This document has been adapted from the 1986 Report of the Committee on Gene Symbolization, Nomenclature and Linkage Groups and includes suggestions for updates and modifications.

[THIS DOCUMENT IS OPEN FOR COMMENTS AND SUGGESTIONS. WE HAVE INCORPORATED COMMENTS FROM THE LAST VERSION AND REVISED THIS DOCUMENT ACCORDINGLY. AREAS HIGHLIGHTED IN YELLOW INDICATE SUGGESTIONS ON POLICY THAT WE WOULD LIKE TO HAVE CONSIDERED BY THE COMMITTEE ON GENE SYMBOLLIZATION, NOMENCLATURE AND LINKAGE GROUPS (CGSNL). THE EDITORS OF THE JOURNAL, GENOME RESEARCH, HAVE REQUESTED A SUMMARY OF GENE NOMENCLATURE RULES FOR RICE, INCLUDING NAMING OF SEQUENCED GENES. WE HOPE THIS DOCUMENT WILL SERVE AS A STARTING POINT FOR SUCH A PUBLICATION.]

Introduction

The biological community is moving toward a universal system for the naming of genes. The advantage to scientific communication in recognizing a common genetic language is to facilitate structural, functional and evolutionary comparisons of genes and genetic variation among organisms. With increasing emphasis on the molecular and biochemical nature of genes and gene products, it is important that the gene nomenclature system for rice reflect knowledge about both the phenotypic consequences of a particular allele in a given genetic background and the biochemical features of a specific gene or gene family.

The current rules for gene symbols in rice are based on recommendations from the **Committee on Gene Symbolization, Nomenclature and Linkage (CGSNL)** of the Rice Genetics Cooperative (Kinoshita, 1986). Most of the original gene names and symbols are descriptive of visible phenotypes that provided the earliest evidence for the existence of a gene and these names and symbols are widely used by the rice research community. As new methods for detecting and describing genes are increasingly applied to rice, nomenclature rules are needed that outline the standard procedures for describing genes based on DNA, RNA and protein sequence analysis and assays for biochemical function in addition to those previously outlined for phenotypic variants.

The focus of this paper is to summarize the existing rules for gene symbols in rice and to propose a few new rules that will bring the gene nomenclature for rice into agreement with the emerging gene nomenclature system being developed for other model organisms.

- Precedence of publication will be the primary determinant of a gene name and gene symbol (gene identifiers) unless an alternative name is proposed and accepted by the CGSNL.
- If genetic analysis (i.e., allelism tests) and/or sequence identity confirms that more than one name or symbol has been associated with the same gene, the case will be referred to the Committee.
- In cases where multiple names have been used over time so that they are widely recognized, these names can be acknowledged by the assignment of synonyms, with priority being given to the first published record of the gene name.
- To ensure that historical names are not lost, links will be maintained between existing (historical) gene names that are based largely on mutant phenotype and new names that are likely to be based largely on sequence annotation and biochemical assays.

In this way, the entire repertoire of rice gene names and gene symbols, along with all associated information will be readily available to the research community. At the same time, it will move our nomenclature system into greater harmony with that of other plant species.

Table of contents on proposition of rules for gene symbolization in rice.

I	**Genes**		3
	A	Gene full name	3
	B	Gene symbol	3
	C	Allelic variants	4
	D	Protein name and symbol	4
	E	Post translational modification	5
II	**Systematic locus id**		5
	A.	Nuclear genes	5
	B.	Organelle genes	5
	C.	Splice site variants	6
	D.	Genes in incomplete or unanchored sequenced clones	6
III	**Making associations between systematic locus ID, gene names and protein products**		6
	A.	Prediction and validation of Gene models	6
	B.	Gene names based on homology or orthology	6
	C	Discrepancies and use of species name in gene symbol	7
	D.	Gene models with protein matches	7
	E.	Pseudogenes	7
	F.	Expression arrays	7
IV	**Adding, deleting, editing, merging and splitting**		8
	A.	Editing genes	8
	B.	Deleting genes	8
	C.	Splitting genes	8
	D.	Merging Genes	8
V	**Naming of Genetic Stocks involving insertion/deletion events**		9
VI	**Naming of chromosomes and cytoplasms**		9
	A.	Chromosomes	9
	B.	Linkage groups	9
	C.	Structural change of chromosomes	10
	D.	Aneuploids	10
	E.	Cytoplasms	10
VII	**Registration of Gene Names and Symbols**		10
	A	Names and symbols for mutant genes associated with phenotypic variation	10
	B.	Names and symbols for genes predicted by sequence	11
	C.	Suggestion for consideration by the CGSNL	11
VIII	**Seed stocks**		11
IX	**Amendments**		11
X	**References**		12

Note: the text highlighted in yellow is meant for requesting suggestions from the mentioned group, researchers or the projects.

Proposition of rules for Gene Symbolization in rice

I. **Genes:** In the naming of genes, the use of an international language (English) is preferred. The name of a gene should either briefly describe the phenotype and/or convey some meaning as to the function of the gene product, if known. All new gene names should be registered with the CGSNL to avoid duplication and confounding of gene names. The rice community gives priority to the first published name for a gene but it is recognized that names change over time to reflect new knowledge. While we do not propose the adoption of a single, standardized system of gene nomenclature at this time, we do propose that a system of synonyms be adopted to permit the establishment of correspondences between sequence based gene identifiers and names based on biochemical function or phenotypic variation.

A set of rules for naming and identifying genes, loci and alleles based on biological function, mutant phenotype and sequence identity is outlined below, along with suggestions for dealing with multiple names, aliases and synonyms.

Naming genes, loci, and alleles based on biological function and mutant phenotype in rice:

A. **Gene full name**: The *full name* of a gene consists of a *name* and a *locus designator*. The name should briefly describe the salient characteristics of a (mutant) phenotype or the biochemical function of the gene product. The locus designator serves as a place holder for all allelic variants at that locus and differentiates it from other loci with similar phenotype. Historically, the gene full name started with an upper case letter if the first allele described in the literature was dominant, and with a lowercase letter if the first allele described was recessive, followed by lowercase letters. It is suggested [for discussion in the CGSNL] that in the future, new gene names begin with an upper case letter, regardless of the dominance or recessiveness of the first allele described. Gene full names are followed immediately by a locus designator, with a dash or hyphen between the name and the locus number (i.e. Shattering-1). Any gene name that does not have a locus designator is presumed to be the first such gene identified and will be assigned the locus designator, "-1", e.g. "Purple node" will be designated "Purple node-1". Historically, genes have been organized into gene classes where the gene class name designates a set of genes that have a similar phenotype. Gene names that have been commonly used in the past will be retained, unless there is redundancy, overlap or confusion caused by use of the same name for different genes or different names for the same gene. In such cases, the first-published gene name will be retained and the CGSNL will work with the authors of such publications to identify a new gene name and gene symbol for the subsequently reported gene(s) or loci.

B. **Gene symbol:** A gene symbol consists of two parts namely, a *gene class symbol* and a *locus designator*. The gene symbol should consist of three to five letters and should be derived from the full name of the gene as mentioned in section A, followed by a locus designator that serves to differentiate genes at different loci that affect the same phenotype. Capitalization of the first letter of the gene symbol should be consistent with that of the full gene name as described above. Gene class symbol should always be written in *italics*, followed by the locus designator, which is not written in italics. Both parts of the gene symbol should be written together with no space, hyphen or any other symbol between them (e.g. *Glh*1, *Glh*2, *Pi*12, *Piz,*

*Pi*ta). It is suggested for discussion in the CGSNL that the locus designator following a gene symbol should not be italicized to avoid confusion between letters and numbers, e.g. *d11,* dwarf-11, is easily confused with *dl1,* drooping leaf-1. Together, the gene class symbol and locus designator form a gene symbol which must be unique to the locus. Every gene symbol corresponds to a gene full name. Where possible, existing symbols should be retained even if they do not fully conform to this rule. Example: *C* (Chromogen for anthocyanin); *A* (Anthocyanin activator) and *wx* (glutinous endosperm). For any gene symbol that does not have a locus designator, it is presumed that the first such gene identified has the locus designator, "1", e.g. glutinous endosperm (*wx*) should be designated glutinous endosperm-1 (*wx1*). All new genes will be assigned a number as the locus designator by the CGSNL, in order of discovery.

It is suggested for discussion in the CGSNL that the use of the suffix "(t)" and "*", that are currently used as a tentative locus designations when the relationship between a newly described gene and a previously reported gene is not clear, be suspended and new genes be assigned a new locus designation, with the assumption that they are new loci. If the new gene is later demonstrated to be allelic to a previously reported locus, the original gene symbol will either be adopted or indicated as a synonym. No previously assigned gene symbols will be deleted, avoiding confusion resulting from re-usage of the same symbol. In cases where a gene symbol based on gene product or function already exists, both names will be registered and the first published takes priority unless the CGSNL determines otherwise.

C. **Allelic variants:** Different alleles of the same gene are distinguished by adding a numerical suffix, separated by a dash or hyphen, to the gene full name or the gene symbol (e.g., *shattering*-1-1; *Pgi*1-1, *Pgi*1-2). Historically, in a few cases, a letter (t) or asterisk (*), rather than a number, was used to indicate an allele and these few cases will be retained as exceptions. It is suggested for discussion in the CGSNL that alleles identified by sequence alone be treated as allelic variants of an existing locus and be given a separate allele identifier. If a sequenced allele is demonstrated to be equivalent to a previously named allele corresponding to a known phenotype or gene product, it will be assigned the existing allele identifier, based on the convention of precedence, with the other identifier retained as a synonym.

D. **Protein name and symbol:** It is recommended that the name of a protein encoded by a particular gene be consistent with the gene full name in cases where the gene name is based on phenotype or molecular function (ref: gene full name section). However, if at a later stage, the functional assignment of the gene based on phenotypic assay is determined to have a biochemically characterized, molecular function such as an enzyme or a structural component (subunit) of a macromolecular complex, it is suggested that the protein be assigned a synonym consistent with the enzyme nomenclature recommended by the IUPAC Enzyme commission or the macromolecule name adapted by the IUBMB. Despite the fact that there may be several synonyms for the protein name (and similarly, for the full gene name), it is recommended that the protein symbol always be consistent with the adopted gene symbol, with an exception that they are written using all upper case characters in italics, followed by a numeric locus identifier. For example, the glutinous endosperm-1 (*wx*1) gene encodes the granule-bound starch synthase enzyme (EC: 2.4.1.11) but the protein symbol would be "*WX*1", by keeping consistency with the gene symbol, *wx*1.

E. **Post translational modification:** In cases where a post translational modification, such as protein splicing, leads to formation of two or more protein molecules with different activities or functions, the spliced protein molecules will carry a protein name and symbol consistent with their molecular function or an associated phenotype.

II. Systematic locus id

A. **Nuclear genes:** For assigning systematic locus identifiers to genes **(**either predicted or experimentally validated) identified on the pseudomolecules assembled for rice nuclear genome, we propose to follow the recommendations for naming genes in Yeast (*Saccharomyces cereviceae*). A systematic identifier is assigned to protein-coding genes (ORFs), RNA coding genes (snoRNA, snRNA, rRNA, tRNAs) and pseudogenes. A nuclear rice gene name will consist of: a) two uppercase letters, "OS" (to indicate *Oryza sativa*), b) a number to indicate a specific rice chromosome (01-12), c) a letter indicating the polarity of the coding strand, where "W" indicates the Watson strand, which goes 5' to 3' from the telomere of the short arm to the telomere of the long arm of a chromosome, and "C" indicates the Crick strand, which goes 5' to 3' from the telomere of the long arm to the telomere of the short arm of a chromosome; d) a 5-digit number (assuming there will be fewer than 10,000 genes per chromosome) indicating the sequential order of a gene along a chromosome, independent of polarity of the strand, in ascending order from the telomere of the short arm (north side) to the telomere of the long arm (south side). The numbers indicating gene order should be initially assigned in increments of 10, thus giving room for expansion as new genes are discovered. For example, the third and fourth genes on rice chromosome 5, where the third gene is on the Watson strand and the fourth gene is on the Crick strand would be indicated as: OS05W00030 and OS05C00040. If during the course of the sequencing or based on new experimental evidence a new gene is detected on the Watson strand between two annotated genes, the new gene will be assigned a number between the two previously annotated genes. For example, a gene discovered between OS05W00030 and OS05C00040 would be assigned OS05W00035, again leaving room for expansion. In some cases, gene order within a particular chromosomal segment may not follow the ascending/descending order rule based on precedence of gene discovery, but despite this shortcoming, the use of a systematic nomenclature for genes is encouraged

B. **Organelle genes:** The main mitochondrial and chloroplast chromosomes are circular and do not have arms, so the systematic names for genes on these chromosomes will replace the chromosome number and arm designations with an upper case 'M' for mitochondrial genes or an upper case 'P' for plastid (chloroplast) genes, respectively. The letter will be followed by a 4-digit number indicating the gene designation in order along the chromosomes and a 'W' or a 'C' indicating the polarity of the coding strand, where "W" indicates the Watson strand, which goes 5' to 3' from the first base pair of the completely sequenced molecule to the last base pair in the linearized molecule, and "C" indicates the Crick strand, which goes 5' to 3' from the last base pair of the completely sequenced molecule to the first base pair in the linearized molecule. For example, OSPW0350 indicates gene 35 on the Watson strand of the rice chloroplast genome. Linear plasmids in the mitochondria will be indicated with a lower case letter (a – z, in order of decreasing plasmid size or the order of precedence by submission

to GenBank, its open to the members suggestion) following the upper case M. For example, OSMaW002 indicates gene 2 on Watson strand of the longest linear or linearized plasmid "a" (a 4.3 kb mitochondrial plasmid). The number series for the genes will start from the start of the fully assembled, sequenced organelle chromosome/plasmid determined by the sequences (to be) submitted by the author in GenBank

C. **Splice site variants:** Annotation of cDNA (but not genomic DNA) will follow the rules outlined above, with the additional modification: splice variants of a gene that are detected post-transcriptionally will be denoted by adding a suffix to the systematic name of the gene, separated by a dash. For example, splice variants of OS5W00030 will be OS5W00030-1, OS5W00030-2, OS5W00030-3 etc. Assigning number series to the splice variants will depend on the precedence of submission to GenBank. For example, the first submitted variant of OS5W00030 will be OS5W00030-1" being the first submission. PLEASE CONSULT THE cDNA ANNOTATION GROUP, TIGR AND THE IRGSP.

D. **Genes in incomplete or unanchored sequenced clones:** For regions where the genome sequence of rice is incomplete, it is suggested for discussion in the CGSNL that a name space be reserved capable of containing 200 genes per 100 kb of predicted gap. For genes identified in unanchored BAC/PAC clones, continued use of the current nomenclature system where the gene is sequentially designated by a numerical suffix following the BAC/PAC clone name assigned by the sequencing center (e.g., F23H14.13) and that the systematic nomenclature system outlined above will supercede the clone-based name once the sequence in the region is fully assembled and completed.

III. Making associations between systematic locus ID's, gene names and protein products.

A. **Prediction and validation of Gene models:** Many groups and individual researchers are in the process of identifying cDNAs that correspond to gene models predicted by annotation software. A predicted gene model may be verified by the full-length cDNA sequence or be improved and re-annotated based on EST or other available cDNA sequence. Sequence matching of cDNA sequences with predicted genes/genome sequences will also identify 'missed genes' (i.e. genes that were not predicted by automated methods) and identify hypothetical proteins as being real (i.e. they are expressed). Curators will continue to primarily use BLAST to make associations between experimentally verified cDNAs and genes, and it would be helpful if such information were included in the definition lines for the cDNA sequences submitted to GenBank, as described for *Arabidopsis* by TAIR and TIGR http://www.arabidopsis.org/info/tigr_naming.jsp .

B. **Gene names based on homology or orthology:** For example, *Oryza sativa* clone AB041838 corresponds to a rice cDNA whose sequence matches the *CONSTANS* gene (*CO)* from *Arabidopsis*. The gene is associated with flowering time, specifically with the QTL, *Hd*1 (Heading date-1) as well as the previously identified mutant, *Se*1 (Photosensitivity-1) in rice. The product of the gene is a transcription factor with sequence similarity to the CONSTANS B-box zinc finger protein family (http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?pid=11094205). In GenBank, the entry was submitted with the name, *Hd*1. However, a previously named gene, called "Histone

Deacetylase 1" (named for the gene product="histone deacetylase") also carries the gene symbol *Hd*1 (GenBank accession AAK01712 for rice and other model organisms). Thus, using the name *Hd*1 for the *Constans* homolog in rice is confusing and precedence would suggest that the name *Se*1 be reassigned with the mutant gene symbol, *Hd*1 listed as a synonym.

C. **Discrepancies and use of species name in gene symbol:** Additional discrepancies in assigning names and symbols to genes of known function may be noted where people have used different names for *histone deacetylase*-1 in rice. For example, in August, 2002, Song and Goodman submitted a sequence to GenBank for *histone deacetylase*-1 in rice, naming it OsHD1 (GenBank accession: AAK01712 (http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=AAK01712). In February 2003, Jang et al. published a paper in the Plant Journal on "rice class-I type histone deacetylase genes" and referred to a gene called OsHDAC1. (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12581311&dopt=Abstract). The use of organism specific prefixes in the gene name or gene symbol, such as "Os" (*Oryza sativa*) is discouraged, because it is redundant with information associated with submitted/registered genes. Also, it leads to a proliferation of gene names "*Oryza sativa* X". This relationship between the gene and the organism is easily maintained between the published records and sequence databases. The authors may append the organism specific prefixes for clarity in publications, however it should not become part of the adopted gene symbol or gene full name, though it is used in the sytematic_locus_ID (e.g. OS5W00030).

D. **Gene models with protein matches** (BLASTP score >100) and having experimental evidence supporting the protein function may be named after the protein database entries as "putative XXX", . When the match is from beginning to end, the name of the database match is also indicated in the definition (with the version number of the database and version number of the software used, if possible). When the match is restricted to certain domains, a general name for that class of protein is used. Gene models with only EST matches are named as [??? PROTEINS?- TIGR and IRGSP advise needed]. Gene models without any database matches are called "hypothetical proteins". For suggestions, please consult http://www.arabidopsis.org/info/tigr_naming.html

E. **Pseudogenes:** Molecular technology has identified sequences that bear striking homologies to structural gene sequences but are not transcribed. These sequences are termed pseudogenes. In order to show the relatedness of pseudogenes to functional genes, pseudogenes will be identified with the gene symbol of the structural gene followed by a ".P"(symbol "Period" and Capital letter "P") for pseudogene. Pseudogenes may be on different chromosomes or closely linked to the functional gene and may occur in varying numbers. The same is suggested for pseudogenes identified in mitochondrial and plastid (chloroplast) genomes. Examples: *Act*B.P1 (actin beta pseudogene-1); *Act*B.P2 (actin beta pseudogene-2), etc. The pseudogenes will be identified with the gene symbol of the structural/functional gene followed by a ".P" instead of conventionally used Greek symbol for "psi". Examples: *rps*14.P instead of *rps*14ψ for pseudo ribosomal protein S14

F. **Expression arrays:** Caution should be exercised in assigning names to ORF's identified as part of bulk expression experiments. For example, a microarray experiment might identify

23 ORFs that are potentially up-regulated in response to a specific treatment, such as salt stress (see manuscript available at http://www.plantcell.org/cgi/reprint/13/4/889.pdf). Unless sequence identity indicates that an ORF corresponds to a known gene, or there is other experimental data for gene function, it is best to publish the accession number of the cDNA only, or to provide the gene name of the entry that has the maximum similarity score with the ORF in question. For example, GenBank AC No. AF169966, *Oryza sativa*, putative *cycloartenol synthase*-1(*Cas*1) mRNA. Another example is AC No. AP000391, *Oryza sativa*, similar to *ribitol dehydrogenase* isolog. In cases where the ORF does not correspond to any known gene and there is no sequence similarity or experimental data associated with a gene function, the gene should be identified by its systematic locus identifier (assigned by the sequencing and annotation groups). Where possible, curators/authors will make associations between sequences represented in commonly accessed arrays (as they become publicly available) and corresponding gene entries in GenBank, Gramene, INE or other databases.

IV. Adding, deleting, editing, merging and splitting

A. **Editing genes:** Consistent use of the gene identifiers such as gene name/symbol/systematic locus ID is suggested, as long as there are no major changes in the gene model or function. It is applicable as long as the modifications do not lead to a completely new gene or a function. For example in cases where the gene encodes for an ORF, the modifications in annotation can change the exon boundaries or the addition or deletion of exon(s) or intron(s) may occur. In other cases, the open reading frame may change due to updated annotation and in such cases the gene name and symbol should reflect the change in molecule's function. The locus ID may remain the same.

B. **Deleting genes:** Genes identified by computational methods, may prove false positives when confirmed by wet lab experiments. In such cases, all gene records and corresponding identifiers should be preserved with a flag "OBSOLETE" and never "DELETED" from data repositories. The flag "OBSOLETE" ensures that the same identifiers are not used again for a new gene, thus avoiding a situation that would lead to confusion.

C. **Splitting genes:** When it is determined that a locus identifier actually refers to more than one gene (e.g. two genes mistakenly identified as one by prediction method), one of them with majority of the sequence will retain the original identifiers and the other gets a new systematic identifiers. The editing of gene name and symbol applies to accommodate the new function if applicable. However, in cases such as Amy1A/C, where two closely linked loci identified previously as one, now identified as two genes with their respective physical locations. Therefore, such genes should split into two e.g. Amy1A and Amy1C with both carrying Amy1A/C as synonym.

D. **Merging Genes:** In the cases where there is experimental evidence (such as full length cDNA) indicating that the two previously identified genes are actually part of the same gene, the two genes should be merged into one and the locus identifiers/gene name/ gene symbol should be retained from the gene with the majority of the sequence. For the second gene, the locus identifiers/gene name/ gene symbol should be made obsolete and termed as secondary identifiers or synonyms to the first one.

V. Naming of Genetic Stocks involving insertion/deletion events

In accordance with established nomenclature systems in other organisms, a *delta* ("Δ") will be used to indicate a deletion and a *double colon (*"::") for an insertion. For example, where the insertion/deletion is the result of an experiment involving chemical mutagenesis or radiation, a deletion that disrupts the *Acp*1.1 gene in cultivar Nipponbare (wild type) would be represented as: <Nipponbare:Δ*Acp*1-1>. A double mutant with insertions in both the *Acp*1-1 gene and the *Adh*1 gene would be represented as: <Nipponbare:Δ*Acp*1-1; Δ *Adh*1>. In the case of a set of lines in which insertions are caused by tissue-culture activation of the native retrotransposon, *Tos*17, in cultivar Nipponbare (wild type), a line in which the *Acp*1-1 gene is disrupted will be represented as < Nipponbare:*Acp*1-1::*Tos*17>. [Comments from FUNCTIONAL GENOMICS groups would be appreciated here.] Where insertion/deletion stocks are generated via transgenesis, the prefix 'TG' will be used before the name of the wild type cultivar, followed by the conventions described above. Transgenic (TG) lines include both integrative (heritable) and non-integrative (transient) transformation events. Constructs involved in transformation may be designed for co-suppression, promoter traps, enhancer traps, gene traps, activation tags, transposon tags, T-DNA insertions, etc. For example, in a set of transposon tagged lines containing Ac/Ds tags in the cultivar Nipponbare, a line where the *Acp*1-1 gene is disrupted would be represented as: <TG:Nipponbare:*Acp*1-1::Ac/Ds> . A transgenic line designed to produce the *Bt* crystal protein, introduced using a construct consisting of the *nos* promoter, the *gus*A reporter gene, a hygromycin selectable marker, the *Bt* gene and the *nos* terminator, the line would be represented as: <TG:Nipponbare::*nos*P:*gus*A:hygro:*Adh*1:*nos*T>. A line that was transformed for transient expression of the *Adh*1 gene using the plasmid pPBI101 would be indicated as: <TG:Nipponbare::*Adh*1(pPBI101)>.

In cases where the insertion or deletion can be mapped to either an assembled sequence from a chromosome or an ordered BAC/PAC clone, in place of a gene name, the user can assign either an identified gene name (transcribed element) or and already assigned systematic name for the ORF. In cases where the insertion does not fall in the genic region (transcribed region) the user should use the chromosome number/sequenced clone name followed by the base pair position of the point of insertion/deletion.

VI. Naming of chromosomes and cytoplasms

A. **Chromosomes**. The chromosomes are assigned Arabic numerals in descending order of their pachytene length (or centromere position in case of ambiguity of length). Short arms are symbolized by "**S**", long arms by "**L**" (Example:1S, 1L). It is acceptable to abbreviate them as chr. 1S, chr. 2S or Chr. 1S, Chr. 2S. The circular chromosomes found in plastid and mitochondrion are assigned English characters P and M in capitol, instead of the Arabic numerals recommended for nuclear chromosomes. Since these chromosomes are circular and do not have centromere, therefore, they will not carry representation for short or long arms. It is acceptable to abbreviate them as chr. M or chr. P

B. **Linkage groups**. Linkage groups are numbered to correspond with the respective chromosomes. If two linkage groups are known to correspond to the same chromosome,

but cannot otherwise be linked, a small letter is appended to the chromosome number, e.g., 8a and 8b. Marker positions are assigned starting from the distal end of the short arm of each chromosome. Positions start from 0 cM on genetic maps and 1 bp on sequence maps.

C. **Structural change of chromosomes**. Chromosomal changes are denoted by a symbol showing the type of aberration plus the chromosome number(s) involved. The symbols used are: **Dp** for duplication, **In** for inversion, **Tn** for translocation, **Tp** for transposition. To distinguish between similar aberrations involving the same chromosome(s), lower-case letters are used following the chromosome numbers (Example: In(1)a, In(1)b, Tn(1-2)a Tn(1-2)b).

D. **Aneuploids**. Monosomics and primary trisomics are designated according to the additional chromosome (Example: Mono1, Mono2, Triplo1, Triplo2).

E. **Cytoplasms.** The cytoplasm type is represented by an abbreviated name of the cytoplasmic group (as assigned by the CGSNL) in which the cytoplasmic trait was identified and the name will be enclosed in square brackets. In cases where the cytoplasm is conferring cytoplasmic male sterility, the abbreviation "cms" followed by a dash ("-") will be used as a prefix to the cytoplasm name. For example: [cms-WA] (from wild-abortive cytoplasm).

VII. Registration of Gene Names and Symbols:

A. **Names and symbols for mutant genes** associated with phenotypic or anatomical variation should be requested of the Rice Genetics Cooperative Convener of Gene Names and Gene Symbols (Dr. Atsushi Yoshimura/or the subsequent office bearer).

a) Information to be submitted includes all of the following categories, as available:
b) descriptive information about the character (trait) effect
c) inheritance and allelism data
d) sequence data
e) protein family relationship
f) a photograph of the mutant.

After an examination of priority, the Convener will notify the author of verification of the symbol and will include the symbol in the list of allele symbols published in OryzaBase and the Gramene Database (or other suitable sites), as well as in the Rice Genetics Newsletter.

Publication of a research note on all new mutants in the Rice Genetics Newsletter [or some other publication?] is recommended and who ever hosts the rice gene symbol database web page should be notified by the approver/convener. Members of the IRGSP and any other interested parties should also be notified, so that the information can be included in the rice genome annotation effort.

B. **Names and symbols for genes predicted by sequence** will be assigned according to the conventions outlined in this document and adopted by the International Rice Genome Sequencing Initiative (IRGSP). The registration process will be managed by agreement among the sequencing/annotation centers [TAKUJI SASAKI & ROBIN BUELL, PLEASE PROVIDE INPUT HERE]

C. **Suggestion for consideration by the CGSNL**: To request an official gene name and gene symbol, a researcher will be required to prepare a textual description of the mutant phenotype, information about map location, sequence identity and biological function (if known) and send it to the convener via an electronic submission form which can be developed in the future and jointly provided via the OryzaBase database (http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp), the Gramene database (**http://dev.gramene.org/gene_symbol/submission**) or any other resource designated by the CGSNL.

VIII. **Seed stocks**: Authors publishing reports describing genes, genomes and mutants in international journals are required to make sequences, clones and genetic materials available to the public.

Contact information for Rice Germplasm Collections throughout the world is listed below:

International Rice Germplasm Center in the Philippines,
[http://www.irri.org/GRC/grchome/home.htm]

International rice information system
[http://iris.irri.org/]

List of stock centers in Japan
[http://www.shigen.nig.ac.jp/rice/oryzabase/strains/queryForm.jsp]

National Plant Germplasm System (NPGS) in the US
[http://www.ars-grin.gov/npgs/searchgrin.html]

Chinese Crop Germplasm Information System
[http://icgr.caas.net.cn/cgris_english.html],

Directorate of Rice Research, India
[http://www.drrindia.org/]

National Plant Genetic Resources Center in Taiwan
[http://192.192.196.1/index.html]

IX. **Amendments**. It is recommended to the CGSNL that suggestions for amendments of these rules can be suggested using an online "Suggestions" form available on OryzaBase and Gramene database web sites or any other site suggested by the CGSNL [NEED TO

PROVIDE AN ONLINE FORM FOR SUGGESTIONS IN BOTH ORYZABASE AND GRAMENE] . Amendments will be announced in the Rice Genetics Newsletter and via the OryzaBase and Gramene Databases (and/or other suggested sources).

X. **References** (URLs related to Gene Nomenclature)

- (Arabidopsis) http://www.arabidopsis.org/info/guidelines.jsp
- (Yeast) http://genome-www.stanford.edu/Saccharomyces/gene_guidelines.html
- (Alleles) http://hermes.bionet.nsc.ru/pg/30/weeden.htm
- (Proteins) http://www.chem.qmw.ac.uk/iubmb/newsletter/1996/news9.html
- (Others) http://www.genome.ad.jp/kegg/kegg4.html
- (Human) http://www.gene.ucl.ac.uk/nomenclature/guidelines.html
- (Mouse) http://www.informatics.jax.org/mgihome/nomen/
- (Solanaceae) http://www.sgn.cornell.edu/solanaceae-project/SOL_part3_draft2_20040215.pdf
- Kinoshita T., 1986. Report of the Committee on Gene Symbolization, Nomenclature and Linkage Groups. RGN 3:4-8. (http://www.gramene.org/newsletters/rice_genetics/rgn3/v3C.html)