Plant Omics Journal POJ

POJ 4(3):126-135 (2011) ISSN:1836-3644

In silico mining and characterization of novel SSRs and candidate genes within QTLs controlling grain protein contents using MPSS signatures and micro array analysis in rice (*Oryza sativa* L.)

Girish Chandel[*], M. Dubey, P. Samual and R. Meena

Department of Plant Molecular Biology and Biotechnology, College of Agriculture, Indira Gandhi Krishi Vishwavidyalaya, Raipur 492 006 Chhattisgarh India

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

Abstract

Protein content in rice grain (*Oryza sativa* L.) is an important trait from human nutrition perspective, particularly for those having rice as a main food in daily life. Several QTLs identified for the grain protein content (GPC) needs refinement and further genetic dissection to truly understand the trait. In this study we have searched for the putative candidate genes underlying five known QTLs, (AQT033, AQT034, AQT037, AQT039 and AQT040) governing high grain protein content in rice. Important putative candidate genes encoding glutelin precursor, peptide transporter, aminotransferases etc were found underlying selected QTLs. The *in silico* expression analysis of candidate genes by massively parallel signature sequence (MPSS) revealed very strong expression for gutelin precursor gene and higher expressions for phosphoesterase, peptide transporter, aminotransferases and calmudulin dependent protein kinase genes. The tissue library information revealed their higher expression in pollen, stigma, immature panicles, germinating and developing seed tissues at reproductive stage. Further characterization of the candidate genes by digital microarrays in the reproductive development stage resulted in identification of genes showing higher seed specific expression. Further, we assessed the abundance of simple sequence repeats (SSRs) in the candidate genes as well as genomic, ESTs and cDNA sequences underlying QTLs. A total of 483 SSRs including 113 SSRs in the genomic, 133 in the cDNA and 237 SSRs in the EST sequences were identified. According to sequence length, the potentially variable Class II SSRs were the most commonly found microsatellites, followed by hyper variable Class I markers. Seventeen QTL specific microsatellite markers have been developed from the genomic, cDNA and EST regions. The identification of novel microsatellite markers and putative candidate genes based on MPSS signatures and digital micro arrays in this study will help in the development of gene specific markers for marker-assisted mapping as well as discovery of novel GPC related genes in rice.

Keywords: *In silico* mining; candidate gene; digital expression analysis; microsatellites; micro arrays.
Abbreviations: EST- expressed sequence tags; Fl-cDNA- full length complementary DNA; GPC- grain protein content; MPSS- massively parallel signature sequencing; QTL- quantitative trait loci.

Introduction

Rice has been a model plant for almost all genomic and molecular biology research owing to its small and compact genome. This research is important because the fruits of such research are going to affect major shift in food productivity and human nutrition as rice feeds more than half of the world population (Parvez and Rather 2007). Protein content in rice grain (*Oryza sativa* L.) is an important trait for health of people whose main food in daily life is rice (Shi et al., 1999). But poor grain protein contents in rice is an important cause of widespread protein malnutrition among rice eating populations especially those residing in developing nations (Lozoff et al*.,* 2006). In India about 47% of children are suffering from protein energy malnutrition (PEM) with infants suffering more from clinical or sub-clinical levels of protein deficiency (UNICEF India, 2005). Enhancing GPC of rice is a recent food based approach that has gained attention not only of nutritionists and crop biologists but also of renowned economists all over the world (Harvest Plus, 2003). GPC is a complex trait showing additive effects of multiple genes and considerable G×E interaction (Singh & Singh, 1982). Several QTLs for GPC have been identified and mapped on different chromosomes of rice genome using molecular markers (Lang & Buu, 2005 and Yoshida et al., 2002). But as the QTLs refers to the larger genetic region having several genes, fine mapping or high resolution mapping of QTLs is necessary to truly understand the quantitative variation and genes underlying them affecting GPC. Among the several marker systems, simple sequence repeat (SSR) or microsatellite markers are efficient, cost effective and have shown significantly higher degree of polymorphism in rice. Microsatellite combines several features of an ultimate molecular marker and are used increasingly in various plant genetic studies and applications (Rahman et al., 2009). Among these, now a day a new generation of SSR markers, derived from the coding sequence of genes, including EST and cDNA are being used by plant biologists. These markers circumvent other molecular marker related limitations such as large physical distance between genetically close marker, genes and also recombination between them. The candidate genes or DNA

Table1. Features of Quantitative Trait Loci known to govern GPC (grain protein content in rice) trait in rice (Yoshida et al., 2002)

Sr. No.	QTL	Chromosome	Position on rice chromosomes	Size of the QTL (Kb)	Number of ESTs present	Number of FL-cDNA sequences present
1	AQT033	2	7606893-7907107 bp	300.2	78	27
2	AQT034	3	28017654-28097852 bp	80.2	46	15
3	AQT037	1	37042027-37072164 bp	30.13	32	11
4	AQT039	4	30700432-30750584 bp	50.15	6	5
5	AQT040	11	21960818-21990657 bp	29.8	35	3

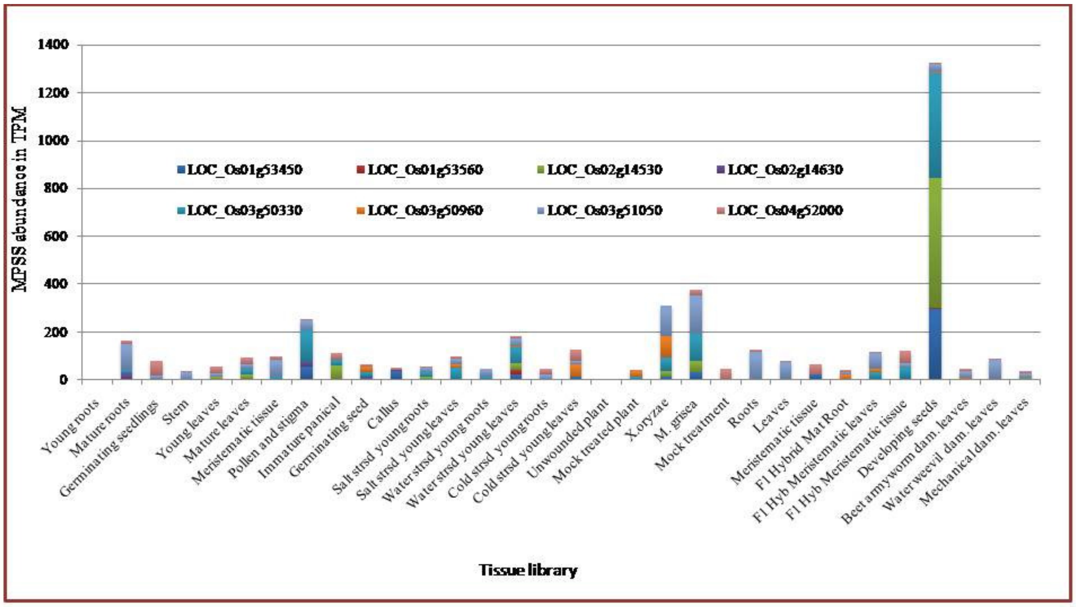

Fig 1. Expression pattern based on MPSS signature abundance in different tissue libraries for putative candidate genes underlying GPC trait related QTLs in rice.

sequences with predicted functions within a given QTL serve as an important source to generate novel and informative molecular markers. The candidate gene based molecular markers are likely to show stable association with the trait across the mapping populations (Thorup et al., 2000). Further the bioinformatics platforms available for transcriptome analysis will help in the digital expression profiling of the putative candidates genes and thus in the identification of most suitable and selective targets for further manipulations. Two advanced transcriptome analysis platforms preferably used today include MPSS and GeneChip arrays or microarrays. MPSS is a sequence tag-based platform which measures spatio-temporal expression of a gene (Nakano et al., 2006). While, DNA microarray or GeneChip array is an ultra high throughput analysis tool which provides digital measurement of time and organ specific gene expression (Mochida and Shinozaki, 2010). In this study we have characterized the sequences underlying five QTLs namely AQT033, AQT034, AQT037, AQT039 and AQT04 (Yoshida et al., 2002), known for governing high grain protein content in rice. The genomics regions specific to all five GPC QTLs were searched initially for the presence of putative candidate genes and later the spatio- temporal expression of these genes has been analyzed using MPSS signatures analysis. The sites of expression of candidate genes in different developmental stages and plant tissues/ organs have also been identified. The findings of MPSS signature analysis were further confirmed with the gene expression data available in rice array database using digital microarray analysis. Finally, seventeen microsatellite markers based on the genomic region, candidate genes, EST's and cDNA sequences were developed for the five GPC QTLs in rice.

Materials and methods

Selection of the target QTLs and in silico analysis of the QTL region

Five QTLs having major effect on grain protein contents in brown and white (polished) rice grains, identified by Yoshida et al. (2002) in a doubled haploid population derived from a cross of Reiho and Yamada-nishiki were selected for the study. These included two QTLs governing high grain protein content in brown rice grains namely AQT033, AQT034 and three QTLs namely AQT037, AQT039 and AQT040 governing high grain protein content in white rice. The physical position of all the QTLs on different rice chromosomes were obtained based on the position of flanking makers from the Gramene annotated Nipponbare Sequence 2009 map set available at gramene website (*www.gramene.org*). The nucleotide sequences varying from 29.8Kb to 300.2Kb underlying QTLs were downloaded as BAC clones and contigs from TIGR, Genome browser (*http://www.tigr.org/tdb/e2k1/osa1/*) & Gramene database (*http://www.gramene.org*) and stored in FASTA formatted text files. The QTL regions were then analyzed for the

Table 2. Putative candidate genes encoding proteins of known function that may possibly be responsible for seed accumulation of proteins in rice

QTL	Name of the gene underlying QTL	Gene ID	Putative functional description	Coordinate (5'-------3')	Protein length (aa)^1	MPSS signatures abundance (TPM)^2
AQT033	Phosphoesterase, putative, expressed	LOC_Os02g14530	Protein modification process	8007608 - 8017026	791	5642
	Glutelin precursor, putative, expressed	LOC_Os02g14600	Nutrient reservoir activity, seed storage protein	8057685 - 8059532	500	22590
	Hydroquinone glucosyltransferase, putative, expressed	LOC_Os02g14630	Biosynthetic process, transferase activity and amino acid and derivative metabolic process	8077469 - 8075973	499	55
AQT034	CAMK_KIN1/SNF1/ CAMK includes calcium/calmodulin depedent protein kinases, expressed	LOC_Os03g50330	Protein modification process	28708440 - 28702287	427	634
	LTPL 118, LTPL family protein	LOC_Os03g50960	Protease inhibitor/seed storage/LTP family protein precursor, expressed	29103474 - 29103845	124	254
	Peptide transporter PTR2	LOC_Os03g51050	Transporter activity	29181687 - 29178349	594	1168
AQT037	Aminotransferase, classes I and II, domain containing protein, expressed	LOC_Os01g53450	Biosynthetic process, protein binding and transferase activity	30713051 - 30716381	451	551
	Aminotransferase, classes I and II, domain containing protein, expressed	LOC_Os01g53560	Protein modification process	30752465 - 30759908	459	26
AQT039	Protein phosphatase 2C, putative, expressed	LOC_Os04g52000	Protein modification process	30694096 - 30690456	322	503
AQT040	-	-	-	-	-	-

1. aa- amino acid residues, 2. TPM- transcript per million

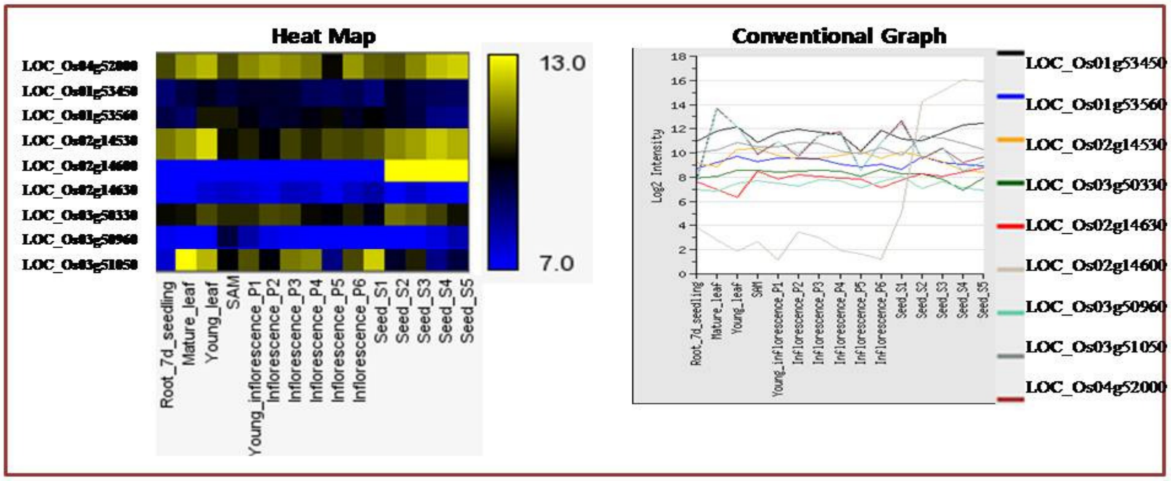

Fig 2. Spatio-temporal expression profiles of putative candidate genes for GPC trait in rice during reproductive development generated at rice array database tool site (www.ricearray.org). The stages of development have been marked at the bottom, SAM: Shoot apical meristem; Inflorescences P1-P6: Temporal stages of inflorescences development; Seed S1-S5 Temporal stages of seed development. Color bar at the right represents $\log_2$ signal values, blue representing low-level expression, black medium and yellow signifies high-level expression.

presence of putative candidate genes, Fl- cDNA (full length cDNA sequence) and co localized ESTs from Rice genome browser at TIGR website (*www.tigr.org*), and these sequences were also downloaded and stored as separate files. The genomic sequences, ESTs and Fl-cDNA sequences were then subjected for SSR identification.

Search for putative candidate genes underlying QTLs controlling grain protein content in rice

The genomic region underlying selected QTLs were searched for putative candidate genes by scanning all the annotated genes present in the target region from TIGR, Genome browser (*http://www.tigr.org/tdb/e2k1/osa1/*). This approach was based on exploiting the information on the role and functions of a particular coding sequence and hypothesizing a plausible cause-effect relationship between the QTL and a feasible candidate gene mapping nearby (Pflieger et al., 2001 & Tuberosa and Salvi, 2007). Putative genes potentially serving for the accumulation/ deposition of proteins in rice grains, mobilization or transport of protein from the source to sink tissues or involved in the activities of modification of proteins for their seed storage were considered as candidate genes for the GPC trait.

In silico analysis of putative candidate gene expression by MPSS signatures and digital microarrays

Further, the identified genes were functionally characterized by analyzing their expression employing two *in silico* based approaches. First approach was by identifying MPSS (massively parallel signature sequencing) tags, co localizing with the candidate genes and analyzing their expression in different tissue libraries. Another approach was the *in silico* micro array based approach. MPSS tag based profiling offers great opportunities for *in silico* applications in functional characterization of genes using web based tools (Dubey and Chandel, 2010). Identifying MPSS signature sequences co localizing with a gene can yield valuable information about putative spatial or temporal expression of that gene (Banerjee et al., 2010). 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 signature tags and information on tag positions, chromosome coordinates *etc*. The sequence of each putative candidate gene was used as query under 'query by sequence' section of rice MPSS database to identify co localized MPSS tags and their expression in 22 diverse tissue libraries from rice MPSS database (this database includes libraries constructed from various developmental stages, tissue types and tissues treated by various biotic and abiotic stresses). The abundance/ frequency of each tag is expressed in TPM (transcript per million) which is considered as the measure of expression in a corresponding tissue library. The second approach was digital expression profiling of putative candidate genes by using microarray data from Rice Array database analysis tool site (*www.ricearray.org*). Based on the spatio- temporal expression data generated from the MPSS analysis, the microarray analysis was carried out particularly under the series accession number GSE6893 comprising expression data for reproductive development in rice. At this analysis tool, the locus identifier of each gene was used as query and 'Affymetrix GeneChip experiment' platform was selected to download the expression data of all the genes for the reproductive development stage as described by Arora et al. (2007). The results were obtained as matching probes from the Affymetrix array. The data were in the form of Log_2 transformed signal values generated from the average of three biological replicates. This data is further used to perform the heat map of normalized signal intensity values, for each gene which provides a quantitative measure of the transcript of a particular gene and hence its expression.

Mining candidate genes and the QTL region for simple sequence repeats

Simple sequence repeats (SSRs) loci were identified in the candidate genes, genomic region of QTLs controlling grain protein content as well as from the regions flanking ESTs and Fl-cDNA sequences present in the QTL region using SSRIT tool available at Gramene database (*http://www.gramene.org/db/searches/ssrtool*). The criteria set for SSR identification was 2-6 nt repeat unit. SSRs were obtained in the query sequence with details of repeat motif, number of repeat units, repeat length, SSR start and SSR end point.

PCR primer design

Genomic DNA based SSR primers were designed from the region flanking SSR repeat motifs present in the candidate gene sequences, other genomic, EST and cDNA sequences underlying QTLs. Similarly, EST derived SSRs (EST-SSRs) were designed from the region flanking SSR repeat motifs present within the EST sequences and cDNA-SSRs from the repeat motifs present within the Fl-cDNA sequences. The primers were designed using Primer-3 software (*http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi*) and the specifications included 18-22 bp of primer length, Tm range of 45-60 °C and 50-400 bp of product size. All other options were left on default value. The primer designing was restricted to the SSRs belonging to hyper variable Class I and potentially variable Class II category consisting of SSRs length ≥20 bp and SSRs length =12 bp to <20 bp respectively (Temnykh et al., 2001). The QTL regions were partitioned in to 3-4 segments of equal sequence length and SSR motifs from all the representative segments were targeted for primer designing to have maximum coverage of the QTL region.

DNA preparation, PCR protocol and DNA marker analysis

DNA was extracted from fresh rice leaves of parents and F_4 population plants as per the method described by Dellaporta et al. (1983). The F_4 population was derived from two parents differing for GPC, namely IR 68144-3B (high GPC) and Swarna (low GPC), developed at Department of Plant Molecular Biology and Biotechnology, IGKV, Raipur for breeding high protein rice (Chandel et al., 2005). The PCR amplification was carried out using a Corbett Research palm cycler PCR System in 20μl reactions. PCR thermal profile included an initial denaturing step of 10 min at 94°C, followed by 35 cycles with denaturation at 94°C for 30 s and extension at 72°C for 30 s. The annealing temperature was set in a range of 45-60 °C depending upon individual primer sequence. After 35 cycles, a final extension step was performed at 72°C for 5 min. The PCR products were then electrophoresed on 2.5% agarose gels containing ethidium bromide, at 80 V for 1 h and observed under a UV transilluminator. Bands were scored as A and B allele on the basis of length differences to generate SSR profile of individual markers.

Table 3. Number of SSRs according to repeat motifs in the genomic, cDNA and EST derived SSRs from the GPC related QTL regions

Repeat motif	Genomic SSRs	EST SSRs	cDNA SSRs
Dinucleotide repeats			
AG / TC	11	10	14
CG / GC	5	6	5
AC / TG	9	7	5
AT / TA	12	8	7
CT / GA	14	9	11
GT / CA	6	11	12
Total dinucleotides	57	51	54
Trinucleotide repeats			
CAG / GTC	7	-	-
TCT / AGA	-	14	11
GAC / CTG	5	12	16
GGC / CCG	10	9	16
CGC / GCG	11	5	12
CAC / GTG	2	11	14
GAT / CTA	4	12	10
Total trinucleotides	41	63	79
Tetranucleotide repeats			
ATAG / CTAT	-	-	-
ATCT / AGAT	1	2	-
CATA / TATG	-	1	7
CCAT / ATGG	1	-	1
CCTC / GAGG	-	-	1
CGAT / ATCG	-	-	-
CGTT / AACG	-	-	9
CTCC / GGAG	-	2	-
CTGG / CCAG	2	1	10
GCAT / ATGC	-	1	
GCCA / TGGC	-	-	10
GTGG / CCAC	1	2	1
TATT / AATA		3	8
TCCC / GGGA	4	1	10
TTCT / AGAA			
Total tetranucleotides	9	13	57
Petanucleotide repeats			
AAAAT / ATTTT	2	2	10
AATGA / TCATT	1	1	9
TGAGC / CGTCA	-	1	9
TCCCC / GGGGA	2	1	10
GCCGC / GCGGC	1	1	9
Total pentanucleotides	6	6	47
Total number of SSRs	113	133	237

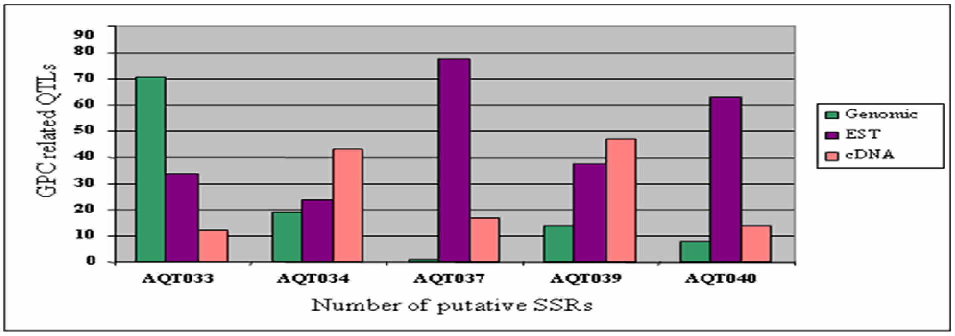

Fig 3. Summary of SSRs present in the region of five known QTLs governing GPC trait in rice.

Results and discussions

In silico analysis of the QTL region

The nucleotide sequences underlying selected QTLs were downloaded as BAC clones and contigs and analyzed *in silico.* Great variation in the size, number of co-localized ESTs and Fl-cDNA sequences was observed among five QTLs. The details of the QTLs including chromosomal location, position and marker interval are presented in Table 1. Mining the region for the ESTs showed the presence of a total of 197 ESTs over five QTLs. Similarly a total of 62 Fl-cDNA sequences were found in the region of five QTLs. The QTL AQT033 showed highest number of co-localized ESTs (78) as well as Fl- cDNA sequences (27), whereas AQT039 and AQT040 showed minimum number of co-localized ESTs (6) and Fl-cDNA sequences (3) respectively (Table 1). The genomic sequences, ESTs and Fl- cDNA sequences were then subjected to SSR identification.

Search for putative candidate genes in the QTL regions

The candidate gene approach exploits information on the role and functions of a particular coding sequence and verifies whether it may represent a feasible candidate for QTL in question or not (Pflieger et al., 2001). If a plausible cause-effect relationship can be hypothesized between a QTL and a candidate gene mapping nearby, then the validation of its role could be attempted (Tuberosa and Salvi, 2007). With this concept, a search was carried out to detect the presence of candidate genes related to grain protein content underlying the target QTL regions. The QTL AQT033 showed the presence of precursor for reported gene for grain protein content in rice 'glutelin'. Protein fraction in rice grains (seed storage proteins) includes glutelins, globulins and prolamins with gutelins forming the major fraction representing 80% of the total seed storage proteins in rice (Shewry and Halford, 2002). Further the glutelin gene sequences were subjected to protein domain search at TIGR website (*www.tigr.org*) which revealed that, it carries a conserved cupin domain belonging to cupin superfamily of proteins. We also retrieved the sequences of other genes encoding globulins and prolamin seed proteins in rice as well as their precursors present anywhere on rice chromosomes and analyzed their protein domains. It was found that prolamins mainly contained LTPL domain belonging to LTP (lipid transfer protein) protein family whereas, globulins mainly contained hair pin induced and membrane protein domains. With these findings, the search was extended to find putative genes carrying cupin, LTPL protein domains/ motifs which can potentially serve as candidate genes for protein related traits. AQT034 showed the presence of a putative gene encoding LTPL118 domain containing protein (Table 2). LTPL118 serves for the protease inhibitor activity, seed storage and as precursor for other LTP family protein. Another gene PTR2, encoding peptide transporter protein involved in the transporter activity was also found mapping nearby underlying the same QTL. Two putative genes encoding aminotransferase, classes I and II, domain containing protein were found in very close proximity underlying QTL AQT037. Similarly, a putative gene encoding protein phosphatase 2C was identified in the region underlying AQT039 (Table 2). No putative candidate genes were found underlying QTL AQT040. This might be due to the smaller span of the QTL on the chromosome or due to the unavailability of the complete annotation of the genes present in this region. Employing a similar approach, Ravel et al. (2006) identified *Glu-B1-1* as a candidate gene underlying QTL related to the quantity of high-molecular-weight glutenin in bread wheat (*Triticum aestivum* L.) by means of an association study using SNP derived markers. Similarly, Wang et al. (2008) characterized the QTL controlling amino acid content in grains of rice and identified various co-localized candidate loci including glutelins, prolamin, globulin precursor and aminotransferases *etc*.

Spatio- temporal expression analysis of putative candidate genes by MPSS signatures

Understanding the molecular mechanism which determines the synthesis of grain proteins (seed storage proteins), their trafficking and accumulation in the grain by directing their deposition in specialized structures, called protein bodies, is important to underpin future attempts to improve grain protein content. This is a complex and tightly regulated mechanism involving many molecular players. Thus, there is a need to selectively target the putative candidates for their characterization. Prior characterization through *in silico* approaches will help in identifying the active players involved in a complex mechanism. The transcript accumulation of candidate genes across a wide range of tissues/organs and developmental stages of rice were analyzed employing two *in silico* based approaches, MPSS and microarray analysis. *In silico* MPSS analysis revealed that all the nine putative candidate genes showed the presence of corresponding 17 base signature tags. This finding suggests that most of the candidates are expressed genes. The number of 17 bp signature tags for the candidate genes varied significantly, similarly the cumulative TPM values for all the tags co-localizing with each gene sequence also varied from moderate (TPM value 26-500) to strong (TPM >500) expression (Meyers et al., 2004). High TPM tags corresponding to six out of nine putative candidate genes were found (Table 2) with MPSS tags corresponding to LOC_Os02g14600 encoding putative precursor for seed storage protein glutelin, showing the highest cumulative TPM value of 22,590. The tissue library wise expression of putative candidate genes revealed that although the genes expressed in diverse tissue libraries, significant expressions were observed in reproductive tissues, mature roots, young leaves under abiotic stress (water, salt and cold stressed) and rice leaves challenged with *Xanthomonas* and *Magnaporthe.* The reproductive stage/ tissue specific expressions were more pronounced in pollen, stigma, immature panicles, germinating and developing seeds. Among the reproductive stages, maximum level of expression based on TPM value was observed in developing seeds which was a common observation for the majority of genes (Fig. 1, showing tissue library wise expression of all the putative candidate genes except glutelin precursor). These genes included LOC_Os02g14530, LOC_Os01g53450, LOC_Os03g50330 and LOC_Os03g51050 encoding hydroquinone glucosyltransferase, aminotransferase, CAMK_KIN1/SNF1/ CAMK like genes and peptide transporter PTR2 genes respectively. In contrast, the gene encoding protein phosphatase 2C (LOC_Os04g52000) showed expression in almost all the tissues libraries of the collection. This finding is attributed to the their property of reversible phosphorylation activity of proteins which is a fundamental mechanism by which living organisms modulate cellular processes including cell cycle, growth factor & hormone and environmental stimuli responses, metabolic control *etc* (Kerk et al., 2002). Exceptionally higher expression level (TPM value 22,590) was observed for glutelin precursor (LOC_Os02g14600) which expressed exclusively in devel-

Table 4. Primer sequences to amplify SSRs derived from the candidate genes and other genomic, cDNA and EST sequences present in QTL regions.

SSR Marker	Primer Forward 5’→ 3’	Primer Reverse 5’ → 3’	Tm (°C)	Expected product size (bp)
cgRM 33-1 (LOC_Os02g14530)	TCGAACATTGAACGATGGAA	AATAAACAAAGGCCCGGTTC	60	226
cgRM 34-1 (LOC_Os03g50960)	GCGATTAGCCAGAGATCGAC	TCCACGGCAAACTTACACAC	59	176
cgRM 34-2 (LOC_Os03g51050)	TGGCCTGCTGTACTCTTTCC	CAGTTCGGTTCAGCAGTTCA	60	162
cgRM 37-1 (LOC_Os04g52000)	GTGGAAGAATCGGATCAGGA	CAATCACCACGGACACAGAC	58	151
cgRM 39-1 (LOC_Os01g53450)	CTCTCTCTTCGCGTCCTGTC	CTGCCCTAATCCAAGCAAAC	59	208
gRM 33-1	TCGTTCTGACATGTTGAGG	CCCGACAAAGTCAACGTC	53	250
gRM 33-2	AACGACATGCAAAATGAGAG	AAGAGGAGATTCCATGTTCA	51	249
gRM 33-3	TGTCTAATTGCAGAATGCAG	CACTAGGACATGGTACGATT	51	230
gRM 34-1	ATAGTATGCCCAGCATTAGC	TGTTCTCTCTGCACTTGTTG	52	209
gRM 37-1	TCACGGAGCTCGTACTTG	ACCTTCCGATCTGGAGTC	55	245
gRM 39-1	TTGCTCCCTGTCACTTAGAT	TCTTTTGCCCACTTCAATAC	52	290
gRM 39-2	GTGATGTGATGTGATGGAAA	CACCTCCAGGATCTCGTC	52	300
gRM 40-1	AACATTTTCAACCCAAGACAA	TCCCTCCCTTTTCAGGCTAT	54	232
eRM 33-1	GCCGACGTTGACTTTGTC	AAAGCGAGACACCTTTTCTT	53	260
eRM 34-1	AATCATCAGGGAACACACTT	GAAAGGAAAAAGGACAGGT	51	189
cRM 34-1	ATGTCTAACATGGTGGCTTG	CGCTTTGAAGGATTTGAATA	54	176
eRM 37-1	AGTGGGAAGATGACTTCTTT	CGGTGTTCTACAACGTGAC	53	367
eRM 39-1	GAGCCAAGAGATGAGTTTCA	AGGACGAATCAGACAAACAG	52	289
cRM 33-1	ATAAGTGGCATCCTTTGGTT	ACAGGACAAGCGTTACAAGA	53	200
cRM 37-1	TTCCTCTGTCAGTGAAATGG	GGACCATGAACATTCAGAAG	54	190
cRM 39-1	CATGCTCTGAAAGGTTCTTG	CAGCCAGATGTACTCTCCAG	53	282
cRM 40-1	CTTGTGTTTTGGACTGCTTC	CCACTTTCTGCTGACACTC	53	215

cgRM- candidate gene based SSR, gRM- genomic region SSR, eRM- EST derived SSR and cRM- cDNA derived SSR markers.

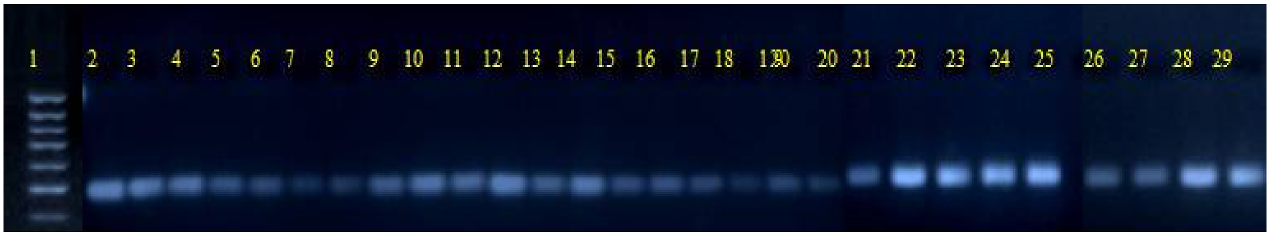

Fig 4. PCR products of segregating individuals showing polymorphism in amplicon size, amplified by using eRM 37-1 marker. Lane 1: 100 bp ladder, Lane 2: IR68144-3B, Lane 3: Swarna and Lane 4-29: F_4 individuals

oping seeds of rice. Using similar approach, significantly higher expression of a metal transporter gene OsZIP9 have been observed in reproductive plant parts of *indica* rice by Chandel et al. (2010). MPSS signature analysis has also been used for the spatio-temporal expression analysis of number of metal homeostasis related candidate genes in rice to identify the putative site of expression of these genes (Banerjee et al., 2010).

In silico expression analysis of putative candidate genes for grain protein content by microarrays

Microarray represents a high throughput means to analyze the expression of a gene and to identify genes involved in a particular biological process. In order to confirm the higher reproductive stage specific expression of the putative candidate genes observed in MPSS signature analysis, the study was further extended to analyze the expression by a second approach based on digital microarrays at Rice array database (*www.ricearray.org*). The series accession number GSE6893 which includes microarray data from 45 hybridizations in the reproductive development stages and organs of rice was selected as the experiment type. The transcript/ expression levels were generated as Log_2 transformed signal values generated from the average of three biological replicates for each tissue library and a heat map of normalized signal intensity values, corresponding to the different organs of the plant, for each gene (Jung et al., 2008) were obtained. Distinct transcript abundance patterns of putative candidate genes were readily identified in the microarray data analyzed (Fig. 2). Many of the genes showed preferential accumulation of transcripts in a given tissue/organ or developmental stage. The analysis revealed that all the genes except LOC_Os02g14630 and LOC_Os03g50960, showed moderate to higher levels of expression based on their signal intensity values. Among the different plant organs, it was observed that majority of the candidate genes expressed more in various temporal stages of seed development, young and mature leaves and various stages of inflorescences development. In this analysis a peculiar expression pattern was observed for the gene encoding glutelin precursor (LOC_Os02g14600), showing exclusively strong seed specific expressions. This finding is in exact confirmation with the MPSS results in which very higher MPSS abundance (TPM) were observed exclusively in developing seeds of rice. The finding is quite obvious as glutelin is a nutrient reservoir protein in rice seeds and is expressed during the seed development stage. The microarray expression results of all the putative candidate genes have been depicted as Heat map generated by the Rice array database tool in figure 2 and alternatively the same has also been displayed as conventional expression graph. Similar findings have been observed by Kawaura et al. (2005) where seed specific expression of two major storage proteins gliadins and glutelins have been observed in wheat, when the expression patterns of these genes were estimated based on the frequencies of ESTs. Using similar approach, Ray et al. (2007) analyzed the expression of calcium-dependent protein kinase gene family during reproductive development and abiotic stress conditions in *indica* rice by Affymetrix GeneChip hybridization experiments and observed differential expression levels based on Log_2 signal intensity values. Similarly, Arabidopsis Affymetrix GeneChip® average data available on the Genevestigator analysis tool site has been utilized for the identification of genes specifically expressed in seeds during early development in *Arabidopsis* by Becerra et al. (2006). They reported maximum expression of genes encoding nutrient reservoirs like cruciferin 12S seed storage protein, oleosin, glycine-rich protein and 2S seed storage proteins during seed development particularly in seeds with embryos at the walking-stick to early curled-cotyledon stages and seeds with embryos at the curled-cotyledon to early green-cotyledon stages. The microarray dataset expression results were found in confirmation with the MPSS abundance in terms of TPM values for all the genes except for the gene encoding aminotransferase, classes I and II, domain containing protein (LOC_Os01g53560). This gene was found to show moderate expression levels by microarray analysis but showed lower expression based on low TPM value of 26 in the MPSS dataset. Putative gene encoding hydroquinone glucosyltransferase (LOC_Os02g14630) showed lower level expression both in terms of TPM value and microarray expression results. Thus, excluding these two genes, the remaining seven putative candidate genes were selected for candidate gene based marker identification.

Frequency, type and SSR-sequence distribution in the QTL region

The sequences of the selected putative candidate genes were subjected to SSR identification using SSRIT tool, SSR motifs under set criteria were detected only in the genomic sequences of putative genes encoding LTPL118, aminotransferase, phosphoesterase, CAMK_KIN1/SNF1 /CAMK protein kinases, peptide transporter PTR2 and Protein phosphatase 2C, while other genes showed very short repeat motifs and no flanking primer sequences. Apart from the identification of candidate gene based SSRs, a comprehensive analysis were performed on the frequency, type and SSR-sequence distribution in the entire the selected QTLs. This resulted in identification of a total of 483 putative SSRs in five QTLs and included 113 SSRs in the genomic region, 133 SSRs in the cDNA sequences and 237 SSRs in the EST sequences (Fig. 3). The repeat motif ranged from 2-5 nucleotide repeats (no hexanucleotide and above repeats were found in any of the query sequences and mono nucleotide repeats were not included in the study), whereas the total repeat length i.e. SSR length varied from 4-80 nucleotides. It was also observed that according to sequence length, Class II or potentially variable SSRs were the most common and constituted 39.2% of the total SSRs. This was followed by hyper variable Class I SSRs. The division of microsatellites into classes (Temnyhk et al., 2001) represents their potential as molecular markers. Class I repeats represents the hyper variable loci in the genome and they should be the starting point for the design of molecular markers as they are the most polymorphic SSRs. Class II are less variable (potentially variable) and class III SSRs have a mutation potential similar to most unique sequences. The SSR density was found to be lower in the genomic region as compared to the EST and cDNA sequences. Microsatellites have been reported to be more abundant in the non coding regions as compared to the coding exons. The lower SSR abundance in genomic region under study may be attributed to the occurrence of SSRs containing mononucleotide repeats in the genomic sequences and since mononucleotide repeats have not been included in the study, this might have resulted in considerable reduction in the number of genomic SSRs. Mononucleotides (A/T) are the most frequent repeat units in the genomes of *Arabidopsis* and *Brassica* which imparts most to their SSR density (Lawson & Zhang, 2006 and Hong et al., 2007). Comparing SSR abundance in ESTs and cDNA sequences, EST-SSRs were more abundant. ESTs often represent partial and redundant cDNA sequences, derived from the UTR regions, whereas cDNAs represent the exonic regions of a gene which reduces their frequency of occurrence, as the tandemely repeated microsatellites are less likely to be present in the actual coding sequences (Gupta et al., 2010). Higher SSR density has been reported in the 5' and 3' UTR regions of *Arabidopsis* and rice as compared to the exonic regions (Lawson and Zhang, 2006). Analysis of SSRs from the perspective of repeat number revealed that dinucleotide repeats were the most frequent motifs in genomic DNA based SSRs (Table 3). This finding agrees with Cardle et al. (2000) in a study on *Arabdopsis*, but differs from that of Varshney et al. (2002), who found trinucleotides as the most frequent repeats in cereals, followed by dinucleotide repeats. Trinucleotide repeat frequency was highest among cDNA and EST derived SSRs. The summary of different SSR motifs present in genomic, EST and cDNA derived SSRs are presented in Table 2. The higher occurrence of trinucleotide repeats in EST derived SSRs is in confirmation to Jayshree et al. (2006) where more common frequency of trinucleotide repeats have been reported among EST-SSR database of cereals and legumes. Similar reports have been made by Varshney et al. (2002) where the trinucleotide repeats, in the range of 54% to 78%, have been observed as the most abundant type of microsatellites in the EST sequences of the cereal species including rice. SSRs analysis from the perspective of type of repeat motifs revealed that among dinucleotide repeats, CT/GA was the most frequent repeat motif in the genomic SSRs, GT/CA in

EST-SSRs whereas AG/ TC was the most frequent dinucleotide repeat motif among cDNA derived SSRs. CG/ GC was identified as the least frequent motif which was the common observation for all the SSRs (Table 3). These findings are similar to what had been previously observed in plants such as *Arabidopsis thaliana*, wheat, barley, rice, maize, almond, peach and rose (Miyao et al., 1996; Cardle et al., 2000; Kantety et al., 2002 and Jung et al., 2005). Among the trinucleotide repeats, GCG/ CGC was the most frequent repeat motif in the genomic SSRs, TCT/ AGA in EST derived SSRs whereas the trinucleotide repeat motifs CGC/GCG and GAC/CTG were the most frequent motifs in the cDNA-SSRs and occurred with same frequency. No particular trend in the occurrence of repeat motif frequency was observed for tetra and pentanucleotide repeats.

Generation of novel candidate gene based/ QTL specific microsatellite markers

Twenty two (18-22 nucleotide long) primers were designed targeting the repeat motifs present in the putative candidate genes as well as other microsatellite loci present in the genomic, EST and cDNA regions corresponding to five QTLs (Table 4). These putative candidate genes were selected on the basis of MPSS and digital microarray expression results. Nineteen (eleven genomic, four cDNA and four ESTs based SSRs) out of twenty two designed primers successfully amplified scorable bands of expected product size while three primers produced poor or non specific amplifications. These primers were then used for genotyping the F_4 population derived from the parent IR68144-3B, identified as high protein rice and a popular *indica* rice variety Swarna, low in grain protein content. Seventeen out of 19 primers amplified single allele per locus while two primers namely gRM 37-2 and gRM 40-1 amplified two alleles per locus. The polymorphisms of individual primers based on length difference of amplified products were analyzed (Fig. 4). Thus, using the huge datasets available in the public domain, we have characterized the genomic region underlying five known QTLs governing GPC in rice for putative candidate genes and novel microsatellite loci. Identification of putative candidate genes underlying target QTLs and their digital expression analysis provides insight in to their functionality and putative sight of expression. Higher level expression of majority of the putative candidate genes during the reproductive phase and tissues and more pronounced expression in the developing seeds show their functional involvement in the complex phenomenon of biosynthesis, modification, transport and accumulation of proteins in the rice grains. Further, the generation of new generation candidate gene based and QTL specific microsatellite markers can be used for the saturation/ fine mapping of these QTLs and effective MAS programs for grain nutritive traits. Since microsatellites combines several features of an ultimate molecular marker including high information content and ease of genotyping (Rahman et al., 2009 and Prathepha, 2011), the markers along with the expression information generated in this study will not only give better understanding of the role of candidate genes in grain accumulation of proteins but, will also serve as platform for selection of genes. This will help in the further characterization of the genes and planning strategies for functional genomics approaches for grain improvement.

Acknowledgements

We acknowledge MSU Rice genome annotation project and Rice Array Database for the availability of the sequences and expression data information of rice. The financial support to undertake this research by Department of Information Technology (DIT), Government of India, New Delhi is thankfully acknowledged.

References

Arora R, Agarwal P, Ray S, Singh AK, Singh VP, Tyagi AK, Kapoor S (2007) MADS-box gene family in rice: genome-wide identification, organization and expression profiling during reproductive development and stress. BMC Genomics 8**:** 242

Banerjee S, Sharma DJ, Verulkar SB and Chandel G (2010) Use of *In-silico* and semiquantitative RT-PCR approaches to develop nutrient rich rice (*Oryza sativa* L.). Indian J Biotechnol 9: 203-212

Becerra C, Puigdomenech P and Vicient MC (2006) Computational and experimental analysis identifies Arabidopsis genes specifically expressed during early seed development. BMC Genomic 7:38

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D and Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156: 847-854

Chandel G, Banerjee S, Vasconcelos M and Grusak MA (2010) Characterization of the Root Transcriptome for Iron and Zinc Homeostasis-related Genes in Indica rice (*Oryza sativa* L). J Plant Biochem Biotech 19(2): 145-152

Chandel G, Dudhare M S, Saluja T, Shiva SM, Sharma Y, Geda AK, Sahu GR, Mishra V N and Katiyar SK (2005) Screening rice accessions for nutritional quality traits to achieve nutritionally balanced rice. In 5th International Rice Genetics Symposium. Nov. 19-23: 60-61

Dellaporta SL, Wood J and Hicks JB (1983) A plant DNA miniprepration: version II. Plant Mol Biol Rep 1: 19-21

Dubey M and Chandel G (2010) *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.). Plant Omics J 3(5):140-148

Gupta S, Shukla R, Roy S, Sen N and Sharma A (2010) *In silico* SSR and FDM analysis through EST sequences in *Ocimum basilicum*. Plant Omics J 3(4):121-128

Harvest Plus. Rice Processing protocol (2003) (*www. harvesplus.org*).

Hong CP, Piao ZY, Kang TY, Batley J, Yang TJ, Hur YK, Bhak J, Park BS, Edwards D and Lim YP (2007) Genomic distribution of simple sequence repeats in *Brassica Rapa.* Mol. Cells 23 (3): 349-356

Jung KH, Dardick C, Bartley LE, Cao P, Phetsom J, Canlas P, Seo YS, Shultz M, Ouyang S, Yuan Q, Frank BC, Ly E, Zheng L, Jia Y, Hsia AP, An K, Chou HH, Rocke D, Lee GC, Schnable PS, An G, Buell CR, Ronald PC (2008) Refinement of light-responsive transcript lists using rice oligonucleotide arrays: evaluation of gene-redundancy. PLoS ONE 3(10):e3337. PubMed

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2005) Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genomics 5: 136-143

Kantety RV, Rota ML, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48: 501-510

Kawaura K, Mochida K, and Ogihara Y (2005) Expression profile of two storage-protein gene families in hexaploid wheat revealed by large-scale analysis of expressed sequence tags. Plant Physiol 139: 1870-1880

Kerk D, Bulgrien J, Smith DW, Barsam B, Veretnik S, Gribskov M (2002) The complement of protein phosphatase catalytic subunits encoded in the genome of Arabidopsis. Plant Physiol 129(2): 908-925

Lang NT and Buu BC (2005) Genetic analysis of grain protein content in rice. Tap chi Nong Nghiep & PTNT 12: 14-20

Lawson MJ and Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes**.** Genome Biol **7:** R14

Lozoff B, Beard JL, Connor JR, Felt BT, Georjieff MK and Schallert T (2006) Long-lasting neural and behavioral effects of iron deficiency in infancy. Nutr Rev 64: S34-S43

Meyers BC, Tej SS, Vu TH and Haudenschild CD, Agrawal V, Edberg SB, Ghazal H and Decola D (2004) The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis.* Genome Res 14: 1641-1653

Miyao A, Zhong HS, Monna L, Yano M, Yamamoto K, Havukkala I, Minobe Y, Sasaki T (1996) Characterization and genetic mapping of simple sequence repeats in the rice genome. DNA Res 3: 233-238

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

Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW and Meyers BC (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 34: D731-D735

Parvez S and Rather AG (2007) QTL analysis in rice improvement: Concept, methodology and application. Biotechnology 6: 1-13

Pflieger S, Lefebvre V and Causse M (2001) The candidate gene approach in plant genetics: a review. Mol Breeding 7 (4): 275-291

Prathepha P (2011) Microsatellite analysis of weedy rice (Oryza sativa f. spontanea) from Thailand and Lao PDR. AJCS 5(1): 49-54

Rahman MS, Molla R, Md. Alam S and Rahman L (2009) DNA fingerprinting of rice (*Oryza sativa* L.) cultivars using microsatellite markers. Aus J Crop Sci 3(3): 122-128

Ravel C, Praud S, Murigneux A, Linossier L, Dardevet M, Balfourier F, Dufour F, Brunel D and Charmet G (2006) Identification of *Glu-B1-1* as a candidate gene for the quantity of high-molecular-weight glutenin in bread wheat (*Triticum aestivum* L.) by means of an association study. Theor_Appl Genet 112 (4): 738-743

Ray S, Agarwal P, Arora R, Kapoor S and Tyagi AK (2007) Expression analysis of calcium-dependent protein kinase gene family during reproductive development and abiotic stress conditions in rice (*Oryza sativa* L. ssp. *indica*). Mol Genet Genomics 278:493-505

Shewry PR and Halford NG (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot 53: 947–958

Shi C, Zhu J, Yang X, Yu Y, and Wu J (1999) Genetic analysis for protein content in indica rice. Euphytica 107: 135-140

Singh NB and Singh HG (1982) Gene action for quality components in rice. Indian J Agric Sci 52(8): 485-488

Temnykh S, DeClerk G, Lukashowa A, Lipovich L, Cartinhour S and McCouch S (2001) Computational and experimental characterization of microsatellites in rice (*Oryza sativa* L.): frequency, length, variation, transposon associations, and genetic marker potential. Genet Res 11: 1441-1452

Thorup GL and Kearsey FD (2000) The principles of QTL analysis (a minimal mathematics approach). J Exp Bot 49:1619-1623

Tuberosa R and Salvi S (2007) From QTLs to genes controlling root traits in maize: Scale and Complexity in Plant Systems Research. Gene-Plant-Crop Relations: 15-24

UNICEF, United Nations Childrens' Emergency Fund (2005) IDA, Prevention, assessment and control, in: Report of a joint WHO/UNICEF/UNU consultation, WHO

Varshney RK (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett 7: 537-546

Wang L, Zhong M, Li X, Yuan D, Xu Y, He Y, Luo L and Zhang Q (2008) The QTL controlling amino acid content in grains of rice (Oryza sativa) are co-localized with the regions involved in the amino acid metabolism pathway. Mol Breeding 21:127-137

Yoshida S, Ikgami M, Kuze J, Sawada K, Hashimoto Z, Ishii T, and Nakamurac Kamijima O (2002) QTL analysis for plant and grain character of sake brewing rice using a double haploid population. Breed Sci 52: 309-317