.1073/pnas.1605509113*

Sandhu D, Pudussery MV, Kumar R, Pallete A, Markley P, Bridges WC, Sekhon RS (2020). Characterization of natural genetic variation identifies multiple genes involved in salt tolerance in maize. Functional & Integrative Genomics 20:261-275. *https://doi.org/10.1007/s10142-019-00707-x*

Schreiber L, Nader-Nieto AC, Schönhals EM, Walkemeier B, Gebhardt C (2014). SNPs in genes functional in starch-sugar interconversion associate with natural variation of tuber starch and sugar content of potato (*Solanum tuberosum* L.). G3 Genes, Genomes, Genetics 4:1797-1811. *https://doi.org/10.1534/g3.114.012377*

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, ... Hassabis D (2020). Improved protein structure prediction using potentials from deep learning. Nature 577:706-710. *https://doi.org/10.1038/s41586-019-1923-7*

Seymour GB, Chapman NH, Chew BL, Rose JK (2013). Regulation of ripening and opportunities for control in tomato and other fruits. Plant Biotechnology Journal 11:269-278. *https://doi.org/10.1111/j.1467-7652.2012.00738.x*

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics 20:467-484. *https://doi.org/10.1038/s41576-019-0127-1*

Tibbs Cortes L, Zhang Z, Yu J (2021). Status and prospects of genome-wide association studies in plants. The Plant Genome 14:e20077. *https://doi.org/10.1002/tpg2.20077*

Wang C-C, Yu H, Huang J, Wang W-S, Faruquee M, Zhang F, ... Zheng T-Q (2020a). Towards a deeper haplotype mining of complex traits in rice with RFGB v2.0. Plant Biotechnology Journal 18:14-16. *https://doi.org/10.1111/pbi.13215*

Wang H, Ham T-H, Im D-E, Lar SM, Jang S-G, Lee J, Mo Y, Jeung J-U, Kim ST, Kwon S-W (2020b). A new SNP in rice gene encoding pyruvate phosphate dikinase (PPDK) associated with floury endosperm. Genes (Basel) 11:465. *https://doi.org/10.3390/genes11040465*

Wang H, Mo Y-J, Im D-E, Jang S-G, Ham T-H, Lee J, Jeung J-U, Kwon S-W (2018). A new SNP in cy OsPPDK gene is associated with floury endosperm in Suweon 542. Molecular Genetics and Genomics 293:1151-1158. *https://doi.org/10.3390/genes11040465*

Wen P-P, Shi S-P, Xu H-D, Wang L-N, Qiu J-D (2016). Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. Bioinformatics 32:3107-3115. *https://doi.org/10.1093/bioinformatics/btw377*

Withana WVE, Kularathna RMRE, Kottearachchi NS, Kekulandara DS, Weerasena J, Steele KA (2020). *In silico* analysis of the fragrance gene (badh2) in Asian rice (*Oryza sativa* L.) germplasm and validation of allele specific markers. Plant Genetic Resources: Characterization and Utilization 18:71-80. *https://doi.org/10.1017/S1479262120000015*

Xia Y, Li R, Ning Z, Bai G, Siddique KH, Yan G, Baum M, Varshney RK, Guo P (2013). Single nucleotide polymorphisms in HSP17. 8 and their association with agronomic traits in barley. PloS One 8:e56816. *https://doi.org/10.1371/journal.pone.0056816*

Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X (2005). GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Research 33:W184-W187. *https://doi.org/10.1093/nar/gki393*

Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D (2020). Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences 117:1496-1503. *https://doi.org/10.1073/pnas.1914677117*

Yang W, Bai X, Kabelka E, Eaton C, Kamoun S, van der Knaap E, David F (2004). Discovery of single nucleotide polymorphisms in Lycopersicon esculentum by computer aided analysis of expressed sequence tags. Molecular Breeding 14:21-34. *https://doi.org/10.1023/B:MOLB.0000037992.03731.a5*

Zaynab M, Fatima M, Abbas S, Sharif Y, Umair M, Zafar MH, Bahadar K (2018). Role of secondary metabolites in plant defense against pathogens. Microbial Pathogenesis 124:198-202. *https://doi.org/10.1016/j.micpath.2018.08.034*

Zhang M, Huang C, Wang Z, Lv H, Li X (2020). *In silico* analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in the human GJA3 gene associated with congenital cataract. BMC Molecular and Cell Biology 21:12. *https://doi.org/10.1186/s12860-020-00252-7*

Zhang W, Mirlohi S, Li X, He Y (2018). Identification of functional single-nucleotide polymorphisms affecting leaf hair number in *Brassica rapa*. Plant Physiology 177:490-503. *https://doi.org/10.1104/pp.18.00025*

Zhang Y, Skolnick J (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33:2302-2309. *https://doi.org/10.1093/nar/gki524*

Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, ... Xie W (2015). RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Research 43:D1018-D1022. *https://doi.org/10.1093/nar/gku894*

Supplementary Files

Table S1. 100 rice accessions from different regions of China using 3000 genome project data

Accession number/ Sequence ID in 3KRG data	Variety name/Source	Indica/Japonica
B060	Aijiaonante	*Indica*
B061	Guangluai 4	*Indica*
B062	Nantehao	*Indica*
B064	Xiangzaoxian 7	*Indica*
B067	Funingzipigengzi	*Indica*
B070	Laoguangtou 83	*japonica*
B072	Qiuqianbai	*Indica*
B073	Jinxibai2	*Indica*
B074	Taishannuo	*Indica*
B075	Jinbaoyin	*Indica*
B076	Minbeiwanxian	*Indica*
B079	Esiniu	*Indica*
B081	Heidu 4	*Indica*
B082	Qiyuexian	*Indica*
B083	Dongtingwanxian	*Indica*
B101	Yangkenuo	*japonica*
B105	Baoxuan 21	*Indica*
B106	Wenxiangnuo	*Indica*
B108	Xianggu	*Indica*
B112	Liusha 1	*Indica*
B113	Chenwan 3	*Indica*
B118	Ergangai	*Indica*
B119	Guangluai 15	*Indica*
B120	Hongwan 1	*Indica*
B123	Zaoshuxiangheimi	*Indica*
B124	Geng 87-304	*japonica*
B126	Zaoxian 240	*Indica*
B141	Geng 7623	*Indica*
B142	Ninghui 21	*japonica*
B143	76-1	*japonica*
B147	Baikehanhe	*Indica*
B149	Haoxiang	*Indica*
B151	Jinnante 43B	*Indica*
B152	Zaoshunonghu 6	*japonica*
B158	Taizhong 65	*japonica*
B162	Baigedao	*japonica*
B163	Liushizao	*Indica*
B164	Qingke	*Aus*
B165	Haohuangla	*Indica*
B196	Taidongludao	*japonica*
B197	Taizhongxianxuan 2	*Indica*
B198	Jiefangxian	*Indica*
B199	Hongmisandan	*japonica*
B200	Jinyou 1	*Indica*
B202	Baoxie 123B	*Indica*
B203	Biwusheng	*Indica*
B204	Longhuamaohu	*japonica*
B205	Cunsanli	*japonica*
B207	Aihechi	*Indica*
B222	Beizinuo	*Indica*
B223	Cungunuo	*japonica*
B224	Younian	*Indica*
B225	Guantuibaihe 1	*japonica*
B226	Heimangdao	*japonica*
B227	Menjiagao 1	*Indica*

B228	Haobayong 1	*japonica*
B229	Menjiading 2	*Indica*
B230	Banjiemang2	*japonica*
B232	Xiangaizao 10hao	*Indica*
B233	Xiangwanxian 1	*Indica*
B235	Zhonghua 8	*Indica*
B236	Jindao 1	*Indica*
B238	Momi	*Indica*
B239	Zhendao 232	*Indica*
B240	Zhengdao 5	*japonica*
B241	Lamujia	*japonica*
B242	Gui 630	*Indica*
B243	Huhui 91269	*Aus*
B244	Xiangdao	*Indica*
B246	Laozaogu	*Indica*
B247	Jinnante B	*Indica*
B248	Zhuzhen B	*Indica*
B249	Chaoyangyihao B	*Indica*
B250	Annongwangeng B	*japonica*
B253	Jiangnongzao 1 B	*Indica*
B252	Xiangai B	*Indica*
IRIS_313-10459	Pi 160862-1	*japonica*
IRIS_313-10562	Seng-Chui-Lin	*japonica*
IRIS_313-11654	PL 3165	*Indica*
IRIS_313-11664	Cun Gu Nuo	*Indica*
IRIS_313-11665	Jin Hua 258	*Indica*
IRIS_313-11666	Long Ge 33	*Indica*
IRIS_313-11667	Luo Ai Zao 3	*Indica*
IRIS_313-11668	Rong Dao 4	*Indica*
IRIS_313-11669	F 478	*Indica*
IRIS_313-11726	Guang Qing 334	*Indica*
IRIS_313-11727	Hong Yang Zao 3	*Indica*
IRIS_313-11728	Luo Si Zhan	*Indica*
IRIS_313-11729	Mei Liu Zao 5	*Indica*
IRIS_313-11730	Qing Er Xiao 2	*Indica*
IRIS_313-11731	Qing Tai Ai	*Indica*
IRIS_313-11732	Qing Zao 3	*Indica*
IRIS_313-11733	Shuang Bai Ai 2	*Indica*
IRIS_313-11734	Si Chao 1	*Indica*
IRIS_313-11735	Yi Li Zhong	*japonica*
IRIS_313-11744	Ai Jiao Ao Fan Zi	*Indica*
IRIS_313-11745	An Fu Zhan	*Indica*
IRIS_313-11746	E 2070	*Indica*
IRIS_313-11747	E 4197	*japonica*
IRIS_313-11748	Gao Jiao Ying Gan Zhan	*Indica*

Table S2. Putative phosphorylation sites in rice *GPAT* genes predicted in both NetPhos 3.1 and GPS 3.0

	GPS 5.0			NetPhos 3.1	
Gene ID	Position	GPS 5.0 (kinase)	Score	NetPhos 3.1(Kinase)	Residue phosphorylated
OsGPAT 1	28	AGC/AKT/AKT1	5.002	0.834 PKA	Serine (S)
	170	AGC/AKT/AKT1	1.263	0.959 unsp	Serine (S)
	451	AGC/AKT/AKT1	0.616	0.968 unsp	Serine (S)
	497	AGC/AKT/AKT1	0.446	0.983 unsp	Serine (S)
OsGPAT 2	173	AGC/AKT/AKT1	3.205	0.975 unsp	Serine (S)
OsGPAT 3	42	AGC/AKT/AKT1	3.919	0.993 unsp	Serine (S)
	43	AGC/AKT/AKT1	0.91	0.917 unsp	Serine (S)
	102	AGC/AKT/AKT1	0.435	0.853 unsp	Threonine(T)
	144	AGC/AKT/AKT1	0.662	0.989 unsp	Serine (S)
	254	AGC/AKT/AKT1	1.457	0.977 unsp	Serine (S)
OsGPAT 4	37	AGC/AKT/AKT1	1.838	0.964 unsp	Serine (S)
	207	AGC/AKT/AKT1	2.327	0.699 PKB	Threonine(T)
	267	AGC/AKT/AKT1	0.857	0.995 unsp	Serine (S)
	438	AGC/AKT/AKT1	0.951	0.968 unsp	Serine (S)
OsGPAT 5	82	AGC/AKT/AKT1	1.127	0.803 unsp	Threonine(T)
OsGPAT 6	35	AGC/AKT/AKT1	2.015	0.686 PKG	Serine (S)
	216	AGC/AKT/AKT1	4.289	0.803 unsp	Serine (S)
	324	AGC/AKT/AKT1	1.68	0.779 PKA	Serine (S)
	478	AGC/AKT/AKT1	3.246	0.772 unsp	Serine (S)
	489	AGC/AKT/AKT1	4.461	0.874 unsp	Serine (S)
OsGPAT 7	56	AGC/AKT/AKT1	0.733	0.975 unsp	Threonine(T)
	61	AGC/AKT/AKT1	1.095	0.919 unsp	Serine (S)
OsGPAT 8	168	AGC/AKT/AKT1	2.299	0.822 unsp	Threonine(T)
	283	AGC/AKT/AKT1	1.257	0.819 PKC	Threonine(T)
	412	AGC/AKT/AKT1	1.166	0.801 unsp	Threonine(T)
	19	AGC/AKT/AKT1	1.897	0.500 EGFR	Tyrosine(Y)
OsGPAT 9	75	AGC/AKT/AKT1	1.17	0.661 unsp	Serine (S)
	343	AGC/AKT/AKT1	0.557	0.632 unsp	Serine (S)
OsGPAT 10	23	AGC/AKT/AKT1	3.744	0.678 PKA	Serine (S)
	43	AGC/AKT/AKT1	1.2	0.995 unsp	Serine (S)
	436	AGC/AKT/AKT1	0.44	0.989 unsp	Serine (S)
OsGPAT 11	12	AGC/AKT/AKT1	0.48	0.701 PKC	Serine (S)
	196	AGC/AKT/AKT1	1.374	0.772 unsp	Threonine(T)
	233	AGC/AKT/AKT1	2.784	0.832 PKA	Serine (S)
	277	AGC/AKT/AKT1	1.982	0.996 unsp	Serine (S)
	292	AGC/AKT/AKT1	4.433	0.725 unsp	Serine (S)
OsGPAT 12	67	AGC/AKT/AKT1	1.015	0.705 unsp	Serine (S)
OsGPAT 13	34	AGC/AKT/AKT1	1.277	0.984 unsp	Serine (S)
	334	AGC/AKT/AKT1	2.973	0.996 unsp	Serine (S)
	423	AGC/AKT/AKT1	0.554	0.968 unsp	Serine (S)
	241	AGC/AKT/AKT1	1.828	0.972 unsp	Tyrosine(Y)
	521	AGC/AKT/AKT1	2.286	0.989 unsp	Tyrosine(Y)
OsGPAT 14	245	AGC/AKT/AKT1	0.726	0.991 unsp	Serine (S)
	314	AGC/AKT/AKT1	0.932	0.996 unsp	Serine (S)
	394	AGC/AKT/AKT1	0.961	0.590 DNAPK	Serine (S)
OsGPAT 15	115	AGC/AKT/AKT1	0.533	0.548 PKA	Serine (S)
	272	AGC/AKT/AKT1	1.904	0.688 PKA	Serine (S)
	393	AGC/AKT/AKT1	1.388	0.835 PKA	Serine (S)
	458	AGC/AKT/AKT1	0.46	0.960 unsp	Serine (S)
	476	AGC/AKT/AKT1	0.781	0.527 CKI	Serine (S)
OsGPAT 16	18	AGC/AKT/AKT1	3.423	0.995 unsp	Serine (S)

	159	AGC/AKT/AKT1	4.659	0.909 unsp	Threonine(T)
	195	AGC/AKT/AKT1	1.719	0.995 unsp	Serine (S)
	392	AGC/AKT/AKT1	0.412	0.632 unsp	Serine (S)
OsGPAT 17	265	AGC/AKT/AKT1	2.07	0.61 PKC	Threonine(T)
	80	AGC/AKT/AKT1	1.012	0.981 unsp	Serine (S)
	397	AGC/AKT/AKT1	1.947	0.711 PKC	Serine (S)
	11	AGC/AKT/AKT1	2.454	0.61 unsp	Tyrosine (Y)
OsGPAT 18	7	AGC/AKT/AKT1	2.052	0.987 unsp	Serine (S)
	74	AGC/AKT/AKT1	2.432	0.998 unsp	Serine (S)
	83	AGC/AKT/AKT1	0.531	0.705 unsp	Tyrosine (Y)
	532	AGC/AKT/AKT1	0.501	0.696 PKC	Serine (S)
OsGPAT 19	348	AGC/AKT/AKT1	4.263	0.998 unsp	Serine (S)
	154	AGC/AKT/AKT1	1.459	0.912 unsp	Serine (S)
	284	AGC/AKT/AKT1	1.462	0.732 PKA	Serine (S)
	362	AGC/AKT/AKT1	0.991	0.742 PKC	Serine (S)
OsGPAT 20	10	AGC/AKT/AKT1	2.668	0.794 PKA	Serine (S)
	70	AGC/AKT/AKT1	0.599	0.917 unsp	Threonine(T)
	293	AGC/AKT/AKT1	1.168	0.776 unsp	Threonine(T)
	376	AGC/AKT/AKT1	0.768	0.968 unsp	Serine (S)
OsGPAT 21	65	AGC/AKT/AKT1	5.53	0.988 unsp	Threonine(T)
	219	AGC/AKT/AKT1	0.983	0.974 unsp	Serine (S)
OsGPAT 22	343	AGC/AKT/AKT1	2.76	0.700 PKA	Serine (S)
	10	AGC/AKT/AKT1	1.458	0.770 PKB	Threonine(T)
	256	AGC/AKT/AKT1	0.588	0.641 PKA	Serine (S)
	368	AGC/AKT/AKT1	0.599	0.968 unsp	Serine (S)
OsGPAT 23	268	AGC/AKT/AKT1	1.005	0.835 unsp	Threonine(T)
	270	AGC/AKT/AKT1	0.736	0.988 unsp	Serine (S)
	282	AGC/AKT/AKT1	1.173	0.802 PKA	Serine (S)
	415	AGC/AKT/AKT1	0.549	0.882 unsp	Serine (S)
	418	AGC/AKT/AKT1	0.873	0.908 unsp	Serine (S)
OsGPAT 24	38	AGC/AKT/AKT1	2.222	0.982 unsp	Serine (S)
	121	AGC/AKT/AKT1	0.474	0.954 unsp	Serine (S)
	362	AGC/AKT/AKT1	2.503	0.950 unsp	Serine (S)
	363	AGC/AKT/AKT1	6.311	0.998 unsp	Serine (S)
OsGPAT 25	7	AGC/AKT/AKT1	1.334	0.951 unsp	Serine (S)
	12	AGC/AKT/AKT1	1.545	0.550 PKC	Serine (S)
	29	AGC/AKT/AKT1	1.819	0.996 unsp	Serine (S)
	249	AGC/AKT/AKT1	0.435	0.706 PKA	Serine (S)
	365	AGC/AKT/AKT1	1.437	0.991 unsp	Serine (S)
	390	AGC/AKT/AKT1	1.017	0.998 unsp	Serine (S)
	423	AGC/AKT/AKT1	0.68	0.873 unsp	Serine (S)
	489	AGC/AKT/AKT1	1.603	0.880 unsp	Serine (S)
	535	AGC/AKT/AKT1	1.174	0.952 unsp	Serine (S)
OsGPAT 26	65	AGC/AKT/AKT1	2.422	0.980 unsp	Serine (S)
	156	AGC/AKT/AKT1	3.563	0.991 unsp	Serine (S)
	157	AGC/AKT/AKT1	3.454	0.998 unsp	Serine (S)
	353	AGC/AKT/AKT1	2.858	0.956 unsp	Threonine(T)
	387	AGC/AKT/AKT1	0.996	0.991 unsp	Serine (S)
	445	AGC/AKT/AKT1	1.17	0.968 unsp	Serine (S)

Table S3. Putative methylation sites of rice *GPAT* genes predicted by PSSMe and iMethyl-PseAAAC toolss

		PSSMe		iMethyl-PseAAC
Gene Id	Position	Flanking residue	SVM Probability	Methylation position site
Os*GPAT*1	292	PPGGECGGV-K-PLVFHDGRL	0.80542	292
	483	HGTSTTPAA-**K**-WMDPFYFMM	0.89011	483
	495	DPFYFMMNP-**K**-PSYRVEFLP	0.8908	495
	545	ELTGMTRKD-**K**-YMMLAGNEG	0.62344	545
Os*GPAT*2	44	MVMVCFFGL-**K**-EKKVARVAR	0.7091	44
	46	MVCFFGLKE-**K**-KVARVARAA	0.56021	46
	91	SRVIPRVMV-**K**-PFLEDYLGV	0.53274	91
	182	SPLPRDQYP-**K**-PLVFHDGRL	0.73319	182
	251	VINSPVGQA-**K**-ADHPRNPKG	0.64146	251
	259	AKADHPRNP-**K**-GHLYVCNHR	0.57032	259
	438	ECTKFTREN-**K**-YLALAGNRG	0.54876	438
Os*GPAT*3	175	FHGTTAGGW-**K**-LLDPLYLLM	0.75545	175
	265	TRTALTRRD-**K**-YLALTGNDG	0.5	265
Os*GPAT*4	9	OMVLPTILP-**K**-IAAHWLFTF	0.6076	9
	65	TVVFPDAAD-**K**-AVVFGFDGA	0.53222	65
	145	DLVARAVLP-**K**-FYMEGLNAQ	0.86769	145
	280	ARLPRDRYP-**K**-PLIFHDGRL	0.78427	280
	325	ISIGILLPY-**K**-ISFGAGALF	0.72122	325
Os*GPAT*5	147	LRKGYLGYI-**K**-YILKSSLMK	0.66405	147
	151	YLGYIKYIL-**K**-SSLMKLPVF	0.81681	151
	175	IFEFIPVER-**K**-WEIDEAIIQ	0.61441	175
	191	IIQNKLSAF-**K**-DPRDPLWLA	0.69038	191
	229	ASEHGLPIL-**K**-NVLLPKTKG	0.86028	229
	362	LSLFSSVWF-**K**-VYVLLSCAY	0.69629	362
Os*GPAT*6	102	EMVARSVLP-**K**-FYAEDVHPE	0.54789	102
	121	SWRVFNSFG-**K**-RYIITASPR	0.64426	121
	155	VVGTELEVG-**K**-NGKATGFMV	0.89347	155
	277	PERIVFYTY-**K**-LMGIRLIVK	0.83839	277
	286	KLMGIRLIV-**K**-GNPPPPPKK	0.9418	286
	294	VKGNPPPPP-**K**-KGHPGVLFV	0.93903	294
	295	KGNPPPPPK-**K**-GHPGVLFVC	0.95441	295
	414	FHGSTVRGF-**K**-LMDPYFFFM	0.71944	414
Os*GPAT*7	97	SSSALRFYR-**K**-KVGKEVDGI	0.63366	97
	98	SSALRFYRK-**K**-VGKEVDGIE	0.78182	98
	213	ILSTFFRSV-**K**-VLPVSRGDG	0.62596	213
	255	EGSRSKDGG-**K**-TVAPAKRGV	0.67056	255
	293	GMQDIMPVG-**K**-RIPRAGKRV	0.81575	293
	404	EPSDVQEPL-**K**-KAKPVLHLE	0.62035	404
Os*GPAT*8	212	SDHDFMAIC-**K**-EAYMVPKNK	0.736	212
	219	ICKEAYMVP-**K**-NKRAPRAAA	0.65467	219
	221	KEAYMVPKN-**K**-RAPRAAADE	0.5874	221
Os*GPAT*9	183	RPVPREELP-**K**-PVVFHDGRL		183
	195	VFHDGRLVQ-**K**-PSPALALLT		195
	374	FHGTTARGW-**K**-ALDPFYFFM		374
	440	SYECTSFTR-**K**-DKYRALAGN		440
Os*GPAT*10	170	PREMVEPFL-**K**-EYLAVDAVV	0.50755	170
	467	FHGTTAGGW-**K**-MLDPFFLLM	0.78342	467
	539	NDGVVANNN-**K**-SNOOOOOOO	0.51865	539
Os*GPAT*11	99	LMWILGNPI-**K**-LEGMENLNT	0.90227	99

	138	APTGTVGIA-**K**-KEIIWYPLF	0.86211	138
	139	PTGTVGIAK-**K**-EIIWYPLFG	0.78888	139
	203	SKTGRLLPF-**K**-KGFVHTALQ	0.60463	203
	279	YADSLPDSQ-**K**-PLEPVNTGK	0.82815	279
Os*GPAT*13	3	OOOOOOOMA-**K**-KPCEFPTAV	0.91101	3
	4	OOOOOOMAK-**K**-PCEFPTAVL	0.79813	4
	47	TGAAIPPAD-**K**-LHNQTVMID	0.91236	47
	119	MVSFFGLPE-**K**-EVVRIGKAV	0.87515	119
	126	PEKEVVRIG-**K**-AVLPKFFLE	0.73036	126
	131	VRIGKAVLP-**K**-FFLEGMAME	0.51934	131
	149	EGLEVVRNA-**K**-KVVVFSPLF	0.76715	149
	228	AVGLAGVGS-**K**-MHHLFSRYC	0.69155	228
Os*GPAT*14	220	SSSFPSFVA-**K**-RSVARLPMV	0.6271	220
	311	KPVILRYPY-**K**-RFSPAWDSM	0.65331	311
	356	PSEQEKEDP-**K**-LYANNVRKL	0.55487	356
Os*GPAT*15	5	OOOOOMSPF-**K**-PIEQCSTEG	0.90928	5
	172	KPGVLIREH-**K**-RNAVVREFG	0.51898	172
	203	SDFDFMAIC-**K**-DAYVVTTSR	0.84008	203
	423	FHGSTARGF-**K**-GMDPYFFFM	0.58306	423
Os*GPAT*16	205	ATRSFLPFC-**K**-KQLRPPFCE	0.65357	205
	206	TRSFLPFCK-**K**-QLRPPFCED	0.59545	206
	423	FHPTTARGW-**K**-AMDPIFFFM	0.5238	423
Os*GPAT*17	101	WPFLFEKIN- **K**-TKFVFSGET	0.77231	101
	146	LRKGRLQCI- **K**-YILKKSLMK	0.60963	146
	150	RLQCIKYIL- **K**-KSLMKLPIF	0.77819	150
	151	LQCIKYILK- **K**-SLMKLPIFN	0.826	151
	174	IIEFIPVER- **K**-WEVDEPLIR	0.69241	174
	361	LTLFSSVWF- **K**-IFVAFSSAF	0.7578	361
Os*GPAT*18	172	SFGYHWIRR-**K**-GKPAPRELA	0.74776	172
	311	NISLGKLMF-**K**-MFTQFHNFM	0.58198	311
	334	LPVVYPPEI-**K**-QENALHFAE	0.84077	334
	447	IFQYFDFEA-**K**-ESITFRQFL	0.52331	447
Os*GPAT*19	118	AFFPVHFLL-**K**-GQKMRSKIE	0.61251	118
	147	FVASWTGVI-**K**-YHGPRPSTR	0.76356	147
	186	MTAFAVIMQ-**K**-HPGWVGFIQ	0.60809	186
	213	CIWFNRNDL-**K**-DREVVAKKL	0.63237	213
	221	LKDREVVAK-**K**-LRDHVQHPD	0.66122	221
	252	VNNQYTVMF-**K**-KGAFELGCA	0.65937	252
	253	NNQYTVMFK-**K**-GAFELGCAV	0.79717	253
	271	VCPIAIKYN-**K**-IFVDAFWNS	0.72962	271
	281	IFVDAFWNS-**K**-KQSFTMHLV	0.78375	281
	334	DMIAARAGL-**K**-KVPWDGYLK	0.60078	334
	343	KKVPWDGYL-**K**-HNRPSPKHT	0.68569	343
Os*GPAT*21	193	LGRCFKFIS- **K**-TSIFMFPII	0.91653	193
	235	LKRCVDLVK- **K**-GASVFFFPE	0.70259	235
	252	PEGTRSKDG- **K**-LGAFKRGAF	0.62248	252
	281	IPITLLGTG- **K**-LMPSGMEGI	0.86417	281
Os*GPAT*23	37	RRGALRLEA- **K**-AAWRPAARG	0.62753	37
	85	ILHIRKEVE- **K**-GKLPADVAA	0.72125	85
	145	NPFTFPPYH- **K**-AVREPFDYY	0.89176	145
	345	CYEVMPPPQ- **K**-VEKEIGEQR	0.78865	345
	348	VMPPPQKVE- **K**-EIGEQRVIS	0.5	348
Os*GPAT*24	39	LFLSIRPFS- **K**-SLYRRINRF	0.89326	39
	146	YLFLERSWA- **K**-DEKTLKWGL	0.62846	146

	149	LERSWAKDE- **K**-TLKWGLQRL	0.5	149
	159	TLKWGLQRL- **K**-DFPRPFWLA	0.72151	159
	180	VEGTRFTPA- **K**-LLAAQEYAV	0.78081	180
	230	IYDTTVIIP- **K**-DSPQPTMLR	0.63331	230
	253	QSSVVHVRM- **K**-RHAMSEMPK	0.57787	253
	278	KWCKDIFVA- **K**-DALLDKHLA	0.80689	278
	333	LWTQLLSTW- **K**-GVGFTGLGL	0.72294	333
Os*GPAT*25	3	OOOOOOOMP- **K**-KKLSHRLFS	0.65634	3
	4	OOOOOOMPK- **K**-KLSHRLFSA	0.73619	4
	5	OOOOOMPKK- **K**-LSHRLFSAL	0.55845	5
	43	TLPHPSLLH- **K**-SSSSFPPME	0.73673	43
	110	CVMGSDMAL- **K**-VMAMVSFCG	0.80442	110
	134	FRAGRAVLP- **K**-WFLEDVGEE	0.57445	134
	184	VEVVSGREM- **K**-VIWGFFTGI	0.83552	184
	255	ARWSALPRD- **K**-YPKPMVFHD	0.84497	255
	258	SALPRDKYP- **K**-PMVFHDGRL	0.75166	258
	319	AATGLSWRL- **K**-GEAPTPLAG	0.64871	319
	514	GYRCTMLTR-K-DKYLMLAGN	0.5	514
Os*GPAT*26	3	OOOOOOOMA-K-TKLFPALFS	0.55163	3
	5	OOOOOMAKT-K-LFPALFSLL	0.57531	5
	200	EVVVAAREM-K-VVWGFYTGV	0.55271	200

Table S4. Putative uiquitylation sites of rice *GPAT* genes predicted by BDM-PUB and Ub-Pred

	BDM-PUB			Ub-Pred		
Gene ID	Position	score	Threshold	Position	score	Ubiquitinated
OsGPAT 2	169	0.4	0.3	169	0.62	Yes-Low confidence
OsGPAT 4	42	0.47	0.3	42	0.68	Yes-Low confidence
	65	1.44	0.3	65	0.65	Yes-Medium confidence
	544	3.93	0.3	544	0.65	Yes-Low confidence
OsGPAT 5	18	1.3	0.3	18	0.69	Yes-Low confidence
	396	4.12	0.3	396	0.64	Yes-Low confidence
OsGPAT 6	7	1.12	0.3	7	0.68	Yes-Low confidence
OsGPAT 7	251	2.26	0.3	251	0.62	Yes-Low confidence
	358	0.55	0.3	358	0.67	Yes-Low confidence
	429	0.48	0.3	429	0.9	Yes-High confidence
OsGPAT 8	502	4.16	0.3	502	0.85	Yes-High confidence
OsGPAT 9	456	0.46	0.3	456	0.62	Yes-Low confidence
	462	1.76	0.3	462	0.67	Yes-Low confidence
	463	1.49	0.3	463	0.73	Yes-Medium confidence
OsGPAT 10	57	0.81	0.3	57	0.62	Yes-Low confidence
	213	0.49	0.3	213	0.65	Yes-Low confidence
	539	2.33	0.3	539	0.71	Yes-Medium confidence
OsGPAT 12	45	2.47	0.3	45	0.46	Yes-Medium confidence
OsGPAT 13	47	2.42	0.3	47	0.67	Yes-Low confidence
	532	3.26	0.3	532	0.64	Yes-Low confidence
OsGPAT 14	32	1.93	0.3	32	0.95	Yes-High confidence
	188	0.71	0.3	188	0.79	Yes-Medium confidence
	246	1.02	0.3	246	0.66	Yes-Low confidence
	402	3.4	0.3	402	0.81	Yes-Medium confidence
OsGPAT 15	5	1.45	0.3	5	0.8	Yes-Medium confidence
	348	2.98	0.3	348	0.65	Yes-Low confidence
	457	0.81	0.3	457	0.81	Yes-Medium confidence
OsGPAT 17	395	4.72	0.3	395	0.62	Yes-Low confidence
OsGPAT 18	376	0.52	0.3	376	0.62	Yes-Low confidence
OsGPAT 20	208	1.13	0.3	208	0.75	Yes-Medium confidence
OsGPAT 22	63	1.41	0.3	208	0.64	Yes-Low confidence
OsGPAT 24	262	1.05	0.3	262	0.67	Yes-Low confidence
OsGPAT 25	43	1.2	0.3	43	0.83	Yes-Medium confidence
	53	2.7	0.3	53	0.85	Yes-High confidence
OsGPAT 26	48	0.83	0.3	48	0.78	Yes-Medium confidence
	508	2.7	0.3	508	0.73	Yes-Medium confidence

Table S5. Highly conserved amino acid positions in Post translation modification sites

Gene ID	Methylation	Phosphorylation	Ubiquitylation
Os*GPAT*1	K483		
	K545		
Os*GPAT*2	K91		
	K438		
Os*GPAT*4	K145		K544
Os*GPAT*5	K151		K396
Os*GPAT*6	K102		
	K414		
Os*GPAT*7	K251		
	K97		
	K213		
	K404		
Os*GPAT*8	K212		
Os*GPAT*9	K374	S343	
Os*GPAT*10	K467		
Os*GPAT*11	K138		
	K139		
	K203		
Os*GPAT*12			K45
Os*GPAT*13	A126	Y521	
	K131	S423	
Os*GPAT*14	K220	S245	K246
		S314	K402
Os*GPAT*15	K203	S393	
Os*GPAT*16	K423		
		S392	
Os*GPAT*17	K150		K395
	K151		
Os*GPAT*18	K311		
Os*GPAT*19	K252		
	K253		
	K343		
Os*GPAT*20		S376	
Os*GPAT*21	K193		
Os*GPAT*22		S368	
Os*GPAT*23	K348		
Os*GPAT*24	K39		
	K146		
	K180		
	k278		
Os*GPAT*25	K134	S365	
	K514	S423	
Os*GPAT*26		S387	
		S445	

Table S6. List of the 26 rice *GPAT* genes and their detailed information including gene ID, locus positions, gene and protein size taken from a previous study. *OsGPAT* denotes rice *GPAT* genes

Gene Name	Locus ID	Chromosomal position	Gene (bp)	Protein
OsGPAT 1	LOC_Os01g14900	Chr01: 8347893-8351006	3114	571
OsGPAT 2	LOC_Os01g19390	Chr01: 10972881 - 10969331	3551	454
OsGPAT 3	LOC_Os01g22560	Chr01: 12678141 - 12677086	1056	265
OsGPAT 4	LOC_Os01g44069	Chr01: 25257001 - 25263981	6981	545
OsGPAT 5	LOC_Os01g57360	Chr01: 33158320 - 33154276	4045	400
OsGPAT 6	LOC_Os01g63580	Chr01: 36866375 - 36863214	3162	498
OsGPAT 7	LOC_Os01g70570	Chr01: 40868356 - 40864908	3449	508
OsGPAT 8	LOC_Os02g02340	Chr02: 778356 - 775494	2863	507
OsGPAT 9	LOC_Os03g52570	Chr03: 30153564 - 30151748	1817	468
OsGPAT 10	LOC_Os03g61720	Chr03: 34985196 - 34990759	5564	542
OsGPAT 11	LOC_Os04g53370	Chr04:31786589 - 31788246	1658	293
OsGPAT 12	LOC_Os04g57150	Chr04:34058149 - 34054300	3850	158
OsGPAT 13	LOC_Os05g20100	Chr05:11756381 - 11753281	3101	538
OsGPAT 14	LOC_Os05g28960	Chr05:16982192 - 16977760	4905	405
OsGPAT 15	LOC_Os05g37600	Chr05:22005196 - 22008976	3781	487
OsGPAT 16	LOC_Os05g38350	Chr05:22481303 - 22484545	3243	523
OsGPAT 17	LOC_Os05g42270	Chr05:24724324 - 24727260	2937	398
OsGPAT 18	LOC_Os06g49790	Chr06:30122819 - 30128027	5209	557
OsGPAT 19	LOC_Os07g34730	Chr07:20812544 - 20808033	4512	371
OsGPAT 20	LOC_Os08g03700	Chr08:1759198 - 1757711	1488	496
OsGPAT 21	LOC_Os10g35390	Chr10:18917046 - 18920354	3309	333
OsGPAT 22	LOC_Os10g41070	Chr10:22059229 - 22060717	1489	480
OsGPAT 23	LOC_Os10g42720	Chr10:23044693 - 23050740	6048	428
OsGPAT 24	LOC_Os11g41900	Chr11:25206038 - 25199831	6208	375
OsGPAT 25	LOC_Os11g45400	Chr11:27486903 - 27490269	3367	544
OsGPAT 26	LOC_Os12g37600	Chr12:23091636 - 23094141	2506	559