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 rice annotation project (RAP-1) meeting held at Tsukuba, Japan from December 16-18th, 2004 and revised this document accordingly. 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 an official document on the gene nomenclature system adopted for rice with a special emphasis on the sequenced genome.]

Introduction

The biological community is moving towards a universal system for the naming of genes as described for number of plants such as *Arabidopsis* (TAIR, 2005), Tomato (Mueller, 2005), *Medicagio* (VandenBosch and Frugoli, 2001), maize (MaizeGDB, 2002), or those of the metazoans such as for mice (MGNC, 2005), human (Wain *et al*., 2004) and yeast (SGD, 2005). 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 (*Oryza*) reflect the knowledge about both the phenotypic consequences of a particular allele in a given genetic background and the biochemical features of a specific gene, gene model 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, characterization and describing genes are increasingly applied to rice and the recent report on the completion of rice genome sequence (Project, 2005), nomenclature rules are needed that outline the standard procedures for describing genes based on DNA, RNA and protein sequence analysis, biochemical characterization (Wu *et al*., 1991). This is in addition to those previously outlined for phenotypic variants (Kinoshita, 1986).

The focus of this publication is to summarize the existing rules for gene nomenclature in rice and to outline a the new rules, recommended by the IRGSP (International Rice Genome Sequencing Project) members present at the RAP-1 meeting held at Tsukuba, that will bring the gene nomenclature for rice into agreement with the emerging nomenclature system being developed for other model organisms.

- Assign a stable unique identifier to the genes/locus, transcripts, proteins and other features like transposable elements (TEs), found on the sequenced genome assembly of the Oryza sativa cultivar Nipponbare.
- Precedence of publication (published either in a peer reviewed journal, Rice Genetics Newsletter or submitted to the GenBank) will be the primary determinant of a gene

name and gene symbol (gene identifiers) unless there is a gene name/symbol conflict and an alternative name is proposed or 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/locus, the case will be referred to the Committee for appropriation. Priority is given to precedence.
- In cases where multiple names have been used over time and they are widely recognized, these names can be acknowledged by the assignment of synonyms, with priority being given to the precedence on registration and /or publication.
- To ensure that historical names are not lost, links will be maintained between existing (historical) gene names that are based largely on mutant phenotype depicting forward genetics approaches and new names that are likely to be based largely on sequence annotation, and biochemical assays, referring to the reverse genetics approaches.

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 the rice gene nomenclature system into greater harmony with that of other model species.

I. Naming of chromosomes

The nuclear chromosomes are assigned Arabic numerals in descending order (1, 2, 3, , ,12) of their pachytene length (or centromere position in case of ambiguity of length). For the purpose of their usage in the locus identifier (see the following sections), the chromosomes will be assigned a two digit number starting with 01 up to 12. If required the 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 plastids (chloroplast) and mitochondrion are assigned English characters “Pt” and “Mt”, 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. Mt or chr. Pt. In the case of genetic experiments on mapping populations, the often used linkage groups are assigned numerals corresponding to the respective chromosomes

II. Systematic locus id

Systematics locus id is assigned to genes, transcripts, proteins and Transposable elements (TEs) found on sequenced genome. This is done to facilitate the better organization of the locus IDs, easy recognition of the locus and its position in the sequenced rice genome. Since majority of the sequenced genes will not have any known functional characterization, name and symbol, this ID will serve as its name and symbol as an interim option until the locus gets a proper name and symbol based on appropriate laboratory experiments. To avoid the existing and future ambiguity in the gene nomenclature, the locus IDs will be the only unique feature of the locus by which it can be identified. This rule must be followed, because despite best efforts of the nomenclature committee and datbases, researchers often name different loci with same name. The following sections will describe the locus ID rules. Database curators and individual researchers must not assign names and symbols unless approved by the CGSNL (or at least registered). The systematic locus ID rules described as follows are for the sequenced genome of *Oryza sativa* ssp japonica Nipponbare cultivar only. The locus IDs identified in the genomes from other cultivars, subspecies and species of genus Oryza, must consult the CGSNL for suggestions.

A. **Nuclear genes:** For assigning systematic locus identifiers to genes (either predicted by gene prediction programs, ortholog alignments or experimentally validated by alignment of expressed sequences ESTs and full length cDNAs) identified on the assembled chromosome contigs, we propose to follow the modified recommendations adopted for Yeast Saccharomyces cereviseae; (SGD, 2005) and Arabidopsis (TAIR, 2005). A systematic identifier is assigned to protein-coding genes (ORFs), RNA coding genes (snoRNA, snRNA, rRNA, tRNAs and miRNAs) and pseudogenes. A nuclear gene locus id will consist of: a) a uppercase letter “O” and lowercase letter “s” to indicate the rice species *Oryza sativa*, b) a two digit number to indicate a specific rice chromosome (01, 02, 03,-12), c) a letter “g” indicating that the locus id is for a gene; 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, 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 that is independent of the polarity of strand (+/- or Watson or Crick), 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 would be indicated as: Os05g00030 and Os05g00040.

If during the course of the sequencing or based on new experimental evidence a new gene is detected between the two already annotated genes, the new gene will be assigned a number between the two previously annotated genes, by using the tenth number space. For example, a gene discovered between Os05g00030 and Os05g00040 would be assigned Os05g00035, again leaving room for expansion. The downside of adopting this strategy is that, 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. The locus ids will be assigned to all the genes including those that are known to be present on the regions of various chromosomes, where an insertion of the portion of the organelle (plastid or mitochondrial genome) has been detected. It is quite possible that such genes are non functional or pseudogenes.

For regions where the genome sequence of rice is incomplete, such as the telomeric ends, centromeric regions and those in the unsequenced regions or gaps in the assembled chromosomes, it is suggested that a name space be reserved capable of containing 1000 genes per telomeric ends, 1000 genes in the centromeric region and 50 genes per 100kb region of a gap.

B. **Organelle genes:** The main mitochondrial and chloroplast chromosomes also called master circle are circular and do not have arms. Therefore, compared to the systematic locus IDs for nuclear genes, the genes found on organellar chromosomes will replace the chromosome number and arm designations with symbols ‘Mt’ for mitochondrion and ‘Pt’ for plastid (chloroplast), respectively. These letter will be followed by a letter “g” indicating that the locus corresponds to a gene, followed a 5-digit number (assuming there will be fewer than 10,000 genes per chromosome) indicating the sequential order of a gene along an organellar chromosome, independent of polarity of the strand, in ascending order from the first base pair of the completely sequenced molecule to the last base pair in the linearized molecule. The base pair position is dependent on the assembled molecule submitted by the author to any of the reference sequence databases namely, NCBI-GenBank, DDBJ or EMBL. For example, OsPtg00010 indicates first gene on the rice plastid genome. Therefore in the GenBank entry NC_001320 for plastid genomes sequenced from Nipponbare cultivar, the locus id OsPtg00010 would be for *psbA* gene (82-1143bp).

In addition to the suggestions for genes found on master circles in the organelle genomes, plasmids, both linear and circular or subgenomics circles in the organelle mitochondria will be indicated with a lower case letter a–z, (in the order of precedence by submission to GenBank) immediately following the organelle symbol Mt or Pt. For example, OsMtag00020 indicates gene 2 on the 2135bp long mitochondrial plasmid B1 (GenBank accession NC_001751). The number series for the genes will start from the start of the fully assembled, sequenced plasmid or subgenomic circle determined by the sequences (to be/were) submitted by the author in GenBank

C. **Transcript id:** Every known or predicted form of transcript of a gene will be assigned the systematic identifier same as the locus identifier except that the letter 'g' for gene will be replaced by letter 't' for transcript. Thus assuring consistency in the gene's locus id and its transcript id. e.g the transcript Os05t00030 is transcribed by the locus Os05g00030 representing gene 3 on chromosome 5. Sometimes the nascent transcript undergoes alternative splicing. In order to clearly identify the alternatively spliced forms of the transcripts a two digit suffix to the systematic transcript id of the gene, separated by a dash e.g. -01, -02, -03, ….-99 in the order of their discovery will be added.. By default the very first transcript or the only transcript identified, irrespective of the locus whether it is associated to alternatively spliced transcript forms or not, the transcript id will always have number "-01" suffixed to the initial transcript id. For example, the transcript id of the locus Os05g00030 with no known splice variants will be Os05t00030-01. Whereas if there is a report suggesting that transcript from this locus undergoes alternative splicing and creates 3 alternative forms. In this case if any of the three forms is identical to the -01, it should reference the same transcript id and the other forms would acquire the ids Os05t00400-02 and Os05t00400-03. Assigning number series to the splice variants will depend on the precedence of identification, the submission to GenBank or by the size of the cDNA. For example, the first submitted alternative transcript of locus Os05g00030 will be Os05t00030-02. Any new alternative forms will be numbered sequentially.

D. **Protein id:** All the peptides deduced experimentally or computationally from a gene sequence/transcript will be assigned the systematic identifier same as the transcript identifier except that the letter 't' for transcript will be replaced by letter 'p' for protein. Thus assuring consistency with the gene's locus id and its transcript id. E.g. the protein Os05p00030-1 translated from transcript Os05t00030-1 that is transcribed by the locus Os05g00030 representing gene 3 on chromosome 5. In order to avoid conflicts with the proteins deduced from alternatively spliced forms of the transcripts from a single locus, the protein id must reflect the corresponding transcript from which it is deduced except for the letter "t".

E. **Genes present on unanchored sequenced clones:** 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 supersede the clone-based name once the sequence in the region is fully assembled and completed. In such cases the earlier clone based locus identifiers must become either the alternate ID or the gene synonym.

F. **Transposable elements:** Similar to the gene locus, it is recommended that the TEs must be assigned a locus ID in the format OsXXe#######, where XX is the chromosome number, followed by “e” for elements (transposable/repeats) and ####### is a 7-digit chromosome-specific ID. The ids are assigned incrementally starting from the telomere end of the short arm of the chromosome to the telomere end of the long arm. We propose incrementing the initial IDs by 100 and leaving large ID spaces about a 1000 per 100kb of physical gap. E.g. the 999th TE present on chromosome 5 will have a locus id Os05e0099900. DO YOU THINK THAT THE SSR LOCI CAN BE PART OF THIS AND AN ID CAN BE ASSIGNED TO THEM? IN SUCH A CASE THE RM IDs CAN BE CONSIDERED AS LOCUS NAME/SYMBOL.

III. Adding, deleting, editing, merging and splitting locus

A. **Editing a locus:** Consistent use of the locus identifiers, full gene name and gene symbol 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 change in the start position of the locus. For example in cases where the gene encodes for an ORF, the modifications in annotation can change the intron-exon boundaries, the addition or deletion of exon(s) or intron(s), change of strand and change or modification of function or an associated phenotype. In other cases, the open reading frame may change due to updated annotation and in such cases the gene’s full name, symbol and the definition line of the GenBank/DDBJ/EMBL records should reflect the change in molecule’s function. In all the above cases the locus id must remain same.

B. **Deleting a locus:** genes or other features such as TEs identified by computational methods, may prove false positives when confirmed by wet lab experiments, thus making it necessary to retire the locus. In such cases, all the 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 locus, thus avoiding a situation that would lead to confusion and if required can still be referenced.

C. **Splitting a locus:** When it is determined that a locus identifier actually refers to more than one object (gene or TEs) (e.g. two genes mistakenly identified as one by prediction method), one of them closer to the locus start position will retain the original locus identifier, gene name and the symbol and the gene present in the newly identified locus gets a new locus identifier, name and a symbol following the recommendations mentioned before. The modification of gene name and symbol applies to accommodate the new function if applicable.

D. **Merging loci:** In the cases where there is experimental evidence (such as full length cDNA) indicating that the two previously identified genes are actually one gene or part of the same locus, the two loci must merge into one, The new locus must retain the locus identifiers, gene name and gene symbol from the locus closer to the start position. For the second gene, the locus identifiers, becomes secondary locus id of the associated to the first one, whereas the gene name and gene symbol should be called as synonyms to the first one.

IV. Rules for Gene Symbolization in rice

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 and approved 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.

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 biochemical function of the gene product or the phenotype rendered due to mutant or allelic forms of this gene. The locus designator serves as a place holder for all allelic variants at that locus and differentiates it from other loci with similar characteristics. 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. In view of the recent advances of identifying genes based on sequences and large scale genomics efforts, it is no more a priority to investigate the dominance or recessiveness of the allele being studied. Therefore, it is recommended that the newly identified genes carry a full name beginning with an upper case letter, regardless of the dominance or recessiveness of the first allele described. In case of the IRGSP's genome assembly of Oryza sativa cultivar Nipponbare, all the loci are referring to the alleles identified in cultivar Nipponbare. 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). This locus identifier is suggesting the first gene identified for a given function or the associated phenotype. By default, 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. If the newly identified genes are proven to be the same as the phenotypically identified genes such as those listed by (Kinoshita, 1986)) the precedence rule applies 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 CSGNL 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. The genes can be assigned a gene name only if there is an experimental evidence reporting a molecular function, role in a biological process, interaction or the phenotype. The names cannot be assigned based on computationally determined sequence similarity to homologs, orthologs or paralogs or those based on the presence of a consensus feature such as an interpro domain, unless there is substantial evidence and experimental proof. In the event of non availability of the gene name, the gene name field is kept empty and it is suggested that the description/definition field is utilized to mention the appropriate characteristics of the gene. RAP2 has decided to use the descriptions in a standardized format (Table-1). The authors of the publication are suggested to use either the 'systematic locus id' or the GenBank accession number, if the appropriate gene name is not available. Use of locus identifier is preferred. Once again databases, curators and individual researchers must not assign gene names unless approved by CGSNL.

Table-1

Description of RAP1 annotation (conclusion)			
	classification	standard	description
Category I	Identical to known rice protein.	Identity >= 98%, length coverage = 100% to known rice protein. [blastx]	receive the same, original gene name
Category II	Similar to known a known protein.	Identity >= 50% to a known protein. [blastx]	receive "original gene name, putative."
Category III	InterPro domain containing protein.	Non of category I or II, and hit to InterPro domain.	receive "InterPro name domain-containing protein."
Category IV	Conserved hypothetical protein.	Hit with hypothetical protein; identity >= 50%, length coverage >= 50%.	receive "Conserved hypothetical protein"
Category V	Hypothetical protein.	Non of category I to IV	receive "Hypothetical protein"

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 the same locus designator that serves to differentiate genes at different loci that affect the same phenotype, mentioned in section A. 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*, Pi*z*, Pi*ta). The locus designator following a gene class 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 and the genome. Every effort should be made to assign gene symbol corresponding 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 with similar characteristics will be assigned a new number as the new locus designator by the CGSNL, in order of discovery.

The suffix "(t)" and "*", that are currently used (Kinoshita, 1986) 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 records of the two should be merged and the original gene symbol will be adopted by precedence rule. The other symbol(s) are termed as synonym(s). No previously assigned gene symbols will be deleted, thus avoiding confusion resulting from re-usage of the same symbol. Assigning a symbol to a gene should be consistent with that of the full gene name as described above

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 (*sh*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 precedence rule, with the other identifier retained as a synonym. By default the allele from the Nipponbare genome assembly from IRGSP is always the first allele.

D. **Use of species name in gene name and symbol:** The use of organism specific prefixes such as "Os" (*Oryza sativa*) in the gene name and/or gene symbol 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. However as an exception, the authors may append the organism specific prefixes for clarity in publications in order to avoid redundancy in usage of species name all the time a gene is referred. In any case the species symbol should not become part of the adopted gene symbol or gene full name. The symbol “Os” is allowed for use in the sytematic_locus_ID (e.g. OS5W00030) as well as the usage of locus id while referring to a gene which does not have a assigned name and symbol.

E. **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 (wx1) gene encodes the granule-bound starch synthase enzyme (EC: 2.4.1.11) but the protein symbol would be “*WX1*”, by keeping consistency with the gene symbol, *wx*1. The protein name would be WAXY-1 and ‘GBSS’ granule-bound starch synthase as synonyms. If the name cannot be assigned based on phenotype, known biochemistry and/or non-availability of gene name and symbol, the situation consistent with description (Table-1) must be used.

F. **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, and will carry the name and symbol from the primary molecule as synonyms.

V. 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 sequences. Sequence matching of cDNA sequences with predicted genes 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 gene models and the cDNA sequences submitted to Ge nBank.

B. **Gene and protein names based on homology or orthology:** As described in table-1, the name and symbol assignment for sequenced genes/locus is classified into five categories. ,. #1 Identical to a known rice gene or it product (protein), where the identity is 98% and covers the full lengthe known. #2 Similar to a known protein from rice or non rice sources with a minimum of 50% identity, #3 Domain containing gene, where the translated protein or the gene may have some known domains found on it either by experimental or prediction based on consensus sequences #4 Conserved hypothetical gene, with a 50% similarity to another hypothetical gene or gene product from rice/non-rice sources and #5 Hypothetical gene, where non of the categories 1-4 qualify. If nothing is known about the gene the experimental supports of full length cDNAs or ESTs can be used to describe it as expressed gene.

C. **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: *ActB*.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

D. **Expression based nomenclature:** Caution should be exercised in assigning names to genes 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 of 98% or higher indicates that a genemodel found on the genome corresponds to a known gene, or there is other experimental data for gene function, it is best to publish either the locus id or the Genbank accession number of the gene/cDNA only, or to provide under the definition section of Genbank entry, the gene name of the entry that has the maximum similarity score with the gene in question. For example, GenBank AC No. AF169966, *Oryza sativa*, putative *cycloartenol synthase*-1(*Cas*1) mRNA. 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 as mentioned in table-1.

VI. 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.

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

C. Sequencing Initiative (IRGSP). The registration process will be managed by agreement among the sequencing/annotation centers [TO BE MANAGED BY RAP DATABASE]

D. Registration process: Information to be submitted includes all of the following categories, as available:
   a. Descriptive information about the characteristics, that may include information on the molecular function, role in a biological process, location in a subcellular component, tissue and growth stage specific transcript and/or protein expression, observed phenotypes.
   b. Inheritance and allelism data
   c. Source germplasm. If from a cross (hybrid) provide information on all the parent germplasms
   d. Source Oryza species. If from germplasm/accession isolated from a interspecific cross, provide species info of all the parents
   e. Sequence data
      i. GenBank accession number or locus id on the sequenced genome
      ii. Sequence
   f. Protein/gene family relationship
   g. A photograph of the mutant phenotype or any other supporting document such as sequence alignments, RNA and/or protein expression, enzymatic direct assays etc.

E. 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.

   After an examination of priority, the Convener will notify the author of verification of the symbol and the ful name. Convenor is required to communicate with following to include the symbol in the list of gene/allele symbols published in OryzaBase , Gramene and IRGSP Databases (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.

VII. **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 (http://www.gramene.org/documentation/nomenclature/)] . Amendments will be announced in the Rice Genetics Newsletter and via the OryzaBase and Gramene Databases (and/or other suggested sources).

References:

1 Kinoshita, T. 1986. Report of the Committee on Gene Symbolization, Nomenclature and Linkage Groups. Rice Genetics Newsletter **3**: 4-8.

2 2002. *A Standard For Maize Genetics Nomenclature*. http://www.maizegdb.org/maize_nomenclature.php.

3 2005. *Mouse Nomenclature Home Page*. http://www.informatics.jax.org/mgihome/nomen/.

4 2005. *SOL Project Sequencing and Bioinformatics Standards and Guidelines*. http://www.sgn.cornell.edu/documents/solanaceae-project/docs/tomato-standards.pdf.

5 Project, I. R. G. S. 2005. The map-based sequence of the rice genome. Nature **436**: 793-800.

6 2005. *SGD Gene Naming Guidelines*. http://www.yeastgenome.org/gene_guidelines.shtml.

7 2005. *Arabidopsis Nomenclature*. http://www.arabidopsis.org/info/guidelines.jsp.

8 VandenBosch, K. A. and J. Frugoli. 2001. Guidelines for genetic nomenclature and community governance for the model legume Medicago truncatula. Mol Plant Microbe Interact **14**: 1364-1367.

9 Wain, H. M., M. J. Lush, F. Ducluzeau, V. K. Khodiyar and S. Povey. 2004. Genew: the Human Gene Nomenclature Database, 2004 updates. Nucleic Acids Res **32**: D255-257.

10 Wu, R., A. Hirai, J. Mundy, R. Nelson and R. Ridriguez. 1991. Guidelines for Nomenclature of Cloned Genes or DNA Fragments in Rice. Rice Genetics Newsletters **8**: 51-53.