Recommendations for assigning symbols and names to *Neurospora crassa* genes now that its genome has been sequenced.

Heather M. Hood[1], Alan Radford[1, 2, 3] and Matthew S. Sachs[3]
Oregon Health & Science University[1], University of Leeds[2], and Texas A&M University[3]

Fungal Genetics Reports 55:29-31

Originally, *Neurospora crassa* genes were named for their mutant phenotypes or natural variant properties. Genes are now increasingly named on the basis of cross-species sequence similarity. These names may also be supported by predicted or experimentally identified molecular function. As a consequence, *N. crassa* gene nomenclature in practice is frequently no longer adequately covered by the established conventions (Perkins *et al.* 2001). Here we provide additional nomenclature guidelines relevant to these new circumstances, and some general guidelines for providing information on the identity of *N. crassa* genes in scientific communications.

The strictest adherence to previously accepted *N. crassa* naming convention would be that genes not receive a symbol and name (*e.g.*, *cot-1* and *colonial temperature sensitive-1*) until a mutant phenotype is described or a function is demonstrated. However, it is not realistic to expect, for example, that the thousands of *N. crassa* genes that have NCU numbers and orthologs in other species be referred to only by their NCU numbers until such time as *N. crassa* experimental data provide the basis for a name. Therefore, we consider how to provide names and symbols for previously undescribed *N. crassa* genes that reflect the emerging practice arising from *in silico* identification. These recommendations do not apply to established genes.

Consider first the naming of a new *N. crassa* gene when its ortholog in another species is known, and when there are experimental data supporting its function in *N. crassa* that are consistent with the function of the known ortholog. As an example, consider the recently identified *N. crassa* gene *spo11*, whose symbol mirrors that of its *Saccharomyces cerevisiae* ortholog, and in which both systems shows a mutant phenotype affecting meiosis (Bowring *et al.* 2006). In this instance, the gene could have been assigned a variety of symbols, but the investigators chose one that indicates its relationship to a well-studied ortholog.

Is it more appropriate to symbolize the gene as *spo-11* or *spo11*? Historically, use of the hyphen indicates that the gene represents one of a series of genetically defined complementation groups named successively from *-1* to *-n*, and in this case would imply that at least ten other *spo* complementation groups had been identified and named *spo-* in *N. crassa*, which is not the case. On the other hand, until recently, few *N. crassa* genes have been named with numbers immediately following letters (*i.e.*, without a hyphen separating them). In the majority of these cases, the numbers represented an estimation of the gene product's molecular mass, [*e.g., hsp70* (Kapoor *et al.* 1995)], following the naming convention (Perkins *et al.* 2001) which stated, "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, unseparated from the letters by a hyphen." We recommend that for situations in which the number is part of the symbol for the ortholog in another organism that there be no hyphen, because the number serves to identify the gene product (*e.g.*, this gene product is the ortholog of a *spo11* gene). Thus, the symbol *spo11* is appropriate. As to the gene name (which in *N. crassa* convention is distinct from the gene symbol), we recommend that, when the original name describes a phenotype, the original name be also explicitly stated, followed by "-*like*", so that *N. crassa spo11* is named "*sporulation 11-like*", and not "*sporulation 11*", or "*spo11-like*".

In a second scenario, there is sequence similarity between the *N. crassa* gene and a gene in another species, and also a mutant phenotype in *N. crassa*, but that phenotype is not particularly informative. This is covered adequately by the 2001 guidelines, which state "If the null mutant is lethal, 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…" There seems no reason to change that guidance.

Scenario three applies to naming an *N. crassa* gene for its ortholog in another species without experimental evidence of its function in *N. crassa*. When orthologs are identified by sequence similarity, they should represent the best bi-directional BLAST hits between *N. crassa* and the other species. Orthologous naming has certain advantages. It provides immediate evidence for evolutionary conservation of sequence and makes it simpler for researchers to consider the potential function of the gene in *N. crassa*. Yet, naming an ortholog by the name given in another organism is potentially problematic because it

may not have similar function(s) in *N. crassa*. With this caveat, what is the best criterion for naming when the only available information is a cross-species sequence similarity? *N. crassa* genes have been named after genes in other species for a variety of reasons. Based on recent literature, we observe that investigators are often choosing the *S. cerevisiae* name, even if the *Schizosaccharomyces pombe* homolog is closer, because the *S. cerevisiae* gene generally has the most experimental evidence associated with it and is known more widely. Choosing a symbol based on a well-known ortholog (which may not necessarily be *S. cerevisiae*) should be the primary criterion if that symbol has not been previously used (pre-empted for use) in *N. crassa*. If the mutant phenotype proves subsequently not to be that anticipated by orthology, then the name and symbol can be revised to more accurately reflect its experimentally determined nature.

An example of a difficulty that can arise by using the symbol for a well-known ortholog is *cdc*, which is widely used across diverse species for genes affecting the eukaryotic cell-division cycle. *N. crassa* encodes many genes with similarity to *cdc* genes in other species. By using the symbol *cdc* for these *N. crassa* genes based on these similarities, their existence in *N. crassa* becomes more universally easy to appreciate. Yet, *N. crassa* lacks a classic cell division cycle, and thus the symbol and name are technically incorrect. The established convention states, "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." On this basis, naming such an *N. crassa* gene *cell division cycle* with the symbol *cdc* would be inappropriate. However, there are perceived advantages to naming genes on the basis of orthology, and therefore in this case, *cdc* orthologs could be named *cell division cycle-like* with the symbol *cdc*. For example, the *N. crassa* ortholog of *S. cerevisiae CDC13* could be symbolized *cdc13* and named *cell division cycle 13-like.*

Scenario four pertains to naming members of gene families where sequence similarity suggests potential function, but making specific one-to-one correspondences with genes from other species is problematic (*e.g.*, cytochrome p450 and glycosyl hydrolase families). The symbol used can then represent the general function of the family (*e.g.*, *cyp450* for cytochrome p450). Such a symbol should be derived from some widely recognized property of the family, but need not be a previously used symbol in another organism. A hyphen is called for to demarcate different members of such a family, such as *cyp450-1* (*cytochrome p450 family -1*) or *gh61-1* (*glycosyl hydrolase 61 family -1*).

Recommendations

New genes identified by orthology and named after the ortholog should have symbols without hyphens, written in lower case italics [*e. g.*, *msh4*; (Conway *et al.* 2006)].

Except where the import of a symbol for an ortholog is not possible because that name/symbol has been pre-empted in *N. crassa* by previous use for another gene and function, names should be based on the function in the best-established system of where the ortholog is studied.

A gene named strictly by *in silico* applications should not be given the same priority in publication as genes named by experimental analyses of functions. If the phenotype and/or function proves different from originally anticipated based on sequence similarity analyses, then the gene may be re-named to reflect its identified phenotype/function.

If there are no experimental studies on the gene in other species to determine its phenotype and/or function, then the gene should not be named based on orthology alone.

Use of complex formatting (*e.g*. subscripts, superscripts, and greek letters) in gene symbols and names should be avoided as this causes problems in electronic database storage and searching.

The standard gene names in the Broad Institute database (http://www.broad.mit.edu/annotation/genome/neurospora/Home.html), the *Neurospora* Compendium (http://www.bioinf.leeds.ac.uk/~gen6ar/newgenelist/genes/gene_list.htm), and the current NCU identifiers should be given in any published work on sequenced genes. This point is important to consider by scientists communicating their results, by reviewers and by editors.

Existing practice for designating an *N. crassa* polypeptide product using the same characters as the corresponding gene symbol calls for it to be written in all roman (no italics) upper case letters. To avoid possible ambiguities, we recommend that if there is a hyphen in the gene symbol, that it be retained in the gene-product designation (for further discussion of peptide nomenclature, see Dunlap *et al.* 1996).

The authors, who serve as curators for the *N. crassa* Community Annotation Project, can assist confidentially with specific questions about gene symbols and names.

Acknowledgements:

HMH and MSS were supported by the NIH program project grant "Functional analysis of a model filamentous fungus" (GM068087) and AR by a Leverhulme Emeritus Fellowship.

References:

Bowring, F. J., P. J. Yeadon, R. G. Stainer and D. E. Catcheside, 2006. Chromosome pairing and meiotic recombination in *Neurospora crassa spo11* mutants. Curr. Genet. 50**:** 115-123.

Conway, S., F. J. Bowring, J. Yeadon and D. Catcheside, 2006. *Neurospora msh4* ortholog confirmed by split-marker deletion. Fungal Genet. Newsl. 53**:** 5-8.

Dunlap, J. C., M. S. Sachs and J. Loros, 1996. A recommendation for naming proteins in Neurospora. Fungal Genet. Newsl. 43**:** 72.

Kapoor, M., C. A. Curle and C. Runham, 1995. The *hsp70* gene family of *Neurospora crassa*: cloning, sequence analysis, expression, and genetic mapping of the major stress-inducible member. J. Bacteriol. 177**:** 212-221.

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