Expert revision of automatic annotations

Alberto Cenci

Protocol for expert annotation with GNPannot/Artemis tools

- **Step 1: Verification of automatic annotation**
  - Is the annotation correct?
  - How to verify?

- **Step 2: Using Artemis to correct the annotation and save the modification**
  - Practical guide to correct the annotation database

Phase 1: Verification of automatic annotation

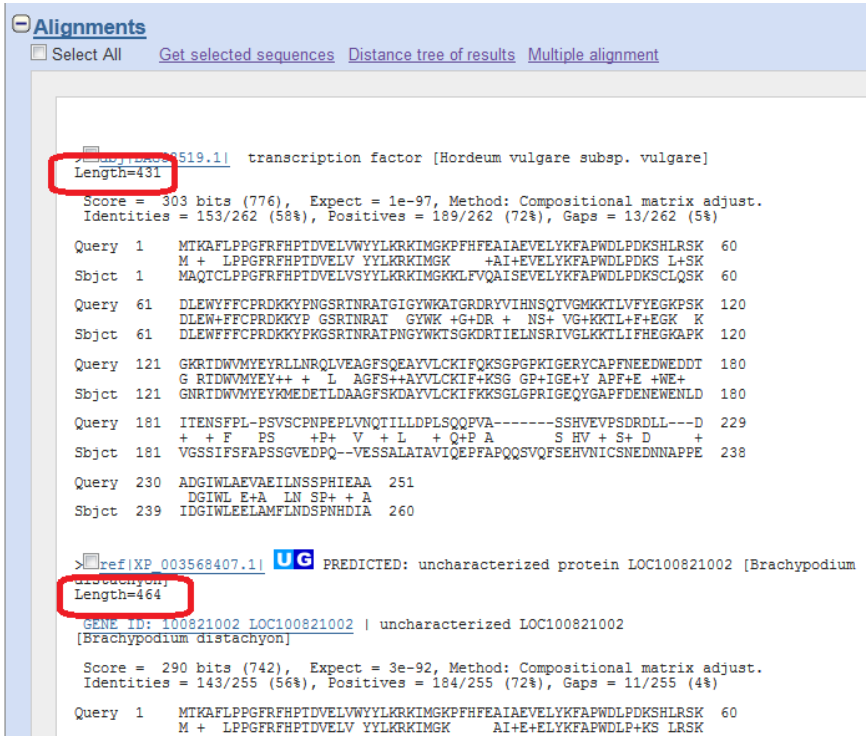
- The amino acid sequence of the analyzed gene (available on the Genome Browser of *Musa acuminata* (GBMa), or on GreenPhyl) is used on a BLASTp query (default parameter) (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins).

- The BLASTp results are analyzed to check that the automatic annotation does not contain major errors (e.g. the length of the sequences found by the BLASTp is clearly longer than the analyzed gene, indicating its possible incompleteness).

- The length of the analyzed gene (Query) is reported in the top of the result page:

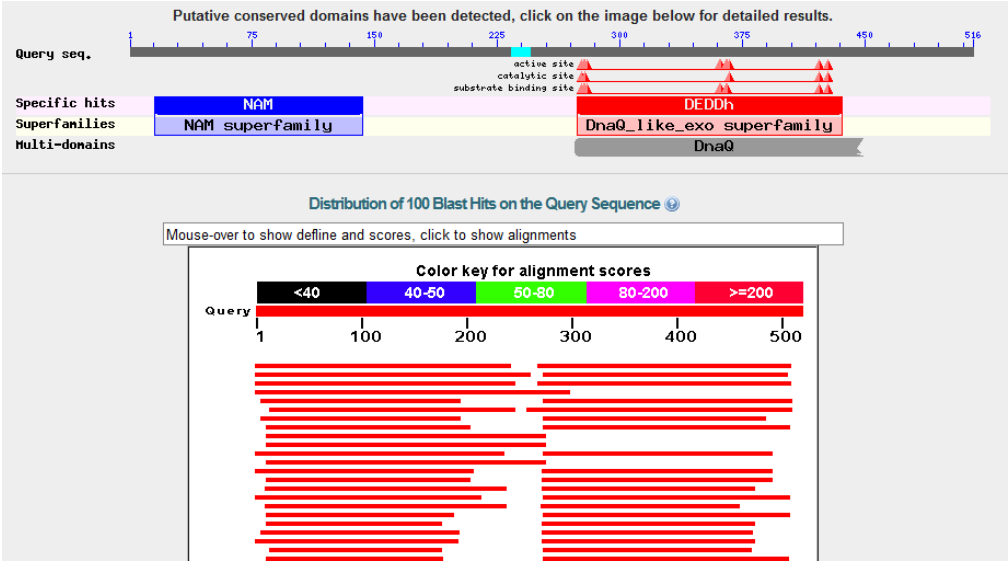
Phase 1: Verification of automatic annotation

- The length of the sequences found by the BLASTp analysis is reported at the top of each sequence alignment:



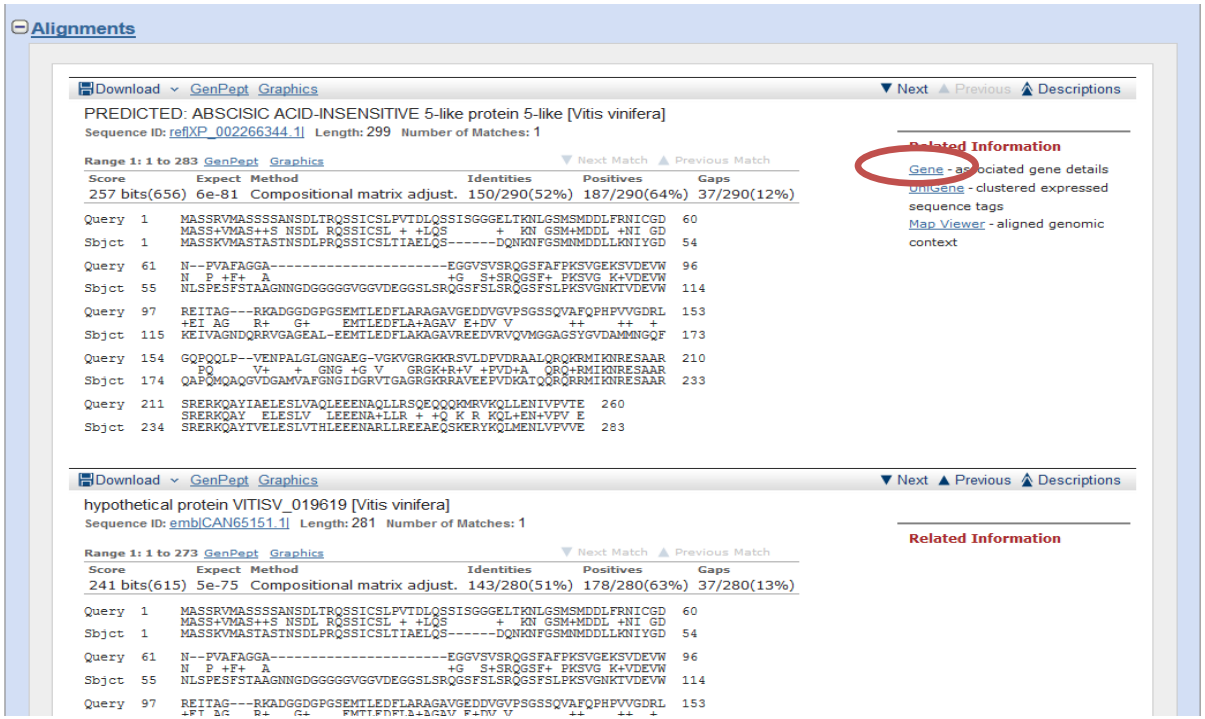
Phase 1: Verification of automatic annotation

- On the other hand, the found sequences could be shorter, indicating that the automatic annotation includes portions that probably do not belong to the gene.
- A particular case is an automatic annotation merging two different consecutive genes (chimerical annotation). This error can be easily detected in the graphical representation of sequence alignments:



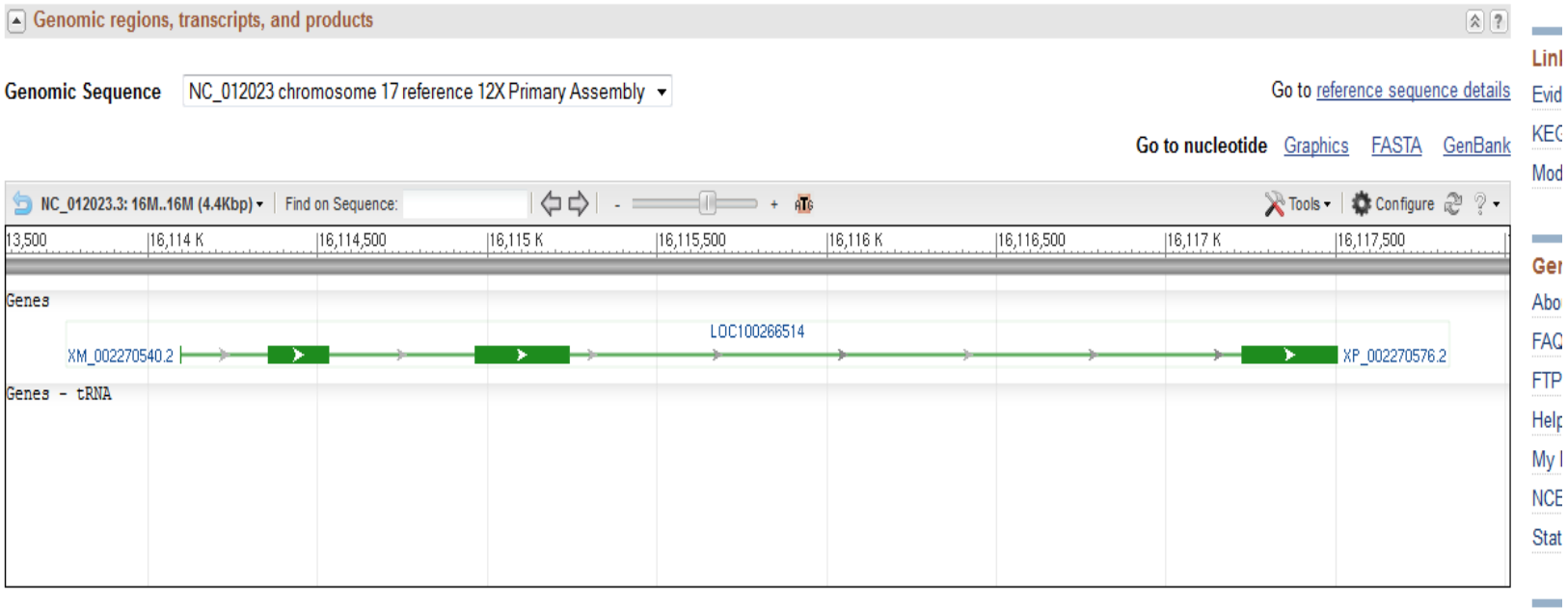
Phase 1: Verification of automatic annotation

- Deeper verifications can be performed by comparing the structure of the analyzed gene and the more similar genes found by BLASTp.
- To access the best BLASTp sequences, open the link "GENE – associated details". I suggest comparisons with the more similar genes of *Vitis vinifera* and *Ricinus communis*, but this is not mandatory.



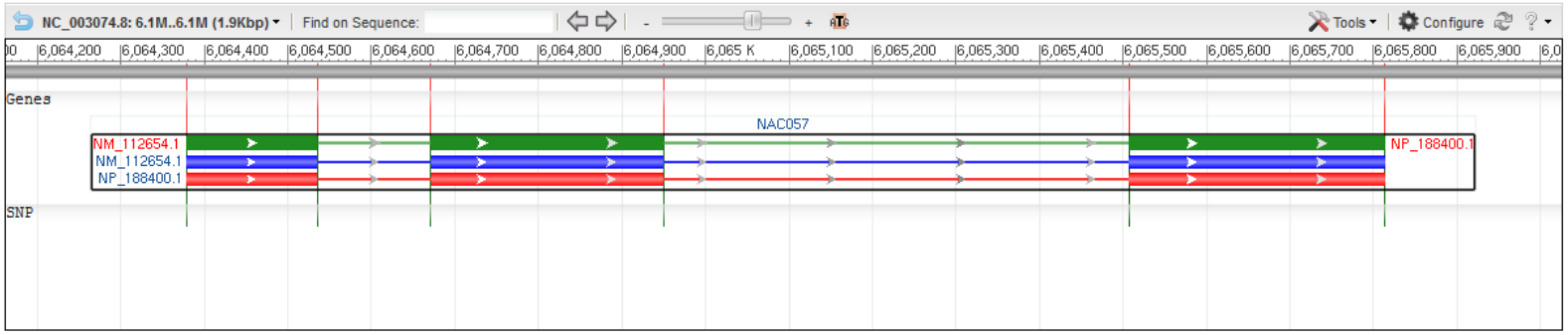
Phase 1: Verification of automatic annotation

- The new web pages contain much information and several links. Among the informative elements, a simplified Genome browser shows the gene structure (number and size of the introns/exons) as annotated in the genome of the selected species.



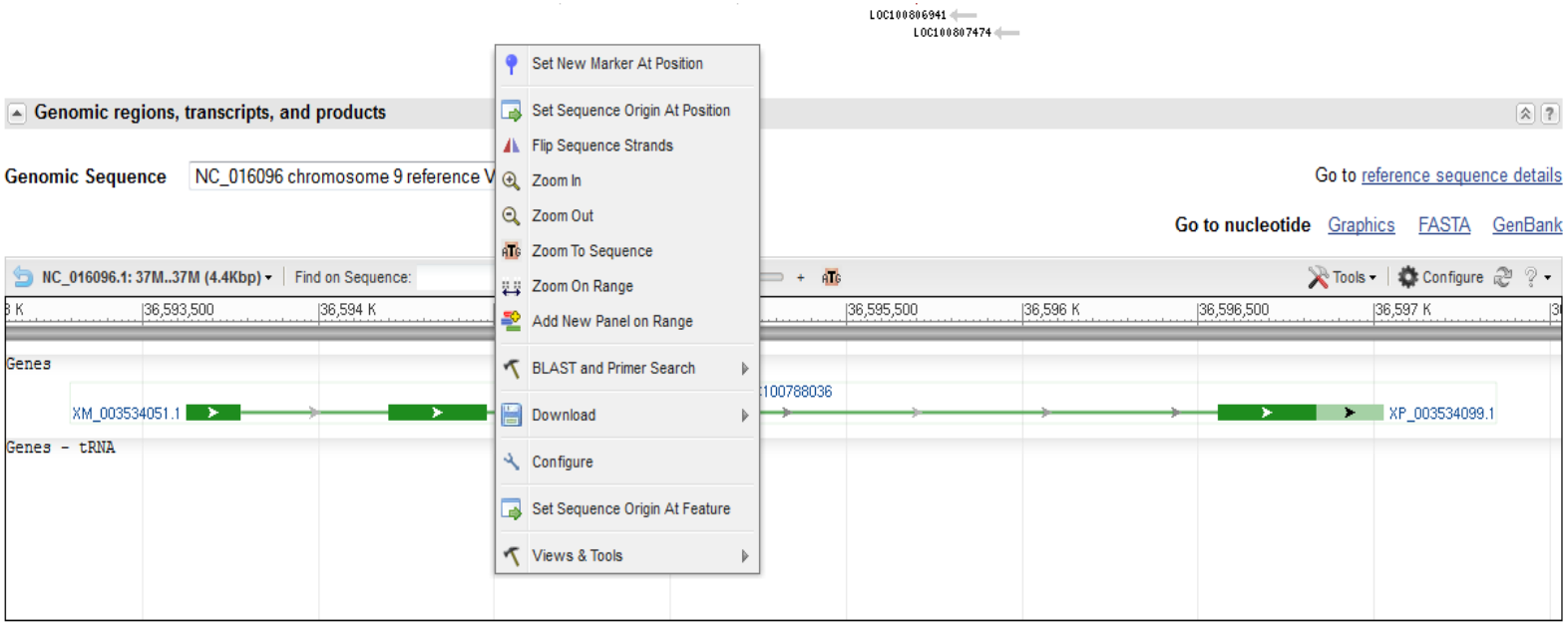
Phase 1: Verification of automatic annotation

- These structures can be compared with that of the analyzed gene, shown in the GBMa.
- 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 more similar genes, it is possible that the automatic annotation needs 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);



Phase 1: Verification of automatic annotation

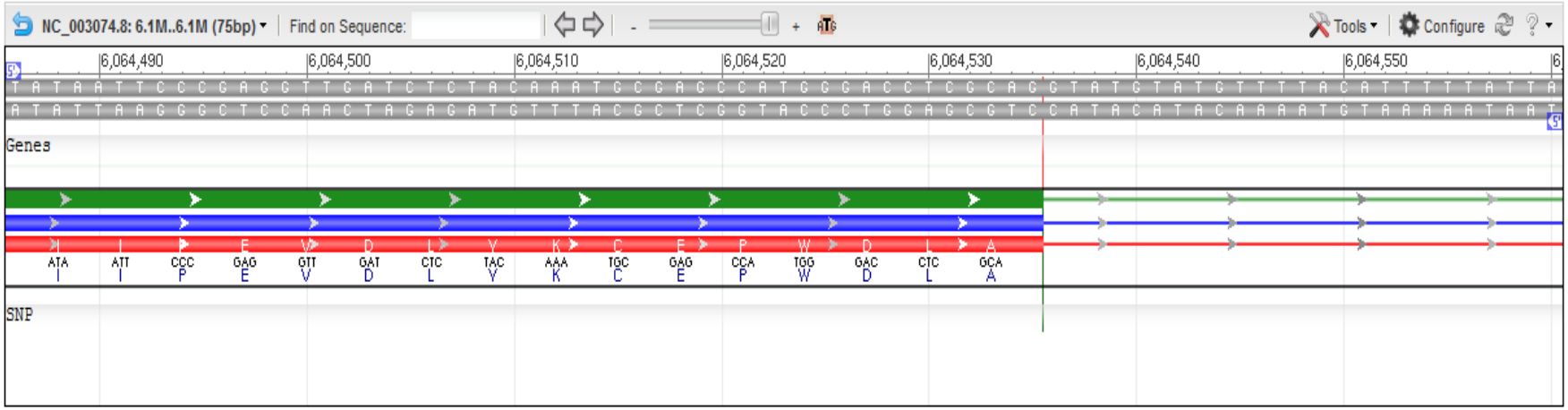
- then click (right mouse button) on the exon end that you want verify and select ‘Zoom to sequence’ in the small menu appearing. These two steps could also be executed in reverse order.



Phase 1: Verification of automatic annotation

- Now it is possible to see the nucleotide and the amino acid sequence of the exon end and compare it to the corresponding one of the Musa gene.

- In 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')).



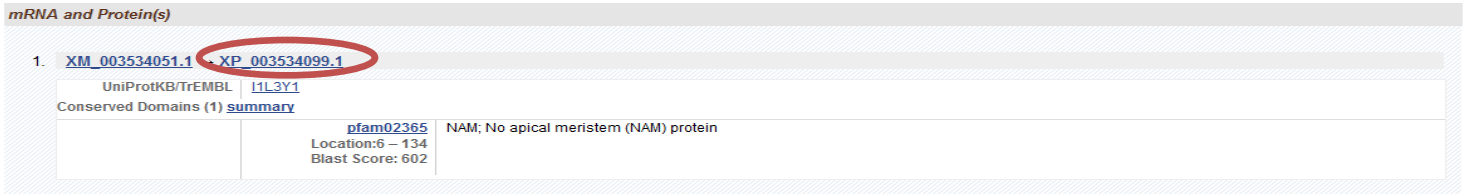
Phase 1: Verification of automatic annotation

- This verification is particularly useful when gaps appeared in the BLASTp alignments or when exon sizes are different, which suggests possible errors in the definition of exon ends.

- In general, when differences are observed between the analyzed gene and one of the most similar genes found by BLASTp, 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 identified and corrections need to be performed in the database containing the annotation.

Phase 1: Verification of automatic annotation

- Summary of errors in structural annotation:
  - Arbitrary intron annotations inside an exon (exon results cut in two or more parts separated by erroneous introns). This error can be detected by the presence of gaps in the Musa sequence in the alignments with the most similar sequences found with BLASTp. This is 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 intron(s) not containing stop codons. The correction consists in the elimination of the erroneous intron(s).
  - Lack of one or more exons in the automatic annotation. This error can be detected by the presence of gaps in the sequence alignments between the analyzed gene and the most similar sequences found with BLASTp or by comparing their gene structures. In order to find the lacking portion of the gene, one of the most similar amino acid sequences (found with BLASTp) can 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 program). The amino acid sequence of a similar sequence can be obtained by clicking in the 'mRNA and Protein(s)' section in the 'GeneID' page.



Phase 1: Verification of automatic annotation

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

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

- 'Artemis' allows the user to handle sequence annotations and to modify them; however, in order to save annotation modifications, a specific protocol needs to be followed.
- Each gene annotation on 'artemis' is composed of four main elements that can be modified. These elements can be visualized in the 'Gene builder' window (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' and 3' untranslated regions (UTR). For poly-exonic genes, the coordinates are composed of the ends of each exon. Its global 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 coincide with the beginning and the end of the translated region (also for poly-exonic genes).
  - At the same level of the 'polypeptide' element, if available, the '**five prime UTR**' and the '**three prime UTR**' elements can be present, with their coordinates at the beginning of 'gene' and the position before the 'ATG' (5' UTR) and at the position following the 'stop codon' and the end of the 'gene' (3' UTR). These elements are calculated automatically by artemis.

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

Structural annotation rules

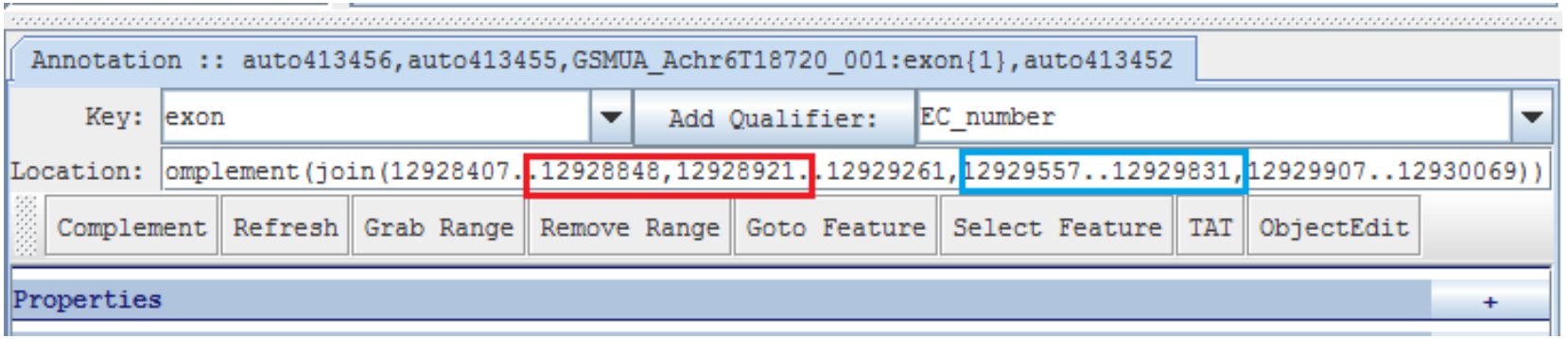
- The structural annotation of a given element is, by convention, marked by its first and its last position in the sequence, separated by two full stops: e.g. '**12928407..12928848**'.

- When an element is composed of more two or more sub-elements (in the case of poly-exonic genes, the element 'exon' is composed of more than one element (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))**'.

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

- Modification of existing elements
- The coordinates of all existing elements can be modified in the 'Location' section or in the window containing the graphic representation of the annotation, by dragging with the mouse the ends of the element to modify.
-
- Intron or exon elimination
- In order to eliminate an intron (i.e. merge two exons), the end position of the first exon and the first position of the other exon are eliminated in the 'Location' section of the 'exon' element along with two full stops (red rectangle in the following figure). In order to eliminate an exon, its coordinates are eliminated with one flanking comma (blue rectangle in the following figure).

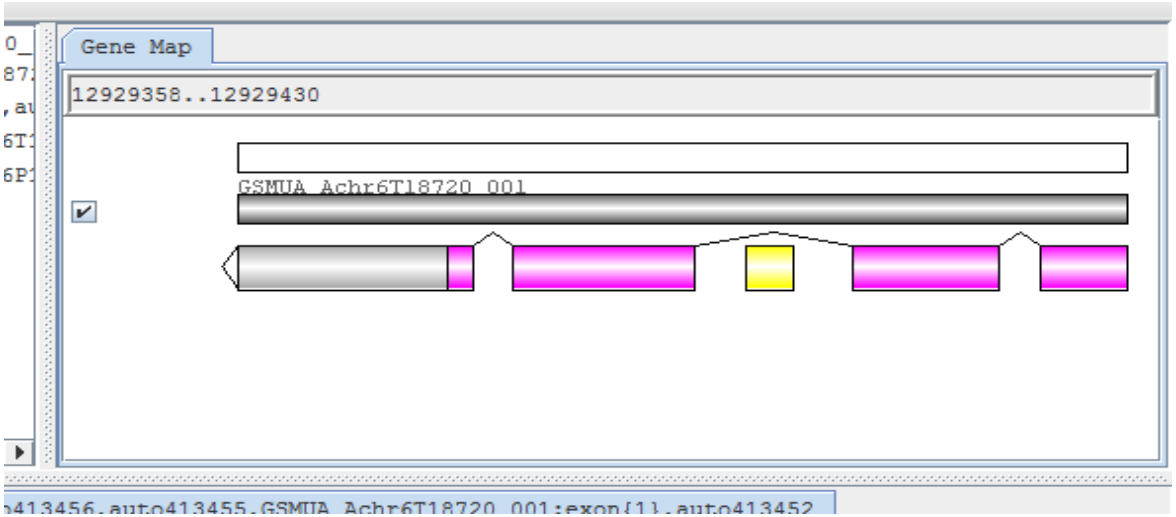


Be careful: each element must be eliminated individually! E.g. if two consecutive introns need to be eliminated, remove one, click «Apply» and then remove the second one.

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

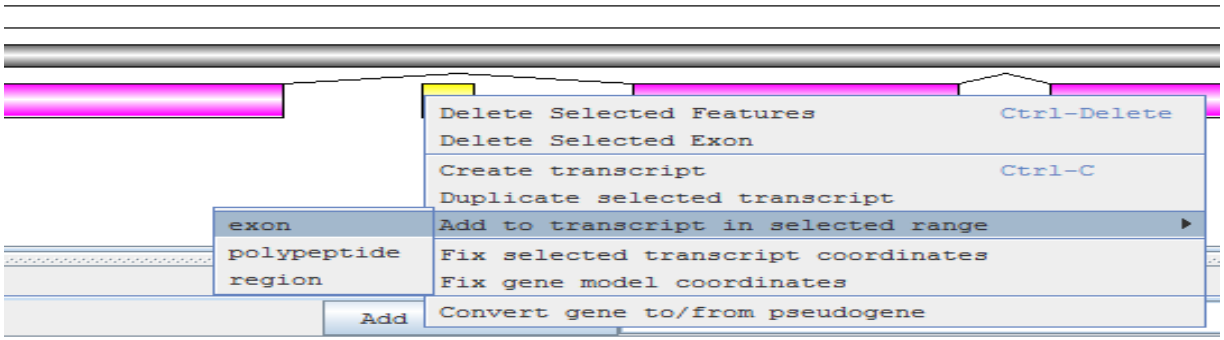
Exon creation

- In order to create an exon, DO NOT introduce its coordinates in the 'Location' section of the 'exon' element, because this modification will not be taken in account during the modification save in the Chado database. 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 upper-right side of the 'gene builder' window. Using the mouse (click the left button), place the next exon in the approximate position (it will appear as a yellow rectangle in the 'Gene Map').



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

- Then, click with the right button in the rectangle to call a pull down menu and select ‘Add to transcript in selected range’ > ‘exon’ (see the 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 in the ‘exon’ element.



Intron creation (to split an exon into two parts)

In order to split an exon into 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.

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

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 easiest way to modify the annotation is to modify first the ‘gene’ element coordinates. This action will reorganize the ‘Gene Map’ window, introducing the place for additional exons outside the region spanned by the original annotation.

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

Extension of a gene annotation (adding UTRs).

- When transcriptomic data are available, it is possible to have information on extent and structure of untranslated regions. Even if these regions are not coding, they are transcribed in the mRNA. To add UTRs, the ends of the ‘gene’, ‘mRNA’ and ‘exon’ elements (but not of the ‘polypeptide’ one) need to be replaced with the ends of the complete transcript.

- Introns can be found also in UTRs; in this case, an exon has to be created in the ‘exon’ element as above.

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

Merging two or more independent annotations

- Sometimes a poly-exonic gene is not correctly recognized and several independent annotations are created that include 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 the exons, for example) whereas the other annotation should be made ‘obsolete’.
- It is preferable to correct the retained annotation with the help of exon information before making the other one obsolete. The annotation correction (addition of exons) needs to be performed as explained for the extension of a gene annotation.

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

Creation of a new gene

- 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 most 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 functional gene (or a 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 the user to create a *de novo* annotation, containing all its elements (i.e. '**gene**', '**CDS**', '**exon**' and '**polypeptide**'). The first step is to provide a new identifying name to the gene, according to the established criteria (*). Then, the exact coordinates can be inserted and, if necessary (poly-exonic gene), new exons can be added 'one by one'), as explained in the previous section.

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

Separation of independent genes merged in a chimerical annotation

- Automatic annotation can also merge independent genes in a chimerical annotation. In order to correct this error, a new annotation needs to be created where one of the merged genes will be re-annotated. The other one will be corrected in the original annotation by the elimination of the alien exons.

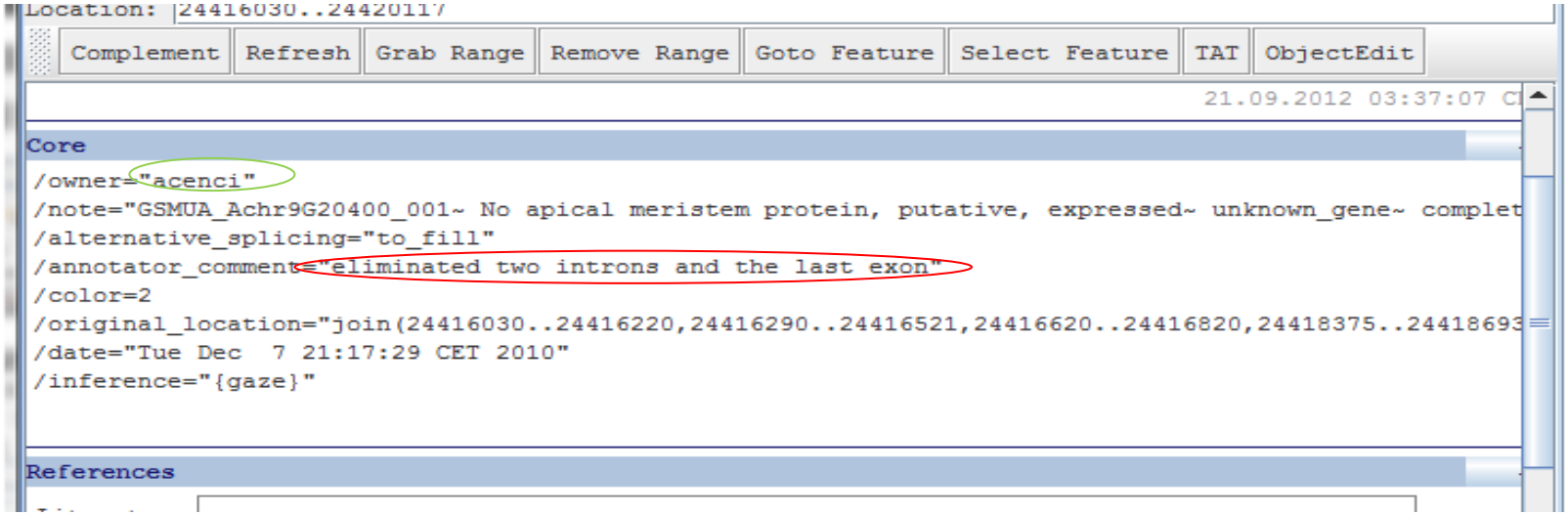
Phase 2: Using Artemis to correct the annotation and save the modification

Saving annotation modifications in the Chado database

- In order to save the corrections performed on the automatic annotations in the 'Chado' database, it is necessary to click on 'Commit' in the upper-right side of the 'artemis' main page (the one containing the graphical representation of the gene). However, before giving 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.

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

1) In the '**Core**' section, modify the feature '**/annotator_comment="to fill"**' with a synthetic comment on the performed modifications. Conversely, no modifications need to be performed at the '**/owner**' feature. The saving system will automatically introduce the 'login' of the last annotator which modified the gene.



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

- 1) In the '**Controlled Vocabulary**' section, modify the following parameters by clicking on the '**ADD**' button:
  - In '**CC_functional_completeness**' select '**complete**' if the revised gene is completely annotated and, a priori, functional. On the contrary select '**pseudogene**' or '**remnant**'.
  - In '**CC_evidence**' select '**curated**' (and remove '**automatic**' in the extant list).
  - In '**CC_evidence_code**' select '**IC**' (and remove '**ISS**' in the extant list).
  - In '**CC_status**' select '**finished**' if the revision of the gene is considered complete or leave '**in_progress**' if additional changes are planned.

Contact:

e.mail: a.cenci@cgiar.org

Skype: **Alberto Cenci**

Guided exercise

http://www.gnpannot.org/content/gnpannot-sandbox-form-2-without-access-restriction