TOGA: integrating gene annotation with orthology inference at scale

Michael Hiller
LOEWE Centre for Translational Biodiversity Genomics, Senckenberg Society for Nature Research & Goethe University, Frankfurt am Main, Germany

Annotating coding genes in newly-sequenced genomes and inferring orthologs are a classical challenges in bioinformatics. Current methods for orthology inference either cluster genes based on sequence similarity and/or determine speciation and duplication nodes in a gene tree. These methods require sets of annotated genes as input, which is why gene annotation typically precedes orthology inference.

In the first part of the talk, I will present TOGA (Tool to infer Orthologs from Genome Alignments), a new method that integrates gene annotation and inferring orthologous genes. TOGA implements a novel paradigm to infer orthologous gene loci that largely relies on intronic and intergenic alignments and uses machine learning to accurately distinguish orthologs from paralogs or processed pseudogenes. TOGA improves ortholog detection and annotation completeness, and handles even highly-fragmented assemblies. As a reference-based method, TOGA is scalable and can be applied to genomes of hundreds of species. TOGA also detects gene losses and facilitates the generation of more accurate codon alignments to screen for selection patterns. With a set of ancestral genes, TOGA can also provide superior measures of genome assembly quality.

In the second part, I will illustrate how TOGA can help to illuminate changes in genes that are linked to interesting phenotypic traits. Using a new haplotype-resolved assembly of the common vampire bat, we uncovered several genes that are specifically lost in vampire bats, which provides insights into the genomic adaptations to blood feeding.