

Prediction of Indels and SNP's in coding regions of glutathione peroxidases – an important enzyme in redox homeostasis of plants

Sayak Ganguli*, Abhijit Datta**

DBT Centre for Bioinformatics, Presidency University, 86/1 College Street, Kolkata - 700073, India

***E-mail address: sayakbif@yahoo.com , abhijit_datta21@yahoo.com

ABSTRACT

Plant glutathione peroxidases are an important class of enzymes which play key roles in the stress adaptability of plants both in context of biotic and abiotic stress pathways. They have been over the years much studied in animals since the catalytic residues are comprised of selenocysteine a variant amino acid which is ribosomally encoded with the help of an RNA structural element known as SECIS. Various workers over the years have shown that plant glutathione peroxidases play active roles in ROS sequestration, lipid hydroperoxidation as well as regulate glutathione levels. However, each plant has various patterns of glutathione peroxidase expression and action and in some plants certain isoforms have not been detected at all. This work focuses on the prediction and identification of single nucleotide polymorphisms (SNPs) and INDELS in the coding regions of plant glutathione peroxidases, with the help of a Bayesian based algorithm subsequently validated. A large number of informative sites were detected 279 of which had variant frequency of $\geq 50\%$. This data should be beneficial for future studies involving genetic manipulation and population based breeding experiments.

Keywords: SNP's; Stress Tolerant Genotypes (STG's); Indels; glutathione peroxidases

1. INTRODUCTION

Single nucleotide polymorphisms (SNPs) and segmental insertions and deletions (indels) represent two major classes of molecular markers, which have attained a large amount of importance in plant population genetics studies. These two classes of markers along with the differences in tandem repeats at a particular locus (microsatellites, SSR's /ISSR's) comprise the three major groups of allelic variations within a particular genome. Among the three major groups SNP's have generated a lot of attraction owing to the fact that they are stable and are the most frequent type of genetic polymorphisms (Syvanen 2001). Various studies have been performed using SNP data, for the analyses of genetic diversity (Varshney 2008), deciphering substructures in populations (Garris et.al 2003, Rakshit 2007, Caicedo 2007); identifying linkage disequilibrium in genomes (Mather 2007, Agrama 2008); and various other screening efforts.

Glutathione peroxidases in plants have been identified to be involved in abiotic stress an responsive pathway which aims to maintain a redox hoemostasis. These enzymes (E.C > 1.11.1.9) have a broad substrate specificity; however, their main affinity is towards H₂O₂.

The main reactions that they catalyze are the reduction of H₂O₂ and lipid hydroperoxide to water and alcohol. Chen et.al. (2004) has also reported the role of GpXs in controlling oxidative burst and programmed cell death in *Arabidopsis*. Phospholipid hydroperoxide glutathione peroxidases (PHGPx) is an unique member of this family of enzymes as it has the ability to catalyze the reduction of phospholipid hydroperoxide and other complex hydroperoxilipids – components of the lipid bilayer.

Computational identification of SNP's in glutathione peroxidases was attempted keeping in mind the importance of this enzyme family as a key abiotic stress regulator which has the potential to serve as an important biomarker for differentiating the levels of ROS – homeostasis in plants. Apart from this several plants with agronomic values can be bred in such a way that their GpX load in the genome is maintained and they can serve as stress tolerant genotypes (STGs).

2. MATERIAL AND METHOD

Sequences were retrieved from the NCBI – GenBank collection and were subsequently curated for obtaining the complete sequences. Partial, hypothetical and incomplete sequences were not considered. Following that an extensive Bayesian based algorithm was used taking into account the depth of the alignment, associated base composition in the region and a standardized priori polymorphism rate. Once the predictions were made the results were validated using the Geneious Pro suite.

3. RESULTS AND DISCUSSION

The final curated sets of sequences were 400 in number which possessed the complete coding sequences. The results indicated 2210 informative sites out of which 1186 were attributable to SNPs whereas 1024 sites were classified as Indels. A total of 279 sites were found to have a variant frequency of greater than equal to 50 out of which 129 were SNPs and 150 sites were Indels. Plant SNP data in context of glutathione peroxidase is very limited in the standard archives such as dbSNP of NCBI; and is at this point restricted to information from *Arabidopsis thaliana* with only 94 entries at this point of time.

Different plants are enriched in different subset of genes and more importantly a same plant may exhibit variant responses in two different stress conditions both biotic and abiotic. Reverse genetics strategies such as post transcriptional gene silencing, insertional mutagenesis, TILLING etc. have been successfully used for identifying single nucleotide polymorphisms and production of adaptable cultivars (Henikoff 2003). Thus the identification of SNPs and correlating that variation with an important agronomic or stress adaptable trait is important for production of better crop species as well as to understand the genetic strategies of the different plant genomes.

The identification of these SNPs and variants such as Indels should be validated in the wet lab through sequencing techniques and subsequent *in silico* analyses. Molecular modelling and subsequent *in silico* mutagenesis (Ganguli et.al. 2013) should also be useful for detecting whether the SNP creates any structural or functional anomaly in the 3D structure of the protein.

Table 1. Data showing those informative sites which have variant frequencies of $\geq 50\%$.

NUCLEOTIDE	MAXIMUM	MINIMUM	COVERAGE	VARIATION TYPE	VARIANT FREQUENCY
A	383	383	20	SNP	50.00%
	553	553	24	Indel	50.00%
	570	570	24	Indel	50.00%
	595	595	24	Indel	50.00%
	596	596	24	Indel	50.00%
	597	597	24	Indel	50.00%
G	746	746	32	SNP	50.00%
C	756	756	32	SNP	50.00%
G	808	808	34	SNP	50.00%
A	1,704	1,704	40	Indel	50.00%
	2,414	2,414	32	Indel	50.00%
	2,452	2,452	32	Indel	50.00%
T	2,453	2,453	32	Indel	50.00%
	2,453	2,453	32	Indel	50.00%
	2,804	2,804	32	Indel	50.00%
	2,881	2,881	32	Indel	50.00%
	2,882	2,882	32	Indel	50.00%
T	2,908	2,908	32	Indel	50.00%
G	3,467	3,467	162	SNP	50.00%
T	3,493	3,493	164	SNP	50.00%
A	4,393	4,393	62	SNP	50.00%
G	4,521	4,521	20	SNP	50.00%
T	4,523	4,523	20	SNP	50.00%
C	3,544	3,544	165	SNP	50.30%
T	4,251	4,251	119	Indel	50.40%
T	3,743	3,743	168	SNP	50.60%
C	3,607	3,607	173	SNP	50.90%

A	3,495	3,495	164	SNP	51.20%
C	3,552	3,552	168	SNP	51.20%
C	3,565	3,565	168	SNP	51.20%
A	959	959	37	SNP	51.40%
C	978	978	37	SNP	51.40%
G	1,045	1,045	37	SNP	51.40%
G	1,068	1,068	37	SNP	51.40%
G	1,098	1,098	37	SNP	51.40%
	1,973	1,973	35	Indel	51.40%
	1,974	1,974	35	Indel	51.40%
G	768	768	33	SNP	51.50%
T	778	778	33	SNP	51.50%
	2,119	2,119	33	Indel	51.50%
	2,171	2,171	33	Indel	51.50%
G	3,509	3,509	165	SNP	51.50%
A	3,791	3,791	171	SNP	51.50%
	2,685	2,685	31	Indel	51.60%
	2,687	2,687	31	Indel	51.60%
	2,689	2,689	31	Indel	51.60%
	2,700	2,700	31	Indel	51.60%
	2,701	2,701	31	Indel	51.60%
	2,702	2,702	31	Indel	51.60%
	2,704	2,704	31	Indel	51.60%
	2,770	2,770	31	Indel	51.60%
	2,771	2,771	31	Indel	51.60%
	2,803	2,803	31	Indel	51.60%
	661	661	29	Indel	51.70%
	672	672	29	Indel	51.70%

	674	674	29	Indel	51.70%
G	701	701	29	SNP	51.70%
G	3,706	3,706	174	SNP	51.70%
C	3,812	3,812	172	SNP	51.70%
	463	463	21	Indel	52.40%
G	3,491	3,491	164	SNP	52.40%
T	4,004	4,004	170	SNP	52.40%
C	3,469	3,469	162	SNP	52.50%
G	1,602	1,602	38	SNP	52.60%
C	1,604	1,604	38	SNP	52.60%
A	3,709	3,709	173	SNP	52.60%
G	3,870	3,870	173	SNP	52.60%
	1,869	1,869	36	Indel	52.80%
	1,872	1,872	36	Indel	52.80%
G	188	188	17	SNP	52.90%
T	286	286	17	SNP	52.90%
C	843	843	34	SNP	52.90%
	2,014	2,014	34	Indel	52.90%
	3,004	3,004	34	Indel	52.90%
T	3,697	3,697	174	SNP	52.90%
A	4,481	4,481	34	SNP	52.90%
C	739	739	32	SNP	53.10%
	2,919	2,919	32	Indel	53.10%
	2,920	2,920	32	Indel	53.10%
A	167	167	15	SNP	53.30%
C	172	172	15	SNP	53.30%
G	184	184	15	SNP	53.30%
	2,585	2,585	30	Indel	53.30%

	2,599	2,599	30	Indel	53.30%
C	3,481	3,481	163	SNP	53.40%
T	4,272	4,272	114	Indel	53.50%
G	3,443	3,443	162	SNP	53.70%
G	3,458	3,458	162	SNP	53.70%
A	4,063	4,063	167	SNP	53.90%
	1,008	1,008	37	Indel	54.10%
	1,105	1,105	37	Indel	54.10%
	1,106	1,106	37	Indel	54.10%
	4,463	4,463	37	Indel	54.10%
A	4,045	4,045	168	SNP	54.20%
	1,929	1,929	35	Indel	54.30%
	1,930	1,930	35	Indel	54.30%
T	3,799	3,799	171	SNP	54.40%
	2,040	2,040	33	Indel	54.50%
	2,084	2,084	33	Indel	54.50%
	2,089	2,089	33	Indel	54.50%
	2,151	2,151	33	Indel	54.50%
	2,152	2,152	33	Indel	54.50%
	2,153	2,153	33	Indel	54.50%
	2,178	2,178	33	Indel	54.50%
	2,978	2,978	33	Indel	54.50%
T	3,536	3,536	165	SNP	54.50%
T	4,312	4,312	95	SNP	54.70%
C	731	731	31	SNP	54.80%
	2,765	2,765	31	Indel	54.80%
	2,792	2,792	31	Indel	54.80%
	2,796	2,796	31	Indel	54.80%

	2,797	2,797	31	Indel	54.80%
	2,798	2,798	31	Indel	54.80%
	2,800	2,800	31	Indel	54.80%
	2,801	2,801	31	Indel	54.80%
	2,802	2,802	31	Indel	54.80%
C	298	298	20	SNP	55.00%
C	299	299	20	SNP	55.00%
	359	359	20	Indel	55.00%
T	3,721	3,721	171	SNP	55.00%
G	4,522	4,522	20	SNP	55.00%
G	3,624	3,624	174	SNP	55.20%
G	893	893	36	SNP	55.60%
T	895	895	36	SNP	55.60%
G	898	898	36	SNP	55.60%
	1,877	1,877	36	Indel	55.60%
G	4,068	4,068	167	SNP	55.70%
G	3,533	3,533	165	SNP	55.80%
C	3,864	3,864	172	SNP	55.80%
G	3,962	3,962	172	SNP	55.80%
C	3,971	3,971	172	SNP	55.80%
T	3,002	3,002	34	Indel	55.90%
C	3,472	3,472	161	SNP	55.90%
	617	617	25	Indel	56.00%
G	3,618	3,618	173	SNP	56.10%
C	3,470	3,470	162	SNP	56.20%
	2,399	2,399	32	Indel	56.30%
	2,479	2,479	32	Indel	56.30%
	1,475	1,475	39	Indel	56.40%

	489	489	23	Indel	56.50%
	551	551	23	Indel	56.50%
A	3,766	3,766	170	SNP	56.50%
T	3,953	3,953	173	SNP	56.60%
C	4,059	4,059	166	SNP	56.60%
	2,584	2,584	30	Indel	56.70%
	3,250	3,250	90	Indel	56.70%
C	927	927	37	SNP	56.80%
G	936	936	37	SNP	56.80%
G	1,034	1,034	37	SNP	56.80%
A	1,050	1,050	37	SNP	56.80%
C	1,095	1,095	37	SNP	56.80%
	1,107	1,107	37	Indel	56.80%
	1,108	1,108	37	Indel	56.80%
C	3,459	3,459	162	SNP	56.80%
C	417	417	21	SNP	57.10%
C	4,041	4,041	168	SNP	57.10%
C	3,588	3,588	171	SNP	57.30%
C	3,704	3,704	174	SNP	57.50%
	2,172	2,172	33	Indel	57.60%
G	4,056	4,056	166	SNP	57.80%
	1,592	1,592	38	Indel	57.90%
	3,120	3,120	57	Indel	57.90%
C	3,865	3,865	173	SNP	58.10%
T	4,081	4,081	161	SNP	58.40%
T	2,994	2,994	34	Indel	58.80%
	3,010	3,010	34	Indel	58.80%
C	3,783	3,783	171	SNP	59.10%

T	4,308	4,308	98	SNP	59.20%
T	3,981	3,981	172	SNP	59.30%
	2,454	2,454	32	Indel	59.40%
	2,482	2,482	32	Indel	59.40%
	2,805	2,805	32	Indel	59.40%
	2,807	2,807	32	Indel	59.40%
	2,808	2,808	32	Indel	59.40%
T	965	965	37	SNP	59.50%
T	1,037	1,037	37	SNP	59.50%
A	1,046	1,046	37	SNP	59.50%
A	1,079	1,079	37	SNP	59.50%
	1,109	1,109	37	Indel	59.50%
T	3,612	3,612	173	SNP	59.50%
T	3,840	3,840	172	SNP	59.60%
A	3,494	3,494	164	SNP	59.80%
G	3,527	3,527	164	SNP	59.80%
C	3,456	3,456	162	SNP	59.90%
G	296	296	20	SNP	60.00%
	614	614	25	Indel	60.00%
	1,920	1,920	35	Indel	60.00%
	1,921	1,921	35	Indel	60.00%
	1,923	1,923	35	Indel	60.00%
	1,961	1,961	35	Indel	60.00%
	3,022	3,022	35	Indel	60.00%
T	3,511	3,511	164	SNP	60.40%
	1,530	1,530	38	Indel	60.50%
	1,532	1,532	38	Indel	60.50%
	1,533	1,533	38	Indel	60.50%

	1,534	1,534	38	Indel	60.50%
	1,541	1,541	38	Indel	60.50%
G	3,841	3,841	172	SNP	60.50%
	2,224	2,224	33	Indel	60.60%
	2,862	2,862	33	Indel	60.60%
	2,976	2,976	33	Indel	60.60%
	508	508	23	Indel	60.90%
	510	510	23	Indel	60.90%
C	3,604	3,604	172	SNP	61.00%
C	4,072	4,072	164	SNP	61.00%
T	4,050	4,050	167	SNP	61.10%
A	3,992	3,992	171	SNP	61.40%
G	3,740	3,740	169	SNP	61.50%
A	3,836	3,836	172	SNP	61.60%
	665	665	29	Indel	62.10%
	666	666	29	Indel	62.10%
G	976	976	37	SNP	62.20%
	568	568	24	Indel	62.50%
	2,343	2,343	32	Indel	62.50%
	1,915	1,915	35	Indel	62.90%
	1,935	1,935	35	Indel	62.90%
	1,943	1,943	35	Indel	62.90%
A	3,462	3,462	162	SNP	63.00%
	1,584	1,584	38	Indel	63.20%
	1,587	1,587	38	Indel	63.20%
	1,588	1,588	38	Indel	63.20%
	1,589	1,589	38	Indel	63.20%
T	3,555	3,555	167	SNP	63.50%

	2,052	2,052	33	Indel	63.60%
G	3,597	3,597	171	SNP	63.70%
C	3,989	3,989	171	SNP	63.70%
A	3,747	3,747	169	SNP	63.90%
G	3,625	3,625	174	SNP	64.40%
A	3,999	3,999	169	SNP	64.50%
	3,008	3,008	34	Indel	64.70%
C	921	921	37	SNP	64.90%
G	3,993	3,993	171	SNP	64.90%
	1,258	1,258	40	Indel	65.00%
	1,263	1,263	40	Indel	65.00%
	1,265	1,265	40	Indel	65.00%
	1,266	1,266	40	Indel	65.00%
	1,303	1,303	40	Indel	65.00%
	1,308	1,308	40	Indel	65.00%
	1,336	1,336	40	Indel	65.00%
	516	516	23	Indel	65.20%
	550	550	23	Indel	65.20%
G	3,724	3,724	169	SNP	65.70%
T	4,016	4,016	170	SNP	65.90%
T	3,831	3,831	172	SNP	66.30%
G	3,875	3,875	173	SNP	66.50%
	567	567	24	Indel	66.70%
	2,846	2,846	33	Indel	66.70%
C	4,011	4,011	169	SNP	66.90%
C	3,788	3,788	171	SNP	67.30%
	1,275	1,275	40	Indel	67.50%
	1,276	1,276	40	Indel	67.50%

	1,317	1,317	40	Indel	67.50%
	1,382	1,382	40	Indel	67.50%
C	3,775	3,775	170	SNP	67.60%
	2,582	2,582	31	Indel	67.70%
	2,679	2,679	31	Indel	67.70%
	2,681	2,681	31	Indel	67.70%
	2,684	2,684	31	Indel	67.70%
	1,567	1,567	38	Indel	68.40%
	1,576	1,576	38	Indel	68.40%
	1,577	1,577	38	Indel	68.40%
	1,962	1,962	35	Indel	68.60%
	2,348	2,348	32	Indel	68.80%
C	3,779	3,779	170	SNP	68.80%
	660	660	29	Indel	69.00%
G	3,977	3,977	172	SNP	69.20%
	519	519	23	Indel	69.60%
	1,395	1,395	40	Indel	70.00%
G	3,756	3,756	170	SNP	70.00%
C	3,631	3,631	174	SNP	70.10%
A	3,567	3,567	168	SNP	70.20%
C	3,980	3,980	172	SNP	71.20%
A	3,523	3,523	165	SNP	71.50%
	1,465	1,465	39	Indel	71.80%
	2,321	2,321	32	Indel	71.90%
	1,163	1,163	40	Indel	72.50%
	1,285	1,285	40	Indel	72.50%
	1,512	1,512	39	Indel	74.40%
	2,318	2,318	32	Indel	75.00%

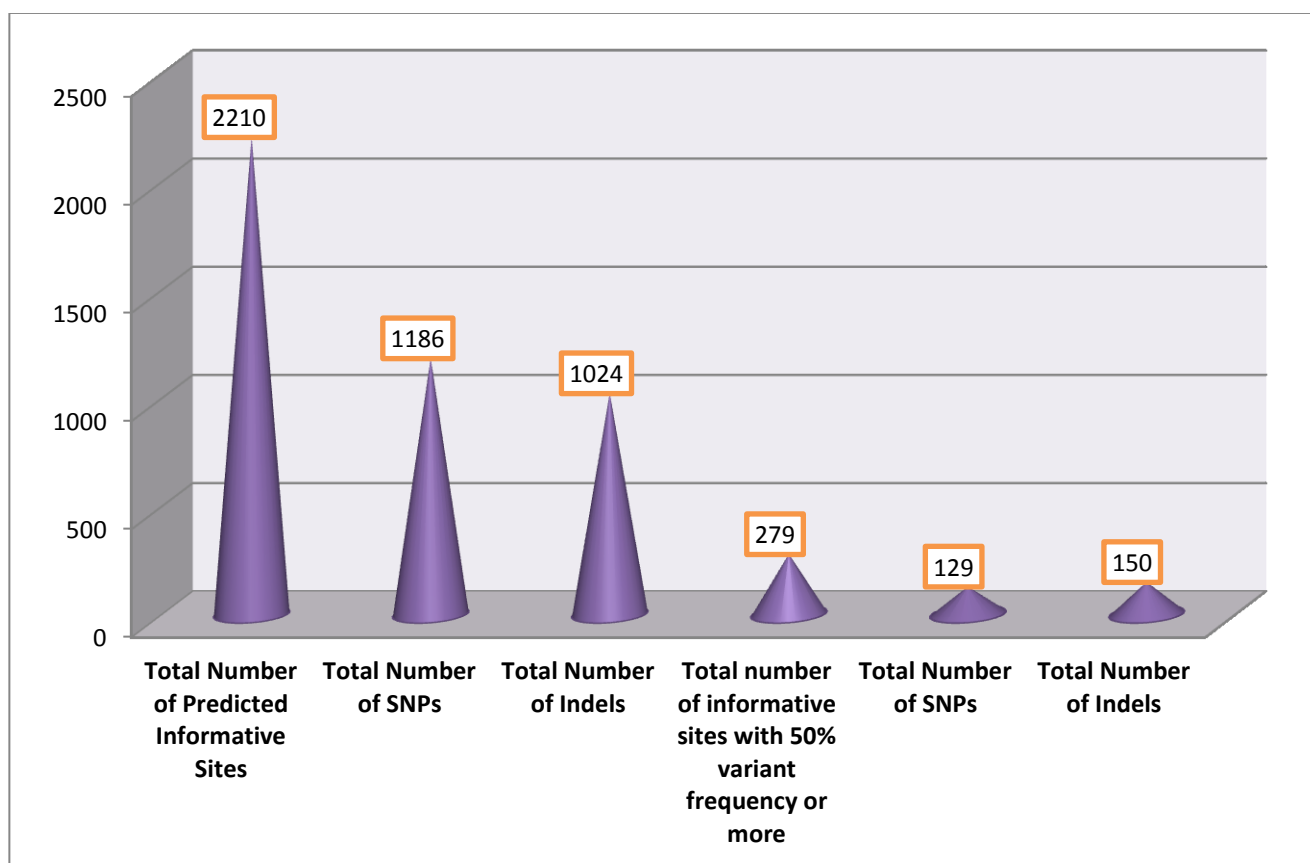


Fig. 1. Graph illustrating the results obtained.

4. CONCLUSIONS

A large number of informative sites were identified in the present study conforming to SNP positions as well as Indels. These informative sites predicted are all in the coding region of the genes and thus possess the ability to alter the function of the encoded protein. Thus these should be validated and reported using NGS methods and subsequent computational analyses.

Acknowledgement

The authors acknowledge the DBT- GoI; BTBI – BIF scheme for the funds provided towards maintenance of the facility.

References

- [1] Syvänen A. C., *Nat Rev Genet.* 2(12) (2001) 930-942.
- [2] Varshney R. K, Thiel T., Sretenovic-Rajicic T., Baum M., Valkoun J., Guo P., Grando S., Ceccarelli S., Graner A., *Mol Breed* 22 (2008) 1-13.
- [3] Rakshit S., Rakshit A., Matsumura H., Takahashi Y., Hasegawa Y., Ito A., Ishii T., Miyashita T., Terauchi R., *Theor Appl Genet* 114 (2007) 731-743.
- [4] Garris A. J., McCouch S. R., Kresovich S., *Genetics* 165 (2003) 759-769.

- [5] Caicedo A. L., et. al., *PLoS Genet* 3 (2007) e163.
- [6] Mather K. A., Caicedo A. L., Polato N. R., Olsen K. M., McCouch S., Purugganan M., D., *Genetics* 177 (2007) 2223-2232.
- [7] Agrama H. A., Eizenga G. E., *Euphytica* 160 (2008) 339-355.
- [8] Henikoff S., Comai L., *Annu. Rev. Plant Biol.* 54 (2003) 375-401.
- [9] Ganguli S., Datta A., *Annu. Res. Rev. Biol.* 4(1) (2013) 143-153.
- [10] NCBI: www.ncbi.nlm.nih.gov
- [11] dbSNP: www.ncbi.nlm.nih.gov/SNP
- [12] GenBank: www.ncbi.nlm.nih.gov/genbank/
- [13] Geneious Pro version 7.0.5 [Windows version]
- [14] Chen S., Vaghchhipawala Z., Li W., Asard H., Dickman M. B., *Plant Physiol* 135 (2004) 1630-1641.

(Received 11 December 2013; accepted 16 December 2013)