Caenorhabditis nomenclature*

Mary Ann Tuli[1,§], Aric Daul[2], Tim Schedl[3,§]

[1]Biology Division, California Institute of Technology, Pasadena CA 91125, USA

[2]Caenorhabditis Genetics Center, University of Minnesota, Minneapolis MN 55455, USA

[3]Department of Genetics, Washington University School of Medicine, St. Louis MO 63110, USA

Table of Contents

* Edited by Paul Sternberg. Last revised November 25, 2015. Published in its final form August 8, 2018. This chapter should be cited as: Tuli M. A., Daul A., Schedl T. *Caenorhabditis* nomenclature. (August 8, 2018), *WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.183.1, http://www.wormbook.org.

§To whom correspondence should be addressed. E-mail: maryann.tuli@wormbase.org or ts@genetics.wustl.edu

Abstract

Genetic nomenclature for *Caenorhabditis* species and other nematodes is supervised by WormBase in collaboration with the *Caenorhabditis* Genetics Center (CGC) and with essential input from the community of scientists working on *C. elegans* and other nematodes.

1. Introduction

Genetic nomenclature allows the genetic features of an organism to be structured and described in a uniform and systematic way. Genetic features, including genes, variations (both natural and induced), and gene products, are assigned descriptors that inform on the nature of the feature. These nomenclature designations facilitate communication among researchers (in publications, presentations, and databases) to advance understanding of the biology of the genetic feature and the experimental utilization of organisms that contain the genetic feature.

The nomenclature system that is used for *C. elegans* was first employed by Sydney Brenner (1974) in his landmark description of the genetics of this model organism, and then substantially extended and modified in Horvitz et al., 1979. The gene, allele, and chromosome rearrangement nomenclature, described below, is an amalgamation of that from bacteria and yeast, with the rearrangement types from *Drosophila*. The nomenclature avoids standard words, subscripts, superscripts, and Greek letters and includes a hyphen (-) to separate the gene name from gene number (distinct genes with similar phenotypes or molecular properties). As described by Jonathan Hodgkin, 'the hyphen is about 1 mm in length in printed text and therefore symbolizes the 1 mm long worm'. These nomenclature properties make *C. elegans* publications highly suitable for informatic text mining, as there is minimal ambiguity. From the founding of the Caenorhabditis Genetics Center (CGC) in 1979 until 1992, Don Riddle and Mark Edgley acted as the central repository for genetic nomenclature. Jonathan Hodgkin was nomenclature czar from 1992 through 2013; this was a pivotal period with the elucidation of the genome sequence of *C. elegans*, and later that of related nematodes, and the inception of WormBase. Thus, under the guidance of Hodgkin, the nomenclature system became a central feature of WormBase and the number and types of genetic features significantly expanded. The nomenclature system remains dynamic, with recent additions including guidelines related to genome engineering, and continued reliance on the community for input.

WormBase assigns specific identifying codes to each laboratory engaged in dedicated long-term genetic research on *C. elegans*. Each laboratory is assigned a laboratory/strain code for naming strains, and an allele code for naming genetic variation (e.g., mutations) and transgenes. These designations are assigned to the laboratory head/PI who is charged with supervising their organization in laboratory databases and their associated biological reagents that are described on WormBase, in publications, and distributed to the scientific community on request. The laboratory/strain code is used: a) to identify the originator of community-supplied information on WormBase, which, in addition to attribution, facilitates communication between the community/curators and the originator if an issue related to the information should arise at a later date; and b) to provide a tracking code for activities at the CGC. The laboratory/strain designation consists of 2-3 uppercase letters while the allele designation has 1-3 lowercase letters. The final letter of a laboratory code should not be an "O" or an "I" so as not to be mistaken for the numbers "0" or "1" respectively. Additionally, allele designations should also not end with the letter "l" which could also be mistaken for the number "1." These codes are listed at the CGC and in WormBase. Investigators generating strains, alleles, transgenes, and/or defining genes require these designations and should apply for them at genenames@wormbase.org.

Information for several other nematode species, in addition to *C. elegans*, is curated at WormBase. All species are referred to by their Linnean binomial names (e.g,. *Caenorhabditis elegans* or *C. elegans*). Details of all the genomes available at WormBase and the degree of their curation can be found at www.wormbase.org/species/all

2. Genes

Three different gene naming schemes are employed in WormBase: GeneIDs, Sequence Names, and Gene Names.

First, for all genes WormBase has assigned a unique accession of the form 'WBGene#', for example 'WBGene00000024'. The accession uniquely refers to a locus at a specific location in the genome: either a protein coding sequence (CDS) and transcript structure, an uncloned gene, a non-coding RNA transcript, a defined sequence element, or pseudogene. This accession follows the gene through any changes that may be made to it based on bioinformatics predictions or experimentation. If the gene object is split into two genes, the original accession will usually apply to the 5' gene and a new accession will be assigned to the other half.

Second, for nearly all *C. elegans* genes WormBase has assigned a Sequence Name, which is derived from the cosmid, fosmid, or YAC clone on which they reside, and which was used to generate the reference Bristol N2 genome sequence (The *C. elegans* Sequencing Consortium, 1998). For example 'AC3.3', indicates that the gene is on cosmid AC3, and that there are at least 2 other genes (.1 and .2) on that cosmid.

Finally, for a number of loci, a standard genetic gene name is employed, which usually refers to mutant phenotype or some aspect of sequence similarity. Researchers who are investigating such genes are encouraged to propose gene names based on the guidelines in Section 2.1 and Section 2.2. Genetic gene names consist of a three or four letter gene class name, e.g., "*abu*" (for **A**ctivated in **B**locked **U**nfolded protein response), followed by a hyphen and a number: *"abu-1"*. A small number of gene objects defined by mutation are currently uncloned, and until associated with a sequence name will appear with just a standard genetic name. Publications commonly use the genetic name, as this designation provides information about function in biological processes, potential molecular function based on sequence similarity, or in some cases known biochemical activity. Sequence names for genes lacking a genetic name, of which there are many, are also used in publications, often with -omic level studies where functional information may be limited.

2.1. Applying for new gene class and/or gene names

Investigators wishing to propose new gene names for *C. elegans* should note the summary guidelines below and apply online via WormBase or by email application to genenames@wormbase.org. It is highly recommended that investigators contact WormBase with their gene name proposal prior to submission of a manuscript or abstract, as proposed gene names are not always used as the main name on WormBase. Reasons for not using a proposed name as the main name include the gene already having a name (usually being held in confidence; see Section 2.2.8.) or that the name is inappropriate based on incorrect sequence analysis or other type of ambiguity.

2.2. Guidelines for proposing new gene names

2.2.1.

Gene names must conform to the standard format of 3 or 4 letters gene class name, hyphen, number.

2.2.2.

Genes can be named on the basis of a mutant phenotype or on the basis of the predicted protein product or RNA product.

2.2.3.

For genes defined by mutation, the approved gene names refer to the mutant phenotype originally detected or most easily scored e.g., dumpy (*d*um*py*) in the case of *dpy-5*, lethal (*let*hal) in the case of *let-37*.

2.2.4.

For genes defined on the basis of sequence similarity or sequence features, the approved gene name refers to the predicted protein product, RNA product, or gene causing a disease, e.g., *my*osin *l*ight *c*hain in the case of *mlc-3*, super*o*xide *d*ismutase in the case of *sod-1*, or *neph*rono*p*hthisis (human kidney disease) homolog, in the case of *nphp-4*.

2.2.5.

Gene names based on homology with a previously named gene in another well-studied organism (e.g., *Homo sapiens*) or that is the standard in the relevant field (e.g., DNA replication) are often appropriate and desirable,

especially where there is convincing orthology between genes (e.g., *wrn-1* is the *C. elegans* ortholog of the human gene WRN1, responsible for Werner's syndrome, or *rnt-1* is the *C. elegans* ortholog of the *Drosophila* gene *runt*).

2.2.6.

If a new gene clearly belongs in an existing gene class, then a new gene number will be assigned after consultation with the laboratory responsible for the gene class in question.

2.2.7.

If the establishment of a new gene class name seems more appropriate, then approval for this gene name must be obtained, preferably online via WormBase or by email application to genenames@wormbase.org. To avoid the situation of multiple names for the same gene in the literature, approval should be sought prior to abstract or manuscript submission.

2.2.8.

New gene name classes can be assigned in confidence, prior to formal publication or disclosure in an abstract.

2.2.9.

Gene names that are memorable, informative, and simply explained are encouraged.

2.2.10.

Gene names and gene numbering schemes that conform to established nomenclature proposals for particular protein classes are desirable.

2.2.11.

Gene names including *c* (for *Caenorhabditis*), *ce* (for *C. elegans*), *n* (for nematode) or *w* (for worm) are discouraged. *C. elegans* as the organism of origin can be specified with a prefix (*Cel-*) if desired, e.g., *Cel-fem-1*.

2.2.12.

Gene names based solely on RNAi phenotypes or high-throughput analysis of gene expression or protein interactions are discouraged.

2.2.13.

A limited number of genes have been given temporary *tag-* names (*tag* = *t*emporarily *a*ssigned *g*ene name). These are genes for which deletion alleles have been generated by reverse genetic methods, but which have not yet been given more informative names based on sequence or mutant phenotype. When sufficient information becomes available, each *tag* name will be replaced by an appropriate standard 3- or 4-letter name.

2.2.14.

If a gene produces a protein that can be classified as a member of a family, the gene may also be assigned an approved name consisting of three or four italicized letters, a hyphen, and an italicized Arabic number, e.g., *nhr-49*, indicating that this is the 49th member of the *nhr* (nuclear hormone receptor) gene family.

2.2.15.

There are a few exceptions to this format. Genes in a paralogous set related to a single named gene are sometimes given the same gene name and number, followed by a distinguishing decimal. For example, four *C. elegans* genes homologous to *SIR2* in *S. cerevisiae* have been given the names *sir-2.1*, *sir-2.2*, *sir-2.3*, *sir-2.4*.

2.2.16.

A limited number of genes, named on the basis of sequence homology, have been given non-standard names ending with alphanumeric identifiers rather than with simple numbers in order to make these names closer to the

generally accepted names used in other organisms. For example, *eif-3.B*, *eif-3.C*, *eif-3.D*, etc. encode proteins of the conserved translation factor eIF3.

2.2.17.

The gene name may, on rare occasions, be followed by an italicized Roman numeral to indicate the chromosome (linkage group) on which the gene resides, e.g., *dpy-5 I* (chromosome *I*) or *let-37 X* or *mlc-3 III*.

2.2.18.

Genes with related properties are often given the same 3 or 4 letter gene class name and different numbers. For example, there are three known myosin light chain genes: *mlc-1*, *mlc-2*, *mlc-3*, and more than twenty different dumpy genes: *dpy-1*, *dpy-2*, *dpy-3*, and so on.

2.2.19.

There is no specific nomenclature for pseudogenes. The Overview section on the WormBase gene page indicates if the gene is designated as a pseudogene.

2.3. Approved gene name conflicts

Approved gene names that have been established in databases and have been used in a body of published literature should preferably not be changed. In cases where a gene has received multiple names, one name will be adopted as the main name for the gene on WormBase. The other names will continue to be listed in WormBase, where relevant associations can be found through search features. Whenever possible, name changes or the adoption of a single main name should be made with the approval of all laboratories concerned.

3. Guidelines for variation data

Variation is defined as a change in genome sequence compared to the reference *C. elegans* Bristol N2 sequence. Variation can be experimentally induced (e.g., chemical mutagenesis, genome engineering) or found in *C. elegans* strains isolated from natural environments. The term allele is applied both to experimentally-induced change and natural variation whether the change produces a phenotype or not. The term mutation is often applied to experimentally-induced variation which can often result in phenotypic deviation. The term polymorphism is applied to natural variation, and often does not result in a phenotype. The Bristol N2 reference strain is defined by convention as wild type. However, during laboratory propagation Bristol N2 can accumulate variation resulting in differences from the reference Bristol N2 sequence. Therefore in cases where an investigator wants to associate phenotype with genotype, multiple lines of evidence are necessary to draw the conclusion that variation x leads to phenotype y. Every variation is assigned an accession of the form WBVariationDDDD. Some variations, for example mutant alleles and polymorphisms of certain types, are assigned an additional allele name.

(Natural variation is discussed below in Section 5. Polymorphisms and other genetic elements.)

3.1. Alleles/mutations from experimentally induced variation

3.1.1.

Every allele has a unique designation. Alleles are given names consisting of one to three italicized letters (the allele designation for the assigning laboratory) followed by an italicized Arabic number, e.g., *e61* or *mn138* or *st5*. The last letter should not be an "l" to avoid confusion with the number "1". For example, *e* refers (originally) to the MRC Laboratory of Molecular Biology (Cambridge, U.K.), (currently) to the laboratory of J. Hodgkin (University of Oxford), and *st* refers to the laboratory of R.H. Waterston (originally at Washington University, St. Louis, MO, currently at the University of Washington, Seattle).

3.1.2.

When gene and allele names are used together, the allele name is included in parentheses after the gene name, e.g., *dpy-5(e61)*, *let-37(mn138)*. When unambiguous (e.g., if only one allele is known for a given gene or if all work

on a gene described in a publication used a single allele cited in a Methods section), gene names can be used in preference to allele names (*let-37* rather than *mn138* or *let-37(mn138)*).

3.1.3.

Optional suffixes indicating characteristics of an allele can follow the allele name. These are usually two-letter non-italicized letters, e.g., *hc17*ts, where ts stands for temperature-sensitive, or *pk15*te, where te stands for transposon-excision.

3.1.4.

The wild type allele (version) of a gene is defined as that present in the Bristol N2 reference strain, stored frozen at the CGC and other locations. The wild type allele can be designated by a plus sign immediately after the gene name, *dpy-5+*, or, more commonly, by including the plus sign in parentheses, *dpy-5(+)*.

3.2. Gene knockouts

3.2.1.

Many gene knockouts constructed to date are small deletions (<5 kb) generated by chemical mutagenesis, by transposon excision, or by genome engineering. These are named as alleles, sometimes with the optional suffix te (transposon-excision) or ko (knockout). Example: *zyx-1(gk190)* is a 777 bp deletion in the *zyx-1* gene.

3.2.2.

Some of the small deletions generated by chemical mutagenesis, by transposon excision, or by genome engineering may remove parts of two adjacent genes. If only two genes appear to be affected, then the deletion is given a single allele name, but the genotype is written with both gene names coupled with an ampersand (&). Example: allele *ok615* is a 1422 bp deletion of two adjacent genes, so it can be written *rad-54&tag-157(ok615)*.

3.2.3.

Deletions that affect more than two genes are named as Deficiencies "(*Df*)", as described in Section 3.5 Chronmosomal aberrations.

3.3. Genome engineering

Genome engineering (CRISPR-Cas9, TALENs, etc.) is increasingly being applied to *C. elegans* and related nematodes. The resulting genomic alterations require some additional recommendations. The aim is to provide compact and unambiguous ways of describing and referring to engineered changes to endogenous loci, as distinct from transgenic constructs that are inserted elsewhere in the genome.

3.3.1.

Each engineered modification to an endogenous locus (point mutations, deletions, insertions, or combinations thereof) should receive a unique allele designation, using the standard allele designation of the originating laboratory. For example: *bus-50(e5000)*. A single allele designation is employed even if multiple changes are made to the locus.

3.3.2.

Optional brackets can be employed to provide additional information. Example: *bus-50(e5000*[T110E]) (an engineered missense mutation).

3.3.3.

A new allele designation is assigned for each independent, yet identical, engineered gene change. For example, if in a structure/function analysis of *gene-x*, the co-conversion strategy used resulted in the R71C change in *dpy-10* in 15 strains containing modifications to *gene-x*, then 15 different allele designations need to be assigned to *dpy-10* (in additional to relevant allele assignments for *gene-x*).

3.3.4.

An engineered fusion of GFP to the C-terminus of BUS-50 would be: *bus-50(e5001[bus-50::gfp])*. Similarly, an N-terminal fusion would be: *bus-50(e5100[gfp:bus-50])*. In the case of an engineered internal GFP fusion, the designation is the same as a C-terminal fusion, with a detailed description of the structure in the relevant publication.

3.3.5.

As a shorter and more convenient form, and where unambiguous, this could be referred to as: *bus-50::gfp*. Such abbreviations should be clearly defined where first used in a publication.

3.3.6.

An engineered insertion of GFP plus the *unc-119*(+) selectable marker, flanked by *loxP* sites, would be: *bus-50(e5002[bus-50::gfp + loxP unc-119(+) loxP])*.

3.3.7.

Each additional engineering of the endogenous locus requires a new allele number. In the example of *bus-50(e5002)*, following Cre-mediated recombinase removal of *unc-119*(+) so that a single loxP site remains, the new genotype would be *bus-50(e5003[bus-50::gfp +loxP])* or *bus-50(e5003)* for short. The original allele can be indicated in brackets with a preceding asterisk (*), in order to allow searches for all derivatives from a given change. The above example would be *bus-50(e5003[*e5002])*.

3.3.8.

Engineered insertions in apparent intergenic regions are given standard *Is* or *Si* (see Section 7.5) insertion names, for example *eIs2002*. Optional descriptors can include the nature of the insertion, e.g., *[unc-119::gfp]* and the position in the genome, e.g., *[III:2992500]*, to give *eIs2002[unc-119::gfp]* or *eIs2002[unc-119::gfp, III:2992500]*.

3.3.9.

Engineered changes to existing integrated transgenes, either *Is* or *Si*, should receive new *Is or Si* numbers using originating lab's prefix. The original *Is* or *Si* insertion can be indicated in brackets with a preceding asterisk (*), in order to allow searches for all derivatives from a given insertion. For example, an engineered change from GFP to mCherry in *eIs2002* might be named as *ozIs909*, or *ozIs909[unc-119::mCherry *eIs2002]*.

3.4. Modifiers: suppressors, revertants, and enhancers

3.4.1.

There is no special nomenclature for modifier mutations. Many extragenic suppressor loci are called *sup* (40 *sup* loci defined so far, with a wide variety of properties and mechanisms). An increasing number of more specific modifier gene classes have been established, such as *smu* (suppressor of *mec* and *unc*), and *smg* (suppressor with *m*orphogenetic effect on *g*enitalia) and *sel* (suppressor/*e*nhancer of *l*in-12).

3.4.2.

Intragenic suppressors or modifiers are indicated by adding a second allele name within parentheses; for example, *unc-17(e245e2608)* is an intragenic partial revertant of *unc-17(e245)*.

3.4.3.

Mutations known to be chromosomal rearrangements, rather than intragenic lesions, are named differently, as described in the Section 3.5.

3.5. Chromosomal aberrations.

*Du*plications (*Dp*), *def*iciencies (*Df*), *in*versions (*In*), and *t*ranslocations (*T*) are known in *C. elegans* genetics and cytogenetics. These are given italicized names consisting of the laboratory mutation prefix, the relevant

abbreviation, and a number, optionally followed by the affected linkage groups in parentheses (e.g., *eT1(III;V)*, *mnDp5(X;f)*, where *f* indicates a free duplication). If linkage groups are indicated in a translocation they should be in the order 'transposed from'; transposed to'. Chromosomal balancers of unknown structure can be designated using the abbreviation *C*, e.g., *mnC1(II)*. See WormBook chapter Genetic balancers for a more detailed explanation.

4. Polymorphisms and other genetic elements

4.1. SNPs and RFLPs

Polymorphic sites, which are mostly SNPs (single *n*ucleotide *p*olymorphisms) or RFLPs (*r*estriction *f*ragment *l*ength *p*olymorphisms) derived from natural isolates of *C. elegans* (and thus differ from the Bristol N2 reference) are designated by an italic letter *P* and an italic number, preceded by the allele designation for the laboratory responsible for identifying the site. For example, *stP17* and *stP196* are RFLPs identified in the laboratory of R.H. Waterston, and *amP6* and *amP15* are SNPs identified in the laboratory of K. Kornfeld.

SNPs identified in whole genome sequencing projects from one or more natural isolates are not assigned a name but may be referred to by their WBVariationID. For example, *WBVar01710822* is an intronic SNP in the *npr-1* gene found in various natural isolates. Other identifiers (such as those used in publications, internal IDs assigned by the project or historical names) are incorporated into WormBase to enable users to recover such entities through searches.

4.2. Natural copy number variants

Hundreds of independent natural isolates of *C. elegans* have been recovered from multiple locations around the world. The genomes of some of these isolates contain large (>10 kb) deletions, duplications, or insertions, relative to the reference wild-type strain, Bristol N2. Deletions are named with the prefix *niDf* (*n*atural *i*solate *def*iciency) followed by a number. Duplications and insertions are named with the prefix *niDp* (*n*atural *i*solate *dup*lication or insertion), followed by a number. Numbers for *niDf* and *niDp* variants are assigned by application to: genenames@wormbase.org

4.3. Introgressed regions in near-isogenic lines.

Genetic regions that have been introgressed from one natural isolate of *C. elegans* into the background of a different isolate are named in a manner similar to that used for deficiencies (*Df*) and duplications (*Dp*). Each *I*ntrogressed *R*egion is given an italicized name consisting of the relevant laboratory allele designation, the letters *IR*, and a number. Thus, a region from the X chromosome of Hawaiian strain CB4856 crossed onto a Bristol N2 background, and created in the Kruglyak lab (allele code *qq*) has been given the name *qqIR1*. Additional information about genetic map location and strain origin can be provided in an optional parenthesis. So this example could be more fully written as *qqIR1(X, CB4856)*, with the implicit assumption that the strain background is Bristol N2. The strain background and the direction of introgression can also be specified, using the symbol >, with this example being written *qqR1(X, CB4856>N2)*.

4.4. Transposons and transposon insertions

Types of *C. elegans* transposons are called Tc1, Tc2, etc., where each number represents a different family. Transposon names are not italicized except when included in a genotype. Different natural isolates of *C. elegans* have different distributions of these transposons, which result in polymorphic differences from the reference wild-type strain Bristol N2. The differences between natural isolates and Bristol N2 are given polymorphism names, as described below.

The endogenous transposons of *C. elegans* can be mobilized to generate new insertional events. In addition, foreign transposons such as Mos1 can be introduced by transformation, and then mobilized to create new insertions. All these newly generated transposon insertions can be named as simple alleles, with an optional suffix indicating the nature of the transposon. They are treated as alleles of named genes if they are located within the boundaries of a gene. Example: *r293* is a Tc1 insertion in the gene *unc-54*. An optional descriptor can also be added after a double colon to indicate the nature of the insertion. Example: *unc-54(r293::Tc1)*. Note that such insertions may often be silent in terms of gene activity, for example if an insertion occurs within an intron and can be spliced out.

Newly generated *T*ransposon *i*nsertions, especially those located in apparently intergenic regions, may also be given *Ti* (transposon insertion) names. These consist of the designation identifying the laboratory of origin, the two letters *Ti*, and a number, all italicized. Example: *eTi13* is an insertion of a Mos transposon into an intergenic region on *LGIII*.

Transposon loci have ID names formed from 'WBTransposon' followed by a unique number, like *WBTransposon00000623*.

Their exon-like structure is curated as a Transposon_CDS (coding sequence) object with a name like *C29E6.6* formed from the YAC or cosmid or clone they are on followed by a number which uniquely identifies it from the other CDS-like objects on that clone, YAC or cosmid.

Transposons and Transposon_CDS are not currently classed as genes in WormBase and so do not have a parent gene object. The WBTransposon and representation on the Genome Browser should be viewed as analogous to the WBGene and how it is displayed.

4.5. Transgenes

Transformation of *C. elegans* with exogenous DNA by microinjection usually leads to the formation of a transmissible extrachromosomal array containing many copies of the introduced DNA. Extrachromosomal arrays differ in their frequency of meiotic and mitotic transmission. Extrachromosomal arrays can subsequently be integrated into the genome by irradiation. Direct integrative transformation with exogenous DNA can be obtained by microparticle bombardment, mosSCI, or miniMos techniques. As these integrative events are not associated with the endogenous locus corresponding to the exogenous DNA they are considered distinct from genome engineered changes to the endogenous locus and thus have different designations.

4.5.1.

Extrachromosomal arrays are given italicized names consisting of the laboratory allele prefix, the two letters *Ex*, and a number.

4.5.2.

Integrated transgenes are designated by italicized names consisting of the laboratory allele prefix, the two letters *Is*, and a number. Single copy integrants, usually generated by the MosSCI or miniMos insertion techniques, are a subset of integrated transgenes and are designated by italicized names consisting of the laboratory allele prefix, the two letters *Si*, and a number.

4.5.3.

Transgene designations *Ex*, *Is,* and *Si* can optionally be followed by genotypic or molecular information describing the transgene in square brackets. For example, *eEx3* or *eIs2* or *stEx5[sup-7(st5) unc-22(+)]*.

4.5.4.

Gene fusions incorporated in transgenes that consist of a *C. elegans* gene or part thereof fused to a reporter such as *lacZ* or *GFP* are indicated by the *C. elegans* gene name followed by two colons and the reporter, all italicized: *pes-1::lacZ, mab-9::GFP*. To distinguish between transcriptional and translational fusions, a lowercase italicized *p* following the gene name has been used to indicate transcriptional fusions, e.g., *mab-9p::GFP*.

5. Genotypes

The genotype of an animal is specified by listing all known differences between its genotype and that of wild type, which is defined by convention as Bristol N2. Each such difference is assigned a unique name. Table 1 lists the currently recognized types of difference that have designations, described at greater length elsewhere in the chapter.

Table 1. Nomenclature terms and usage in *C. elegans* research

R107.8	Systematic gene identification (the 8th predicted gene on cosmid R107)
lin	Gene class, "abnormal cell LINeage"
lin-12	The 12th "abnormal cell LINeage" gene named
ar170	Allele name ("*ar*" allele designation from the Greenwald lab; the 170th allele generated in the Greenwald lab)
LIN-12	Protein name (product of *lin-12* gene)
Lin	Phenotype (abnormal cell **lin**eage phenotype)
lin-12(ar170) or *lin-12(ar170)/lin-12(ar170)*	Homozygous for *lin-12(ar170)* allele
lin-12(ar170)/+	Heterozygous for *lin-12(ar170)* allele
lin-12(ar170)/lin-12(n941)	Heterozygous for two different *lin-12* alleles (also call a compound heterozygote) ("*n*" allele designation from the Horvitz lab)
lin-12(n676n930)	*n930* is an intragenic revertant of the *n676 lin-12* gain of function allele
gk181351	Allele of *lin-12* from the Million mutant project, MMP ("*gk*" MMP allele numbering starts at 100000)
WBVar00070143	Single Nucleotide Polymorphism (SNP) in the *lin-12* 3'UTR found in the natural isolate CB4856 (Hawaiian strain)
stP17	Restriction Fragment Length Polymorphism (RFLP) ("*st*" allele designation from the Waterston lab)
lin-12p::gfp	GFP transcriptional fusion (using only the promoter of the gene)
lin-12::gfp	GFP translational fusion (in which *gfp* is inserted at the C-terminus of the *lin-12* coding sequence)
GS60	Strain name ("GS" laboratory/strain designation from the Greenwald lab); full genotype *unc-32(e189) lin-12(n676n930) III*.
lin-41(tn1490)	*lin-41* allele generated by chemical mutagenesis ("*tn*" allele designation from the Greenstein lab)
lin-41(xe11)	*lin-41* allele generated by genome engineering ("*xe*" allele designation from the Grosshans lab)
lin-41(tn1490[G883E]*)*	Amino acid change in *lin-41* allele *tn1490* indicated
mnDp26	Duplication (*Dp*) ("*mn*" allele designation from the Herman lab)
nDf17	Deficiency (*Df*) (multi-gene deletion)
nT1 (IV; V)	Translocation (*T*) involving chromosomes *IV* and *V*
rtEx726	Extrachromosomal (*Ex*) transgene array ("*rt*" allele designation from the Hart lab)
arIs80	Integrated (*Is*) transgene
dotSi110	Single copy insertion (*Si*) ("*dot*" allele designation from the J. Chen lab)
glc-1(pk54::Tc1)	Transposon (Tc1) insertion in *glc-1* gene ("*pk*" allele designation from the Plasterk lab)

pgIR2	Introgressed region (*IR*); full genotype *(II, CB4856>N2)*, which indicates that a region from chromosome *II* of the Hawaiian strain CB4856 has been crossed into the Bristol N2 background ("*pg*" allele designation from the M. Goodman lab).

Where necessary, wild type sequence can be indicated using the symbol +. Because every genetic "feature" (i.e., difference from Bristol N2) has a unique name, an animal's genotype is fully specified by listing all the named features that it carries. Example: *e2123; mdIs18*.

For clarity and convenience, additional information about genes, chromosomes, transgene contents, etc., can be added as described elsewhere in this document, to produce a more informative genotype. Example: *pha-1(e2123ts) III; mdIs18[pha-1(+) unc-17::GFP]*

Strains carrying more than one mutation are designated by sequentially listing mutant genes or alleles according to the left-right (= up-down) order on the genetic map/genome sequence. Different chromosomes (linkage groups) are separated by a semicolon and given in the order *I, II, III, IV, V, X, f, M. I-V* are the five autosomes, *X* is the *X* chromosome, *f* refers to free duplications or chromosomal fragments, and *M* is the mitochondrial genome. For example: *dpy-5(e61) I; bli-2(e768) II; unc-32(e189) III*.

Integrated transgenes (*Is* and *Si*) should be grouped with other mutant genes or alleles on the chromosome, if it is known into which chromosome the transgene is integrated. Extrachromosomal arrays (*Ex*) and unmapped integrated arrays should be included at the end of the genotype. Example: *oxTi330 III; gaIs283*.

Heterozygotes, with allelic differences between chromosomes are designated by separating mutations on the two homologous chromosomes with a slash. For example, the compound heterozygote *lin-12(n941)/lin-12(n137)*. Where unambiguous, wild type alleles can be designated by a plus sign alone, or even omitted. For example, *dpy-5(e61) unc-13(+)/dpy-5(+) unc-13(e51) I* can also be written *dpy-5 +/+ unc-13 or dpy-5/unc-13.*

6. Phenotypes

Phenotypic characteristics can be described in words, e.g., dumpy animals or uncoordinated animals. If more convenient, a nonitalicized 3-letter or 4-letter abbreviation, which usually corresponds to a gene class or gene name, may be used. The first letter of a phenotypic abbreviation is capitalized, e.g., Unc for uncoordinated, Dpy for dumpy. If it is necessary to distinguish among related but distinguishable phenotypes, the relevant gene number can be added, e.g., Unc-4 and Unc-13 to differentiate the distinct phenotypes produced by mutations in the two genes *unc-4* and *unc-13*. WormBase maintains a standard set of defined phenotype descriptors (the WormBase Phenotype Ontology).

Abbreviations that do not correspond to a gene class or gene name can also be used, e.g., Muv for *mu*ltiple *v*ulval development, and Daf-c for *da*uer-*f*ormation-*c*onstitutive. To avoid ambiguity and conflicts, phenotype abbreviations not corresponding to a gene name are controlled by WormBase and requests for names should be made, prior to abstract or manuscript submission, via email to: genenames@wormbase.org.

A common and accepted convention, when comparing a mutant with the wild type, is to use the prefix non- to refer to the wild type phenotype, for example, non-Lin (= wild type cell lineage) or Dpy non-Unc (= wild type with respect to movement, but dumpy with respect to body shape).

7. Proteins

The protein product of a gene can be referred to by the relevant gene name, written in non-italic capitals, e.g., the protein encoded by *unc-13* can be called UNC-13.

In some cases the gene name and the protein products have distinct names, often in situations where the gene name is based on phenotype and the protein product is named based on sequence similarity or biochemical activity. For example, for the gene *let-60* and the corresponding protein RAS, designation *let-60*/RAS or *let-60* RAS can be used.

Where more than one protein product is predicted for a gene (usually as a result of alternative message processing), the different proteins are distinguished by adding 'isoform' and then the isoform letter derived from the isoform letter of the name of the WormBase CDS, e.g., the gene 'tra-1' has two CDS isoforms: 'Y47D3A.6a' and 'Y47D3A.6b' which give rise to the protein isoforms: 'TRA-1, isoform a' and 'TRA-1, isoform b'.

Mutant protein products can be named by the missense change, for example a mutant 'TRA-1, isoform a' protein with a Pro to Leu change at codon 79 would be written: 'TRA-1, isoform a (P79L)'.

8. Strains

A strain is a set of individuals of a particular genotype with the capacity to produce more individuals of the same genotype. Strains are given non-italicized names consisting of two or three uppercase letters followed by a number. The strain letter prefixes refer to the laboratory of origin and are distinct from the allele (mutation) letter prefixes. Examples: CB1833 is a strain of genotype *dpy-5(e61) unc-13(e51)*, originally constructed by S. Brenner at the MRC Laboratory of Molecular Biology (strain designation CB, allele designation *e*), and MT688 is a strain of genotype *unc-32(e189) +/+ lin-12(n137) III; him-5(e1467) V*, constructed in the laboratory of H.R. Horvitz at M.I.T. (strain designation MT, allele designation *n*).

Strains can and should be preserved as frozen stocks at −70 °C, or ideally in liquid nitrogen, in order to ensure long-term maintenance and to avoid drift or accumulation of modifier mutations.

Bacterial strain names employ the two or three letter Laboratory/Strain designation, followed by "b". For example, CBb#. This facilitates distinguishing nematode strains from bacterial strains. Please provide full information on species and relevant genotype of the bacteria.

9. Other nematodes

Research and genomic analysis of non-*C. elegans* species is increasing rapidly. An important mission of WormBase is to make available information for each species listed in the Overview section within the database structure developed for *C. elegans*. For these organisms WormBase will also supervise gene naming in order to maximize consistency with *C. elegans*. It is recommended that nomenclature in general should follow the principles used for *C. elegans*, as far as possible. Gene name proposals and queries should be made online via WormBase or sent to genenames@wormbase.org

9.1. Species prefixes

In order to specify unambiguously the nematode species-of-origin, an optional 3-letter standard prefix and hyphen can be added to the gene name (Table 2). Examples: the *C. briggsae* and *Pristionchus pacificus* orthologs of *C. elegans tra-1* are called *Cbr-tra-1* and *Ppa-tra-1*, respectively. WormBase coordinates the species prefix designations to avoid the use of the same designation for more than one species; contact genenames@wormbase.org for prefix proposals.

Table 2. Species prefixes currently in use

Species	Prefix
C. elegans	*Cel-*
C. briggsae	*Cbr-*
C. remanei	*Cre-*
C. brenneri	*Cbn-*
C. japonica	*Cpj-*
Heterorhabditis bacteriophora	*Hba-*
Oscheius tipulae	*Oti-*
Pristionchus pacificus	*Ppa-*
For additional prefixes since publication of this article see WormBase	

9.2. Gene naming: homologous genes

Genes predicted from whole genome sequences in other nematode species will, in many cases, have identifiable close homologs in *C. elegans*, for which approved names already exist. In these cases, the same name should be used as in *C. elegans*, with the relevant species identifier. Possible scenarios are listed below.

9.2.1. One-to-one

Where one gene in *C. elegans* corresponds to a single gene in another nematode species, ortholog naming can be applied automatically. Example: *thoc-1* in *C. elegans* has a *C. briggsae* ortholog, *Cbr-thoc-1*.

9.2.2. One-to-many

Where one gene in *C. elegans* is related to multiple genes (paralogs) in another nematode species, these paralogs can be named using additional decimal numbers. Example: *thoc-3* in *C. elegans* has two *C. briggsae* paralogs, *Cbr-thoc-3.1* and *Cbr-thoc-3.2*.

9.2.3. Many-to-one

Where multiple genes exist in *C. elegans* but only a single gene in another nematode species, it is recommended that either the most closely similar, or the lowest numbered *C. elegans* gene, be used to name the single gene, as appropriate.

9.2.4. Many-to-many.

Where multiple closely related genes can be identified in both species but the phylogenetic relationships of the two sets are complex, new gene numbers can be assigned to the set of genes in the other nematode species, after consultation with genenames@wormbase.org

9.2.5. A standard gene name has not yet been assigned in *C. elegans*

The gene can be referred to using the sequence name, such as *Hba-W01B11.3*. However, in such cases it will usually be both feasible and desirable to assign a standard name to the *C. elegans* gene as well, at the same time.

9.3. Gene naming: non-homologous genes

It is expected that many genes in other nematode species will lack obvious close homologs in *C. elegans*, because of loss or substantial divergence during the evolution of *C. elegans*. These genes can be given new gene numbers if they belong to an identifiable named class in *C. elegans*, or else new gene name classes can be established for them. In either case, assignment of an approved name should be made after consultation with genenames@wormbase.org

9.4. Gene naming: forward genetics

A significant amount of mutation-based forward genetic analysis is being pursued in nematodes other than *C. elegans*, in particular using other species of *Caenorhabditis* (*C. briggsae, C. remanei, C. brenneri*, and others), as well as species of *Oscheius* and *Pristionchus*. It is expected that most, but not all, of the mutationally-defined genes discovered in these species will prove to have orthologs with equivalent or similar function in *C. elegans*, and hence that standard genetic names will have been approved already. Possible situations are described below.

9.4.1. The molecular identity is known and orthology is obvious

It is recommended that the *C. elegans* name be used, with the appropriate species identifier prefix. Example: *Ppa-mab-5* is the *Pristionchus pacificus* ortholog of *C. elegans mab-5*.

9.4.2. The molecular identity is not initially known but the mutant phenotype corresponds to a known *C. elegans* mutant phenotype

To reduce gene renaming as much as possible, it is recommended that: a) new mutations should be mapped as precisely as possible, b) all relevant complementation tests should be done, and c) if the genome assembly for that

species is sufficiently accurate, potential *C. elegans* orthologs should be tested by sequencing or RNAi. If the gene still seems to be novel, then a new gene class name can be established, following consultation with genenames@wormbase.org in order to ensure that the new name is available and appropriate. An appropriate name should describe the phenotype yet be distinct from the corresponding *C. elegans* name. If the molecularly identity is subsequently found to be conserved in *C. elegans*, then the name would typically revert to the *C. elegans* name, but this would be determined on a case-by-case basis.

9.4.3. The molecular identity is unknown and the mutant phenotype does not correspond to a known *C. elegans* mutant phenotype

A new gene class name can be established, following consultation with genenames@wormbase.org in order to ensure that the new name is available and appropriate. Example: *cov* = *C*ompetence and/or centering *O*f *V*ulva abnormal.

10. Acknowledgements

This work is funded by U41 HG002223D to WormBase (M.A.T and T.S.) and NIH-ORIP (P40 OD010440) to the CGC (A.D.). The following individuals made substantial contributions to the *C. elegans* nomenclature system and its utilization: Sydney Brenner, Robert Horvitz, Jonathan Hodgkin, Robert Herman, Phil Anderson, Mark Edgley and Don Riddle.

11. References

Brenner S. (1974). The genetics of *Caenorhabditis elegans*. Genetics *77*, 71-94. Abstract Article

Harris T.W., Baran, J., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., Done, J., Grove, C., Howe, K., et al. (2014). WormBase 2014: new views of curated biology. Nucleic Acids Res. *42* (Database Issue) D789-793. Abstract Article

Horvitz H.R., Brenner, S., Hodgkin, J., and Herman, R.K. (1979). A uniform genetic nomenclature for the nematode *Caenorhabditis elegans*. Mol. Gen. Genet. *175*, 129-133. Abstract Article

The *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science *282*, 2012-2018. Abstract Article