HiMAP2 tutorial

In the tutorial below, we have provided a step-by-step descriptions of commands and outputs (prefixed with "###") from the HiMAP2 installation to exon selection and visualization:

```
### Download and extract HiMAP2 repo from github.
git clone https://github.com/popphylotools/HiMAP_v2.git
mv HiMAP_v2 HiMAP2
find HiMAP2 -type f \( -name "*.py" -o -name "*.sh" \) -exec chmod +x {} \;
or
chmod -R +x HiMAP2
cd HiMAP2

### Optional: if you plan to generate species tree from exon trees, install ASTRAL and specify full path to the ASTRAL jar file in config.ini ("astral_path" setting in the config.ini). Check https://github.com/smirarab/ASTRAL/ for more information about installation and usage.
wget https://github.com/smirarab/ASTRAL/raw/master/Astral.5.7.8.zip
unzip Astral.5.7.8.zip

### Create HiMAP2 conda environment from the environment.yml file for linux or Mac OSX operating systems.
./create_anaconda_environments.sh

### Activate HiMAP2 environment.
conda activate himap2
```

```
### Set up directory system by running initialization script.
./bin/00_initialize_working_dir.py

### Provide initial input for the pipeline by copying dataset fasta and gff
files from ./toy_data to ./data/00_fasta and ./data/00_gff, respectively.
cp ./toy_data/*.fasta ./data/00_fasta
cp ./toy_data/*.gff ./data/00_gff

### Check specifications in the config.ini file. The config.ini file provided
in the HiMAP2 package is by default set up to run the pipeline for the toy
dataset so no changes are necessary to test the pipeline, except for setting
the path of ASTRAL as mentioned above. Note, relative paths to data
directories assume that your current directory is HiMAP2 (cd HiMAP2).
nano ./config.ini

### Detailed information about the settings and custom run configuration in
the configuration file can be found in Appendix 2.

### Run pipeline starting with the 01_sequence_extraction.py script that uses
provided fasta and gff files to extract coding sequences, select longest
isoforms, and prepare nucleotide and peptide sequences for ortholog
prediction.
./bin/01_sequence_extraction.py

### Once this step is done, you will see message:
###     Step_01 is complete
###     Processed  10 input fasta/gff files
### Check content of the directories ./data/01_gff_db, ./data/01_pep_fasta
and ./data/01_nuc_fasta. They should now contain database and sequence files
```

```
generated for each taxon. These files are the input for the next step of the
pipeline.

### Run the 02a_orthofinder.py script to conduct ortholog prediction using
OrthoFinder v2.5.4.
./bin/02a_orthofinder.py

### Output from this step are full OrthoFinder results written to the
./data/02_orthofinder directory. Once this step is complete, double-check
that the folder now also contains Orthogroups.tsv that is used as input for
ortholog filtering in the next step of the pipeline.

### Perform ortholog filtering by running the 02b_ortho_selection.py script.
This script uses specifications in the configuration file to filter orthologs
predicted by OrthoFinder and outputs orthologs passing the filters to the
./data/02_orthofinder/keep_orthos.csv file. Default settings provided in the
config.ini file filter single-copy orthologs present in all core taxa and in
at least three supplemental taxa.
./bin/02b_ortho_selection.py

### Generate final filtered exon sequences using the
03_alignments_and_filtering.py script. This script will first generate
ortholog alignments for the core species (written to the
./data/03a_core_alignment directory). Next, these alignments are split by
exon and sequences from the supplemental taxa are added and aligned to the
core sequences. The intermediate results are written to the
./data/03b_supplementary_alignment directory. Finally, aligned exon sequences
are filtered according to the specifications in the configuration file and
final filtered exons are output to the ./data/03c_final_exons directory.
```

```
Outgroup sequences can be added to core or supplemental alignments or both by
editing configuration file (see Appendix 2 for details).
./bin/03_alignments_and_filtering.py

### Generate exon trees for the final filtered alignments using RAxML-NG.
./bin/04_exon_phylogeny.py

### This script determines the best-fit substitution model using ModelTest-NG
and reconstruct a maximum likelihood phylogeny using RAxML-NG. Output from
this step is written to the ./data/04_exon_phylogeny directory with results
for each final filtered exon alignment placed in subdirectories with the
corresponding names. The final exon trees with bootstrap support values are
written to the files with the suffix *.raxml.support.

### Infer a species tree from exon trees using ASTRAL.
./bin/05_speciestree.py

### Output from this step is written to the ./data/05_speciestree directory
and includes the speciestree_quartetsupport.tre file which contains the
species tree with quartet branch support values.

### Visualize results of the exon selection pipeline and further refine your
selection.
bokeh serve --show HiMAP2_viz

### The visualization application requires final filtered exon alignments in
the ./data/03c_final_exons directory. This minimum input will produce a
heatmap displaying exon alignments colored and organized by the amount of
missing taxa and sequence similarity (second panel from the left in Figure
```

2). The multiselect taxon panel (left-most panel in Figure 2) allows users to pick a subset of taxa to generate a heatmap of the exon alignments. If exon phylogenies from Step 04 are available, the visualization app can also produce a MDS plot for the selected groups of exons in the heatmap (second panel from the right in Figure 2). Individual exon tree points in the MDS plot are colored according to the average bootstrap support of that exon tree while diameter of the point is scaled relative to the number of taxa in the exon alignment. When groups or individual exon tree points are selected in the MDS plot, a species tree is inferred from the selected exon trees and displayed in the right-most tree view panel. The species tree can be rooted to a user-specified taxon or group of taxa using the multiselect tool below the tree view panel. For the exons selected in the heatmap and/or MDS plot, summary information as well as fasta files can be exported using the export buttons below the corresponding plots.