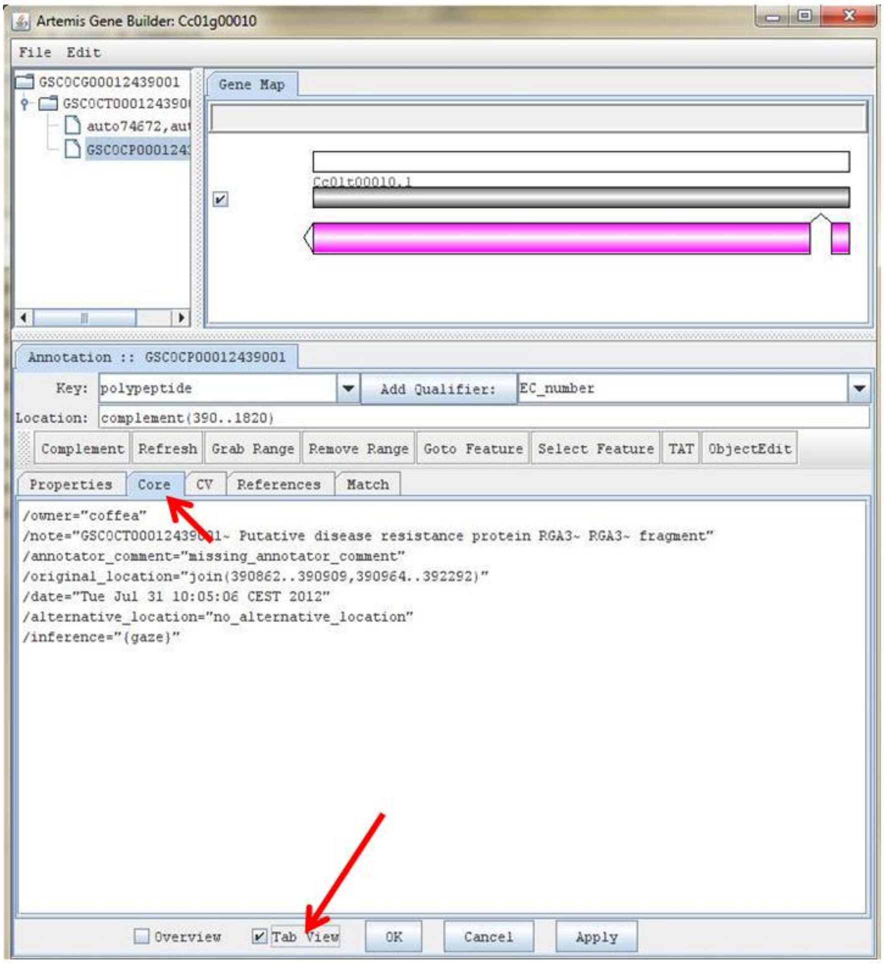
Functional annotation rules

This document is a guide for the manual annotation of the *M. acuminata* genome. It aims at defining some rules and procedures to standardize the modifications and enhancements of the genome automatic annotation that has been performed.

Before initiating your work, you must check that you have the rights to edit the annotation with Artemis as an expert annotator.

The manual annotation consists of editing different qualifiers associated to the gene/protein in the Chado database. To do that, you have to select the gene to annotate in Artemis, select "polypeptide" and type Ctrl+E to edit a functional annotation.
Check "tab view".



Key: polypeptide

Add Qualifier: EC_number

Location: complement(390..1820)

```
/owner="coffea"
/note="GSC0CT00012439001~ Putative disease resistance protein RGA3~ RGA3~ fragment"
/annotator_comment="missing_annotator_comment"
/original_location="join(390862..390909,390964..392292)"
/date="Tue Jul 31 10:05:06 CEST 2012"
/alternative_location="no_alternative_location"
/inference="(gaze)"
```

Then you will be able to modify the qualifiers in different tabs.

There is no way to erase the modifications, therefore keep the original information as a screenshot (protein structure, intron- exon limits, functional annotation…) and then keep the modified structure / functional annotation as another screenshot for your record. Insert your screenshots in a word file. Copy - paste PMID, inference and product in the same word file.

After every change, don't forget to click on "commit" and wait for the inspector comments.

Here is the list of qualifiers to fill, some qualifiers are mandatory others are optional:

- **Core**
  - Inference **(mandatory)**. By default, it corresponds to Gaze. You can add new inferences.
    It must be written following the rule: "Database_name:identifier"
    ex: gaze:GSCOCT00026571001, SwissProt:Q9LRM7
  - Annotator_comment (**optional**)
    ex: The two first exons have been merged due to similarity results (Scov < 0.8)
- **CV (Controlled Vocabulary)**
  Note: This part of annotation is based on Controlled Vocabulary. Some CV terms might be already assigned by automatic functional annotation and can be observed by passing the mouse on them. You can add easily new CV terms by clicking "Add". Previous ones need then to be deleted from the table.
  - Product (**mandatory**). To add it, click on "Add" button and see the different *product* options already available. If you don't find your product (named similarly as in SWISSPROT) you can add a new term by clicking "add term" in the CV term selector, and then adding your new term in the corresponding area (and an optional definition).
    ex: Dehydration-responsive element-binding protein
    NB: When identified by similarity in another genus, don't be too specific (i.e. prefer DREB to DREB1D…) because of potential co-orthology.
  - CC_Evidence_code (**mandatory**). To add them, click on "Add" button and see the different *CC_evidence_code* options
    ex: IC_2a (NB: IC means "inferred by curator". You will need then to delete "ISS" (inferred from sequence homology) from the table
  - CC_Functional_completness **(mandatory)**. To add them, click on "Add" button and see the different *CC_functional_completeness* options
    ex: complete / partial…
  - *CC_Gene* ***(mandatory if Evidence_Code is 1 or 2)***
    ex: NCED
  - CC_Ec_number (**mandatory if an enzyme**)
    ex: 3.1.3.16
  - Status (**optional**)
    ex: in_progress / finished …
- **References**

- PMID (PubMed ID) (mandatory if Evidence_Code is 1 or 2).
  Note: PMID must be added in the section "Dbxref" and not in "Litterature" section.
  PMID can be searched and found through the SwissProt database (on SwissProt entry, you can get the PMID by right clicking on it) or directly from PubMed.
  ex: PMID:23483889

When you're done (and have committed your modifications), you can check on the Genome browser the track named "Track modified by annotators (mRNA)". Through the link "detailed report", you should find all modifications made on that gene.

==

How to evaluate the value to assign to the qualifier "Evidence Code" to a gene for functional annotation?

Here is a general classification:

1: The function of the cognate polypeptide was validated from experiment

2a: The function of the predicted orthologous polypeptide was validated from experiment

2b: The function of the predicted orthologous polypeptide was infered from computational analyses

3: The function of the similar polypeptide was infered from computational analyses

4: The function of the similar polypeptide is hypothetical

5: No significant blast hit but a high ab-initio score

6: No significant blast hit & low ab-initio score

7: Very partial match & strong anomalies of the gene structure

The table presented below indicates more precisely the different conditions to meet to assign an "Evidence Code" value.

Conditions	*Gene*	PMID	Product	GO	Code	%id	Cov	
							Q	S
Similarity with a polypeptide whose function has been experimentally demonstrated *in the studied organism* OR *in the same genus*	*Gene*, syn	yes	Description of the cognate polypeptide	yes	**1**	>=90%	>=0.95	
High similarity with a polypeptide of known function								
Experimentally demonstrated in an other genus	*Gene*, syn	yes	Description of the orthologous polypeptide	yes	**2a**	>=45%	>=0.8	
Strong orthologous gene (without experiment)		yes			**2b**	>=50%	>=0.8	
In case of partial match…								
If LengthQuery > LengthSubject*	no	yes	Desc. of the different modules (PROTEIN1-PROTEIN2 modules)	yes	**2a or 2b**	>= 45-50%	**<0.8**	>=0.8
If LengthQuery < LengthSubject**	no	yes	Desc. of the similar polypeptide (PROTEIN fragment) (N-terminal fragment)				>=0.8	**<0.8**
Similarity with			Putative …					
Swissprot/TrEMBL polypeptide	no	no	desc. of the similar polypeptide (PROTEIN)	yes	**3**	>=25%	>=0.8	
InterPro family	no	no	desc. ot the similar family (without "domain|superfamily")	yes				
Similarity with polypeptide of unknown function OR interspecies EST	no	no	Conserved hypothetical protein	ND	**4**	>=25%	>=0.8	
SignalPHMM result***	no	no	CHP; putative exported protein	ND				
TMHMM results (> to 3 results)	no	no	CHP; putative membrane protein	ND				
InterPro domain	no	no	CHP; putative domain desc.	ND				
No significant blast hit	no	no	Hypothetical protein	ND	**5**	<25%	<0.8	
InterPro domain, TMHMM, SignalPHMM	no	no	HP; putative domain desc., membrane or exported protein	ND				
Doubtful CDS****	no	no	HP (one short exon, weak Pc, CDS overlap, change of coding strand, atypical codon usage)	ND	**6**			
Very partial match and strong anomalies of the gene structure*** **	no	no	desc. of the similar polypeptide (remnant) (one small fragment or more than 3 fragments)	ND	**7**	>=35%	>=0.8	<0.5

*: Check if the true start codon is not further downstream
**: Check for a possible gene fission, check if the true start codon is not further upstream (type fragment)
***: Check if start codon is correct
****: short CDS, weak coding probability, CDS overlap, change of coding strand, atypical synonymous codon usage
****: Gene with no CDS (type fragment or pseudogene)

Glossary:

Gene: gene symbol or a synonymous gene symbol
GO: Gene Ontology
%id: identity percentage
Qcov: query covering (match/query ratio)
Scov: subject covering (match/subject ratio)
PROTEIN: protein symbol
PMID: PubMed ID
ND: Not determined
CC: Chado Controler

How to get Scov and Qcov?

It can be retrieved by looking at “Match” tab in Artemis, and by asking for details.
In this section, you will get % id, Scov and Qcov.