Liftoff moth annotation

G$

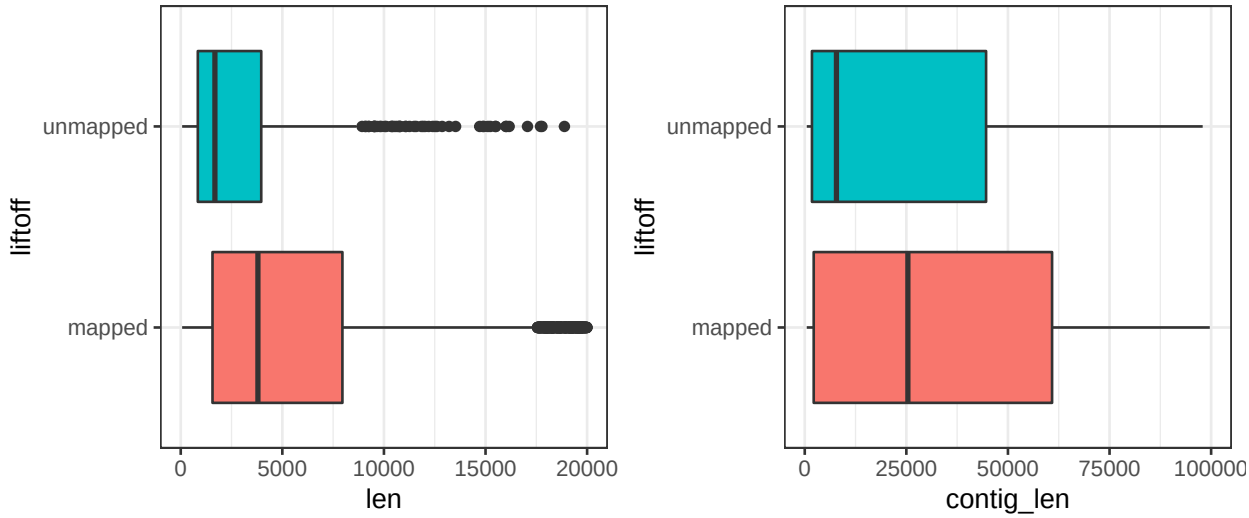
July 21, 2020

Running liftoff

1. Run Liftoff on moth annotations total features in reference gtf: 41110
2. liftoff gtf to new asm total features: 39957
3. There were 482 “unmapped” features

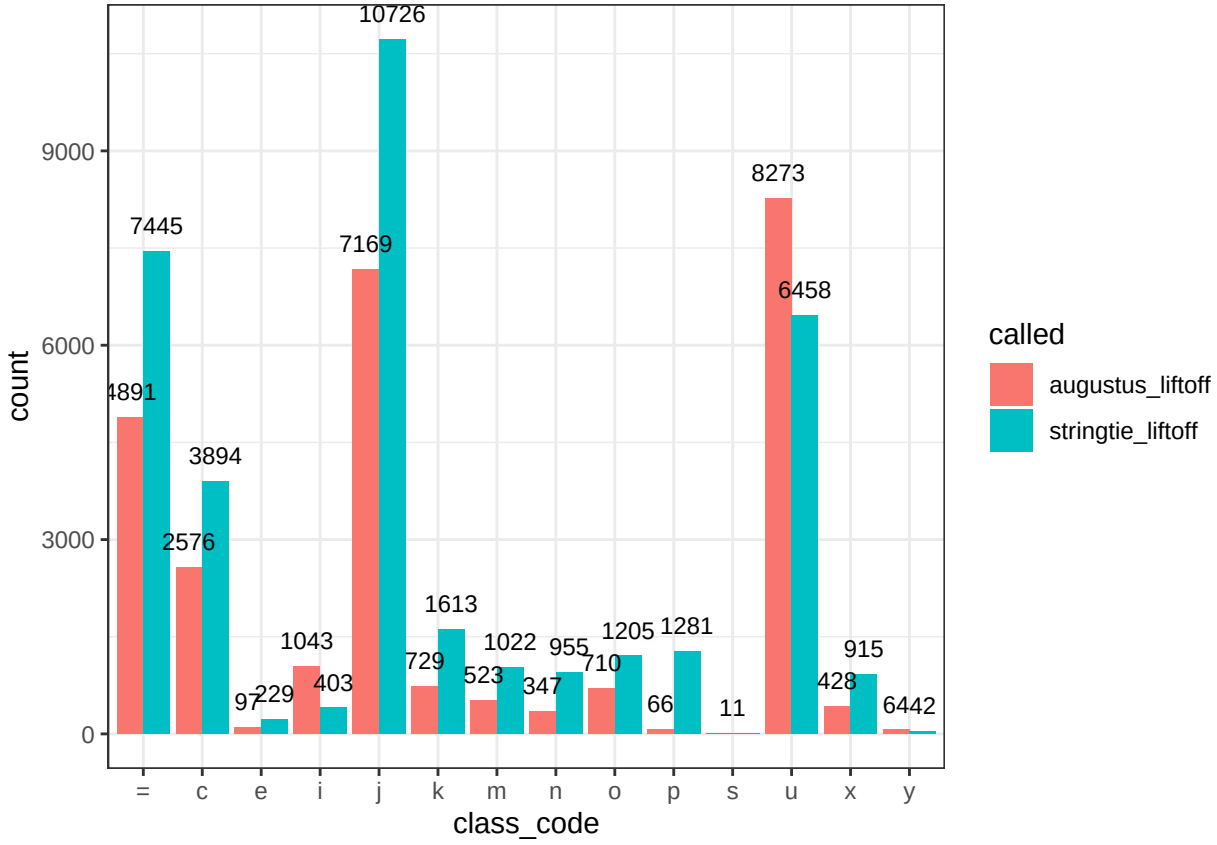
Unmapped features

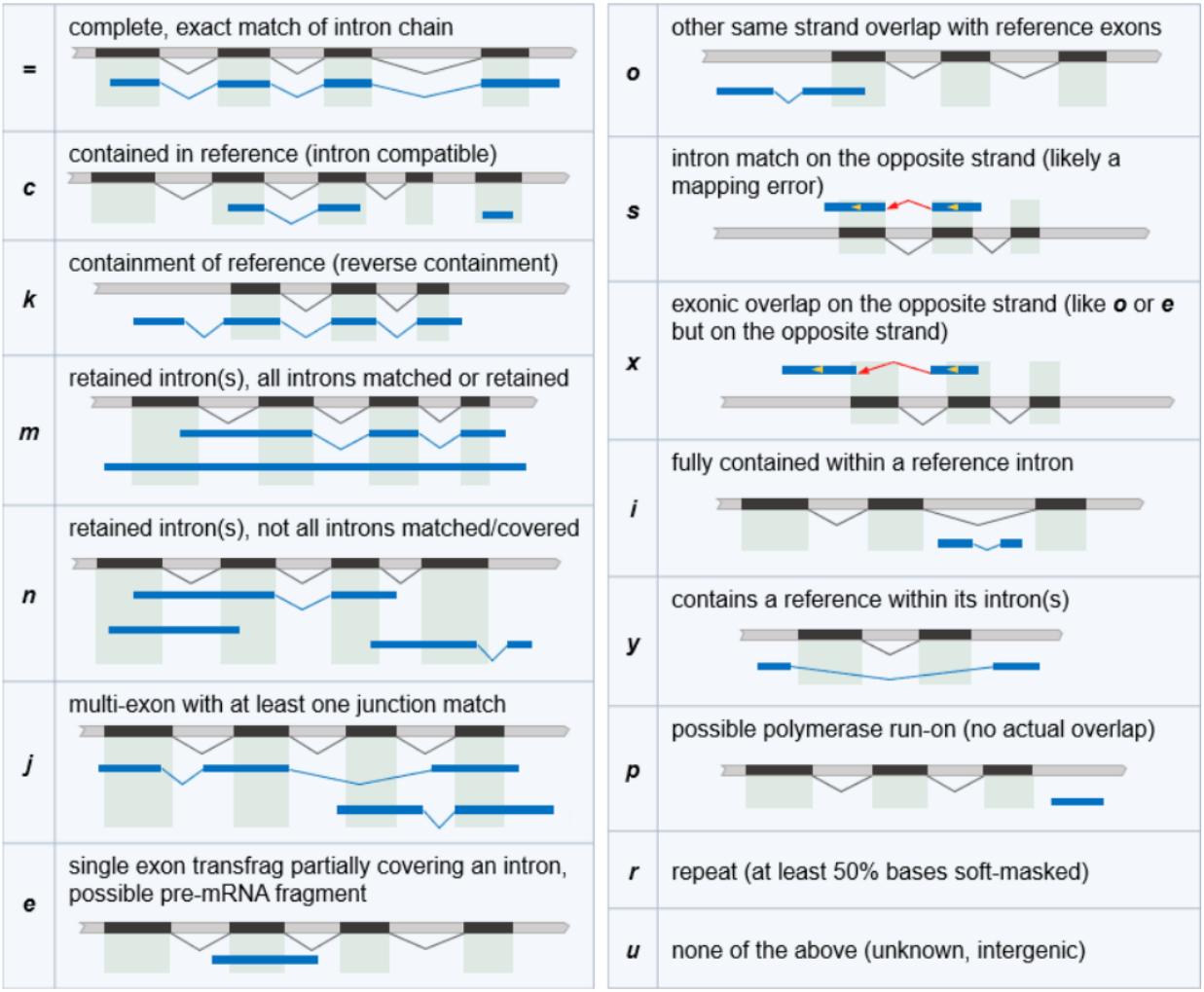
Overall the unmapped transcripts were shorter than the mapped transcripts and the contig sizes they were on in the original assembly were smaller. This leads us to believe they were not full length transcripts and were fragmented in some way.



Compare liftoff and augustus to Stringtie2 assemblies from RNA-seq

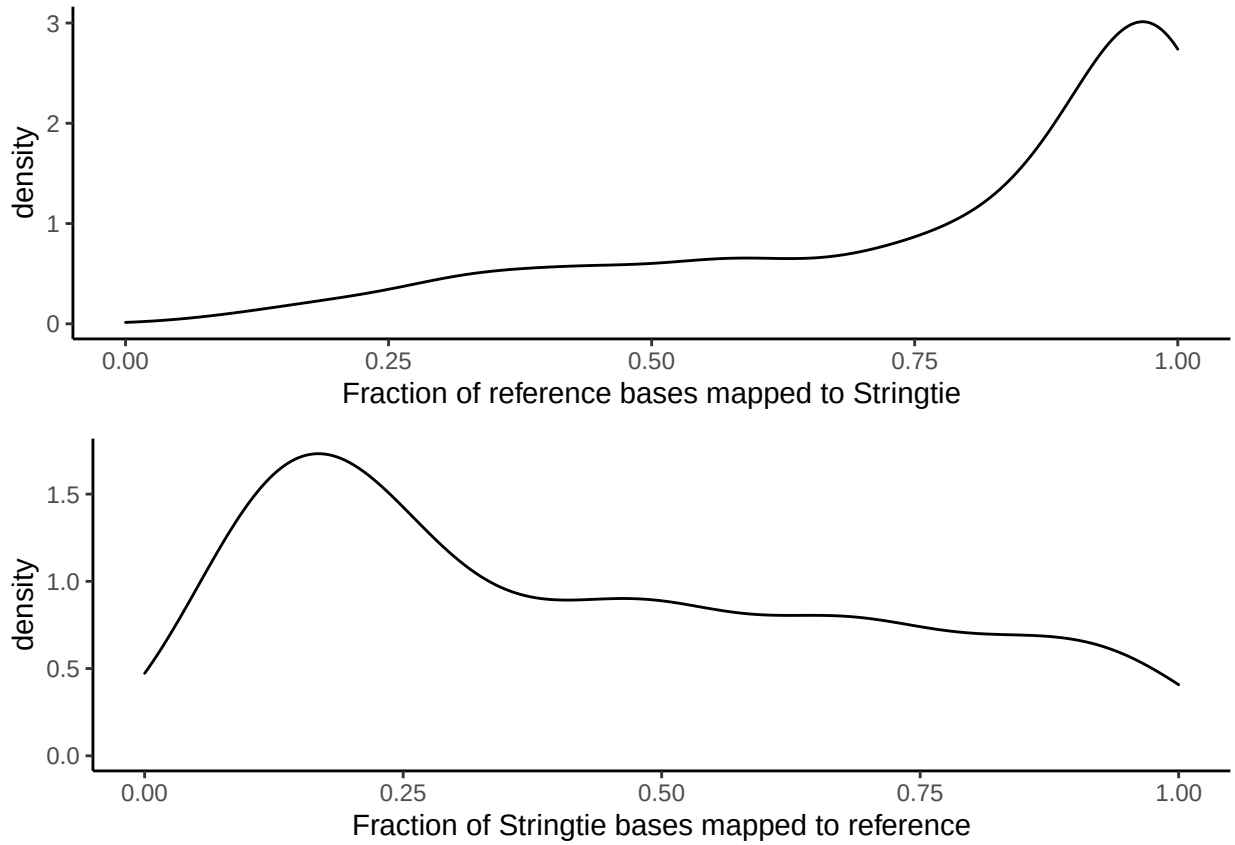
Liftoff annotations match RNA-seq data better than augustus. If Liftoff transcript is contained within the Stringtie assembled transcript (k) include Stringtie coordinates in final annotation with liftoff functional annotation. Look more into the “u”, stringtie assembled transcripts that are not in the liftoff annotations





Look into features that Liftoff could not map (n=482)

1. Take unmapped features from reference transcript sequences
2. Align Stringtie2 assembled transcripts to the unmapped features
3. Filter for MAPQ = 60



545 transcripts map to 297 of the unmapped features. By plotting fraction of the length that mapped we can tell that the unmapped features are fragmented because a majority of the sequence maps to a portion of the Stringtie transcript.

Look into the Stringtie assembled transcripts that are not in the liftoff annotations

1. Filter the Stringtie transcripts that are “u” when compared to liftoff transcripts
2. Compare the augustus gene predictions with the Stringtie “u” transcripts
3. 794 Stringtie “u” transcripts overlap exactly with augustus gene predictions

total feature count

from liftoff: 39957

from Stringtie mapped to unmapped reference features: 545

from Stringtie/augustus exact overlaps not in reference: 794

Total features: 41296

Total features in old asm: 41110