

Bioinfo standard : GFF File

GFF : Generic Feature Format Version 3

Specifications

<https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>

Samples

S1

```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + .
ID=tfbs00001;Parent=gene00001
ctg123 . mRNA 1050 9000 . + .
ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA 1050 9000 . + .
ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA 1300 9000 . + .
ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon 1300 1500 . + . ID=exon00001;Parent=mRNA00003
ctg123 . exon 1050 1500 . + .
ID=exon00002;Parent=mRNA00001,mRNA00002
```

S2

```
ctg123 . cDNA_match 1050 9000 6.2e-45 + .
ID=match00001;Target=cdna0123 12 2964;Gap=M451 D3499 M501
D1499 M2001
```

Recommended library

?

Comments

Element length

In S1, in element mRNA00001, what is the length of the element ? does the base 9000 part of the element ?

The specification is not that clear on this. But other samples found in the specification make the answer to be :

$L = \text{end} - \text{start} + 1$, the base 'end' is part of the element.

Target length = target end - target start + 1

Ref length = ref end - ref start + 1

Argument

In S2, target length is the sum of gap M bases : $M451 + M501 + M2001 = 2953$ bases = $2964 - 12 + 1$

In S2, ref length is the sum of gap M and D bases : $M451 + D3499 + M501 + D1499 + M2001 = 7951$ bases = $9000 - 1050 + 1$