Annex B.1.

Controlled vocabularies

Contents

1 List of nucleotides

The nucleotide base codes to be used in sequence listings are presented in Table 1. 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
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

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

Abbreviation	Modified Nucleotide
q	queosine
s2c	2-thiocytidine
s2t	5-methyl-2-thiouridine
s2u	2-thiouridine
s4u	4-thiouridine
t	5-methyluridine
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)

3 List of amino acids

The amino acid codes to be used in sequence are presented in Table 3.

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

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

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 are 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 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 includes amino acid conceptual translation

Optional qualifiers
allele
artificial_location
codon_start
EC_number
exception
function
gene
gene_synonym
map
note
number
operon
product
pseudo
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 oruniversal 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

---

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 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 standard_name

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 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 ofseveral 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 standard_name

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

		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 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	feature sequence is different from that presented in the entry 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 that arises as a result of genetic manipulation (e.g. site directed mutagenesis); 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 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 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 paragraph 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 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 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 standard_name
	Parent Key	CDS
	Organism scope	eukaryotes

5.29.	Feature Key	operon
	Definition	region containing polycistronic transcript containing genes that encode enzymes that are in the same metabolic pathway and regulatory sequences
	Mandatory qualifiers	operon
	Optional qualifiers	allele function map note phenotype pseudo 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
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
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
standard_name

5.44. Feature Key source

Definition identifies the biological source of the specified span of the sequence; this key is mandatory; more than one source key per sequence is allowed; every entry/record will have, as a minimum, either a single source key spanning the entire sequence or multiple source keys, which together, span the entire sequence

Mandatory qualifiers organism
mol_type

Optional qualifiers cell_line
cell_type
chromosome
clone
clone_lib
collected_by
collection_date
country
cultivar
dev_stage
ecotype
environmental_sample
focus

frequency
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
transgenic
variety

Molecule scope any

Comment transgenic sequences must have at least two source feature keys; in a transgenic sequence the source feature key describing the organism that is the recipient of the DNA must span the entire sequence.

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 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 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 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

		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 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 term “3’UTR” must be represented as the term “3&apos;UTR” in the XML, 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 term “5’UTR” must be represented as the term “5&apos;UTR” in the XML, 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)

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:<base_range>,aa:<amino_acid>) where <base_range> is the position of the anticodon and <amino_acid> is the abbreviation for the amino acid encoded
	Example	<INSDQualifier_value>(pos:34..36,aa:Phe)</INSDQualifier_value>

6.3.	Qualifier	artificial_location
	Definition	indicates that location of the CDS or mRNA is modified to adjust for the presence of a frameshift or internal stop codon and not because of biological processing between the regions
	Value format	"heterogeneous population sequenced", "low-quality sequence region"
	Example	<INSDQualifier_value>heterogeneous population sequenced</INSDQualifier_value> <INSDQualifier_value>low-quality sequence region</INSDQualifier_value>
	Comment	expected to be used only for genome-scale annotation

6.4.	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.

6.5.	Qualifier	cell_line
	Definition	cell line from which the sequence was obtained

	Value format	free text
	Example	<INSDQualifier_value>MCF7</INSDQualifier_value>
6.6.	Qualifier	cell_type
	Definition	cell type from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>leukocyte</INSDQualifier_value>
6.7.	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>
6.8.	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; to indicate that the sequence was obtained from multiple clones, multiple source features should be given.
6.9.	Qualifier	clone_lib
	Definition	clone library from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>lambda-hIL7</INSDQualifier_value>
6.10.	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>
6.11.	Qualifier	collected_by
	Definition	name of the person who collected the specimen
	Value format	free text
	Example	<INSDQualifier_value>Dan Janzen</INSDQualifier_value>

6.12.	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.

6.13.	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.

6.14.	Qualifier	country
	Definition	locality of isolation of the sequenced organism indicated in terms of political names for nations, oceans or seas
	Value format	<country_value> where <country_value> is any value from the controlled vocabulary in paragraph 10 of this Annex
	Example	<INSDQualifier_value>Canada</INSDQualifier_value> <INSDQualifier_value>France</INSDQualifier_value> <INSDQualifier_value>Atlantic Ocean</INSDQualifier_value>
	Comment	Intended to provide a reference to the site where the source organism was isolated or sampled. Regions and localities may be indicated in a note qualifier. Note that the physical geography of the isolation or sampling site should be represented in an isolation_source qualifier.

6.15.	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;

6.16. 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>

6.17. Qualifier direction

Definition direction of DNA replication or transfer

Value format left, right, or both
where left indicates toward the 5’ end of the entry 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

6.18. 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.

6.19. 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.

6.20. 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. entries including environmental_sample must not include the strain qualifier

6.21. 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

Example <INSDQualifier_value>RNA editing</INSDQualifier_value>
<INSDQualifier_value>reasons given in citation</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.

6.22. Qualifier focus

Definition identifies the source feature of primary biological interest for records that have multiple source features originating from different organisms and that are not transgenic

Value format none

Comment the source feature carrying the focus qualifier identifies the main organism of the entry; only one source feature with a focus qualifier is allowed in an entry; the focus and transgenic qualifiers are mutually exclusive in an entry.

6.23. 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>

6.24. 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.

6.25. 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.

6.26. 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

---

6.27. Qualifier germline

Definition the sequence presented in the entry 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)

6.28. 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>

6.29. 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>

6.30. 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>

6.31. Qualifier identified_by

Definition name of the taxonomist who identified the specimen

Value format free text
Example <INSDQualifier_value>John Burns</INSDQualifier_value>

6.32. 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>

6.33. 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; the country qualifier should be used to describe the country and major geographical sub-region.

6.34. 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

6.35.	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>45.01 S 4.12 E</INSDQualifier_value>

6.36.	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

6.37.	Qualifier	map
	Definition	genomic map position of feature
	Value format	free text
	Example	<INSDQualifier_value>8q12-13</INSDQualifier_value>

6.38.	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.

6.39.	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>

6.40. 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 paragraph 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

---

6.41. 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 every source feature key; all mol_type values within one entry/record must be the same;; 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.

6.42. 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
RNase_P_RNA

Comment [A1]: Clarification is needed from INSDC whether “ribozyme” should be included as a value. INSDC vers. 9.0 of the feature table definition lists “ribozyme” as a value and indicates “TYPE is a term taken from the INSDC controlled vocabulary for ncRNA classes (http://www.insdc.org/rna_vocab.html); on 15-Oct-2008, the following terms were valid” . However, the value “ribozyme” is not listed on the controlled vocabulary list for ncRNA classes indicated on this webpage.

NCBI is investigating making the appropriate change.

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

6.43.	Qualifier	note
	Definition	any comment or additional information
	Value format	free text
	Example	<INSDQualifier_value>A comment.about the feature</INSDQualifier_value>

6.44.	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

6.45.	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

6.46.	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>

6.47. 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>

6.48. Qualifier

PCR_conditions

Definition

description of reaction conditions and components for PCR

Value format

free text

Example

<INSDQualifier_value>Initial denaturation:94degC,1.5min</INSDQualifier_value>

Comment

used with primer_bind feature key only

6.49. 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 Annex B.1, paragraph 1, 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.

6.50. 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>

6.51. 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>

6.52. Qualifier pop_variant

Definition name of a variation that characterizes a particular sub-population within a given species. The variation could be in the genotype or the phenotype

Value format free text

Example <INSDQualifier_value>pop1</INSDQualifier_value>
<INSDQualifier_value>Bear Paw</INSDQualifier_value>

6.53. 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)

6.54. 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

6.55. Qualifier pseudo

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

Comment [A2]: Note that INSD 10 includes the following comment:
not to be used for new submissions from 15-APR-2012; After 15-APR-2012 a new qualifier /pseudogene will become valid.

the feature key

Value format none

Comment only one of the qualifiers translation and pseudo are permitted to further annotate a CDS feature

6.56. 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)

6.57. 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

6.58. 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

6.59. Qualifier rpt_family

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

Value format free text

Example <INSDQualifier_value>Alu</INSDQualifier_value>

6.60. 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.

6.61. 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.

6.62. 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

---

6.63. 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.

6.64.	Qualifier	segment
	Definition	name of viral or phage segment sequenced
	Value format	free text
	Example	<INSDQualifier_value>6</INSDQualifier_value>

6.65.	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".

6.66.	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".

6.67.	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

6.68.	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).

6.69.	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

6.70.	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

6.71.	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>

6.72.	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

6.73.	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

6.74.	Qualifier	tissue_lib
	Definition	tissue library from which sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>tissue library 772</INSDQualifier_value>

6.75.	Qualifier	tissue_type
	Definition	tissue type from which the sequence was obtained
	Value format	free text
	Example	<INSDQualifier_value>liver</INSDQualifier_value>

6.76.	Qualifier	transgenic
	Definition	identifies the source feature of the organism which was the recipient of transgenic DNA
	Value format	none
	Comment	transgenic sequences must have at least two source feature keys; the source feature key having the transgenic qualifier must span the whole sequence; the source feature carrying the transgenic qualifier identifies the main organism of the entry, this determines: a) the name displayed in the organism lines, b) if no translation table is specified, the translation table; only one source feature with a transgenic qualifier is allowed in an entry; the focus and transgenic qualifiers are mutually exclusive in an entry

6.77.	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 Annex B.1, paragraph 3, 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’.

6.78. 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 range of specified tables are reported in transl_except qualifiers.

6.79. 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)

6.80. 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 exceptions in the transl_except qualifier

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

Example <INSDQualifier_value>MASTFPPWYRGCASTPSLKGLIMCTW</INSDQualifier_value>

Comment to be used with CDS feature only; see transl_table for definition and location of genetic code Tables; only one of the qualifiers translation and pseudo are permitted to further annotate a CDS feature.

6.81. 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>

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 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 NOTE

7.10. 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)”

7.11. 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 disukfide bonds, the NOTE qualifier indicates the nature of the cross-link, by identifying the other protein, for example, “Interchain (between A and B chains)”

7.12. Feature Key D_SEGMENT

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

Optional qualifiers NOTE

7.13. 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”

7.14. 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”

7.15. Feature Key HELIX

Definition: Secondary structure: Helices, for example, Alpha-helix; 3 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.

7.16. 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, and is not used when the initiator methionine is not cleaved off.

7.17. Feature Key INTRAMEM

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

Optional qualifiers: NOTE

7.18. Feature Key J_SEGMENT

Definition: Joining segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains

Optional qualifiers: NOTE

7.19. 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.”

7.20. Feature Key METAL

Definition Binding site for a metal ion. The description field indicates the nature of the metal

Mandatory qualifiers NOTE

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

7.21. 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

7.22. Feature Key MOTIF

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

Optional qualifiers NOTE

7.23. Feature Key MUTAGEN

Definition Site which has been experimentally altered by mutagenesis

Optional qualifiers NOTE

7.24. 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) present in the amino acid sequence.

7.25. 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.

7.26. 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”.
7.27.	Feature Key	PEPTIDE
	Definition	Extent of a released active peptide
	Optional qualifiers	NOTE
7.28.	Feature Key	PROPEP
	Definition	Extent of a propeptide
	Optional qualifiers	NOTE
7.29.	Feature Key	REGION
	Definition	Extent of a region of interest in the sequence
	Optional qualifiers	NOTE
7.30.	Feature Key	REPEAT
	Definition	Extent of an internal sequence repetition
	Optional qualifiers	NOTE
7.31.	Feature Key	SIGNAL
	Definition	Extent of a signal sequence (prepeptide)
	Optional qualifiers	NOTE
7.32.	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
	Optional* qualifiers	NOTE
	Comment	*The “NOTE” qualifier is mandatory when SITE is used to annotate an “other” amino acid as per this standard, and must contain the full name of the amino acid. Otherwise, the “NOTE” qualifier is optional.
7.33.	Feature Key	SOURCE
	Definition	Identifies the biological source of the specified span of the sequence; this key is mandatory; more than one source key per sequence is allowed; every entry/record will have, as minimum, either a single source key spanning the entire sequence or multiple source keys, which together, span the entire sequence
	Mandatory qualifiers	MOL_TYPE
	ORGANISM	
	Optional qualifiers	NOTE

7.34.	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.

7.35.	Feature Key	TOPO_DOM
	Definition	Topological domain
	Optional qualifiers	NOTE

7.36.	Feature Key	TRANSMEM
	Definition	Extent of a transmembrane region
	Optional qualifiers	NOTE

7.37.	Feature Key	TRANSIT
	Definition	Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.)
	Optional qualifiers	NOTE

7.38.	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.

7.39.	Feature Key	UNSURE
	Definition	Uncertainties in the amino acid sequence
	Optional qualifiers	NOTE
	Comment	Used to describe region(s) of a an amino acid sequence for which the authors are unsure about the sequence presentation.

7.40.	Feature Key	V_REGION
	Definition	Variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; the variable amino terminal portion; can be composed of V_segments, D_segments, N_regions, and J_segments
	Optional qualifiers	NOTE

7.41.	Feature Key	V_SEGMENT
	Definition	Variable segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; most of the variable region (V_region) and the last few amino acids of the leader peptide
	Optional qualifiers	NOTE

7.42.	Feature Key	VARIANT
	Definition	Authors report that sequence variants exist.
	Optional qualifiers	NOTE

7.43.	Feature Key	VAR_SEQ
	Definition	Description of sequence variants produced by Alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting
	Optional qualifiers	NOTE

7.44.	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”

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>polypeptide</INSDQualifier_value>
	Comment	mol_type qualifier is mandatory on every 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 every SOURCE feature key.

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 17-20 do not exist, therefore these numbers do not appear in Table 5.)

Table 5: Genetic Code Tables

1 - Standard Code

```
    AAs  =  FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  ---M---------------M---------------M----------------------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

2 - Vertebrate Mitochondrial Code

```
    AAs  =  FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG
  Starts =  --------------------------------MMMM---------------M------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

3 - Yeast Mitochondrial Code

```
    AAs  =   FFLLSSSSYY**CCWWTTTTPPPPHHQQRRRRIIMMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =   ----------------------------------MM----------------------------
  Base1  =   ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =   ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =   tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

4 - Mold, Protozoan, Coelenterate Mitochondrial Code & Mycoplasma/Spiroplasma Code

```
    AAs  =  FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  --MM---------------M------------MMMM---------------M------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

5 - Invertebrate Mitochondrial Code

```
    AAs  =   FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSSSVVVVAAAADDEEGGGG
  Starts =   ---M----------------------------MMMM---------------M------------
  Base1  =   ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =   ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =   tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

6 - Ciliate, Dasycladacean and Hexamita Nuclear Code

```
    AAs  =  FFLLSSSSYYQQCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =   -----------------------------------M----------------------------
  Base1  =   ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =   ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =   tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

9 - Echinoderm and Flatworm Mitochondrial Code

```
    AAs  =   FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGG
  Starts =   -----------------------------------M---------------M------------
  Base1  =   ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =   ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =   tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

10 - Euplotid Nuclear Code

```
    AAs  =  FFLLSSSSYY**CCCWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  -----------------------------------M----------------------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

11 - Bacterial and Plant Plastid Code

```
    AAs  =  FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  ---M---------------M------------MMMM---------------M------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

12 - Alternative Yeast Nuclear Code

```
    AAs  =  FFLLSSSSYY**CC*WLLLSPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  -------------------M---------------M----------------------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

13- Ascidian Mitochondrial Code

```
    AAs  =  FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSSGGVVVVAAAADDEEGGGG
  Starts =  ---M------------------------------MM---------------M------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

14 - Alternative Flatworm Mitochondrial Code

```
    AAs  =  FFLLSSSSYYY*CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNNKSSSSVVVVAAAADDEEGGGG
  Starts =  -----------------------------------M----------------------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

15 - Blepharisma Nuclear Code

```
    AAs  =  FFLLSSSSYY*QCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  -----------------------------------M----------------------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

16 - Chlorophycean Mitochondrial Code

```
    AAs  =  FFLLSSSSYY*LCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  -----------------------------------M----------------------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

21 - Trematode Mitochondrial Code

```
    AAs  =  FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNNKSSSSVVVVAAAADDEEGGGG
  Starts =  -----------------------------------M---------------M------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

22 - Scenedesmus obliquus mitochondrial

```
    AAs  =  FFLLSS*SYY*LCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  -----------------------------------M----------------------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

23 - Thraustochytrium Mitochondrial Code

```
    AAs  =  FF*LSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
  Starts =  --------------------------------M--M---------------M------------
  Base1  =  ttttttttttttttttccccccccccccccccaaaaaaaaaaaaaaaagggggggggggggggg
  Base2  =  ttttccccaaaaggggttttccccaaaaggggttttccccaaaaggggttttccccaaaagggg
  Base3  =  tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcagtcag
```

10 “Country” qualifier values

The qualifier “country” requires a controlled vocabulary to indicate the country of origin of a DNA sample. This listing was revised on 15 June 2010.

A

- Afghanistan
- Albania
- Algeria
- American Samoa
- Andorra
- Angola
- Anguilla
- Antarctica
- Antigua and Barbuda
- Arctic Ocean
- Argentina
- Armenia
- Aruba
- Ashmore and Cartier Islands
- Atlantic Ocean
- Australia
- Austria
- Azerbaijan

B

- Bahamas
- Bahrain
- Baltic Sea
- Baker Island
- Bangladesh
- Barbados
- Bassas da India
- Belarus
- Belgium
- Belize
- Benin
- Bermuda
- Bhutan
- Bolivia
- Borneo
- Bosnia and Herzegovina
- Botswana
- Bouvet Island
- Brazil
- British Virgin Islands
- Brunei
- Bulgaria
- Burkina Faso
- Burundi

C

- Cambodia
- Cameroon
- Canada
- Cape Verde
- Cayman Islands
- Central African Republic
- Chad
- Chile
- China
- Christmas Island
- Clipperton Island
- Cocos Islands
- Colombia
- Comoros
- Cook Islands
- Coral Sea Islands
- Costa Rica
- Cote d'Ivoire
- Croatia
- Cuba
- Cyprus
- Czech Republic

D

- Democratic Republic of the Congo
- Denmark
- Djibouti
- Dominica
- Dominican Republic

E

- East Timor
- Ecuador
- Egypt
- El Salvador
- Equatorial Guinea
- Eritrea
- Estonia
- Ethiopia
- Europa Island

F

- Falkland Islands (Islas Malvinas)
- Faroe Islands
- Fiji
- Finland
- France
- French Guiana
- French Polynesia
- French Southern and Antarctic Lands

G

- Gabon
- Gambia
- Gaza Strip
- Georgia
- Germany
- Ghana
- Gibraltar
- Glorioso Islands
- Greece

- Greenland
- Grenada
- Guadeloupe
- Guam
- Guatemala
- Guernsey
- Guinea
- Guinea-Bissau
- Guyana

H

- Haiti
- Heard Island and McDonald Islands
- Honduras
- Hong Kong
- Howland Island
- Hungary

I

- Iceland
- India
- Indian Ocean
- Indonesia
- Iran
- Iraq
- Ireland
- Isle of Man
- Israel
- Italy

J

- Jamaica
- Jan Mayen
- Japan
- Jarvis Island
- Jersey
- Johnston Atoll
- Jordan
- Juan de Nova Island

K

- Kazakhstan
- Kenya
- Kerguelen Archipelago
- Kingman Reef
- Kiribati
- Kosovo
- Kuwait
- Kyrgyzstan

L

- Laos
- Latvia
- Lebanon
- Lesotho
- Liberia
- Libya
- Liechtenstein
- Lithuania
- Luxembourg

M

- Macau
- Macedonia
- Madagascar
- Malawi
- Malaysia
- Maldives
- Mali
- Malta
- Marshall Islands
- Martinique
- Mauritania
- Mauritius
- Mayotte
- Mediterranean Sea
- Mexico
- Micronesia
- Midway Islands
- Moldova
- Monaco
- Mongolia
- Montenegro
- Montserrat
- Morocco
- Mozambique
- Myanmar

N

- Namibia
- Nauru
- Navassa Island
- Nepal
- Netherlands
- Netherlands Antilles
- New Caledonia
- New Zealand
- Nicaragua
- Niger
- Nigeria
- Niue
- Norfolk Island
- North Korea
- North Sea
- Northern Mariana Islands
- Norway

O

- Oman

P

- Pacific Ocean
- Pakistan
- Palau
- Palmyra Atoll
- Panama
- Papua New Guinea
- Paracel Islands
- Paraguay
- Peru
- Philippines
- Pitcairn Islands
- Poland
- Portugal
- Puerto Rico

Q

- Qatar

R

- Republic of the Congo
- Reunion
- Romania
- Ross Sea
- Russia
- Rwanda

S

- Saint Helena
- Saint Kitts and Nevis
- Saint Lucia
- Saint Pierre and Miquelon
- Saint Vincent and the Grenadines
- Samoa
- San Marino
- Sao Tome and Principe
- Saudi Arabia
- Senegal
- Serbia
- Seychelles
- Sierra Leone
- Singapore
- Slovakia
- Slovenia
- Solomon Islands
- Somalia
- South Africa
- South Georgia and the South Sandwich Islands
- South Korea
- Southern Ocean
- Spain
- Spratly Islands
- Sri Lanka
- Sudan
- Suriname
- Svalbard
- Swaziland
- Sweden
- Switzerland
- Syria

T

- Taiwan
- Tajikistan
- Tanzania
- Tasman Sea
- Thailand
- Togo
- Tokelau
- Tonga
- Trinidad and Tobago
- Tromelin Island
- Tunisia
- Turkey
- Turkmenistan
- Turks and Caicos Islands
- Tuvalu

U

- USA
- Uganda
- Ukraine
- United Arab Emirates
- United Kingdom
- Uruguay
- Uzbekistan

V

- Vanuatu
- Venezuela
- Viet Nam
- Virgin Islands

W

- Wake Island
- Wallis and Futuna
- West Bank
- Western Sahara

Y

- Yemen

Z

- Zambia
- Zimbabwe

Historical Country Names

- Belgian Congo
- British Guiana
- Burma
- Czechoslovakia
- Former Yugoslav Republic of Macedonia
- Korea
- Serbia and Montenegro
- Siam
- USSR
- Yugoslavia
- Zaire