Next-generation **G**enomic **F**eature **F**ormat Specification

version v202210

The main version of this document can be found at https://github.com/dengcao3/ngfftools.

Special Notes:

(1) Our format is not totally fresh designed, on the contrary, considering the success of gff3 and gtf, we reserved their very user-friendly designs and add new feature. So do not expect a totally difference and it's not strange that you can see some similar structure or words used by the GFF3/GTF specification. Here, we thank the Lincoln Stein, other contributors and users of gff3/gtf format, you opened the door!

(2) Our goal of this document is merely to make user understand our format, and our design optimizes the format to storage annotation information and we do not invent new terms, so we may direct copy the descriptions of these term into our document without reference to the original (we may complete them in the future). As we used a large number of biological terms to explain the format, we have not tried to remove duplication of the sentences. We do not claim any rights to this document. Here, we acknowledge the contributions of scientists and document creators, you make things more easily!

(3) This specification is still in its babyhood, we will will continue to update. Any suggestions and cooperations are welcomed.

Contents

1. Introduction

Next-generation **G**enomic **F**eature **F**ormat (NGFF) is nine-column, tab-delimited, plain text file format, including two sections: **header section** and **data section**. Undefined fields are replaced with the "." character. The ‘feature’ here includes:

（1） sequence：DNA, RNA, Protein。

（2） one continuous fragment of one sequence, such as a exon;

（3） joint sequence from incontinuous fragments of one sequence, such as a processed transcript that have more than one exon in eukaryote;

（4） joint sequence from multiple sequences, such as the sequence with chromosome translocation, the trans-spliced transcript, the trans-spliced functional polypeptide.

（5） point position, a sequence with one length, which is used to record sequence modification or change (such as RNA editing site, RNA methylation site, protein post-translation modification site)

2. Header section

```
##NGFF version 1.0
##species NCBI_Taxonomy_URI
##genome-build dbName buildName
##reference seqId.1 start end
##reference seqId.2 start end
##…
##reference seqId.n start end
```

Example：

```
#comments
##NGFF version 1.0
##species Arabidopsis thaliana
##genome-build TAIR 10.1
##reference Chr1 1 30427671
##reference Chr2 1 19698289
##reference Chr3 1 23459830
##reference Chr4 1 18585056
##reference Chr5 1 26975502
##reference ChrM 1 367808
##reference ChrC 1 154478
```

3. Data section: definitions of the nine fields

3.1. seqId

The ID of the reference sequence used to establish the coordinate system for the current feature. IDs may contain any characters, but must escape any characters not in the set [a-zA-Z0-9.:^*$@!+_?-|]. In particular, IDs may not contain unescaped whitespace and must not begin with an unescaped ">".

3.2. start and end

The start and end coordinates of the feature are given in 1-based non-zero integer coordinates, relative to the reference sequence.

If the integer is non-zero positive, the reference is given in column one. The ‘start’ is always less than or equal to the ‘end’. For features that cross the origin of a circular sequence (e.g. most bacterial genomes, plasmids, and some viral genomes), the requirement for ‘start’ to be less than or equal to ‘end’ is satisfied by making ‘end’ = ‘the real position of the end’ + ‘the length of the reference sequence’. For example, the mitogenome has length of 16010 bp, the ‘start’ and ‘end’ of the region 16000 to 200 should be recorded as 16000 and 16210 respectively.

If the integer is non-zero negative, the reference is given in the column seven and must have one parentId.

3.3. strand

The strand of the feature. The ‘+’ is for positive strand (relative to the reference), the ‘-‘ is for minus strand, the ‘.’ is for features that are not stranded, and the ‘?’ is used for features whose strand is relevant but unknown.

3.4. level

The levels of the features belong to. The hierarchical structure of the levels:

chromosome -> locus –> primary -> processed -> product

3.5. featureType

The type of the features. Each ‘***level***’ has a shared but optional featureType, the ‘***regulator***’ and ‘***Non regulator***’. ‘***Non regulator***’ must contribute to the formation of sequence. ‘***regulator***’ has nothing to do with the formation of sequence (about the type of regulator please see the **section 3.8.2**).

Level	featureType
chromosome	reference
locus	regulator, locus
primary	regulator, region
processed	regulator, cap, exon, intron, polyA_sequence
product	regulator, CDS, mature

3.6. id

The ID(s) of the sequence(s) this feature produces. If have multiple ID values, the number of attributes must correspond with them, unless all the attributes refer to the same value.

For example, if this feature is an ‘***exon***’, the ‘***id***’ here is the IDs of the processed transcripts.

3.7. parentId

The ID(s) of the parent level sequence that generates this feature (When start and end is absolute position, parentId can be multiple values (e.g. fusion gene)).

For example, if this feature is an ‘***exon***’, the ‘***id***’ here is the IDs of the processed transcripts, while the ‘***parentId***’ here is the ID(s) of the primary transcript(s) which produce this exon.

The ‘***genome***’ is the reserved ***parentId*** for all the chromosome level sequences.

3.8. attributes

- if the feature has no any mandatory attributes, this column could be replaced with the "." character.

- format to record atrributes (JSON style): attribute1=single_word_content; attribute2="long_contents"; attribute3=multi1,multi2; attribute4="long_content1"," long_content2"; attribute5=[nested1,nested2,nested3],[nested4,nexted5,nested6]; attribute6={"name1":"value1","name2":"value2","name3":{"inner1":"innervalue1"}};
- order-irrelevance among attributes
- names of attributes are case sensitive.
- attributes must end with a semicolon which must then be separated from the start of any subsequent attribute by zero or more spaces character (NOT the tab character).
- textual attribute values should be surrounded by double quotes if the text has space. Literal use of tab, newline, carriage return, the percent (%) sign, and control characters must be encoded using RFC 3986 Percent-Encoding; no other characters may be encoded. Backslash and other ad-hoc escaping conventions that have been added to the NGFF format are not allowed. The file contents may include any character in the set supported by the operating environment, although for portability with other systems, use of Latin-1 or Unicode are recommended.
    - tab (%09)
    - newline (%0A)
    - carriage return (%0D)
    - % percent (%25)
    - control characters (%00 through %1F, %7F)

  In addition, the following characters have reserved meanings in this column and must be escaped when used in the values:
    - ; semicolon (%3B)
    - = equals (%3D)
    - & ampersand (%26)
    - , comma (%2C)

3.8.1. featureType

featureType	mandatory attributes
reference	seqSource, circular
locus	none
regulator	regulator_biotype
region	primary_biotype, transcribed
cap, exon, intron, polyA_sequence	processed_biotype, product_biotype
CDS	product_biotype, phase
mature	product_biotype

3.8.2. regulator_biotype

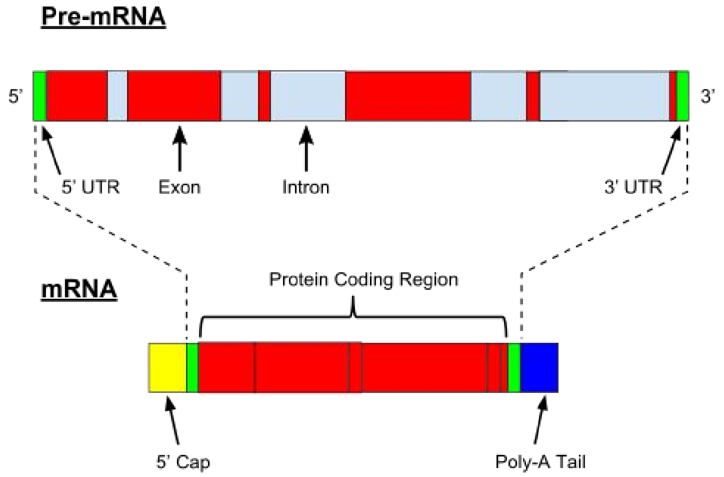
Each '***level***' has a shared but optional featureType, the '***regulator***', which is the regulated region on the sequence. The '***regulator***' featureType must have a '***regulator_biotype***' in attributes (9th column).

level	Biotype value	regulator_biotype value
locus	-	enhancer
locus	-	promoter
locus	-	insulator
locus	-	unknown
primary	primary_transcript	TSS
primary	primary_transcript	TSS_region
primary	primary_transcript	transcription_pause_site
primary	primary_transcript	transcription_termination_signal
primary	primary_transcript	…
primary	transposable_element	?
primary	tandem_repeat	?
primary	CNS	?
primary	complex	?

processed	processed_transcript	edited_site
processed	processed_transcript	modified_site
processed	processed_transcript	polyA_signal_sequence
processed	processed_transcript	polyA_site
processed	processed_transcript	repression_signal_sequence
processed	processed_transcript	degradation_signal_sequence
processed	processed_transcript	attenuator
processed	processed_transcript	…
processed	processed_TE	?
processed	processed_TR	?
processed	processed_CNS	?
product	ORF/sORF/teORF	UTR5
product	ORF/sORF/teORF	UTR3
product	ORF/sORF/teORF	start_codon
product	ORF/sORF/teORF	stop_codon
product	…	…

terminology interpretation:

primary_transcript: A primary transcript is the single-stranded ribonucleic acid (RNA) product synthesized by transcription of DNA, and processed to yield various mature RNA products such as mRNAs,tRNAs,and rRNAs(in our format,the biotype value for these transcripts is **processed_transcript**). The primary transcripts designated to be mRNAs are modified in preparation for translation. For example, a precursor mRNA (pre-mRNA) is a type of primary transcript that becomes a messenger RNA (mRNA) after processing.



transposable_element: A transposable element (TE, transposon, or jumping gene) is a DNA sequence that can change its position within a genome, sometimes creating or reversing mutations and altering the cell's genetic identity and genome size. Transposition often results in duplication of the same genetic material.

tandem_repeat: Tandem repeats occur in DNA when a pattern of one or more nucleotides is repeated and the repetitions are directly adjacent to each other. Several protein domains also form tandem repeats within their amino acid primary structure, such as armadillo repeats.

CNS: A conserved non-coding sequence (CNS) is a DNA sequence of noncoding DNA that is evolutionarily conserved. These sequences are of interest for their potential to regulate gene production. CNSs in plants and animals are highly associated with transcription factor binding sites and other cis-acting regulatory elements.

enhancer: In genetics, an enhancer is a short (50–1500 bp) region of DNA that can be bound by proteins (activators) to increase the likelihood that transcription of a particular gene will occur. Enhancers are cis-acting. They can be located up to 1 Mbp (1,000,000 bp) away from the gene, upstream or downstream from the start site.

Promoter: A promoter is a sequence of DNA to which proteins bind to initiate transcription of a single RNA transcript from the DNA downstream of the promoter. Promoters are located near the transcription start sites of genes, upstream on the DNA (towards the 5' region of the sense strand). Promoters can be about 100–1000 base pairs long, the sequence of which is highly dependent on the gene and product of transcription, type or class of RNA polymerase recruited to the site, and species of organism.

Insulator: An insulator is a type of cis-regulatory element known as a long-range regulatory element. Found in multicellular eukaryotes and working over distances from the promoter element of the target gene, an insulator is typically 300 bp to 2000 bp in length. Insulators contain clustered binding sites for sequence specific DNA-binding proteins and mediate intra- and inter-chromosomal interactions

TSS: transcription start site. A transcription start site is the location where the first DNA nucleotide is transcribed into RNA.

transcription_pause_site: RNA polymerase (RNAP) reads the DNA segment and copies the information to form a strand of RNA. However, RNAP does not usually read a whole gene in one go: there are several ‘pause sites’ in the sequence where it stops and waits for instruction. These ‘pause sites’ are transcription pause sites.

transcription_termination_signal: a transcription termination signal is a sequence that signals the end of transcription.

edited_site: It is RNA editing site. RNA editing (also RNA modification) is a molecular process through which some cells can make discrete changes to specific nucleotide sequences within an RNA molecule after it has been generated by RNA polymerase. It occurs in all living organisms and is one of the most evolutionarily conserved properties of RNAs. RNA editing may include the insertion, deletion, and base substitution of nucleotides within the RNA molecule.

start_codon: The start codon is the first codon of a messenger RNA (mRNA) transcript translated by a ribosome. The start codon always codes for methionine in eukaryotes and Archaea and a N-formylmethionine (fMet) in bacteria, mitochondria and plastids. The most common start codon is AUG (i.e., ATG in the corresponding DNA sequence).

stop_codon: In molecular biology (specifically protein biosynthesis), a stop codon (or termination codon) is a codon (nucleotide triplet within messenger RNA) that signals the termination of the translation process of the current protein.

polyA_signal_sequence, polyA_site: Polyadenylation is the addition of a poly(A) tail to an RNA transcript, typically a messenger RNA (mRNA). The poly(A) tail consists of multiple adenosine monophosphates.

UTR5: The 5’ untranslated region (also known as 5’ UTR, leader sequence, transcript leader, or leader RNA) is the region of a messenger RNA (mRNA) that is directly upstream from the initiation codon. This region is important for the regulation of translation of a transcript by differing mechanisms in viruses, prokaryotes and eukaryotes. While called untranslated, the 5’ UTR or a portion of it is sometimes translated into a protein product. This product can then regulate the translation of the main coding sequence of the mRNA. In many organisms, however, the 5’ UTR is completely untranslated, instead forming complex secondary structure to regulate translation.

UTR3: In molecular genetics, the three prime untranslated region (3’ -UTR) is the section of messenger RNA (mRNA) that immediately follows the translation termination codon. The 3’ -UTR often contains regulatory regions that post-transcriptionally influence gene expression.

3.8.3. primary_biotype, processed_biotype, product_biotype

primary_biotype, processed_biotype, product_biotype is an attribute of ***primary***, ***processed***, ***product*** level, respectively.

Value: string
Required: NO
Multivalued: YES

The alternative attributes value of ***primary_biotype, processed_biotype, product_biotype***:

primary_biotype	processed_biotype	product_biotype
primary_transcript	processed_transcript	ORF
		sORF
		teORF
		miRNA
		circleRNA
		lncRNA
		snRNA
		rRNA
		tRNA
		snoRNA
		tmRNA
		unknown
transposable_elements	processed_TE	(share others if exist)
		…
tandem_repeat	processed_TR	(share others if exist)
		…
CNS	processed_CNS	CNS
complex	processed_TE	(share others if exist)
	processed_TR	(share others if exist)
	processed_CNS	CNS

terminology interpretation:

ORF: In molecular biology, open reading frames (ORFs) are defined as spans of DNA sequence between the start and stop codons. Usually, this is considered within a studied region of a prokaryotic DNA sequence, where only one of the six possible reading frames will be 'open' (the 'reading', however, refers to the RNA produced by transcription of the DNA and its subsequent interaction with the ribosome in translation). Such an ORF may contain a start codon (usually AUG in terms of RNA) and by definition cannot extend beyond a stop codon (usually UAA, UAG UGA in RNA).

sORF: The most common definition of an sORF is simply an ORF of less than 50 amino acids (aa). These sORFs can be located within coding transcripts (5′ UTR, CDS or 3′ UTR) or even within non-coding RNAs such as long noncoding RNAs (lncRNAs), circular RNAs, and mitochondrial RNAs.

teORF: ORF from transposable elements.

miRNA: A microRNA (abbreviated miRNA) is a small single-stranded non-coding RNA molecule (containing about 22 nucleotides) found in plants, animals and some viruses, that functions in RNA silencing and post-transcriptional regulation of gene expression.

circleRNA: Circular RNA (or circRNA) is a type of single-stranded RNA which, unlike linear RNA, forms a covalently closed continuous loop. In circular RNA, the 3' and 5' ends normally present in an RNA molecule have been joined together. This feature confers numerous properties to circular RNA, many of which have only recently been identified.

lncRNA: Long non-coding RNAs (long ncRNAs, lncRNA) are a type of RNA, generally defined as transcripts more than 200 nucleotides that are not translated into protein.

snRNA: Small nuclear RNA (snRNA) is a class of small RNA molecules that are found within the splicing speckles and Cajal bodies of the cell nucleus in eukaryotic cells. The length of an average snRNA is approximately 150 nucleotides. They are transcribed by either RNA polymerase II or RNA polymerase III. Their primary function is in the processing of pre-messenger RNA (hnRNA) in the nucleus. They have also been shown to aid in the regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA), and maintaining the telomeres.

rRNA: Ribosomal ribonucleic acid (rRNA) is a type of non-coding RNA which is the primary component of ribosomes, essential to all cells. rRNA is a ribozyme which carries out protein synthesis in ribosomes.

tRNA: Transfer RNA (abbreviated tRNA and formerly referred to as sRNA, for soluble RNA) is an adaptor molecule composed of RNA, typically 76 to 90 nucleotides in length (in eukaryotes), that serves as the physical link between the mRNA and the amino acid sequence of proteins.

snoRNA: In molecular biology, Small nucleolar RNAs (snoRNAs) are a class of small RNA molecules that primarily guide chemical modifications of other RNAs, mainly ribosomal RNAs, transfer RNAs and small nuclear RNAs.

tmRNA: Transfer-messenger RNA (abbreviated tmRNA, also known as 10Sa RNA and by its genetic name SsrA) is a bacterial RNA molecule with dual tRNA-like and messenger RNA-like properties.

3.8.4. product_cluster/gene/gene_name

Value: string
Required: YES
Multivalued: NO
See detail please refer to 4.1(2).

3.8.5. seqSource

Values：string (major, plasmid, B, mitochondrion, kinetoplast, mitosome, plastid, chloroplast, chromoplast, apicoplast)
Default: unknown
Required: NO
Multivalued: NO

- **major**:
- **plasmid**: https://en.wikipedia.org/wiki/Plasmid
- **B**, B chromosome. These chromosomes are not essential for the life of a species, and are lacking in some (usually most) of the individuals. Thus, a population would consist of individuals with 0, 1, 2, 3 (etc.) supernumeraries. https://en.wikipedia.org/wiki/B_chromosome
- **mitochondrion**: https://en.wikipedia.org/wiki/Mitochondrion
- **kinetoplast**: https://en.wikipedia.org/wiki/Kinetoplast
- **mitosome**: https://en.wikipedia.org/wiki/Mitosome
- hydrogenosome: https://en.wikipedia.org/wiki/Hydrogenosome
- **plastid**: https://en.wikipedia.org/wiki/Plastid. In plants, plastids may differentiate into several forms, depending upon which function they play in the cell. Undifferentiated plastids (proplastids) may develop into any of the following variants:
    1. **Chloroplast**: green plastids for photosynthesis;
    2. **Chromoplast**: coloured plastids for pigment synthesis and storage
    3. Gerontoplast: control the dismantling of the photosynthetic apparatus during plant senescence
    4. Leucoplast: colourless plastids and sometimes differentiate into more specialized plastids:
        a) Amyloplast: for starch storage and detecting gravity (for geotropism)
        b) Elaioplast: for storing fat
        c) Proteinoplast: for storing and modifying protein
        d) Tannosome: for synthesizing and producing tannins and polyphenols

The DNA in chloroplasts and chromoplasts is identical. One subtle difference in DNA was found after a liquid chromatography analysis of tomato chromoplasts was conducted, revealing increased cytosine methylation.

- **apicoplast**: https://en.wikipedia.org/wiki/Apicoplast

3.8.6. circular

Value: FALSE, TRUE, ND (not determined)
Default: ND
Required: NO
Multivalued: NO
The sequence is circular or not or unknown currently.

3.8.7. phase

Value: 0,1,2
Required: YES (for CDS)
Multivalued: NO

The phase is REQUIRED for all CDS features.

For features of type "CDS", the phase indicates where the feature begins with reference to the reading frame. The phase is one of the integers 0, 1, or 2, indicating the number of bases that should be removed from the beginning of this feature to reach the first base of the next codon. In other words, a phase of "0" indicates that the next codon begins at the first base of the region described by the current line, a phase of "1" indicates that the next codon begins at the second base of this region, and a phase of "2" indicates that the codon begins at the third base of this region.

For forward strand features, phase is counted from the start field. For reverse strand features, phase is counted from the end field.

Phase of current CDS segment could be calculated as (3 - ((length - frame) mod 3)) mod 3.

- (length-frame) is the length of the previous feature starting at the first whole codon (and thus the frame subtracted out).
- (length-frame) mod 3 is the number of bases on the 3' end beyond the last whole codon of the previous feature.
- 3-((length-frame) mod 3) is the number of bases left in the codon after removing those that are represented at the 3' end of the feature.
- (3-((length-frame) mod 3)) mod 3 changes a 3 to a 0, since three bases makes a whole codon, and 1 and 2 are left unchanged.

The CDS segment that represent the new reading frame in programmed frameshift will always has a phase of 0 since the ribosome is moving and thus redefining the codon.

3.8.8. order

Value: int, “ ”
Default: “ ”
Required: NO
Multivalued: NO

The joint sequence order number of trans-spliced transcripts (for example: fusion gene).

3.8.9. sequence

Value: string
Default: “ ”
Required: NO
Multivalued: NO

3.8.10. edited_site

Value: string
Default: “ ”
Required: NO
Multivalued: YES

The format for this value is “[seqId]:[site_in_genome][position_in_genome][edited_site_in_genome]”, such as ‘Chr1:A6036C’, which means the nucleotide at Chr1, 6036bp in genome, A, is edited to C based on genome system. If the strand of the transcript is minus, the processed transcript of this position will convert to G, as C is based on genome system.
Multiple edited sites is separated by ‘,’, such as ‘Chr1:A6036C,Chr1:A6049C’.

3.8.11. frameshift

Value: string
Default: “ ”

Required: NO
Multivalued: NO

The format for the value is “X:Y”. The ‘X’ is the position, and the ‘Y’ is the shift value. The possible shift values include:

1. -2: The ribosome retreats (to 5’) two bases when move to position X;
2. -1: The ribosome retreats (to 5’) one base when move to position X;
3. +1: The ribosome moves forward (to 3’) one base when move to position X;
4. +2: The ribosome moves forward (to 3’) two bases when move to position X;

3.8.12. featureUid

Value: string
Default: seqId_start_end_strand
Required: NO
Multivalued: NO

The unique ID of the feature. Here, unique is defined as that the element has same ***seqId***, ***start***, ***end*** and ***strand***, namely the first four columns are the same. If empty, ngfftools will assign a string ***seqId_start_end_strand*** to it.

3.8.13. name

Value: string
Default: “ ”
Required: NO
Multivalued: NO

3.8.14. alias

Value: string
Default: [“ ”]
Required: NO
Multivalued: YES

3.8.15. annotationSource

Value: string
Default: “ ”
Required: NO
Multivalued: YES

The ***annotationSource*** is a free text qualifier intended to describe the algorithm or operating procedure that generated this feature. Typically this is the name of a piece of software, such as "Genescan" or a database name, such as "Genbank."

3.8.16. score

Value: floating point number
Default: NA
Required: NO
Multivalued: NO

The score field indicates a degree of confidence in the feature's existence and coordinates. It may be a floating point number or integer.

3.8.17. note

Value: string
Default: “ ”
Required: NO
Multivalued: YES

Any free text.

3.8.18. seqType

Values：DNA, RNA, Protein, ND
Default: ND
Required: NO
Multivalued: NO

3.8.19. locus_type

Value: gene_free, gene_containing, ND
Default: ND
Required: NO
Multivalued: NO
The final products of the locus contain or do not contain genes.

3.8.20. gene_number

Value: integer
Default: NA
Required: NO
Multivalued: NO
The number of genes in this locus.

3.8.21. pseudo

Value: FALSE, TRUE, ND
Default: ND
Required: NO
Multivalued: NO

Indicating after variation whether the sequence become a dysfunctional sequence containing pseudogenization through mutation or transcript pseudogenization through transcript splicing.

3.8.22. family

Value: string
Required: NO
Multivalued: YES

The family of TE or proteins. For example, the family of At1g18580 maybe ‘GT/GT8’, and the family value of a Copia TE could be ‘LTR/Copia’.

Some other type of TE: microsatellite, minisatellite, satellite:

A **microsatellite** is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 1–6 or more base pairs) are repeated, typically 5–50 times. Microsatellites occur at thousands of locations within an organism's genome. They have a higher mutation rate than other areas of DNA leading to high genetic diversity. Microsatellites are often referred to as short tandem repeats (STRs) by forensic geneticists and in genetic genealogy, or as simple sequence repeats (SSRs) by plant geneticists. https://en.wikipedia.org/wiki/Microsatellite

A **minisatellite** is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 10–60 base pairs) are typically repeated 5-50 times. https://en.wikipedia.org/wiki/Minisatellite

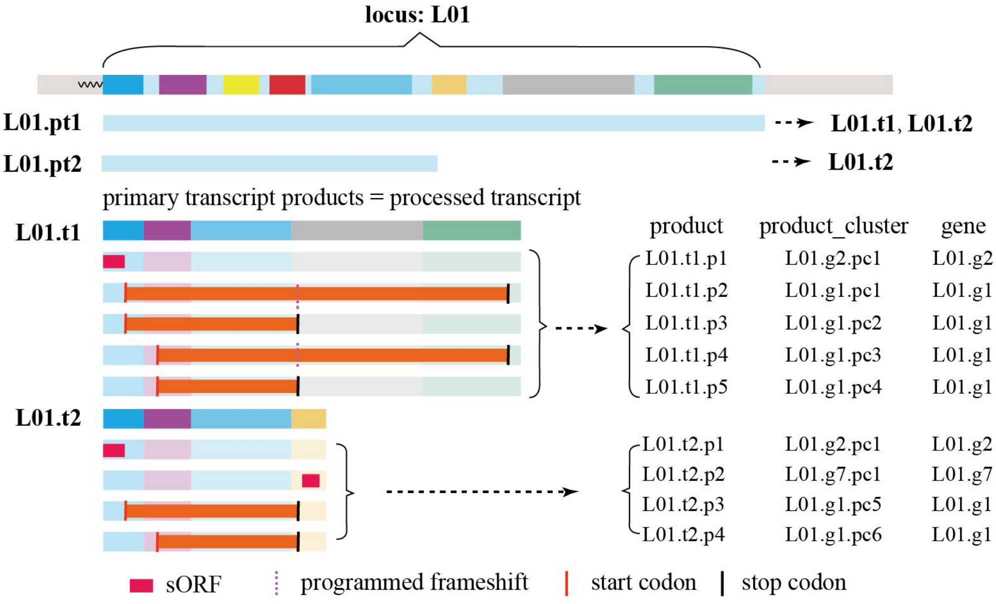
The **satellite** DNA consists of very large arrays of tandemly repeating, non-coding DNA. Satellite DNA is the main component of functional centromeres, and form the main structural constituent of heterochromatin. https://en.wikipedia.org/wiki/Satellite_DNA

3.8.23. expression

Value:
Required: YES (for CDS)
Multivalued: NO

4. Principles and Examples

4.1. organization and inheritance



product	product_cluster	gene
L01.t1.p1	L01.g2.pc1	L01.g2
L01.t1.p2	L01.g1.pc1	L01.g1
L01.t1.p3	L01.g1.pc2	L01.g1
L01.t1.p4	L01.g1.pc3	L01.g1
L01.t1.p5	L01.g1.pc4	L01.g1
L01.t2.p1	L01.g2.pc1	L01.g2
L01.t2.p2	L01.g7.pc1	L01.g7
L01.t2.p3	L01.g1.pc5	L01.g1
L01.t2.p4	L01.g1.pc6	L01.g1

Figure1. Example for description the organization of NGFF format

The ngff file corresponding to Figure 1 is: *Example for description the organization of NGFF format.xls*.

As produced according to special biological process, different kinds of sequences have the structural originated correlation between each other. On the other hand, sequences have different biological function so that effective hierarchical classification can help us understand the functions of sequence. That's why NGFF has two organized and inheritance relationship.

4.1.1. According to the biological process

It is accomplished through the relation among ***level***, ***id*** and ***parentId***. The source sequence of ***id*** is the sequence of ***parentId***. A certain locus can be successfully transcribed to primary transcript(s) on condition that various regulator elements orderly work together. Primary transcript structurally may contain different substructure, such as ***cap***, ***exon***, ***intron***, ***polyA_sequence***, and different regulator elements including ***TSS***, ***transcription_pause_site***, ***transcription_termination_signal*** and so on. After post-translational modification, primary transcript becomes processed transcript. It may contain different substructure, such as ***CDS*** and ***miRNA***, and the regulator elements including ***edited_site***, ***modified_site***, ***polyA_signal_sequence***, ***polyA_site***, ***repression_signal_sequence***, ***degradation_signal_sequence***, ***attenuator***. The final product is produced through processed transcript, such as ***ORF, sORF, miRNA, tRNA, rRNA***. It also needs many regulators, for instance, ***start codon*** and ***stop codon***. These inheritance relations of sequences are stored in ***id*** and ***parentId***.

Here we interpret how this gene biological process should be described using NGFF (see '*Example for description the organization of NGFF format.xls*'), Figure 1 shows the process according to the file. We can see ***Locus*** L01 produced two ***primary*** transcripts L01.pt1 and L01.pt2. Then L01.pt1 produced two ***processed*** transcripts L01.t1 and L01.t2, while L01.pt2 produced L01.t2, which is also the ***product*** of L01.pt1. Finally, ***processed*** transcripts L01.t1 produced L01.t1.p1- L01.t1.p5, while L01.t2 produced L01.t2.p1- L01.t2.p4. Some regulator elements during the process have been recorded, such as ***enhancer*** and ***promoter*** of Locus L01, and the ***start codon***, ***end codon*** of L01.t1.p1.

4.1.2. According to biological function

These relations are stored in some special attribute type in attribute column. For the products in the same ***locus***, although they are from different processed sequence, may have the same sequence. Logically, these are regarded as one ***product_cluster*** (In Figure 1, product L01.t1.p1 and L01.t2.p1, they have the same sequence in the same ***locus***, so they have the same ***product_cluster*** id L01.g2.pc1). Since the common sequence segments tend to have the same or close biological function, we treat these sequence shared ***product_cluster*** of one locus as one ***gene*** (In Figure 1, L01.g1.pc1 and L01.g1.pc2 are of the same ***gene*** L01.g1 because of the shared sequence. However, L01.g2.pc1 is an independent ***gene*** L01.g2 because of the independent sequence source). One the other hand, the sequences of genes of the same ***product_biotype*** from different locus may be remarkably similar to each other (such as the mature miRNAs of short length and paralogues). Therefore, we assign these genes being high similarity of different locus with the same ***gene_name***. In logical, the order is ***product -> product_cluster -> gene -> gene_name***. We set these classification names as the member of attribute names.

4.2. Examples

4.2.1. coding RNA (linear)

ORF and ***sORF***: if the length of amino acid less than 50, the polypeptide is defined as ***sORF***, otherwise, is defined as ***ORF***.

4.2.2. non-coding RNA (linear)

NGFF format can describe the structure of linear non-coding RNA. Primary transcript L01.pt6 produces one lncRNA L01.t10, which is composed of four exons. Primary transcript L01.pt7 produces one pre-miRNA L01.t11 and further produces two miRNA product, mature miRNA L01.t11.p1 and L01.t11.p2. The featureType of linear non-coding RNA should be "***mature***". Especially, here we attempt to describe the product lncRNA L01.t10.p1 with the relative position from parent sequence (processed transcripts, the sequence joined by multiple exon segments 6124-6342,6398-6478,6675-6787,6987-7063). And the start, end numbers are of the non-zero negative integer.
One of the miRNA product L01.t11.p2 is derived from two exons, so it shows two lines of the product description.

The ngff file corresponding to Figure 2 is *Linear non-coding transcripts.xls*.

Following table is the simple version (skip attributes).

#seqId	start	end	strand	level	featureType	id	parentId	attributes
Chr1	1	30427671	+	chromosome	reference	Chr1	genome	.
Chr1	5928	8737	+	locus	locus	L01	Chr1	.
Chr1	6124	7663	+	primary	region	L01.pt6	L01	.
Chr1	6124	6578	+	primary	region	L01.pt7	L01	.
Chr1	6124	6342	+	processed	exon	L01.t10	L01.pt6	.
Chr1	6398	6478	+	processed	exon	L01.t10	L01.pt6	.
Chr1	6675	6787	+	processed	exon	L01.t10	L01.pt6	.
Chr1	6987	7063	+	processed	exon	L01.t10	L01.pt6	.
Chr1	6124	6342	+	processed	exon	L01.t11	L01.pt7	.
Chr1	6398	6478	+	processed	exon	L01.t11	L01.pt7	.
Chr1	-1	-490	+	product	mature	L01.t10.p1	L01.t10	.
Chr1	6234	6252	+	product	mature	L01.t11.p1	L01.t11	.
Chr1	6340	6342	+	product	mature	L01.t11.p2	L01.t11	.
Chr1	6398	6413	+	product	mature	L01.t11.p2	L01.t11	.

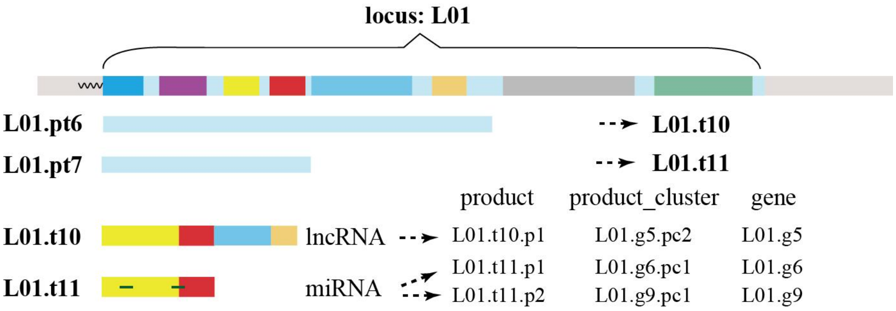


Figure2. Linear non-coding transcripts

4.2.3. Circular RNA

NGFF format can describe the structure of circular RNA. Primary transcript L01.pt4 produced one linear processed transcript L01.t5 and three circle processed transcripts: L01.t6, L01.t7, L01.t8. L01.t6 is composed of two exons, while L01.t7 is composed of two exons and one intron, L01.t8 is composed of only one intron. The featureType of circleRNA should be "***mature***". Especially, here we attempt to describe the final products circle RNA with the relative position from parent sequence (processed transcripts) and the start, end numbers are of the non-zero negative integer.

The ngff file corresponding to Figure 3 is: *circle RNA transcripts.xls*

Following table is the simple version (skip attributes).

#seqId	start	end	strand	level	featureType	id	parentId	attributes
Chr1	5928	8737	+	locus	locus	L01	Chr1	.
Chr1	6036	7663	+	primary	region	L01.pt4	L01	.
Chr1	6036	6098	+	processed	exon	L01.t5	L01.pt3,L01.pt4	.
Chr1	6675	6787	+	processed	exon	L01.t5	L01.pt3,L01.pt4	.
Chr1	6987	7063	+	processed	exon	L01.t5	L01.pt3,L01.pt4	.
Chr1	6124	6342	+	processed	exon	L01.t6	L01.pt4	.
Chr1	6398	6478	+	processed	exon	L01.t6	L01.pt4	.
Chr1	6124	6342	+	processed	exon	L01.t7	L01.pt4	.
Chr1	6343	6397	+	processed	intron	L01.t7	L01.pt4	.
Chr1	6398	6478	+	processed	exon	L01.t7	L01.pt4	.
Chr1	6343	6397	+	processed	intron	L01.t8	L01.pt4	.
Chr1	-1	-300	+	product	mature	L01.t6.p1	L01.t6	.
Chr1	-1	-355	+	product	mature	L01.t7.p1	L01.t7	.
Chr1	-1	-55	+	product	mature	L01.t8.p1	L01.t8	.

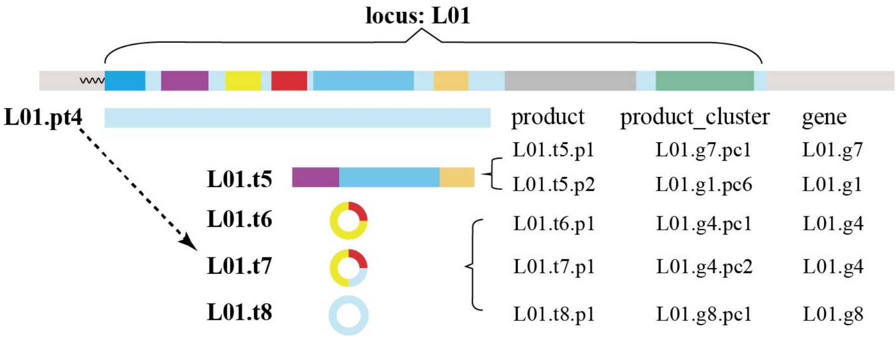


Figure3. circle RNA transcripts

4.2.4. polycistronic transcripts

This is the case in which a single (possibly spliced) transcript encodes multiple open reading frames that generate independent protein products.

The ngff file corresponding to Figure 4 is: *polycistronic transcript.xls*

Following table is the simple version (abbreviated level: ***chro (chromosome); prim (primary); proc (processed); prod (product)***).

#seqId	start	end	strand	level	featureType	id	parentId	attributes
Chr1	XX	XX	+	chro	reference	Chr1	genome	seqSource=major;circular=FALSE;
Chr1	XX	XX	+	locus	locus	L01	Chr1	.
Chr1	XX	XX	+	prim	region	L01.pt1	L01	primary_biotype=primary_transcript;transcribed=TRUE;
Chr1	XX	XX	+	proc	exon	L01.t1	L01.pt1	processed_biotype=processed_transcript;product_biotype=ORF;
Chr1	XX	XX	+	proc	exon	L01.t1	L01.pt1	processed_biotype=processed_transcript;product_biotype=ORF;
Chr1	XX	XX	+	proc	exon	L01.t1	L01.pt1	processed_biotype=processed_transcript;product_biotype=ORF;
Chr1	XX	XX	+	proc	exon	L01.t1	L01.pt1	processed_biotype=processed_transcript;product_biotype=ORF;
Chr1	XX	XX	+	proc	exon	L01.t1	L01.pt1	processed_biotype=processed_transcript;product_biotype=ORF;
Chr1	XX	XX	+	prod	CDS	L01.t1.p1	L01.t1	product_biotype=ORF;product_cluster=L01.g1.pc1;gene=L01.g1;gene_name=L01.gn1;phase=0;
Chr1	XX	XX	+	prod	CDS	L01.t1.p2	L01.t1	product_biotype=ORF;product_cluster=L01.g2.pc1;gene=L01.g2;gene_name=L01.gn2;phase=0;
Chr1	XX	XX	+	prod	CDS	L01.t1.p3	L01.t1	product_biotype=ORF;product_cluster=L01.g3.pc1;gene=L01.g3;gene_name=L01.gn3;phase=0;
Chr1	XX	XX	+	prod	CDS	L01.t1.p4	L01.t1	product_biotype=ORF;product_cluster=L01.g4.pc1;gene=L01.g4;gene_name=L01.gn4;phase=0;

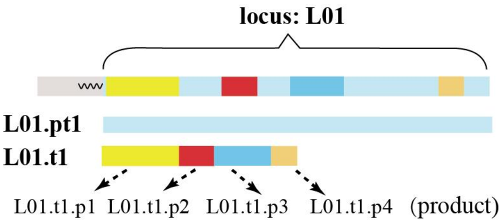


Figure4. polycistronic transcript

4.2.5. Trans-spliced transcript – intergenic and intragenic

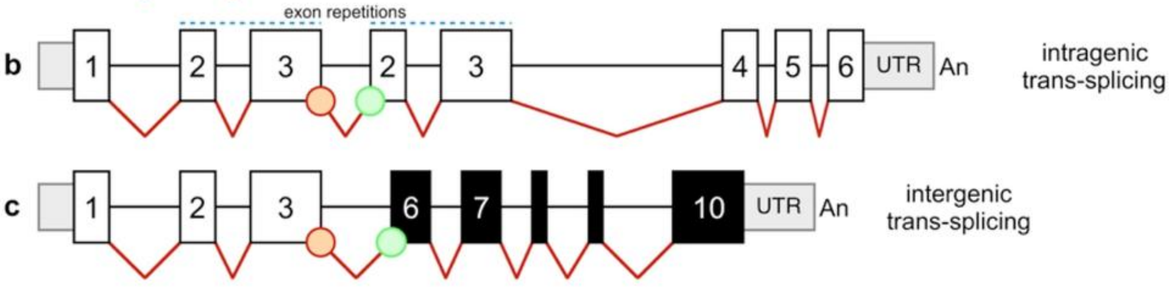


Figure5. intergenic and intragenic trans-spliced transcript

This occurs when multiple genes contribute to a processed transcript via a trans-splicing reaction. In this situation, transcripts are formed according to the order in attribute order value.

Following is one example of fusion gene. Especially, in these lines with level of processed and product, '***order'*** need to be set. The first segment has a value of '1', while the latter segment has a value of 2 and so on. The sequences of fusion gene are joined as this order [1-2-3…], see Figure 6.

The ngff file corresponding to Figure 6 is: *fusion gene transcripts.xls*.

Following table is the simple version (skip some attributes).

#seqId	start	end	strand	level	featureType	id	parentId	attributes
Chr1	1	30427671	+	chromosome	reference	Chr1	genome	.
Chr2	1	19698289	+	chromosome	reference	Chr2	genome	.
Chr3	1	23459830	+	chromosome	reference	Chr3	genome	.
Chr4	1	18585056	+	chromosome	reference	Chr4	genome	.
Chr5	1	26975502	+	chromosome	reference	Chr5	genome	.
ChrC	1	154478	+	chromosome	reference	ChrC	genome	.
ChrM	1	366924	+	chromosome	reference	ChrM	genome	.
Chr1	5928	8737	+	locus	locus	L01	Chr1	.
Chr1	3212	3234	-	locus	regulator	L01	Chr1	.
Chr1	5917	5927	+	locus	regulator	L01	Chr1	.
Chr1	6036	8737	+	primary	region	L01.pt3	L01	.
Chr1	6036	6098	+	processed	exon	L01.t4	L01.pt3	.
Chr1	6675	6787	+	processed	exon	L01.t4	L01.pt3	.
Chr1	7985	8243	+	processed	exon	L01.t4	L01.pt3	.
Chr1	8486	8645	+	processed	exon	L01.t4	L01.pt3	.
Chr1	6036	6098	+	processed	exon	L01.t5	L01.pt3,L02.pt1	order=1;
Chr1	6675	6787	+	processed	exon	L01.t5	L01.pt3,L02.pt1	order=1;
Chr1	7985	8243	+	processed	exon	L01.t5	L01.pt3,L02.pt1	order=1;
Chr1	8486	8644	+	processed	exon	L01.t5	L01.pt3,L02.pt1	order=1;
Chr2	7000	7299	-	processed	exon	L01.t5	L01.pt3,L02.pt1	order=2;
Chr2	9001	9330	-	processed	exon	L01.t5	L01.pt3,L02.pt1	order=2;
Chr1	6049	6098	+	product	CDS	L01.t4.p2	L01.t4	.
Chr1	6675	6787	+	product	CDS	L01.t4.p2	L01.t4	.
Chr1	7985	7998	+	product	CDS	L01.t4.p2	L01.t4	.
Chr1	6036	6098	+	product	CDS	L01.t5.p1	L01.t5	order=1;
Chr1	6675	6787	+	product	CDS	L01.t5.p1	L01.t5	order=1;
Chr1	7985	8243	+	product	CDS	L01.t5.p1	L01.t5	order=1;
Chr1	8486	8644	+	product	CDS	L01.t5.p1	L01.t5	order=1;
Chr2	7001	7299	-	product	CDS	L01.t5.p1	L01.t5	order=2;
Chr2	9000	9330	-	product	CDS	L01.t5.p1	L01.t5	order=2;
Chr2	7000	9377	-	locus	locus	L02	Chr2	.
Chr2	7000	9370	-	primary	region	L02.pt1	L02	.
Chr2	7001	7299	-	processed	exon	L02.t1	L02.pt1	.
Chr2	9000	9330	-	processed	exon	L02.t1	L02.pt1	.
Chr2	7001	7299	-	product	CDS	L02.t1.p1	L02.t1	.
Chr2	9000	9330	-	product	CDS	L02.t1.p1	L02.t1	.

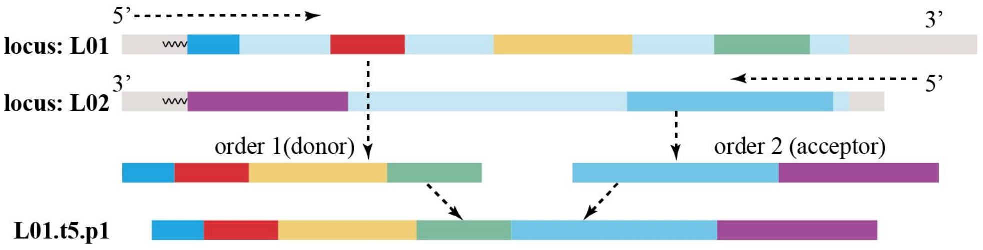


Figure6. fusion gene transcripts

4.2.6. RNA editing

RNA editing annotation is accomplished through attribution ***edited_site***. You can see detail in 3.8.10 about the ***edited_site*** value:
The format for this value is “[seqId]:[site_in_genome][position_in_genome][edited_site_in_genome]”, such as ‘Chr1:A6036C’, which means the nucleotide at Chr1, 6036bp in genome, A, is edited to C based on genome system. If the strand of the transcript is minus, the processed transcript of this position will convert to G, as C is based on genome system.
Following is the example ngff annotation contained RNA editing. When use ngfftools to retrieve sequence (module: ***seq***), it will consider edited site and produce edited sequences.

The ngff file corresponding to this part is: *RNA editing.xls*.

Following table is the simple version (skip some attributes).

#seqId	start	end	strand	level	featureType	id	parentId	attributes
Chr1	1	30427671	+	chromosome	reference	Chr1	genome	.
Chr2	1	19698289	+	chromosome	reference	Chr2	genome	.
Chr3	1	23459830	+	chromosome	reference	Chr3	genome	.
Chr4	1	18585056	+	chromosome	reference	Chr4	genome	.
Chr5	1	26975502	+	chromosome	reference	Chr5	genome	.
ChrC	1	154478	+	chromosome	reference	ChrC	genome	.
ChrM	1	366924	+	chromosome	reference	ChrM	genome	.
Chr1	5928	8737	+	locus	locus	L01	Chr1	.
Chr1	3212	3234	-	locus	regulator	L01	Chr1	.
Chr1	5917	5927	+	locus	regulator	L01	Chr1	.
Chr1	6036	8737	+	primary	region	L01.pt3	L01	.
Chr1	6036	6098	+	processed	exon	L01.t4	L01.pt3	edited_site=Chr1:A6036C,Chr1:A6049C;
Chr1	6675	6787	+	processed	exon	L01.t4	L01.pt3	edited_site=Chr1:C6675T;
Chr1	7985	8243	+	processed	exon	L01.t4	L01.pt3	.
Chr1	8486	8645	+	processed	exon	L01.t4	L01.pt3	.
Chr1	6049	6098	+	product	CDS	L01.t4.p2	L01.t4	.
Chr1	6675	6787	+	product	CDS	L01.t4.p2	L01.t4	.
Chr1	7985	7998	+	product	CDS	L01.t4.p2	L01.t4	.
Chr1	6036	6098	-	locus	locus	L02	Chr1	.
Chr1	6036	6098	-	primary	region	L02.pt1	L02	.
Chr1	6036	6098	-	processed	exon	L02.t1	L02.pt1	edited_site=Chr1:A6036G,Chr1:A6049G;
Chr1	6048	6098	-	product	CDS	L02.t1.p1	L02.t1	.

4.2.7. RNA with programmed frameshift

This event occurs when the ribosome performs a programmed frameshift during translation in order to skip over an in-frame stop codon. The frameshift may occur forward or backward. The representation of this is to make the CDS discontinuous.

The CDS segment that represent the new reading frame will always has a phase of 0 since the ribosome is moving and thus redefining the codon. For details and example, please see 3.8.11 frameshift and Figure 1

It is suggested that the mRNA be tagged with the appropriate SO transcript attributes such as "minus_1_translational_frameshift" (SO:1000069). This will allow all such programmed frameshift mRNAs to be recovered with a query. The accession for "plus_1_translational_frameshift" is SO:1001263.

4.2.8. A single locus including multiply genes

See the example of Figure 1. A single ***locus*** has three genes: L01.g1, L01.g2, L01.g7.

4.2.9. A single gene from multiply loci

This is a situation of a single gene from multiply loci, for example, the paralogous genes. Figure 7 shows two paralogous genes L01.t1.p2 and L02.t1.p2, and they have the same ***gene*** string ‘L01.g1’.

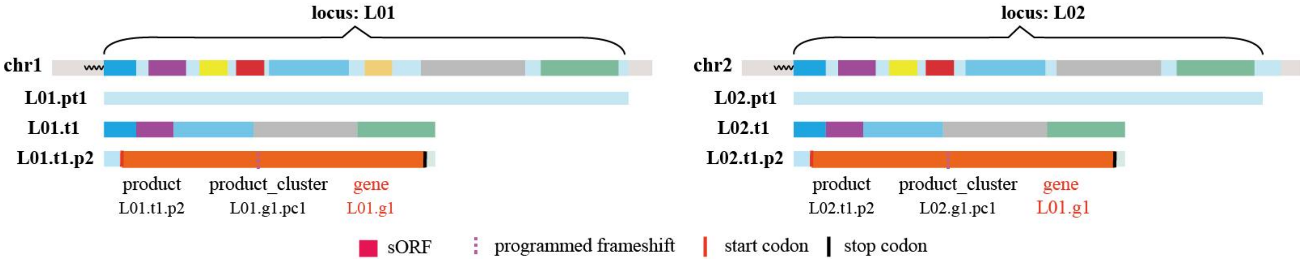


Figure7. A single gene from multiply loci

5. Comparison to other format

term	BED	GTF	GFF3	NGFF
parent (part-of) relationships	-	+	+	+
canonical genes non-coding transcripts	-	+	+	+
polycistronic transcripts	-	-	+	+
genes containing smORF	-	-	-	+
RNA editing	-	-	-	+
programmed frameshifts	-	-	+	+
Organization according to biological process	-	+	+	+
Organization according to biological function	-	-	-	+
Circular RNA	-	-	-	+
Trans-spliced transcript (two)	-	-	+	+
Trans-spliced transcript (multiple)	-	-	-	+
Relative position of parent sequence	-	-	-	+
Multiple regulator biotype	-	-	-	+
alignments	-	-	+	-