Conventions For Naming Zebrafish Genes

(Source: The Zebrafish Nomenclature Committee from Zebrafish Book 5th Edition)

Genetic designations vary immensely from organism to organism, and the zebrafish community recognizes the importance of agreeing upon conventions for naming mutants and genes. The following conventions have been chosen by most labs to minimize confusion and maximize the utility of the nomenclature and the ease with which people outside (as well as those within) the field can follow the field. It is very important that the entire zebrafish community adopt one set of conventions. The current nomenclature guidelines are available from ZFIN.

Contents

1. GENE NAMES AND SYMBOLS

Full gene names are lowercase italic, and gene symbols are three or more lowercase letters and are also italicized. The letters should be unique with respect to other named zebrafish mutants and genes and except in cases of established orthology, where the gene symbol should match that of the orthologue. Zebrafish gene designations should not be preceded by 'Z' or 'Zf'. The use of punctuation such as period and hyphens in gene names or symbols is discouraged, except under specific circumstances described below.

Gene names should be registered at https://zfin.org/action/nomenclature/gene-name

1.1 Genes identified by cloning Genes should be named after the mammalian orthologue whenever possible. When mammalian orthologues are known, the same name and abbreviation should be used, except all letters are italicized and lower case. Members of a gene family are sequentially numbered.

Examples:
Names - *engrailed 1a, engrailed 2b* Symbols - *eng1a, eng2b*

In some cases when a zebrafish gene has been renamed to the mammalian orthologue from an older zebrafish name, it is still preferable within a publication to refer to the previous name. Refer to the previous name by appending the previous name in parentheses. Previous names are searchable at ZFIN.

Examples:*shha (syu), bmp2b (swr)*

1.2 *Duplicated genes:*The zebrafish genome contains duplicated segments that resulted from a genome-wide duplication in the ray fin fish lineage after it diverged from the lobe fin lineage (that included avian and mammalian species). For this reason, zebrafish often have two copies of a gene that is present as a single copy in mammals.

In these cases, symbols for the two zebrafish genes should be the same as the approved symbol of the human or mouse orthologue followed by "*a*" or "*b*" to indicate that they are duplicate copies. Before these symbols are assigned, it is important to provide evidence by mapping that the two copies reside on duplicated chromosome segments. It is preferable that all copies in one of the duplicate chromosome segments use the same "*a*" or "*b*" suffix, although this will not always be possible for historical reasons. This terminology should not be used for duplicates that resulted prior to the divergence of ray fin and lobe fin fish. In these cases it is preferable to use terminology that is most consistent with the mammalian nomenclature.

Examples: *hoxa13a, hoxa13b*

In some cases when there is a unique mammalian orthologue, but addition of the *a, b*suffixes would conflict with a different mammalian gene symbol, then numerical suffixes *.1, .2* should be appended to the orthologous mammalian gene symbol instead of *a, b*. Tandem duplicates with a single mammalian orthologue may also be appended with a *.1, .2*, using the same symbol as the mammalian orthologue.

Examples:*stat5.1, stat5.2*

When mammalian gene duplications prevent identification of a unique mammalian orthologue, then an alternate gene symbol should be chosen. A possible choice would be an approved gene symbol from a unique non-mammalian orthologue. When a gene is homologous to a human gene, but orthology is ambiguous, the gene should be named after the closest mammalian homologue with the word 'like' appended to the name of the homologue. In some cases, a gene family described in zebrafish is homologous to a mammalian gene family but the evolution of the gene family is ambiguous. Under these circumstances the zebrafish gene family should be named with the same stem as the mammalian gene family with the gene number beginning after the end of the mammalian numbering and continuing sequentially throughout the gene family. If the members of the gene family are on the same chromosome, the adjacent genes should be given sequential numbers.

Mutant loci with unidentified genes

Mutant loci for which the gene has not yet been identified are given placeholder gene names. When the gene is identified, it is renamed following standard nomenclature guidelines as described above. Genes identified by mutation are typically named to reflect the mutant phenotype. The symbol should be derived from the full name. Numbers should generally not be used in naming a mutant.

Example: *touchy feely, tuf*

Mutant names should be registered at https://zfin.org/action/nomenclature/line-name.

1.4 Genes identified only by genomic sequencing projects
Large-scale genome sequencing projects use a variety of prediction methods to identify both open reading frames and genes. Some of these genes are already known, while others are new. Novel genes identified by these means often cannot be identified and are assigned a name comprised of a prefix, a clone name, and an integer. The prefix is used to specify the research institution that identified the gene (e.g., "*si*" for the Sanger Institute). A colon separates the prefix from the clone identifier. In many cases, there are multiple predicted reading frames in a single clone. These genes are distinguished with a full stop (period) between the clone name and an integer. Integers are assigned to genes in the clone as they are identified and do not indicate the order of genes. If part of a gene is found in more than one clone, the name of the first clone in which the 5' portion of the gene is found takes precedence.

Examples: *si:bz3c13.1, si:bz3c13.2, si:bz3c13.3*

Genes initially identified by genomic sequencing projects are renamed using standard nomenclature guidelines (described above) as more information about them becomes available.

1.5 Genes identified only by other large scale projects
Large-scale sequencing of ESTs or full length cDNA clone sets often result in large numbers of unidentified genes. These are given placeholder names with the project prefix, a colon and a clone number, similar to genes identified by genomic sequencing projects. In these cases, the clones usually contain only one or a fragment of a single gene.

Examples: *im:7044540,* zgc:165514

1.6 Transcript variants
Transcript variants that originate from the same gene are not normally given different gene symbols and names. However, variants from a single gene can be distinguished in publications by adding to the end of the full name a comma, "*transcript variant*", and a serial number; and by adding to the end of the symbol an underscore, "*tv*", and a serial number.

Examples:
Names -
myosin Vla, transcript variant 1 *myosin Vla, transcript variant 2* Symbols -
myo6a_tv1 *myo6a_tv2*

---

2. PROTEINS

The protein symbol is the same as the gene symbol, but non-italic and the first letter is uppercase.

Examples: Ndrw, Brs, Eng1a, Eng2b, Ntl

Note the differences between zebrafish and mammalian naming conventions:

species / gene / proteinzebrafish / *shha*/ Shha
human / *SHH* / SHH
mouse / *Shh* / SHH

In publications, it is sometimes convenient to refer to a protein which has been renamed based on orthology using the more commonly known name in parentheses following the current name.

Examples: Shha (Syu), Bmp2b (Swr)

---

3. ALLELES and GENOTYPES

3.1 Allele designations

When describing genes wild-type alleles are indicated using a superscript "+", while mutant alleles are indicated using a superscript allele designation. Allele designations are composed of a laboratory-specific designation followed by a number. The full list of laboratory designations can be found at ZFIN. Other letters should not immediately follow the laboratory designation but may be appended to the end of the allele designation to make them unique. Dominant alleles have a *d* in the first position of the superscript to distinguish them from recessive alleles.

Examples:

b is the Eugene designation; *m* is for MGH, Boston; *t* is Tuebingen, Germany
wild type_: lof{}{}_ ndr2, brs^{+} mutant: lof^{dt2}, $ndr2^{b16}$, $ndr2^{m101}$, $ndr2^{t219}$

3.2 Genotype nomenclature for publications

For unlinked loci, heterozygotes and homozygotes are distinguished by having each allele separated by a slash "/".

Examples:
$ednrb1^{b140}/ednrb1^{+}$ (heterozygote, can be abbreviated $ednrb1^{b140/+}$)
$ednrb1^{b140}/ednrb1^{b140}$ (homozygote, can be abbreviated $ednrb1^{b140/b140}$ or $ednrb1^{b140}$)

For homozygous genotypes, the genotype at each locus is listed in order according to Chromosome number, from 1 to 25, with a semicolon to separate loci on different chromosomes.

Examples: $ednrb1^{b140}$; $slc24a5^{b16}$

For heterozygous genotypes, loci on homologous chromosomes are separated by a slash.

Examples: $ednrb1^{b140}/ednrb1^{+}$; $slc24a5^{b16}/slc24a5^{m592}$

For linked loci, the haplotype on each chromosome is written as a sequence, with a space separating syntenic loci, and loci are placed in the order they appear on the Linkage Group, top to bottom. Homologous chromosomes are separated by a slash, and non-homologous chromosomes are separated by semicolons.

Examples: $ednrb1^{b140}cx41.8^{t1}$; $slc24a5^{b16}$

For unmapped loci, genotypes of unmapped loci are listed alphabetically within braces following genotypes of mapped loci on different chromosomes.

Examples: $ednrb1^{b140}$; $mycbp2^{tj236}$ {$edil^{tz253}$color:lack}} (*edi* is unmapped, all three loci are on different chromosomes)

Poorly resolved loci on same chromosome are listed alphabetically within braces.

Examples:

{$abc^{b000}def^{m000}$color:lack}} (poorly resolved loci on same chromosome)$ednrb1^{b140}$ {$abc^{b000}def^{m000}$color:lack}} $cx41.8^{t1}$ (poorly resolved loci in a known interval between mapped loci, all on same chromosome)

3.3 Genotype display at ZFIN

Due to technical constraints, genotypes at ZFIN are shown in alphabetical order by gene, and then by allele designation. See below for display of complex genotypes involving transgenic or chromosomal rearrangements.

4. CHROMOSOMES AND ABERRATIONS

The chromosome numbering system corresponds to the old Linkage Group designations with what was LG 1 now named Chr 1. Chromosomes are designated by non-italic numerals, 1 to 25. Chromosome differences have not been observed between males and females.

Examples:
Chr1 to Chr25

Chromosome rearrangements are indicated with the following prefixes, followed by the details within parentheses. See below for specific examples. Common prefixes include:
Df, deficiency
Dp, duplication
In, inversion
Is, insertion
T, translocation
Tg, transgene

4.1 Deficiencies: The general format for naming a deficiency is:

Df(Chr##)xxx^{allele}

*Df*indicated deficiency. The term *xxx*should describe the salient features of the deficiency, as determined by the investigator. In cases where the deficiency removes sequences from a named gene, it should contain the standard symbol for that gene. The allele name should follow standard nomenclature conventions (laboratory designation followed by allele name).

The chromosome where the deficiency maps should be specified by its number (##) using two digits (i.e., 03 for Chr 3) so that computers will order them properly.

Example: *Df(Chr12)dlx3*^{b380}

Deficiencies are written as an allele of a gene when the gene is disrupted at one of the two breakpoints of the deficiency. The omitted genes are listed as missing.

Example:

ndr2^{b16}is a deficiency in which a portion of *ndr2* is missing, causing the deficiency to be an allele of *ndr2*.

4.2 Translocations: The general format for naming translocations depends upon the type of translocation:

Resiprocal translocations have two separate chromosomal elements, and each element has a distinct name:

T(Chr##;Chr##)xxx^{allele}*,##U.##L* and *T(Chr##;Chr##)xxx*^{allele}*,##U.##L*

*T*indicates translocation. The elements in the parentheses are the chromosomes involved, the lower numbered chromosome is listed first, and the chromosomes are separated by a semicolon. The chromosomes should be specified by their numbers (##) using two digits (i.e., 03 for Chr 3) so that computers will order them properly.

The term *xxx*should describe some salient feature of the translocation, as determined by the investigator. In cases where the translocation moves a named gene primarily studied by the investigator, *xxx* would usually be the standard symbolfor that gene. Alternatively, *xxx* could just be an experimental series number.

The allele nameshould follow standard nomenclature conventions (laboratory designation followed by allele name, superscripted). After the allele name comes a comma, and then a phrase that indicates the new order of the chromosomes, starting from the top of the chromosome as displayed by convention. The first number (##) is the Chr number, followed by upper case *U* to indicate the upper arm of a chromosome or by upper case*L* to indicate the lower arm of a chromosome. The location of the centromere is indicated by a period. No spaces. Translocations are written as an allele of a gene when the gene is disrupted at one of the breakpoints of the translocation. There can be as many as four alleles of a translocation.

Example:
T(Chr02;Chr12)cyc^{b213}*,02U.12L02L* and *T(Chr02;Chr12)cyc*^{b213}*,12U.12L02L*

This example illustrates a reciprocal translocation where a portion of the lower arm of Chr12 was translocated interstitially into the proximal lower arm of Chr2 and a portion of the lower arm of Chr2 was translocated to the distal lower arm of Chr12.

Resolved translocations are where the two elements of the translocation separate and a mutant line has just one of the elements. This results in the animal being monosomic for some chromosome regions and trisomic for others. In these cases, the mutant line would be designated with just one of the elements rather than two as in the reciprocal designation above. The allele name would remain the same to indicate their common origin and common breakpoint.

Example: *T(Chr02;Chr12)cyc*^{b213}*,02U.12L02L*

4.3 Transgenic lines and constructs

Transgenic constructs now have their own pages in ZFIN. Transgenic construct names are important because the construct name is used in the transgenic line nomenclature when the insertion is NOT an allele of a gene (see below).

4.3.1 Transgenic constructs

Tg(promoter:gene)

*Tg*indicates transgene. Within the parentheses, the most salient features of the transgene should be described. Brevity and clarity in the transgene name are favored, in general, over exhaustive detail. Regulatory sequences should be listed to the left of the colon, and coding sequences to the right of the colon. Not all transgenic constructs will have both promoter and coding elements, and in this case, the colon will not be used. In cases where a construct utilizes sequences from a named gene, it should contain the standard zebrafish lowercase symbol for that gene. Human genes may be listed in all capital letters, mouse genes with the first letter capitalized, and other non-zebrafish genes with the two letter species abbreviation (Ss for Salmo Salar) and the first letter of the gene capitalized. For commonly used reporter elements (i.e. GFP, DsRed) use standard naming conventions.

Example: *Tg(SsNdr2:GFP)*

Regulatory sequence could be derived from either an enhancer or promoter, and are denoted by the symbol of the regulated gene. Regulatory or coding sequence fusions should be separated by hyphens. In cases where a number of constructs are generated with differing sizes of promoter elements, these may be specified within the parentheses as follows:

Examples:

Tg(-3.5hhex-ERE:sptb-GFP)

Tg(-6.0hhex-ERE:sptb-GFP)

Here, two constructs having a fusion protein of *spectrin beta (sptb)* and GFP driven by a combination of an estrogen response element and an upstream enhancer containing either 3.5kb or 6.0kb 5' to the *hhex*gene.

However, in a number of cases, the changes within the construct may be too small to change the number of kb. In this case, the constructs will be appended with a period and a number within the parentheses, referring to the element that has changed, instead of including further details in the name. Alternatively, if the .1, .2 nomenclature conflicts with a gene name, then a number may be placed at the beginning of the construct name. The numbering should start with a 1 and increment by one for each different construct. The details of construct differences will be available on the construct pages.

Examples:

Tg(-1.7shha.1:GFP)

Tg(-1.7shha.2:GFP)

Tg(-1.7shha.3:GFP)

Sometimes within a single construct, there are multiple cassettes, each containing regulatory and coding sequences. In this case, it is necessary to distinguish between what is coding in the first cassette with what is regulatory in the second. Multiple cassettes may be distinguished using a comma. In the following example, isl3 promoter and retinoic acid response element drive GAL4, and UAS drives GFP.

Example: *Tg(isl3-RARE:GAL4,UAS:GFP)*

Enhancer, promoter, and gene-trap constructs may use Et, Pt, or Gt, all of which are considered types of transgenic constructs.

4.3.2 Enhancer trap, promoter trap, gene trap constructs

These all use the same nomenclature convention as described for transgenic constructs above, substituting Pt, Gt, Et as necessary.

4.3.3 Transgenic lines:

Transgenic lines are of two types, those that are known to create alleles of genes and those that are not known to create alleles of genes.

For a line that does not create an allele of a gene, the feature name consists of the construct name appended with a unique line number with no superscript. The line number should begin with the laboratory designation followed by a unique number.

Example: *Tg(hsp70l:GFP)mik6*

For lines that do create alleles of a gene, a standard genetic representation is used, where the allele designation is superscripted above the gene, but is appended with a Tg to indicate that it is a transgenic insertion allele. Details regarding the construct used will be available on the genotype page. Gene traps and enhancer traps known to create alleles of a gene are handled in a similar fashion, appending Gt or Et to the allele designation.

Examples:

arnt^{hi2639cTg}

parg^{mn2Et}

4.3.4 Display of complex genotypes at ZFIN

Genotypes at ZFIN are shown in alphabetical order with transgenic lines that are not alleles of genes first, then other alleles.

Example: *Tg(-0.7her5:EGFP)ne2067;hmgcrbs*$^{617/s617}$

---

5. PRIORITY IN NAMES

As described above, zebrafish genes are named based on orthology to a human or mouse gene. If an ortholog cannot be identified, then the name that appears first in the literature will be given priority assuming it follows other nomenclature guidelines. ZFIN recommends submission of proposed gene names via the ZFIN form or consultation with the zebrafish nomenclature committee (nomenclature@zfin.org) for nomenclature assignment.

When a mutation is found in a previously cloned zebrafish gene, then the mutant will be referred to as an allele of the gene. If both the cloned gene and the mutation are known by different names and later found to be the same gene, then the name of the gene usually takes priority. The exception to this rule is when the mammalian gene has a gene symbol that is less than two characters such as the mouse gene brachyury which has the symbol T. In this case the zebrafish gene retained the original name *no tail, ntl*.

---

6. MAPPING AND SEQUENCING INFORMATION

The genome project began in 1994, and by 1996 the genetic map was closed. NIH funded major programs to develop a doubled haploid meiotic mapping panel, deficiency strains and expressed sequence tags (ESTs), The ESTs and anonymous markers have been mapped on two radiation-hybrid panels. The Sanger Institutebegan full genome squencing in 2001. A physical map is being constructed from the BAC libraries used for sequencing. Genomic information is updated regularly on ZFIN.

---

7. CONTRIBUTORS

Marc Ekker (marc.ekker@science.uottawa.ca), Center for Advanced Research in Environmental Genomics, University of Ottawa, Ontario, Canada

Mary Mullins (mullins@mail.med.upenn.edu), Department of Cell and Developmental Biology, University of Pennsylvania, USA

John Postlethwait (jpostle@oregon.uoregon.edu), Institute of Neuroscience, University of Oregon, USA

Monte Westerfield (monte@uoneuro.uoregon.edu), Institute of Neuroscience, University of Oregon, USA

Erik Segerdell, XenBase, University of Calgary, Canada

Melissa Haendel (haendel@ohsu.edu), Oregon Health and Sciences University, Portland, OR.

Ceri Van Slyke (van_slyke@uoneuro.uoregon.edu), Zebrafish Information Network, University of Oregon, USA

---

8. REFERENCES

1. The Zebrafish Science Monitor (1992) Sept. 21.

2. Mullins, M. (1995) Genetic methods: conventions for naming zebrafish genes in The Zebrafish Book (3rd edition, Westerfield, M., ed.), pp 7.1-7.4, University of Oregon Press.

3. Genetic Nomenclature Guide, Trends in Genetics (1998).