Plant Omics Journal POJ

POJ 3(5):140-148 (2010) ISSN:1836-3644

In silico survey and characterization of Resistance Gene Analogues (RGAs) in the genomic regions encompassing gall midge resistance genes *Gm4* and *Gm5* in rice (*Oryza sativa* L.)

Mahima Dubey[1] and Girish Chandel[1*]

Department of Plant Molecular Biology and Biotechnology[1], College of Agriculture, Krishak Nagar, Indira Gandhi Agricultural University, Raipur 492006, C. G. India

***Corresponding author: ghchandel@gmail.com**

Abstract

With recent advances in the field of rice genome analysis and availability of large genomic data, we have surveyed and characterized resistance gene analogues (RGAs) in the genomic region of two gall midge resistance genes *Gm4* and *Gm5* which confer resistance against Asian rice gall midge biotypes 1, 2, 4 and 5. Both gall midge resistance genes, *Gm4* and *Gm5* have been mapped on Chromosomes 8 and 12 of rice (*Oryza sativa* L.). Here, we have investigated the presence of RGAs in the rice genomic region between the co-dominant SSR markers RM210 and RM256 flanking *Gm4* gene, spanning a region of 2.056 Mb on long arm of chromosome 8 and the genomic region between RM101 and RM309 flanking *Gm5* gene and spanning 13.2 Mb regions on chromosome 12. By scanning Nipponbare sequences (*japonica* rice), we found 11 and 30 RGAs in the genomic regions of *Gm4* and *Gm5* genes respectively and further confirmed the sequences of each of the RGAs in *indica* (93-11) genome sequences. The RGAs were also characterized by sequence tag based methods of expression profiling *viz* ESTs and MPPS signature analysis to understand the functions of putative candidate genes. A total of 174 ESTs and 178 MPSS tags co-localized with the RGAs present in the genomic region of gall midge resistance gene *Gm4*. Similarly, 284 ESTs and 586 MPSS signatures were found to co-localize with the RGAs present in the genomic region of *Gm5* gene. Based on the higher frequencies of ESTs (≥7 matches) and MPSS tags (>500 TPM value), three RGAs (RGA-*Gm4*-04, RGA-*Gm4*-05 and RGA-*Gm4*-08) associated with *Gm4* and five RGAs (RGA-*Gm5*-02, RGA-*Gm5*-10, RGA-*Gm5*-14, RGA-*Gm5*-27 and RGA-*Gm5*-30) associated with *Gm5* genes were identified as functional RGAs. The identification of functionally associated RGAs for two gall midge resistance genes forms the basis for the development of DNA markers for the marker assisted selection of gall midge resistance in rice.

Keywords: Expression analysis; Functional categorization; Functional R-genes; *In silico*; Rice gall midge; Resistance gene analogues

Abbreviations: EST- expressed sequence tags; HR- hypersensitive reaction; MPSS- massively parallel signature sequencing; NBS-LRR- nucleotide- binding site leucine- rich repeats; RGA- resistance gene analogues; R-genes- resistance genes

Introduction

Asian rice gall midge is one of major pests seriously affecting the rice cultivation in South East Asian countries including India. Resistance against this pest in most cases is controlled by a single dominant locus (Katiyar et al., 2001) and till date, ten non allelic resistance genes (*Gm1-Gm10*) conferring resistance to various gall midge biotypes have been identified. Molecular markers have been used to map these gall midge resistance genes and eight gall midge resistance genes *Gm-1 to Gm-8* are either mapped to rice genome or tagged with various molecular markers. The tedious work over years has resulted in the tagging and mapping of *Gm1* on chromosome 9, *Gm2*, *gm3*, *Gm 6(t)* & *Gm7* on chromosome 4, *Gm4* & *Gm8* on chromosome 8 and *Gm5* on chromosome 12 of rice (Biradar et al., 2004; Mohan et al., 1994; Katiyar et al., 2000; Katiyar et al., 2001; Sardesai et al., 2002 and Jain et al., 2004). Based on the genetic diversity studies among the differentials and their reactions to Indian biotypes of gall midge, it has been reported that many of the gall midge resistance genes induce hypersensitive type of reactions in resistant genotypes and their reactions are classified as HR+ type. Two genes *Gm4* and *Gm5* are known to evoke such response (Bentur and Kalode, 1996 and Bentur et al., 2003). HR+ is a well characterized phenomenon in a plant pathogen interaction. Therefore, the reaction induced by *Gm4* and *Gm5* genes has been assumed to be similar to the one induced by pathogens. Extensive research carried out in the area of plant- pathogen interaction and disease resistance governing genes (R-genes) has led to the isolation and cloning of number of resistance genes in various plant species. Studies gained momentum with the identification of resistance gene analogues (RGAs) in other crop species and exploitation of these RGAs as molecular markers to fine map and clone resistance loci as well as to serve as candidate resistance genes (Maleki et al., 2003, Wang et al., 2005). RGAs are widely distributed in the genome, often organized in clusters and sometimes strongly linked to known resistance loci

(Kanazin et al., 1996 and Meyers et al., 1999). These markers serve as landmark for fine mapping of R genes and also speed up the process of positional cloning of these genes. The RGA based markers have been greatly exploited to isolate resistance gene analogues (RGAs) for disease in a number of crop species *viz* soybean, rice, maize and wheat (Collins et al., 1998; Kanazin et al., 1996; Mago et al., 1999 and Maleki et al., 2003) but have not been utilized for the mapping of insect resistance loci. Although insect 'R' genes may be rare for chewing insects, they are not rare in insect-plant interactions that involve less mobile insects that feed by other mechanisms, like aphids, plant hoppers and gall midges. Gall midges classified as internal feeders by entomologists feed by unknown mechanisms which bear resemblance to the feeding mechanism of sedentary pathogens (Berzonsky et al., 2002 and Yencho and Byrne, 2000). Owing to the similarity between the reaction patterns induced by gall midge and pathogens and keeping in mind the invaluable role of RGAs in mapping and cloning of resistance loci, they can be potentially used to identify candidate genes and to design RGA based markers for fine mapping of gall midge resistance genes *Gm4* and *Gm5*. Rice being the most tractable species for genomic applications among monocots is privileged at all the fronts of genomics including sequencing, computational resources, annotation *etc* (Yuan et al., 2001). High throughput techniques developed for transcriptome profiling and sequencing have paved way to characterize function of any gene under question. One of the latest techniques is the sequence tag-based platforms in transcriptomics which includes ESTs (expressed sequence tags), SAGE (serial analysis of gene expression) and MPSS (massively parallel signature sequencing). ESTs are randomly sequenced in an unbiased cDNA library and are classified into clusters of transcript sequences using sequence-clustering and assembling methods. The abundance of transcripts expressed in each tissue is estimated as EST count for each cDNA library (Mochida and Shinozaki, 2010). Another sequencing-based technology is the MPSS which uses a unique method to quantify gene expression levels by generating millions of short sequence tags per library by sequencing 16–20 bp from the 3' side of cDNA using a micro bead array. The abundance of each signature represents and measures the gene expression levels in the sample tissue (Brenner et al., 2000). While MPSS, SAGE, and expressed sequence tags (ESTs) are all sequence-based technologies for transcriptional profiling, MPSS provides more thorough qualitative and quantitative description of gene expression due to the tremendous depth of sampling it offers (Iandolino et al., 2008). EST and MPSS tag based profiling offers great opportunities for *in silico* applications using web based tools. A huge collection of EST and MPSS tags in various tissue libraries for rice are available at EST anatomy viewer at TIGR database *(http://www.tigr.org/tdb/e2k/osa1/ dnav/)* and rice MPSS database *(http://mpss.udel.edu/rice)*. These open access databases provide information about abundance of an EST or MPSS tag in a specific tissue library. Thus identification of ESTs or MPSS tag sequences in a gene can yield valuable information about putative spatial or temporal expression of that gene. With the completion of rice genome sequencing, anchoring of these sequences into a fine genetic map (Sasaki et al., 2005) and availability of advanced techniques of expression profiling it is now possible to search for and characterize the resistance gene analogues using bioinformatics tools, in the mapped gall midge resistance genomic regions. We searched the genomic regions of two gall midge resistance genes *Gm4* and *Gm5* for RGAs and further characterized them for ESTs and MPSS tags for expression of putative genes located on the genomic regions. Our analysis revealed the presence of 11 and 30 RGAs for *Gm4* and *Gm5* genes respectively. Three RGAs for *Gm4* gene and 5 RGAs for *Gm5* gene were also identified as functional resistance gene analogs based on the ESTs and MPSS tag analysis.

Materials and methods

In silico mining of the genomic region encompassing Gm4 and Gm5 genes for resistance gene analogues (RGAs)

The genomic region encompassing *Gm4* and *Gm5* genes were searched for the presence of resistance gene analogues (RGAs) which are known to be involved in the defense related mechanisms. The plant R genes cloned so far show striking similarities and are known to contain conserved protein motifs of nucleotide-binding site leucine-rich repeats (NBS-LRR), *Cf2/Cf5*, kinase proteins (serine/ threonine protein kinase, receptor kinases *etc*), LRRs and putative disease resistance genes. The genes carrying these conserved protein motifs are considered as RGAs (Ghazi et al., 2009). The genomic regions underlying *Gm4* gene flanked by RM210 and RM256 and *Gm5* gene defined by RM101 and RM309 were taken from the latest sequence map set of Gramene Annotated Nipponbare Sequence 2009 *(http://www.gramene.org)*. *Gm4* gene was found to span a region of 2.056 Mb on long arm of chromosome 8 whereas *Gm5* gene was found to span 13.2 Mb genomic region covering parts of both the arms of chromosome 12 across the centromere. The putative genes carrying the above mentioned conserved protein motifs and putative disease resistance genes (RGAs) were identified in the region of *Gm4* and *Gm5* genes from the Rice Annotation Project (RAP) annotated genes (*http://www.tigr.org/*). The Genomic, c-DNA and protein sequences of all the RGAs were downloaded in FASTA format and stored as separate files.

Analysis of sequence similarity of RGAs between japonica and indica subspecies of rice

DNA level homology was performed for all the RGAs present in both the resistance gene region (*Gm4* and *Gm5*) of *japonica* (Nipponbare) by BLASTN search **(***http://www.gramene.org/ multi/blastview*) to the genome sequences of *indica (*93-11). Further, the protein sequences of all the RGAs were subjected to T-BLASTN protein homology search to identify their R-gene homologs and alleles in *indica* genomic sequences.

Functional categorization of the genes present in the genomic regions encompassing gall midge resistance genes Gm4 and Gm5

The region encompassing *Gm4* and *Gm5* genes were searched to estimate the gene content and types of annotated genes using FTP browser available at *(ftp://ftp.tigr.org/pub/data/ Eukaryotic _Projects/o_sativa/annotation_dbs/*) and were classified to different functional categories as per the MIPS functional catalog using web browser *(http://mips.gsf.de/ proj/thal/db/)*.

Table 1. List of different classes of RGAs found in the specific genomic regions encompassing gall midge resistance genes *Gm4* (chromosome 8) and *Gm5* (chromosome 12) in rice

RGA ID	Physical position (in bp)	Size of CDS	Dir	Predicted protein function
In the genomic region encompassing *Gm4* flanked by RM 210 and RM 256				
RGA-gm4-01	22415752 - 22413423	1437	-	Putative serine/threonine protein kinase
RGA-gm4-02	22526032 - 22519854	2706	-	NBS-LRR disease resistance protein
RGA-gm4-03	22699445 - 22704392	1194	+	Putative Disease resistance protein
RGA-gm4-04	24376462 - 24380046	3315	+	Receptor-like protein kinase
RGA-gm4-05	24537434 - 24535449	1986	-	Transmembrane receptor kinase
RGA-gm4-06	24631273 - 24634233	2961	+	Putative transmembrane receptor kinase
RGA-gm4-07	25032026 - 25034512	2487	+	Leucine Rich Repeat family protein
RGA-gm4-08	25377878 - 25375775	1731	-	Leucine rich repeat containing protein
RGA-gm4-09	25495562 - 25499103	2202	+	Putative protein kinase protein
RGA-gm4-10	25931158 - 25928444	2154	-	Receptor-like protein kinase
RGA-gm4-11	26157158 - 26158875	537	+	Putative LRR receptor kinase
In the genomic region encompassing *Gm5* flanked by RM 101 and RM 309				
RGA-gm5-01	7600188 - 7604161	2883	+	Putative NBS-LRR disease resistance protein
RGA-gm5-02	9783824 - 9780610	2700	-	stripe rust resistance protein Yr10, putative, expressed
RGA-gm5-03	9926400 - 9923982	1116	-	CC-NBS-LRR resistance protein MLA13, putative
RGA-gm5-04	9985325 - 9983692	846	-	NBS-LRR disease resistance protein, putative, expressed
RGA-gm5-05	9993456 - 9992611	1545	-	resistance protein, putative
RGA-gm5-06	10011703 - 10007910	2706	-	MLA12, putative
RGA-gm5-07	10018771 - 10014595	2646	-	stripe rust resistance protein Yr10, putative, expressed
RGA-gm5-08	13598474 - 13597179	1296	-	protein kinase domain containing protein, expressed
RGA-gm5-09	13600373 - 13601551	1179	+	protein kinase domain containing protein, expressed
RGA-gm5-10	16196655 - 16193635	1272	-	serine/threonine-protein kinase AFC2, putative, expressed
RGA-gm5-11	16537165 - 16534848	2196	-	NBS-LRR disease resistance protein, putative
RGA-gm5-12	16542080 - 16540735	840	-	NBS-LRR disease resistance protein, putative
RGA-gm5-13	16575498 - 16569228	2361	-	NBS-LRR disease resistance protein, putative
RGA-gm5-14	16592225 - 16588003	2844	-	NBS-LRR disease resistance protein, putative, expressed
RGA-gm5-15	16685599 - 16680936	2304	-	disease resistance protein RPM1, putative, expressed
RGA-gm5-16	17356925 – 17359483	2559	+	disease resistance protein RGA3, putative, expressed
RGA-gm5-17	17564502 - 17512177	4515	-	NBS-LRR disease resistance protein, putative
RGA-gm5-18	17734518 - 17729944	4575	-	NBS-LRR disease resistance protein, putative
RGA-gm5-19	17981689 - 17978520	2208	-	disease resistance protein RPM1, putative
RGA-gm5-20	17987203 - 17984533	1386	-	disease resistance protein RPM1, putative
RGA-gm5-21	17989303 - 17999433	1437	+	protein kinase domain containing protein, expressed
RGA-gm5-22	18342033 - 18343310	1278	+	disease resistance protein RGA2, putative, expressed
RGA-gm5-23	18445885 - 18440649	3168	-	disease resistance protein, putative
RGA-gm5-24	19018267 - 19015562	2706	-	disease resistance protein RPM1, putative, expressed
RGA-gm5-25	18956069 - 18958083	1569		MLA10, putative, expressed
RGA-gm5-26	18956069 - 18958083	2808	+	protein kinase, putative
RGA-gm5-27	19681164 - 19676790	3075	-	NBS-LRR domain containing protein, expressed
RGA-gm5-28	19690657 - 19689176	1482	-	CC-NBS-LRR resistance protein, putative, expressed
RGA-gm5-29	19777524 - 19778819	1296	+	disease resistance RPP13-like protein 1, putative, expressed
RGA-gm5-30	20037778 - 20034570	2793	-	RGH1A, putative, expressed

In silico expression profiling of RGAs by frequency analysis of ESTs and MPSS signatures

The RGA sequences were predicted for putative temporal and spatial pattern of expression through signature tag based approaches. To assess functional pattern of genes for spatial and temporal expression, the gene sequences i.e. RGA sequences were searched for co-localized ESTs. The locus ID of each RGA was used as query to search for the ESTs mapped over these genes and further expression pattern was predicted on the basis of respective tissue expression library using Rice Gene Expression Anatomy Viewer and Digital Northern tools available at TIGR database *(http://www.tigr.org/tdb/e2k/osa1/dnav/)*. ESTs corresponding to a tissue library provided information about putative site of expression of the RGAs in which it was identified. Mahalingam et al., (2003) have described the criteria for transcriptome analysis of stress modulated genes by digital northern technique in model genome *Arabidopsis*. As per their criteria, the genes were classified to three different categories based on the number of

EST matches to a gene. Minimally expressed category was assigned if the EST matches to a gene were <7, relatively highly expressed for 7-200 EST matches and highly expressed for > 200 EST matches to a gene sequence. A similar criterion was followed by Ameline-Torregrosa et al. (2008) for characterization of NBS-LRR genes in the model tree species Medicago by *in silico* expression analysis using EST library. Based on this criterion, the RGAs were sorted to minimally expressed, relatively highly expressed and highly expressed category. As the MPSS provides more thorough qualitative and quantitative description of gene expression, the further characterization of RGAs was done by MPSS analysis. The rice MPSS database includes a comprehensive set of libraries which can be accessed at site, *http://mpss.udel.edu/rice*. The tool provides 17 and 20 nucleotide long tags, tag positions, chromosome coordinates *etc*. The sequence of each RGA was used as query under 'query by sequence' section of rice MPSS database to identify MPSS tags corresponding to the RGA query sequence as well as their abundance in 22 diverse tissue libraries constructed from various developmental stages, tissue types and pathogen challenged or non challenged tissues. The abundance/ frequency of each tag is expressed in TPM (transcript per million) and the TPM value under 'Norm Abund' category is considered as the measure of expression in a corresponding tissue library.

Results

Number, type and distribution of RGAs in the genomic region encompassing gall midge resistance genes Gm4 and Gm5

A total of 11 RGAs for the region encompassing gall midge resistance gene *Gm4* and 30 RGAs for the region encompassing gall midge resistance gene *Gm5* were identified on rice chromosomes 8 and 12 respectively and are presented systematically in Table 1. Based on sequence motif/ domains, these RGAs belonged to nucleotide-binding site leucine-rich repeats (NBS-LRR), *Cf2/Cf5*, kinase proteins (serine/ threonine protein kinase, receptor kinases *etc*), LRR classes and putative disease-resistance genes. The RGAs are clustered as groups on rice chromosomes. For the region encompassing *Gm4* gene, most of the RGAs were found in small clusters of 2-4 RGA per cluster. Out of 11 RGAs, 5 RGAs belonged to receptor like kinase class, 5 belonged to LRR class including one NBS-LRR disease resistance gene and the remaining one was of serine/ threonine protein kinase type. Similarly in the region encompassing *Gm5* gene, 30 RGAs were present in 7 clusters with 3-6 RGA per cluster. The biggest cluster was of 6 RGAs found between 17.5-18.7 Mb regions on rice chromosome 12. RGAs from all the major classes of R genes except *cf2/cf5* disease resistance class were represented in the genomic region of *Gm5* gene. In contrast to the *Gm4*, 11 out of 30 RGAs were of NBS LRR type.

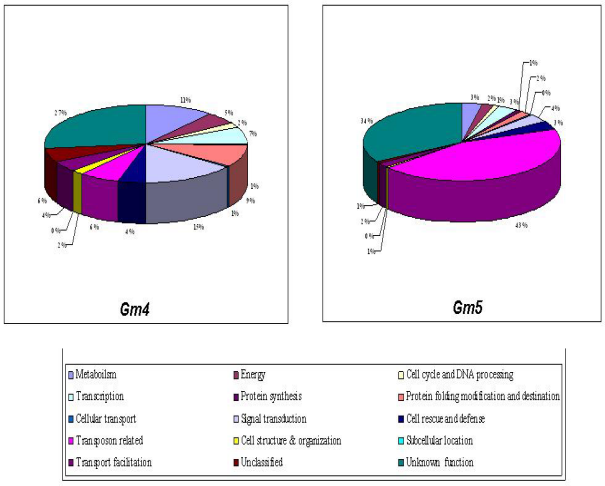

Fig1. Distribution of genes present in the region encompassing *Gm4* and *Gm5* genes to different functional categories.

Sequence similarity of RGAs between the two subspecies of rice

The RGAs identified for both the gall midge resistance genes showed significant DNA level homology with *indica* genome sequences (score ≥1000 and zero e-value). For the genomic region underlying gall midge resistance gene *Gm4*, RGA-*Gm4*-02 encoding NBS-LRR disease resistance gene exhibited maximum identity (100% identity with scores > 4000 and zero e-value.). Similarly the three RGAs namely RGA-*Gm5*-11, RGA-*Gm5*-16 and RGA-*Gm5*-18 of the region encompassing *Gm5* gene also showed maximum identity to the *indica* genome sequences. In contrast, the RGA, RGA-*Gm5*-02 did not show significant DNA level homology at the set BLAST parameters (score> 1000 and zero e value). Further the RGAs with maximum sequence homology (one for the *Gm4* gene and 3 for *Gm5* gene) were subjected to domain search (www.tigr.org) which resulted in the identification of NB-ARC domain as the common protein domain among all these RGAs. This indicates that the putative genes carrying NB-ARC domain are highly conserved between these two sub-species of rice especially in the genomic regions under study. It was interesting to report that the RGAs showing high sequence homology to *indica* for both the genomic regions were also positioned within the same flanking microsatelite markers (BGI 93-11 Sequence 2005 map set) as stated for the Nipponbare. The protein sequences of all the identified RGAs were subjected to T-BLASTN protein level homology search which showed that 10 out of 11 RGAs in region of gall midge resistance gene *Gm4* had their homologs in *indica* genome. Among these, RGA-*Gm4*-03 appeared non allelic to *japonica* because of being located on a different chromosome instead of chromosome 8. Similarly 26 out of 30 RGAs in region of *Gm5* gene had their homologs in *indica* genome. Three of these homologs (RGA-*Gm5*-06, RGA-*Gm5*-15, RGA-*Gm5*-17) were present on chromosome 3, 5 and 11 respectively and hence were non allelic to *indica* rice.

Table 2. Expression pattern of RGAs based on EST count and MPSS signature abundance

Sr. No.	RGA ID	EST count	MPSS signature abundance in TPM
For RGAs in the genomic region encompassing *Gm4*			
1	RGA-gm4-01	6	62
2	RGA-gm4-02	6	46
3	RGA-gm4-03	11	145
4	RGA-gm4-04*	28	755
5	RGA-gm4-05*	58	1160
6	RGA-gm4-06	17	285
7	RGA-gm4-07	4	209
8	RGA-gm4-08*	20	506
9	RGA-gm4-09	9	201
10	RGA-gm4-10	14	298
For RGAs in the genomic region encompassing *Gm5*			
1	RGA-gm5-01	5	119
2	RGA-gm5-02*	29	684
3	RGA-gm5-04	6	502
4	RGA-gm5-05	0	21
5	RGA-gm5-07	2	49
6	RGA-gm5-08	0	26
7	RGA-gm5-10*	78	12201
8	RGA-gm5-11	3	270
9	RGA-gm5-12	2	20
10	RGA-gm5-13	2	276
11	RGA-gm5-14*	26	705
12	RGA-gm5-16	1	17
13	RGA-gm5-18	3	49
14	RGA-gm5-21	0	17
15	RGA-gm5-22	4	92
16	RGA-gm5-25	5	87
17	RGA-gm5-27*	36	1183
18	RGA-gm5-28	5	58
19	RGA-gm5-29	3	38
20	RGA-gm5-30*	73	7212

* Resistance gene analogs identified commonly as functional RGAs based on the combined expression results of ESTs and MPSS datasets (EST matches 7 and above, Mahalingam et al., 2003 and MPSS TPM value > 500, Meyers et al., 2004b)

Functional categorization of genes present in the region encompassing gall midge resistance genes Gm4 and Gm5

All the functionally annotated genes present in the genomic region encompassing *Gm4* and *Gm5* genes were downloaded from the FTP browser *(ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/)* and were sorted to 15 primary functional categories as per the MIPS functional catalog. A total of 156 and 1,762 functionally annotated genes were found in the genomic regions of *Gm4* and *Gm5* genes respectively. The distribution of the genes to different functional categories is represented as pie chart in Figure 1. The largest category of genes was formed by genes encoding proteins involved in signal transduction comprising 15% of the total genes present in the region underlying *Gm4* gene. About 11% of the genes in this region were classified under metabolism category forming the second largest group followed by transposon related genes. The genes involved in transcriptional activity constituted 7% of the genes. Cell rescue and defense related genes constituted 4% of total genes which included genes encoding proteins of antioxidant response like peroxidase, genes induced in response to other stresses, heat shock, pathogens elicitors, chaperones and hairpins. In the region of *Gm5* gene, transposon related genes formed the largest category comprising 43% of the total genes. This may be attributed to high-copy-number of transposable elements that comprise the majority of eukaryotic genomes (Ken et al., 2009). Rice genome particularly contains 16,220 transposable element related genes/loci **(**MSU Rice Genome Annotation, Release 6.1**)** and also many R-gene loci have been reported to contain interspersed transposable elements (Noel et al., 1999). Genes related to signal transduction, transcription and cell rescue and defense were present in significant numbers in this region. We have found that both the regions are transcriptionally active and are rich in signal transduction and defense related genes, which is indicative of their active involvement in plant defense related mechanism.

In silico expression profiling of RGAs by frequency analysis of Expressed Sequence Tags (ESTs)

In silico expression profiling of RGAs was carried out to predict the putative temporal and spatial pattern of expression by analyzing the co-localization of identified ESTs and MPSS tags with the RGA sequences. The expression pattern was predicted based on the frequency of ESTs and MPSS signatures

abundance. The EST anatomy viewer and digital northern analysis generated high quality ESTs that co localized with RGAs present in the regions of *Gm4* and *Gm5* genes. A total of 174 ESTs co localized with 10 out of the 11 RGAs present in the genomic region of *Gm4* gene and EST number ranged from 4-58 EST per expressed RGA. RGA-*Gm4*-05 showed the highest number of ESTs (58) followed by RGA-*Gm4*-04 which showed 28 ESTs (Table 2). These ESTs expressed in diverse tissue libraries including shoot, root, leaf, panicle, flower, callus, whole plant, mixed tissues *etc*. The tissue library information revealed that shoot was the common tissue type in which ESTs corresponding to all the 10 RGAs expressed (Figure 2a). Based on the criteria of Mahalingam et al (2003), 7 out of 10 RGAs present in this region showed EST matches of 7 and above and thus were in relatively highly expressed category. Similarly for the region encompassing *Gm5* gene, 17 out of 30 RGAs present in this region showed EST support. A total of 284 ESTs were mapped over 17 RGAs with 1-78 EST per expressed RGA. In this region, only 5 RGAs were categorized to relatively highly expressed category. RGA-*Gm5*-10 encoding serine/ threonine-protein kinase showed highest number of ESTs (78) followed by RGA-*Gm5*-30 having 73 co localized ESTs (Table 2). These ESTs expressed in different tissue libraries *viz* shoot, root, leaf, panicle, immature seed, flower and callus (Fig 2b). Similar to the genomic region of *Gm4* gene, the tissue type shoot was predominant in which all the ESTs corresponding to all the 17 RGAs expressed in significant frequency. Based on the EST frequency data, 7 RGAs *viz* RGA-*Gm4*-03, RGA-*Gm4*-04, RGA-*Gm4*-05, RGA-*Gm4*-06, RGA-*Gm4*-08, RGA-*Gm4*-09, RGA-*Gm4*-10 in the region encompassing *Gm4* gene and 5 RGAs *viz* RGA-*Gm5*-02, RGA-*Gm5*-10, RGA-*Gm5*-14, RGA-*Gm5*-27, RGA-*Gm5*-30 in the region of *Gm5* gene showed higher level expression.

MPSS signature analysis

We also performed *in silico* expression analysis of RGAs by measuring the abundance of MPSS signatures co localizing with the RGA sequences. The abundance of MPSS tags identified in each sequence is depicted by its TPM value (transcript per million), which is an exact digital representation of number of copies of the transcript in a tissue which indicates expression level of corresponding gene quantitatively. Great variation in TPM values of MPSS tags were observed for resistance gene analogues ranging from 5 to 12,200, but only those MPSS tags having TPM > 15 in atleast one tissue library were considered. The TPM value below 15 is indicative of very low and basal levels of expression (Meyers et al., 2004a and b). A total of 178 MPSS tags (17 bp) were found corresponding to RGAs in the genomic region encompassing *Gm4* gene and the tag number ranged from 2-31. Out of 178 MPSS tags, maximum number of signature tags belonged to class I (those present within the exonic region of the gene sequence) and rest belonged to II, IV and V class of MPSS signature tags (Table 3). High TPM tags corresponding to 3 RGAs in this region were found (Table 2) with MPSS tag GATCCTCTCCG TGAGAT corresponding to RGA-*Gm4*-05 showing the highest cumulative TPM value of 2,525 in all the libraries. The tissue library wise expression of RGAs based on TPM value is shown in Fig. 3a. High TPM tags showed significantly higher expression in meristematic tissues, reproductive parts (pollen and stigma), *Xanthomonas oryzae, M. grisea* challenged rice leaves, beet army worm and water weevil damaged leaves,

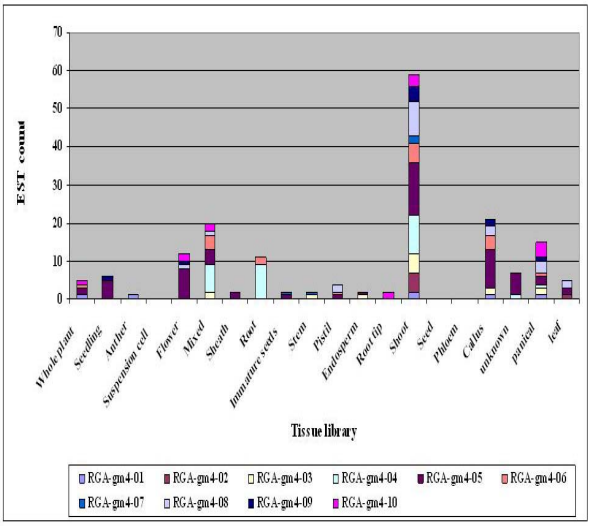

Fig 2a: EST count in different tissue libraries for RGAs present in the genomic region encompassing *Gm4* gene

water stressed young roots, callus, F_1 hybrid meristematic leaves, germinating seed and developing seeds. We have also observed that among the stressed libraries, these RGAs expressed significantly higher in pathogen and insect challenged rice leaves (Fig. 3a) Similarly a total of 586 MPSS tags were found corresponding to RGAs present in the region encompassing *Gm5* gene. MPSS tags belonging to all the seven classes were found. Maximum number of signature tags belonged to class I and rest belonged to class II, IV, III, V, VI and class VII (Table 3). Out of 586 MPSS tags, 73 were high TPM tags. The tag GATCTGTATCAGTCTGA corresponding to RGA-*Gm5*-10 showed highest TPM value of 8,243 followed by tag GATCGCCTCGCTGAGCT corresponding to RGA-*Gm5*-30 with TPM value of 6,577 which expressed in all the tissue libraries. The RGAs with high TPM value tags showed comparatively higher expression in reproductive parts like pollen and stigma, immature panicle, germinating seed, callus, *X. oryzae, M. grisea* challenged rice leaves, meristematic tissues, developing seeds, beet army worm and water weevil challenged and mechanically damaged rice leaves as compared to other libraries. Their tissue library information indicated the higher expression of RGAs in leaves under stressed conditions particularly under biotic stress induced by pathogen challenge and insect damage. So on the basis of MPSS analysis a total of 3 RGAs within *Gm4* region and 6 RGAs within *Gm5* region (Table 2) were identified as highly expressed resistance gene analogues (TPM> 500, Meyers et al., 2004b). Based on the ESTs and MPSS signature analysis, the three RGAs *viz* RGA-*Gm4*-04, RGA-*Gm4*-05 and RGA-*Gm4*-08 in the region encompassing *Gm4* gene and 5 RGAs *viz* RGA-*Gm5*-02, RGA-*Gm5*-10, RGA-*Gm5*-14, RGA-*Gm5*-27 and RGA-*Gm5*-30 were found to be highly expressed in the tissue libraries challenged by pathogens and insects. These RGAs are identified as the functional resistance gene analogues and can be used for the development of RGA based markers.

Discussion

The *in silico* survey of the genomic regions encompassing gall midge resistance genes *Gm4* and *Gm5* for the resistance gene

Table 3. Number of 17-base MPSS signatures of different classes corresponding to RGAs present in the region encompassing *Gm4* and *Gm5*

Classes of MPSS tags	Position of MPSS tag	Number of signature tags	
		Gm4 region	*Gm5* region
1	Within exon, same strand	71	204
2	Within 500 bp potential 3_UTR	39	115
3	Antisense to exon	-	82
4	Unannotated	41	91
5	Within intron, sense strand	27	65
6	Within intron, antisense strand	-	26
7	Spans an exon/intron splice site	-	03
	Total	178	586

- MPSS tags classified to seven different classes based on the location of the signature sequence relative to the gene position (Meyers et al., 2004b)

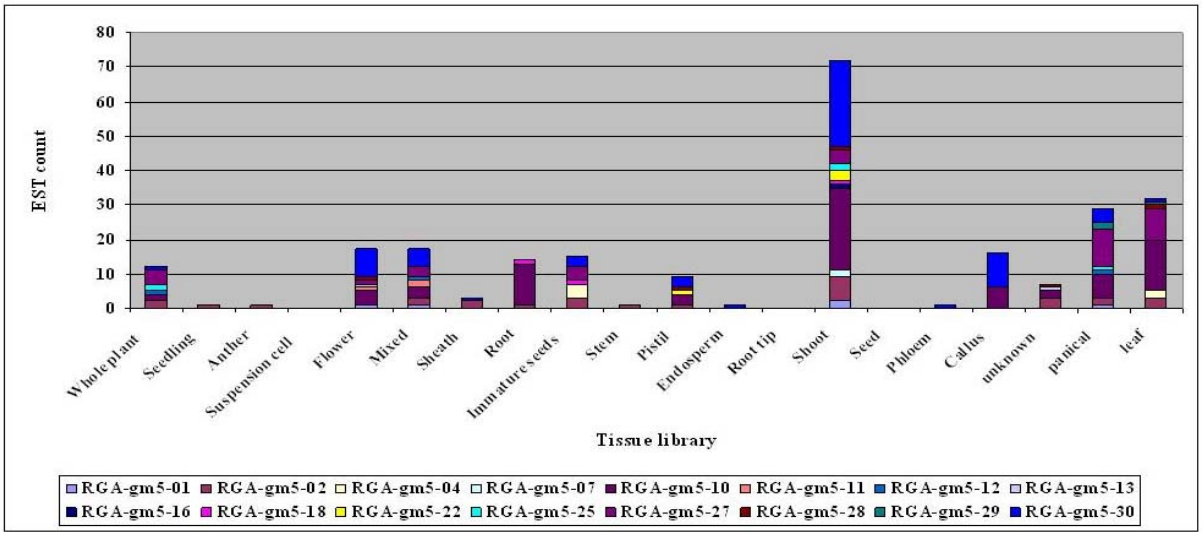

Fig 2b: EST count in different tissue libraries for RGAs present in the genomic region encompassing *Gm5* gene.

analogues revealed that both the regions are fairly rich in RGAs with 11 RGAs in a 2.056 Mb region of *Gm4* gene on chromosome 8 and 30 RGAs in the region of *Gm5* gene, which was a bigger segment of 13.2 Mb on chromosome 12 of rice. These RGAs represent 7% of a total of 158 RGAs present on chromosome 8 and 25% of a total of 126 RGAs present on chromosome 12 of rice as reported by Wang et al (2005). On the basis of sequence similarity analysis between the RGAs of *japonica* and *indica* genomes, significant DNA level homology was observed. Further the putative genes carrying NB-ARC domain were found to be highly conserved between these two sub-species of rice especially in the genomic regions under study. The overall sequence homology of the resistance genes may vary significantly but several short motifs of their encoding proteins, such as NBS and ARC motifs were reported to be highly conserved between the plant species (He et al., 2004 and Graham et al., 2000). The protein level homology of all the RGAs present in the region encompassing both the gall midge resistance genes showed 90% homology in our study confirming the DNA level homology findings. Out of a total of 41 RGAs, 37 showed allelism with the indica sequences and the remaining 4 were found to be non allelic. A considerable level of allelic differences and polymorphism does exist between the R gene homologs of these two sub species (Ghazi et al., 2009). Similar findings have been observed by Wang et al (2005) who reported that out of a total of 861 *indica* R-gene homologs identified and screened by aligning the *japonica* RGA sequences to the *indica* genomic sequence using program T-BLASTN, 702 of them showed allelism and 159 were found to be non allelic. Functional characterization of RGAs based on the ESTs analysis revealed that 50% of the RGAs identified in the genomic region of both the gall midge resistance genes had expression support with majority of them falling under relatively highly expressed category. While working with *Medicago*, Ameline-Torregrosa et al., (2008) reported that 50.5% of the predicted NBS-LRR genes had EST support with an average of 3.1 ESTs per expressed NBS gene using EST libraries. Out of a total of 178 and 586 MPSS tags that co localized with RGAs present in the region of *Gm4* and *Gm5* genes respectively, 71 and 204 for both the genes respectively co localized in the exonic region of the same strand of the gene sequence and hence are the strong and reliable source for inferring expression levels. Class I MPSS signature tags have been used for the characterization of several developmental

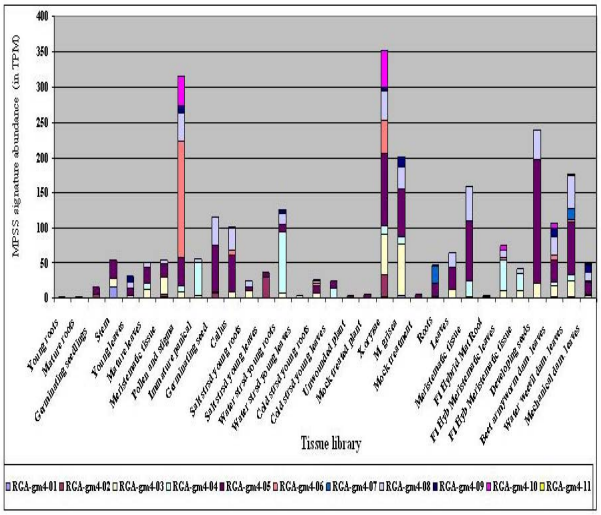

Fig 3a. MPSS signature abundance in different tissue libraries for RGAs present in the genomic region encompassing *Gm4* gene.

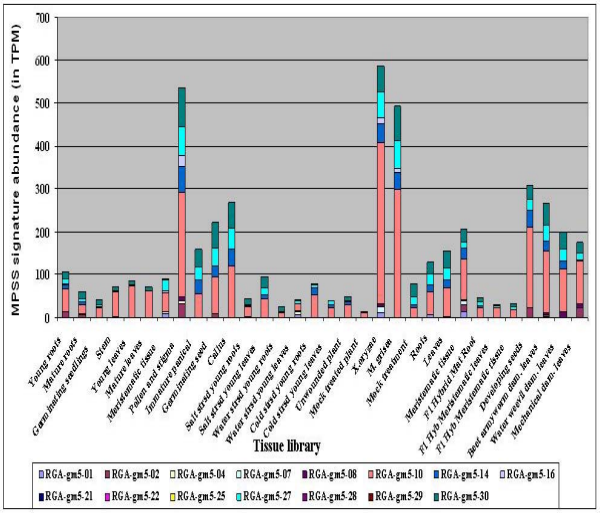

Fig 3b. MPSS signature abundance in different tissue libraries for RGAs present in the genomic region encompassing *Gm5* gene.

libraries and stress transcriptomes by many workers (Goring et al., 2004). Computational expression profiling through EST and MPSS signatures revealed a considerable degree of commonalties in the two sequence tag based approaches. RGA-*Gm4*-05 was the top ranking gene of the region encompassing *Gm4* gene based on its higher EST frequency as well as MPSS signature abundance. Similarly RGA-*Gm5*-10 was the most highly expressed gene followed by RGA-*Gm5*-30 in the region encompassing *Gm5* gene by EST count and MPSS signature abundance analysis approaches indicating the common features of both the datasets. Collectively RGA-*Gm4*-05, RGA-*Gm4*-04, RGA-*Gm5*-10, RGA-*Gm5*-30, RGA-*Gm5*-27 are the top ranking genes in both the datasets. At the same time small differences in expression of few LRR class RGAs (RGA-*Gm4*-07, RGA-*Gm5*-11 and RGA-*Gm5*-13) were also observed. These RGAs showed moderately higher expression levels based on higher MPSS signature abundance but were not supported by comparable EST matches. Similar results have been observed by Iandolino et al (2008) while comparative expression profiling in grape by frequency analysis of ESTs and MPSS signatures. They observed both commonalities and differences in the expression of some storage proteins, putative transcription factor and plasma membrane aquaporin related genes. The tissue expression library information of ESTs revealed that the tissue type shoot is the putative site of expression of RGAs, which was a common observation for the RGAs present in the genomic regions of both *Gm4* and *Gm5* genes. These genes have also showed comparatively higher expression in the library comprising of stressed rice leaves particularly in pathogen and insect challenged libraries based on their higher MPSS signature abundance (TPM values). This indicates that they are the functional resistance gene analogues. RGAs have long been recognized to be linked to disease resistance and pathogen defense (also supported by their higher expression in *M. grisea* and *X. oryzae* challenged leaves, Fig. 3a and 3b), but their significant expression in insect (beet army worm and water weevil) challenged rice leaves indicate their involvement in insect resistance as well. These can be potentially used for the development of RGA based markers as a new marker system for fine mapping of insect resistance genes. Since the initial identification of RGAs and their use to identify resistance loci, the difficulty remained in the identification of the functional resistance genes. This difficulty can be solved to an extent by prior analysis and characterization of the RGAs. The identification of functional RGAs for both the regions encompassing *Gm4* and *Gm5* using *in silico* functional characterization provides the basis for the development of candidate gall midge resistance gene markers for fine mapping of gall midge resistance genes *Gm4* and *Gm5* for effective use in MAS and as a starting point towards map based cloning of these two genes.

Acknowledgements

We gratefully acknowledge the Department of Information Technology (DIT), New Delhi and Department of Biotechnology, New Delhi for funding the research work.

References:

Ameline-Torregrosa CA, Wang BB, Majesta S. O'Bleness, Deshpande S, Zhu H, Roe B, Young ND and Cannon SB (2008) Identification and Characterization of Nucleotide-Binding Site-Leucine-Rich Repeat Genes in the Model Plant *Medicago truncatula*. Plant Physiol 146: 5-21

Bentur JS and Kalode M (1996) Hypersensitive reaction and induced resistance in rice against the Asian rice gall midge *Orseolia oryzae*. Indian J of Agric Sci 66: 197-199

Bentur JS, Pasalu IC, Sharma NP, Prasada Rao U and Mishra B (2003) Gall midge resistance in Rice: Current Status in India and Future Strategies. DRR Research Paper Series 01/2003

Berzonsky W, Shanower T, Lamb R, McKenzie R and Ding H (2002) Breeding wheat for resistance to insects. Plant Breed Rev 22: 221-97

Biradar SK, Sundaram RM, Thirumurugan T, Bentur JS, Amudhan S, Shenoy VV, Mishra B, Bennett J and Sharma

NP (2004) Identification of flanking SSR markers for a major rice gall midge resistance gene *Gm-1* and their validation. Theor Appl Genet 109: 1468-1473

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M and Ewan M (2000) Gene expression analysis of massively parallel signature sequencing (MPPS) on microbead arrays. Nat Biotechnol 18: 630-634.

Collins NC, Webb CA, Seah S, Ellis JG, Hulbert SH and Pryor A (1998) The isolation and mapping of disease resistance gene analogues in maize. Mol Plant-Microbe Interaction 11:968-978

Ghazi IA, Srivastava PS, Dalal V, Gaikwad K, Singh AK, Sharma TR, Singh NK and Mohapatra T (2009) Physical mapping, expression analysis and polymorphism survey of resistance gene analogues on chromosome 11 of rice. J Biosci 34: 251-261

Goring DR, Salt JN, Stone SL, Shiu SH and Mudgil Y (2004) A Large Complement of the Predicted Arabidopsis ARM Repeat Proteins Are Members of the U-Box E3 Ubiquitin Ligase Family1**.** Plant Physiol 134: 59-66

Graham MA, Marek LF, Lohnes D, Cregan P and Shoemaker RC (2000) Expression and genome organization of resistance gene analogues in soybean. Genome 43: 86-93

He L, Du C, Covaleda L, Xu Z, Robinson AF, Yu JZ, Kohel RK and Zhang HB (2004) Cloning, characterization and evolution of the NBS-LRR-Encoding resistance gene analogue family in polyploid cotton (*Gossypium hirsutum* L.). Mol Plant Mic. Int 17:123-1241

Iandolino A, Nobuta K, Goes da Silva F, Cook DR and Meyers BC (2008) Comparative expression profiling in grape (*Vitis vinifera*) berries derived from frequency analysis of ESTs and MPSS signatures. BMC Plant Biology 8:53-72

Jain A, Ariyadasa R, Kumar A, Srivastava MN, Mohan M and Nair S (2004) Tagging and mapping of a rice gall midge resistance gene *Gm8*, and development of SCARs for use in marker-aided selection and gene pyramiding. Theor Appl Genet 109: 1377-1384

Kanazin V, Frederick ML and Shoemaker RC (1996) Resistance gene analogues are conserved and clustered in soybean. Boc Natl Acad Sci USA 93: 11746-11750

Katiyar SK, Tan Y, Huang B, Chandel G, Xu Y, Zhang Y, Xie Z and Bennett J (2001) Molecular mapping of gene *Gm6 (t)* which confers resistance against four biotypes of Asian rice gall midge in China. Theor Appl Genet 103: 953-961

Katiyar, S.K., Verulkar, S.B., Adsul, G., Dudhare, M., Chandel, G. and Bennet, J. (2000) Molecular markers for gall midge resistance genes in rice: Stage set for MAS and map based cloning. Paper presented at 4th International symposium on rice genetics, International Rice Research Institute, Los Banos, Philippines, 22-27 October 2000

Ken N, Feng Z, Takuji T, Hiroki SC, Nathan H, Richardson AO, Okumoto Y, Tanisaka T & Wessler SR (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461: 1130-1134

Mago R, Nair S and Mohan M (1999) Resistance gene analogues from rice: Cloning, sequencing and mapping. Theor Appl Genet 99: 50-57

Mahalingam R, Gomez-Buitrago A, Eckardt N, Shah N, Guevara A, Day P, Raina R and Fedoroff NV (2003) Characterizing the stress/defense transcriptome of *Arabidopsis*. Genome Biology 4: R20.1-R20.14

Maleki L, Faris DS, Bowden RL, Gill BS and Fellers JP (2003) Physical and Genetic Mapping of Wheat Kinase Analogues and NBS-LRR Resistance Gene Analogues. Crop Science 43:660-670

Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW and Young ND (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily**.** Plant J 20: 317-332

Meyers BC, Galbraith DW, Nelson T and Agarwal V (2004a) Mehods of transcriptional profiling in plants: Be fruitful and replicate. Plant Physiol 135: 637-652

Meyers BC, Tej SS, Vu TH and Haudenschild CD**,** Agrawal V, Edberg SB, Ghazal H and Decola D (2004b) The Use of MPSS for Whole-Genome Transcriptional Analysis in *Arabidopsis.* Genome Res 14: 1641-1653

Mochida K and Shinozaki K (2010) Genomics and Bioinformatics Resources for Crop Improvement. Plant Cell Physiol 51(4): 497-523

Mohan M, Nair S, Bentur JS, Prasada Rao U and Bennett J (1994) RFLP and RAPD mapping of the rice *Gm-2* gene that confers resistance to biotype 1 of gall midge (*Orseolia oryzae*). Theor Appl Genet 87: 782-788

Noel L, Moores TL, van Der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JD (1999) Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. Plant Cell 11:2099-2112

Sardesai N, Kumar A, Rajyashri KR, Nair S and Mohan M (2002) Identification and mapping of an AFLP marker linked to *Gm-7*, a gall midge resistance gene and its conversion to a SCAR marker for its utility in marker aided selection in rice. Theor Appl Genet 105: 691-698

Sasaki T, Matsumoto T, Antonio BA and Nagamura Y (2005) From Mapping to Sequencing, Post-sequencing and Beyond. Plant and Cell Physiology 46(1):3-13

Wang X, Wu W, Jin G & Zhu J (2005) Genome-wide identification of R genes and exploitation of candidate RGA markers in rice**.** Chinese Science Bulletin 50: 11 1120-1125

Yencho GM and Byrne PF (2000) Application of tagging and mapping insect resistance loci in plants. Annu Rev Entom 45: 393-422

Yuan Q, Quackenbush J, Sultana R, Pertea M, Salzberg SL and Buell CR (2001) Rice Bioinformatics: Analysis of Rice Sequence Data and Leveraging the Data to Other Plant Species. Plant Physiol 125: 1166-1174