SOP Re-Annotation

http://ann-nblrrome.tuebingen.mpg.de

Warnings

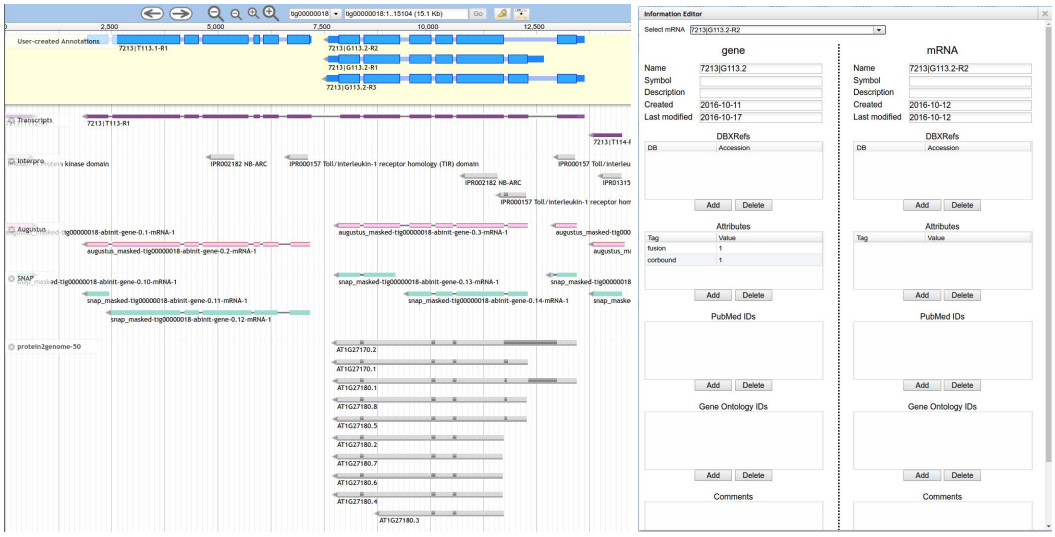
WebApollo does allow duplicate names!

Comments

We only re-annotate genes with NB- or TIR- domains

Normal Re-annotation

1. Check if regions with evidence (prefer ESTs > Proteins > Gene Predictions) have a gene model (if not add it)
2. Check the gene model (exon/intron structure) against protein and EST evidence (always compare SNAP and Augustus in case of likely fusions)
3. Re-annotate if necessary
    a. Add UTR information from est2genome track if it can be easily incorporated into the gene model
    b. Compare exon-intron structures from transcript and protein evidence to maker annotation and adjust where needed
4. Rename to original transcript annotation of MAKER (example: gene name=6909|G040, transcript name=6909|T040-R1)
5. If several transcripts are annotated, use -R1, -R2, etc
   Example: Gene 7213|G113
   Additionally to being split (see ‘Gene Fusion Handling’), three transcripts are annotated: 7213|T113.2-R1, 7213|T113.2-R2 and 7213|T113.2-R3



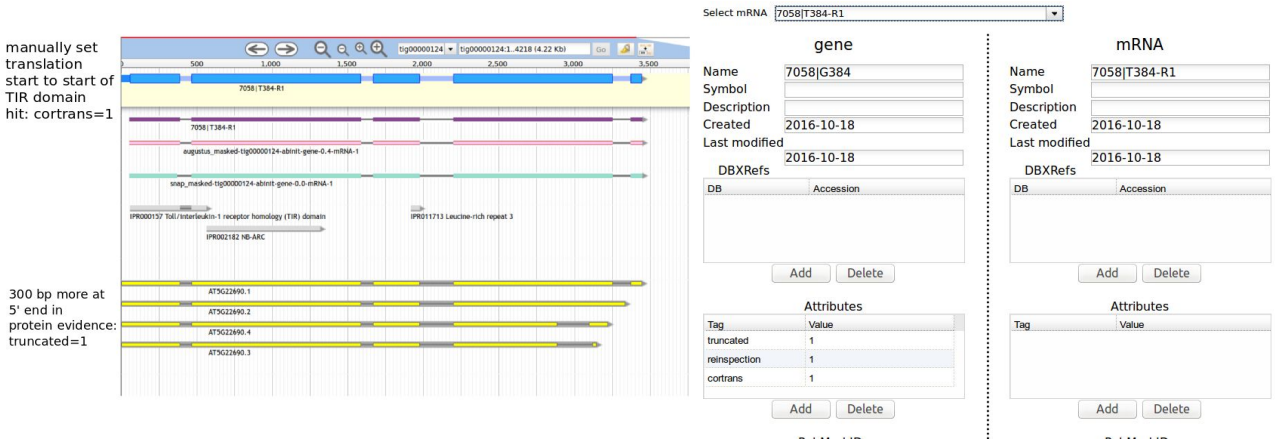
6. Set appropriate tag-value pairs (see ‘Attribute Settings’)

Truncated Genes

1. Check NLR genes at the beginning or end of contigs (especially if the distance to the contig border is < 500bp)
2. Check if available evidence is longer than annotated (causes alignment overhangs)
3. If possible, correct the annotation using the evidence
    a. If you correct exon-/intron- boundaries, set attribute corbound=1
    b. If you correct translation start or end, set attribute cortrans=1

    (!) Be careful with those adjustments and do only when evidence is bulletproof.
4. Add attribute truncated=1

Example: Gene 7058|G384



Splitting Gene Fusions

1. Split gene according to evidence
2. Rename the new genes:
   General Format: ACC|[G/T]*XXX*.*Y*-RZ
    - ACC=Accession ID
    - [G/T]*XXX*=Gene/Transcript ID
    - .*Y*=Split Gene/Transcript Sub-ID
    - -RZ=Isoform ID

   Example: Gene 6909|G040
    - Original gene name 6909|G040 → 6909|G040.1 and 6909|G040.2 (the leftmost gene gets appendix '.1', then '.2', '.3' etc follow while moving right)
    - Original transcript name 6909|T040-R1 → 6909|T040.1-R1 and 6909|T040.2-R1
3. Add attribute fusion=1

Genes Without Evidence

- This situation is common for non-reference datasets. You might observe many predictions without EST or Protein evidences. Besides using BLAST, the Araport JBrowser (https://apps.araport.org/jbrowse/?data=arabidopsis) is handy to verify/compare gene architectures. In some cases these instances are pseudogenes, and so the Araport information is useful. Sequences, or gene identifiers can be searched directly by using the field left to the Go button in the top of the browser.

- If you find a pseudogene (based on Araport annotation and synteny), you can flag it with pseudogene=ATG_identifier.
- The noevidence=1 parameter is only set if we choose to reannotate something without evidence (evidence here is everything that gives you a hint for this reannotation to be necessary, so no evidence at all will probably nearly never be the case)

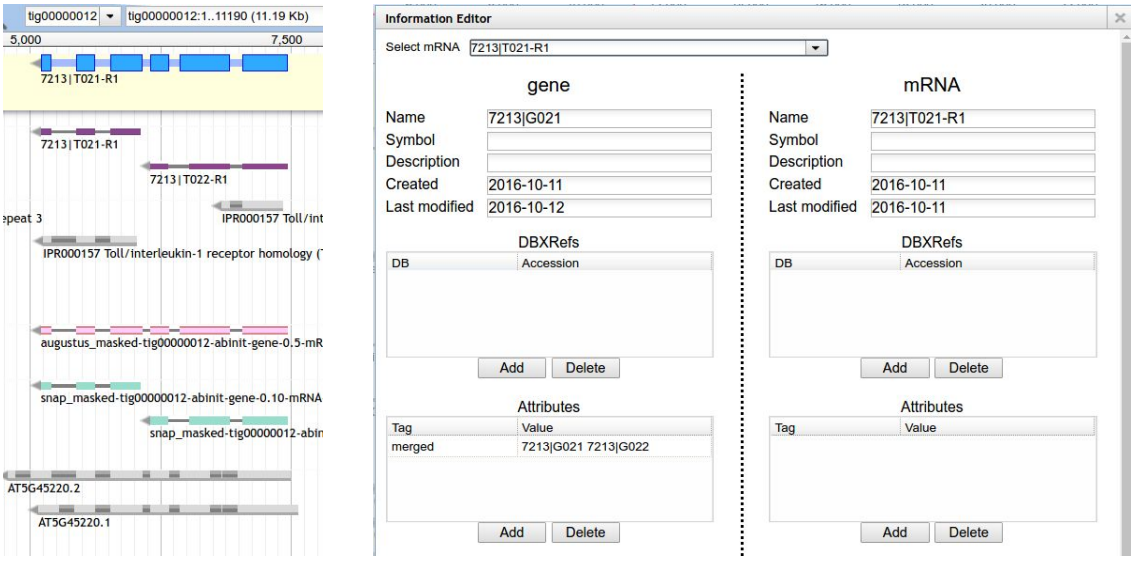
Merging Split Genes

Be very careful when merging genes. It is worse for orthogroup predictions to have a wrongly merged gene, than to have two erroneously split genes.

1. Merge
2. Rename the merged gene according to the 'leftmost' gene that is contained
3. Add the tag 'merged' with all the gene names that are merged. Separate by space (do not use comma)

Example: Genes 7213|G021 and 7213|G022

- New gene name: 7213|G021
- New transcript name: 7213|T021-R1
- add tag: merged='7213|G021 7213|G022' (Do not use comma as separator)



Example: Merging a gene with another one that had been flagged as part of a fusion

- Original gene names: 7058|G069 and 7058|G070
- After splitting: 7058|G069.1 (fusion=1), 7058|G069.2 (fusion=1), and 7058|G070
- After merging: 7058|G069.1 (fusion=1) and 7058|G069.2 (fusion=1; merged=7058|G069.2 7058|G070)

Annotating New Genes

In case we want to add an additional new annotation.

1. Name the new gene according to the gene to its left, and add ".N<number>"
   Example: New gene next to 1925|G530
   - Gene Name 1925|G530.N1
   - mRNA Name 1925|T530.N1-R1
-

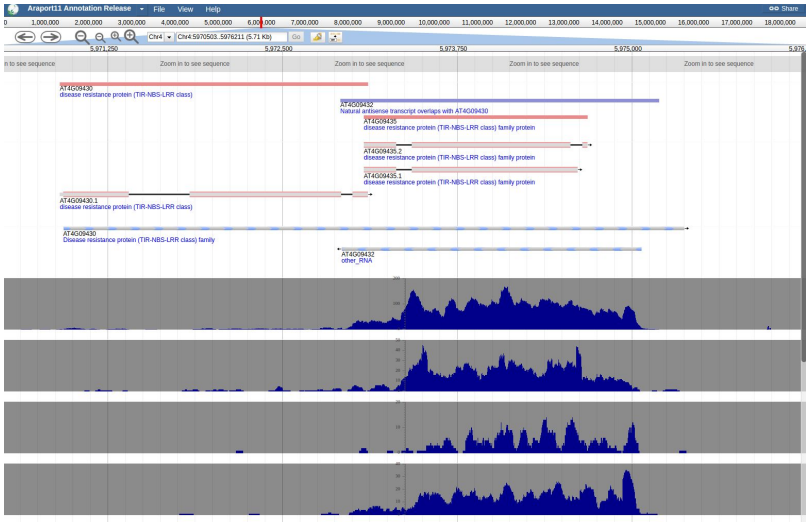
Attribute Settings

Tag	Value	
reinspection	1	set if reinspection of the gene model is needed
fusion	1	set for split genes. Each gets the fusion tag.
truncated	1	set if a gene seems to be truncated
pair	GeneID of “partner”	set if evidence is in “Col-0 Pairs” from pairs_and_putpairs list
putpair	gene ID of “partner” gene	set if head-to-head orientation is given and evidence is not in the “Col-0 Pairs” from pairs_and_putpairs list
pseudogene	Araport11 Identifier	set if there is pseudogene evidence from Col-0
noevidence	1	set if the reannotation was done without any source of evidence (nearly never the case)
merged	GeneID<space>GeneID(<space>GeneID...)	set for merged gene
corbound	1	set if exon-/intron- boundaries were changed without direct evidence
cortrans	1	set if translation start or end was set without direct evidence
misassembly	1	set if a misassembled contig is suspected
delete	1	use if a gene model should not be replaced, but deleted completely
mod	1	use if gene model was extensively changed (mostly without evidence from gene predictors or transcript/protein mappings) in order to rescue the domain structure. Genes that were re-annotated this way probably need to be excluded from some analyses.

Further Remarks:

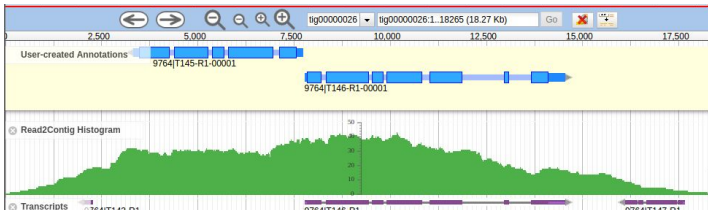
- We found a misannotation in Araport11. They split the tair10 gene AT4G09430 into AT4G09430 and AT4G09435, both annotated as TNLs. This is wrong. The first gene contains the NB- and the TIR- domain, the second one only has a LRR annotated. The misannotation is driven by a natural antisense gene AT4G09432 that overlaps AT4G09430. This antisense gene is expressed, whereas AT4G09430 might not be expressed (at least not in their data). Araport11 mistreats the expression as belonging to the TNL, and splits the gene. Long story short, don’t split a gene if your evidence is from AT4G09430 and/or AT4G09435 :)
  For a visual reference check: 6909|G370

- Be careful if you have to re-annotate two overlapping genes. They always get treated as two isoforms from the same gene. We agreed on removing UTR that is overlapping.



- Change default view of **Read2Contig Histogram** track.

change *"bicolor_pivot": "mean"* to *"bicolor_pivot": "0"*



- Noncanonical splice sites:
  found e.g. in AT5G47280, AT4G33300