# Relax NG schema for LRG

# Version $Revision: 1.9 $

# Last modified: 2014-10-01

start = LRG

LRG = element lrg {

attribute schema\_version {

xsd:decimal

},

FIXED\_ANNOTATION,

UPDATABLE\_ANNOTATION

}

# fixed annotation section will remain unchanged for lifespan of LRG

FIXED\_ANNOTATION = element fixed\_annotation {

# main LRG identifier e.g. LRG\_1

element id {

xsd:string {

pattern = 'LRG\_[0-9]+'

}

},

element hgnc\_id {

xsd:integer

}?,

element sequence\_source {

text #seq-id of the sequence that LRG is based on, normally RefSeqGene

}?,

element organism {

attribute taxon {

xsd:integer

},

text

},

SOURCE\*,

element mol\_type {

text

},

element creation\_date {

xsd:date

},

COMMENT?,

DNA\_SEQUENCE,

FX\_TRANSCRIPT\*

} # /fixed\_annotation

SOURCE = element source {

element name {

text

},

element url {

text

}\*,

# multiple contacts may be specified within a source

CONTACT\*

}

CONTACT = element contact {

element name {

text

}?,

element url {

text

}\*,

element address {

text

}?,

element email {

text

}?

}

DNA\_SEQUENCE = element sequence {

xsd:string {

pattern = '[ATGC]+'

}

}

FX\_TRANSCRIPT = element transcript {

attribute name {

xsd:string {

pattern = 't[0-9]+'

}

}, # e.g. "t1", "t2"

element creation\_date {

xsd:date

}?, # e.g. when a new transcript is added to a public LRG

COMMENT\*,

LRG\_COORDS, # location of transcript in LRG coordinates

element cdna {

DNA\_SEQUENCE

},

# The coding region is optional to accommodate e.g. non-protein-coding genes

CODING\_REGION\*,

# All transcripts have at least one exon

FX\_EXON,

# Additional (intron)-exon blocks

INTRONEXON\*

}

# LRG genomic coordinates element

LRG\_COORDS = element coordinates {

attribute coord\_system {

xsd:string {

pattern = 'LRG\_[0-9]+'

}

},

COORD\_ATTRIBS

}

# Common coordinate attributes

COORD\_ATTRIBS =

attribute start { xsd:integer },

attribute end { xsd:integer },

attribute start\_ext { xsd:integer }?,

attribute end\_ext { xsd:integer }?,

attribute strand { '-1' | '1' }?,

attribute mapped\_from { text }?

CODING\_REGION = element coding\_region {

# start phase of coding sequence when coding sequence is incomplete

attribute codon\_start { xsd:integer }?,

# location of coding region in LRG and the parent transcript coordinates

LRG\_COORDS,

# amino-acid replacement, e.g. canonical stop codon read through, in p# coordinates

element translation\_exception {

attribute codon { xsd:integer },

PEPTIDE\_SEQUENCE

}\*,

#ribosomal slippage

element translation\_frameshift {

attribute cdna\_position {xsd:integer}, #location in the parent t# coordinates

attribute shift { xsd:integer } #typically -1 or 1

}\*,

TRANSLATION

}

TRANSLATION = element translation {

attribute name {

xsd:string {

pattern = 'p[0-9]+[A-Za-z]\*'

}

}, # e.g. "p1", "p2"

PEPTIDE\_SEQUENCE

}

PEPTIDE\_SEQUENCE = element sequence {

xsd:string {

pattern = '[ACDEFGHIKLMNOPQRSTUVWY]+'

}

}

# exon element - coordinates given in three systems

FX\_EXON = element exon {

attribute label { xsd:string }, #Exon number based on LRG's t# transcripts

LRG\_COORDS,

LRG\_CDNA\_COORDS,

LRG\_PEPTIDE\_COORDS\*

}

# LRG cDNA coordinates element

LRG\_CDNA\_COORDS = element coordinates {

attribute coord\_system {

xsd:string {

pattern = 'LRG\_[0-9]+t[0-9]+'

}

},

COORD\_ATTRIBS

}

# LRG peptide coordinates element

LRG\_PEPTIDE\_COORDS = element coordinates {

attribute coord\_system {

xsd:string {

pattern = 'LRG\_[0-9]+p[0-9]+[A-Za-z]\*'

}

},

COORD\_ATTRIBS

}

# Compound element containing one (optional) intron and one (required) exon

INTRONEXON = (

FX\_INTRON?,

FX\_EXON

)

FX\_INTRON = element intron {

# intron phase

# 0 : intron falls between codons

# 1 : intron falls between 1st and 2nd base of codon

# 2 : intron falls between 2nd and 3rd base of codon

attribute phase { '0' | '1' | '2' }

}

# updatable annotation can be updated with e.g. new annotations, mappings, labels etc

UPDATABLE\_ANNOTATION = element updatable\_annotation {

# an annotation set comes from one source - allows multiple sets of annotations from e.g. Ensembl/NCBI

ANNOTATION\_SET+

} # /updatable\_annotation

ANNOTATION\_SET = element annotation\_set {

attribute type { text }?,

SOURCE+,

COMMENT?,

element modification\_date { xsd:date },

# One annotation per fixed transcript

# allows exons to be named; in fixed layer exons are defined only by their coordinates

# allows for legacy systems of amino acid numbering

# allows for comments on fixed layer transcripts

FIXED\_TRANSCRIPT\_ANNOTATION\*,

# multiple mappings allowed e.g. to different assemblies

MAPPING\*,

element lrg\_locus {

FEATURE\_SOURCE,

text

}?,

FEATURES?,

# Optional extra note(s)/information from the community

NOTE\*

}

FIXED\_TRANSCRIPT\_ANNOTATION = element fixed\_transcript\_annotation {

# fixed\_id corresponding to the transcript name used in the fixed section

attribute name {

xsd:string {

pattern = 't[0-9]+'

}

},

OTHER\_EXON\_NAMING\*,

ALT\_AA\_NUMBERING\*

}

COMMENT = element comment { text }

OTHER\_EXON\_NAMING = element other\_exon\_naming {

attribute description { text },

element url {

text

}?,

COMMENT?,

element exon {

LRG\_COORDS,

element label { text }

}+

}

ALT\_AA\_NUMBERING = element alternate\_amino\_acid\_numbering {

attribute description { text },

element url {

text

}?,

COMMENT?,

element align {

attribute lrg\_start { xsd:integer },

attribute lrg\_end {xsd:integer },

attribute start { xsd:integer },

attribute end { xsd:integer }

}+

}

MAPPING = element mapping {

# Coord\_system can be e.g. GRCh37, NM\_000088.3

attribute coord\_system { text },

# Name is typically the chromosome name

attribute other\_name { text }?,

# Id can be e.g. the NCBI accession for a chromosome

attribute other\_id { text },

attribute other\_id\_syn { text }?,

attribute other\_start { xsd:integer },

attribute other\_end { xsd:integer },

# Attribute giving the information about the type of sequence aligned

attribute type { 'main\_assembly' | 'other\_assembly' | 'patch' | 'haplotype' | 'transcript' }?,

# a mapping span corresponds to a block of the LRG sequence that aligns with an assembly

MAPPING\_SPAN+

}

MAPPING\_SPAN = element mapping\_span {

attribute lrg\_start { xsd:integer },

attribute lrg\_end { xsd:integer },

attribute strand { '-1' | '1' },

attribute other\_start { xsd:integer },

attribute other\_end { xsd:integer },

# diff elements allow small-scale differences between the LRG sequence and assembly to be detailed

DIFF\*

}

DIFF = element diff {

attribute type {'mismatch' | 'lrg\_ins' | 'other\_ins' },

attribute lrg\_start { xsd:integer }?,

attribute lrg\_end { xsd:integer }?,

attribute other\_start { xsd:integer }?,

attribute other\_end { xsd:integer }?,

attribute lrg\_sequence {text}?, #for mismatches or lrg\_ins

attribute other\_sequence {text}? #for mismatches or genome\_ins

}

# Optional extra note/information from the community

NOTE = element note {

attribute name { text }?,

text

}

FEATURE\_SOURCE = attribute source {

text

}

FEATURES = element features {

UP\_GENE\*

}

UP\_GENE = element gene {

FEATURE\_SOURCE,

FEATURE\_ACCESSION,

SYMBOL,

# LRG coordinates should always be provided

LRG\_COORDS,

# In addition, coordinates can be given relative to other coordinate systems as well

COORDS\*,

# indicates if the gene only partially overlaps the LRG

PARTIAL\*,

LONG\_NAME\*,

COMMENT\*,

DB\_XREF\*,

# genes may contain zero or more transcripts

UP\_TRANSCRIPT\*

}

FEATURE\_ACCESSION = attribute accession {

text

}

# General coordinate element

COORDS = element coordinates {

attribute coord\_system { text },

attribute name { text }?,

COORD\_ATTRIBS

}

PARTIAL = element partial {

'5-prime' | '3-prime'

}

SYMBOL = element symbol {

attribute name { text },

FEATURE\_SOURCE?,

SYNONYM\*

}

LONG\_NAME = element long\_name {

text

}

DB\_XREF = element db\_xref {

attribute source { 'GeneID' | 'HGNC' | 'MIM' | 'GI' | 'RefSeq' | 'Ensembl' | 'CCDS' | 'UniProtKB' | 'RFAM' | 'miRBase' | 'pseudogene.org' },

FEATURE\_ACCESSION,

SYNONYM\*

}

SYNONYM = element synonym {

text

}

UP\_TRANSCRIPT = element transcript {

FEATURE\_SOURCE,

FEATURE\_ACCESSION,

attribute fixed\_id {

xsd:string {

pattern = 't[0-9]+'

}

} ?,

LRG\_COORDS,

COORDS\*,

PARTIAL\*,

LONG\_NAME\*,

COMMENT\*,

DB\_XREF\*,

UP\_EXON\*,

UP\_PROTEIN\_PRODUCT\*

}

UP\_EXON = element exon {

FEATURE\_SOURCE,

FEATURE\_ACCESSION,

LRG\_COORDS,

COORDS\*,

PARTIAL\*,

COMMENT\*

}

UP\_PROTEIN\_PRODUCT = element protein\_product {

FEATURE\_SOURCE,

FEATURE\_ACCESSION,

attribute fixed\_id {

xsd:string {

pattern = 'p[0-9]+[A-Za-z]\*'

}

} ?,

attribute codon\_start { xsd:integer }?,

LRG\_COORDS,

COORDS\*,

PARTIAL\*,

LONG\_NAME\*,

COMMENT\*,

DB\_XREF\*

}