Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL
August 18th, 2020
Agenda

• Basic RNA-Seq evaluation
• Basic structural changes- splitting and merging a model, adding and removing exons
• UTRs –when and how to add and adjust
• Changing translation start and stop sites, and open reading frames
• Non-canonical splice sites
• Annotating isoforms
• Sequence alterations and stop-codon readthroughs
• Annotating Non-coding features
Other resources

• An additional Apollo webinar with more background: https://www.youtube.com/watch?v=dol99KExLgY&feature=youtu.be

• Monica Munoz-Torres from the Apollo group has a number of comprehensive tutorials:
  • https://www.slideshare.net/MonicaMunozTorres/presentations
    • I recommend these slides if you need more background:
      • https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015
    • If you are new to Apollo, or need a refresher, I highly recommend that you review one of her presentations

• The official Apollo annotation guide:
  • http://genomearchitect.org/users-guide/

• i5k Workspace manual annotation landing page: https://i5k.nal.usda.gov/manual-annotation-and-apollo

• Other manual curation tutorials: http://genomecuration.github.io/genometrain/d-feature-curation-crossing/

• VEuPathDB Apollo training webinar: https://eupathdb.org/eupathdb/webinars.jsp#apollo
Basic RNA-Seq evaluation
RNA-Seq tracks

- **Coverage plots**: Histogram of the number of mappings at each nucleotide; hover over the blue area to see the value.

- **Mapped reads**: Individual glyphs of each mapped read. Show mapped and spliced areas, and SNPs/indels. Informative, but hard to work with when zoomed out.

- **Junction reads**: Useful combined with coverage plots; show where mapped reads are spliced. Control-click on read and look under ‘score’ to see how many mapped reads support the splice junction.
A simple case

RNA-Seq coverage

Mapped reads

Splice junctions
A simple case

Select ‘About this track’ from drop-down menu

Information about methods
A more complex case

A second isoform?

A missing exon?
A really messy case

- Mappings support models on forward and reverse strands
- Lots of mapped reads not reflected in gene model
Basic structural changes – splitting and merging a model, adding and removing exons
Annotation Example

• Glycerate kinase-like in the trap-jaw ant *Odontomachus brunneus*

• More information about the trap-jaw ant genome assembly: [https://i5k.nal.usda.gov/odontomachus-brunneus](https://i5k.nal.usda.gov/odontomachus-brunneus)

• *Odontomachus brunneus* Apollo URL: [https://apollo.nal.usda.gov/apollo/4006447/jbrowse/index.html?loc=NW_022639451.1%3A5084490..5093717&tracks=DNA%2CAnnotations%2CNCBI_Annotation_Release_100_Gene-CBT&highlight=]
RNA-Seq evaluation

Very different coverage between UTR and CDS
RNA-Seq evaluation

Junction reads don’t support connection between the two expressed regions
Create new model in user-created annotations track

Drag evidence to UcA track (or right-click and select “create annotation”)
Split model

Select exons on which to split the model using the ‘shift’ key

Right-click on the model while continuing to hold shift to get the drop-down menu

Select ‘split’
Split model

You now have 2 models! Let’s start fixing the model on the right – it needs a 3’ exon.
Add an exon

Zoom in, select the missing exon, drag up to UcA track
Merge exons

Shift-select both exons, shift-right click, then select ‘merge’ from the dropdown menu
UTRs – how and when to add or adjust
Adding or adjusting UTR boundaries

• When should you add or change UTRs?
  • Only if you have RNA-Seq evidence with sufficient coverage (e.g. > 50 reads)
  • Adding or changing UTRs is helpful, but not necessary if you’re only interested in the protein sequence
  • Deciding where the UTR ends is usually a judgement call

• Apollo tools for gene boundary changes:
  • Manual edge-matching to available evidence
  • Automated edge-matching to available evidence
Adjusting gene boundaries

RNA-Seq evidence ends in different places for each track – how do you decide?
Adjusting gene boundaries

Pick the longest boundary available, and note which track you used in the ‘Comments’ section.
Adjusting gene boundaries

One way to change the boundary: find a mapped read on the same strand as the model; hold shift and click on the read and the model to highlight them both.
Adjusting gene boundaries

Right-click on model in user-created annotations track, and select ‘Set as 3’ end from the drop-down menu.
Adjusting gene boundaries

New UTR is there!
Adjusting gene boundaries

Add comment explaining UTR addition
Starts, stops, open reading frames
Setting the sequence start, stop, and open reading frame (ORF)


- However, in some fringe cases, you will need to double-check

- You can change this if needed
Starts, stops, ORFs

Open reading frame (ORF): translated region

Translation start at Methionine

Translation stop at stop codon
This is a ‘low quality’ protein coding gene from NCBI – it will likely show some problems in Apollo
We can see a non-canonical splice site in the UcA (more on that later). Let’s zoom to the start of the model.
Starts, stops, ORFs

Apollo shows this model in the green reading frame – however, we can see a stop pretty early on in the genome sequence - but that’s not reflected in the Apollo model! It looks like the pink reading frame doesn’t have stops.
Starts, stops, ORFs

Sure enough, the protein sequence is suspiciously short.
Starts, stops, ORFs

Let’s set the translation start in the pink reading frame – click on the 3rd nucleotide in the UcA, right-click, and select ‘Set Translation Start’
Starts, stops, ORFs

We’re in the pink reading frame now – let’s check the protein sequence
Starts, stops, ORFs

That looks better.
Sometimes it can be hard to tell what the protein sequence should be – in that case you can right-click and select ‘Set Longest ORF’
Starts, stops, ORFs

This also fixed the reading frame.
Similarly, if you have evidence to change the translation end, you can click on the corresponding nucleotide, right-click, and select ‘Set Translation End’
Starts, stops, ORFs

Now the sequence after the translation end is 3’ UTR.
Non-canonical splice sites
Splice sites

Introns are removed from primary transcripts by cleavage at conserved sequences called **splice sites**. These sites are found at the 5’ and 3’ ends of introns. (https://www.nature.com/scitable/topicpage/rna-splicing-introns-exons-and-spliceosome-12375/)

5’

3’

Donor splice site (GT/U)

Acceptor splice site (AG)

’Non-canonical’ splice sites – non-conserved, and possibly erroneous sites – are marked by an exclamation point in Apollo.
Fixing non-canonical splice sites

Looks like there’s another isoform here – let’s add an exon
Fixing non-canonical splice sites

Drag model to UcA track
Fixing non-canonical splice sites

Drag evidence to UcA track to add 5’ exon
Fixing non-canonical splice sites

Oops, wrong strand – our model is on the forward strand
Fixing non-canonical splice sites

Let’s flip it to the opposite strand
Fixing non-canonical splice sites

That’s better.
Fixing non-canonical splice sites

Merge the new exon to the gene model
Fixing non-canonical splice sites

Non-canonical splice sites in merged model
Fixing non-canonical splice sites

The splice junction reads don’t support the 2nd exon with the new 5’ exon, so let’s remove it.
Fixing non-canonical splice sites

Only 1 non-canonical splice site left to fix
Fixing non-canonical splice sites

Extend exon to RNA-Seq boundary
Fixing non-canonical splice sites

Still not quite right
Fixing non-canonically spliced sites

Let’s have Apollo find the downstream splice site donor for us.
Fixing non-canonical splice sites
Annotating isoforms
Isoform annotation example

• In our experience, lots of mapped RNA-Seq reads are critical for good manual isoform annotation

• Before evaluating RNA-Seq for isoforms, it helps to understand how to interpret gradual and abrupt drops in coverage
  • Gradual – usually means 5’ start or 3’ end of expression
  • Abrupt – usually means splice junction

• Checking junction reads (if available) is incredibly useful
Isoform annotation example

5’ end of MAKER tyrosine protein kinase gene prediction

RNA-Seq coverage plot

Abrupt coverage dropoff

Gradual coverage increase

RNA-Seq junctions

Pooled data from 60 adult 1st generation female beetles from Wisconsin coverage

Pooled data from 60 adult 1st generation female beetles from Wisconsin - Splice Junctions
Isoform annotation example

Mapped RNA-Seq reads

Splice junction

Splice junction
Isoform annotation example

Create 2 isoforms from Maker model
Isoform annotation example

Select and delete 5’ exon from one of the isoforms
Isoform annotation example

Add a new 5’ exon from mapped RNA-Seq evidence
Isoform annotation example

Merge the new 5‘ exon with the rest of the model
Isoform annotation example

2 isoforms supported by RNA-Seq evidence
Sequence alterations and stop-codon readthroughs
Sequence alterations

• Apollo supports annotating the genome sequence with insertions, deletions, and substitutions
• Note that this will not change the genome fasta in the sequence export – but it will allow Apollo to re-calculate a gene model’s sequence
• Only add a sequence alteration in Apollo if there is evidence for it – e.g. SNPs in mapped RNA-Seq
Stop-codon readthroughs

• Apollo allows you to annotate stop-codon readthrough features on the coding sequence of a gene model.
• This is a special case for selenocysteine-containing proteins.
• This feature can be used in other cases — e.g. if you have evidence of errors in the genome assembly — but we don’t recommend it.
Sequence alterations

RNA-Seq suggests that this intron may not exist
Sequence alterations

Adding the gene model to Apollo reveals non-canonical splice sites
Sequence alterations

There is some support for a spliced isoform, but the RNA-Seq also suggests contiguous coding sequence.
Zooming in, we see a stop codon in the ‘pink’ frame on the reverse strand, but SNPs in all the RNA-Seq reads.
Sequence alterations

Right-click on the corresponding nucleotide in the *genome assembly* and select ‘Create Genomic Substitution’
Add the substitution, and a (required) justification
Sequence alterations

Apollo added the substitution! Now, let’s merge the CDS regions
Sequence alterations

The sequence merged!
Non-coding features
Non-coding features

• Apollo supports protein-coding and non-coding features
• By default, Apollo will create protein-coding features
• For non-coding features, you can set the feature type before or after setting up the model
Non-coding features

This feature is a tRNA

Let’s drag it to the UcA track to modify it
Non-coding features

Apollo turned it into an mRNA – let’s fix that
Non-coding features

Right-click on feature, select ‘Change annotation type’, then ‘tRNA’
Non-coding features

Now we have the correct feature type
Non-coding features

Another way – right-click on model before adding it to the UcA track, select ‘Create new annotation’, then select ‘tRNA’
Non-coding features

Now it’s a tRNA!
Thank you!

The NAL Team
• Chris Childers
• Vern Chapman
• Ming Chen
• Susan McCarthy
• Shang-Yu Chang
• Hsiu-Kang Huang

• i5k Coordinating Committee
• i5k Pilot Project
• Apollo & JBrowse Development Teams
• GMOD/Tripal community
• All of our users and contributors!

Contact us:
- https://i5k.nal.usda.gov/contact
- i5k@ars.usda.gov
- Monica.Poelchau@usda.gov
- Christopher.Childers@usda.gov