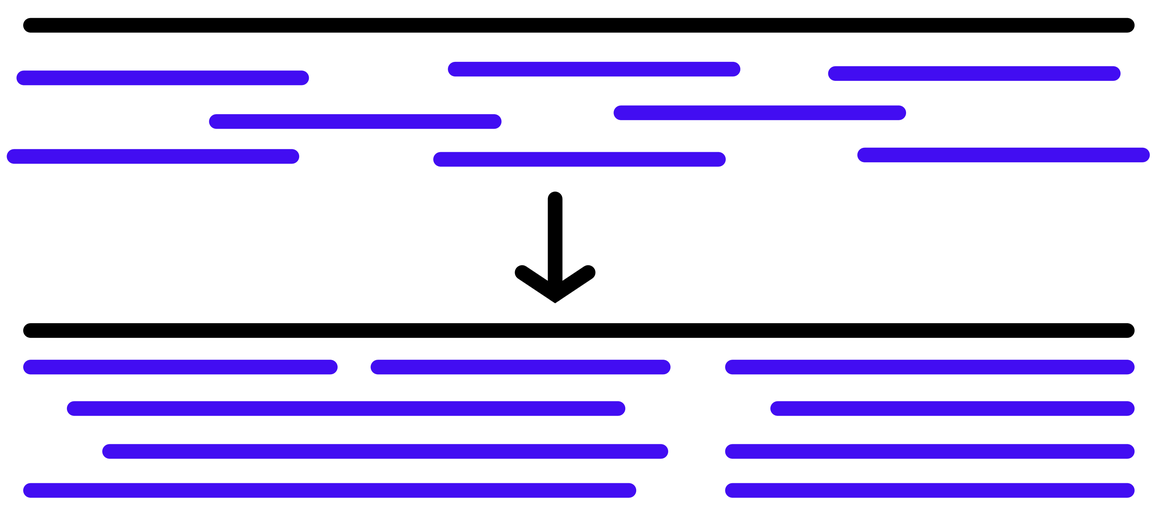
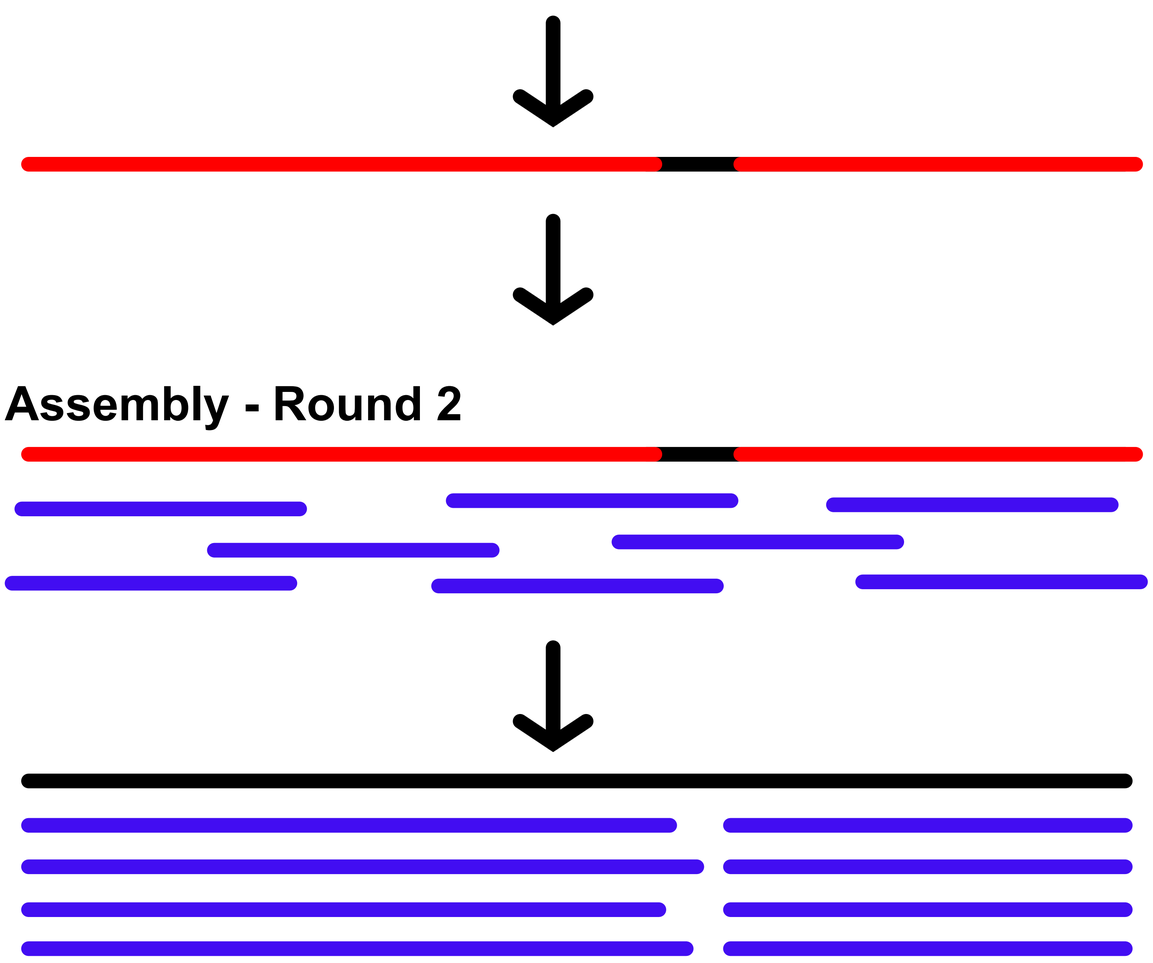
Protein-Coding Sequence Assembly

Assembly - Round 1



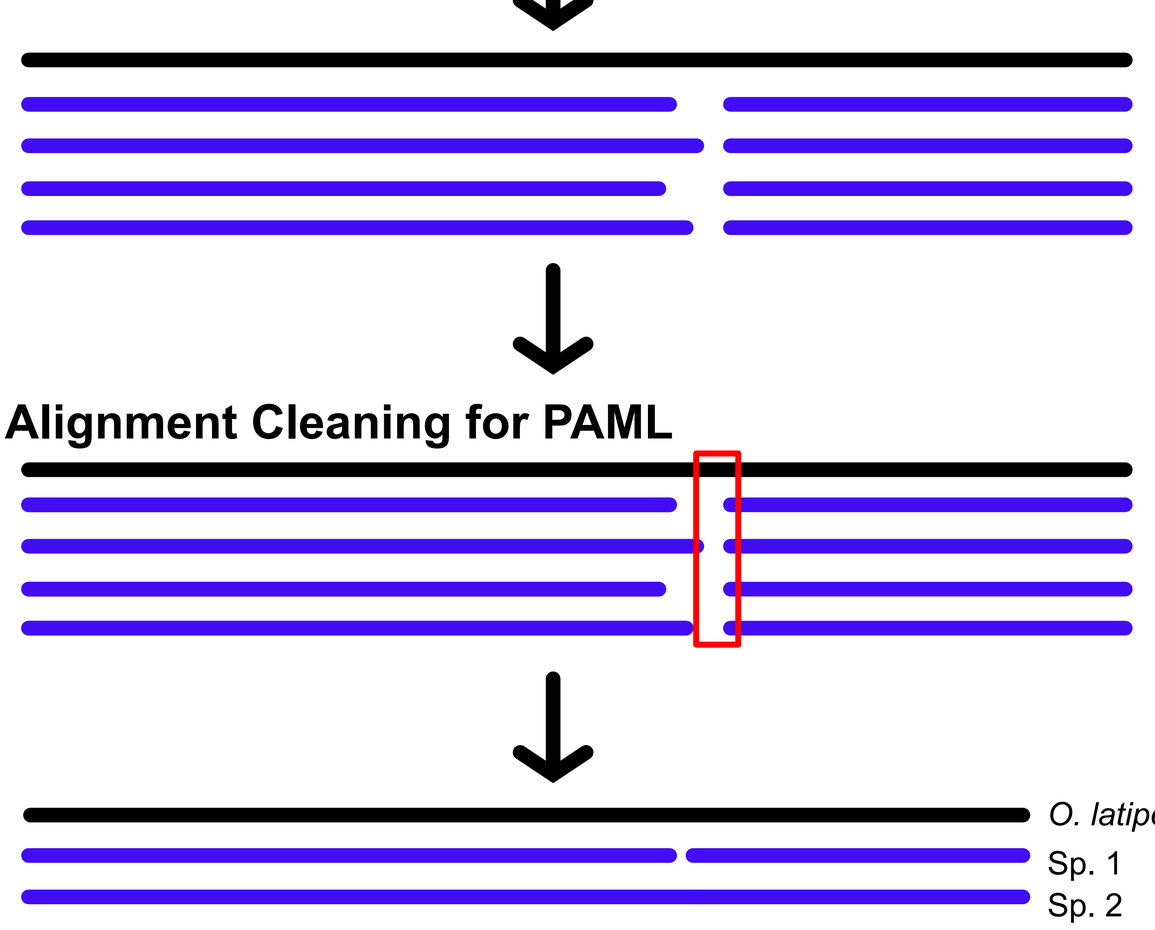
***Oryzias latipes* reference CDS**

Align raw reads to reference sequence using BWA-MEM.



Multiple sequence alignment (gapped)

Generate consensus sequences for each species. Gaps present due to poor assembly in some regions of the CDS.



Generate new reference CDS

New reference sequence is based on most complete CDS assembled for a species. Gaps are filled in with sequence from *Oryzias latipes* to avoid introducing gaps into the new alignment.

Assembly - Round 2

Align raw reads to new CDS

Align raw reads to new "makeshift" reference sequence.

Multiple sequence alignment (less gapped)

New MSA is less gapped, the new reference sequence is replaced by the original CDS of *Oryzias latipes* to obtain a complete MSA

Alignment Cleaning for PAML

Clean multiple sequence alignment

If 30%+ of a codon position has gaps, or 10%+ of premature stop codons, those sites are removed to ensure gaps/stop codons do not interfere with PAML analyses.

Final multiple sequence alignment

Final, edited MSA has fewer gaps for each species.