Plant Mol Biol (2007) 63:625–635
DOI 10.1007/s11103-006-9118-7

A large-scale collection of phenotypic data describing an insertional mutant population to facilitate functional analysis of rice genes

Akio Miyao · Yukimoto Iwasaki · Hidemi Kitano · Jun-Ichi Itoh · Masahiko Maekawa · Kazumasa Murata · Osamu Yatou · Yasuo Nagato · Hirohiko Hirochika

Received: 9 November 2006 / Accepted: 16 November 2006 / Published online: 19 December 2006

Abstract In order to facilitate the functional analysis of rice genes, we produced about 50,000 insertion lines with the endogenous retrotransposon *Tos17*. Phenotypes of these lines in the M_2 generation were observed in the field and characterized based on 53 phenotype descriptors. Nearly half of the lines showed more than one mutant phenotype. The most frequently observed phenotype was low fertility, followed by dwarfism. Phenotype data with photographs of each line are stored in the *Tos17* mutant panel web-based database with a dataset of sequences flanking *Tos17* insertion points in the rice genome (http://tos.nias.affrc.go.jp/). This combination of phenotypic and flanking sequence data will stimulate the functional analysis of rice genes.

Keywords *Oryza sativa* · Insertion mutagenesis · Phenotyping · Retrotransposon · Database

Introduction

Rice is the most important staple crop for half the world's population. Improvements in rice yield and quality beyond the benefits of the green revolution of 30 years ago are required to meet the demands of an increasing global population. At the beginning of the 21st century, with the hope of finding creative solutions to the problems of food production, nutrition and transportation, nearly the entire nucleotide sequence

A. Miyao (✉) · H. Hirochika
Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, 2-1-2, Kannondai, Tsukuba, Ibaraki 305-8602, Japan
e-mail: miyao@affrc.go.jp

H. Hirochika
e-mail: hirohiko@affrc.go.jp

Y. Iwasaki
Faculty of Bioscience, Fukui Prefectural University, 4-1-1, Kenjyojima, Matsuoka, Eiheiji-cho, Yoshida-gun, Fukui 910-1195, Japan
e-mail: iwasaki@fpu.ac.jp

H. Kitano
Bioscience and Biotechnology Center, Nagoya University, Furocho, Chikusa, Nagoya 464-8601, Japan
e-mail: hdkitano@nuagr1.agr.nagoya-u.ac.jp

J. Itoh · Y. Nagato
Graduate School of Agricultural and Life Sciences, University of Tokyo, Tokyo 113-8657, Japan

J. Itoh
e-mail: ajunito@mail.ecc.u-tokyo.ac.jp

Y. Nagato
e-mail: anagato@mail.ecc.u-tokyo.ac.jp

M. Maekawa
Research Institute for Bioresources, Okayama University, Kurashiki 710-0046, Japan
e-mail: mmaekawa@rib.okayama-u.ac.jp

K. Murata
Agricultural Experiment Station, Toyama Agricultural Research Center, 1124-1 Yoshioka, Toyama 939-8153, Japan
e-mail: kmurata@agri.pref.toyama.jp

O. Yatou
Rice Biotechnology Research Subteam, Hokuriku Research Center, National Agricultural Research Center (NARC), 1-2-1 Inada, Joetsu, Niigata 943-0193, Japan
e-mail: yatou@affrc.go.jp

of the rice genome, *Oryza sativa* L. cv. Nipponbare, was determined through a world-wide collaborative effort (International Rice Genome Sequencing Project 2005). The sequence data provide an important resource for promoting the discovery of important genes for crop improvement. Currently, a Rice Annotation Project (RAP) using nucleotide sequences of full-length cDNAs (The Rice Full-Length cDNA Consortium 2003) is in progress to position functional genes on the rice genome map (Ohyanagi et al. 2006); however, many genes remain ''unknown'' due to lack of experimental evidence or sufficient similarity with characterized genes from other organisms. The next great challenge after completion of genome sequencing is the functional characterization of these genes and discovery of genes that affect vital developmental, agronomic or biochemical plant functions (Hirochika et al. 2004).

Gene annotation is mostly based on the sequence similarity to known genes from other species. The limitations of this method are that every organism may have unique genes that do not have homologues even in closely related species, and assignment of a protein's function based on similarity may only give a partial description. For example, a gene may contain a domain that is conserved among protein kinases, but the actual substrate of the enzyme would be difficult or impossible to determine without experimental evidence. Categorization using cladistic associations (i.e. a phylogenetic tree) is more sensitive and is able to detect BLAST mis-hits or false positives (Sjolander 2004). However, characteristics supported by experimental data, e.g., correlation between the phenotype and the function of the protein kinase, are indispensable to the exact annotation, because all of the associative algorithms ultimately depend on experimental data.

Mutational analysis through gene disruption is one of the most efficient methods for identifying gene function. The related approaches of interrupting gene expression with RNAi (Miki and Shimamoto 2004), and overexpression are also effective because the functions of genes can be determined by the correlation of disrupted genes and their associated phenotypes. Currently, more than 130,000 T-DNA insertion lines of *Arabidopsis thaliana* have been created and are publicly available (Alonso et al. 2003). Phenotyping of *Ds* insertion lines in Arabidopsis is also in progress (Kuromori et al. 2006). For characterizing monocot genomes, we have produced more than 50,000 disruption lines of *Oryza sativa* cv. Nipponbare (*Japonica*), using the endogenous retrotransposon *Tos17*, which has a ''copy and paste'' type of transposition activity (Miyao et al. 2003). There are two native copies of *Tos17* in the Nipponbare genome that are activated specifically in cultured cells (Hirochika et al. 1996). On average, ten new copies of *Tos17* are transposed in each cell during 5 months in culture. When plants have been regenerated from the cultured cells, *Tos17* retrotransposition is immediately inactivated, and *Tos17* copies become fixed and segregate in a Mendelian fashion in the next generation.

There are many advantages to the *Tos17* disruption system for mutational analyses. Because the flanking regions of *Tos*17 insertion points are easily amplified by TAIL-PCR or a suppression PCR method using a 3′-end primer of *Tos17*, insertion sites in the rice genome can be easily determined. Furthermore, mutants can be screened by PCR using the *Tos*17 end primer and a primer from any desired genomic sequence. The distribution of transposed *Tos17s* is not random, and a large number of transpositional ''hot spots'' are detected throughout the rice genome. The insertion frequency of *Tos17* into genic regions is three-fold higher than that into non-coding regions. Owing to this polarization, *Tos17* insertion lines have great advantages for the functional analysis of rice genes (Miyao et al. 2003). Because *Tos17* is an endogenous retrotransposon, regenerated lines can be grown in the field, and seeds can be exported without regulations associated with genetically modified organisms such as rice lines that have undergone ''transgenic-'' insertion of T-DNA (Jeon et al. 2000; Sallaud et al. 2003; Wu et al. 2003), *Ds* (Greco et al. 2003; Kim et al. 2004), or *En/Spm* (Kumar et al. 2005).

Mutant lines generated by *Tos*17 retrotransposition have already been used for the functional analysis of rice genes. For example, *Tos17* insertion mutants with a dwarf or viviparous phenotype were used to identify and analyze genes for gibberellin and abscisic acid metabolism (Sakamoto et al. 2004; Agrawal et al. 2001). Phenotypic characteristics of most of the *Tos17* insertion lines, however, remain to be described. A large number of plant scientists working with rice or other monocotyledonous species could benefit from a systematic phenotypic analysis of many *Tos17* insertion lines and the creation of a public database. To promote the functional analysis of rice genes, the phenotypes of all of our mutant lines have been observed in rice fields through the collaboration of seven laboratories. Collected phenotypic data are useful for predicting the function of a disrupted gene. In this paper, we report the phenotypic statistics of a *Tos17* insertion mutant population for the discovery of agronomically important genes.

Materials and methods

Plant materials

Nipponbare calli derived from embryos were grown in N6 liquid medium (Otsuki 1990) containing 1 mg/l 2,4-D for 5 months. In total, about 50,000 plants were regenerated. Seeds of the M_2 generation were harvested from each M_1 plant. To check the activity of *Tos17* under various hormone conditions, N6 liquid media containing 1, 2, 5, 10 or 20 mg/l 2,4-D with or without 0.1 mg/l BA were used.

Phenotyping

Ten to twenty-five seeds were planted per line. Germination rates and seedling phenotypes were observed in the nursery. After 1 month from seeding, the seedlings were transferred to the paddy field. Phenotypes in the field were observed at the vegetative stage, near the heading stage, at the seed maturation stage, and at harvest. Lines segregating abnormal plants at about 25% frequency were digitally photographed and assigned a phenotype ID.

Database

All phenotype data were stored into a relational database on the PostgreSQL relational database managing system (http://www.postgresql.org/) with a FreeBSD 5.5 operating system (http://www.freebsd.org/). FreeBSD 6.1 was used for the statistical analysis. Tabular structure for phenotypic description consists of the line name, plant serial number, observation date, observing person, a link to the photograph file, phenotype ID, and detailed descriptions. Additional data for each line are inserted as a new row. Data can be modified only by the observing person. With this table structure, observation logs for each line by many persons over the course of many years can be stored without conflict. For counting lines of each mutant phenotype, Perl script, which can connect with the PostgreSQL server with a Pg.pm module, reads the phenotype table, stores line names and phenotype IDs into an associative array (''hash''), to convert from a redundant number of phenotype IDs to a unique number for each line. The number of lines showing each phenotype ID in the hash was counted. Flanking sequence data of *Tos17* insertion sites are also stored in the same database. Loci of *Tos17* insertions are determined by BLASTN searches against rice genome sequences of the International Rice Genome Sequencing Project (IRGSP) Build3. Loci of annotated genes are also stored in the relational database.

SOM analysis

Software for SOM analysis was downloaded from the Neural Networks Research Center of Helsinki University of Technology (http://www.cis.hut.fi/research/som_pak/). For the initialization program, randinit, parameters: 24 for x dimension, 16 for y dimension, hexa for topology, bubble for neighborhood function, and 123 for seeds were used. For the map training program, vsom, parameters for first learning: 1000 for learning length in training, 0.05 for initial learning rate, 10 for initial radius of the training area in som-algorithm, parameters for second learning: 10000 for learning length, 0.02 for initial learning rate, 3 for initial radius. The analyzed data were visualized using the umat program.

Results

Flow of phenotypic analysis

Nipponbare calli from 92 seeds were independently cultured for 5 months, and about 50,000 plants were regenerated. Seeds (M_2 generation) from the regenerated plants were independently harvested and labeled with a line name. Ten to twenty-five M_2 plants of each line were observed over a full developmental cycle in the nursery and paddy.

Each line was designated by ''N'' for ''Nipponbare'' followed by a letter A-G that indicates the yearly lot, and four figures, e.g. NA1234. Lines named with ''NA'' and ''NB'' were used in a small-scale pilot study and ''NC'' to ''NG'' were used in large-scale studies.

To evaluate the effect of hormone (auxin and cytokinin) concentration in the medium on *Tos*17 retrotransposition activity, the NC line was subcultured in media containing 1, 2, 5, 10, or 20 mg/l of 2,4-dichlorophenoxyacetic acid (2,4-D) with or without 0.1 mg/l benzyl adenine (BA). There was no significant difference in observed transposition events among lines derived from the subcultures. In lots, ND, NE, NF, and NG, calli were induced in a medium containing 2 mg/l 2,4-D, and maintained in a medium supplemented with 1 mg/l 2,4-D for 5 months. The cultural conditions for each line are available with the phenotypic description list from the mutant panel database.

Co-segregation of mutant phenotypes with *Tos17* insertion can be detected by DNA blot hybridization,

which will help the further analysis of gene function. If co-segregation is detected, the flanking region of co-segregated *Tos17* can be isolated by TAIL-PCR or suppression PCR (Miyao et al. 1998).

Classification of phenotypes

To provide a classification system and a database useful for most rice scientists, phenotype scoring was limited to 53 phenotype descriptors belonging to 12 classes. Because environmental conditions such as day length, temperature and soil conditions differ significantly among the seven fields in which this project was conducted, some variability in traits such as heading date or leaf color may have occurred between different fields. Thus, some of the present data may vary with environmental conditions, although most of mutant phenotypes are stably expressed. Phenotype classifications and a summary of observations are shown in Table 1. Each subclass has a phenotype ID code to enable data entry as a barcode with a portable recording device and to enable data compilation from all seven laboratories.

1. Germination. This trait was evaluated by measuring germination rate under defined conditions. Since wild type (cv. Nipponbare) showed a germination rate higher than 95%, lines that showed germination rates less than 75% were recorded in the present project. Since several laboratories measured germination rates of all lines, all primary data are also stored in the database. In total, 3489 lines showed a low germination rate, and 525 lines showed germination rates less than 50%. Some of the lines with low germination rates may be embryo mutants.
2. Growth. Growth was observed at the seedling stage in the rice nursery or at an early stage after transfer to the field. ''Weak'' refers to mutants that formed slim seedlings with retarded growth, probably caused by a deficiency in some housekeeping gene product (Fig. 1, NG0352). Some of the ''Weak'' lines were reclassified eventually to ''Lethal''. An example of ''Abnormal shoot'' is shown in Fig. 1 (NG0356). Another example is represented by NE3024. This seedling is smaller than wild type and its leaves are short and wide, resembling mutants defective in genes associated with gibberellin biosynthesis or signaling pathways (Sakamoto et al. 2004; Uegchi-Tanaka et al. 2005).
3. Leaf color. Frequently appearing pigmentation phenotypes are ''Albino'' (Fig. 1, NG1048) and ''Virescent'' (Fig. 1, NE1517). Completely white or yellow (Fig. 1, NG1469) seedlings died within 3 weeks after seeding. If green and white segments coexisted on leaves, a condition called virescent, the seedlings survived in the field. Zebra mutants that show repetition of a white or a pale green band and a green band in the longitudinal direction were also observed (Fig. 1, NF6044). However, in most lines, the ''Zebra'' phenotype was limited to young stages. The ''Stripe'' phenotype (Fig. 1, NE4001) often showed extremely biased segregation, e.g., only one plant in 25, and often was not stably inherited.
4. Leaf shape. Many different kinds of abnormally shaped leaves were observed. ''Short Leaf'' and some ''Wide leaf'' phenotypes were eventually reclassified as ''Dwarf'' or ''Severely dwarf'' phenotypes (Fig. 1, NE8114). The ''Short leaf'' phenotype is similar to mutants defective in gibberellin biosynthesis or signal transduction. The line NE5022 could not develop a normal, flat leaf blade and eventually died (Fig. 1). Several lines showed pleiotropic phenotypes of the shoot. The line NE8329 did not develop tillers, showed dwarfism and formed rolled or twisted leaves (Fig. 1). Since these phenotypes are difficult to characterize exactly by only phenotype codes, additional remarks were presented in the comment column of the database. Frequency distribution of the leaf width mutants was biased toward the narrow type (Fig. 1, NG0754), rather than the wide type.
5. Culm shape. Dwarfism is the most abundant mutant phenotype, along with sterility. The ''Semi-dwarf'' condition is characterized by plant heights that are 70–80% of wild type values. ''Dwarf'' describes plants with heights smaller than ''Semi-dwarf'' but larger than ''Severely dwarf''. Plants are classified ''Severely dwarf'' when the plant height is smaller than 30 cm at maturity. Some dwarf phenotypes often co-segregated with other abnormal phenotypes such as ''Fine leaf'', ''Wide leaf'', ''Spiral leaf'', or ''Abnormal panicle shape''. The dwarf mutants accompanying other shoot/panicle abnormalities are expected to be involved in hormonal signaling or synthesis pathways. Other culm phenotypes such as ''Thick culm'' (Fig. 1, NG9874) appeared infrequently.
6. Spotted leaf/lesion mimic. Various types of ''Spotted leaf/lesion mimic'' phenotypes, e.g., small and scattered spots, large and dispersed spots, were observed. Heavy lesions caused the early death of leaves (Fig. 1, NG0752). Mutants with large white lesions were also observed.

Table 1 Summary of phenotype data

	Class	Phenotype	ID code	NC	ND	NE	NF	NG	Total
1	Germination	Low germination rate	1	327	531	1326	1005	300	3489
2	Growth	Lethal	2	135	300	689	271	234	1629
		Abnormal shoot	3	189	172	219	823	384	1787
		Weak	4	114	204	341	556	390	1605
3	Leaf color	Albino	11	264	384	230	275	254	1407
		Yellow	12	96	139	186	317	71	809
		Dark green	13	295	285	354	58	69	1061
		Pale green	14	270	429	372	265	395	1731
		Virescent	15	35	278	206	131	184	834
		Stripe	16	46	40	64	132	102	384
		Zebra	17	12	15	13	39	29	108
		Others	18	9	15	23	22	16	85
4	Leaf shape	Wide leaf	21	38	28	51	4	21	142
		Narrow leaf	22	165	324	432	204	251	1376
		Long leaf	23	3	3	14	4	5	29
		Short leaf	24	2	4	25	8	1	40
		Drooping leaf	25	8	140	35	12	31	226
		Rolled leaf	26	31	52	153	23	83	342
		Spiral leaf	27	10	23	55	3	18	109
		Brittle leaf/culm	28	5	14	40	53	10	122
		Abnormal lamina joint angle	29	14	27	13	14	28	96
		Withering	30	112	270	251	88	147	868
		Others	31	37	54	46	113	93	343
5	Culm shape	Semi-dwarf	41	645	822	661	633	900	3661
		Dwarf	42	803	1550	1351	882	1123	5709
		Severely dwarf	43	248	411	355	143	217	1374
		Long culm	44	258	217	50	68	83	676
		Fine culm	45			1	16	1	18
		Thick culm	46	31	24	2	1	4	62
		Others	47		1	1	4	2	8
6	Spotted leaf/lesion mimic	Spotted leaf/lesion mimic	51	115	211	197	269	302	1094
7	Tillering	High tillering	61	11	27	25	21	27	111
		Low tillering	62	418	678	696	609	421	2822
		Lazy	63	70	115	53	80	56	374
8	Heading date	Early heading	65	352	976	99	297	65	1789
		Late heading	66	111	561	247	215	110	1244
		Non-heading?	67	2	2	42	32	18	96
9	Spikelet	Abnormal hull	71	41	268	36	52	111	508
		Abnormal floral organ	72	12	128	43	11	17	211
10	Panicle	Long panicle	81	1	3	19	3	3	29
		Short panicle	82	51	79	426	92	104	752
		Lax panicle	83	18	16	15	23	32	104
		Dense panicle	84	41	18	97	65	43	264
		Viviparous	85	85	473	267	142	138	1105
		Shattering	86	1	3	2			6
		Neck leaf	87	38	51	48	10	24	171
		Abnormal panicle shape	88	21	43	265	79	25	433
11	Sterility	Sterile	91	441	1026	818	658	882	3825
		Low fertility	92	2877	3127	935	2991	2612	12542
12	Seed	Large grain	101	55	78	3	34	25	195
		Small grain	102	47	95	96	145	44	427
		Slender grain	103	20	28	18	21	14	101
		Others	104	979	1117	883	143	347	3469

Lots NC through NG were harvested in 1997 through 2001, respectively. Numbers of mutant lines with the corresponding phenotypes are listed

Brown spots were more frequently observed than white ones.

7. Tillering. In the tillering class, ''Low tillering'' mutants were abundant. The ''High tillering'' phenotype tends to co-segregate with the ''Fine leaf'' phenotype. The ''Lazy'' (Fig. 1, NG0667) mutants have open and recumbent tillers. A quite unusual phenotype among ''High tillering'' lines

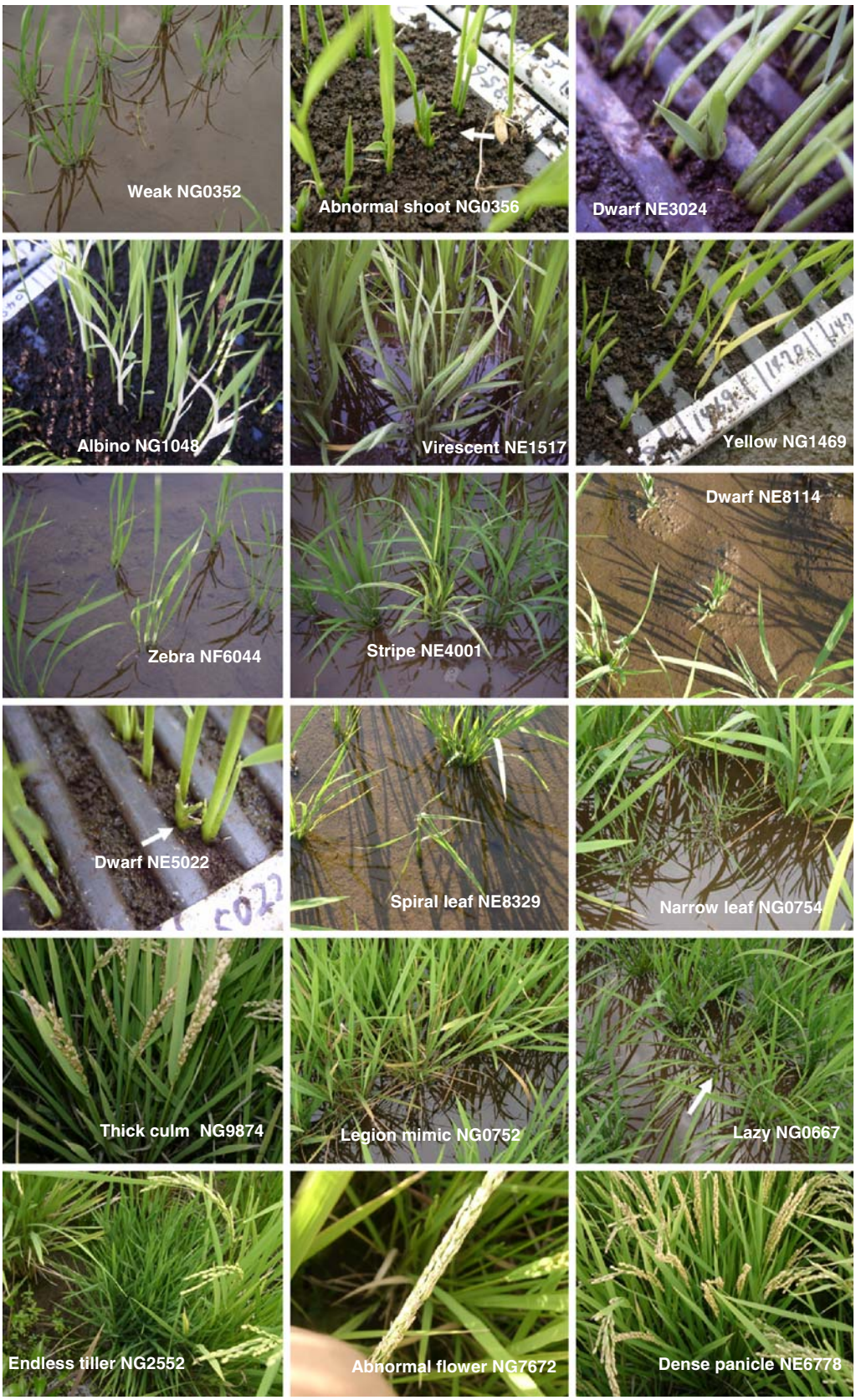

Fig. 1 Representative phenotypes of *Tos*17 insertional mutants. Phenotype and line designation are indicated on each photograph

is NG2552 that produced tillers reiteratively from upper internodes and failed to produce panicles.

8. Heading date. In this category, lines whose heading date deviated more than 7 days from normal were classified as heading date mutants. This category includes ''Early heading'', ''Late heading'', and ''Non-heading?'' mutants. A small number of lines segregated as ''Non-heading?'' mutants that did not form panicles even at harvest (five to 6 months after sowing). The number of ''Early heading'' lines (1797) is a little larger than that of ''Late heading'' lines (1244).
9. Spikelet. Each spikelet of rice is composed of two rudimentary glumes, two empty glumes and one floret comprised of one lemma, one palea and three kinds of floral organs (two lodicules, six stamens and one pistil). ''Abnormal hull'' refers to any glume abnormality and ''Abnormal floral organ'' refers to any abnormality in floral organs (e.g. an abnormal pistil). In the ''Abnormal hull'' class, there is a mutant that failed to close hulls after flower opening. On the other hand, mutants that could not open flowers due to underdeveloped stamens were often re-classified as ''Completely sterile'' mutants. Another mutant produced an extra glume. Abnormal flower phenotypes showing shoots growing from floral organs as in Fig. 1 (NC7672), were also detected at relatively low frequency.
10. Panicle. The most frequently observed abnormal panicle phenotype was the precocious germination of seeds while still attached to the maturing panicle, or ''Viviparous''. As for panicle shape, ''Dense panicle'' and ''Short panicle'' mutants (Fig. 1, NE6778) were also frequently observed, and ''Long panicle'' and ''Lax panicle'' were somewhat rare. Incomplete emergence of the panicle from the flag leaf sheath was categorized as ''Neck leaf''.
11. Sterility. ''Sterile'' is the third most frequent mutant in this *Tos17* insertion mutant population after low seed fertility and dwarf mutants. The ''Sterile'' and ''Low fertility'' conditions correspond to lines with less than 2% and ca. 50% seed fertility, respectively.
12. Seed. Only ''Large grain'', ''Small grain'', and ''Slender grain'' were distinguished in this category. The frequency of ''Small grain'' was two and four times higher than the frequency of ''Large grain'' and ''Slender grain'', respectively. Among the ''Others'', white or dull kernel phenotypes were often observed.

Correlation between phenotypes

To understand how the 53 phenotypic descriptors are related, lines showing two or more abnormal phenotypes were selected, and a matrix with 53 rows and 53 columns of phenotypes with a number of lines showing each pair of phenotypes was devised. Values for the number of respective phenotypes were changed to ''x'' in the matrix, (''x'' is ignored by the SOM program), and data were subjected to self-organizing map (SOM) analysis (Kohonen 1995). The SOM algorithm is used for visualization of multidimensional complex data using an unsupervised learning method based on a grid of artificial neurons. The 53-dimensional correlation data of phenotypes was reflected in a two-dimensional map (Fig. 2).

Topologies (not distances) of phenotypes on the SOM coordinates coincide with the correlation between phenotypes. Distance between two phenotypes is indicated by grayscale color on the SOM. For example, assume that phenotype A, B, and C appeared in 5000, 50, and 10 lines, respectively. Seven lines showed phenotype A and C. Three-lines showed phenotype B and C. The color between phenotype A and C should be darker than the color between phenotype B and C, because the line number of phenotype A lines is much greater than that of phenotype B lines, although the phenotype correlation of C with A is stronger than that with B. It is difficult to explain multidimensional data completely on a two dimensional map, but a SOM map indicates generally that neighboring phenotypes have relatively tight correlations. The SOM analysis in Fig. 2 shows that there are many pairs of phenotypes that are apparently correlated. That is, among 53 phenotypes, several phenotypes have a high probability to emerge simultaneously with other specific phenotypes. This phenotypic correlation could be caused by pleitropic expression of a single gene, or could reflect developmental causality of the two abnormalities. For example, ''Long leaf'' and ''Long panicle'' (lg_lf, lg_pa) located in the upper left corner on the SOM map in Fig. 2 have a very strong correlation. This result suggests that the gene involved in leaf length also affects panicle length. On the contrary, long leaf and dwarfism, at the opposite corner of the map, do not have any correlation. Both dwarf and semi-dwarf phenotypes have a relatively strong correlation with sterile and low fertile phenotypes. Since this correlation is not so strong, a part of ''Dwarf'' and ''Semi-dwarf'' mutants do not affect seed fertility, but other ''Dwarf'' and ''Semi-dwarf'' mutants may be regu-

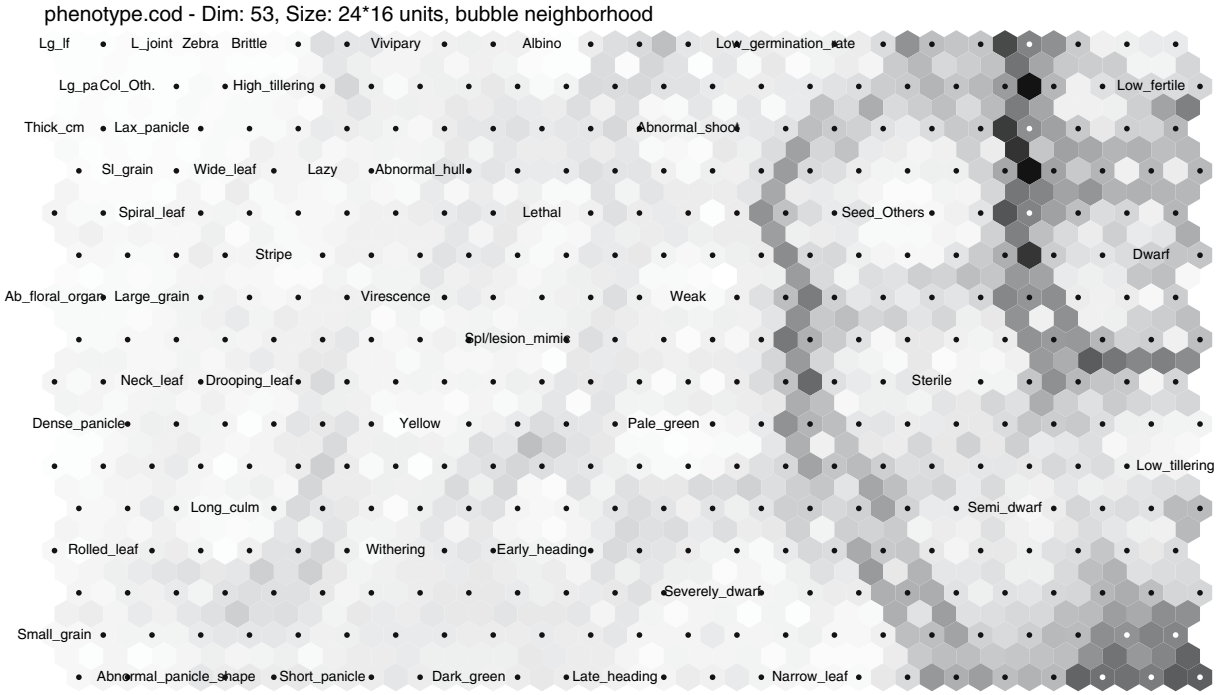

Fig. 2 Self-organizing map for correlation of phenotypes. Lines that showed two or more abnormal phenotypes were analyzed by SOM. Distances between phenotypes indicate the magnitude of correlation of phenotypes. Col_Oth., Color Others; Lg_lf, Long leaf; L_joint, Abnormal lamina joint angle; Lg_pa, Long panicle; Ab_floral_organ, Abnormal floral organ; Sl_grain, Slender grain; Thick_cm, Thick culm. Grayscale levels of each node represents the distance between references

lated by genes associated with some housekeeping functions.

Correlations between mutant phenotype and *Tos17* insertion

If allelic insertion lines show the same phenotype, it is highly probable that the mutant phenotype is caused by disruption of the target gene by a *Tos17* insertion. We searched for genes that were disrupted by *Tos17* in at least two lines. A total of 391 genes were detected based on the Rice Annotation Project (Ohyanagi et al. 2006). Table 2 is a partial list of such loci and their mutant phenotypes. For example, five lines have insertions of *Tos17* in the *Os06g027500* gene (Fig. 3A), of which two lines have the *Tos17* insertions in exons and show an early heading phenotype. On the other hand, the other three lines have *Tos17* insertions in introns and are not early heading. (see http://tos.nias.affrc.go.jp/). The *Os06g027500* gene is *Hd1*, a key gene for determining heading date (Yano et al. 2000). Another example is the magnesium chelatase subunit gene, *Os07g0656500* in Fig. 3B (Jung et al. 2003). *Tos17* insertion into this gene was correlated with an albino phenotype, because mutants defective in this gene lack active chlorophyll and eventually become albino. All correlations between phenotypes and disrupted genes by *Tos17* insertion are displayed on the mutant panel database website (http://tos.nias.affrc.go.jp/).

A large-scale phenotypic characterization of an M_2 population generated by *Tos17* retrotransposition has revealed that this population contains a large number of mutants covering many easily scored phenotypes. In addition, linkage between the mutant phenotype and a specific *Tos17* insertion facilitates greatly the isolation of the causal genes and the elucidation of the gene functions.

Discussion

Three-years of collaboration among seven laboratories has led to a large-scale phenotypic characterization of about 50,000 M_2 plants generated by *Tos17* retrotransposition. We have examined 53 kinds of abnormal phenotypes in rice from the seedling to harvest stages that are easily evaluated in the field. This project has revealed that this population contains a large number of mutants covering a wide range of phenotypes. Although several mutant phenotypes may be environment-sensitive and their expression may be unstable, most of the mutant phenotypes were stable. We also

Table 2 List of loci that have *Tos17* insertions in exons in at least two lines

Locus name	Description	Phenotypes
Os01g0113200	LRK14	Pale green leaf, low fertility
Os01g0113300	Receptor-like kinase ARK1AS	Dwarf, spotted leaf/lesion mimic
Os01g0147800	Protein of unknown function DUF547 domain containing protein	Pale green leaf, semi-dwarf, long culm, short panicle, sterile, low fertile
Os01g0685900	65 kD Microtubule associated protein	Narrow leaf, semi-dwarf
Os02g0552600	8-Oxoguanine DNA glycosylase	Lethal
Os04g0464200	Betaine-aldehyde dehydrogenase (EC 1.2.1.8) (BADH)	Low fertility
Os04g0680400	Allantoinase (EC 3.5.2.5)	Dwarf, low fertility
Os05g0318600	Protein kinase domain containing protein	Narrow leaf
Os05g0548900	Phosphoethanolamine methyltransferase	Early heading
Os05g0552400	Zn-finger, RING domain containing protein	Early heading
Os06g0176800	2OG-Fe(II) oxygenase domain containing protein	Dark green leaf, dwarf, severely dwarf
Os06g0275000	Hd1	Dwarf, early heading
Os06g0680500	Glutamate receptor 3.1 precursor (Ligand-gated ion channel 3.1) (AtGLR2)	Low fertility
Os07g0197100	Hexokinase	Dwarf, sterile
Os07g0646500	SWIM Zn-finger domain containing protein	Late heading
Os07g0656500	Protoporphyrin IX Mg-chelatase subunit precursor	Lethal, albino, dwarf
Os09g0278300	Phosphatidylinositol-4-phosphate 5-kinase family protein	Semi-dwarf, dwarf, low tillering, sterile, low fertility
Os10g0567100	Chlorophyll b synthase (Fragment)	Dark green leaf, pale green leaf, withering, semi-dwarf, severely dwarf, early heading, late heading, low fertility
Os12g0127600	WRKY transcription factor 57	Sterile, low fertility
Os12g0566000	HCO3-Transporter domain containing protein	Severely dwarf
Os12g0572500	Protein of unknown function XH domain containing protein	Early heading

Locus name and description are from RAP data

deposited additional information about certain phenotypes in a comment field within the phenotype description table, if this information was provided. A text search box on the phenotype list web page is envisioned to facilitate use of the resource.

Because M_2 plants were observed, inserted *Tos17s* segregate among individuals. An average of 10 new copies of *Tos17* are inserted into each regenerated plant. If the phenotype co-segregates with an insertion of *Tos17*, the phenotype is likely to be caused by disruption of the target gene. Of course, there are lines whose mutant phenotypes are not correlated with *Tos17* insertion. These mutants might be caused by insertions of other native transposons, by chromosomal aberrations, or by other mutations during tissue culture. In this case, their causal genes could be isolated by usual positional cloning methods.

There were some difficulties in clustering phenotypes using standard relational algorithms, because each of the 53 phenotypes has as many as 52 kinds of frequencies against other phenotypes, respectively. SOM analysis is a kind of clustering method suitable for such multi-dimensional non-linear data (Kohonen 1995). SOM analysis confirmed some associations due to hormonal or developmental constraint, but some additional intriguing correlations were also detected, e.g., between leaf color/shape and heading date, between dwarfism and sterility. In Fig. 2, the ''Long leaf'' phenotype is strongly correlated with the ''Long panicle'' phenotype. The ''Weak'' phenotype at the seedling stage has a relatively strong correlation with ''Lethal'', ''Abnormal shoot'', ''Pale green'' Leaf and ''Spotted leaf/lesion mimic'' phenotypes, indicating that the ''Weak'' plants at the seedling stage likely show ''Lethal'', ''Abnormal shoot'', ''Pale green'' and/or ''Spotted leaf/lesion mimic'' phenotypes at later stages. ''Dwarf'', ''Semi-dwarf'', ''Sterile'', ''Low fertile'' and ''Low tillering'' phenotypes showed relatively high correlations. The SOM map is quite useful for the overview of phenotype correlations. This method of analysis will be useful not only for considering the major and side-effects of gene disruption but also for re-categorization of phenotypes. The advantage of SOM analysis is the ability to analyze many different kinds of data simultaneously. Currently, metabolome data have been obtained from Arabidopsis (Hirai et al. 2004), tomato (Schauer et al. 2006) and other plants. It is reasonable to predict that

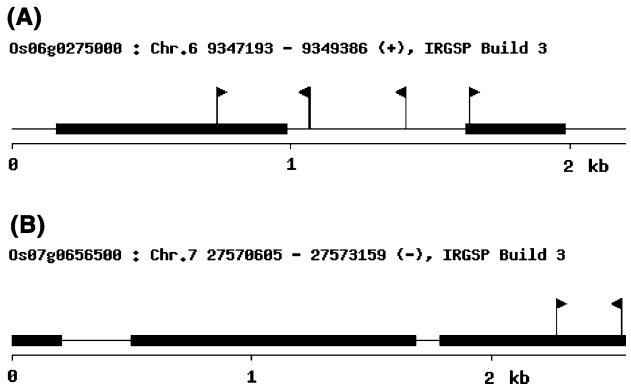

Fig. 3 Examples of genes that have insertions of *Tos17*. (**A**) Os06g0275000, *Hd1* gene. (**B**) Os07g0656500, Magnesium chelatase subunit gene. The structure and insertion points of these genes are drawn on demand from our web site (http://tos.nias.affrc.go.jp/). Coding regions are indicated with broad lines. Positions of *Tos17* insertions are indicated by flags. The direction of the flag is the direction of the *Tos17* insertion. Detailed descriptions of insertion points, sequence names, accession numbers, and line names will appear on the web page. When a sequence name is clicked, the figure will be re-drawn with a flag labeling the sequence name to distinguish closely inserted *Tos17s*

the combinatorial analysis of phenotype and metabolome data of rice will reveal correlations between metabolic pathways and phenotypes without additional genetic information.

Phenotype databases of rice have been developed in many countries, e.g., Oryzabase (Kurata and Yamazaki 2006), RMD (Zhang et al. 2006), and IRRS (Wu et al. 2005). Our system includes both phenotypic and insertion data on the rice genome in a relational database. The number of photographs on file is more than 58,000, and the number of phenotype description records is more than 158,000. Our database structure enables direct access of Perl script to the database and the extraction of many kinds of correlations. Flanking sequences of insertion points and phenotypes of the insertion lines provide a provisional assignment of the function of a disrupted gene. We have already sequenced more than 25,000 flanking sequences from 20% of the insertion lines, and the number of flanking sequences is still increasing. Of the 27,448 total annotated loci based on RAP1/IRGSP Build3, 391 loci have more than two insertions in exons. When two or more lines have *Tos17* insertions in a common gene, they usually exhibited similar phenotypes. These data are a strong indication of the function of a disrupted gene. However, several phenotypes such as ''Dwarf'' and ''Low fertile'' were observed in many lines. It might be difficult to assign a gene function based solely on correlation for these frequently observed phenotypes, and a complementation test would be required to confirm the correlation. If flanking sequence data from all insertion lines can be obtained, this correlation will be more useful for the annotation of genes. We are continuing to sequence the flanking regions of all our mutant lines.

In this study, we collected a large amount of phenotypic data in seven fields under natural conditions. If phenotypes are observed under other conditions such as drought or temperature stress conditions or under pathogen pressure, phenotypic description of this population would be much more enriched. In addition, the present study evaluated only a limited number of traits easily scored on above-ground organs. Thus, other traits such as roots and seed storage composition remain to be investigated in the future. Expansion of this study to include new traits would enable investigators to find new correlations with disrupted genes. Furthermore, integration of phenotype data with those of microarray experiments, metabolic profiling, and other approaches will be a powerful tool for revealing new aspects of plant physiology.

All phenotype and flanking sequence data can be obtained via http://tos.nias.affrc.go.jp/. At this site, all annotated rice genes and locations of *Tos17* insertions are shown on a clickable chromosome map. Details containing illustrated gene structures with insertion points, phenotypes of corresponding lines, nucleotide sequences flanking the disrupted region, candidate primer sequences for segregation analysis, and annotation from RAP data corresponding to the selected position are displayed. Mutant seeds are available for scientific use from the Genome Resource Center at NIAS (http://www.rgrc.dna.affrc.go.jp/).

Acknowledgements This work was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Project MP-2101, MP-2103, MP-2198, MP-2129, MP-2130, MP-2131, MP-Seeds8).

References

Agrawal GK, Yamazaki M, Kobayashi M, Hirochika R, Miyao A, Hirochika H (2001) Screening of the rice viviparous mutants generated by endogenous retrotransposon *Tos17* insertion. Tagging of a zeaxanthin epoxidase gene and a novel OsTATC gene. Plant Physiol 125:1248–1257

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, Gadrinab C, Heller C, Jeske A, Koesema E, Meyers CC, Parker H, Prednis L, Ansari Y, Choy N, Deen H, Geralt M, Hazari N, Hom E, Karnes M, Mulholland C, Ndubaku R, Schmidt I, Guzman P, Aguilar-Henonin L, Schmid M, Weigel D,

Carter DE, Marchand T, Risseeuw E, Brogden D, Zeko A, Crosby WL, Berry CC, Ecker JR (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science 301:653–657

Greco R, Ouwerkerk PB, De Kam RJ, Sallaud C, Favalli C, Colombo L, Guiderdoni E, Meijer AH, Hoge Dagger JH, Pereira A (2003) Transpositional behaviour of an *Ac/Ds* system for reverse genetics in rice. Theor Appl Genet 108:10–24

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 101:10205–10210

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci USA 93:7783–7788

Hirochika H, Guiderdoni E, An G, Hsing YI, Eun MY, Han CD, Upadhyaya N, Ramachandran S, Zhang Q, Pereira A, Sundaresan V, Leung H (2004) Rice mutant resources for gene discovery. Plant Mol Biol 54:325–334

Jeon JS, Lee S, Jung KH, Jun SH, Jeong DH, Lee J, Kim C, Jang S, Yang K, Nam J, An K, Han MJ, Sung RJ, Choi HS, Yu JH, Choi JH, Cho SY, Cha SS, Kim SI, An G (2000) T-DNA insertional mutagenesis for functional genomics in rice. Plant J 22:561–570

Jung KH, Hur J, Ryu CH, Choi Y, Chung YY, Miyao A, Hirochika H, An G (2003) Characterization of a rice chlorophyll-deficient mutant using the T-DNA gene-trap system. Plant Cell Physiol 44:463–472

Kim CM, Piao HL, Park SJ, Chon NS, Je BI, Sun B, Park SH, Park JY, Lee EJ, Kim MJ, Chung WS, Lee KH, Lee YS, Lee JJ, Won YJ, Yi G, Nam MH, Cha YS, Yun DW, Eun MY, Han CD (2004) Rapid, large-scale generation of *Ds* transposant lines and analysis of the *Ds* insertion sites in rice. Plant J 39:252–263

Kohonen T (1995) The self-organizing map. Springer-Verlag, Heidelberg

Kumar CS, Wing RA, Sundaresan V (2005) Efficient insertional mutagenesis in rice using the maize *En/Spm* elements. Plant J 44:879–892

Kurata N, Yamazaki Y (2006) Oryzabase. An integrated biological and genome information database for rice. Plant Physiol 140:12–17

Kuromori T, Wada T, Kamiya A, Yuguchi M, Yokouchi T, Imura Y, Takabe H, Sakurai T, Akiyama K, Hirayama T, Okada K, Shinozaki K (2006) A trial of phenome analysis using 4000 *Ds*-insertional mutants in gene-coding regions of Arabidopsis. Plant J 47:640–651

Miki D, Shimamoto K (2004) Simple RNAi vectors for stable and transient suppression of gene function in rice. Plant Cell Physiol 45:490–495

Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, Shinozuka Y, Onosato K, Hirochika H (2003) Target site specificity of the *Tos17* retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. Plant Cell 15:1771–1780

Miyao A, Yamazaki M, Hirochika H (1998) Systematic screening of mutants of rice by sequencing retrotransposon-insertion sites. Plant Biotechnol 15:253–256

Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, Fujii Y, Antonio BA, Nagamura Y, Imanishi T, Ikeo K, Itoh T, Gojobori T, Sasaki T (2006) The Rice Annotation Project Database (RAP-DB): hub for *Oryza sativa* ssp. japonica genome information. Nucleic Acids Res 34:D741–D744

Otsuki Y (1990) A visual manual for the protoplast culture system of rice. Food and Agriculture Research Development Association, Tokyo

Sakamoto T, Miura K, Itoh H, Tatsumi T, Ueguchi-Tanaka M, Ishiyama K, Kobayashi M, Agrawal GK, Takeda S, Abe K, Miyao A, Hirochika H, Kitano H, Ashikari M, Matsuoka M (2004) An overview of gibberellin metabolism enzyme genes and their related mutants in rice. Plant Physiol 134:1642–1653

Sallaud C, Meynard D, van Boxtel J, Gay C, Bes M, Brizard JP, Larmande P, Ortega D, Raynal M, Portefaix M, Ouwerkerk PB, Rueb S, Delseny M, Guiderdoni E (2003) Highly efficient production and characterization of T-DNA plants for rice *(Oryza sativa* L.) functional genomics. Theor Appl Genet 106:1396–1408

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, Willmitzer L, Zamir D, Fernie AR (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 24:447–454

Sjolander K (2004) Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20:170–179

The Rice Full-Length cDNA Consortium (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. Science 301:376–379

Ueguchi-Tanaka M, Ashikari M, Nakajima M, Itoh H, Katoh E, Kobayashi M, Chow TY, Hsing YI, Kitano H, Yamaguchi I, Matsuoka M (2005) *GIBBERELLIN INSENSITIVE DWARF1* encodes a soluble receptor for gibberellin. Nature 29:693–698

Wu C, Li X, Yuan W, Chen G, Kilian A, Li J, Xu C, Li X, Zhou DX, Wang S, Zhang Q (2003) Development of enhancer trap lines for functional analysis of the rice genome. Plant J 35:418–427

Wu JL, Wu C, Lei C, Baraoidan M, Bordeos A, Madamba MR, Ramos-Pamplona M, Mauleon R, Portugal A, Ulat VJ, Bruskiewich R, Wang G, Leach J, Khush G, Leung H (2005) Chemical- and irradiation-induced mutants of indica rice IR64 for forward and reverse genetics. Plant Mol Biol 59:85–97

Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, Fuse T, Baba T, Yamamoto K, Umehara Y, Nagamura Y, Sasaki T (2000) *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene *CONSTANS*. Plant Cell 12:2473–2484

Zhang J, Li C, Wu C, Xiong L, Chen G, Zhang Q, Wang S (2006) RMD: a rice mutant database for functional analysis of the rice genome. Nucleic Acids Res 34:745–748