The PN40024.v4 *Vitis vinifera* Genome

An updated reference assembly and annotation

Tutorial aim: The aim of this tutorial is to guide users through the process of manual editing using the range of functionalities Apollo provides to update the PN40024.v4.1 annotation which will accompany the new PN40024.v4 genome assembly on *Vitis vinifera* cv. PN40024. Manual curation is an ongoing task so the idea would be that the grapevine community continues to use Apollo beyond this workshop.

Introduction

The PN40024.v4 genome assembly is built upon the previous 12X.v2 PN40024 assembly. The strategy aimed to improve the previous assembly without discarding the existing Sanger assembly which is still a high quality sequence in terms of base pair accuracy. The new assembly includes new PacBio and Illumina reads which allow the filling of gaps, improvement of existing scaffolds and sequence correction of newly incorporated PacBio regions. The update shows promising improvements including a significant increase in the average size of scaffolds (235,000 to 741,572 bp), a significant decrease in the number of Ns (~16 to ~2 million bp) and an improved assignment of unknown contigs to chromosomes (~27 to ~12 million bp). The new assembly supports better alignment of RNA-Seq reads across a range of cultivars. In addition, heterozygous regions have been delimited and provided with the assembly as smaller alternative chromosomes containing only those regions. One interesting finding is that the PN40024 cultivar was obtained through self-crossing of cv. Helfensteiner and not cv. Pinot Noir as previously thought. Cultivar Helfensteiner was initially created in 1931 from the crossing cv. Pinot Précoce Noir (early ripening PN) and cv. Schiava Grossa.

The PN40024.v4 annotation was created with the aim of obtaining a more reasonable number of protein-coding genes (35,197 as opposed to 41,733) with respect to VCOST.v3 since this version contained many extra protein-coding genes which are unlikely to be real. In an effort to preserve VCOST.v3 gene IDs where possible (i.e. in cases where chromosome numbers are unchanged) a reciprocal best-hits strategy between VCOST.v3 and PN40024.v4.1 annotations was carried out. This allowed for the transfer of 67.27% of VCOST.v3 gene IDs. In general, the new *de novo* annotation performed better than the previous annotation in terms of BUSCO stats.

Motives for a curated manual annotation

In silico strategies for genome annotation have their limitations especially with regards to the actual expression of individual transcripts, which can be shown by transcriptomic data (i.e. RNA-Seq). Although the new annotation is already an improvement with respect to the previous version, some gene families such as those derived from recent gene expansions have not been well annotated. For a better informed annotation, several data sources have been integrated within Apollo to help the user choose the best gene model for a particular transcript. A summary of the tracks are outlined below:

1. **PN40024.v4.1 annotation** (v4 *de novo* annotation)
    a. **REF**
    b. **ALT** (annotation corresponding to alternative heterozygous regions only)
2. **Previous annotation versions** (Transfer of previous gene model coordinates to the new assembly)
    a. **CRIBI V1**
    b. **VCOSTv3**
3. **Gene catalogue**
    a. **Gene catalogue** (Integrape v1.1 catalogue, P450s, RGAs and more)

4. **RNAseq**
    a. **Mixed samples** (12 RNA-Seq studies from a range of tissues)
    b. **Mixed samples - insert size <= 300 - MQ >= 40** (filtered by intron length and read quality)
    c. **Mixed samples with multimapping**
    d. **PN40024 RNAseq on REF only** (PN40024 mixed tissues)
    e. **CS IsoSeq max introns 20kb** (cv. Cabernet Sauvignon showing full transcripts in fruit)
5. **Separate RNAseq** (11 runs from **mixed samples**; the remaining run was already a blend of tissues)
    a. **RCDN1_S1** (Vitis sylvestris C1-2, green berries)
    b. **RCDN5_S2** (Vitis sylvestris C1-2, ripening berries)
    c. **SRR1502882** (cv. Carignan, leaves)
    d. **SRR2015361** (cv. Riesling, Shoot)
    e. **SRR4447140** (cv. Jingxiangyu, leaves)
    f. **SRR5435950** (cv. Pinot Noir, leaves)
    g. **SRR6195043** (cv. Vlaska, petioles)
    h. **SRR6365708** (cv. Bangalore blue, leaves)
    i. **SRR7192360** (cv. Cabernet Franc, leaves)
    j. **SRR7451534** (cv. Victoria, leaves)
    k. **SRR7768546** (cv. Rosario Bianco, buds)
6. **Orthologs** (Transferred gene models from different databases)
    a. **Arabidopsis**
    b. **ORTHODB** (orthologous protein-coding genes across vertebrates, arthropods, fungi, plants, and bacteria.)
    c. **SWISSPROT** (protein database)
    d. **Vitales**
7. **SNPs**
    a. **PN40024 Illumina SNPs** (SNPs identified by Illumina reads)
8. **Others** (Useful during annotation as a gene that crosses a scaffold or an N gap is problematic)
    a. **Heterozygous region**
    b. **Scaffolds**
    c. **N gaps**
    d. **Repeat regions**

Apollo: A tool for multi-user genome annotation

Apollo is a browser-based tool for visualisation and manual curation of sequence annotations. It is especially suited for collaborative efforts since every user gets real-time updates of ongoing editing. In short, it is a modified genome browser (JBrowse) with an extra track for user-created annotations which eventually may be exported in the standard “GFF3” annotation format. It is the Google Drive of annotation!

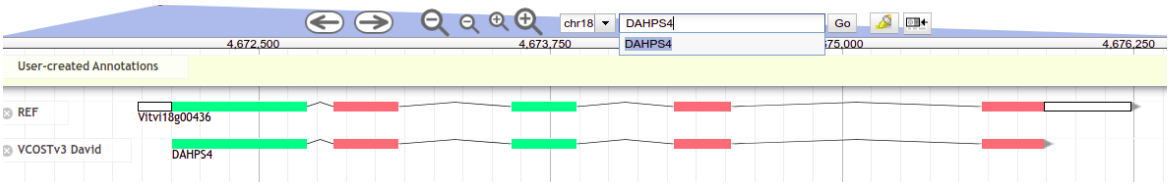
1. **Log in:**

Address: http://138.102.159.70:8080/apollo/

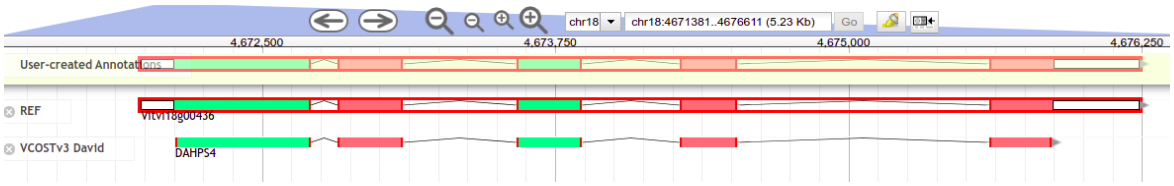
Enter your username and password to access the Apollo browser. Once you login you will see a panel on the right showing three tabs; annotations (showing user-created annotations), tracks (showing the information described above) and Ref Sequence (showing all the different reference sequence tracks and their stats). The left panel shows the different active tracks, a dropdown menu for choosing a particular reference sequence, zoom buttons and a search bar. Most importantly you have the user-created annotations track (in yellow) below the search bar. This is the interactive panel where you will drag and drop genes of interest.

2. Quick guide: annotating a gene

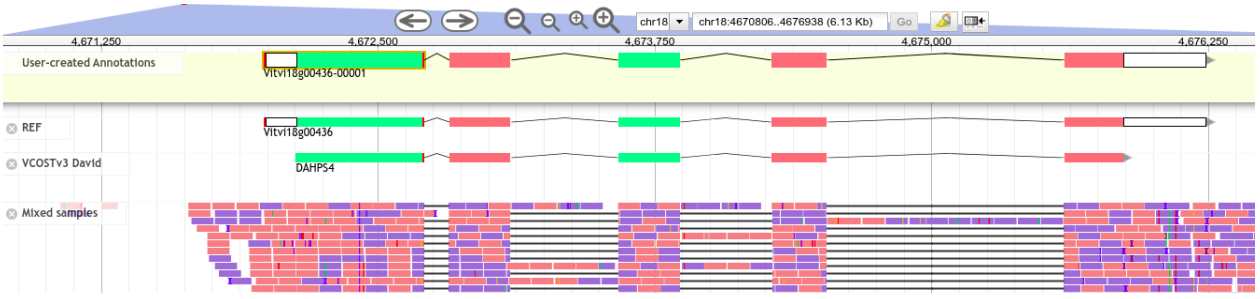
- Activate at least the following tracks: REF, VCOSTv3 and mixed samples.
- Find a gene of interest using its gene symbol (e.g. DAHPS4) or CRIBI V1, VCOSTv3, v4.1 gene IDs or a region of interest (format chrXX:start..end)



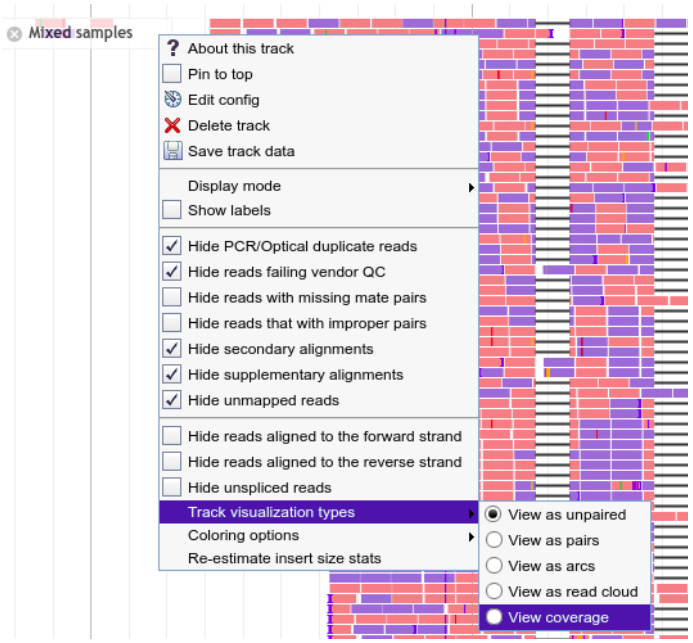
- Drag and drop a gene model of choice into the user-created annotations track (remember to select the full gene model by double-clicking, or else just a subfeature will be dragged)



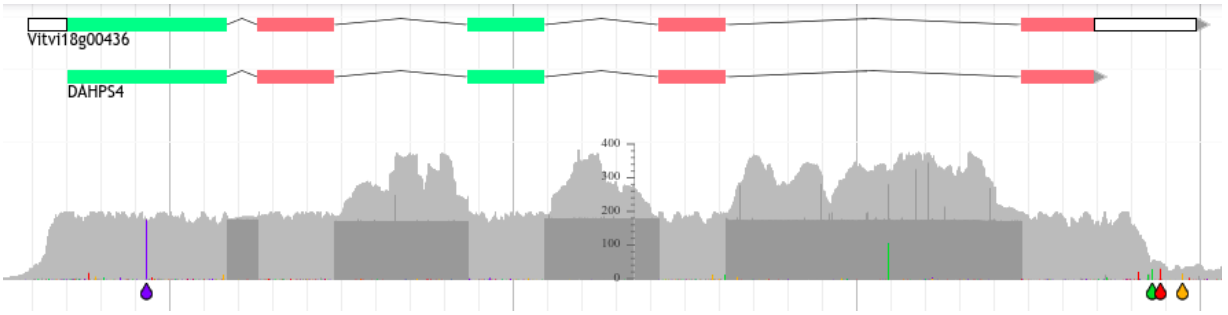
- Notice that UTRs are marked in white whilst CDSs are coloured in one of three colours
  - The arrow at the 3' end of a gene model marks the coding strand
  - There are three possible reading frames on each strand of the genome and each one has one colour (red, green or blue)
    - This is useful especially when considering merging two exons; if they share the same colour the reading frame wouldn't be changed in the merge and there is a lower chance of resulting in a truncated protein
- Apollo allows you to intuitively modify your user-created gene model -> Try the following
  - Drag exon boundaries to change their length (first select an exon)
  - Select two consecutive exons (hold shift) and right click "Merge"
  - Select a single exon and right click "Make intron" -> an intron will be predicted within
  - Right click on a gene model -> "Undo" and "Redo"
- Does the gene have RNA-Seq support from the mixed samples track or any other? If it does,
  - Use split alignments (horizontal black lines) to define introns and regions with read alignments (FW in red and RV in blue) to define exons



- An alternative is to use the coverage view which can be especially useful to show drops in coverage where intronic regions are otherwise unclear

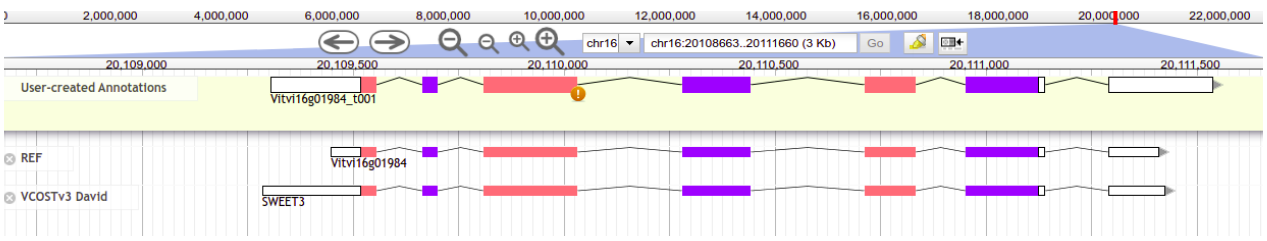


- Dark grey regions represent drops in read coverage which should correspond to introns.

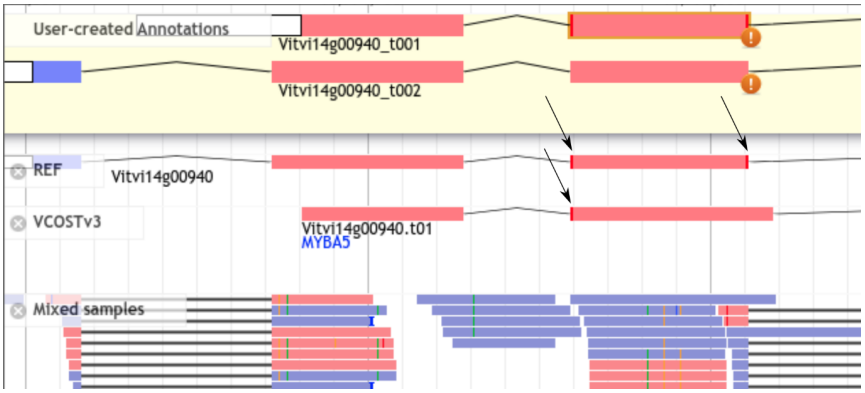


- If the gene of choice is part of a recently expanded gene family with high sequence similarity and coverage is low in the "mixed samples" track you can load the "mixed samples with multimapping" track while making sure that the "hide secondary alignments" and "hide supplementary alignments" options are unticked (they are ticked by default).

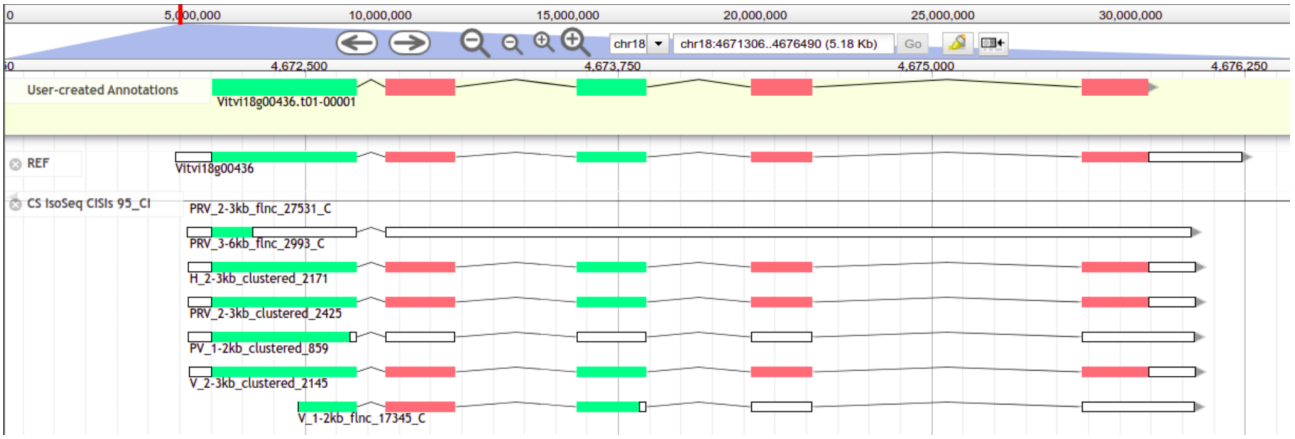
- Check intron-exon boundaries. Canonical boundaries for introns are GT followed by AG. Non canonical but still common boundaries (about 5% of the time) are G**C** and AG. These will be flagged with an orange exclamation mark.



- Select an exon when defining it since other exon boundaries (from other tracks) will light up in red should they coincide with the selected exon (the same would happen if you select the whole gene)
    - In this example it is clear that the *de novo* REF annotation is correct and the user-created gene model matches the REF exon and not the VCOSTv3 (which has no RNA-Seq support)

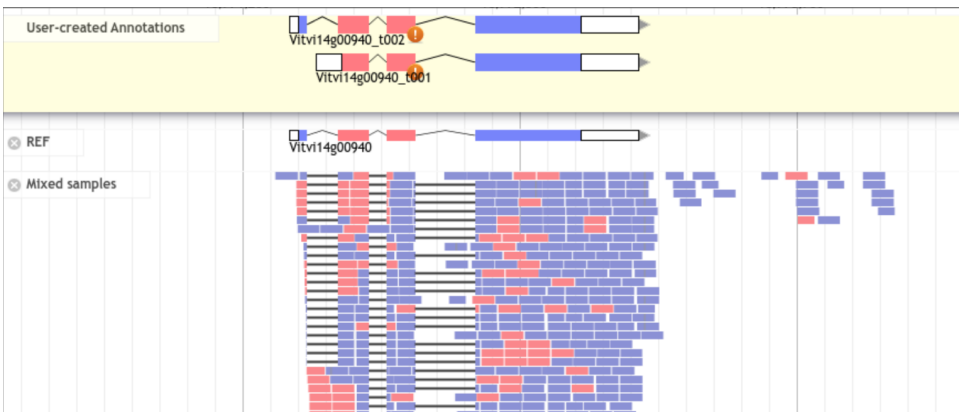


- Refine UTR boundaries:
    - Display the IsoSeq track (full length transcripts in fruit) and use the 5' and 3' ends of the transcript as evidence to refine the UTRs. This track is also very useful to annotate additional transcripts (i.e. alternative splice variants).

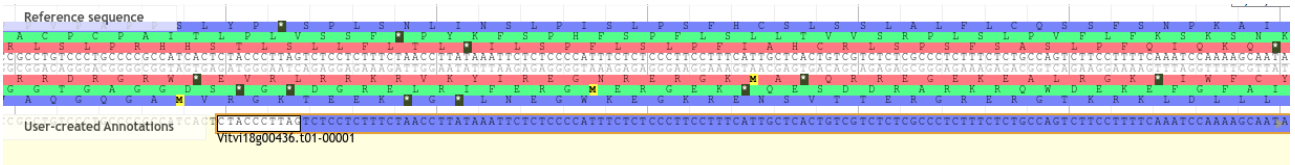


    - If no IsoSeq information is available, use the RNAseq track and delimit UTR boundaries where coverage begins to drop considerably (use both the read alignments and the coverage view to determine this)

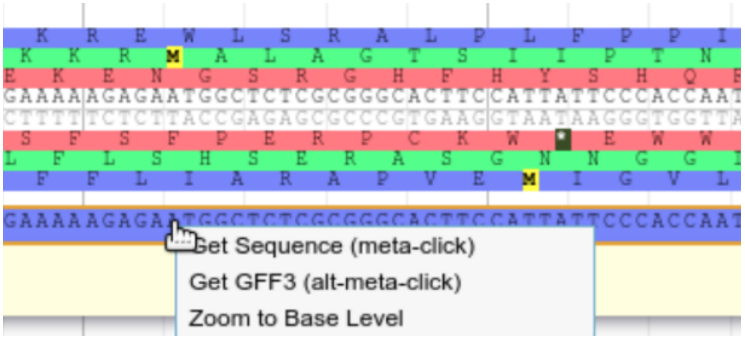
- To annotate alternative splice variants create a second transcript (drag and drop or duplicate). Make sure that multiple transcripts for the same gene have a different ID end (e.g. _t001 and _t002)



- Define the protein coding sequence:
    - **BEWARE**: Apollo has a habit of changing CDSs (specially the start) when transferring a gene model from another track into the user-created annotations or when resizing a UTR (this often means that the changed CDS doesn't even start with an ATG)
    - It doesn't always happen, but it happens enough times that you should double check your translation start and stop sites
    - Right click on the exon (first coding exon to search for the translation start) anc click on zoom to base level, this view is also recommended for defining exon boundaries.
        - The reference sequence is shown in black for the forward strand and grey for the reverse -> the arrow marks on gene models mark the sense of transcription
        - Protein sequences corresponding to the forward strand are shown above (the reference DNA sequence) -> the colours match the three possible reading frames depicted on CDSs. Reverse protein sequences are shown below the reference DNA.



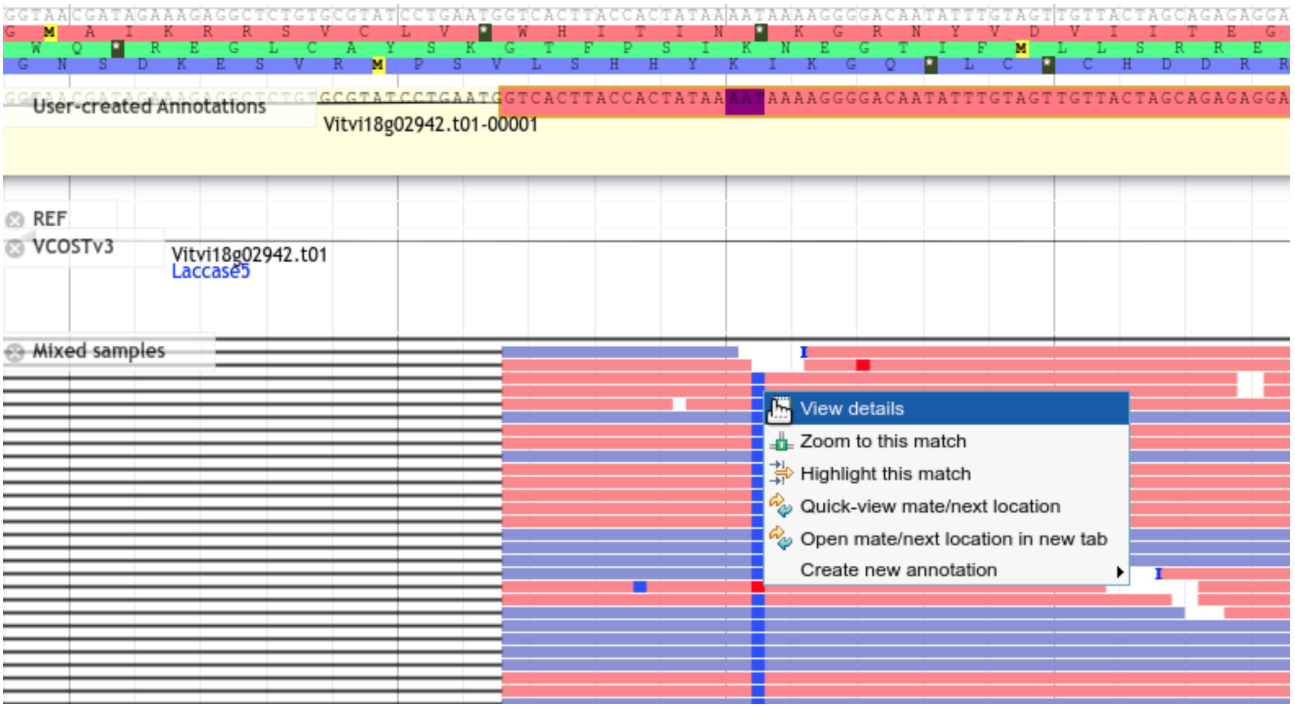
        - Notice that the translation start site in the previous example is incorrect (it doesn't start with a Methionine -> All Methionines in the genome are highlighted in yellow whilst stop codons are marked with *
    - Right click on the A from the ATG within the user-created sequence and select Set Translation Start. This can also be done for stop codons (Set Translation End). In the case of additional ATGs or alternative start sites -> add the default "Unclear translation start" comment to the transcript



    - Does your gene encode a reasonable complete protein? Obtain its protein sequence and blast it on NCBI or the specifically adapted blast server (Sequence Server) with all the Gene catalogue sequences and PN40024 proteins/CDS/transcripts from previous annotations: http://138.102.159.70:4567/
        - You can right click your user-created model (remember to select the whole transcript) and select Get Sequence -> you can then obtain the peptide, cDNA, CDS sequences or even upstream/downstream bps

- Occasionally (quite rarely) there may be a stop codon in the reference sequence, which is not supported by RNA-Seq. Discrepancies between read and assembly sequences are marked in different colours -> you can see below that the T from the TAA stop codon is in fact a different base as suggested by almost all RNA-Seq reads (right click on a read and view details to see its full sequence)

    - If there are discrepancies, check the SNP track to see if they are due to an assembly error and not cultivar differences. If it is clearly due to an assembly error you may consider the Set Readthrough Stop Codon option

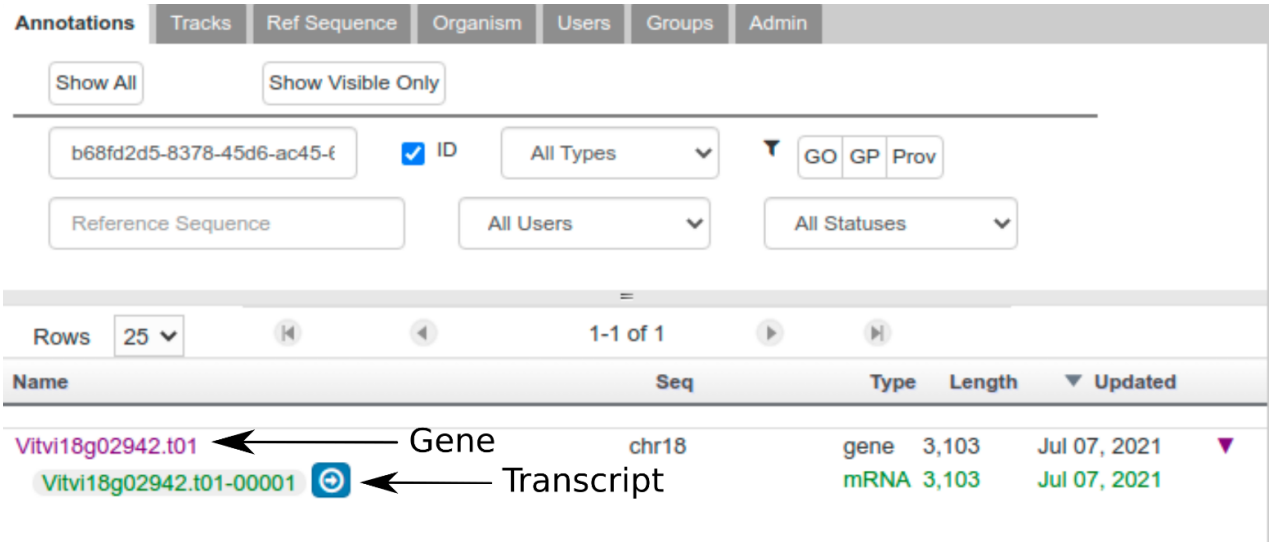


- Confirm the gene structure with other species (Arabidopsis, ORTHODB and SWISSPROT tracks)
- Confirm the gene structure using Gramene Genetree application (change V3 ID in address):

  http://curate.gramene.org/grapevine/?gene=Vitvi17g00614

- The default type for an annotation is gene, however, right-clicking on a gene allows you to change annotation types:
    - Change to **pseudogene** (e.g. truncated protein) if you think the gene structure can't provide a functional protein
    - If blasting the predicted sequence shows results in transposases or gag proteins, set the annotation status "to delete" as we are focusing on gene annotation
    - There are also a range of non-coding RNA options such as **tRNA**
- Is there a problem with a frameshift or a very short intron? Check the PN40024 Illumina SNPs track for errors in the assembly
    - If you can see an InDel on this track, create a "fake" intron restoring the correct frame and make sure to add the default "Assembly error" comment to your transcript
- Do not try to curate your gene across two scaffolds or over N gaps. Keep the annotations on a single scaffold and make sure to comment the transcript with the default comment "Partial"

- **Before moving to another gene, it is vital to check the metadata at the gene and transcript levels**
- Select your curated gene by double-clicking on it. Right click and select “Open Annotation”. You can then select gene and transcript features



  - **WARNING**: Searching for a gene in Apollo will not update the annotations panel automatically so make sure the ID in the right-hand panel matches the gene you are working with -> Always reopen the annotations panel with every new gene/transcript
  - If a gene symbol (i.e. name) is associated to your curated gene (e.g. DAHPS4) -> add it at the **gene level** under the “Symbol” field under the “details” tab
  - A range of predefined comments have been provided under the comments tab of **transcripts** as a dropdown menu, some of which have been described already
    - An interesting comment is the “**Pending annotation of other splicing variants**” flag which is meant to be added when you have spotted alternative splicing variants but you have not annotated them
- **Finally define the annotation status for all transcripts and their corresponding gene:**
  - **Approved** -> If your annotation edit is complete
  - In Progress -> If editing is ongoing
  - Problematic -> If you are unsure about the gene model and need help
  - **To delete** -> If you think this annotated region does not belong to any gene or shares similarity with transposable element sequences
- Only “approved” and “to delete” features at both gene and transcript level will be considered for the updated annotation (.gff file)
- Take a look at this Google document for keeping track of user created annotations:

https://docs.google.com/spreadsheets/d/1BxRbtkJFdvo4ObV8jU7nRSUt6u9jrzcSvLSRUtV7mZo/edit#gid=0

- Once your gene model is ready please update the “catalogue_annotation” tab from the Google document or the “Other curated genes” tab if your gene is not part of the Gene catalogue