GUIDELINES FOR THE NOMENCLATURE OF GENETIC ELEMENTS IN TUNICATE GENOMES.

Preliminary version July 22, 2013 D. Dauga, P. Lemaire

Definitions of genomic sequence features for which the guidelines apply:

The following definitions are mostly derived from Sequence Ontology (SO) and extended.

Gene (SO:0000704):

"A region (or regions) that includes all of the sequence elements necessary to encode a functional transcript. A gene may include regulatory regions, transcribed regions and/or other functional sequence regions".

Types of genes: Genes can be encoded by the nuclear or the mitochondrial genomes. They can also correspond to Transposable element sequences. They can code for proteins, be non-coding or antisense with respect to another gene. A gene expresses one or more non-coding RNA genes (ncRNA) or protein-coding sequences (CDS).

- Protein coding gene (SO:0001217) : "A gene which codes for a protein."
- rRNA gene (SO:0001637) : A gene that encodes for ribosomal RNA.
  tRNA gene (SO:0001272) : A gene that encodes for transfer RNA.
  lncRNA gene (SO:0001877): A non-coding RNA over 200nucleotides in length.
  snRNA gene (SO:0001268) : A gene that encodes for small nuclear RNAS.
  snoRNA gene (SO:0001267) : A gene that encodes for small nucleolar RNAs.
  miRNA gene (SO:0001265) : A gene that encodes for microRNAs.
  mt_gene (SO:0000088): A gene located in mitochondrial sequence.

Pseudogene (SO:0000336):

A sequence that closely resembles a known functional gene, at another locus within a genome, that is non-functional as a consequence of (usually several) mutations that prevent either its transcription or translation (or both). In general, pseudogenes result from either reverse transcription of a transcript of their "normal" paralog (SO:0000043) (in which case the pseudogene typically lacks introns and includes a poly(A) tail) or from recombination (SO:0000044) (in which case the pseudogene is typically a tandem duplication of its "normal" paralog).

Transcript (SO:0000673):

An RNA synthesized on a DNA or RNA template by an RNA polymerase. Several transcripts with alternative structures can be produced by a single gene though the process of alternative splicing.

Transcriptional *cis*-regulatory region (SO:0001055):

A regulatory region that modulates the transcription of a gene or genes.

Operons (or polycistronic genes) (SO:0000178):

A group of contiguous genes transcribed as a single (polycistronic) mRNA from a single regulatory region.

Transposable element (SO:0000101):

A transposon or insertion sequence. An element that can insert in a variety of DNA sequences. A transposon may contain the genes necessary for its transposition. For example gag, int, env and pol are the transposable element genes of the TY element in yeast.

Transgene (SO:0000902):

A transgene is a gene that has been transferred naturally or by any of a number of genetic engineering techniques from one organism to another.

Assembly (SO:0001248):

A region of the genome of known length that is composed by ordering and aligning two or more different regions. An assembly is characterized by the lab that made it, a date and possibly a publication.

In addition to genetic elements, the document specifies naming rules for **transgenic lines and constructs** and for **mutant lines** :

Transgenic line:

A line of organisms derived from a common parent that has been modified by heritable transgenic insertion (**SO:0000781):** An insertion that derives from another organism, via the use of recombinant DNA technology.

Mutant line:

A line of organisms derived from a common genetically modified parent, through classical mutagenesis or transgenic insertion. Mutant lines are defined by their genetic alteration, their genetic background and their phenotype.

Constructs

A construct is an engineered plasmid (SO:0000637) that carries various features including for example a transgene (SO:0000902), a targetting vector (SO:0001644), a reporter gene, etc.

Not covered in this document: alleles and natural variation, mutations either natural or experimental.

GENE NOMENCLATURE

Functions of gene nomenclature are to:

1. Unambiguously identify a gene.
2. Identify the gene as a member of a family, which may give further information about gene function by reference to other family members
3. Identify the gene as the ortholog of a gene in another mammal (usually human)
4. Identify the gene as involved in a novel phenotype or trait

Gene nomenclature should be registered at XXX (NISEED ? need to centralize all these infos).

Elements in a gene name

- Gene identifier: This is a stable identifier distinct from the internal DB ID and which is guaranteed to follow the gene throughout any changes that may be made to its structure. A possible syntax would be for a *Ciona intestinalis* coding gene (CG): CiCG0004567. As these genes are generated automatically and independently for each species, there is no reason that CiCG0004567 and PmCG0004567 should be orthologous. Non coding genes should be identified according to the same syntax, but changing the middle characers: eg CitRNA0000085, CimiRNA0000054, …..

- Gene model name: Derived from the contig/scaffold on which they reside (e.g. in *Ciona intestinalis* KH2012:KH.C1.841). This name can change with time as novel information/analyses becomes available. Different species can have different syntaxes, reflecting the state of advancement of the assembly.
- Gene symbol: A short-form representation/abbreviation of the descriptive gene name, unique within the species. Usually 3-5 characters long, no more than 10. (e.g. *CYB5D1*). Capital letters should be used. Gene symbols are italicized. The same symbols not italicized represent proteins. Orthologous ascidian genes have the same symbol.

  Ascidian gene symbols should not be preceded by a mention of the species (*'Ci'*, '*Pm'*, '*Hr'*) in current use, except when the distinction between orthologous genes in different ascidians needs to be made. In this case, a species prefix (e.g. *Ci-*) is added in front of the gene symbol.

  The use of punctuation such as period and hyphens in gene symbols is discouraged, except under specific circumstances described below.

  Ascidian symbols follow the following conventions. When a *Ciona* gene has a 1-to-1 mammalian ortholog, the mammalian symbol is adopted. In case of 1-to-many orthology relationships, the *Ciona* gene symbol reflects all human paralogs (e.g. *FGF9/16/20* is the single *Ciona* ortholog of human genes *FGF9*, *FGF16* and *FGF20*). For simplicity, when the mammalian paralogs have very differing symbols, a single "class" symbol is used for the *Ciona* gene (eg *Ciona CREB1* is orthologous to human CREB1, ATF1 and CREM). Finally in case of many-to-many orthology relationships, when independent duplication events have taken place in the two animal lineages, the *Ciona* paralogs were distinguished by –a, -b, etc suffixes (eg. *Ciona* genes *HES-a*, *HES-b*, *HES-c* and *HES-d* are orthologous to all Human *HES* genes).

  When a *Ciona* gene shows preferential similarity to a mammalian gene without robust orthology association, it inherits the vertebrate symbol name followed by the suffixes -*related* or *–like* (eg *Ciona PROP-related* is similar to mammalian *Prop* without being a confident ortholog).

  Genes that belong to a given protein family, but show only limited similarity to non-tunicate genes are considered tunicate-specific (TS). Their symbol is TS followed by the protein family name, followed by a number (e.g. *TS-bHLH2* is a tunicate specific bHLH gene).

Genes with no protein sequence similarity to any gene outside tunicates do not receive a symbol or name, unless they have been functionally characterized (e. g. *PEM* for *Posterior End Mark*).

- Gene name: A short sentence describing gene function or structure (e. g. cytochrome b5 domain containing 1).

- Gene synonyms: A gene can have several synonyms, which are names or symbols that have been applied to the gene at various times. These synonyms may be associated with the gene in databases and publications, but the established gene name and symbol should always be used as the primary identifier.

PSEUDOGENE NOMENCLATURE

Pseudogenes should be suffixed by -ps and a serial number if there are multiple pseudogenes. If only one pseudogene copy of a particular gene has been found, it should be given the suffix -ps1.

Examples (find ascidian specific example) :

In mouse, phosphoglycerate kinase 1, pseudogenes 1 to 7, Pgk1-ps1 to Pgk1-ps7

In rat, calmodulin pseudogene 1, Calm-ps1

TRANSCRIPT NOMENCLATURE

No absolute identifier

Name : GENE-SYMBOL.suffix, the suffix varies between gene model sets and species. It generally refers to the sequence accession ID or to the gene model set of reference (ex : otx. KH.C4.84.v1.A.SL5-1)

➔ What it defined in other species (zebrafish) and could be interesting for us :

[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 VIa, transcript variant 1, myosin VIa, transcript variant 2,

Symbols -myo6a_tv1 myo6a_tv2]

TRANSCRIPTIONAL *CIS*-REGULATORY REGION NOMENCLATURE

➔ Cis-regulatory region named in the tunicate community :

ANISEED :

ci-"common gene name" "start of region/end of region e.g. ci-Otx -4037/+31

The problem of this nomenclature is that the sequence can regulate multiple genes + do we have to name an hypothetical *cis*-regulatory region that, after experiment, has no effect on gene expression ? We currently name these regions in ANISEED with the nearest gene.

In addition, the problem with the ANISEED nomenclature is the TSS definition and corresponding relatives coordinates (problem when assembly is updated).

➔ What it defined in other species (rat) and could be interesting for us :

Enhancers, promoters, and regulatory regions can influence multiple genes. In addition, they can be localized far away from the gene(s) that they affect. Thus, it is misleading to name them based on the gene for which regulation was first recognized.

Enhancers, promoters, and regulatory regions are to be symbolized as:

Rr# regulatory region #

where # indicates the next number in the series

OPERONS (OR POLYCISTRONIC GENES) NOMENCLATURE

➔ Operons named in the tunicate community :

Ghost :

KHOP.807

➔ What it defined in other species (worm) and could be interesting for us :

- Groups of genes which are co-transcribed as operons are curated as Operon objects.
- These have names formed from 'CEOP' followed by a value for the chromosome (1,2,3,4,5,X) and a unique 3 digit number like CEOP5460 and are manually curated using evidence from the SL2 trans-spliced leader sequence sites.
  *CEOP : C. Elegans OPeron

TRANSPOSABLE ELEMENT NOMENCLATURE

➔ To discuss with specialist in the field

➔ Transposable element named in the tunicate community :

Cigr-1

Cili-1

Cili-2

Cics-1

Cimi-1

(Martin W. Simmen and Adrian Bird)

Tc1/mariner superfamily transposon, sleeping beauty (SB)

(Sasakura et al.)

➔ What it is defined in other species (mouse) and could be interesting for us :

The transgenic transposable elements are identified with a standard prefix Tg (for transgenic) and Tn (for transposable element). The class of transposable element may be included in parentheses. The general format of the symbol is:

TgTn(transposon_class_abbreviation-vector)#Labcode

Example: TgTn(sb-T2/GT2/tTA)1Dla

The symbol consists of:

- Tg denoting transgenic
- Tn denoting transposon
- In parentheses, a lowercase abbreviation of the transposon class (in this case sb for Sleeping Beauty), followed by a hyphen and the vector designation
- The laboratory's line or founder designation or a serial number
- The Laboratory Code of the originating lab

TRANSGENE NOMENCLATURE

➔ Transgene named in the tunicate community :

ANISEED :

pGene regulatory region pBasal promoter::Reporter gene (ex : pOtx -1541/-1417 pBra::NLSLacZ)

DBTGR :

pCi-G[alpha]i1-EGFP

pCi-synaptotagmin-EGFP

➔ What it is defined in other species (zebrafish) and could be interesting for us :

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.

Example: Tg(pitx2-002:GFP) In this case an internal pitx2 gene promoter that generates the pitx2-002 transcript is driving expression of GFP.

Regulatory sequence could be derived from either an enhancer or promoter, and is denoted by the symbol of the regulated gene or gene transcript. Regulatory or coding sequence fusions should be separated by hyphens.

Example: Tg(TetRE:Mmu.Axin1-YFP) In this case the construct has a fusion protein of mouse Axin1 and YFP under the control of a tetracycline response element.

Example: Tg(EPV.Tp1-Ocu.Hbb2:hmgb1-mCherry) In this case the construct utilizes six copies of the promoter from the Terminal protein 1 gene (Tp1) from the Epstein-Barr Virus (EPV), upstream of the rabbit (Ocu)
beta-globin (Hbb2) minimal promoter driving hmgb1 fused to mCherry.

Example: Tg(actb2:stk11-mCherry) In this case the construct has a fusion protein of stk11 and mCherry under the control of the actb2 promoter.

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:sptb-GFP)

Tg(-6.0hhex:sptb-GFP)

These examples represent two constructs containing a fusion protein of spectrin beta (sptb) and GFP driven by 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 kbp. 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.

Example:

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 drives GAL4, and UAS drives GFP.

Example: Tg(isl2b:GAL4,UAS:GFP)

When one promoter is used to drive more than one gene, a comma is used to separate the gene names. This includes uni- & bidirectional promoters.

Example: Tg(abhd2a:YFP,mCherry)

For those situations where a construct utilizes enhancers or promoters from genes that regulate two or more genes, only one of the genes should be represented in the name such that the gene with the lowest number or gene closest to the promoter is listed.

Example: Tg(dlx1aIG:GFP) This construct utilizes intergenic (IG) regulatory elements of dlx1a and dlx2a to drive expression of GFP. In this case the lower numbered gene was listed in the name.

Example: Tg(zic4:Gal4TA4, UAS:mCherry) This construct utilizes an enhancer of the zic1 and zic4 genes to drive expression of Gal4TA4, with an additional cassette that has UAS driving mCherry expression. In this case the gene closest to the enhancer was listed in the name.

For those cases where a gene from a different species is used, the three letter abbreviation should be used (Homo sapien (Hsa), Mus musculus (Mmu), Salmo salar(Ssa)) followed by a period then the gene symbol. For human genes use the standard gene symbol conventions of all capital letters. For mouse and other species the first letter of the gene is capitalized.

Example: Tg(Hsa.FGF8:GFP) Here the promoter of the human FGF8 gene is driving expression of GFP.

Example: Tg(Ssa.Ndr2:GFP) Here the promoter of the salmon Ndr2 gene is driving expression of GFP.

Transgenic constructs using modified clones such as BACs and PACs should be named with the clone type inserted between the "Tg" and the "("

Example: GFP is inserted to replace the coding sequence of tal1 in the PAC with the GenBank# AL592495. The construct name would be TgPAC(tal1:GFP). The accession number of the clone must be included in the publication so that it can be associated with the construct. A link to the appropriate clone can be found on the construct page.
Enhancer, promoter, and gene-trap constructs may use Et, Pt, or Gt, all of which are considered types of transgenic constructs.