Readme: Pipeline used in "Targeted NGS for species level phylogenomics:
"made to measure" or "one size fits all"? (
Kadlec, M., Bellstedt, D. U., Le Maitre, N. C., and Pirie, M. D. (2017). Peer J Prepr. doi:10.7287/peerj.preprints.2763v1.)

The pipeline consists of 2 scripts: AllMarkers.py and BestMarkers.py

AllMarkers.py (version 1.5)

This script selects single copy sequences suitable for hybrid capture probe design from two or more transcriptomes
by comparing sequences either with Whole Genome Sequence (WGS) data or with a database of single copy genes in flowering plants (De Smet et al, 2013, doi:10.1073/pnas.1300127110) and can be used with python 2 and python 3

Input: two or more transcriptomes in fasta format in a directory. Optional: WGS data.
Output: selected sequences for each transcriptome in fasta format + homologues in WGS or database data

Command line examples:

Show help
./AllMarkers.py -h

1: Select sequences with a minimum length of 1000 bp from 4 transcriptomes, using WGS data for copy number status,
and using 6 CPUs
Command line:
./AllMarkers.py -Nbt 4 -np 6 -l 1000 TranscriptomesDirectory WGSData nucl
Output: In Transcriptome Directory, 6 directories:
Blast: Blast files produced by comparing transcriptomes
BlastWGS: Blast files produced by comparing transcriptomes and WGS data
Databases: Databases used for BLAST
SeqAboveFilterLength: Sequences in transcriptomes above 1000 bp
TranscriptomeHomologues: Homologous sequences found in Transcriptomes
Final_AllSelectedSequences: Selected sequences (for hybrid capture probe design)

2: Select sequences in from 2 or more transcriptomes using WGS data using 4 CPUs of your computer
Commandline:
./AllMarkers.py -Nbt 2 -np 4 TranscriptomesDirectory WGSData nucl
Output: In Transcriptome Directory, 5 directories :
Blast: Blast files produced by comparing transcriptomes
BlastWGS: Blast files produced by comparing transcriptomes and WGS data
Databases: Databases used for BLAST
TranscriptomeHomologues: Homologous sequences found in Transcriptomes
Final_AllSelectedSequences: Selected sequences for target

capture

BestMarkers.py (version 1.0)

BestMarkers.py selects best markers to be used in hybrid capture for phylogenetic analysis from a pool of single copy markers (such as produced using AllMarkers.py; file in fasta format) by optimising either sequence length (with or without introns),or similarity given a predefined number of probes, probe-length and coverage.

To select the best markers according to similarity, the script needs a multiresult blast file in xml format.

Command line examples:

Show help

./BestMarkers.py -h

Select longest sequence markers >1200 bp that might be captured with 6000 unique baits (length: 120 bp, coverage: x4), assuming an arbitrary length of 200 bp for introns

Command line:

./BestMarkers.py -l 1200 -intron? Y -blast_file BLASTFILE -Length_intron 200 -Number_baits 6000 -Length_baits 120 -Coverage_baits 4 -Coverage_baits 4 Sequences.fasta

Output: 4 files

- Summary.txt
- AllCandidatesInfo.txt (exon length, predicted length including introns, percentage identity per candidate marker)
- Candidates.fasta: Candidate markers (complete sequences) in fasta format
- CandidateCodingRegionSe.fasta: Candidate marker coding regions in fasta format