Arabidopsis Nomenclature

This page provides information about and guidelines for nomenclature used for Arabidopsis loci, genes,markers, polymorphisms, clones,vectors and stocks from the ABRC. Use the outline below to find information for nomenclature of specific data types in TAIR.

Please READ: Community Standards for Arabidopsis Genetics [download PDF]. Standards of gene nomenclature have been adopted by the Arabidopsis community and should be followed in publications and presentations.

TAIR provides a Gene Class Symbol Registry. You can register a gene symbol currently in use by your lab (e.g. *STM*) or reserve a gene class symbol (e.g. *CYP*) here.

Please contact TAIR to request a new locus identifier (i.e. AGI Locus code). Consistency of locus identifiers and avoidance of duplication can only be achieved if individuals do not assign locus identifiers on their own. Once you have the new identifier please contact TAIR to provide any information you have about the function or expression pattern of the new gene. If you are registered at TAIR, you can submit functional annotation information directly or re gister a new gene symbol. Registration is free.

Locus Nomenclature (Modified from MIPS AGI-codes section)

Designation of unique locus identifiers is performed as part of the genome sequence annotation at TAIR. The following section describes the syntax of chromosome based locus nomenclature and how locus identifiers are assigned. In some cases locus identifiers have been made obsolete. If you have information about a sequenced locus that has not been given a locus identifier, please contact curator@arabidopsis.org.

Guidelines for use of unique gene ids (modified from MIPS)

- Format of chromosomal based nomenclature
  - AT (Arabidopsis thaliana)
  - 1,2,3,4,5 (chromosome number) or M for mitochondrial or C for chloroplast.
  - G (gene), other letters possible for repeats etc.)
  - 12300 (five-digit code, numbered from top/north to bottom/south of chromosome)
- Chromosome based locus identifiers are assigned to
  - protein-coding genes
  - RNA coding genes (sn, r, tRNAs)
  - pseudogenes
- Chromosome based locus identifiers are not assigned to
  - transposons
- Usage
  - The first AGI locus identifier release made use of locus identifiers ending in zero, eg 10010, 10020, 10030 and so on so that intervening numbers could be used for newly discovered genes.
  - Where there are gaps in the sequence, the first release skipped at least 200 codes for each 100 kb of gap.
  - In the first release, some genes were present as fragments as they lie across the boundary of two BACS. Each fragment got its own locus identifier if there was no way to represent the whole gene. There gene fragments were merged into a single locus in later releases, and one of the AGI locus identifiers became obsolete.

Adding, deleting, editing,merging and splitting

- Adding new genes
  - If there are free ATxGxxxx0 locus identifiers, those will be assigned first as in the rules above. If not, the last digit will be used, leaving space as appropriate, i.e. ...5 if the new gene is in the middle or ...8 if it is close to the neighbor with higher identifier. If there are no free

identifiers between the neighboring genes at all, the nearest free identifier will be used. As often as possible, the sequential numbering of genes along the chromosome will not be disturbed, but users should be aware that adjacent loci are often not in sequential order. This may be due to reorientation of BACS, or if genes are added in an interval in which no sequential identifiers remain.

- Deleting genes
  - Deleted genes are kept in the database so they can be retrieved by searching for the identifier, but are marked "obsolete" and do not appear in database displays. Identifiers from deleted genes are not used again.
- Editing genes
  - Consensus in the AGI was that identifiers should be kept constant as long as there are no major changes in the gene model. As long as modifications in the gene model do not lead to a completely new protein (e.g. through use of a different reading frame), the identifier will be kept, even if exon boundaries change or individual exons are added/removed.
- Merging and splitting genes
  - Splitting Genes: When it is determined that a locus identifier actually refers to more than one gene (e.g. two genes were mistakenly predicted to be one gene), one of the genes will retain the original gene name and the second will get a new gene name. Rules for deciding which gene retains the original identifier is based on which gene contains the majority of sequence from the original locus.
  - Merging Genes: In the case where experimental evidence is found to indicate that two genes are actually a single locus (e.g. a full length cDNA is obtained) the two locus entries will be merged into one and the name that corresponds to the locus with the majority of sequence will be retained. The second locus identifier will be made obsolete (but kept associated to the locus identifier of the merged gene.

Tracking history

- Notes about splits and merges will be kept as well as the different versions of the locus sequence. Versions are identified by locus identifier, source, and date. For example AT2G18190 later becomes split into two entries AT2G18190 and AT2G18193 with a note that indicates that the second entry resulted from a split of AT2G18190. You can search TAIR for the annotation Locus Histories and download lists of locus names that are obsolete or in use.
- What terms in history tracking refer to:
  - delete means a gene model has been eliminated
  - merge means a gene model has been merged with another gene but retained old name
  - mergedelete means a gene model has been merged but its name has not been retained
  - insert means a gene model has been inserted from scratch
  - split means a gene model has been split but has retained its name
  - splitinsert means a gene model has been split and has a new name
  - new means a gene model has been generated
  - obsoleted means a gene model has disappeared
- The terms new and obsoleted may describe MIPS data when it is unknown if an insert or delete was due to a splitinsert or mergedelete.

Other Notes:
Generally, the idea is to be as conservative as possible. The identifiers should identify a specific chromosome locus, not a particular protein, and even if this identifier is used in an old publication, it should still direct a user to the current annotation for that locus, so that he will be able to see that the protein sequence has changed in the meantime. This is preferable to having a new identifier after modifications, where the user will first have to look up what is the current annotation for this locus. Keeping backwards-compatible versions of all entries cannot be achieved, and identifiers should not be a way of "versioning" genes.

Important notes from MIPS

Most people assume that if they sort the identifiers by ascending numbers they get a list of genes that represents the order along the chromosome. This was true originally, but no longer: Some BACS needed to be flipped, i.e. their orientation reversed, as new data on overlaps was generated. So all genes on these BACS now number the wrong way round. At MIPS, we decided it is more important to conserve the identifiers than the order, as the order can also be sorted by coordinates. Generally, the identifier still gives a good idea of the location on the chromosomes, only local reversals are expected. If you need a list of identifiers in the order along the chromosome, contact us. Once the orientation of BACS seems stable, this may be corrected by assigning new identifiers to the affected genes, as this will be more intuitive for users (This would be a breach of our "be conservative" rule, but the "be user-friendly" rule is more important).

Original document on creating AGI gene codes.

A uniform gene nomenclature system for Arabidopsis was discussed at an impromptu meeting at GSAC in Miami attended by Daphne Preuss, Chris Somerville, Claire Fraser, Xiaoying Lin and Mike Bevan on Sept. 18th.

It was decided that the following uniform system will be used in the forthcoming publication of the sequence of chr 2 and chr 4. A rapid decision was needed due to the time needed to implement the new names.

AT =organism 1,2,3,4,5 =chromosome G =gene 00010 =gene id

The 'G' convention is useful as repeats (r) will soon be annotated, initially as markers. Pseudogenes will be numbered like functional genes.
Gene are numbered in order from the top to bottom of the chromosomes. In the case of chr 2 and 4 this boundary is known due to the presence of rDNA clusters. Gene AT4G00010 is the first gene south of the cluster. Gene order is defined in units of 10 ie. 00010, 00020, 00030, etc allowing 9000 genes per chromosome.
If new genes are found between two annotated genes, either by experiment or improved gene finding programs, these will be numbered as: 00010, 00012,3,4,-9. This give plenty of room for expansion.
Different versions of a gene product, eg a differentially spliced gene , are denoted as 00010.1,2,3 etc.
Where there are sequence gaps, often of uncertain size and content (eg CEN2 and CEN4), the sequence groups will leave a space the equivalent of 100 - 200 genes. Where the top arm telomeres have not yet been reached, a gap equivalent to about 50 genes should be left, ie numbering will start 05000, 05010, etc.
The numbering of repeats will follow an independent system, where repeat ids are not interpolated between gene identities.
Please don't worry that the BAC naming conventions will be lost or erased from the records. We realize these are presently the most commonly used names, therefore the databases will have a simple way of relating the two naming conventions. Note that a single "AT4G00650" gene can have two BAC names, due to overlaps, and this is one of the reasons for implementing the new nomenclature. You will be able to search for an individual gene with this new name.
We believe this system conforms to that used in other organisms, and will be very useful to the community.

Gene Nomenclature

This section addresses how gene names/symbols are assigned and also illustrates some of the problems associated with gene nomenclature. We suggest guidelines for naming genes to avoid propagation of duplicated or misleading names.

Gene names should convey some meaning as to the function of the gene product. Names based upon a quantifiable feature such as biochemical assay, protein-protein or genetic interaction, or mutant phenotype are preferred to names based upon sequence similarity alone. Regardless of the derivation, it is not likely that a name can convey all that is known about a particular gene and names have changed to reflect new knowledge. Many of the problems associated with gene nomenclature may eventually be resolved by adoption of a standard rule for gene nomenclature that is accepted by the research community. At this time, at TAIR, we are not adopting a standardized system of gene nomenclature. We will concentrate our efforts on making associations of gene names and aliases so that information relating to each gene can be obtained regardless of the variable nomenclature. One of our goals is to relate all the diverse information about an object, such as a genetic locus, so that it can be ! accessed by researchers using a simple query. Furthermore, the data and annotations are ascribed to their sources so that the experimental evidence can be evaluated.

Problems with gene nomenclature

A major source of problems occurs when more than one published name is associated with the same gene or when the same gene symbol is assigned to more than one gene. An example of the former is *EMB30* which is also known as *GNOM*,and of the latter is the symbol *FDH* which has been used for both *FORMATE DEHYDROGENASE* and *FIDDLEHEAD*.These problems have been addressed, in part, by the establishment of a gene name registry for genes identified by mutation (Meinke and Koornneef, 1997). As there are many cases where same gene has been published under many names, TAIR maintains a list of aliases associated with each gene (see below). See section:Choosing a unique gene symbol.

Another problem that occurs is the tendency for error propagation with names based upon sequence similarity alone. For example, a gene is named *YFG2* based upon sequence similarity to *YFG1*, gene *YFG3* is then named based on similarity to *YFG2* and *YFG4* is named based upon similarity to *YFG3*. *YFG3* and *YFG4* may be quite distantly related to *YFG1* so in this case, the relationship inferred by the name is misleading. Gene names that imply functional or process equivalence should be avoided unless that function is experimentally verified.If a name is assigned based upon sequence similarity alone, the appropriateness of a name should be carefully considered. It is better to use the name *NITRATE REDUCTASE-LIKE*, than to use the name *NITRATE REDUCTASE5* if there is no experimental evidence to support the biological function the name implies.

Caution should be exercised in assigning names to ORFs 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. Unless sequence identity indicates an ORF corresponds to a known gene, or there is other experimental data for gene function, it is better to publish using the standard ORF name(s) along with a description (for example: At1g02730, cellulose synthase catalytic subunit, putative). Where possible, TAIR is making associations between sequences represented in commonly accessed arrays (e.g. AFGC arrays and Affymetrix chips) and their corresponding gene entries in TAIR.

Gene names should not be assigned to ORFs whose expression has not been proven (i.e. hypothetical proteins). At a minimum, full-length cDNA sequence should be obtained and analyzed to confirm its expression and gene structure. An alternative strategy might be to perform RT-PCR analysis of expression and sequence the RT-PCR product in cases where a cDNA sequence proves elusive (i.e. for genes expressed at very low levels).

Naming genes based upon mutant phenotype

Please refer to Meinke and Koornneef, 1997 for a discussion and examples of naming genes based upon mutant phenotype. This manuscript provides instructions for developing mutant gene names/symbols, proper nomenclature for publication and community standards for genetic analysis of mutant phenotypes. Mutant gene names are generally based upon one or more aspects of the mutant phenotype (e.g.*NON-PHOTOTROPIC HYPOCOTYL1)* or a genetic interaction such as *SUPPRESSOR OF PHYA-105*. Gene symbols are three letters and may or may not derive from the full name (e.g. *NON-PHOTOTROPIC HYPOCOTYL1*; *NPH1* or *ENHANCER OF AGAMOUS*; *HUA*). For publications and presentations, mutant gene names and symbols are lowercase and italicized and wild type alleles are uppercase and italicized. Protein products of genes are uppercase and not italicized. To help alleviate the problems associated with duplication of gene names, a mutant gene name registry has been created. Names and symbols for mutant genes should be registered with the curator of mutant gene names (Dr. David Meinke) along with map location and a description of the mutant phenotype (http://mutant.lse.okstate.edu/genepage/genepage.html).

Naming genes based upon function

Ideally gene names based upon function should indicate something about the quantifiable feature of the gene product. For example, *FATTY ACID DESATURASE* is a molecular function as determined by a biochemical assay or *PHYTOCHROME INTERACTING FACTOR* can be shown by protein-protein interaction. Enzyme names should be the standard name as defined by the Enzyme Commission. Names that correspond to a function should be reserved for those genes that have been experimentally proven to have that biological activity.

Gene name and symbol selection:

Meinke and Koorneeff suggest a 3-letter code for gene symbols whereas the Commission on Plant Gene Nomenclature (CPGN) allows for up to 8 characters (although currently only up to 5 characters are used). Although a number of classical genetic loci were assigned 2-letter symbols years ago, the continued use of 2-letter symbols to name new loci is strongly discouraged except in cases where there is a compelling reason based on the underlying science. A similar justification should be provided for the use of gene class symbols with more than 3 letters. This number should be more than enough coverage for all ~25,000 genes, provide a greater range of letter combinations for creating mnemonic symbols, and will not require renaming of CPGN designated gene names. Meinke and Koorneeff also describe numbering conventions to be used for genes and alleles. However, if a published gene name does not include a numeric suffix, it is inferred to be one (e.g. *sqt* = *sqt1*).The use of organism specific prefixes such as *At* or *Ath* is discouraged as this is redundant and leads to a lot of genes named *'Arabidopsis thaliana X'*. Relationships between genes and the organisms they are derived from are easily maintained within databases and do not need to! be reflected in the name. Organism specific prefixes that are appended for clarity in publications should not be part of the gene name.

Choosing a unique gene symbol

Before selecting a gene name/symbol check for name/symbol on the Mutant Gene Symbol list or use Arabidopsis GeneHunter. The Gene Hunter program is a text based searching tool that scans TAIR, the Mutant Gene Name Registry, GenBank, PubMed, Swiss-Pro, PIR, MIPS, AGR, Mendel-CPGN and the journals, Plant Cell and Plant Physiology for the input string (e.g. gene name or symbol) and, where appropriate, the term Arabidopsis thaliana. Do not use names or symbols for Arabidopsis genes that are already in use by other researchers.

Gene names and aliases in TAIR

In TAIR, a gene is referred to by the symbol (gene name) or the full name (e.g. *DYP*=gene name;*DYAD POLLEN*=gene full name). Many genes are associated with more than one name. For example,*FRUITFULL* is also known as *AGAMOUS-LIKE8 (AGL8)*. In TAIR, a gene name is determined by precedence of publication (see below) and all other names are maintained as aliases. As with the names, aliases are searchable within the database. In this way, TAIR maintains the connection between different publications that have referred to the same gene under various names. Therefore, a search for *AGL8* would retrieve the record for *FRUITFULL* and *AGL10* will bring up the record for *CAULIFLOWER*. We encourage the community to contact us when aliases are missing from the database. Authors should use the standard gene name in publications.

Making associations to BAC-based and chromosome-based names

Many groups and individual researchers are in the process of identifying cDNAs that may correspond to gene models predicted by the AGI. A predicted gene model might be verified by the complete cDNA sequence, or be improved and re-annotated based upon the cloned cDNA sequence. Sequence matching of cDNA sequences with the predicted ORFs/genome sequences will also identify 'missed ORFs' (i.e. ones that were not predicted by automated methods) and identify hypothetical proteins as being real (i.e.they are expressed). While TAIR will primarily use BLAST to make the associations between experimentally verified cDNAs and ORFs, it would be helpful if such information were included in the definition lines for the cDNA sequences submitted to Genbank. For example: Arabidopsis thaliana clone R09083 (FL5-10-H3)unknown protein (F14H20.12/AT2G02050) mRNA, complete cds, includes the BAC-based and chromosome-based names associated with this cDNA. In the event that a cDNA corresponds to mo! re than one predicted ORF, both standard ORF names should be included in the definition line.

Priority of gene names in TAIR

Precedence of publication is the primary determinant of a gene name unless the community chooses otherwise. If sequence (i.e. sequence identity) or genetic (i.e. allelism) analysis confirms that more than one name has been associated with a gene, the individual researchers should contact each other and agree upon a mutually acceptable name. Names of published mutants will have precedence over cloned gene names (based upon prior publication) unless the alternative (cloned) name is preferred by the community.

Suggestions for Genetic Markers, Polymorphisms and transgenic strain nomenclature:

Suggestions for Genetic Markers and Polymorphisms

The tremendous variability in naming markers creates problems for both research community and database curators. In many cases more than one type of marker has the same name as other objects in the database. For example, CTR1 is a gene, a CAPS and an SSLP marker. Ideally, a marker name should be; 1) unique, and 2) contain information about its type (e.g. SSLP, CAPS). We would like to propose a system of nomenclature that uses a lab organization specific prefix to designate marker origin followed by a simple accession number. This system is already in use for naming SNP polymorphisms (e.g. SGCSNP1-9299 = Stanford Genome Center SNP 1-9299, and the CER SNPs). An organization may be a lab, university, or a company. A three letter prefix should be more than enough to provide a unique code (>17000 possible combinations) for every lab. For example, markers developed by researchers at the Carnegie Institution of Washington (Stanford, CA) have the prefix CIW followed by some unique integer (e.g. CIWCAPS12, CIWSSLP34, etc.). Potential new marker names should be searched for in TAIR to avoid duplication of names.

Naming transgenic lines

Nomenclature for transgenic strains is also quite variable and not always informative. It would be helpful if the name would convey some information about the lines. We propose a system of prefixes and accessions similar to that described for markers except that an additional prefix would be added to indicate the type of transgene. In addition to the lab designation, each transgenic strain name would have a symbol indicating the type of construct.

- TG: Transgenes such as co-suppression constructs, promoter:GUS/GFP construct.
- ET: Enhancer trap.
- GT: Gene trap.
- AT: Activation tag.
- TP: Transposon tag (e.g. Ac/Ds, Spm/dSpm insertions).
- TD: T-DNA insertion.
- PT: Promoter trap.

Each unique transgenic germplasm would be given a new accession (unique integer). For example, the insertion lines from the IMA are designated as SET# (enhancer trap) or SGT# (gene trap). If it turns out that two different names are associated with the same insertion event, appropriate aliases will be made in the database.

Names used for large sets of SNPs and insertion/deletions

The following table lists prefixes used by functional genomics projects for naming T-DNA insertions. This information can be used to search for all of the . For example, to find all of the deletions identified by the Stanford Genome Sequencing Center you can search by Polymorphism name starts with SGC and type is deletion.

Prefix	Source	Comment
SGC	Stanford Genome Center	Includes insertions, deletions and single nucleotide polymorphisms
CER	Cereon Genomics	Includes insertions, deletions and single nucleotide polymorphisms. Available only to registered users from non-profit and academic institutions.

Names used for large sets of T-DNA or transposon insertions

The following table lists prefixes used by functional genomics projects for naming T-DNA insertions. This information can be used to search for all of the insertion lines generated by a project. For example, to find all of the T-DNA insertion lines generated by Joe Ecker's group at the SALK institute you can search by Polymorphism name starts with SALK.

Prefix	Source	Comment
SALK	Joe Ecker et.al.	Sequence indexed library of insertion mutations generated using the pROK2 T-DNA vector.
SGT	V.Sundareson et.al.	Gene trap lines from the Institute for Molecular Agrobiology (IMA)
SET	V.Sundareson et.al.	Enhancer trap lines from Institute for Molecular Agrobiology (IMA)

Clone and Vector Names

Arabidopsis clones are usually named with the acronym of the vector followed by the plate and row numbers of the isolated clone. For example, CIC (YAC), T (TAMU BAC), F (IGF BAC) are some common vector acronyms.The following table gives information about nomenclature for clones and vectors in TAIR. You can use the prefix in a wild card search for all clones from a particular source or vector. For example, you can use the DNA search to find all TAMU clones by choosing clone name [starts with] T.

Vector type	Clone Prefix	Vector Name	Source	Description
BAC	T	pBeLoBAC11	TAMU (Texas A&M University)	from bacterial artificial chromosome library used for genomic sequencing
BAC	F	pBELoBACkan	IGF (Institut fur Genbiologische)	from bacterial artificial chromosome library used for genomic sequencing
P1	M	pAd10sacBII	Mitsui et.al.	from Bacteriophage P1 library used for genomic sequencing
TAC	K	pTAC-YL7	Kazuza	transformation-competent bacterial artificial chromosome vector
YAC	CIC	pYAC4	CEPH/INRA/CNRS	From yeast artificial chromosome libary
YAC	EG	pYAC41	Grill and Somerville	from EG1 yeast artificial chromosome library
YAC	EW	pYAC3	E. Ward et.al.	from yeast artificial chromosome library
YAC	yUP	pYAC4	Joe Ecker et.al.	yeast artificial chromosome library
Cosmid	G		Howard Goodman et.al.	From cosmid library prepared by H.Goodman et.al.

GenBank Accessions

Certain objects such as genes, clones, clone ends and some insertions in TAIRs database can be accessed by searching with the associated Genbank accession number. Each accession number in GenBank is unique. See http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#AccessionB for information about GenBank accession numbers.