**Pipeline for finding contaminant transcripts**

Commands used for the selection of plant contaminants transcripts from *T. diversipes* transcriptome. The same commands were used for contaminants from generation 1 and 2. **#** - Indicate comment lines; **>** - indicate commands given in R; **$** - indicates commands in bash.

***1-*** *Selection of possible contaminant transcripts using the taxonomy of the best blast hit from the annotation data table.*

> geral.annotation <- read.delim("Annocript/output/Td\_allx50-trans\_seqs\_uniref\_2015\_09\_filt\_ann\_out.txt", header = TRUE)

> uniref\_species <- geral.annotation$Taxonomy

> write.table(uniref\_species, "~/uniref\_blast\_species.txt", sep = "\n", row.names=FALSE,col.names=FALSE)

$ python single\_entry.py -i uniref\_blast\_id2.txt -o uniref\_blast\_id.txt

$ sort uniref\_blast\_id.txt > contaminant\_species.txt

$ python Find\_Contaminant.py -i contaminant\_species.txt –t Td\_Ads\_transcriptome.Annocript/output/Td\_AdG1xG2\_def-trans\_uniref\_2015\_08\_filt\_ann\_out.txt -o contaminant\_transcripts.txt

# This will generate the “contaminant\_transcripts.txt” output, which is a list with the transcripts ID of all the possible contaminants to be used in the next step.

***2-*** *Double check the list of contaminants based on the gene/ isoform classification.*

$ python single\_transcriptInCluster.py -i contaminantG1\_transcripts.txt -c Td\_AdG1-clusters.txt -o multiple\_iso.contaminatsG1.tx

# Manually check each remaining transcript ID from genes with multiple isoforms to verify possible mistakes in contaminant identification. Keep only the transcripts from genes where all the isoforms blasted against plants and create a new list with the ID of the contaminants transcripts (“contaminantsG1\_ids.txt”).

$ sort contaminantsG1\_ids.txt > contaminantsG1\_final.txt

$ uniq contaminantsG1\_final.txt > contaminatsG1\_ids\_final.txt

# This will generate the “contaminatsG1\_ids\_final.txt” output, which is the final ID list of all the contaminant transcripts.

***3-*** *Create a fasta file with all the contaminant transcripts.*

$ python FastaChomper.py -i -f Td\_AdG1-trans.fasta -l contaminatsG1\_ids\_final.txt -o Td\_AdG1\_plantcontaminants