GV-SOLAS
Gesellschaft für Versuchstierkunde

Specialist information

From the Committee for Genetics and Laboratory Animal Breeding

Introduction to the nomenclature of genes, mutations and transgenes of mouse and rat

September 2018 – translated February 2021

Authors: Jutta Davidson,
Dirk Wedekind, Johannes Schenkel

Table of contents

Introduction

Nomenclature rules were developed in the past for the exact identification of animal strains and mutations. While these nomenclature rules are in principle desirable for all laboratory animal species used, they have only been clearly defined for the mouse and rat to date.

Overall, the nomenclature rules very comprehensive. In total, there are four “rulebooks” that apply to the nomenclature of mouse and rat:

1. Rules for the nomenclature of mouse and rat strains
2. Rules for genes, genetic markers, alleles and mutations in mouse and rat
3. Rules for the nomenclature of chromosomal aberrations
4. Rules for the nomenclature of mutant alleles that have been generated in embryonic stem cells (ES cells) by the International knock-out Mouse Consortium (IKMC)

The purpose of this publication is to explain the most important rules from **rulebooks 2 and 4** in view of the sharp increase that has occurred in the large number of mutant mouse and rat strains, especially with the use of methods for the genetic modification of the genome..

Not all the individual rules are comprehensively addressed here, but the rules that most users have to apply are listed. For the practical purposes, these are explained in abridged form at certain points. Detailed explanations can be found online in the Mouse Genome Informatics (MGI) database at www.informatics.jax.org.

General information

For the avoidance of misunderstandings, please note that this publication is limited to the nomenclature of mouse and rat genes and their mutations and does not cover the nomenclature of the proteins formed.

The use of upper case and lower case in abbreviations of genes helps to differentiate mouse and rat genes from the gene abbreviations of other species. Abbreviations of mouse and rat genes begin with an upper-case letter, followed by lower-case letters, while the genes of other species are abbreviated entirely in upper case, e.g. for humans, or entirely in lower case, e.g. for bacteria.

The proteins produced are abbreviated using only upper case, also if they are expressed in mouse or rat.

Laboratory codes

An important component of the nomenclature rules is the laboratory code (lab code), which uniquely identifies certain people, working groups, laboratories, institutes, or other institutions.

Laboratory codes usually consist of 3-4 letters, beginning with an upper-case letter followed by lower-case letters.

Examples

J The Jackson Laboratory

Unc University of North Carolina

Mpin Max Planck Institute for Neurobiology

Laboratory codes can be requested and obtained from the Institute for Laboratory Animal Research (ILAR) in Washington DC
(https://www.nationalacademies.org/ilar/lab-code-database)

Nomenclature of gene symbols and gene names

Gene:

A gene is a functional unit that typically codes for a protein or RNA, the heredity of which can be tracked experimentally.

The most important rules state that a gene symbol must

- consist of three to five characters
- only contain Latin letters and/or Arabic numerals
- begin with an upper-case letter, followed by lower-case letters (for rodents)
- not include tissue specification or molecular weights
- be italicized in publications

Note: italics and superscripts are lost in many databases.

Examples:

Plaur urokinase plasminogen activator receptor

Lep leptin

Lepr leptin receptor

With specific mutants, genes of other species are also frequently included. The most important databases for genes are:

- Mouse: www.informatics.jax.org
- Human (upper case): http://www.ncbi.nlm.nih.gov/omim
- Bacteria (lower case): http://bacteria.ensembl.org/index.html
- Rat: http://rgd.mcw.edu/
- Rabbit: http://www.ensembl.org/Oryctolagus_cuniculus/Info/Index
- Vertebrates: http://vega.sanger.ac.uk/index.html
- Chicken: http://www.ensembl.org/Gallus_gallus/Info/Index
- mi-RNA: www.mirbase.org

Nomenclature for names and symbols of alleles

Allele:

The two homologous genes of male and female autosomes are each referred to as an allele. An individual chromosome can only carry one allele (exceptions: deletions and duplications). Each autosomal gene has two alleles (exception: trisomies).

The following terms are used to define the zygosity of alleles:

Both alleles identical = homozygous (homo = same)

Both alleles different = heterozygous (hetero = different)

Only one allele present = hemizygous (hemi = half)

Attention: the term “hemizygous” is used both for alleles on sex chromosomes and for transgenes, because the donor organism of transgene inserts differs usually from the recipient organism and the sequences cannot therefore be classified as either “identical” to or “different” from the recipient. On the other hand, knock-out/knock-in (ko/ki) alleles can be heterozygous or homozygous.

The zygosity of a given allele is therefore not indicated in the nomenclature of the strain.

Different alleles of a gene can be differentiated using a range of methods, including those which are most frequently used at the DNA level:

- Restriction fragment length polymorphism (RFLP)
- Microsatellite polymorphism (STR or MIT)
- Single nucleotide polymorphism (SNP)

The most important rules for the nomenclature of alleles state that alleles must

- begin with letters
- contain superscript alphanumeric characters
- after the gene denomination
- be italicized
- reflect/present the dominant/recessive mode of inheritance using upper- and lower-case letters (dominant: uppercase; recessive: lower case)

Nomenclature of spontaneous mutations

Spontaneous mutations lead to alleles that differ from the wild type. These are identified as soon as they show a visible phenotype. This phenotype is often attributable to an initially unidentified mutation in an individual gene. Since the gene is unknown, a descriptive gene symbol is initially allocated. When the gene is identified, the descriptive denomination is withdrawn as a gene symbol and assigned as an allele.

Mutations that are spontaneous, induced or generated using targeted genetic modification lead to alleles of the gene concerned, i.e. they are identified using superscript characters placed after the gene locus.

If the wild type is to be explicitly named, the wild-type allele is represented by a plus sign (+).

Example:

Lepr^{db} = mutated allele (“db”) of the gene locus leptin receptor

Lepr^{+} = wild-type allele of the gene locus leptin receptor

When spontaneous mutations are named directly in the strain in which they occurred, the corresponding gene locus is added directly to the name of the strain in which the mutation occurred with a hyphen, followed by the mutant allele.

Example:

129P3-Lepr$^{db\text{-}3J}$/J

- 129P3: genetic background strain
- Lepr: gene locus, symbol name
- db-3J: mutated allele, 3rd remutation, occurred at Jackson Laboratory
- J: laboratory code (J = The Jackson Laboratory)

Nomenclature of mutants created by genetic engineering techniques

Mutations created by genetic engineering techniques include transgenic animals and targeted mutants.

Put simply, classical transgenesis involves the introduction of additional genetic material to the recipient genome by pronuclear injection, whereas targeted mutations involve the modification of an endogenous gene by homologous recombination in ES cell lines.

Genetic material is inserted into the target genome randomly by pronuclear injection.

The two paths for generating mutants through genetic engineering techniques (transgenic and homologous recombinants) are differentiated in the nomenclature.

Nomenclature of transgene inserts

The denomination of the transgene insert consists of four parts:

- The abbreviation Tg for “transgene”
- The denomination of the inserted gene or DNA segment in parentheses
- The founder line number or a serial number, in rare cases a letter
- The laboratory code of the original laboratory.

Example:

Tg(CAG-EGFP)1Osb

- **Tg**: transgene
- **(CAG-EGFP)**: official gene symbol of the inserted DNA construct

Note: the denomination of the promoter (in the above example: CAG) must be added in the case of lines in which specific expression is achieved through a promoter (as gene construct). If the complete gene is inserted with its regulatory sequences/elements, it is not necessary to mention the promoter.

1: Founder line number

It can be assumed that in each (founder) animal that develops after the pronuclear injection and carries the injected transgene, the transgene is integrated at a different site, because integration of the foreign DNA in unforeseeable DNA segments of the recipient zygote is purely random. This means that endogenous sequences can occur with differing influences described as position effects and that the transgene expression pattern differs in each of these animals. Each animal thus establishes an individual founder line, which is bred separately. Since a new line is established (founder line) after pronuclear injection, these lines must be bred separately from one another and characterized in parallel.

Osb: Laboratory code of original laboratory (Dr Masaru Okabe, Osaka University)

Parallel founder lines are differentiated with reference to the differing founder numbers.

Example:

All founders carry the same (human) transgene insert (BCL2) with the Emu enhancer, SV40 Promoter, but show different patterns of expression:

- C57BL/6-Tg(BCL2)22Wehi/J B cell lines
- C57BL/6-Tg(BCL2)25Wehi/J T cell lines
- C57BL/6-Tg(BCL2)36Wehi/J B and T cell lines

Insertion in the X chromosome, for example, has a strong position effect, leading to mosaic patterns of expression in female offspring because an X chromosome is randomly inactivated in somatic cells.

Differing transgene founders and resulting transgenic models that contain the same promoter and the same gene should also be denominated by the same gene symbol in parentheses and differentiated by means of a different founder or serial number. In cases where the transgene model was generated in a different laboratory of origin, the lines are additionally differentiated by a different laboratory code.

Full information on the structure of the transgene insert should be documented in publications or appropriate databases.

With the insertion of **fusion genes** in which both genes have approximately the same proportions of the translated protein, both genes are shown joined by a forward slash.

Example:

Tg(CAG-cre/Esr1)5Amc describes a transgene insert in which a fusion gene has been inserted that consists of cre recombinase and modified mouse “estrogen receptor ligand binding domain” (Esr) controlled by the CAG promoter.

When two transgene inserts are co-inserted, the two transgene inserts are shown separated by a comma.

Example:

Tg(HLA-B*2705, B2M)33-3Trg describes a transgene insert in which the human genes HLA-B*2705 and B2M have been co-injected.

With the random insertion of transgenic inserts, there is a possibility that they may be inserted into a gene resulting into its inactivation (ko). In these cases, the transgenic insert is shown as an allele of the (defective) gene concerned.

Example:

awg$^{Tg(GBtslenv)832Pkw}$ describes the phenotypically (awg beginning with lower-case letters!) observed mutation of an abnormal “wobbly gait”, which occurred as the result of a transgene insertion into founder line 832 in the laboratory of Paul Wong. Provided the abbreviation remains unequivocally clear, the abbreviated form awg^{Tg832Pkw} may be used.

Nomenclature of targeted mutations

In the case of targeted mutations, an endogenous gene is modified by homologous recombination in ES cell lines. In functional terms, very different modifications may be undertaken here. Targeted mutations include classical *knock-out* (ko), *knock-in* (ki) and conditional mutants.

It is crucial to the use of the nomenclature for targeted mutations that the technique of homologous recombination in ES cells has been used.

The denomination of a targeted mutation generated by means of ES cell technology consists of three parts:

- the letters tm for “targeted mutation”
- a serial number denoting the laboratory of origin
- the lab code of the laboratory of origin

Since it is not possible to differentiate on the level of the nomenclature used, whether the allele concerned is a knock-out, a knock-in or a conditional allele, the full information on the targeted mutation should be documented in publications or appropriate databases.

Both of the alleles below are targeted modifications in the murine Bcl2 gene, but were generated by different laboratories and have different functional modifications. The short descriptions can be found in full in the MGI database:

Examples:

$Bcl2^{tm1Sjk}$	Sjk = Stanley J. Korsemeyer
$Bcl2^{tm1Mpin}$	Mpin = Max Planck Institute for Neurobiology
^{tm1Sjk}	targeted mutation 1, Stanley J Korsmeyer

A 1.1 kb genomic fragment with the complete coding region of **Exon 3 was replaced by a neomycin selection cassette**. The authors report that an incomplete protein was found *in vitro* or *in vivo*.

$tm1^{Mpin}$	targeted mutation 1, Max Planck Institute for Neurobiology

Exon 2 was interrupted by the insertion of a lacZ gene and a neomycin cassette. Western Blot analysis of the hippocampus samples from mice with homozygous mutations showed that no stable Bcl2 protein from this allele was translated. Instead, **beta-galactosidase is transcribed under the control of the endogenous promotor of this allele**.

Further examples:

Bcl2^{tm1Mpin} This is a knock-out allele

- ***Bcl2***: gene locus, symbol name
- ***tm1***: targeted mutation, allele #1
- ***Mpin***: lab code of the laboratory of origin (Max Planck Institute for Neurobiology)

Hdh^{tm4Mem} This is a knock-in allele

- ***Hdh***: gene locus, symbol name (Huntington's disease gene homologue)
- ***tm4***: targeted mutation, allele #4 (92 CAG repeat units in the first exon)
- ***Mem***: lab code of the laboratory of origin (Dr Marcy MacDonald, Massachusetts General Hospital)

Scn5a$^{tm1(SCN5A)Rdn}$ This is a knock-in allele

$^{tm1(SCN5A)Rdn}$. targeted mutation, allele #1. A recombinase-mediated cassette exchange was used to replace exon 2 with the full human SCN5A sequence, which is shown in parentheses. By means of RFLP genotyping it was confirmed that only a human transcript was expressed in the hearts of the mutants.

Pparg^{tm2Rev} This is a conditional allele

- Pparg: gene locus
- tm2: targeted mutation, allele #2
- Rev: lab code of the laboratory of origin

Nomenclature of conditional models, with recombinase systems, such as cre/loxP or Flp-FRT

These systems are two-component systems:

One mouse line contains the genes for the expression of cre recombinase (causes recombination) from the bacteriophage P1 under the control of a given promoter. This is frequently inserted into the genome as a transgene (random integration!) and thus follows the "Tg" rules. Alternatively, cre-expressing lines may also be generated using knock-in technology.

In the second mouse, the target gene is flanked by recognition sites for the respective recombinase (l*oxP*, **l**ocus **o**f cross-over (**x**), in the bacteriophage, **P**1). These recognition sequences are inserted into the target gene by homologous recombination. The nomenclature thus follows the rules for targeted mutations (tm). Depending on the positioning of the recognition sites, mating with a cre reaction will result in the deletion or inversion of the flanked gene segment in the F1 generation.

By mating of both mice, new models can be produced. In this case,

the knock-in of the recognition site follows the rules for targeted mutations (tm).

If a new, heritable allele was generated through mating with a cre recombinase-expressing mouse line, the tm nomenclature is retained and a serial number is added.

If mating with a recombinase-expressing mouse leads to somatic changes in the offspring with no germline transmission, the tm nomenclature is retained unchanged. No new nomenclature is issued for the tm allele.

Example: *Tfam^{tm1Lrsn}* and *Tfam$^{tm1.1Lrsn}$*.

In this example, *Tfam^{tm1Lrsn}* denotes the targeted mutation in which loxP recognition sites were inserted into the Tfam gene.

Tfam$^{tm1.1Lrsn}$ denotes a new allele showing <u>germline transmission</u> to offspring, which was generated through mating with a recombinase-expressing mouse.

Analogous procedures are used with other recombinase systems, e.g. Flp-FRT.

Nomenclature of endonuclease-induced mutations

Endonuclease-induced mutations are targeted mutations produced in pluripotent or totipotent cells using e.g. zinc finger, TALEN or CRISPR/Cas technology. The mutation arises as a result of homologous or also non-homologous DNA repair after induced DNA strand breaks.

The denomination of a mutation generated by means of endonuclease consists of three parts:

- The letters "*em*" for "endonuclease mutation"
- A serial number for the laboratory of origin
- The lab code for the laboratory of origin

Example:

Fgf1^{em1Mcw} denotes the first endonuclease-induced mutation of the fibroblast growth factor 1 (Fgf1) gene, produced at the Medical College of Wisconsin.

Since it is crucial for the assessment according to Germany's Gene Technology Act (GenTG) to establish which Endonuclease technology was used and whether any foreign DNA remains in the organism (-> genetically modified organism [GMO]) or not (-> no GMO), it is to be expected that the official international nomenclature rules will be adapted. Until such time, it is recommended for organisms in which there is any remaining foreign DNA be identified by including the remaining DNA in parentheses, similar to the practice commonly used today in gene trap alleles, e.g. *Fgf1*$^{em1(EGFP)lab\ code}$; see also below.

Nomenclature of gene trap mutations

Gene trapping is a high-throughput procedure to introduce insertion mutations into the mammalian genome. These mutants are often developed for test procedures.

The denomination of a gene trap mutation consists of four parts:

- The abbreviation *Gt* for gene trap
- Indication in parentheses of the vector used
- A serial number for the laboratory of origin
- The lab code for the laboratory of origin

In those cases where the gene with the gene trap mutation is known, the gene trap mutation is written as an allele of the gene concerned.

Example:

Akap12$^{Gt(ble\text{-}lacZ)15Brr}$ denotes the 15th gene trap allele of the Akap gene, generated in the laboratory of Jaqueline Barra (Brr). The gene trap vector contains a phleomycin resistance gene (ble) and the LacZ gene.

In cases where the gene with the gene trap mutation is not known, the denomination consists only of the four elements mentioned above.

Example:

Gt(ROSA)26Sor denotes the 26th gene trap mutation with the vector ROSA, generated in the laboratory of P. Soriano.

Nomenclature for targeted mutations of genetically modified mice generated in high-throughput procedures

In 2007, the International knock-out Mouse Consortium (IKMC) was formed with the aim of generating knock-out mutants of all known mouse genes and regulatory control elements.

To obtain the most precisely defined genetic background possible, it was stipulated that only specific ES cell lines and mouse strains should be used, including the **ES cell line E14TG2a (from 129P2/OlaHsd)** and **ES cell lines from C57BL/6N (e.g. JM8.F6)**.

The strategy was developed to generate a “starting allele”, from which further alleles can be produced through mating with corresponding recombinase-expressing strains.

The “starting” alleles generated are therefore complex and usually contain several elements, in most cases a combination of FRT and loxP recognition sites, so that various specific sequences can be selectively cut out of the genome by mating with cre or Flp-expressing strains.

This complexity of the further modification options of the “starting” allele means that the conventional nomenclature used e.g. in the mating of a conditional mutant with a ubiquitously cre-expressing mouse does not offer a sufficiently nuanced denomination system to allow a differentiation of the various recombined alleles that are produced.

Nomenclature rules have therefore been defined that apply specifically to the alleles generated by the project groups involved in the International knock-out Mouse Consortium (IKMC).

http://www.informatics.jax.org/mgihome/nomen/IKMCnomen.shtml

The International knock-out Mouse Consortium (IKMC) includes the following projects:

- Knock-out Mouse Program (KOMP)
- Texas Institute for Genomic Medicine (TIGM)
- North American Conditional Mouse Mutagenesis (NorCOMM)
- European Conditional Mouse Mutagenesis (EUCOMM)

Each project involves several institutions. Alleles produced by them are assigned to the respective institutions, as also with the conventional nomenclature rules, based on the lab codes.

The following lab code abbreviations are currently listed:

IKMC abbreviations

Nomenclature abbreviation	IKMC project	Assigned lab codes
KOMP	Knock-out Mouse Project (KOMP, USA)	Mbp, Wtsi, Vlcg
EUCOMM	European Conditional Mouse Mutagenesis (EUCOMM) & EUCOMMMtools (Europe)	Hmgu, Wtsi
NCOM	North American Conditional Mouse Mutagenesis (NorCOMM, Canada)	Cmhd, Mfgc
TIGM	Texas A & M Institute for Genomic Medicine (TIGM)	Tigm

Double and multiple mutants

Multiple mutants are often generated by mating different single mutants. It must be kept in mind here that the individual models used for these matings might differ in their genetic background, and their contribution to the new model should be reflected in the resulting genetic background denomination of the newly created model, at the beginning of the full model name.

The individual mutations or transgenic inserts are then placed one behind the other without further punctuation, and the lab code the generating laboratory is also retained.

Since these new mutants were obtained by mating different pre-existing models, the laboratory that performed the mating and hence generated the new mutants is named at the end after a slash.

Example:

B6;D2-Tg(HPV11-lacZ)1704Aal Tg(UBC-HPV11E2)613Josc/Josc

This shows that the new double mutant was generated with mutants Tg(HPV11-lacZ)1704Aal and Tg(UBC- HPV11E2)613Josc in the laboratory of Johannes Schenkel (Josc).

The genetic background is still mixed and composed of the genetic background of the two individual models, used to generate the double mutant model: B6 and D2.

Unusual denominations

The nomenclature rules have been repeatedly changed over the years. The denominations mentioned here are normally not used any longer, but they remain in the literature and therefore it is helpful to be familiar with them:

- TgN for overexpressors
- TgH for homologous recombinants

- e for embryo transfer
- f for foster breeding
- h for in-hand breeding
- o for ovary transplantation
- p for cryo-preservation

Gene ID

A great many mutant lines are filed at www.informatics.jax.org which also have a Gene ID in addition to the correct nomenclature. This Gene ID is not an integral part of the nomenclature rules but should be mentioned in more detailed records for the purpose of unequivocal identifiability of the mutants.

Synonyms

Over the years and with the continual updating of the nomenclature rules, there have been repeated changes in the way certain genes are denoted. When new denominations are generated, it is important to use current names and abbreviations. In many databases, including the Mouse Genome Informatics (MGI) database, versions that are no longer in use can be found as synonyms.

Disclaimer

Any use of GV-SOLAS publications (specialist information, statements, booklets, recommendations, etc.) and application of the information contained therein are at the express risk of the user. Neither GV-SOLAS nor also the authors can accept liability for any accidents or damages of any kind arising from the use of a publication (e.g. resulting from the absence of safety instructions), irrespective of legal grounds. Liability claims against GV-SOLAS and the author for damages of a material or non-material nature caused by the use or non-use of the information or by the use of erroneous and/or incomplete information are in principle excluded. Legal claims and claims for damages are therefore excluded. The work, including all content, was compiled with utmost care. However, GV-SOLAS and the authors assume no responsibility and no liability for the currentness, correctness, completeness or quality of the information provided or for printing errors. GV-SOLAS and the authors accept no legal responsibility or liability in any form for incorrect statements and consequences arising therefrom. Responsibility for the content of the internet pages printed in these publications lies solely with the owner of the websites concerned. GV-SOLAS and the authors have no influence on the design and content of third-party websites and therefore distance themselves from all third-party content. Responsibility within the meaning of press legislation lies with the board of GV-SOLAS.