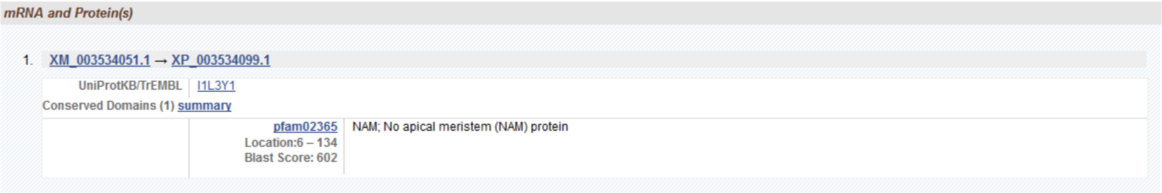
Tutorial for expert gene structure annotation using Artemis in the Banana Genome Hub

Contents

Step 1: Check of automatic annotation

- ### Summary of errors in structural annotation

1) Arbitrary intron annotation inside an exon (exon results cut in two or more part separated by erroneous introns). This error has been frequently observed in the automatic annotations obtained by GAZE pipeline). In this case, in the 'artemis' gene representation, two or more consecutive exons, sharing the same reading frame (i.e. in the same line) are separated by introns(s) not containing stop codons. The correction consists in the elimination of the erroneous intron(s).
2) Lack of one or more exons in the automatic annotation. This error can be detected as gap in the sequence alignments between the analyzed gene and the most similar ones or by comparing their gene structures. In order to detect le lacking portion of the gene, one of the most similar amino acid sequences found by Blastp could be used to perform a tBLASTn on the genomic sequence of the genome/chromosome/scaffold (using the tools of the web site hosting the genome sequence or, in local, for example by using BioEdit). The amino acid sequence of a similar sequence can be obtained by clicking in the 'mRNA and Protein(s)' section in the 'GeneID' page.


mRNA and Protein(s)

1. XM_003534051.1 → XP_003534099.1

UniProtKB/TrEMBL | I1L3Y1

Conserved Domains (1) summary

	pfam02365 Location:6 – 134 Blast Score: 602	NAM; No apical meristem (NAM) protein

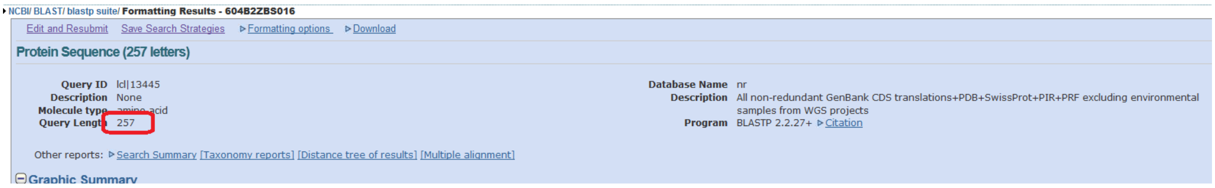
3) Partial annotation of the gene (detected by difference of size between analyzed and similar genes in other species). In order to detect le lacking portion of the gene, perform a tBLASTn as in point 2.
4) Subdivision of a gene in two or more independent annotations (the most of coding exons are detected, but separated in different genes).
5) Merging of two independent genes in a unique annotation (chimerical artifact).
6) Wrong definition of exon ends.

- ### Suggested protocol to check annotation

Perform a Blastp search (default parameter) at NCBI with the amino acid sequence of a protein-coding gene retrieved via the Banana Genome Hub (http://banana-genome.cirad.fr/). To retrieve banana sequences, you can read the documentation section (http://banana-genome.cirad.fr/documentation).

The examination of the Blastp results help to determine whether automatic annotation contain major errors e.g. length of the subject sequences found by the Blastp significantly larger than the query sequence, suggesting its possible incompleteness. Conversely, shorter subject sequences suggest that automatic annotation includes genome region not belonging to the gene.

As underlined in red, the length of the query sequence is reported at the top of the result page:



▸ NCBI/ BLAST/ blastp suite/ **Formatting Results - 604B2ZBS016**

Edit and Resubmit   Save Search Strategies   ▷ Formatting options   ▷ Download

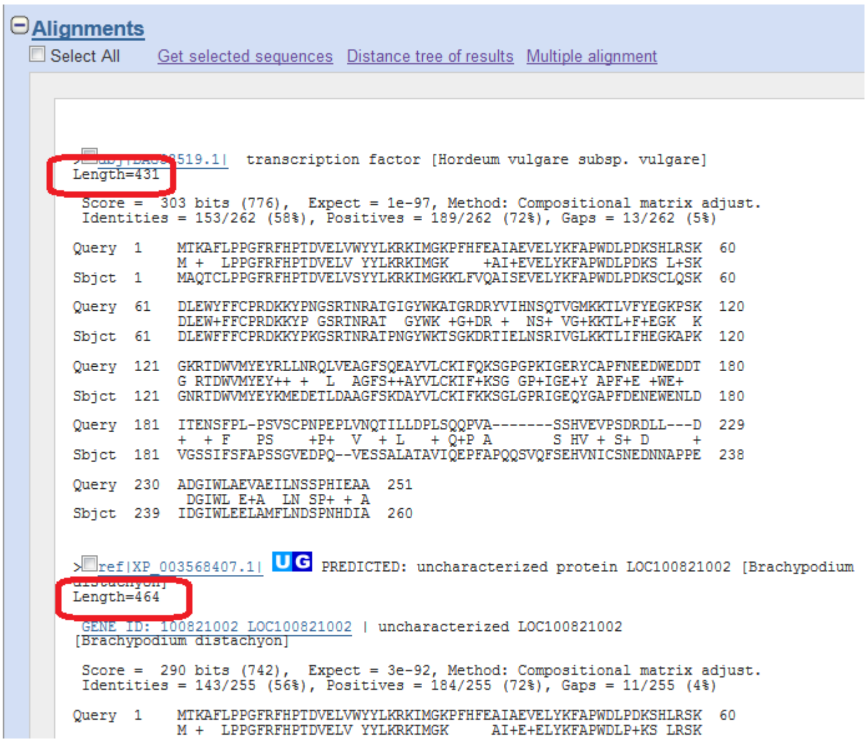
Protein Sequence (257 letters)

Query ID	lcl|13445	**Database Name**	nr
Description	None	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid	**Program**	BLASTP 2.2.27+ ▷ Citation
Query Length	257		

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

⊟ Graphic Summary

The length of the subject sequence is reported at the top of each sequence alignment:

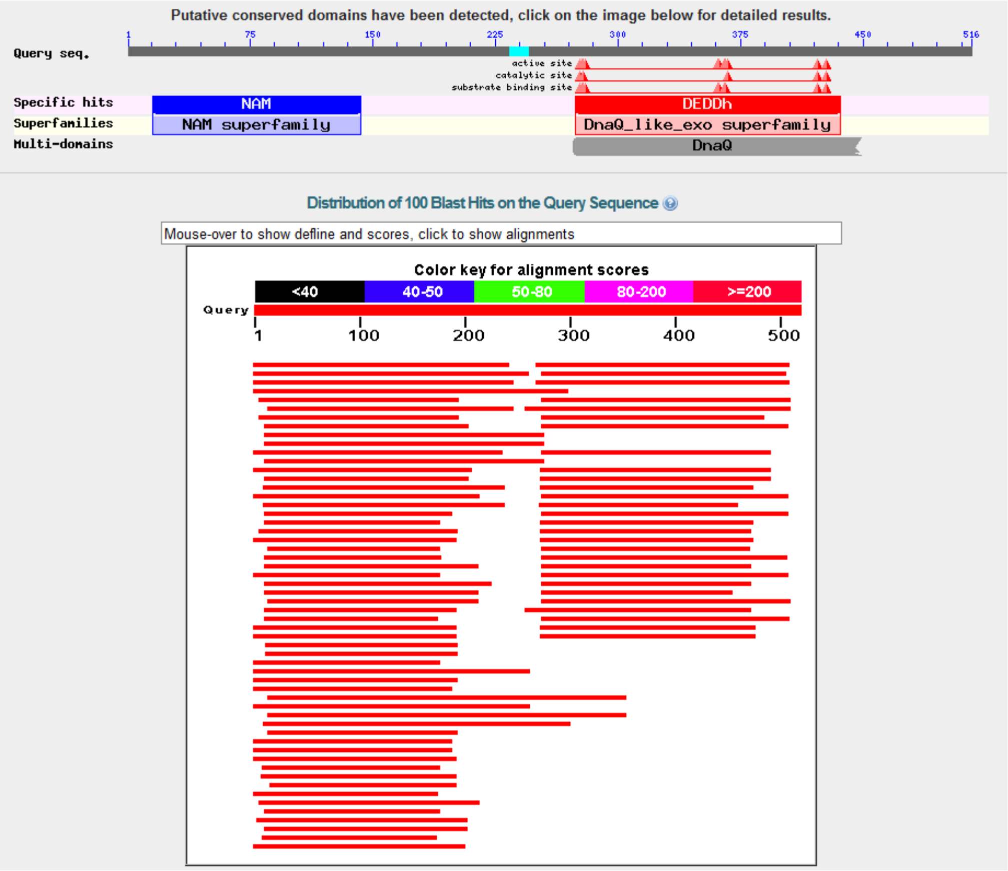


⊟ Alignments

□ Select All   Get selected sequences   Distance tree of results   Multiple alignment

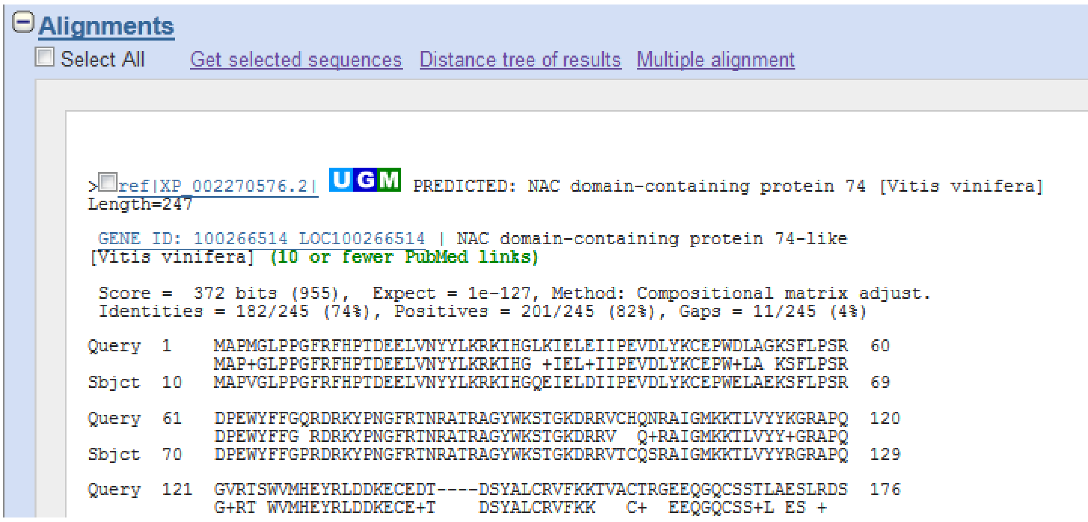
```
>□dbj|BA[illegible]519.1|  transcription factor [Hordeum vulgare subsp. vulgare]
Length=431

 Score =  303 bits (776),  Expect = 1e-97, Method: Compositional matrix adjust.
 Identities = 153/262 (58%), Positives = 189/262 (72%), Gaps = 13/262 (5%)

Query  1    MTKAFLPPGFRFHPTDVELVWYYLKRKIMGKPFHFEAIAEVELYKFAPWDLPDKSHLRSK  60
            M +  LPPGFRFHPTDVELV YYLKRKIMGK    +AI+EVELYKFAPWDLPDKS L+SK
Sbjct  1    MAQTCLPPGFRFHPTDVELVSYYLKRKIMGKKLFVQAISEVELYKFAPWDLPDKSCLQSK  60

Query  61   DLEWYFFCPRDKKYPNGSRTNRATGIGYWKATGRDRYVIHNSQTVGMKKTLVFYEGKPSK  120
            DLEW+FFCPRDKKYP GSRTNRAT  GYWK +G+DR +  NS+ VG+KKTL+F+EGK  K
Sbjct  61   DLEWFFFCPRDKKYPKGSRTNRATPNGYWKTSGKDRTIELNSRIVGLKKTLIFHEGKAPK  120

Query  121  GKRTDWVMYEYRLLNRQLVEAGFSQEAYVLCKIFQKSGPGPKIGERYCAPFNEEDWEDDT  180
            G RTDWVMYEY++ +  L  AGFS++AYVLCKIF+KSG GP+IGE+Y APF+E +WE+
Sbjct  121  GNRTDWVMYEYKMEDETLDAAGFSKDAYVLCKIFKKSGLGPRIGEQYGAPFDENEWENLD  180

Query  181  ITENSFPL-PSVSCPNPEPLVNQTILLDPLSQQPVA-------SSHVEVPSDRDLL---D  229
            +  + F   PS    +P+  V  + L   + Q+P A       S HV + S+ D     +
Sbjct  181  VGSSIFSFAPSSGVEDPQ--VESSALATAVIQEPFAPQQSVQFSEHVNICSNEDNNAPPE  238

Query  230  ADGIWLAEVAEILNSSPHIEAA  251
             DGIWL E+A  LN SP+ + A
Sbjct  239  IDGIWLEELAMFLNDSPNHDIA  260


>□ref|XP_003568407.1| UG PREDICTED: uncharacterized protein LOC100821002 [Brachypodium
distachyon]
Length=464

 GENE ID: 100821002 LOC100821002 | uncharacterized LOC100821002
[Brachypodium distachyon]

 Score =  290 bits (742),  Expect = 3e-92, Method: Compositional matrix adjust.
 Identities = 143/255 (56%), Positives = 184/255 (72%), Gaps = 11/255 (4%)

Query  1    MTKAFLPPGFRFHPTDVELVWYYLKRKIMGKPFHFEAIAEVELYKFAPWDLPDKSHLRSK  60
            M +  LPPGFRFHPTDVELV YYLKRKIMGK     AI+E+ELYKFAPWDLP+KS LRSK
```

A recurrent case is an automatic annotation merging two neighbor genes (chimerical annotation). This error can be easily detected by looking at the graphical representation of sequence alignments:


Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
1
75
150
225
300
375
450
516
active site
catalytic site
substrate binding site
Specific hits
Superfamilies
Multi-domains
NAM
NAM superfamily
DEDDh
DnaQ_like_exo superfamily
DnaQ
Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500

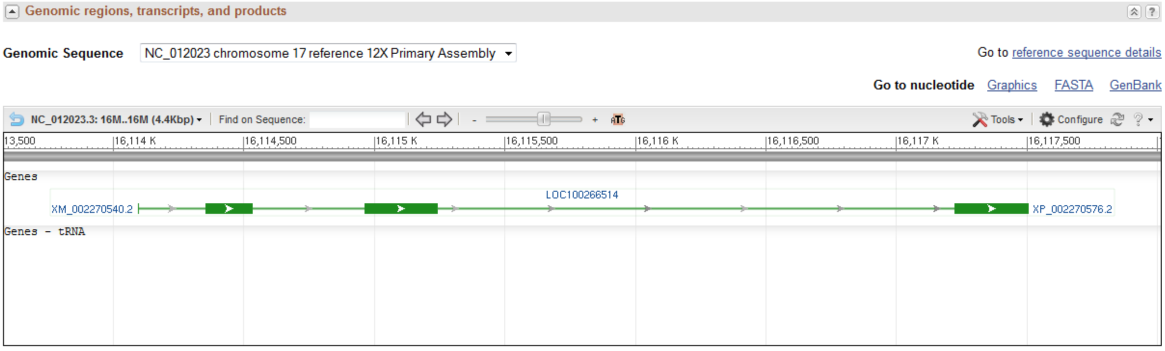
In-depth verifications can be performed by comparing the gene structure of the query sequence and the most similar genes found by Blastp.

To do that, click on the link "GENE ID: XXXX" among the best Blastp results. *Empirical results showed that Vitis vinifera and Ricinus communis can sometimes be good models although being not monocots plants.*



```
>□ref|XP_002270576.2| UGM PREDICTED: NAC domain-containing protein 74 [Vitis vinifera]
Length=247

 GENE ID: 100266514 LOC100266514 | NAC domain-containing protein 74-like
[Vitis vinifera] (10 or fewer PubMed links)

 Score =  372 bits (955),  Expect = 1e-127, Method: Compositional matrix adjust.
 Identities = 182/245 (74%), Positives = 201/245 (82%), Gaps = 11/245 (4%)

Query  1    MAPMGLPPGFRFHPTDEELVNYYLKRKIHGLKIELEIIPEVDLYKCEPWDLAGKSFLPSR  60
            MAP+GLPPGFRFHPTDEELVNYYLKRKIHG +IEL+IIPEVDLYKCEPW+LA KSFLPSR
Sbjct  10   MAPVGLPPGFRFHPTDEELVNYYLKRKIHGQEIELDIIPEVDLYKCEPWELAEKSFLPSR  69

Query  61   DPEWYFFGQRDRKYPNGFRTNRATRAGYWKSTGKDRRVCHQNRAIGMKKTLVYYKGRAPQ  120
            DPEWYFFG RDRKYPNGFRTNRATRAGYWKSTGKDRRV  Q+RAIGMKKTLVYY+GRAPQ
Sbjct  70   DPEWYFFGPRDRKYPNGFRTNRATRAGYWKSTGKDRRVTCQSRAIGMKKTLVYYRGRAPQ  129

Query  121  GVRTSWVMHEYRLDDKECEDT----DSYALCRVFKKTVACTRGEEQGQCSSTLAESLRDS  176
            G+RT WVMHEYRLDDKECE+T    DSYALCRVFKK   C+  EEQGQCSS+L ES +
```

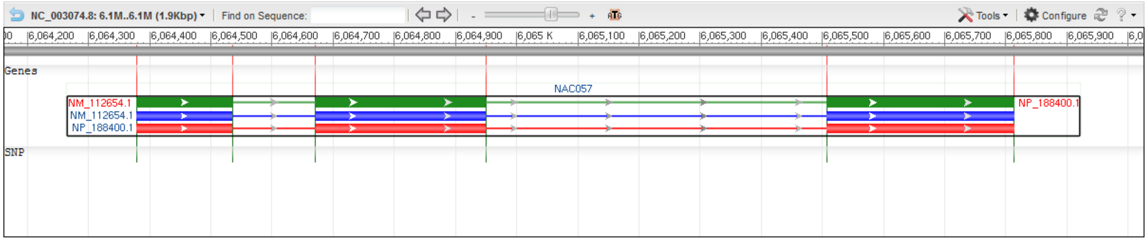
The gene id page of the NCBI contains informative elements such as a simplified Genome browser which shows the gene structure (number and size of the introns/exons) as annotated in their respective genomes. These structures can be compared with the query gene, shown in the GBMa.


Genomic regions, transcripts, and products

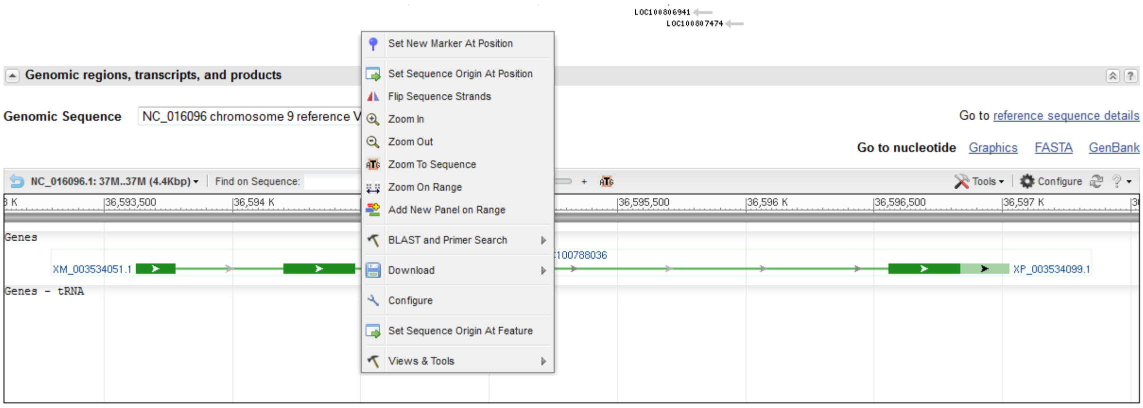
Genomic Sequence: NC_012023 chromosome 17 reference 12X Primary Assembly

If no clear structural differences can be detected, it is likely that the Musa gene was correctly annotated. On the contrary, if major differences are detected in comparisons with some more similar genes, it is probable that the automatic annotation need some adjustments.

It is even possible to verify the consistence of the exon ends between the Musa gene and the more similar genes detected by Blastp. To perform this verification, click (left mouse button) on an exon (green bar) in the simplified genome browser (a red and a blue bar will appear)

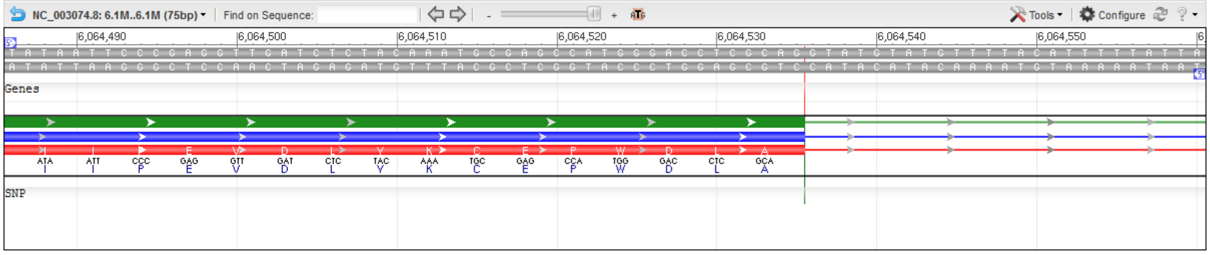


Then click (right mouse button) on the exon end that you want to verify and select “Zoom to sequence” in the contextual menu. These two steps could also be executed in inversed order.


Genomic regions, transcripts, and products

Genomic Sequence: NC_016096 chromosome 9 reference V

Now, it is possible to see the nucleotide and the amino acid sequences of the exon end and compare it to the corresponding one of the Musa gene. In the most cases, both corresponding exons end in the same way (i.e. similar sequence and identical reading phase: in the example of the following figure, the last exon nucleotide has the +1 position (i.e. it is the first nucleotide of the codon following the one coding for ‘A’).


NC_003074.8: 6.1M..6.1M (75bp)
Find on Sequence:
Tools
Configure
6,064,490
6,064,500
6,064,510
6,064,520
6,064,530
6,064,540
6,064,550
5'
T A T A A T T C C C G A G G T T G A T C T C T A C A A A T G C G A G C C A T G G G A C C T C G C A G G T A T G T A T G T T T T A C A T T T T T A T T A
A T A T T A A G G G C T C C A A C T A G A G A T G T T T A C G C T C G G T A C C C T G G A G C G T C C A T A C A T A C A A A A T G T A A A A A T A A T
5'
Genes
ATA I
ATT I
CCC P
GAG E
GTT V
GAT D
CTC L
TAC Y
AAA K
TGC C
GAG E
CCA P
TGG W
GAC D
CTC L
GCA A
SNP

This verification is useful when gaps appeared in the Blastp alignments or when exon sizes are different, that suggests possible errors in the definition of exon ends.

In general, when differences were observed between the query gene and one of the most similar subject genes, two or three additional similar genes need to be compared. If similar inconsistencies are observed, it is likely that errors were produced by the automatic annotation. A corrected version of the gene annotation needs to be made and corrections need to be performed in the database containing the annotation.

Step 2: Using Artemis to correct the annotation and save the modification

The Artemis software allows to handle sequence annotations and to modify them; however, in order to save annotation modifications, a specific protocol must to be followed. Each gene annotation on 'artemis' is composed by 4 main elements that can be modified. These elements can be visualized in a Gene builder (opened by the shortcut 'Ctrl + e' after the selection of an element).

The elements are:

gene: a continuous region included between the beginning and the end of the transcription (coordinates provided in the 'Location' section).

mRNA: similar to 'gene', but, for poly-exonic genes, in the GBMa it appears as group of regions joined by traits.

exon: it corresponds to the spliced mRNA, i.e. the coding region (CDS) plus , if present, the 5' et 3' untranslated regions (UTR). For poly-exonic genes, the coordinates are composed by the ends of each exon. Its whole ends have to coincide with the ones if 'gene'.

polypeptide: it corresponds to the coding portion (from the ATG to the 'stop codon'). Its coordinates coincides with the beginning and the end of the translated region (also for poly-exonic genes).

UTR: At the same level of the 'polypeptide' element, if available, the '**five prime UTR**' and the '**three prime UTR**' elements can be present, being their coordinates the beginning of 'gene' and the position before the 'ATG' (5' UTR) and the position following the 'stop codon' and the end of the 'gene' (3' UTR).

Structural annotation rules

The structural annotation of a given genomic feature is, by convention, marked by its first and its last positions in the sequence, separated by two full stops: e.g. '**12928407.. 12928848**'.

When an element is composed by more two or more sub-elements (in the case of poly-exonic genes, the element 'exon' is composed by more than one elements (the exons)), the structural annotation will be indicated by 'join' followed by, between brackets, the coordinates of each sub-element separated by a comma (','): e.g.

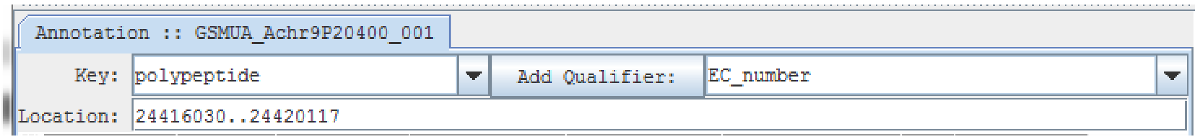
```
join(12928407..12928848,12928921..12929261,12929557..12929831,12929907..12930069)
```

Finally, if the element is reverse-oriented, the structural annotation will be indicated by 'complement' and, between brackets, the coordinates of the element: e.g.

```
complement(join(12928407..12928848,12928921..12929261,12929557..12929831,12929907..12930069))
```

Modification of existing elements

The coordinates of all existing features can be modified in the 'Location' section or in the window containing the graphic representation of the annotation, by drifting the ends of the element to modify.


Annotation :: GSMUA_Achr9P20400_001

Key: polypeptide

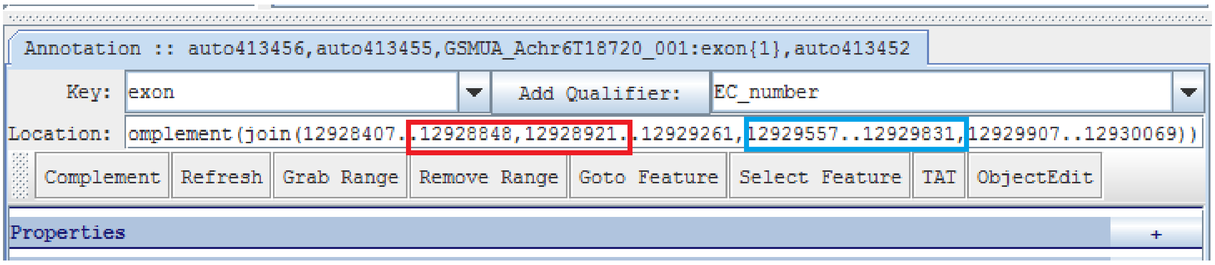
EC_number

Location: 24416030..24420117

- ### Intron or exon elimination

In order to eliminate an intron (i.e. merge two exons) the end position of the first exon to merge and the first position of the other element to merge have to be eliminated in the 'Location' section; along with two full stops (red rectangle in the following figure).

In order to eliminate an exon, its coordinates have to be eliminated (blue rectangle in the following figure).


Annotation :: auto413456,auto413455,GSMUA_Achr6T18720_001:exon{1},auto413452

Key: exon

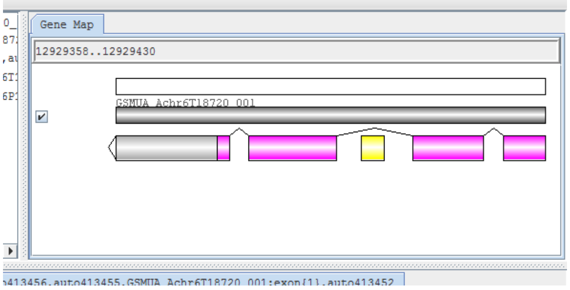
EC_number

Location: omplement(join(12928407..12928848,12928921..12929261,12929557..12929831,12929907..12930069))

- ### Exon creation

In order to create an exon, do NOT introduce its coordinates in the 'Location' section because this modification will not be saved in the database. In fact, exon creation needs a specific and mandatory protocol.

The exon has to be created in the 'Gene Map' section (see following figure) located in the up-right side of the 'gene builder' window. Using the mouse (pushing in the left button) place the next exon in the approximate position (it will appear as a yellow rectangle in the 'Gene Map').

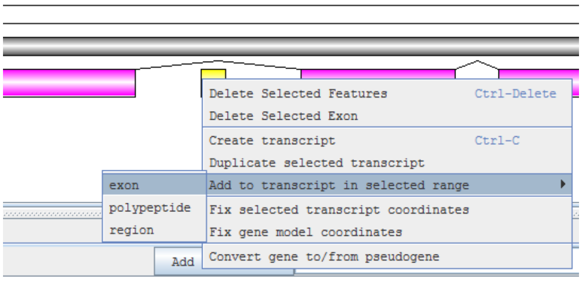


12929358..12929430

☑

413456,auto413455,GSMUA_Achr6T18720_001:exon{1},auto413452

Then, click with the right button in the rectangle and select 'Add to transcript in selected range' > 'exon' (see following figure). The new exon will be added to the gene structure and it will be possible to replace the approximate coordinates with the exact ones.


Delete Selected Features
Ctrl-Delete
Delete Selected Exon
Create transcript
Ctrl-C
Duplicate selected transcript
exon
Add to transcript in selected range
polypeptide
Fix selected transcript coordinates
region
Fix gene model coordinates
Add
Convert gene to/from pseudogene

- **Intron creation (to split an exon in two parts)**

In order to split an exon in two parts separated by an intron, a new exon has to be created flanking the exon to split. Then, the coordinates of both involved exons have to be corrected.

- **Extension of a gene annotation (whose automatic annotation is truncated)**

Automatic annotation could miss the detection of exons at the beginning or the end of a given gene. E.g., exon 4 and 5 are not detected and annotation of CDS is terminated at the first stop codon following the exon 3.

Since the exons to add are placed outside the region spanned by the original annotation, it is difficult to add new exons. After determining the correct gene structure (the coordinates of all its elements) the easier way to modify the annotation is to modify first the 'gene' coordinates. This action will reorganize the 'Gene Map' window introducing the place for additional exons outside the region spanned by the original annotation.

- **Merging two or more independent annotations**

Sometimes a poly-exonic gene is not correctly recognized and several independent annotations are created that includes only a portion of the exons of the whole gene. In order to correct this annotation, only one annotation should be arbitrarily retained (the one including the most of exons, for example) whereas the other should be made 'obsolete'. However, it is preferable to correct the retained annotation by the help of exon information before to make obsolete the not retained ones. Cependant, avant d'effectuer cette dernière opération, il est mieux corriger l'annotation retenue à l'aide des coordonnées des autres. The annotation correction will be performed as an extension of a gene annotation.

- **New gene creation**

Sometimes genes are not detected by the automatic annotation pipeline. Undetected genes can be found by tBLASTn analysis (protein vs nucleotide sequence) on the complete genome. Even if in the most of the cases the undetected genes are just remnants or pseudogenes, undetected functional genes could be still detected.

In order to perform a *de novo* annotation of a gene (functional or pseudogene), a new annotation element has to be created. After selecting the approximate region containing the new gene in the graphic representation of ‘artemis’, the shortcut ‘Ctrl+c’ will allow to create a *de novo* gene structure, containing all associated elements (i.e. ‘**gene**’, ‘**CDS**’, ‘**exon**’ and ‘**polypeptide**’).

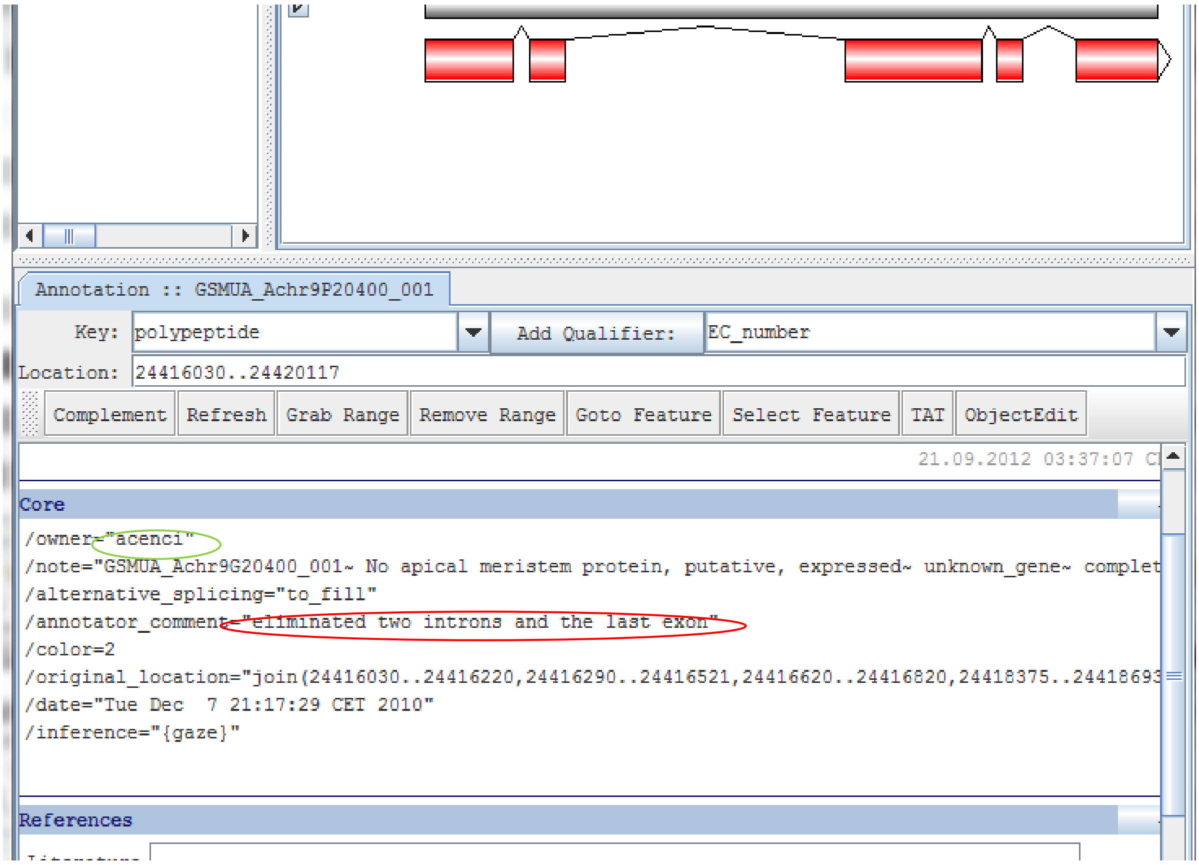
The first step is to provide a new identifier to the gene, according to established criteria (*). Then, the exact coordinates could be inserted and, if necessary (poly-exonic gene), new exons could be added as explained in the above section.

- **Separation of independent genes merged in a chimerical annotation**

Automatic annotation sometimes merges independent genes in a chimerical annotation. In order to correct this error, a new gene structure needs to be created (see new gene creation). Then, one of the merged genes will be re-annotated (see appropriate section) whereas the other one will be corrected in the original annotation by the elimination of the alien exons (see intron or exon elimination).

- **Saving annotations into the database**

To save your modifications into the database, it is required to click on ‘Commit’ in the up-right side of the main Artemis window (the one containing the graphical representation of the gene). However, before performing the commit command, some annotation parameters need to be modified in the ‘polypeptide’ page of the ‘gene builder’, otherwise an error will be signaled by the automatic controller that filters the database modifications.



Annotation :: GSMUA_Achr9P20400_001

Key: polypeptide

Add Qualifier: EC_number

Location: 24416030..24420117

21.09.2012 03:37:07 C

Core

```
/owner="acenci"
/note="GSMUA_Achr9G20400_001~ No apical meristem protein, putative, expressed~ unknown_gene~ complet
/alternative_splicing="to_fill"
/annotator_comment="eliminated two introns and the last exon"
/color=2
/original_location="join(24416030..24416220,24416290..24416521,24416620..24416820,24418375..24418693
/date="Tue Dec  7 21:17:29 CET 2010"
/inference="{gaze}"
```

References

Parameters to be modified:

1) In the ‘**Core**’ section, modify the feature ‘**/annotator_comment="to fill"**’ with a summary of modifications done.
2) In the ‘**Controlled Vocabulary**’ section, modify the following parameters by clicking first onto the ‘**ADD**’ button:
    a) In ‘**CC_functional_completeness**’ select ‘**complete**’ if the gene is completely annotated and, *a priori* functional. On the contrary, select ‘**pseudogene**’ ou ‘**remnant**’ whenever necessary.
    b) In ‘**CC_evidence**’ select ‘**curated**’.
    c) In ‘**CC_evidence_code**’ select ‘**IC**’.
    d) In ‘**CC_status**’ select ‘**finished**’ if the revision of the gene is considered complete or ‘**in progress**’ if additional changes are planned.

Notes: no modification need to be done for the ‘**/owner’** feature. The system will introduce automatically the ‘login’ of the last annotator which modified the gene.