**ST.26 - ANNEX I**

CONTROLLED VOCABULARY

Final Draft

*Proposal presented by the SEQL Task Force for consideration and adoption at the CWS/4*

TABLE OF CONTENTS

SECTION 1: LIST OF NUCLEOTIDES 23

SECTION 2: LIST OF MODIFIED NUCLEOTIDES 23

SECTION 3: LIST OF AMINO ACIDS 25

SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS 26

SECTION 5: FEATURES KEYS FOR NUCLEIC SEQUENCES 27

SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC SEQUENCES 47

SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES 65

SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES 71

SECTION 9: GENETIC CODES TABLES 72

## SECTION 1: LIST OF NUCLEOTIDES

The nucleotide base codes to be used in sequence listings are presented in Table 1. The symbol “t” will be construed as thymine in DNA and uracil in RNA when it is used with no further description. Where an ambiguity symbol (representing two or more bases in the alternative) is appropriate, the most restrictive symbol should be used. For example, if a base in a given position could be “a or g,” then “r” should be used, rather than “n”. The symbol “n” will be construed as “a or c or g or t/u” when it is used with no further description.

Table 1: List of nucleotides

|  |  |
| --- | --- |
| **Symbol** | **Nucleotide** |
| a | adenine |
| c | cytosine |
| g | guanine |
| t | thymine in DNA/uracil in RNA (t/u) |
| m | a or c |
| r | a or g |
| w | a or t/u |
| s | c or g |
| y | c or t/u |
| k | g or t/u |
| v | a or c or g; not t/u |
| h | a or c or t/u; not g |
| d | a or g or t/u; not c |
| b | c or g or t/u; not a |
| n | a or c or g or t/u; unknown or other |

## SECTION 2: LIST OF MODIFIED NUCLEOTIDES

The abbreviations listed in Table 2 are the only permitted values for the mod\_base qualifier. Where a specific modified nucleotide is not present in the table below, then the abbreviation “OTHER” must be used as its value. If the abbreviation is “OTHER,” then the complete unabbreviated name of the modified base must be provided in a note qualifier. The abbreviations provided in Table 2 must not be used in the sequence itself.

Table 2: List of modified nucleotides

| **Abbreviation** | **Modified Nucleotide** |
| --- | --- |
| ac4c | 4-acetylcytidine |
| chm5u | 5-(carboxyhydroxylmethyl)uridine |
| cm | 2’-O-methylcytidine |
| cmnm5s2u | 5-carboxymethylaminomethyl-2-thiouridine |
| cmnm5u | 5-carboxymethylaminomethyluridine |
| d | dihydrouridine |
| fm | 2’-O-methylpseudouridine |
| gal q | beta,D-galactosylqueosine |
| gm | 2’-O-methylguanosine |
| i | inosine |
| i6a | N6-isopentenyladenosine |
| m1a | 1-methyladenosine |
| m1f | 1-methylpseudouridine |
| m1g | 1-methylguanosine |
| m1i | 1-methylinosine |
| m22g | 2,2-dimethylguanosine |
| m2a | 2-methyladenosine |
| m2g | 2-methylguanosine |
| m3c | 3-methylcytidine |
| m5c | 5-methylcytidine |
| m6a | N6-methyladenosine |
| m7g | 7-methylguanosine |
| mam5u | 5-methylaminomethyluridine |
| mam5s2u | 5-methoxyaminomethyl-2-thiouridine |
| man q | beta,D-mannosylqueosine |
| mcm5s2u | 5-methoxycarbonylmethyl-2-thiouridine |
| mcm5u | 5-methoxycarbonylmethyluridine |
| mo5u | 5-methoxyuridine |
| ms2i6a | 2-methylthio-N6-isopentenyladenosine |
| ms2t6a | N-((9-beta-D-ribofuranosyl-2-methyltiopurine-6-yl)carbamoyl)threonine |
| mt6a | N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine |
| mv | uridine-5-oxyacetic acid-methylester |
| o5u | uridine-5-oxyacetic acid (v) |
| osyw | wybutoxosine |
| p | pseudouridine |
| q | queosine |
| s2c | 2-thiocytidine |
| s2t | 5-methyl-2-thiouridine |
| s2u | 2-thiouridine |
| s4u | 4-thiouridine |
| t6a | N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine |
| tm | 2’-O-methyl-5-methyluridine |
| um | 2’-O-methyluridine |
| yw | wybutosine |
| x | 3-(3-amino-3-carboxypropyl)uridine, (acp3)u |
| OTHER | (requires note qualifier) |

## SECTION 3: LIST OF AMINO ACIDS

The amino acid codes to be used in sequence are presented in Table 3. Where an ambiguity symbol (representing two or more amino acids in the alternative) is appropriate, the most restrictive symbol should be used. For example, if an amino acid in a given position could be aspartic acid or asparagine, the symbol “B” should be used, rather than “X”. The symbol “X” will be construed as any one of “A”, “R”, “N”, “D”, “C”, “Q”, “E”, “G”, “H”, “I”, “L”, “K”, “M”, “F”, “P”, “O”, “S”, “U”, “T”, “W”, “Y”, or “V”, when it is used with no further description.

Table 3: List of amino acids

|  |  |
| --- | --- |
| **Symbol** | **Amino acid** |
| A | Alanine |
| R | Arginine |
| N | Asparagine |
| D | Aspartic acid (Aspartate) |
| C | Cysteine |
| Q | Glutamine |
| E | Glutamic acid (Glutamate) |
| G | Glycine |
| H | Histidine |
| I | Isoleucine |
| L | Leucine |
| K | Lysine |
| M | Methionine |
| F | Phenylalanine |
| P | Proline |
| O | Pyrrolysine |
| S | Serine |
| U | Selenocysteine |
| T | Threonine |
| W | Tryptophan |
| Y | Tyrosine |
| V | Valine |
| B | Aspartic acid or Asparagine |
| Z | Glutamine or Glutamic acid |
| J | Leucine or Isoleucine |
| X | unknown or other |

## SECTION 4: LIST OF MODIFIED AND UNUSUAL AMINO ACIDS

Table 4 lists the only permitted abbreviations for a modified or unusual amino acid in the mandatory qualifier “NOTE” for feature keys “MOD\_RES” or “SITE”. The value for the qualifier “NOTE” must be either an abbreviation from this table, where appropriate, or the complete, unabbreviated name of the modified amino acid. The abbreviations (or full names) provided in this table must not be used in the sequence itself.

Table 4: List of modified and unusual amino acids

|  |  |
| --- | --- |
| **Abbreviation** | **Modified or Unusual Amino acid** |
| Aad | 2-Aminoadipic acid |
| bAad | 3-Aminoadipic acid |
| bAla | beta-Alanine, beta-Aminoproprionic acid |
| Abu | 2-Aminobutyric acid |
| 4Abu | 4-Aminobutyric acid, piperidinic acid |
| Acp | 6-Aminocaproic acid |
| Ahe | 2-Aminoheptanoic acid |
| Aib | 2-Aminoisobutyric acid |
| bAib | 3-Aminoisobutyric acid |
| Apm | 2-Aminopimelic acid |
| Dbu | 2,4-Diaminobutyric acid |
| Des | Desmosine |
| Dpm | 2,2’-Diaminopimelic acid |
| Dpr | 2,3-Diaminoproprionic acid |
| EtGly | N-Ethylglycine |
| EtAsn | N-Ethylasparagine |
| Hyl | Hydroxylysine |
| aHyl | allo-Hydroxylysine |
| 3Hyp | 3-Hydroxyproline |
| 4Hyp | 4-Hydroxyproline |
| Ide | Isodesmosine |
| aIle | allo-Isoleucine |
| MeGly | N-Methylglycine, sarcosine |
| MeIle | N-Methylisoleucine |
| MeLys | 6-N-Methyllysine |
| MeVal | N-Methylvaline |
| Nva | Norvaline |
| Nle | Norleucine |
| Orn | Ornithine |

## SECTION 5: FEATURES KEYS FOR NUCLEIC SEQUENCES

This paragraph contains the list of allowed feature keys to be used for nucleotide sequences, and lists mandatory and optional qualifiers. The feature keys are listed in alphabetic order. The feature keys can be used for either DNA or RNA unless otherwise indicated under “Molecule scope”. Some feature keys include a ‘Parent Key’ designation; when a parent key is indicated in the description of a feature key, it is mandatory that the designated parent key be used. Certain Feature Keys may be appropriate for use with artificial sequences in addition to the specified “organism scope”.

Feature key names must be used in the XML instance of the sequence listing exactly as they appear following “Feature key” in the descriptions below, except for the feature keys 3’UTR and 5’UTR. See “Comment” in the description for the 3’UTR and 5’UTR feature keys.

5.1. Feature Key attenuator

Definition 1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons;

2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription

Optional qualifiers allele

gene

gene\_synonym

map

note

operon

phenotype

Organism scope prokaryotes

Molecule scope DNA

5.2. Feature Key C\_region

Definition constant region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; includes one or more exons depending on the particular chain

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Parent Key CDS

Organism scope eukaryotes

5.3. Feature Key CAAT\_signal

Definition CAAT box; part of a conserved sequence located about 75 bp up-stream of the start point of eukaryotic transcription units which may be involved in RNA polymerase binding; consensus=GG(C or T)CAATCT [1,2]

Optional qualifiers allele

gene

gene\_synonym

map

note

Organism scope eukaryotes and eukaryotic viruses

Molecule scope DNA

References [1] Efstratiadis, A. et al. Cell 21, 653-668 (1980)

[2] Nevins, J.R. "The pathway of eukaryotic mRNA formation" Ann Rev Biochem 52, 441-466 (1983)

5.4. Feature Key CDS

Definition coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature may include amino acid conceptual translation

Optional qualifiers allele

artificial\_location

codon\_start

EC\_number

exception

function

gene

gene\_synonym

map

note

number

operon

product

protein\_id

pseudo

pseudogene

ribosomal\_slippage

standard\_name

translation

transl\_except

transl\_table

trans\_splicing

Comment codon\_start qualifier has valid value of 1 or 2 or 3, indicating the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature; transl\_table defines the genetic code table used if other than the Standard or universal genetic code table; genetic code exceptions outside the range of the specified tables are reported in transl\_except qualifier; only one of the qualifiers translation and pseudo are permitted with a CDS feature key; when the translation qualifier is used, the protein\_id qualifier is mandatory if the translation product contains four or more amino acids

5.5. Feature Key centromere

Definition region of biological interest indentified as a centromere and which has been experimentally characterized

Optional qualifiers note

standard\_name

Comment the centromere feature describes the interval of DNA that corresponds to a region where chromatids are held and a kinetochore is formed

5.6. Feature Key D-loop

Definition displacement loop; a region within mitochondrial DNA in which a short stretch of RNA is paired with one strand of DNA, displacing the original partner DNA strand in this region; also used to describe the displacement of a region of one strand of duplex DNA by a single stranded invader in the reaction catalyzed by RecA protein

Optional qualifiers allele

gene

gene\_synonym

map

note

Molecule scope DNA

5.7. Feature Key D\_segment

Definition Diversity segment of immunoglobulin heavy chain, and T-cell receptor beta chain

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Parent Key CDS

Organism scope eukaryotes

5.8. Feature Key enhancer

Definition a cis-acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter

Optional qualifiers allele

bound\_moiety

gene

gene\_synonym

map

note

standard\_name

Organism scope eukaryotes and eukaryotic viruses

5.9. Feature Key exon

Definition region of genome that codes for portion of spliced mRNA,rRNA and tRNA; may contain 5’UTR, all CDSs and 3’ UTR

Optional qualifiers allele

EC\_number

function

gene

gene\_synonym

map

note

number

product

pseudo

pseudogene

standard\_name

trans\_splicing

5.10. Feature Key GC\_signal

Definition GC box; a conserved GC-rich region located upstream of the start point of eukaryotic transcription units which may occur in multiple copies or in either orientation; consensus=GGGCGG

Optional qualifiers allele

gene

gene\_synonym

map

note

Organism scope eukaryotes and eukaryotic viruses

5.11. Feature Key gene

Definition region of biological interest identified as a gene and for which a name has been assigned

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

product

pseudo

pseudogene

phenotype

standard\_name

trans\_splicing

Comment the gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to its positions at the ends; it is meant to represent a region where the gene is located.

5.12. Feature Key iDNA

Definition intervening DNA; DNA which is eliminated through any of several kinds of recombination

Optional qualifiers allele

function

gene

gene\_synonym

map

note

number

standard\_name

Molecule scope DNA

Comment e.g., in the somatic processing of immunoglobulin genes.

5.13. Feature Key intron

Definition a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it

Optional qualifiers allele

function

gene

gene\_synonym

map

note

number

pseudo

pseudogene

standard\_name

trans\_splicing

5.14. Feature Key J\_segment

Definition joining segment of immunoglobulin light and heavy

chains, and T-cell receptor alpha, beta, and gamma chains

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Parent Key CDS

Organism scope eukaryotes

5.15. Feature Key LTR

Definition long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses

Optional qualifiers allele

function

gene

gene\_synonym

map

note

standard\_name

5.16. Feature Key mat\_peptide

Definition mature peptide or protein coding sequence; coding sequence for the mature or final peptide or protein product following post-translational modification; the location does not include the stop codon (unlike the corresponding CDS)

Optional qualifiers allele

EC\_number

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

5.17. Feature Key misc\_binding

Definition site in nucleic acid which covalently or non-covalently binds another moiety that cannot be described by any other binding key (primer\_bind or protein\_bind)

Mandatory qualifiers bound\_moiety

Optional qualifiers allele

function

gene

gene\_synonym

map

note

Comment note that the feature key RBS is used for ribosome binding sites

5.18. Feature Key misc\_difference

Definition featured sequence differs from the presented sequence at this location and cannot be described by any other Difference key (unsure, variation, or modified\_base)

Optional qualifiers allele

clone

compare

gene

gene\_synonym

map

note

phenotype

replace

standard\_name

Comment the misc\_difference feature key should be used to describe variability introduced artificially, e.g. by genetic manipulation or by chemical synthesis; use the replace qualifier to annotate a deletion, insertion, or substitution.

5.19. Feature Key misc\_feature

Definition region of biological interest which cannot be described by any other feature key; a new or rare feature

Optional qualifiers allele

function

gene

gene\_synonym

map

note

number

phenotype

product

pseudo

pseudogene

standard\_name

Comment this key should not be used when the need is merely to mark a region in order to comment on it or to use it in another feature’s location

5.20. Feature Key misc\_recomb

Definition site of any generalized, site-specific or replicative recombination event where there is a breakage and reunion of duplex DNA that cannot be described by other recombination keys or qualifiers of source key (proviral)

Optional qualifiers allele

gene

gene\_synonym

map

note

standard\_name

Molecule scope DNA

5.21. Feature Key misc\_RNA

Definition any transcript or RNA product that cannot be defined by other RNA keys (prim\_transcript, precursor\_RNA, mRNA, 5’UTR, 3’UTR, exon, CDS, sig\_peptide, transit\_peptide, mat\_peptide, intron, polyA\_site, ncRNA, rRNA and tRNA)

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

product

pseudo

pseudogene

standard\_name

trans\_splicing

5.22. Feature Key misc\_signal

Definition any region containing a signal controlling or altering gene function or expression that cannot be described by other signal keys (promoter, CAAT\_signal, TATA\_signal, -35\_signal, -10\_signal, GC\_signal, RBS, polyA\_signal, enhancer, attenuator, terminator, and rep\_origin)

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

phenotype

standard\_name

5.23. Feature Key misc\_structure

Definition any secondary or tertiary nucleotide structure or conformation that cannot be described by other Structure keys (stem\_loop and D-loop)

Optional qualifiers allele

function

gene

gene\_synonym

map

note

standard\_name

5.24. Feature Key mobile\_element

Definition region of genome containing mobile elements

Mandatory qualifiers mobile\_element\_type

Optional qualifiers allele

function

gene

gene\_synonym

map

note

rpt\_family

rpt\_type

standard\_name

5.25. Feature Key modified\_base

Definition the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod\_base qualifier value)

Mandatory qualifiers mod\_base

Optional qualifiers allele

frequency

gene

gene\_synonym

map

note

Comment value for the mandatory mod\_base qualifier is limited to the restricted vocabulary for modified base abbreviations in Section 2 of this Annex.

5.26. Feature Key mRNA

Definition messenger RNA; includes 5’ untranslated region (5’UTR), coding sequences (CDS, exon) and 3’ untranslated region (3’UTR)

Optional qualifiers allele

artificial\_location

function

gene

gene\_synonym

map

note

operon

product

pseudo

pseudogene

standard\_name

trans\_splicing

5.27. Feature Key ncRNA

Definition a non-protein-coding gene, other than ribosomal RNA and transfer RNA, the functional molecule of which is the RNA transcript

Mandatory qualifiers ncRNA\_class

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

product

pseudo

pseudogene

standard\_name

trans\_splicing

Comment the ncRNA feature is not used for ribosomal and transfer RNA annotation, for which the rRNA and tRNA feature keys should be used, respectively

5.28. Feature Key N\_region

Definition extra nucleotides inserted between rearranged immunoglobulin segments

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Parent Key CDS

Organism scope eukaryotes

5.29. Feature Key operon

Definition region containing polycistronic transcript including a cluster of genes that are under the control of the same regulatory sequences/promotor and in the same biological pathway

Mandatory qualifiers operon

Optional qualifiers allele

function

map

note

phenotype

pseudo

pseudogene

standard\_name

5.30. Feature Key oriT

Definition origin of transfer; region of a DNA molecule where transfer is initiated during the process of conjugation or mobilization

Optional qualifiers allele

bound\_moiety

direction

gene

gene\_synonym

map

note

rpt\_family

rpt\_type

rpt\_unit\_range

rpt\_unit\_seq

standard\_name

Molecule Scope DNA

Comment rep\_origin should be used for origins of replication; direction qualifier has legal values RIGHT, LEFT and BOTH, however only RIGHT and LEFT are valid when used in conjunction with the oriT feature; origins of transfer can be present in the chromosome; plasmids can contain multiple origins of transfer

5.31. Feature Key polyA\_signal

Definition recognition region necessary for endonuclease cleavage of an RNA transcript that is followed by polyadenylation; consensus=AATAAA [1]

Optional qualifiers allele

gene

gene\_synonym

map

note

Organism scope eukaryotes and eukaryotic viruses

References [1] Proudfoot, N. and Brownlee, G.G. Nature 263, 211-214 (1976)

5.32. Feature Key polyA\_site

Definition site on an RNA transcript to which will be added adenine residues by post-transcriptional polyadenylation

Optional qualifiers allele

gene

gene\_synonym

map

note

Organism scope eukaryotes and eukaryotic viruses

5.33. Feature Key precursor\_RNA

Definition any RNA species that is not yet the mature RNA product; may include 5’ untranslated region (5’UTR), coding sequences (CDS, exon), intervening sequences (intron) and 3’ untranslated region (3’UTR)

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

product

standard\_name

trans\_splicing

Comment used for RNA which may be the result of post-transcriptional processing; if the RNA in question is known not to have been processed, use the prim\_transcript key

5.34. Feature Key prim\_transcript

Definition primary (initial, unprocessed) transcript; includes 5’ untranslated region (5’UTR), coding sequences (CDS, exon), intervening sequences (intron) and 3’ untranslated region (3’UTR)

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

standard\_name

5.35. Feature Key primer\_bind

Definition non-covalent primer binding site for initiation of replication, transcription, or reverse transcription; includes site(s) for synthetic e.g., PCR primer elements

Optional qualifiers allele

gene

gene\_synonym

map

note

standard\_name

PCR\_conditions

Comment used to annotate the site on a given sequence to which a primer molecule binds - not intended to represent the sequence of the primer molecule itself; PCR components and reaction times may be stored under the PCR\_conditions qualifier; since PCR reactions most often involve pairs of primers, a single primer\_bind key may use the order(location,location) operator with two locations, or a pair of primer\_bind keys may be used

5.36. Feature Key promoter

Definition region on a DNA molecule involved in RNA polymerase binding to initiate transcription

Optional qualifiers allele

bound\_moiety

function

gene

gene\_synonym

map

note

operon

phenotype

pseudo

pseudogene

standard\_name

Molecule scope DNA

5.37. Feature Key protein\_bind

Definition non-covalent protein binding site on nucleic acid

Mandatory qualifiers bound\_moiety

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

standard\_name

Comment note that RBS is used for ribosome binding sites

5.38. Feature Key RBS

Definition ribosome binding site

Optional qualifiers allele

gene

gene\_synonym

map

note

standard\_name

References [1] Shine, J. and Dalgarno, L. Proc Natl Acad Sci USA 71, 1342-1346 (1974)

[2] Gold, L. et al. Ann Rev Microb 35, 365-403 (1981)

Comment in prokaryotes, known as the Shine-Dalgarno sequence: is located 5 to 9 bases upstream of the initiation codon; consensus GGAGGT [1,2]

5.39. Feature Key repeat\_region

Definition region of genome containing repeating units

Optional qualifiers allele

function

gene

gene\_synonym

map

note

rpt\_family

rpt\_type

rpt\_unit\_range

rpt\_unit\_seq

satellite

standard\_name

5.40. Feature Key rep\_origin

Definition origin of replication; starting site for duplication of nucleic acid to give two identical copies

Optional Qualifiers allele

direction

gene

gene\_synonym

map

note

standard\_name

Comment direction qualifier has valid values: RIGHT, LEFT, or BOTH

5.41. Feature Key rRNA

Definition mature ribosomal RNA; RNA component of the ribonucleoprotein particle (ribosome) which assembles amino acids into proteins

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

product

pseudo

standard\_name

Comment rRNA sizes should be annotated with the product qualifier

5.42. Feature Key S\_region

Definition switch region of immunoglobulin heavy chains; involved in the rearrangement of heavy chain DNA leading to the expression of a different immunoglobulin class from the same B-cell

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Parent Key misc\_signal

Organism scope eukaryotes

5.43. Feature Key sig\_peptide

Definition signal peptide coding sequence; coding sequence for an N-terminal domain of a secreted protein; this domain is involved in attaching nascent polypeptide to the membrane leader sequence

Optional qualifiers allele

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

5.44. Feature Key source

Definition identifies the source of the sequence; this key is mandatory; every sequence will have a single source key spanning the entire sequence

Mandatory qualifiers organism

mol\_type

Optional qualifiers cell\_line

cell\_type

chromosome

clone

clone\_lib

collected\_by

collection\_date

cultivar

dev\_stage

ecotype

environmental\_sample

germline

haplogroup

haplotype

host

identified\_by

isolate

isolation\_source

lab\_host

lat\_lon

macronuclear

map

mating\_type

note

organelle

PCR\_primers

plasmid

pop\_variant

proviral

rearranged

segment

serotype

serovar

sex

strain

sub\_clone

sub\_species

sub\_strain

tissue\_lib

tissue\_type

variety

Molecule scope any

5.45. Feature Key stem\_loop

Definition hairpin; a double-helical region formed by base-pairing between adjacent (inverted) complementary sequences in a single strand of RNA or DNA

Optional qualifiers allele

function

gene

gene\_synonym

map

note

operon

standard\_name

5.46. Feature Key STS

Definition sequence tagged site; short, single-copy DNA sequence that characterizes a mapping landmark on the genome and can be detected by PCR; a region of the genome can be mapped by determining the order of a series of STSs

Optional qualifiers allele

gene

gene\_synonym

map

note

standard\_name

Molecule scope DNA

Parent key misc\_binding

Comment STS location to include primer(s) in primer\_bind key or primers

5.47. Feature Key TATA\_signal

Definition TATA box; Goldberg-Hogness box; a conserved AT-rich septamer found about 25 bp before the start point of each eukaryotic RNA polymerase II transcript unit which may be involved in positioning the enzyme for correct initiation; consensus=TATA(A or T)A(A or T) [1,2]

Optional qualifiers allele

gene

gene\_synonym

map

note

Organism scope eukaryotes and eukaryotic viruses

Molecule scope DNA

References [1] Efstratiadis, A. et al. Cell 21, 653-668 (1980)

[2] Corden, J., et al. "Promoter sequences of eukaryotic protein-encoding genes" Science 209, 1406-1414 (1980)

5.48. Feature Key telomere

Definition region of biological interest identified as a telomere and which has been experimentally characterized

Optional qualifiers note

rpt\_type

rpt\_unit\_range

rpt\_unit\_seq

standard\_name

Comment the telomere feature describes the interval of DNA that corresponds to a specific structure at the end of the linear eukaryotic chromosome which is required for the integrity and maintenance of the end; this region is unique compared to the rest of the chromosome and represents the physical end of the chromosome

5.49. Feature Key terminator

Definition sequence of DNA located either at the end of the transcript that causes RNA polymerase to terminate transcription

Optional qualifiers allele

gene

gene\_synonym

map

note

operon

standard\_name

Molecule scope DNA

5.50. Feature Key tmRNA

Definition transfer messenger RNA; tmRNA acts as a tRNA first, and then as an mRNA that encodes a peptide tag; the ribosome translates this mRNA region of tmRNA and attaches the encoded peptide tag to the C-terminus of the unfinished protein; this attached tag targets the protein for destruction or proteolysis

Optional qualifiers allele

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

tag\_peptide

5.51. Feature Key transit\_peptide

Definition transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organellar protein; this domain is involved in post-translational import of the protein into the organelle

Optional qualifiers allele

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

5.52. Feature Key tRNA

Definition mature transfer RNA, a small RNA molecule (75-85 bases long) that mediates the translation of a nucleic acid sequence into an amino acid sequence

Optional qualifiers allele

anticodon

function

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

trans\_splicing

5.53. Feature Key unsure

Definition author is unsure of exact sequence in this region

Optional qualifiers allele

compare

gene

gene\_synonym

map

note

replace

Comment use the replace qualifier to annotate a deletion, insertion, or substitution.

5.54. Feature Key V\_region

Definition variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for the variable amino terminal portion; can be composed of V\_segments, D\_segments, N\_regions, and J\_segments

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Parent Key CDS

Organism scope eukaryotes

5.55. Feature Key V\_segment

Definition variable segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for most of the variable region (V\_region) and the last few amino acids of the leader peptide

Optional qualifiers allele

gene

gene\_synonym

map

note

product

pseudo

pseudogene

standard\_name

Parent Key CDS

Organism scope eukaryotes

5.56. Feature Key variation

Definition a related strain contains stable mutations from the same gene (e.g., RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others)

Optional qualifiers allele

compare

frequency

gene

gene\_synonym

map

note

phenotype

product

replace

standard\_name

Comment used to describe alleles, RFLP’s,and other naturally occurring mutations and polymorphisms; variability arising as a result of genetic manipulation (e.g. site directed mutagenesis) should be described with the misc\_difference feature; use the replace qualifier to annotate a deletion, insertion, or substitution

5.57. Feature Key 3’UTR

Definition region at the 3’ end of a mature transcript (following the stop codon) that is not translated into a protein

Optional qualifiers allele

function

gene

gene\_synonym

map

note

standard\_name

trans\_splicing

Comment The apostrophe character has special meaning in XML, and must be substituted with “&apos;” in the value of an element. Thus “3’UTR” must be represented as “3&apos;UTR” in the XML file, i.e., <INSDFeature\_key>3&apos;UTR</INSDFeature\_key>.

5.58. Feature Key 5’UTR

Definition region at the 5’ end of a mature transcript (preceding the initiation codon) that is not translated into a protein

Optional qualifiers allele

function

gene

gene\_synonym

map

note

standard\_name

trans\_splicing

Comment The apostrophe character has special meaning in XML, and must be substituted with “&apos;” in the value of an element. Thus “5’UTR” must be represented as “5&apos;UTR” in the XML file, i.e., <INSDFeature\_key>5&apos;UTR</INSDFeature\_key>.

5.59. Feature Key -10\_signal

Definition Pribnow box; a conserved region about 10 bp upstream of the start-point of bacterial transcription units which may be involved in binding RNA polymerase; consensus=TAtAaT [1,2,3,4]

Optional qualifiers allele

gene

gene\_synonym

map

note

operon

standard\_name

Organism scope prokaryotes

Molecule scope DNA

References [1] Schaller, H., Gray, C., and Hermann, K. Proc Natl Acad Sci USA 72, 737-741 (1974)

[2] Pribnow, D. Proc Natl Acad Sci USA 72, 784-788 (1974)

[3] Hawley, D.K. and McClure, W.R. "Compilation and analysis of Escherichia coli promoter DNA sequences" Nucl Acid Res 11, 2237-2255 (1983)

[4] Rosenberg, M. and Court, D. "Regulatory sequences involved in the promotion and termination of RNA transcription" Ann Rev Genet 13, 319-353 (1979)

5.60. Feature Key -35\_signal

Definition a conserved hexamer about 35 bp upstream of the start.point of bacterial transcription units; consensus=TTGACa or TGTTGACA

Optional qualifiers allele

gene

gene\_synonym

map

note

operon

standard\_name

Organism scope prokaryotes

Molecule scope DNA

References [1] Takanami, M., et al. Nature 260, 297-302 (1976)

[2] Moran, C.P., Jr., et al. Molec Gen Genet 186, 339-346 (1982)

[3] Maniatis, T., et al. Cell 5, 109-113 (1975)

## SECTION 6: DESCRIPTION OF QUALIFIERS FOR NUCLEIC SEQUENCES

This section contains the list of qualifiers to be used for features in nucleotide sequences. The qualifiers are listed in alphabetic order.

Where a Value format of “none” is indicated in the description of a qualifier (e.g. germline), the INSDQualifier\_value element must not be used.

6.1. Qualifier allele

Definition name of the allele for the given gene

Value format free text

Example <INSDQualifier\_value>adh1-1</INSDQualifier\_value>

Comment all gene-related features (exon, CDS etc) for a given gene should share the same allele qualifier value; the allele qualifier value must, by definition, be different from the gene qualifier value; when used with the variation feature key, the allele qualifier value should be that of the variant.

6.2. Qualifier anticodon

Definition location of the anticodon of tRNA and the amino acid for which it codes

Value format (pos:<location>,aa:<amino\_acid>,seq<text>) where location is the position of the anticodon and <amino\_acid> is the abbreviation for the amino acid encoded and seq is the sequence of the anticodon

Example <INSDQualifier\_value>(pos:34..36,aa:Phe,seq:aaa)</INSDQualifier\_value>

<INSDQualifier\_value>(pos:join(5,495..496,aa:Leu,seq:taa)</INSDQualifier\_value>

<INSDQualifier\_value>(pos:complement(4156..4158),aa:Glu,seq:ttg)</INSDQualifier\_value>

6.3. Qualifier bound\_moiety

Definition name of the molecule/complex that may bind to the given feature

Value format free text

Example <INSDQualifier\_value>GAL4</INSDQualifier\_value>

Comment Multiple bound\_moiety qualifiers are legal on "promoter" and "enhancer" features. A single bound\_moiety qualifier is legal on the "misc\_binding", "oriT" and "protein\_bind" features.

6.4. Qualifier cell\_line

Definition cell line from which the sequence was obtained

Value format free text

Example <INSDQualifier\_value>MCF7</INSDQualifier\_value>

6.5. Qualifier cell\_type

Definition cell type from which the sequence was obtained

Value format free text

Example <INSDQualifier\_value>leukocyte</INSDQualifier\_value>

6.6. Qualifier chromosome

Definition chromosome (e.g. Chromosome number) from which the sequence was obtained

Value format free text

Example <INSDQualifier\_value>1</INSDQualifier\_value>

<INSDQualifier\_value>X</INSDQualifier\_value>

6.7. Qualifier clone

Definition clone from which the sequence was obtained

Value format free text

Example <INSDQualifier\_value>lambda-hIL7.3</INSDQualifier\_value>

Comment not more than one clone should be specified for a given source feature; where the sequence was obtained from multiple clones it may be further described in the feature table using the feature key misc\_feature and a note qualifier to specify the multiple clones.

6.8. Qualifier clone\_lib

Definition clone library from which the sequence was obtained

Value format free text

Example <INSDQualifier\_value>lambda-hIL7</INSDQualifier\_value>

6.9. Qualifier codon\_start

Definition indicates the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature.

Value format 1 or 2 or 3

Example <INSDQualifier\_value>2</INSDQualifier\_value>

6.10. Qualifier collected\_by

Definition name of persons or institute who collected the specimen

Value format free text

Example <INSDQualifier\_value>Dan Janzen</INSDQualifier\_value>

6.11. Qualifier collection\_date

Definition date that the specimen was collected

Value format DD-Mmm-YYYY, Mmm-YYYY or YYYY

Example <INSDQualifier\_value>21-Oct-1952</INSDQualifier\_value>

<INSDQualifier\_value>Oct-1952</INSDQualifier\_value>

<INSDQualifier\_value>1952</INSDQualifier\_value>

Comment full date format DD-Mmm-YYYY is preferred; where day and/or month of collection is not known either "Mmm-YYYY" or "YYYY" can be used; three-letter month abbreviation can be one of the following: Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec.

6.12. Qualifier compare

Definition Reference details of an existing public INSD entry to which a comparison is made

Value format [accession-number.sequence-version]

Example <INSDQualifier\_value>AJ634337.1</INSDQualifier\_value>

Comment This qualifier may be used on the following features: misc\_difference, unsure, and variation. Multiple compare qualifiers with different contents are allowed within a single feature. This qualifier is not intended for large-scale annotation of variations, such as SNPs.

6.13. Qualifier cultivar

Definition cultivar (cultivated variety) of plant from which sequence was obtained

Value format free text

Example <INSDQualifier\_value>Nipponbare</INSDQualifier\_value>

<INSDQualifier\_value>Tenuifolius</INSDQualifier\_value>

<INSDQualifier\_value>Candy Cane</INSDQualifier\_value>

<INSDQualifier\_value>IR36</INSDQualifier\_value>

Comment ’cultivar’ is applied solely to products of artificial selection; use the variety qualifier for natural, named plant and fungal varieties.

6.14. Qualifier dev\_stage

Definition if the sequence was obtained from an organism in a specific developmental stage, it is specified with this qualifier

Value format free text

Example <INSDQualifier\_value>fourth instar larva</INSDQualifier\_value>

6.15. Qualifier direction

Definition direction of DNA replication

Value format left, right, or both
where left indicates toward the 5’ end of the sequence (as presented) and right indicates toward the 3’ end

Example <INSDQualifier\_value>LEFT</INSDQualifier\_value>

Comment The values left, right, and both are permitted when the direction qualifier is used to annotate a rep\_origin feature key. However, only left and right values are permitted when the direction qualifier is used to annotate an oriT feature key. The values are case-insensitive, i.e. both "RIGHT" and "right" are valid.

6.16. Qualifier EC\_number

Definition Enzyme Commission number for enzyme product of sequence

Value format free text

Example <INSDQualifier\_value>1.1.2.4</INSDQualifier\_value>

<INSDQualifier\_value>1.1.2.-</INSDQualifier\_value>

<INSDQualifier\_value>1.1.2.n</INSDQualifier\_value>

Comment valid values for EC numbers are defined in the list prepared by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (published in Enzyme Nomenclature 1992, Academic Press, San Diego, or a more recent revision thereof).The format represents a string of four numbers separated by full stops; up to three numbers starting from the end of the string can be replaced by dash "." to indicate uncertain assignment. Symbol "n" can be used in the last position instead of a number where the EC number is awaiting assignment. Please note that such incomplete EC numbers are not approved by NC-IUBMB.

6.17. Qualifier ecotype

Definition a population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat

Value Format free text

Example <INSDQualifier\_value>Columbia</INSDQualifier\_value>

Comment an example of such a population is one that has adapted hairier than normal leaves as a response to an especially sunny habitat. ’Ecotype’ is often applied to standard genetic stocks of Arabidopsis thaliana, but it can be applied to any sessile organism.

6.18. Qualifier environmental\_sample

Definition identifies sequences derived by direct molecular isolation from a bulk environmental DNA sample (by PCR with or without subsequent cloning of the product, DGGE, or other anonymous methods) with no reliable identification of the source organism. Environmental samples include clinical samples, gut contents, and other sequences from anonymous organisms that may be associated with a particular host. They do not include endosymbionts that can be reliably recovered from a particular host, organisms from a readily identifiable but uncultured field sample (e.g., many cyanobacteria), or phytoplasmas that can be reliably recovered from diseased plants (even though these cannot be grown in axenic culture)

Value format none

Comment used only with the source feature key; source feature keys containing the environmental\_sample qualifier should also contain the isolation\_source qualifier. Sequences including environmental\_sample must not include the strain qualifier.

6.19. Qualifier exception

Definition indicates that the coding region cannot be translated using standard biological rules

Value format One of the following controlled vocabulary phrases:

RNA editing

rearrangement required for product

annotated by transcript or proteomic data

Example <INSDQualifier\_value>RNA editing</INSDQualifier\_value>

<INSDQualifier\_value>rearrangement required for product</INSDQualifier\_value>

Comment only to be used to describe biological mechanisms such as RNA editing; protein translation of a CDS with an exception qualifier will be different from the according conceptual translation; must not be used where transl\_except qualifier would be adequate, e.g. in case of stop codon completion use.

6.20. Qualifier frequency

Definition frequency of the occurrence of a feature

Value format free text representing the proportion of a population carrying the feature expressed as a fraction

Example <INSDQualifier\_value>23/108</INSDQualifier\_value>

<INSDQualifier\_value>1 in 12</INSDQualifier\_value>

<INSDQualifier\_value>0.85</INSDQualifier\_value>

6.21. Qualifier function

Definition function attributed to a sequence

Value format free text

Example <INSDQualifier\_value>essential for recognition of cofactor </INSDQualifier\_value>

Comment The function qualifier is used when the gene name and/or product name do not convey the function attributable to a sequence.

6.22. Qualifier gene

Definition symbol of the gene corresponding to a sequence region

Value format free text

Example <INSDQualifier\_value>ilvE</INSDQualifier\_value>

Comment Use gene qualifier to provide the gene symbol; use standard\_name qualifier to provide the full gene name.

6.23. Qualifier gene\_synonym

Definition synonymous, replaced, obsolete or former gene symbol

Value format free text

Example <INSDQualifier\_value>Hox-3.3</INSDQualifier\_value>

in a feature where the gene qualifier value is Hoxc6

Comment used where it is helpful to indicate a gene symbol synonym; when used, a primary gene symbol must always be indicated in a gene qualifier

6.24. Qualifier germline

Definition the sequence presented has not undergone somatic rearrangement as part of an adaptive immune response; it is the unrearranged sequence that was inherited from the parental germline

Value format none

Comment germline qualifier should not be used to indicate that the source of the sequence is a gamete or germ cell; germline and rearranged qualifiers cannot be used in the same source feature; germline and rearranged qualifiers should only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)

6.25. Qualifier haplogroup

Definition name for a group of similar haplotypes that share some sequence variation. Haplogroups are often used to track migration of population groups.

Value format free text

Example <INSDQualifier\_value>H\*</INSDQualifier\_value>

6.26. Qualifier haplotype

Definition name for a specific set of alleles that are linked together on the same physical chromosome. In the absence of recombination, each haplotype is inherited as a unit, and may be used to track gene flow in populations.

Value format free text

Example <INSDQualifier\_value>Dw3 B5 Cw1 A1</INSDQualifier\_value>

6.27. Qualifier host

Definition natural (as opposed to laboratory) host to the organism from which sequenced molecule was obtained

Value format free text

Example <INSDQualifier\_value>Homo sapiens</INSDQualifier\_value>

<INSDQualifier\_value>Homo sapiens 12 year old girl</INSDQualifier\_value>

<INSDQualifier\_value>Rhizobium NGR234</INSDQualifier\_value>

6.28. Qualifier identified\_by

Definition name of the expert who identified the specimen taxonomically

Value format free text

Example <INSDQualifier\_value>John Burns</INSDQualifier\_value>

6.29. Qualifier isolate

Definition individual isolate from which the sequence was obtained

Value format free text

Example <INSDQualifier\_value>Patient #152</INSDQualifier\_value>

<INSDQualifier\_value>DGGE band PSBAC-13</INSDQualifier\_value>

6.30. Qualifier isolation\_source

Definition describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived

Value format free text

Examples <INSDQualifier\_value>rumen isolates from standard Pelleted ration-fed steer #67</INSDQualifier\_value>

<INSDQualifier\_value>permanent Antarctic sea ice</INSDQualifier\_value>

<INSDQualifier\_value>denitrifying activated sludge from carbon\_limited continuous reactor</INSDQualifier\_value>

Comment used only with the source feature key; source feature keys containing an environmental\_sample qualifier should also contain an isolation\_source qualifier

6.31. Qualifier lab\_host

Definition scientific name of the laboratory host used to propagate the source organism from which the sequenced molecule was obtained

Value format free text

Example <INSDQualifier\_value>Gallus gallus</INSDQualifier\_value>

<INSDQualifier\_value>Gallus gallus embryo</INSDQualifier\_value>

<INSDQualifier\_value>Escherichia coli strain DH5 alpha</INSDQualifier\_value>

<INSDQualifier\_value>Homo sapiens HeLa cells</INSDQualifier\_value>

Comment the full binomial scientific name of the host organism should be used when known; extra conditional information relating to the host may also be included

6.32. Qualifier lat\_lon

Definition geographical coordinates of the location where the specimen was collected

Value format free text - degrees latitude and longitude in format "d[d.dddd] N|S d[dd.dddd] W|E"

Example <INSDQualifier\_value>47.94 N 28.12 W</INSDQualifier\_value>

<INSDQualifier\_value>45.0123 S 4.1234 E</INSDQualifier\_value>

6.33. Qualifier macronuclear

Definition if the sequence shown is DNA and from an organism which undergoes chromosomal differentiation between macronuclear and micronuclear stages, this qualifier is used to denote that the sequence is from macronuclear DNA

Value format none

6.34. Qualifier map

Definition genomic map position of feature

Value format free text

Example <INSDQualifier\_value>8q12-13</INSDQualifier\_value>

6.35. Qualifier mating\_type

Definition mating type of the organism from which the sequence was obtained; mating type is used for prokaryotes, and for eukaryotes that undergo meiosis without sexually dimorphic gametes

Value format free text

Examples <INSDQualifier\_value>MAT-1</INSDQualifier\_value>

<INSDQualifier\_value>plus</INSDQualifier\_value>

<INSDQualifier\_value>-</INSDQualifier\_value>

<INSDQualifier\_value>odd</INSDQualifier\_value>

<INSDQualifier\_value>even</INSDQualifier\_value>"

Comment mating\_type qualifier values male and female are valid in the prokaryotes, but not in the eukaryotes;

for more information, see the entry for the sex qualifier.

6.36. Qualifier mobile\_element\_type

Definition type and name or identifier of the mobile element which is described by the parent feature

Value format <mobile\_element\_type>[:<mobile\_element\_name>]
where <mobile\_element\_type> is one of the following:

transposon

retrotransposon

integron

insertion sequence

non-LTR retrotransposon

SINE

MITE

LINE

other

Example <INSDQualifier\_value>transposon:Tnp9</INSDQualifier\_value>

Comment mobile\_element\_type is legal on mobile\_element feature key only. Mobile element should be used to represent both elements which are currently mobile, and those which were mobile in the past. Value "other" for <mobile\_element\_type> requires a <mobile\_element\_name>

6.37. Qualifier mod\_base

Definition abbreviation for a modified nucleotide base

Value format modified base abbreviation chosen from this Annex, Table 2

Example <INSDQualifier\_value>m5c</INSDQualifier\_value>

<INSDQualifier\_value>OTHER</INSDQualifier\_value>

Comment specific modified nucleotides not found in Section 2 of this Annex are annotated by entering OTHER as the value for the mod\_base qualifier and including a note qualifier with the full name of the modified base as its value

6.38. Qualifier mol\_type

Definition molecule type of sequence

Value format One chosen from the following:

genomic DNA

genomic RNA

mRNA

tRNA

rRNA

other RNA

other DNA

transcribed RNA

viral cRNA

unassigned DNA

unassigned RNA

Example <INSDQualifier\_value>genomic DNA</INSDQualifier\_value>

<INSDQualifier\_value>other RNA</INSDQualifier\_value>

Comment mol\_type qualifier is mandatory on the source feature key; the value "genomic DNA" does not imply that the molecule is nuclear (e.g. organelle and plasmid DNA should be described using "genomic DNA"); ribosomal RNA genes should be described using "genomic DNA"; "rRNA" should only be used if the ribosomal RNA molecule itself has been sequenced; values "other RNA" and "other DNA" should be applied to synthetic molecules, values "unassigned DNA", "unassigned RNA" should be applied where in vivo molecule is unknown.

6.39. Qualifier ncRNA\_class

Definition a structured description of the classification of the non-coding RNA described by the ncRNA parent key

Value format TYPE
where TYPE is one of the following controlled vocabulary terms or phrases:

antisense\_RNA

autocatalytically\_spliced\_intron

ribozyme

hammerhead\_ribozyme

lncRNA

RNase\_P\_RNA

RNase\_MRP\_RNA

telomerase\_RNA

guide\_RNA

rasiRNA

scRNA

siRNA

miRNA

piRNA

snoRNA

snRNA

SRP\_RNA"

vault\_RNA

Y\_RNA

other

Example <INSDQualifier\_value>autocatalytically\_spliced\_intron </INSDQualifier\_value>

<INSDQualifier\_value>siRNA</INSDQualifier\_value>

<INSDQualifier\_value>scRNA</INSDQualifier\_value>

<INSDQualifier\_value>other</INSDQualifier\_value>

Comment specific ncRNA types not yet in the ncRNA\_class controlled vocabulary can be annotated by entering "other" as the ncRNA\_class qualifier value, and providing a brief explanation of novel ncRNA\_class in a note qualifier

6.40. Qualifier note

Definition any comment or additional information

Value format free text

Example <INSDQualifier\_value>A comment about the feature</INSDQualifier\_value>

6.41. Qualifier number

Definition a number to indicate the order of genetic elements (e.g. exons or introns) in the 5’ to 3’ direction

Value format free text (with no whitespace characters)

Example <INSDQualifier\_value>4</INSDQualifier\_value>

<INSDQualifier\_value>6B</INSDQualifier\_value>

Comment text limited to integers, letters or combination of integers and/or letters represented as a data value that contains no whitespace characters; any additional terms should be included in a standard\_name qualifier. Example: a number qualifier with a value of 2A and a standard\_name qualifier with a value of long

6.42. Qualifier operon

Definition name of the group of contiguous genes transcribed into a single transcript to which that feature belongs

Value format free text

Example <INSDQualifier\_value>lac</INSDQualifier\_value>

Comment valid only on Prokaryota-specific features

6.43. Qualifier organelle

Definition type of membrane-bound intracellular structure from which the sequence was obtained

Value format One of the following controlled vocabulary terms and phrases:

chromatophore

hyrogenosome

mitochondrion

nucleomorph

plastid

mitochondrion:kinetoplast

plastid:chloroplast

plastid:apicoplast

plastid:chromoplast

plastid:cyanelle

plastid:leucoplast

plastid:proplastid,

Examples <INSDQualifier\_value>chromatophore</INSDQualifier\_value>

<INSDQualifier\_value>hydrogenosome</INSDQualifier\_value>

<INSDQualifier\_value>mitochondrion</INSDQualifier\_value>

<INSDQualifier\_value>nucleomorph</INSDQualifier\_value>

<INSDQualifier\_value>plastid</INSDQualifier\_value>

<INSDQualifier\_value>mitochondrion:kinetoplast</INSDQualifier\_value>

<INSDQualifier\_value>plastid:chloroplast</INSDQualifier\_value>

<INSDQualifier\_value>plastid:apicoplast</INSDQualifier\_value>

<INSDQualifier\_value>plastid:chromoplast</INSDQualifier\_value>

<INSDQualifier\_value>plastid:cyanelle</INSDQualifier\_value>

<INSDQualifier\_value>plastid:leucoplast</INSDQualifier\_value>

<INSDQualifier\_value>plastid:proplastid</INSDQualifier\_value>

6.44. Qualifier organism

Definition scientific name of the organism that provided the sequenced genetic material, if known, or the available taxonomic information if the organism is unclassified; or an indication that the sequence is a synthetic construct

Value format free text

Example <INSDQualifier\_value>Homo sapiens</INSDQualifier\_value>

6.45. Qualifier PCR\_primers

Definition PCR primers that were used to amplify the sequence. A single /PCR\_primers qualifier should contain all the primers used for a single PCR reaction. If multiple forward or reverse primers are present in a single PCR reaction, multiple sets of fwd\_name/fwd\_seq or rev\_name/rev\_seq values will be present

Value format [fwd\_name: XXX1, ]fwd\_seq: xxxxx1,[fwd\_name: XXX2, ]fwd\_seq: xxxxx2, [rev\_name: YYY1, ]rev\_seq: yyyyy1,[rev\_name: YYY2, ]rev\_seq: yyyyy2</INSDQualifier\_value>

Example <INSDQualifier\_value>fwd\_name: CO1P1, fwd\_seq: ttgattttttggtcayccwgaagt,rev\_name: CO1R4, rev\_seq: ccwvytardcctarraartgttg</INSDQualifier\_value>

<INSDQualifier\_value>fwd\_name: hoge1, fwd\_seq: cgkgtgtatcttact, rev\_name: hoge2, rev\_seq: cg&lt;i&gt;gtgtatcttact</INSDQualifier\_value>

<INSDQualifier\_value>fwd\_name: CO1P1, fwd\_seq: ttgattttttggtcayccwgaagt, fwd\_name: CO1P2, fwd\_seq: gatacacaggtcayccwgaagt, rev\_name: CO1R4, rev\_seq: ccwvytardcctarraartgttg</INSDQualifier\_value>

Comment fwd\_seq and rev\_seq are both mandatory; fwd\_name and rev\_name are both optional.

Both sequences should be presented in 5’>3’ order. The sequences should be given in the symbols from Section 1 of this Annex, except for the modified bases; those must be enclosed within angle brackets < >. In XML, the angle brackets < and > must be substituted with &lt; and &gt; since they are reserved characters in XML.

6.46. Qualifier phenotype

Definition phenotype conferred by the feature, where phenotype is defined as a physical, biochemical or behavioural characteristic or set of characteristics

Value format free text

Example <INSDQualifier\_value>erythromycin resistance</INSDQualifier\_value>

6.47. Qualifier plasmid

Definition name of naturally occurring plasmid from which the sequence was obtained, where plasmid is defined as an independently replicating genetic unit that cannot be described by chromosome or segment qualifiers

Value format free text

Example <INSDQualifier\_value>pC589</INSDQualifier\_value>

6.48. Qualifier pop\_variant

Definition name of subpopulation or phenotype of the sample from which the sequence was derived

Value format free text

Example <INSDQualifier\_value>pop1</INSDQualifier\_value>

<INSDQualifier\_value>Bear Paw</INSDQualifier\_value>

6.49. Qualifier product

Definition name of the product associated with the feature, e.g. the mRNA of an mRNA feature, the polypeptide of a CDS, the mature peptide of a mat\_peptide, etc.

Value format free text

Example <INSDQualifier\_value>trypsinogen</INSDQualifier\_value> (when qualifier appears in CDS feature)

<INSDQualifier\_value>trypsin</INSDQualifier\_value> (when qualifier appears in mat\_peptide feature)

<INSDQualifier\_value>XYZ neural-specific transcript</INSDQualifier\_value> (when qualifier appears in mRNA feature)

6.50. Qualifier protein\_id

Definition protein sequence identification number, an integer used in a sequence listing to designate the protein sequence encoded by the coding sequence identified in the corresponding CDS feature key

Value format an integer greater than zero

Example <INSDQualifier\_value>89</INSDQualifier\_value>

6.51. Qualifier proviral

Definition this qualifier is used to flag sequence obtained from a virus or phage that is integrated into the genome of another organism

Value format none

6.52. Qualifier pseudo

Definition indicates that this feature is a non-functional version of the element named by the feature key

Value format none

Comment The qualifier pseudo should be used to describe non-functional genes that are not formally described as pseudogenes, e.g. CDS has no translation due to other reasons than pseudogenisation events. Other reasons may include sequencing or assembly errors. In order to annotate pseudogenes the qualifier pseudogene must be used, indicating the TYPE of pseudogene.

6.53. Qualifier pseudogene

Definition indicates that this feature is a pseudogene of the element named by the feature key

Value format TYPE

where TYPE is one of the following controlled vocabulary terms or phrases:
processed
unprocessed
unitary
allelic
unknown

Example <INSDQualifier\_value>processed</INSDQualifier\_value>

<INSDQualifier\_value>unprocessed</INSDQualifier\_value>

<INSDQualifier\_value>unitary</INSDQualifier\_value>

<INSDQualifier\_value>allelic</INSDQualifier\_value>

<INSDQualifier\_value>unknown</INSDQualifier\_value>

Comment Definitions of TYPE values:

processed - the pseudogene has arisen by reverse transcription of a mRNA into cDNA, followed by reintegration into the genome. Therefore, it has lost any intron/exon structure, and it might have a pseudo-polyA-tail.

unprocessed - the pseudogene has arisen from a copy of the parent gene by duplication followed by accumulation of random mutation. The changes, compared to their functional homolog, include insertions, deletions, premature stop codons, frameshifts and a higher proportion of non-synonymous versus synonymous substitutions.

unitary - the pseudogene has no parent. It is the original gene, which is functional is some species but disrupted in some way (indels, mutation, recombination) in another species or strain.

allelic - a (unitary) pseudogene that is stable in the population but importantly it has a functional alternative allele also in the population. i.e., one strain may have the gene, another strain may have the pseudogene. MHC haplotypes have allelic pseudogenes.

unknown - the submitter does not know the method of pseudogenisation.

6.54. Qualifier rearranged

Definition the sequence presented in the entry has undergone somatic rearrangement as part of an adaptive immune response; it is not the unrearranged sequence that was inherited from the parental germline

Value format none

Comment The rearranged qualifier should not be used to annotate chromosome rearrangements that are not involved in an adaptive immune response; germline and rearranged qualifiers cannot be used in the same source feature; germline and rearranged qualifiers should only be used for molecules that can undergo somatic rearrangements as part of an adaptive immune response; these are the T-cell receptor (TCR) and immunoglobulin loci in the jawed vertebrates, and the unrelated variable lymphocyte receptor (VLR) locus in the jawless fish (lampreys and hagfish); germline and rearranged qualifiers should not be used outside of the Craniata (taxid=89593)

6.55. Qualifier replace

Definition indicates that the sequence identified in a feature’s location is replaced by the sequence shown in the qualifier’s value; if no sequence (i.e., no value) is contained within the qualifier, this indicates a deletion

Value format free text

Example <INSDQualifier\_value>a</INSDQualifier\_value>

<INSDQualifier\_value></INSDQualifier\_value> - for a deletion

6.56. Qualifier ribosomal\_slippage

Definition during protein translation, certain sequences can program ribosomes to change to an alternative reading frame by a mechanism known as ribosomal slippage

Value format none

Comment a join operator, e.g.: [join(486..1784,1787..4810)] should be used in the CDS spans to indicate the location of ribosomal\_slippage

6.57. Qualifier rpt\_family

Definition type of repeated sequence; "Alu" or "Kpn", for example

Value format free text

Example <INSDQualifier\_value>Alu</INSDQualifier\_value>

6.58. Qualifier rpt\_type

Definition organization of repeated sequence

Value format One of the following controlled vocabulary terms:

tandem

inverted

flanking

terminal

direct

dispersed

other

Example <INSDQualifier\_value>INVERTED</INSDQualifier\_value>

Comment the values are case-insensitive, i.e. both "INVERTED" and "inverted" are valid; Definitions of the values:

tandem - a repeat that exists adjacent to another in the same orientation;

inverted - a repeat which occurs as part of as set (normally a part) organized in the reverse orientation;

flanking - a repeat lying outside the sequence for which it has functional significance (eg. transposon insertion target sites);

terminal - a repeat at the ends of and within the sequence for which it has functional significance (eg. transposon LTRs);

direct - a repeat that exists not always adjacent but is in the same orientation;

dispersed - a repeat that is found dispersed throughout the genome;

other - a repeat exhibiting important attributes that cannot be described by other values.

6.59. Qualifier rpt\_unit\_range

Definition location (range) of a repeating unit

Value format <base\_range> - where <base\_range> is the first and last base (separated by two dots) of a repeating unit

Example <INSDQualifier\_value>202..245</INSDQualifier\_value>

Comment used to indicate the base range of the sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat\_region.

6.60. Qualifier rpt\_unit\_seq

Definition identity of a repeat sequence

Value format free text

Example <INSDQualifier\_value>aagggc</INSDQualifier\_value>

<INSDQualifier\_value>ag(5)tg(8)</INSDQualifier\_value>

<INSDQualifier\_value>(AAAGA)6(AAAA)1(AAAGA)12</INSDQualifier\_value>

Comment used to indicate the literal sequence that constitutes a repeating unit within the region specified by the feature keys oriT and repeat\_region

6.61. Qualifier satellite

Definition identifier for a satellite DNA marker, compose of many tandem repeats (identical or related) of a short basic repeated unit

Value format <satellite\_type>[:<class>][ <identifier>] - where <satellite\_type> is one of the following:

satellite;

microsatellite;

minisatellite

Example <INSDQualifier\_value>satellite: S1a</INSDQualifier\_value>

<INSDQualifier\_value>satellite: alpha</INSDQualifier\_value>

<INSDQualifier\_value>satellite: gamma III</INSDQualifier\_value>

<INSDQualifier\_value>microsatellite: DC130</INSDQualifier\_value>

Comment many satellites have base composition or other properties that differ from those of the rest of the genome that allows them to be identified.

6.62. Qualifier segment

Definition name of viral or phage segment sequenced

Value format free text

Example <INSDQualifier\_value>6</INSDQualifier\_value>

6.63. Qualifier serotype

Definition serological variety of a species characterized by its antigenic properties

Value format free text

Example <INSDQualifier\_value>B1</INSDQualifier\_value>

Comment used only with the source feature key; the Bacteriological Code recommends the use of the term ’serovar’ instead of ’serotype’ for the prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B "Infraspecific Terms".

6.64. Qualifier serovar

Definition serological variety of a species (usually a prokaryote) characterized by its antigenic properties

Value format free text

Example <INSDQualifier\_value>O157:H7</INSDQualifier\_value>

Comment used only with the source feature key; the Bacteriological Code recommends the use of the term ’serovar’ instead of ’serotype’ for prokaryotes; see the International Code of Nomenclature of Bacteria (1990 Revision) Appendix 10.B "Infraspecific Terms".

6.65. Qualifier sex

Definition sex of the organism from which the sequence was obtained; sex is used for eukaryotic organisms that undergo meiosis and have sexually dimorphic gametes

Value format free text

Examples <INSDQualifier\_value>female</INSDQualifier\_value>

<INSDQualifier\_value>male</INSDQualifier\_value>

<INSDQualifier\_value>hermaphrodite</INSDQualifier\_value>

<INSDQualifier\_value>unisexual</INSDQualifier\_value>

<INSDQualifier\_value>bisexual</INSDQualifier\_value>

<INSDQualifier\_value>asexual</INSDQualifier\_value>

<INSDQualifier\_value>monoecious</INSDQualifier\_value> [or monecious]

<INSDQualifier\_value>dioecious</INSDQualifier\_value> [or diecious]

Comment The sex qualifier should be used (instead of mating\_type qualifier) in the Metazoa, Embryophyta, Rhodophyta & Phaeophyceae; mating\_type qualifier should be used (instead of sex qualifier) in the Bacteria, Archaea & Fungi; neither sex nor mating\_type qualifiers should be used in the viruses; outside of the taxa listed above, mating\_type qualifier should be used unless the value of the qualifier is taken from the vocabulary given in the examples above

6.66. Qualifier standard\_name

Definition accepted standard name for this feature

Value format free text

Example <INSDQualifier\_value>dotted</INSDQualifier\_value>

Comment use standard\_name qualifier to give full gene name, but use gene qualifier to give gene symbol (in the above example gene qualifier value is Dt).

6.67. Qualifier strain

Definition strain from which sequence was obtained

Value format free text

Example <INSDQualifier\_value>BALB/c</INSDQualifier\_value>

Comment entries including strain qualifier must not include the environmental\_sample qualifier

6.68. Qualifier sub\_clone

Definition sub-clone from which sequence was obtained

Value format free text

Example <INSDQualifier\_value>lambda-hIL7.20g</INSDQualifier\_value>

Comment not more than one sub\_clone should be specified for a given source feature; to indicate that the sequence was obtained from multiple sub\_clones, multiple source features should be given

6.69. Qualifier sub\_species

Definition name of sub-species of organism from which sequence was obtained

Value format free text

Example <INSDQualifier\_value>lactis</INSDQualifier\_value>

6.70. Qualifier sub\_strain

Definition name or identifier of a genetically or otherwise modified strain from which sequence was obtained, derived from a parental strain (which should be annotated in the strain qualifier). sub\_strain from which sequence was obtained

Value format free text

Example <INSDQualifier\_value>abis</INSDQualifier\_value>

Comment If the parental strain is not given, this should be annotated in the strain qualifier instead of sub\_strain. For example, either a strain qualifier with the value K-12 and a substrain qualifier with the value MG1655 or a strain qualifier with the value MG1655

6.71. Qualifier tag\_peptide

Definition base location encoding the polypeptide for proteolysis tag of tmRNA and its termination codon

Value format <base\_range> - where <base\_range> provides the first and last base (separated by two dots) of the location for the proteolysis tag

Example <INSDQualifier\_value>90..122</INSDQualifier\_value>

Comment it is recommended that the amino acid sequence corresponding to the tag\_peptide be annotated by describing a 5’ partial CDS feature; e.g. CDS with a location of <90..122

6.72. Qualifier tissue\_lib

Definition tissue library from which sequence was obtained

Value format free text

Example <INSDQualifier\_value>tissue library 772</INSDQualifier\_value>

6.73. Qualifier tissue\_type

Definition tissue type from which the sequence was obtained

Value format free text

Example <INSDQualifier\_value>liver</INSDQualifier\_value>

6.74. Qualifier transl\_except

Definition translational exception: single codon the translation of which does not conform to genetic code defined by organism or transl\_table.

Value format (pos:location,aa:<amino\_acid>) where <amino\_acid> is the amino acid coded by the codon at the base\_range position

Example <INSDQualifier\_value>(pos:213..215,aa:Trp) </INSDQualifier\_value>

<INSDQualifier\_value>(pos:462..464,aa:OTHER) </INSDQualifier\_value>

<INSDQualifier\_value>(pos:1017,aa:TERM) </INSDQualifier\_value>

<INSDQualifier\_value>(pos:2000..2001,aa:TERM) </INSDQualifier\_value>

<INSDQualifier\_value>(pos:X22222:15..17,aa:Ala) </INSDQualifier\_value>

Comment if the amino acid is not one of the specific amino acids listed in Section 3 of this Annex, use OTHER as <amino\_acid> and provide the name of the unusual amino acid in a note qualifier; for modified amino-acid selenocysteine use three letter code ’Sec’ (one letter code ’U’ in amino-acid sequence) for <amino \_acid>; for partial termination codons where TAA stop codon is completed by the addition of 3’ A residues to the mRNA either a single base\_position or a base\_range is used for the location, see the third and fourth examples above, in conjunction with a note qualifier indicating ‘stop codon completed by the addition of 3’ A residues to the mRNA’.

6.75. Qualifier transl\_table

Definition definition of genetic code table used if other than universal or standard genetic code table. Tables used are described in this Annex

Value format <integer>
where <integer> is the number assigned to the genetic code table

Example <INSDQualifier\_value>3</INSDQualifier\_value> - example where the yeast mitochondrial code is to be used

Comment if the transl\_table qualifier is not used to further annotate a CDS feature key, then the CDS is translated using the Standard Code (i.e. Universal Genetic Code). Genetic code exceptions outside the range of specified tables are reported in transl\_except qualifiers.

6.76. Qualifier trans\_splicing

Definition indicates that exons from two RNA molecules are ligated in intermolecular reaction to form mature RNA

Value format none

Comment should be used on features such as CDS, mRNA and other features that are produced as a result of a trans-splicing event. This qualifier should be used only when the splice event is indicated in the "join" operator, e.g. join(complement(69611..69724),139856..140087)

6.77. Qualifier translation

Definition one-letter abbreviated amino acid sequence derived from either the standard (or universal) genetic code or the table as specified in a transl\_table qualifier and as determined by an exception in the transl\_except qualifier

Value format contiguous string of one-letter amino acid abbreviations from Section 3 of this Annex, "X" is to be used for AA exceptions.

Example <INSDQualifier\_value>MASTFPPWYRGCASTPSLKGLIMCTW</INSDQualifier\_value>

Comment to be used with CDS feature only; must be accompanied by protein\_id qualifier when the translation product contains four or more amino acids; see transl\_table for definition and location of genetic code Tables; only one of the qualifiers translation, pseudo and pseudogene are permitted to further annotate a CDS feature.

6.78. Qualifier variety

Definition variety (= varietas, a formal Linnaean rank) of organism from which sequence was derived.

Value format free text

Example <INSDQualifier\_value>insularis</INSDQualifier\_value>

Comment use the cultivar qualifier for cultivated plant varieties, i.e., products of artificial selection; varieties other than plant and fungal variatas should be annotated via a note qualifier, e.g. with the value <INSDQualifier\_value>breed:Cukorova</INSDQualifier\_value>

## SECTION 7: FEATURE KEYS FOR AMINO ACID SEQUENCES

This section contains the list of allowed feature keys to be used for amino acid sequences. The feature keys are listed in alphabetic order.

7.1. Feature Key ACT\_SITE

Definition Amino acid(s) involved in the activity of an enzyme

Optional qualifiers NOTE

Comment Each amino acid resdidue of the active site should be annotated separately with the ACT\_SITE feature key. The corresponding amino acid residue number should be provided as the location descriptor in the feature location element.

7.2. Feature Key BINDING

Definition Binding site for any chemical group (co-enzyme, prosthetic group, etc.). The chemical nature of the group is indicated in the NOTE qualifier

Mandatory qualifiers NOTE

Comment Examples of values for the “NOTE” qualifier: “Heme (covalent)” and “Chloride.” Where appropriate, the features keys CA\_BIND, DNA\_BIND, METAL,and NP\_BIND should be used rather than BINDING.

7.3. Feature Key CA\_BIND

Definition Extent of a calcium-binding region

Optional qualifiers NOTE

7.4. Feature Key CARBOHYD

Definition Glycosylation site

Mandatory qualifiers NOTE

Comment This key describes the occurrence of the attachment of a glycan (mono- or polysaccharide) to a residue of the protein. If the nature of the reducing terminal sugar is known, its abbreviation is shown between parentheses. If three dots ’...’ follow the abbreviation this indicates an extension of the carbohydrate chain. Conversely no dots means that a monosaccharide is linked. The type of linkage (C-, N- or O-linked) to the protein is indicated in the “NOTE” qualifier. Examples of values used in the “NOTE” qualifier: O-linked (GlcNAc); C-linked (Man); N-linked (GlcNAc...); and O-linked (Glc...).

7.5. Feature Key CHAIN

Definition Extent of a polypeptide chain in the mature protein

Optional qualifiers NOTE

7.6. Feature Key COILED

Definition Extent of a coiled-coil region

Optional qualifiers NOTE

7.7. Feature Key COMPBIAS

Definition Extent of a compositionally biased region

Optional qualifiers NOTE

7.8. Feature Key CONFLICT

Definition Different sources report differing sequences.

Optional qualifiers NOTE

7.9. Feature Key CROSSLNK

Definition Post translationally formed amino acid bonds.

Mandatory qualifiers NOTE

Comment Covalent linkages of various types formed between two proteins (interchain cross-links) or between two parts of the same protein (intrachain cross-links); except for cross-links formed by disulfide bonds, for which the “DISULFID” feature key is to be used. For an interchain cross-link, the location descriptor in the feature location element is the residue number of the amino acid cross-linked to the other protein. For an intrachain cross-link, the location descriptors in the feature location element are the residue numbers of the cross-linked amino acids in conjunction with the “join” location operator, e.g. “join(42,50).” The NOTE qualifier indicates the nature of the cross-link; at least specifying the name of the conjugate and the identity of the two amino acids involved. Examples of values for the “NOTE” qualifier: “Isoglutamyl cysteine thioester (Cys-Gln);” “Beta-methyllanthionine (Cys-Thr);” and “Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in ubiquitin)”

7.10. Feature Key DISULFID

Definition Disulfide bond

Optional qualifiers NOTE

Comment For an interchain disulfide bond, the location descriptor in the feature location element is the residue number of the cysteine linked to the other protein. For an intrachain cross-link, the location descriptors in the feature location element are the residue numbers of the linked cysteines in conjunction with the “join” location operator, e.g. “join(42,50)”. For interchain disulfide bonds, the NOTE qualifier indicates the nature of the cross-link, by identifying the other protein, for example, “Interchain (between A and B chains)”

7.11. Feature Key DNA\_BIND

Definition Extent of a DNA-binding region

Mandatory qualifiers NOTE

Comment The nature of the DNA-binding region is given in the NOTE qualifier. Examples of values for the “NOTE” qualifier: “Homeobox” and “Myb 2”

7.12. Feature Key DOMAIN

Definition Extent of a domain, which is defined as a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold

Mandatory qualifiers NOTE

Comment The domain type is given in the NOTE qualifier. Where several copies of a domain are present, the domains are numbered. Examples of values for the “NOTE” qualifier: “Ras-GAP” and “Cadherin 1”

7.13. Feature Key HELIX

Definition Secondary structure: Helices, for example, Alpha-helix; 3(10) helix; or Pi-helix

Optional qualifiers NOTE

Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a ’loop’ or ’random-coil’ structure.

7.14. Feature Key INIT\_MET

Definition Initiator methionine

Optional qualifiers NOTE

Comment The location descriptor in the feature location element is “1”. This feature key indicates the N-terminal methionine is cleaved off. This feature is not used when the initiator methionine is not cleaved off.

7.15. Feature Key INTRAMEM

Definition Extent of a region located in a membrane without crossing it

Optional qualifiers NOTE

7.16. Feature Key LIPID

Definition Covalent binding of a lipid moiety

Mandatory qualifiers NOTE

Comment The chemical nature of the bound lipid moiety is given in the NOTE qualifier, indicating at least the name of the lipidated amino acid. Examples of values for the “NOTE” qualifier: “N-myristoyl glycine”; “GPI-anchor amidated serine” and “S-diacylglycerol cysteine.”

7.17. Feature Key METAL

Definition Binding site for a metal ion.

Mandatory qualifiers NOTE

Comment The NOTE qualifier indicates the nature of the metal. Examples of values for the “NOTE” qualifier: “Iron; catalytic” and “Copper”.

7.18. Feature Key MOD\_RES

Definition Posttranslational modification of a residue

Mandatory qualifiers NOTE

Comment The chemical nature of the modified residue is given in the NOTE qualifier, indicating at least the name of the post-translationally modified amino acid. If the modified amino acid is listed in Table 4 of this Annex, the abbreviation may be used in place of the the full name. Examples of values for the “NOTE” qualifier: “N-acetylalanine”; “3-Hyp”; and “MeLys” or “N-6-methyllysine"

7.19. Feature Key MOTIF

Definition Short (up to 20 amino acids) sequence motif of biological interest

Optional qualifiers NOTE

7.20. Feature Key MUTAGEN

Definition Site which has been experimentally altered by mutagenesis

Optional qualifiers NOTE

7.21. Feature Key NON\_STD

Definition Non-standard amino acid

Optional qualifiers NOTE

Comment This key describes the occurrence of non-standard amino acids selenocysteine (U) and pyrrolysine (O) in the amino acid sequence.

7.22. Feature Key NON\_TER

Definition The residue at an extremity of the sequence is not the terminal residue

Optional qualifiers NOTE

Comment If applied to position 1, this means that the first position is not the N-terminus of the complete molecule. If applied to the last position, it means that this position is not the C-terminus of the complete molecule.

7.23. Feature Key NP\_BIND

Definition Extent of a nucleotide phosphate-binding region

Mandatory qualifiers NOTE

Comment The nature of the nucleotide phosphate is indicated in the NOTE qualifier. Examples of values for the “NOTE” qualifier: “ATP” and “FAD”.

7.24. Feature Key PEPTIDE

Definition Extent of a released active peptide

Optional qualifiers NOTE

7.25. Feature Key PROPEP

Definition Extent of a propeptide

Optional qualifiers NOTE

7.26. Feature Key REGION

Definition Extent of a region of interest in the sequence

Optional qualifiers NOTE

7.27. Feature Key REPEAT

Definition Extent of an internal sequence repetition

Optional qualifiers NOTE

7.28. Feature Key SIGNAL

Definition Extent of a signal sequence (prepeptide)

Optional qualifiers NOTE

7.29. Feature Key SITE

Definition Any interesting single amino-acid site on the sequence that is not defined by another feature key. It can also apply to an amino acid bond which is represented by the positions of the two flanking amino acids

Mandatory qualifier NOTE

Comment When SITE is used to annotate a modified amino acid the value for the qualifier “NOTE” must either be an abbreviation set forth in Section 4 of this Annex, Table 4, or the complete, unabbreviated name of the modified amino acid.

7.30. Feature Key SOURCE

Definition Identifies the source of the sequence; this key is mandatory; every sequence will have a single SOURCE feature spanning the entire sequence

Mandatory qualifiers MOL\_TYPE

ORGANISM

Optional qualifiers NOTE

7.31. Feature Key STRAND

Definition Secondary structure: Beta-strand; for example Hydrogen bonded beta-strand or residue in an isolated beta-bridge

Optional qualifiers NOTE

Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a ’loop’ or ’random-coil’ structure.

7.32. Feature Key TOPO\_DOM

Definition Topological domain

Optional qualifiers NOTE

7.33. Feature Key TRANSMEM

Definition Extent of a transmembrane region

Optional qualifiers NOTE

7.34. Feature Key TRANSIT

Definition Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.)

Optional qualifiers NOTE

7.35. Feature Key TURN

Definition Secondary structure Turns, for example, H-bonded turn (3-turn, 4-turn or 5-turn)

Optional qualifiers NOTE

Comment This feature is used only for proteins whose tertiary structure is known. Only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a ’loop’ or ’random-coil’ structure.

7.36. Feature Key UNSURE

Definition Uncertainties in the amino acid sequence

Optional qualifiers NOTE

Comment Used to describe region(s) of an amino acid sequence for which the authors are unsure about the sequence presentation.

7.37. Feature Key VARIANT

Definition Authors report that sequence variants exist.

Optional qualifiers NOTE

7.38. Feature Key VAR\_SEQ

Definition Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting

Optional qualifiers NOTE

7.39. Feature Key ZN\_FING

Definition Extent of a zinc finger region

Mandatory qualifiers NOTE

Comment The type of zinc finger is indicated in the NOTE qualifier. For example: “GATA-type” and “NR C4-type”

##

## SECTION 8: QUALIFIERS FOR AMINO ACID SEQUENCES

This section contains the list of allowed qualifiers to be used for amino acid sequences.

8.1. Qualifier MOL\_TYPE

Definition In vivo molecule type of sequence

Value format protein

Example <INSDQualifier\_value>protein</INSDQualifier\_value>

Comment The "MOL\_TYPE" qualifier is mandatory on the SOURCE feature key.

8.2. Qualifier NOTE

Definition Any comment or additional information

Value format free text

Example <INSDQualifier\_value>Heme (covalent)</INSDQualifier\_value>

Comment The “NOTE” qualifier is mandatory for the feature keys: BINDING; CARBOHYD; CROSSLNK; DISULFID; DNA\_BIND; DOMAIN; LIPID; METAL; MOD\_RES; NP\_BIND and ZN\_FING

8.3. Qualifier ORGANISM

Definition Scientific name of the organism that provided the peptide

Value format free text

Example <INSDQualifier\_value>Homo sapiens</INSDQualifier\_value>

Comment The “ORGANISM” qualifier is mandatory for the SOURCE feature key.

## SECTION 9: GENETIC CODES TABLES

Table 5 reproduces Genetic Code Tables to be used for translating coding sequences. The value for the trans\_table qualifier is the number assigned to the corresponding genetic code table. Where a CDS feature is described with a translation qualifier but not a transl\_table qualifier, the 1 - Standard Code is used by default for translation. (Note: Genetic code tables 7, 8, and 17 to 20 do not exist, therefore these numbers do not appear in Table 5.)

Table 5: Genetic Code Tables

|  |
| --- |
| **1 - Standard Code** |
|  AAs = FFLLSSSSYY\*\*CC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = ---M---------------M---------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **2 - Vertebrate Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS\*\*VVVVAAAADDEEGGGGStarts = --------------------------------MMMM---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **3 - Yeast Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*\*CCWWTTTTPPPPHHQQRRRRIIMMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = ----------------------------------MM----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **4 - Mold, Protozoan, Coelenterate Mitochondrial Code &****Mycoplasma/Spiroplasma Code** |
|  AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = --MM---------------M------------MMMM---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **5 - Invertebrate Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSSSVVVVAAAADDEEGGGGStarts = ---M----------------------------MMMM---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **6 - Ciliate, Dasycladacean and Hexamita Nuclear Code** |
|  AAs = FFLLSSSSYYQQCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = -----------------------------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **9 - Echinoderm and Flatworm Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGGStarts = -----------------------------------M---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |

|  |
| --- |
| **10 - Euplotid Nuclear Code** |
|  AAs = FFLLSSSSYY\*\*CCCWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = -----------------------------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **11 - Bacterial and Plant Plastid Code** |
|  AAs = FFLLSSSSYY\*\*CC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = ---M---------------M------------MMMM---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **12 - Alternative Yeast Nuclear Code** |
|  AAs = FFLLSSSSYY\*\*CC\*WLLLSPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = -------------------M---------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **13- Ascidian Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSGGVVVVAAAADDEEGGGGStarts = ---M------------------------------MM---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **14 - Alternative Flatworm Mitochondrial Code** |
|  AAs = FFLLSSSSYYY\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGGStarts = -----------------------------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **15 - Blepharisma Nuclear Code** |
|  AAs = FFLLSSSSYY\*QCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = -----------------------------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **16 - Chlorophycean Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*LCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = -----------------------------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **21 - Trematode Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNNKSSSSVVVVAAAADDEEGGGGStarts = -----------------------------------M---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |

|  |
| --- |
| **22 - Scenedesmus obliquus Mitochondrial Code** |
|  AAs = FFLLSS\*SYY\*LCC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = -----------------------------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **23 - Thraustochytrium Mitochondrial Code** |
|  AAs = FF\*LSSSSYY\*\*CC\*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = --------------------------------M--M---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **24 - Pterobranchia Mitochondrial Code** |
|  AAs = FFLLSSSSYY\*\*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSSKVVVVAAAADDEEGGGGStarts = ---M---------------M---------------M---------------M------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |
| **25 - Candidate Division SR1 and Gracilibacteria Code** |
|  AAs = FFLLSSSSYY\*\*CCGWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGGStarts = ---M---------------M---------------M----------------------------Base1 = ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaaggggggggggggggggBase2 = ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggBase3 = tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag |

[Annex II to ST.26 follows]