How to name and symbolize genes at previously unnamed loci.

David D. Perkins

Names of genes and gene loci have traditionally been based on a word that describes the mutant or variant phenotype. With polymorphic genes, the name may be based on function (e.g., *het*: *heterokaryon incompatibility*). Abbreviations (usually 3-letter) of the name have been used as gene symbols. In case of conflict, priority is given to the name that was published first after the gene was shown to be at a new locus. Subsequent names are considered inactive synonyms. (See Perkins *et al.* 2001 and Radford's on-line gene list http://www.bioinf.leeds.ac.uk/~gen6ar/newgenelist/genes/gene_list.htm for established gene names and symbols. See Table 2 in Perkins *et al.* 2001 for a list of synonymous gene symbols). It is still desirable to base names on mutant phenotypes, but now this is frequently not feasible because gene identification may be based on DNA sequence, long before the phenotype of a null mutant becomes known. In this situation, a provisional name may be given that is based on homology to sequenced genes of known function in the same organism (paralogous) or in other organisms (orthologous). The following statement is from Perkins (1999), which should be consulted for references and further details on nomenclature.

> *Names.* Neurospora resembles Drosophila in having a relatively well-defined wild-type phenotype. In the formative years with both organisms, existence of a gene was recognized when a mutation occurred that deviated from the wild type. The gene was then named using a word that described the mutant phenotype. Gene loci recognized on the basis of naturally occurring variants (e.g., mating type idiomorphs, vegetative incompatibility genes, isozyme markers) were named according to the phenotype affected. Descriptive gene names were given in preference to using numbers or nondescriptive names. They are informative, easier to remember, and less likely to result in confusion with other loci.
>
> In choosing what aspect of the phenotype to use as a basis for naming a mutant gene, preference was given to the most convenient and useful manifestation. For example, all arginine auxotrophs were named "arginine" rather than being given different names based on the earliest utilizable precursor (citrulline, ornithine, etc.) or on the enzyme that was rendered nonfunctional. A gene specifying the molybdenum cofactor that is shared by nitrate reductase and xanthine dehydrogenase was named "nitrate-8" rather than "molybdenum cofactor" or "xanthine dehydrogenase" because the mutant is scored as a nitrate nonutilizer. These considerations still hold.
>
> Gene names should be concise and informative. Each name must be unique and must not have been used previously for a Neurospora gene. Gene names or symbols should not be prefixed with the word *Neurospora* or the letters *n* or *nc* to indicate that a gene is from Neurospora. To do so would be redundant. Sequence-database identification-code entries, which often begin with NC or NEU, are not gene names or symbols, nor do they establish priority.
>
> Different loci bearing the same name and the same base symbol should be numbered sequentially beginning with one, e.g., *arg-1, arg-2, arg-3,* etc. If a name applies to only one locus, use of the number 1 is optional. For example, the gene that specifies invertase is symbolized *inv* rather than *inv-1*. Arbitrary strain-identification numbers should not be converted into locus numbers.

Regulatory genes have usually been given the same name and symbol as the structural genes they regulate (e.g., *nit-2, leu-3, cys-3*), but this is not always true (e.g., *pcon, pgov, scon, ty*).

When new names, symbols, locus numbers, or allele-number prefixes are to be assigned, it is essential to avoid duplication by consulting the most recent FGSC stock list and the lists that accompany the current genetic maps.

Symbols. Symbols are preferably three-letter abbreviations of the gene name, but they may consist of two letters or (rarely) one or four. Symbols are written in lower case italics (e.g., *inv*) except when the name is based on a mutant allele that is dominant. The first letter is then capitalized (e.g., *Asm*). Nonallelic genes that have the same descriptive name and symbol are distinguished from one another by numbers that are separated from the base symbol by a hyphen (e.g., *al-1, al-2, al-3*). This use of hyphens in Neurospora and Drosophila differs sharply from the convention in many other organisms, where a locus number (or letter) is not separated from the base symbol. Hyphens are used only to separate the locus number from the base symbol to which it is appended.

When a gene name contains a number that is necessary for identifying the product or phenotype, the product-identifying number is included as an integral part of the base symbol, with digits unseparated from the letters by a hyphen (e.g., *tom22*; *nuo78*). A hyphen can then be used if it is needed to distinguish locus numbers from numerals belonging to the gene name (e.g., *hsp70-1, hsp70-2*).

Roman numerals should be avoided in gene symbols.

Suppressors are symbolized using the letters *su,* followed immediately by the symbol of the suppressed gene in parentheses. If nonallelic suppressors of the same gene are known, locus numbers follow the parentheses (e.g., *su(met-7)-1, su(met-7)-2*). As in Drosophila, su^{+} designates the wild-type gene, *su* the mutant suppressor allele. For allele-specific suppressors, the allele number is included as a superscript of the locus symbol (e.g., $su(trp\text{-}3^{td201})\text{-}2$). Enhancers are symbolized in a similar way (e.g., *en(am)-1*).

Chromosomal loci other than genes usually have the initial letter of the symbol capitalized (e.g., *Cen, Tel, In, T*). Also, the initial letter is usually capitalized in symbols for active or relic transposons (e.g., *Tad, Pogo*).

Dominance and recessiveness. When a gene is named for a mutant phenotype that is recessive to the wild type, the name and symbol are written in lower case letters (e.g., *al*: *albino*). The initial letter is capitalized when the mutant phenotype is dominant (e.g., *Ban*: *Banana*). The initial letter is not capitalized when a gene is named for alleles that show codominance (e.g., *het*: *heterokaryon incompatibility*).

Mutant phenotypes may be expressed either in the vegetative phase or in the sexual phase, or in both. Some mutant genes are known to be dominant in the sexual phase but recessive in vegetative tissues. The initial letter of the name and symbol is then capitalized if the gene name is based on the dominant mutant phenotype (e.g., *R, Asm*) but the initial letter is not capitalized if the name is based on the recessive mutant phenotype (e.g., *mei-3,* pk^{D}).

Dominance or recessiveness is usually not known at the time new vegetative-phase mutants are named. In the absence of that information, lower case symbols are routinely used because recessive loss-of-function mutations are the most common type to be detected phenotypically. Tests for dominance in the vegetative phase may employ either heterokaryons or heterozygous partial diploids. Partial diploids are preferred because they ensure a 1:1 allele ratio, whereas the ratio of nuclear types in heterokaryons may depart widely from equality.

Partial diploids are obtained as duplication progeny from crosses heterozygous for insertional or quasiterminal rearrangements.

Mutant genes that are recognized by their expression in the perithecia of heterozygous crosses are immediately known to be dominant (e.g, *R, Asm*). Recessive sexual-phase mutations are less likely to be detected because they must be present in both parents of a cross in order to be expressed. Many of the known sexual-phase recessives were recognized in crosses homozygous for mutant genes affecting mutagen sensitivity and DNA repair (e.g., *uvs, mus*). These had already been detected and named as recessive vegetative-phase mutants. Other recessive sexual-phase mutants have come from backcrosses in experiments specifically designed to detect them. Still others were discovered accidentally in crosses between inbred parents (e.g., *mei-1, mei-3*).

Gene loci recognized by DNA sequence. We need no longer depend on mutant differences. cDNA libraries and sequencing now make it possible to recognize genes for which no variant product or phenotype has been detected. These "anonymous" genes can be placed on the genetic map by using them as probes in RFLP mapping. In absence of a known mutant phenotype, gene names may be based on the time or site of expression (e.g. *con*). The null mutant of such a gene may (e.g., *asd-1*) or may not (e.g., *con-11*) reveal a conveniently recognizable mutant phenotype on which to base a descriptive name. If the null mutant is lethal (as with *tom19* and *tom22*, for example), or if it is phenotypically wild type, or if the mutant phenotype remains undetermined, it is appropriate and informative to base the name on sequence-homology with a gene or gene family the function of which is known in another organism (e.g., *ras*: *ras-like, pzl*: *phosphatase-z-like*). This should be done, however, only if the sequence makes a strong prediction of function. A Neurospora gene should not be named for the overt phenotype of its homolog in another organism if that phenotype is developmentally complex and far removed from the primary gene product. Manifestation of the genes may have diverged in the two organisms, resulting in quite different phenotypes. For example, mutations in homologous genes appear to be responsible for cerebrohepatorenal anomalies in humans and for failure of premeiotic nuclear fusion in the croziers of Podospora and Neurospora.

If neither phenotype nor homology is known, a gene may be given a generic symbol indicating anonymity. The symbol *anon* is used in Drosophila, with some distinguishing suffix, and this is recommended for Neurospora. An alternative that has been proposed is *eat* (*encodes anonymous transcript*). The meaning of *eat* is not obvious from the symbol, however. Generic names and symbols of this type, that represent a category of mutants rather than a specific mutant, have a long history of use in Neurospora. Best known is the use of *un* for temperature sensitive genes of unknown function. Other generic categories are *ccg* for clock-controlled genes, *con* for genes expressed during conidiation, and *sdv* for genes expressed under conditions favoring sexual development

When a mutant phenotype or a definitive sequence-homology is discovered for an anonymous mutant, the option exists of changing the name to something more definitive. For example, if the null allele of a gene initially called *anon (NP6C9)* were found to result in restricted colonial growth, the name could be changed to *col-x*.

Different *anon* genes are best distinguished using isolation numbers, as in the example, because if the genes were numbered serially, a clearing house would be needed to avoid using the same number repetitively.

References

Davis, R. H. 2000. *Neurospora: Contributions of a Model Organism.* Oxford University Press, New York.

Perkins, D.D. 1999. Neurospora genetic nomenclature. Fungal Genet. Newslett. 4 34-41. (Reprinted as Appendix 1 of Perkins *et al.* 2001. Also in Davis, R. H. 2000.

Perkins, D. D., A. Radford, and M. S. Sachs 2001. *The Neurospora Compendium: Chromosomal Loci.* Academic Press, San Diego.

Radford, A. 2006 and ongoing. The *Neurospora crassa* "gene list" e-Compendium. http://www.bioinf.leeds.ac.uk/~gen6ar/newgenelist/genes/gene_list.htm.

DDP