Table 1
Classes of Genes Defined in Sequence Ontology and Corresponding Class Descriptors in Unique Gene Identifiers

Class of gene	Sequence ontology ID	Sequence ontology definition	Class descriptor
Protein-coding gene	SO:0001217	A gene that codes for a protein	CG
Ribosomal RNA gene	SO:0001637	A gene that encodes a ribosomal RNA	rRNA
Transfer RNA gene	SO:0001272	A gene that encodes a transfer RNA	tRNA
Non-coding RNA gene			ncRNA
Long non-coding RNA gene	SO:0001877	A non-coding RNA over 200 nucleotides in length.	lncRNA
Long intergenic non-coding RNA gene	SO:0001463	A multiexonic non-coding RNA transcribed by RNA polymerase II.	lincRNA
Small nuclear RNA gene	SO:0001268	A gene that encodes a small nuclear RNA	snRNA
Small nucleolar RNA gene	SO:0001267	A gene that encodes a small nucleolar RNA	snoRNA
Micro RNA gene	SO:0001265	A gene that encodes a microRNA	miRNA
piwi-associated RNA gene	SO:0001035	A small non coding RNA, part of a silencing system that prevents the spreading of selfish genetic elements.	piRNA
Enhancer RNA	SO:0001870	A short ncRNA that is transcribed from an enhancer. May have a regulatory function.	eRNA
Mitochondrial gene	SO:0000088	A gene located in mitochondrial sequence	mt
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).	ps

discriminative letter from the generic name (Pv for *Pycnoclavella*). In rare cases, there are too many genera starting with the same first letter to define unambiguous two-letter abbreviations for all genera (e.g., the names of 42 genera start with a P). In such case, some genera receive a three-letter abbreviation (e.g., Prm for *Protomolgula*). There should be no ambiguities between two-letter and three-letter abbreviations. While there can be Prm for *Protomolgula* and Prh for *Protoholozoa*, there is no genus whose two-letter abbreviation is simply Pr, as this would cause confusion as to whether a binomial abbreviation follows the 2G4S or 3G3S formula.

Abbreviations of specific epithets are usually four letters long (2G4S rule), except if they belong to a genus abbreviated with three letters, in which case their abbreviations are also three letters long (3G3S rule). Specific epithet abbreviations are usually built from the first three or four letters of the specific name (e.g., inte for *intestinalis*), except when this results in ambiguities within a genus (e.g., *Pyura squamata* vs. *Pyura squamulosa*). In these cases, a unique species abbreviation can incorporate discriminative letters from the specific epithet, which maximize intuitive reading (e.g., sqma for *squamata* vs. sqml for *squamulosa*). In cases of 2G4S abbreviations, when the specific epithet only has three letters, a letter from the generic epithet is added before the species abbreviation (e.g., *Didemnum ahu* = Didahu). Some prefixes are frequently used in specific epithet, and can be systematically abbreviated in a standard way. These prefixes are currently used: poly- (abbr.: py-); psam- (psm-); pseudo- (ps-); longi- (l-); multi-(m-); trans- (tr-). Examples: *pseudogrisiatum* is abbreviated as psgr, *translucidum* becomes trlu; and *polyducta* becomes pydu. These abbreviations of frequently used prefixes do not apply to generic epithets.

Identical or very similar specific epithets in different genera can share a common abbreviation (or at least the first three letters of the abbreviation, for species following the 3G3S rule), as long as this does not result in identical, conflicting specific abbreviations within a genus. In this case, one of the conflicting species abbreviations should be modified accordingly.

GENES

A gene is defined in Sequence Ontology (Eilbeck *et al.*, 2005) as "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" (Sequence Ontology term SO:0000704, Table 1). A gene can encode one or more proteins, or one or more non-coding RNAs.

A precise gene nomenclature system should be able to:

1. Unambiguously and stably identify a gene, across the successive releases of genome assemblies and gene builds for a given species.
2. Track the history of successive gene models.
3. Identify the gene as the ortholog, or a close relative, of a Human gene in order to facilitate a connection

to available information on gene function, gathered from the larger corpus of biomedical research.
4. Reflect the belonging to a structural gene family or the phenotype obtained following perturbation of the gene activity.

Finally the evolution of a gene name should be traceable, when additional knowledge builds up on the function of the gene in tunicate or other species.

To achieve these goals, a tunicate gene is defined by a combination of a unique gene locus identifier, a gene model identifier, a primary gene name and primary symbol, and synonymic gene names and symbols.

Unique Gene Locus Identifier

This is a stable identifier, which is guaranteed to follow the gene throughout any changes that may be made to its structure. Locus identifiers are composed of a species prefix and an eight-digit number. The unique identifier does not provide information on the type of the gene (coding, non-coding, etc.).

Example:

- Cisavi.00009682 defines the *Ciona savignyi* gene number 00009682, which, in this case, is a coding gene.

As these identifiers are generated automatically and independently for each species, there is no reason that the genes Ciinte.00004567 and Phmamm.00004567 should be orthologous.

Genes are initially identified on the basis of ab initio prediction and experimental evidence. As new evidence accumulates, the structure of a predicted gene can evolve with time, leading to the fusion of two predicted genes, to the split of a gene into two or more genes, or to the disappearance of the gene. Genes resulting from fusions or splits receive new unique gene identifiers, and a tracking system will link the new genes to their precursors. In such cases, a suffix is added to the unique gene identifier, indicating a change of status. The suffix _suppr[year] indicates that the gene has been suppressed from the species gene list in the indicated year. The suffix _split[year] indicates that the gene has been split into two or more genes in the indicated year. The suffix _fused[year] indicates that the gene has been fused with one or more other genes in the indicated year. In case of complex patterns of splits and fusions, fusions are considered dominant.

Examples:

- Ciinte.00000657_suppr2013 means that this *Ciona intestinalis* gene has ceased to be considered a gene in 2013.
- Cisavi.00008765_split2011 indicates that this *Ciona savignyi* gene has been split into two or more genes in 2011.
- Boschl.00012998_fused2014 indicates that this *Botryllus schlosseri* gene has been fused with one or more other genes in 2014.

Gene Model Identifier

This identifier characterizes a particular instance (model) of a gene in a given assembly. It is composed of a species prefix, a descriptor of the class of gene considered (CG for a protein-coding gene, tRNA, miRNA, eRNA. Table 1 lists accepted Sequence Ontology gene class descriptors), a reference of the gene build and assembly considered, and a unique identifier that specifies on which contig/scaffold/chromosome the gene is located. This identifier can change with time as novel information/analyses become available. Different species can initially adopt different syntaxes for the gene build and assembly identifiers as exemplified below, although an effort should be made to unite these syntaxes in genomes that are considered stable.

Examples:

- Ciinte.CG.KH2012.C1.841 defines the *Ciona intestinalis* coding gene of KH assembly, gene build 2012, "ranked" 841 on Chromosome 1. Note that in *C. intestinalis*, genes on a given chromosome/scaffold are "ranked" in arbitrary order, and genes with successive ranking numbers are not necessarily neighbors on the chromosome.
- Cisavi.miRNA.ENS75.R16.3924783-3924863 defines the *Ciona savignyi* miRNA gene of built ENSEMBL release 75 located on Reftig (scaffold) 16 between coordinates 3,924,783 and 3,924,863. Note that in *C. savignyi*, genes are identified by coordinates on a scaffold.
- Boschl.CG.Botznik2013.botctg020918.g2519 defines the *Botryllus schlosseri* protein coding gene reference g2519, located on the contig 020918 of assembly Botznik2013. In *B. schlosseri*, genes are identified by their rank in the global gene list.

Primary Coding Gene Name and Symbol

The name of a gene is a word or short phrase describing the structure of a gene (e.g., *Cytochrome b5 domain containing 1*), its orthology group (*Orthodenticle homeobox*), its function (*Mitogen-activated protein kinase kinase 1/2*), or its expression pattern (*Posterior end mark*). The first letter of the first word is capitalized and names are italicized. The 1-to-1 tunicate orthologs should share a common gene name and symbol.

The primary symbol of a gene is a short-form representation/abbreviation of the primary gene name, unique within the species. Usually 3–5 characters long, no more than 10 (e.g., *Cyb5d1* is the gene symbol for *Cytochrome b5 domain containing 1*). The first letter of the symbol should be capitalized, and the other letters should be in lower case. The use of punctuation,

such as period and hyphens, within gene symbols is discouraged, except under specific circumstances described below. Gene symbols are italicized.

Tunicate gene symbols should generally not be preceded by a mention of the species symbol (e.g., Ciinte, Harore.), except when the distinction between orthologous genes in different tunicates needs to be made. In such cases, the species abbreviation (e.g., Ciinte), is added in front of the symbol and separated from it by a dot (".").

To facilitate comparison to vertebrates, tunicate gene names and symbols are preferentially named after their mouse or human orthologs, as defined by the HGNC project (Gray *et al.*, 2013), with a preference for the human ortholog(s) when they exist.

In the case of one-to-many orthology relationships between tunicate and human orthologs, the tunicate gene name/symbol reflects all human (or mouse) paralogs.

Example:

- *Fibroblast growth factor 9/16/20*—symbol *Fgf9/16/20*—is the single tunicate ortholog of the three human genes *FGF9*, *FGF16*, and *FGF20*.

To avoid excessively long names or symbols, the numbers are omitted if a single tunicate gene is orthologous to all human members of a subfamily.

Example:

- *Pitx* is the single tunicate ortholog of Human *PITX1, 2*, and *3*.

For simplicity, when the human (or mouse) paralogs have very differing names, a single "class" or widely accepted "family" name is used for the tunicate gene. The *Ciona intestinalis* choice for this name determines the names in subsequently sequenced tunicate genomes.

Examples:

- *Ciona intestinalis cAMP response element-binding protein 1* (*Creb1*) is orthologous to the three human genes *cAMP response element-binding* protein 1 (*CREB1*), *Activating transcription factor 1* (*ATF1*), and *cAMP Response Element Modulator* (*CREM*).
- In *Ciona intestinalis*, *Dan domain family member* (*Dand*) is the single ortholog of Human *Cerberus* (*CER1*), *Gremlin* (*GREM1*), and *Dan domain 5* (*DAND5*) genes.
- *Ciona intestinalis Otx* is the single ortholog of Human *OTX1*, *OTX2*, and *CRX* genes.
- *Ciona intestinalis Myogenic regulatory factor (Mrf)* is orthologous to the human Myogenic regulatory factor family members *Myogenic differentiation 1 (MYOD1), Myogenin (MYOG), Myogenic factor 5 (MYF5),* and *Myogenic factor 6 (MYF6).*

In rare cases, tunicate gene names and symbols can deviate from those officially used for the orthologous human genes, and reflect more closely the terms used to refer to the encoded protein in humans and/or orthologous genes in other vertebrates.

Example:

- The human gene named *"T, brachyury homolog (Mouse)"* encodes a protein whose recommended name is "Brachyury," according to the UniProt database (Jain *et al.*, 2009). The orthologous tunicate gene name is thus *Brachyury*, and its symbol is *Bra*.

In case of many-to-one or many-to-many orthology relationships, the tunicate paralogs are distinguished by ".{a–z}" suffixes. In some cases, the tunicate duplication will have taken place at the root of either the tunicate phylogeny or a specific sublineage (ascidians, salps, stolidobranchs, etc.) and the same suffices should be used for all orthologs within this branch. Because of the rapid evolution of tunicate genomes, precise orthology relationships are sometimes difficult to draw. Hence, while we suggest that orthologous tunicate genes receive the same suffix when possible, non-orthologous genes may receive the same suffix, especially in distant species. This rule may be subject to revision when a higher coverage of the taxon with sequenced genomes is achieved, and we have a better vision of the diversity and complexity of orthology relationships within families, orders, and genera.

Examples:

- The *Ciona intestinalis Ephrin-A* genes have undergone independent duplications in the vertebrate and ascidian lineages (Fig. 1A), giving rise to *Ciona intestinalis Ephrin-A.a* to *.d*, (symbols: *Efna.a to .d*) which are all orthologous to human *Ephrin-A1* to *5* (*EFNA1 to 5*, Fig. 1A).
- Hedgehog.*a* and *.b* are the two paralogous *Ciona intestinalis* orthologs of human *sonic*, *Indian*, and *desert hedgehog* genes.

When a tunicate gene shows preferential similarity to a mammalian gene without robust orthology association, it inherits the vertebrate name followed by the suffixes *-related* (*-r* in the symbol). We discourage the use of the suffix "-like" in this case because it occurs in some human gene names (e.g., *Fer3-like BHLH transcription factor*). Should several tunicate genes be preferentially similar to the same mammalian gene, they are distinguished by adding a ".{a–z}" suffix to their names and symbols.

Examples:

- *Ciona savignyi* prophet of Pit-1-related (Prop-r) is similar to human prophet of Pit-1 (PROP) without being a confident ortholog.
- In *Ciona intestinalis* several gene duplications have given rise to four paralogous *T-box six-related* (*Tbx6-r*) genes, all related to the single human *TBX6* gene. These genes are named *Tbx6-r.a, .b, .c, .d*. There are also four *Halocynthia roretzi Tbx6-r* genes, which

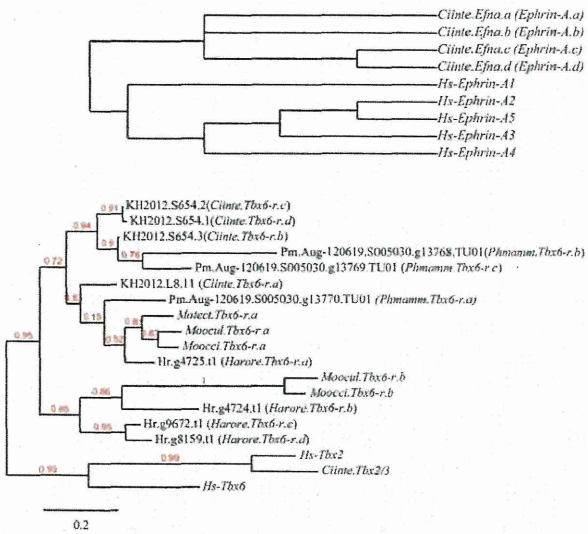

FIG. 1. (**A**) Phylogenetic tree for the *Ciona intestinalis* and Human *Ephrin A* gene family. (**B**) Phylogenetic tree for the tunicate *Tbx6-related* class of genes in *Ciona intestinalis*, *Halocynthia roretzi*, *Phallusia mammillata*, and *Molgula* spp., and their human (Hs-) relatives.

however do not have clear 1-to-1 *Ciona* orthologs. These genes are also named *Tbx6-r.a, .b, .c, .d*, but there is no strong inference that *Harore.Tbx6-r.x* is orthologous to *Ciinte.Tbx6-r.x*. (Fig. 1B).

- *Ciona intestinalis*, *Ciona savignyi*, and *Halocynthia* have two groups of *Zic-related* genes, which are called *Macho-1* and *ZicL* (or *ZicR*). All of these species have one copy of the gene initially called *Macho-1*, and this gene is named *Zic-r.a* (*Macho-1* remains as a secondary synonym, see below). *Ciona intestinalis* has five copies of *ZicL*, and they are named *Zic-r.b* to *Zic-r.f*. Similarly, *Halocynthia* has a single copy of *ZicR*, named *Zic-r.b*. Note that *Ciinte.Zic-r.a* and *Harore.Zic-r.a* are orthologous, but *Ciinte.Zic-r.b* and *Harore.Zic-r.b* might not be orthologous in a strict sense.

Genes that belong to a protein family, but show only limited similarity to non-tunicate genes are considered "tunicate-specific" members of this family. Their name is "Tunicate," followed by the protein family name, and by a numerical identifier. Numerical identifiers are unique within a protein family. A tunicate-specific gene may be present in some but not all tunicate genomes, in which case its numerical identifier remains unused in species that do not have this gene. Use of the term "orphan" in gene names and symbol is discouraged for genes without non-tunicate orthologs. The symbols of "tunicate-specific" genes are of the form [Family]tun[#].

Example:

- *Tunicate bhlh 2* (*Bhlhtun2*) is the tunicate bHLH gene number 2 with no characterized ortholog outside of tunicates. This gene is present in *Ciona intestinalis*, but not *Halocynthia roretzi*, and in this species tunicate bhlh identifiers skip 2 and jump from *Bhlhtun1* to *Bhlhtun3*.

Genes without significant protein sequence similarity (BlastP E-value >1e-5), or conserved domain, with any non-tunicate species do not receive a descriptive name, unless they have been functionally characterized in tunicates (e.g., *Posterior end mark*). They are named using their stable gene identifier.

Pseudogenes inherit the name of the functional gene from which they are derived, followed by the suffix -ps and a serial number if there are multiple pseudogenes. If only one pseudogene copy of a particular gene exists, it is given the suffix -ps1.

Gene Synonyms

A gene can have several "secondary" synonymous names or symbols, which differ from those that have been applied to the gene at various times. Such names, which often reflect the specialized function of a gene in tunicates, can be used in publications, provided the primary symbol is also mentioned. Secondary synonyms are important to ensure that databases will trace deprecated names to their current primary name and symbol. The current primary gene name and symbol should always be mentioned in new work. Primary gene names and symbols can evolve in time, for instance to reflect more accurate orthology relationships. In such cases, the modified primary names and symbols become synonyms to the new primary name and symbols.

Examples:

- *Macho-1* is a secondary synonym for *Zic-r.a*.
- *Hepatocyte nuclear factor 6 (Hnf6) is a secondary synonym for One cut homeobox (Onecut).*

Overlapping, Antisense, and Opposite Strand Genes

In general, genes whose transcripts partially overlap on opposite strands (for instance overlapping 3′ ends) should be given distinct names and symbols. This rule also applies to genes encoded at the same locus, in the same orientation, but using different reading frames.

A gene of unknown function running on the opposite strand to a coding gene, and included within this gene locus, should receive the name and symbol of the coding gene with the suffix "-os" for opposite strand.

Example:

- In *Ciona intestinalis* the *Bra-os* gene (KH2012:KH.S1404.3) is running on the opposite strand, and included within the *Ciona intestinalis Brachyury* locus (KH2012:KH.S1404.1).

A gene running on the opposite strand of a coding gene and known to regulate the function of this gene receives the name of the affected gene, with a suffix "-as" for antisense.

Non-Coding Gene Primary Names and Symbols

In addition to coding genes, many genes encode functional short and long non-coding RNAs, including tRNAs, snoRNAs, rRNAs, snRNAs, miRNAs, piRNAs... In human, these genes have recently been the focus of efforts to design a unique nomenclature (Wright and Bruford, 2011), from which the tunicate nomenclature is adapted. Primary names and symbols can be accompanied by synonyms, to relate the gene to more ancient nomenclatures. Our current knowledge on tunicate non-coding genes is partial, and some of the following rules will need refinement as this knowledge increases. Genes with clear 1-to-1 orthologs share the same name and symbol, one-to-many and many-to-many orthology rules are the same as for coding genes.

Tunicate nuclear tRNA gene names are of the form [Species] transfer RNA [amino acid] [#] (anticodon), where [Species] is the symbol of the species, usually omitted if the context is unambiguous, [amino acid] is the name of the amino acid, [#] is the serial number for the transfer RNA, and (anticodon) is the sequence of the anticodon. The corresponding symbol is of the form *trna[aa][#]*, where [aa] is the single letter abbreviation for the amino acid.

Example:

- *trnas1* is the symbol for the transfer RNA serine 1 (UGA) gene.

*Mito*chondrial tRNA gene are named according to the same logic, with the prefix "mt-."

Example:

- *mt-trnag* is the symbol for the mitochondrial transfer RNA Glycine (UCU) gene.

Tunicate small nucleolar RNA genes follow the nomenclature of snoRNABase (Xie *et al.*, 2007) (www-snorna.biotoul.fr) and are split into small nucleolar RNA, C/D Box (SNORD), small nucleolar RNA, H/ACA Box (SNORA), and small Cajal body-specific (SCARNA). In each species, their symbol follows the syntax: *{snord, snora, scarna}[#]{a-z}*, where [#] is a serial number describing the snoRNA family. SnoRNA genes are frequently duplicated and the different members of each family in a tunicate genome are distinguished by the *.{a-z}* suffix.

Example:

- *Ciinte.snord18.a* is the symbol of one of the five *Ciona intestinalis* small nucleolar RNA, C/D Box 18 gene, orthologous to all human SNORD18 (many-to-many orthology relationship). The others are *Ciinte.snord18.b*, *.c*, *.d*, and *.e*.

Tunicate ribosomal RNA gene names have the syntax [Species] (mitochondrial) ribosomal RNA {5, 5.8, 12, 16, 18, 26}S [#], where [#] is a serial number that reflects the presence of many copies of each ribosomal RNA gene. The corresponding symbol is *(mt-)rn{5, 5.8, 12, 16, 18, 26}s [#]*.

Examples:

- *Ciinte.rn18s2* is the symbol of the second nuclear ribosomal 18S RNA gene in *Ciona intestinalis*.
- *Harore.mt-rn16s* is the symbol of the *Halocynthia roretzi* mitochondrial ribosomal 16S RNA gene (synonym used in previous work: *mt-lrRNA* for large mitochondrial rRNA gene).

Small nuclear RNAs (snRNA or U-RNAs) are Uridine-rich small RNAs found in the nucleus of eukaryotic cells. The syntax of the symbol of spliceosomal snRNAs (U1, 2, 3, 4, 4atac, 5, 6, 6atac, 11, and 12) and of the U7 snRNA is *rnu{1, 2, 3, 4, 4atac, 5, 6, 6atac, 7, 11, 12}-[#]*, where [#] is a numerical identifier.

Example:

- *rnu1-1* is the symbol of the U1 snRNA 1.

Other small nuclear RNAs include vault RNAs (symbol *vtrna[#]*), 7SK RNA (*rn7sk*), 7SL RNAs (*rn7sl[#]*),Y RNAs that form part of the Ro RNP (*rny[#]*), and telomerase RNA component (*Terc*).

Piwi-interacting RNA form the largest family of expressed ncRNAs. piRNAs are designed by symbols with the syntax *pirna[#]*, where *[#]* is a serial number. piRNAs are often grouped in large clusters, which are themselves identified through symbols with the syntax: *pirc[#]* for PiRNA cluster [#], where [#] is a serial number.

Finally, tunicate miRNA genes are named according to accepted international standards used by MiRBase (Kozomara and Griffiths-Jones, 2014) and first published in 2003 (Ambros *et al.*, 2003). Briefly, full names are of the form Species *miRNA[#]-[#]*, where [#] is the accepted MiRBase number for this class of miRNA. Symbols of miRNA genes are of the form *mir[#]*. Mature miRNAs are distinguished from the gene they originate from with a capital R (miR[#]). miRNAs that encode homologous mature transcripts share the same mir number, with differing suffixes. If the mature miRNAs differ by only one or two nucleotides, they are allocated letter suffixes (e.g., mir10A and mir10B), whereas if the mature miRNAs are identical, the genes are given hyphenated numerical suffixes (e.g., *mir1-1* and *mir1-2*) (Wright and Bruford, 2011).

Examples:

- In *Phallusia mammillata*, *Phmamm.mir121* is the symbol of the *miRNA 121* gene, orthologous to the Human *miRNA 121* gene.
- *miR121* is the mature miRNA produced by the gene *mir121* in a tunicate species.

TRANSCRIPTS

A transcript is "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 and/or alternative promoter usage" (SO:0000673). Transcripts are defined by a transcript model identifier, a transcript name, and a symbol.

Transcript Model Identifier

The transcript model identifier is generally composed of its gene model identifier, a suffix that uniquely identifies each transcript model variant, and the ".t" suffix. The syntax of unique transcript suffixes may differ between species.

Examples:

- Ciinte.CG.KH2012.C4.84.v1.A.SL5-1.t is the transcript model variant v1.A.SL5-1 of *Ciona intestinalis* coding gene Ciinte.CG.KH2012.C4.84.
- Cisavi.CG.ENS75.R90.454100-458898.16640.t is the transcript model variant 16,640 of *Ciona savignyi* coding gene Cisavi.CG.ENS75.R90.454100-458898.

Transcript Model Name and Symbol

The transcript model inherits its gene name and symbols, followed by a suffix that distinguishes the transcript from the gene. No specific rules are established for the syntax of transcript suffixes in names and symbols, which can thus differ between species.

PROTEINS

Protein Model Identifiers

Proteins receive the same identifiers as the transcript they are produced from, not italicized, except that the ".t" suffix is replaced by ".p."

Example:

- Ciinte.CG.KH2012.C4.84.v1.A.nonSL3-1.p is the protein produced by the transcript variant Ciinte.CG.KH2012.C4.84.v1.A.nonSL3-1.t of the *Ciona intestinalis Otx* gene.

Protein Names and Symbols

When a single protein is produced by a gene, this protein inherits the name and symbol of the corresponding gene, without italics.

Example:

- T-box 6-related.d (Tbx6-r.d) is the sole protein produced from by the gene *T-box 6-related.d* (*Tbx6-r.d*).

As a simplified convention to refer to the different protein isoforms translated from different transcripts of the same gene, one may use the gene symbol followed by a suffix of syntax -i{#}, where i stands for isoform and # is an integer.

Example:

- In both *Ciona intestinalis* and *Halocynthia roretzi*, the *Lhx3/4* gene produces two protein isoforms (Christiaen *et al.*, 2009; Kobayashi *et al.*, 2010), which can be named Lhx3/4-i1 and -i2.

TRANSCRIPTIONAL *CIS*-REGULATORY REGIONS

A transcriptional *cis*-regulatory region is a segment of DNA, usually non-coding, that modulates the level of activity of one or more genes (SO:0001055). It is defined by a unique identifier, by its coordinates in a given genome assembly, by its type of activity and by optional target gene or genes, when known. Note that only experimentally tested *cis*-regulatory elements are identified and named, and that no inference is made about the precise boundaries of the functional *cis*-regulatory region in the genomic context.

Cis-Regulatory Region Unique Identifier

This is a stable identifier, which is guaranteed to follow the region in successive genome assemblies. It is composed of the species symbol, followed by REG and an eight-digit unique number (e.g., Ciinte.REG00000034). Regulatory regions with the same unique number in different species are not inferred to be orthologous.

Cis-Regulatory Region Identifier Within an Assembly

As regulatory sequences can regulate multiple genes, or the target genes may not be known, target gene name is not necessarily included to define a *cis*-regulatory region within an assembly. A *cis*-regulatory region identifier is of the form: [Species].REG.[assembly].[start-end](|[target gene symbol1]|[Target gene symbol2]|.), where [Species] is the species abbreviation, ".REG." indicates the class of element, [assembly] identifies the assembly within the species, [start-end] gives the coordinate of the region, and (|[target gene symbol1]|.) is optional and list experimentally determined targets of the *cis*-regulatory region by alphabetical order.

Examples:

- Ciinte.REG.KH2012.C1.289567-289760 is a regulatory sequence located on the Chromosome C1 of KH assembly of *Ciona intestinalis*, between positions 289,567 and 289,760. Note that the coordinates do not include commas (,) to avoid difficulties when parsing .csv formatted files.
- Ciinte.REG.KH2012.C4.4313996-4315697|*Otx* is a *Ciona intestinalis* regulatory sequence for the *Otx*

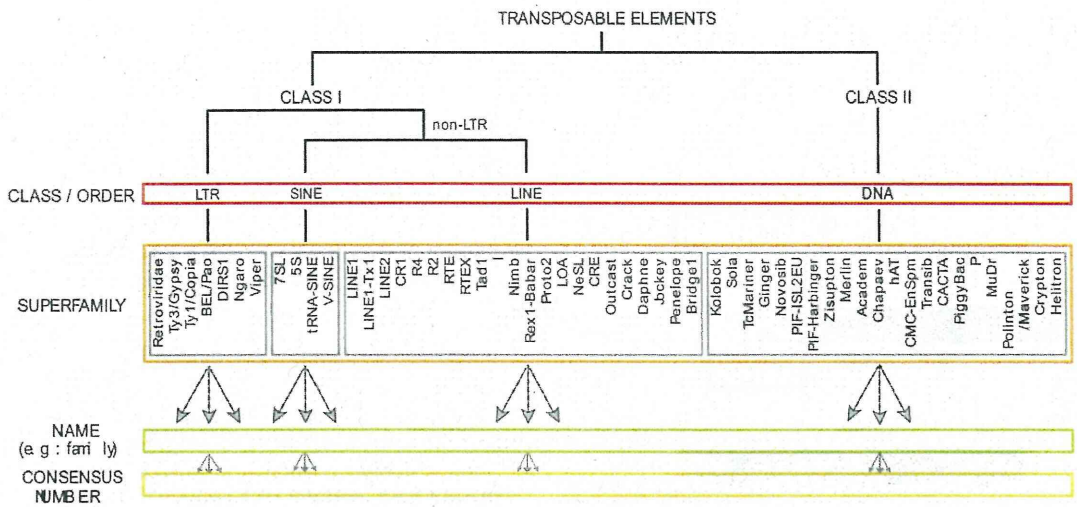

FIG. 2. Classification of tunicate transposable elements.

gene, located on Chromosome C4 between positions 4,313,996 and 4,315,697.

- Ciinte.REG.KH2012.C2.1981987-1982339|*Admp*|*Pinhead* is a *Ciona intestinalis* regulatory sequence controlling the expression of both *Admp* and *Pinhead* (Imai *et al.*, 2012).

Note that *cis*-regulatory element names found in scientific articles are usually shorter and more "biologically relevant" than the proposed nomenclature, but short names are difficult to generalize and are often legitimately biased to illustrate the main message of the article. The following guidelines can help authors build such short names in their manuscripts. Naming *cis*-regulatory sequences with respect to a gene transcription start site is discouraged unless the transcription start site has been precisely mapped in the tissue in which the *cis*-regulatory region is active. Naming regulatory sequences with respect to the start of the transcript extending furthest 5′ of the gene can be used in articles, provided the identity of the gene build is mentioned. In all cases, the materials and methods of the article should provide the *cis*-regulatory region identifier defined above.

Example:

- The name *Otx*[-1541/-1417] can be used in the main text of a scientific article dealing with *Ciona intestinalis*, but the full identifier of this sequence, Ciinte.REG.KH2012.C4.4315574-4315697|*Otx* should be mentioned in the materials and methods.

Often a "minimal" or "basal" promoter of one gene is used in combination with distal *cis*-regulatory elements from the same or a different locus. The minimal promoter is often defined as the sequence immediately flanking the site of initiation of transcription by RNA polymerase and is necessary, but not sufficient, for transcription. In scientific articles, minimal promoters may be indicated by the prefix "pr" attached to a gene symbol (e.g., *prOtx* or *prCiinte.Otx*), but the standard rules for naming of *cis*-regulatory region identifiers still apply to them.

OPERONS

An operon (or polycistronic gene) is defined as "A group of contiguous genes transcribed as a single (polycistronic) mRNA from a single regulatory region" (SO:0000178). As this may be difficult to assess, a relaxed definition can be used: two or more genes transcribed in the same orientation with no or very short intergenic sequences and evidence for trans-splicing of the downstream gene(s) (Satou *et al.*, 2008). Each Operon receives a unique Operon identifier.

Unique Operon Identifier

This identifier is stable across successive genome assemblies and gene builds. It is composed of a species identifier, followed by "-OP" and an eight-digit number.

Example:

- Ciinte.OP00000807. There is no assumption of orthology between operons with the same numerical identifier in different species.

TRANSPOSABLE ELEMENTS (TE)

A TE is "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 TE genes of the TY element in yeast" (SO:0000101).

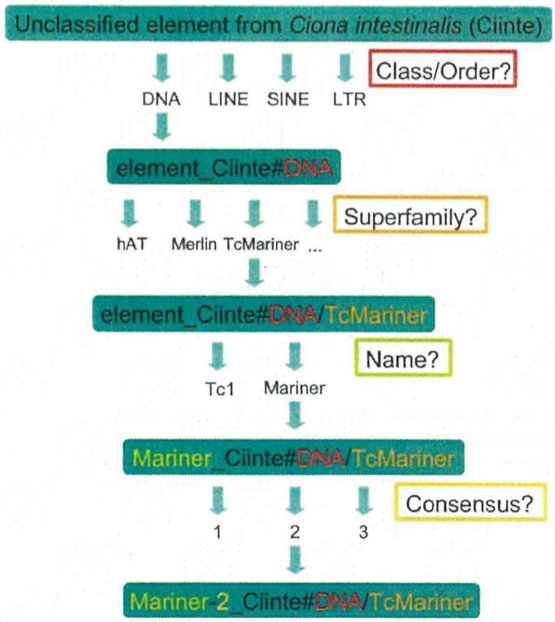

FIG. 3. Flowchart for the classification process for *Ciona intestinalis* transposable elements.

The nomenclature of tunicate TEs follows the international nomenclature (Kapitonov and Jurka, 2008; Wicker *et al.*, 2007), allowing TE researchers to share data and use the common Repbase repeat database (Jurka *et al.*, 2005). This nomenclature is also used by classical repeat-detection software such as RepeatMasker (http://www.repeatmasker.org/). Figure 2 represents the different levels of TE classification and Figure 3 represents a flowchart for their nomenclature in *Ciona intestinalis*.

At the higher level (class), TEs are classified based on their transposition mechanism. Class I elements (retrotransposons) transpose via the reverse transcription of an RNA intermediate through a mechanism commonly called "copy-and-paste." Class II elements (DNA transposons) directly transpose from DNA to DNA sequences using a "cut-and-paste" mechanism. Within both classes, there are several levels of classification (subclasses, orders, superfamilies, and families) that include elements sharing common structural features and transposition mechanisms, and forming phylogenetically distinct groups of sequences.

Retrotransposons (class I elements) are subdivided in two subclasses, depending on the presence or absence of Long Terminal Repeat (LTR) flanking the element: LTR retrotransposons include the LTR and DIRS orders, while non-LTR retrotransposons comprises the Penelope, LINE and SINE orders. Within orders, elements are classified into superfamilies. For instance, the LTR order contains four superfamilies: Ty3/Gypsy, BEL/Pao, Ty1/Copia, and retroviridae.

DNA transposons (class II elements) are subdivided into two subclasses, depending on the number of DNA strands that are cut at the TE donor site for transposition. Subclass I (single strand cut) includes two orders (TIRs for "Terminal Inverted Repeats" and Crypton). Subclass II (double strand cut) includes two orders, Helitron and Polinton.

TE subfamilies are named according to (Kapitonov and Jurka, 2008): **NAME#(SUB)CLASS/SUPERFAMILY**,

Where the **NAME** contains Prefix-Infix1[-Infix2]_Suffix

Prefix is the name of the element, in general the family or superfamily name.

Infix 1 is the subfamily identifier, in general a number.

Infix 2 defines particular structural features such as LTR, or I for Internal part (region of the element comprised between the two LTR or TIR).

Suffix is the species identifier (4–5 letters).

Where the **(SUB)CLASS** can be (strict choice)

LTR for the following orders: LTR, DIRS.

LINE for the non-LTR retrotransposons and Penelope orders.

SINE for the non-LTR non-autonomous SINE order.

DNA for all DNA subclasses.

Where **SUPERFAMILY** corresponds to a superfamily, when identified, belonging to one of the classes or subclasses if the superfamily has been identified.

Examples:

- Tc1-4_Phmamm#DNA/TcMariner is the consensus sequence of a Tc1 element in *Phallusia mammillata* (Phmamm).
- Gypsy-3-LTR_Ciinte#LTR/Gypsy is the consensus sequence of a Solo-LTR of a Gypsy retrotransposon in *Ciona intestinalis* (Ciinte).
- LTR-10_Harore#LTR/Unknown or LTR-10_Harore#LTR is the consensus sequence of a LTR retrotransposon but unclear superfamily in *Halocynthia roretzi* (Harore).
- LINE1_Moocci#LINE/L1 is the consensus sequence of a LINE1 non-LTR retrotransposon in *Molgula occidentalis* (Moocci).

TRANSGENIC LINES AND TRANSGENIC CONSTRUCTS

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" (SO:0000902). Transgenes are obtained by the introduction into an organism of a transgenic construct, an engineered plasmid or other recombinant DNA molecule that carries various features including, but not limited to, *cis-*

regulatory sequences, reporter genes, etc. Transgenic lines are named according to the construct(s) used to generate them.

Transgenic Constructs

Most transgenic constructs are defined by a *cis*-regulatory sequence, and a protein-coding cDNA sequence (or sequence that is transcribed as a non-coding RNA). The general syntax of transgenic constructs is: p{Mi., SB., SBT2., (blank)}-{driver}>{functional RNA- or protein-coding gene}.

The "p" indicates that the construct has a plasmid background, the first bracket indicates the type of transposon used, if any: Mi for Minos transposon, SB for sleeping beauty transposon, and SBT2 for the T2 variant of SB used in *Ciona intestinalis*. This field can be left in blank if no transposon system is used to facilitate transgene integration.

The driver is of the form: {Species}.{*cis*-reg}. When the driver consists of the fusion of an enhancer and a promoter, these two elements are separated by a colon (":"). *Cis*-regulatory and functional RNA or protein names are separated by the symbol ">."

Transgenes can be transcribed into protein-coding or non-coding RNAs. The symbol of the sequence to be transcribed can adopt various syntaxes, according to their source. It can code for a non-tunicate protein, in which case the symbol used should be the one frequently used in the literature (e.g., *EGFP*). If it encodes a tunicate protein, the gene symbol conventions established in these guidelines should be used (e.g., *Creb1*). If the distinction between different tunicate species needs to be made, then the convention is {species abbreviation}.{gene symbol} (e.g., *Ciinte.Creb1*).

Additionally, the transcribed sequence can also encode a fusion between two different proteins or parts thereof. A double colon ("::") indicates an in-frame fusion between two protein-coding sequences.

Example:

- *EGFP::Ciinte.Creb1* corresponds to a fusion between the full length EGFP placed N-terminally in-frame with the Creb1 protein from *Ciona intestinalis*.

The names of both *cis*-regulatory sequence and coding or non-coding RNAs should be brief and need not provide exhaustive details. When the name of a construct is too long to be used in the main text of a manuscript, an abbreviated form can be used, provided the full name is mentioned in the materials and methods of the article.

Examples:

- pCiinte.REG000005>*NLSLacZ*, is a plasmid construct that drives a nuclear form of β-Galactosidase (encoded by the *NLSLacZ* gene) under control of the *cis*-regulatory sequence 000005 from *Ciona intestinalis*.
- pMi-Ciinte.REG.KH2012.C4.4315574-4315697|*Otx:prCiinte.Bra>Ciinte.Syt1/2::EGFP* is a plasmid construct made in a Minos transposon vector that drives a fusion between *Ciona intestinalis* synaptotagmin 1/2 (symbol Syt1/2) and enhanced GFP (placed in C-terminal position), under control of a composite *Ciona intestinalis cis*-regulatory sequence made of an *Otx* enhancer and the *Ciona intestinalis Brachyury* minimal (Corbo *et al.* 1997). This *Otx* regulatory sequence is often referred to as the a-element in the literature, and the abbreviated form of this construct, if the species is not ambiguous, could be pMi-*Otx-a-elt:prBra>Syt1/2::EGFP*, or even pMi-*Otx>Syt::EGFP* if there is no ambiguity about which *Otx* element and *Syt* gene is being used.
- pPhmamm.REG000056>*tdTomato::Ciinte.H2b* is a plasmid construct that drives a fusion of the *Ciona intestinalis* histone 2b and the tdTomato tandem fluorescent protein (placed in N-terminal position), under control of the *Phallusia mammillata cis*-regulatory sequence 000056.

In some cases (e.g., Gal4/UAS system), a single construct can harbor several cassettes, each containing a regulatory and a coding or functional sequence. In this case, multiple cassettes are distinguished by a semicolon inside parentheses, appended by "p" to denote the fact that both cassettes are on the same plasmid.

Example:

- p(*Ciinte.Isl>GAL4*;6xUAS:*prCiinte.Bra>GFP*): this construct drives Gal4 under the control of a *Ciona intestinalis Islet (Isl) cis*-regulatory region. It also harbors a second cassette, with a 6-mer of the UAS sequence, placed in front of the *prBra* minimal promoter, and driving GFP. Note that short construct names should be completed in materials and methods by the full name including coordinates, in a specified assembly, of *Isl* and *prBra*.

Transgenic Line

A transgenic line is 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. A transgenic line is defined by a species, one or more transgenic constructs and, when known, the insertion locus of the construct in the species genome. The general syntax of a line is Species.Tg[-construct]{n}.Insertion_site, where [construct] is built from the above rules and where the insertion descriptor is omitted if the insertion site is unknown. The number at the {n} part is used to distinguish different transgenic lines harboring the same transgenic construct. This number does not always coincide with the number of available transgenic lines produced with the construct,

because some of the initial transgenic lines may not have been kept.

Examples:

- Ciinte.Tg[pMi-Ciinte.REG.KH2012.I41.267342-270949|*Zip:prFoxa.a>Kaede*]1 (note that *prFoxa.a* should be formally Ciinte.REG.KH2012.C11.4404828-4404710), is the first transgenic line of *Ciona intestinalis* obtained by *Minos* transposon-mediated transgenesis with a vector that drives the fluorescent protein Kaede under control of a *Ciona intestinalis Zip* enhancer placed in front of the minimal promoter of *Foxa.a* (Di Gregorio *et al.* 2001). This name could be abbreviated as Ciinte.Tg[pMi-*Zip>Kaede*]1. Transgenic lines can have synonymous names to reflect previous names. For instance, this construct was initially named Tg[MiCiZipCifkhK]1 in the first report (Nakazawa *et al.*, 2013).
- Ciinte.Tg[pSB-Ciinte.REG.KH2012.C10.4438567-4440059|*Msi:prTpo>NLSDsRed*;Ciinte.REG.KH2012.C14.1414843-1413822|*Nut>MiTP*]1 is a *Ciona intestinalis* transgenic line generated using a vector including a sleeping beauty transposon element. The vector contains two expression cassettes (separated by a semi colon). One drives a nuclear form of DsRed under control of the *cis* element of an enhancer from *Musashi (Msi)* and a promoter of a gene encoding thyroid peroxidase (*Tpo*) of *Ciona intestinalis*. The other cassette drives *Minos* transposase (MiTP) under control of a *cis* element from *Nut* gene of *Ciona intestinalis*. This transgenic line was initially named Ju[SBFr3dTPORCiNutMiTP]1 (Hozumi *et al.*, 2010).

Enhancer trap lines should be identified by the capital letter affix E instead of Tg, because expression pattern of transgenes cannot be deduced from the transgenic line name. Gene trap and promoter trap lines could be identified by the capital letter affixes G and P, respectively.

Example:

- Ciinte.E[pMi-TSA-Ciinte.REG.KH2012.L3.178445-177583|*Tpo>NLSEGFP*]124.KH2012.L171.188604 is an enhancer trap line of *Ciona intestinalis* transformed by the *Minos*-mediated transgenesis that drives nuclear-localized EGFP under control of the *cis* element of *Tpo* of *Ciona intestinalis*. In this line, the insertion site of the vector has been identified, and the information follows the line name. "TSA" indicates the splicing acceptor and transcription termination sequence cassette used in the construct. This transgenic line was initially named EJ[MiTSAdT-POG]124 (Sasakura *et al.*, 2012).

MUTANT LINES

A mutant line is a line of organisms showing a deviation from the wild-type phenotype and derived from a single genetically modified parent, through classical mutagenesis or transgenic insertion. Mutant lines are defined by their genetic alteration, their genetic background and their phenotype. Nomenclatures differ for mutants isolated by forward and reverse genetic means.

Mutants obtained by forward genetic means are defined by their phenotypes, and their name is left to the imagination of the creator of the mutant, who should also provide an abbreviated symbol, which will subsequently be used. Mutant allele names of causative genes are specified by gene names followed by a superscription of abbreviated mutant names.

Example:

- *swimming juvenile* (*sj*) (Sasakura *et al.*, 2005) is a *Ciona intestinalis* mutant in which metamorphic events occur in an irregular order. The causative gene of this mutant is the *Cellulose synthase A* gene (*Cesa*), and the allele name of mutated *Cesa* in this mutant line is $Cesa^{sj}$.

The name of mutants or mutant alleles obtained by reverse genetic means, the mutant lines can be specified by the target gene and the method used to produce mutants. The general syntax is: {gene symbol}$^{\{TAL,\ CAS,\ INS\}}$, where TAL stands for TALEN, CAS for CRISPR/Cas9, INS for insertion, of a transgene for instance. If more than two different mutant alleles are established by the same technology, they can be distinguished by numbers, like Bra^{TAL2} and Bra^{TAL3}.

Example:

- The Bra^{TAL} mutant designates a mutant for the *Brachyury* gene created with the help of TALEN technology.
- The animal specified by the name Bra^{TAL2}/Bra^{CAS} is a heterozygous animal that harbors one *Brachyury* mutant allele created by TALEN (the second such allele to be reported) and one *Brachyury* mutant allele created by CRISPR/Cas9.

CONCLUDING REMARKS: INSTRUCTIONS FOR AUTHORS AND FOR BIOINFORMATICIANS

The rules listed in this article were designed to balance the occasionally antagonistic wishes of experimental and computational biologists. Experimentalists want compact names that make biological sense to them. The precise formatting of these names is frequently seen as a hurdle that muddles biological messages. By contrast, bioinformaticians' priorities are reproducible standardized and traceable names, with a syntax facilitating the efficient parsing of large files. As a result, some of the consensus rules listed here will probably be problematic for either categories of users. This final paragraph lists some difficulties for experimental and