

UCSC_Hcal_v1 annotation guidelines

Darrin T. Schultz
dts@ucsc.edu
26 December 2019

Motivation

This document serves as guidelines for selecting a hand-curated annotation for the *Hormiphora californensis* genome (UCSC_Hcal_v1). It contains descriptions of the requisite files, and how to select transcripts.

Files

All of the files are contained in the `files_for_annotation` directory on the google drive.

- Iso-Seq HQ collapsed and filtered transcripts - `GL064_isoseq.collapsed.filtered.gff`
 - Contains HQ (high-copy-number) Iso-Seq transcripts, clustered by IsoSeq3, then collapsed using Liz Tseng's `cDNA_cupcake` tools. Easy to look at and a lot of evidence for each transcript in this set.
- Iso-Seq singleton-inclusive transcripts - `GL064_singletons.collapsed.gff`
 - Contains HQ and low-copy-number transcripts. **Often has transcripts that are not present in any other set.**
- Pinfish transcripts - `UCSC_Hcal_v1_B1_LR.pinfish_clusters_c2p20.gff.gz`
 - Contains transcripts from the output of the pinfish program - ran with Iso-Seq CCS reads as the input. **Sometimes has things that the stringtie transcripts lack. Often has multi-mapped genes that are missing from stringtie.**
- stringtie transcripts - `UCSC_Hcal_v1_B1_LR.stringtie_f01.gff.gz`
 - Contains transcripts from stringtie, ran with the Iso-Seq CCS reads as the input. This is the base set of transcripts.
- iso-seq HQ bam - `UCSC_Hcal_v1_B1_LR_HQ.sorted.bam`
 - contains the HQ (high-copy-number) and partly collapsed Iso-Seq reads.
- iso-seq CCS bam - `LR_to_Hcv1.sorted.bam`
 - contains all the CCS reads from the Iso-Seq output. **Has many reads for genes that are not present in stringtie, pinfish, or Iso-Seq HQ gffs. Always look at this track to check for missing transcripts.**
- genome assembly - `UCSC_Hcal_v1.fa`
 - The genome assembly that is on NCBI.

The annotation strategy - summarized

The stringtie transcript set has many of the transcripts correct, based on the long PacBio Iso-Seq data. However, sometimes transcripts are missing, and sometimes there are transcripts fused. In these cases we can remove the stringtie transcript, and/or add a transcript from another source.

The general procedure for finding the correct transcripts will be to manually scan the entire genome (all 110 million bases), and update a google docs spreadsheet by hand to pick the final annotation. The spreadsheet will start by being populated with the stringtie transcripts as a guide, and new rows will be added as needed as transcripts are added from other annotation sources. Stringtie transcripts will not be removed if they are incorrect, but instead will be flagged for removal later on.

After producing the set of manual annotations, additional isoforms will be added by searching the genes against a DB of Iso-Seq reads.

Number of stringtie transcripts per scaffold


Below is a table of number of stringtie transcripts per scaffold. One person will annotate one half. The second person will annotate the other half. Altogether there are 10458 genes in the stringtie dataset.

1208	c1
1149	c2
906	c3
606	c4
946	c5
881	c6
----- Half of transcripts above, The other half below	
820	c7
783	c8
726	c9
569	c10
811	c11
472	c12
544	c13
37	sca1 - sca31+ M

Description of columns in annotation file

The annotation spreadsheet is simple. It needs to be simple given that there are 10000 genes to review. The columns are as follows:

- A. chromosome
 - contains the header of the genome fasta file. c1 is chromosome 1, sca1 is scaffold 1 not placed on a chromosome, M is mitochondrion
- B. stringtie_id
 - The id of the stringtie gene. There can be multiple isoforms for each gene, but the ID here is just the base name for each gene. This string is grep-able from the gtf file.
- C. DTS_checked
 - just used to keep track of position while annotating and a record of completion.
- D. WRF_checked
 - used to keep track of position while annotating.
- E. pinfish_id
 - the pinfish id for a new transcript. The ID is some long nonsensical number that is in no way related to other isoforms of the same gene. For example this is a pinfish_id for one gene:
135c34c5-f6fb-4240-8ce2-57aa64e16785
- F. isoseq_hq_id
 - The id for genes coming from the isoseq_hq gtf file. The names will be in the form "PB.19" for the 19th gene of this file.
- G. isoseq_singleton_id
 - Same as isoseq_hq_id. The id for genes coming from the isoseq_singletons gtf file. The names will be in the form "PB.19" for the 19th gene of this file.
- H. remove_st
 - This is a flag to remove this stringtie gene (because we will not remove stringtie genes from the spreadsheet). The text in this column can be ["yes", "YES", "x", "true", "True"] or any caps variant.
- I. interesting
 - This is a flag to pull out interesting genes. Could be whatever reason. ["yes", "YES", "x", "true", "True"] or any caps variant.
- J. spliced_in_intron
 - A flag to pull out spliced genes located inside introns of other genes. ["yes", "YES", "x", "true", "True"]
- K. comment
 - whatever comment you would like to add

 Hcal_annotation ☆ 📁

File Edit View Insert Format Data Tools Add-ons Help [All changes saved in Drive](#)

100% \$ % .0 .00 123 Default (Ari... 10 B I S A 🔍 📏 📐 📊 📋 📌 📍 📎 📏 📐 📊 📋 📌 📍 📎

	A	B	C	D	E	F	G	H	I
1	chromosome	stringtie_id	pinfish_id	isoseq_hq_id	isoseq_singleton_id	remove_st	interesting	spliced_in_intron	comment
2	c1	B1_LR.1							
3	c1	B1_LR.2							
4	c1	B1_LR.3							
5	c1	B1_LR.4							

Rules for Annotation

There are several DOs and DON'Ts for annotation to make sure that it is consistent between multiple people annotating different regions of the genome.

DOs

1. DO look at both the HQ Iso-Seq reads and unfiltered Iso-Seq reads bam files. There will be reads in the unfiltered bam file that are true genes, but didn't make it into the stringtie transcripts.
2. DO look at all four gene sources when looking for the best gene for a locus.
3. DO keep an eye out for weird features and mark them in the spreadsheet.
4. DO make sure that transcripts in stringtie are not inappropriately fused. Keep an eye on the reads and possible transcripts that are really close to one another. There is a tendency for there to be short, single-exon genes to be immediately 5' or 3' of another transcript.
5. DO flag incorrect stringtie genes in the `remove_st` column.
6. DO look out for spliced genes located in the intron of another gene. Mark this in `spliced_in_intron` column.

DON'Ts

1. DON'T delete any rows in the spreadsheet. Incorrect stringtie genes should be flagged in the `remove_st` column.
2. DON'T worry about a little bit of 5' or 3' UTR missing from the gene model if most of it is there. Later on in the annotation process we will use this gene set as a base, then pull out high quality transcripts to fill in additional isoforms.

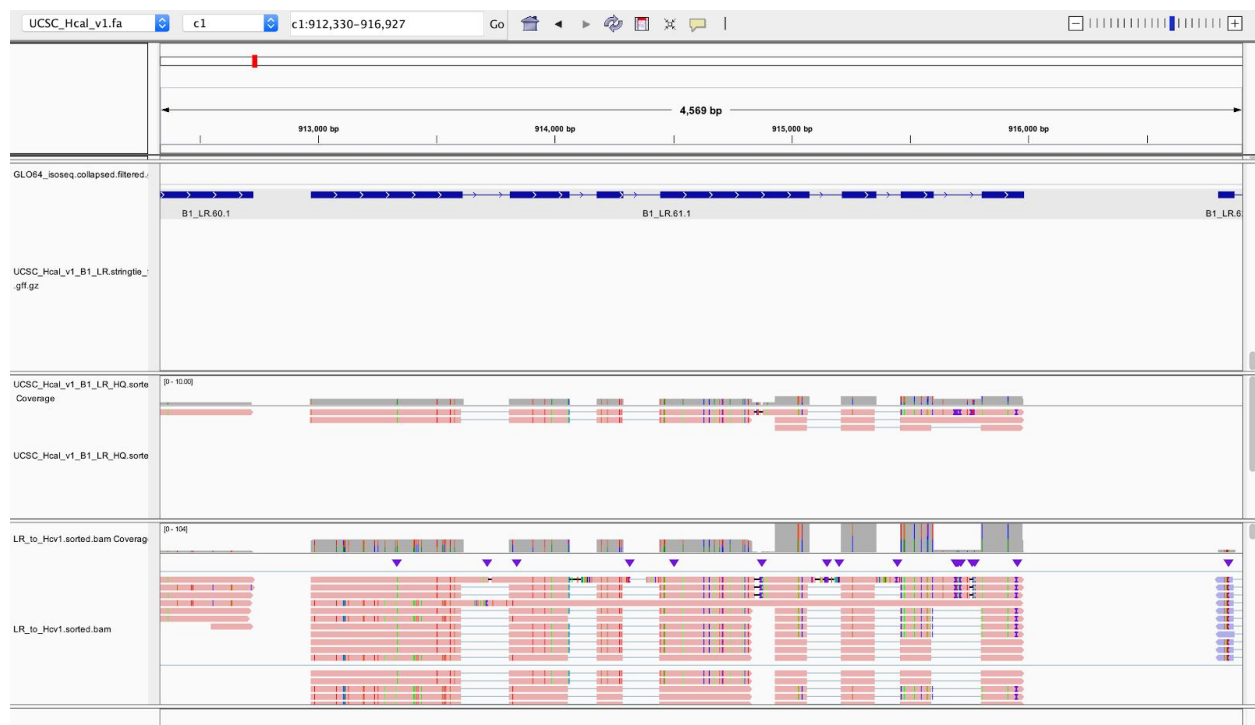
Annotation examples

Annotation Example 1 - Copacetic annotation

In this case the gene appears to have been correctly annotated by stringtie. We don't really do anything, except maybe mark the spreadsheet that we checked the stringtie annotation and it looks fine.

In IGV:

The annotation for B1_LR.61.1 matches the reads.



On the Spreadsheet:

We just mark that we checked the annotation (DTS_checked) and that it looks fine.

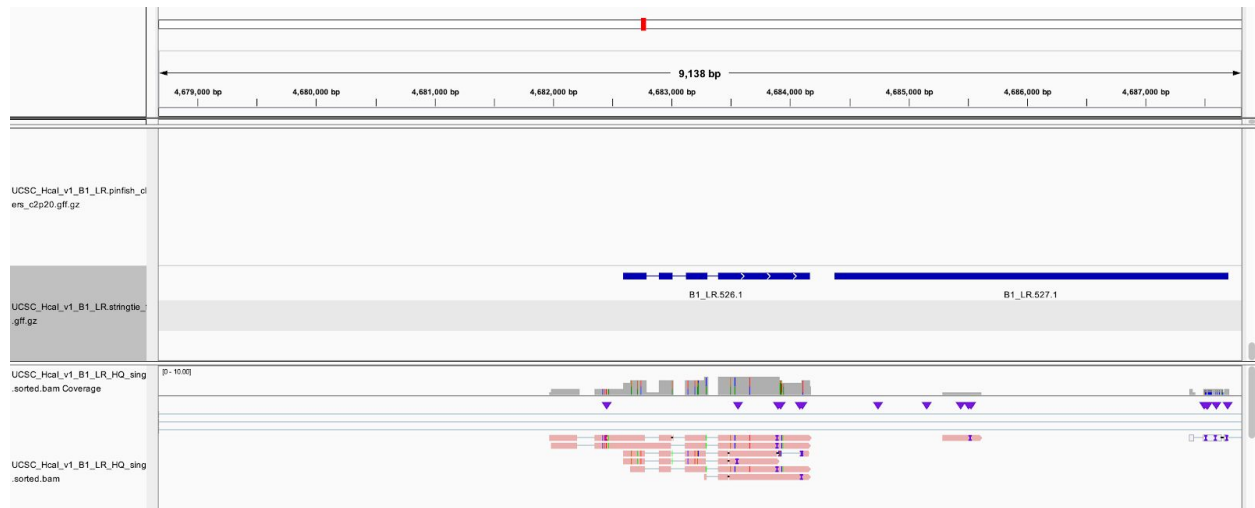
	A	B	C	D	E	F	G	H	I	J	K	L
1	chrom	stringtie_id	DTS_checked	spliced_in_intron	WRF_	isoseq_hq_id	pinfish_id	isoseq_singleton_id	remove_st	interesting	comment	time
110	c1	B1_LR.61	y									

Annotation Example 2 - The annotation is missing an exon

If the annotation looks correct, but is just missing an exon or two, then just mark the annotation as correct. The correct annotation with all of the isoforms can be pulled out later with BLAST.

In IGV:

The annotation for B1_LR.526.1 is correct, but there an isoform is missing on the 5' end.



On the Spreadsheet:

We just mark that we checked the annotation (DTS_checked) and that it looks fine.

	A	B	C	D	E	F	G	H	I
1	chrom	stringtie_id	DTS_checked	spliced_in_intron	WRF_isoseq_hq_id	pinfish_id	isoseq_singleton_id	remove_st	
700	c1	B1_LR.526	y						

Annotation Example 3 - The stringtie annotation doesn't exist

In this case the stringtie annotation is just wrong. It is either in the wrong orientation, or just simply doesn't exist using the read data. We will mark this transcript for deletion and add the correct transcript.

In IGV:

The annotation for B1_LR.395.2 is in the wrong orientation. It is not the same gene as B1_LR.395.1 as it is on the opposite strand.



On the Spreadsheet:

The model B1_LR.395 is wrong, as it has transcripts from both strands. In this case it is just easier to do the following:

1. Delete B1_LR.395 (row 512)
2. Add back the specific transcript B1_LR.395.1 because it is correct.
3. Add the pinfish transcript f2c7a225... as that is also correct.

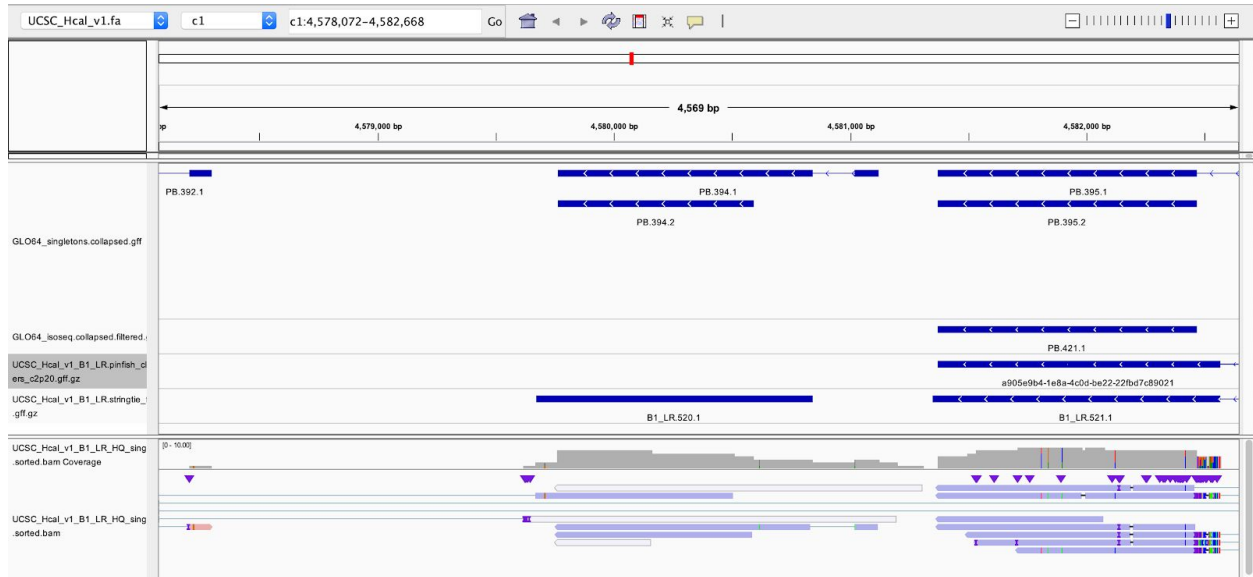
	A	B	C	D	E	F	G	H	I	J	K
1	chrom	stringtie_id	DTS_checked	spliced_in_intron	WRF_	isoseq_hq_id	pinfish_id	isoseq_singleton_id	remove_st	interesting	comment
512	c1	B1_LR.395	y						y		
513	c1	B1_LR.395.1	y								
514	c1		y	y			f2c7a225-2a3b-4884-b754-80628ce3069c		y		

Annotation Example 4 - The annotation has no direction

Sometimes Stringtie doesn't assign the correct direction to a transcript. In IGV this is easily noticeable when the transcript has no arrows.

In IGV:

The annotation for B1_LR.520.1 has no direction. This is not correct, as you can see all the transcripts are on the opposite strand (Iso-Seq data is all directional).



On the Spreadsheet:

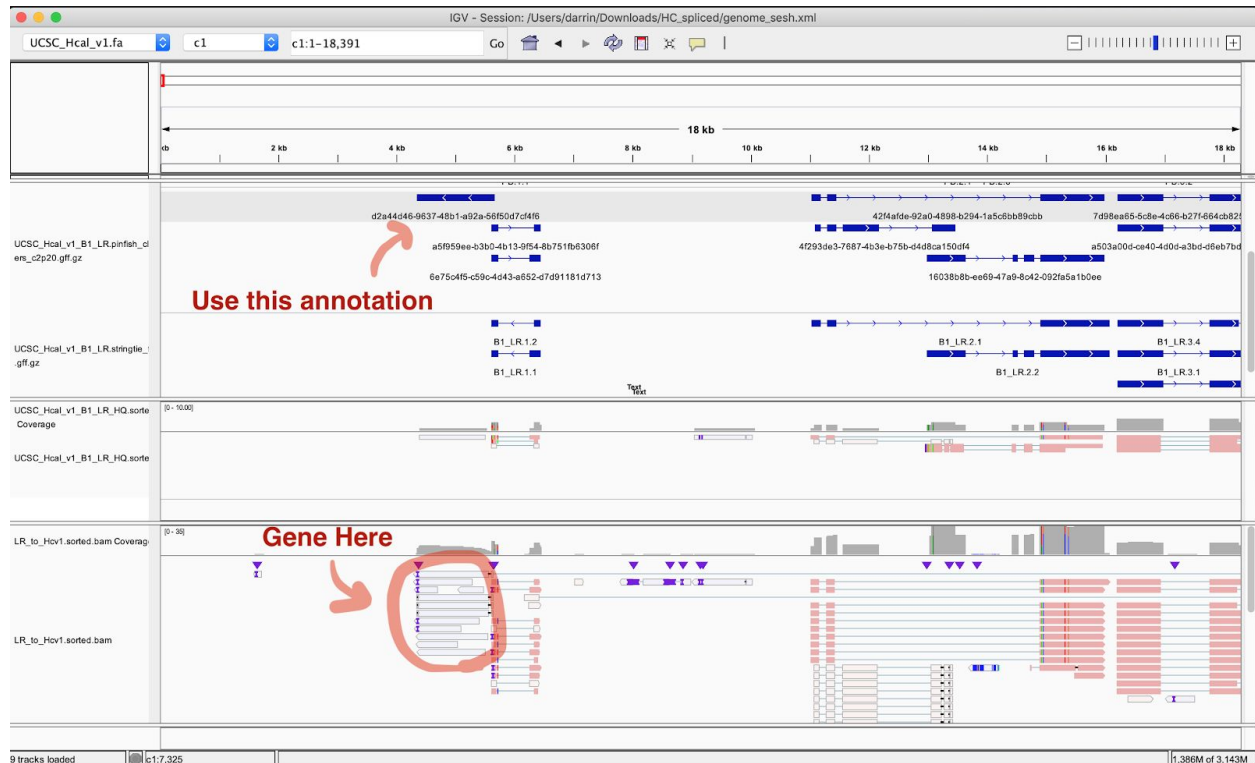
Delete the stringtie transcript that has no direction. Add a new row and add a transcript that has the correct directionality.

	A	B	C	D	E	F	G	H	I
1	chrom	stringtie_id	DTS_checked	spliced_in_intron	WRF_isoseq_hq_id	pinfish_id	isoseq_singleton_id	remove_st	
689	c1	B1_LR.520	y						y
690	c1		y				PB.394.1		

Annotation Example 5 - Multi-mapped gene

In this example, there is a multi-mapped gene that wasn't present in the stringtie geneset. Multi-mapped means that the reads map to more than one place in the genome. In IGV you can tell that this gene maps to multiple locations based on the color in alignment view, and by hovering over the reads (Mapping = Secondary @ MAPQ 0).

In IGV:



On the Spreadsheet:

We made a new row and pasted the pinfish sequence ID to the pinfish_id column. We also noted that this is a multi-mapping gene by putting 'mm' in the comments section.

	A	B	C	D	E	F	G	H	I
1	chromosome	stringtie_id	pinfish_id	isoseq_hq_id	isoseq_singleton_id	remove_st	interesting	spliced_in_intron	comment
2	c1		d2a44d46-9637-48b1-a92a-56f50d7cf4f6						mm
3	c1	B1_LR.1							

Annotation Example 6 - Overlapping gene / Multi-source Annot.

In this case the stringtie annotation missed a few isoforms of a locus with many overlapping isoforms. We had to delete one stringtie annotation and add annotations from other sources to reflect the locus.

In IGV:

There is a single, complex, gene with many overlapping isoforms (B1_LR.58 and B1_LR.59). Stringtie is also missing the isoform that spans both of these models.



On the Spreadsheet:

To correct this, we marked B1_LR.58 for deletion, added it to B1_LR.59, then added isoforms from the Iso-Seq HQ transcripts and the Iso-Seq singleton-inclusive transcripts to complete the gene model

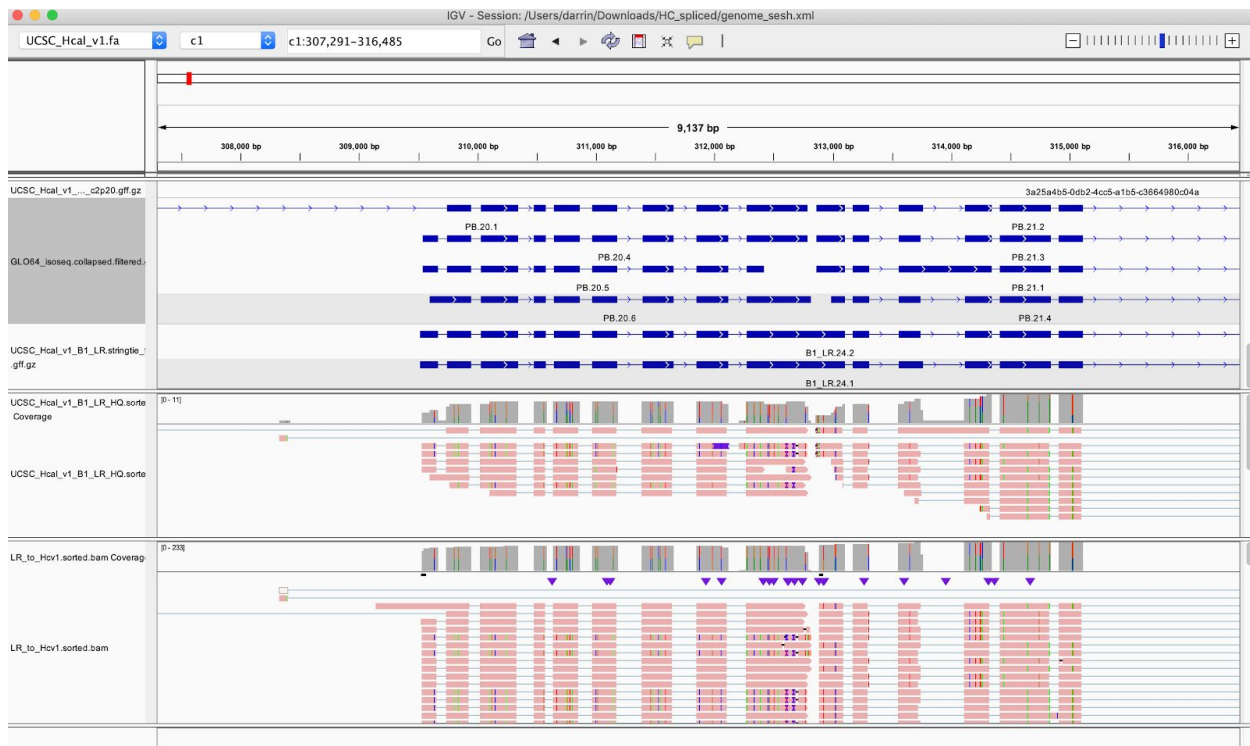
	A	B	C	D	E	F	G	H	I	J	K	L
1	chrom	stringtie_id	DTS_spliced_in_intron	WRF_	isoseq_hq_id	pinfish_id	isoseq_singleton_id	remove_st	interesting	comment	time	
107	c1	B1_LR.58	y					y				
108	c1	B1_LR.59, B1_LR.58			PB.43.3		PB.55.9					

Annotation Example 8 - Stringtie fused two transcripts

Stringtie has a tendency to fuse transcripts when they are very close to one another. In these scenarios, we must delete the stringtie annotation and add the correct annotations separately.

In IGV:

The Iso-Seq data clearly show that there are two transcripts, albeit very close to one another (the split is at 312,500bp). Stringtie transcript B1_LR.24 is a fusion of these two genes. This is incorrect. The correct annotation is PB.20 and PB.21 from the Iso-Seq HQ transcripts.



On the Spreadsheet:

We removed the stringtie transcript B1_LR.24 by adding a “y” to the `remove_st` column. We added two rows below and added PB.20 and PB.21 from Iso-Seq HQ transcripts.

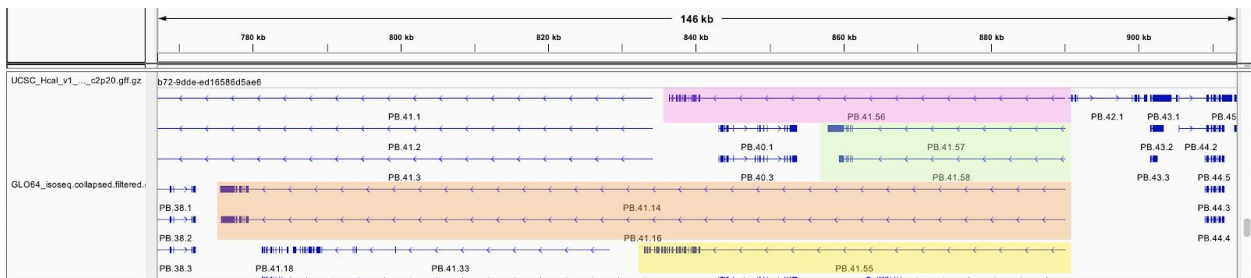
	A	B	C	D	E	F	G	H	I	J	K	L
1	chromosome	stringtie_id	DTS_checked	WRF_checked	pinfish_id	isoseq_hq_id	isoseq_singleton_id	remove_st	interesting	spliced_in_intron	comment	time
42	c1	B1_LR.24	y					y				
43	c1					PB.20						
44	c1					PB.21						
45	c1	R1 IR 25										

Annotation Example 9 - Stringtie fused two transcripts

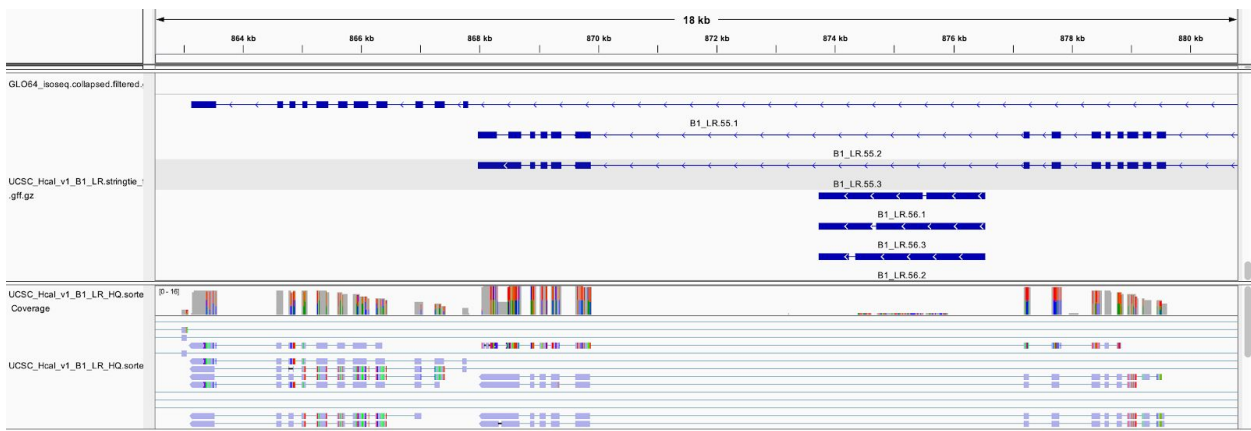
Sometimes there are clusters of genes that share a single first exon, but the “isoforms” actually encode completely different proteins - there is no overlap whatsoever in the protein sequence aside from a small N-terminal portion.

In IGV:

Below we can see that these three genes share the same small start exon, but the actual protein content is completely different. Technically they are alternative splice variants of the same gene, but practically speaking there are four different genes. (yellow and pink are the same).



Below, we see that stringtie annotated B1_LR.55.1, B1_LR.55.2, and B1_LR.55.3 as the same gene, even though they do not overlap in protein-coding sequence.



On the Spreadsheet:

The original gene model needed to be split, so I marked B1_LR.55 for deletion, added two rows below, and designated B1_LR.55.1 as one gene, and the cluster of [B1_LR.55.2, B1_LR.55.3] as another gene.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	chrom	stringtie_id	DTS	spliced_in_intron	WRF	isoseq_hq_id	pinfish_id	isoseq_singleton_id	remove_st	interesting	comment	time	
101	c1	B1_LR.55	y						y				
102		B1_LR.55.1	y	y						y	shares a start exon with other genes		
103		B1_LR.55.2, B1_LR.55.3	y	y						y	shares a start exon with other genes		

Timed runs

To calculate how much time this process will take we timed how many sequences we could annotate when uninterrupted.

Time period	Annotator	Number of stringtie transcripts	seconds per transcript	Projected time for 10458 transcripts
30 m	DTS	70	25.714	74.7 hr
30 m	DTS	59	30.5	88.6 hr
30 m	DTS	57	31.578	91.73
30 m	DTS	59	30.5	88.6 hr

Observations

- There is a large stretch of c4 with very few genes. Around positions c4:5,039,631-5,055,289.
- almost no operon-type gene clusters on chromosome 4