Gene family expansion and contraction analysis

Overview

This document is the pipeline to do gene family expansion and contraction analysis. Before you follow the pipeline, the **MOST IMPORTANT** thing is to choose the species you want to include, because the classfication of orthogroups is always changing if you use differnt data set. Here are some tips to guide you to choose species:
https://davidemms.github.io/orthofinder_tutorials/orthofinder-best-practices.html

If you want to see the function of interseted genefamily, it's better to included Arabidopsis thaliana in your analysis, so that you could check the function of the gene/genefamily easily in TAIR: https://www.arabidopsis.org.

And also here are some blogs about how to do gene family expansion and contraction analysis:

https://www.jianshu.com/p/8c6ef557cc71

https://www.jianshu.com/p/dc75116b0099

http://www.chenlianfu.com/?p=2974

1. Preparing the pep files

Delete invalid transcripts, which include stop condon, lack start codon;

Delete transcripts less than 50 AA (150bp);

Only keep the primary transcripts;

run_pipeline_orthofinder.sh Bash 复制代码

```
#1. extract valid transcripts,filter following transcripts:
##Genes with internal stop codons(-g -V),lack start or end codon(-J);
##Genes which length less than 50 amino acids(-l 150);
##discard redundant transcripts(-M -K -Q)
##remain coding only(-C)

:<<!
for i in `cat list_13sp`
do
gffread -g $i.fa -y $i.ext.merge.pep -V -M -K -J -Q -C -l 150 $i.gff3 &
gffread -g $i.fa -x $i.ext.merge.cds -V -M -K -J -Q -C -l 150 $i.gff3 &
done
!

#:<<!
#2. remain the primary (longest) transcripts as representation of the gene
##the script was downloaded from the website
#make sure the geneID in fasta is uniq for each transcript
#if something wrong when running the script, FIRST check your data: fasta, gff

python3 ../extract_primaryTranscript.py Juglans_mandshurica_NFU.ext.merge.cds Juglans_mandshurica_NFU.gff3 Juglans_mandshurica_NFU.ext.cdsL &
python3 ../extract_primaryTranscript.py Juglans_mandshurica_NFU.ext.merge.pep Juglans_mandshurica_NFU.gff3 Juglans_mandshurica_NFU.ext.pepL &
!
```

2. Gene family cluster using Orthofinder

put all protein files from last step into a directory, for example: 0703_PPJF

run_pipeline_orthofinder.sh Bash 复制代码

```
nohup orthofinder -f 0703_PPJF -t 10 >log 2>err &
```

If everything goes well, you may see file **Orthogroups_SingleCopyOrthologues.txt** in the directory named Orthogroups. These single copy genes in file Orthogroups_SingleCopyOrthologues.txt will be used to reconstruct the phylogeny tree in the following analysis.

3. Reconstruct the ultrametric phylogeny tree with time using MCMCTree

The pipeline here is similar to the preparation the input of partitionFinder, just concentrate codon1, codon2 and codon3 of all the single copy genes into three big matrix. It's not definitely proper way to reconstruct the phylogeny, but for MCMCtree, it's no need to analyse each gene seperately, please search for detail in this blog if you want: http://www.chenlianfu.com/?p=2974

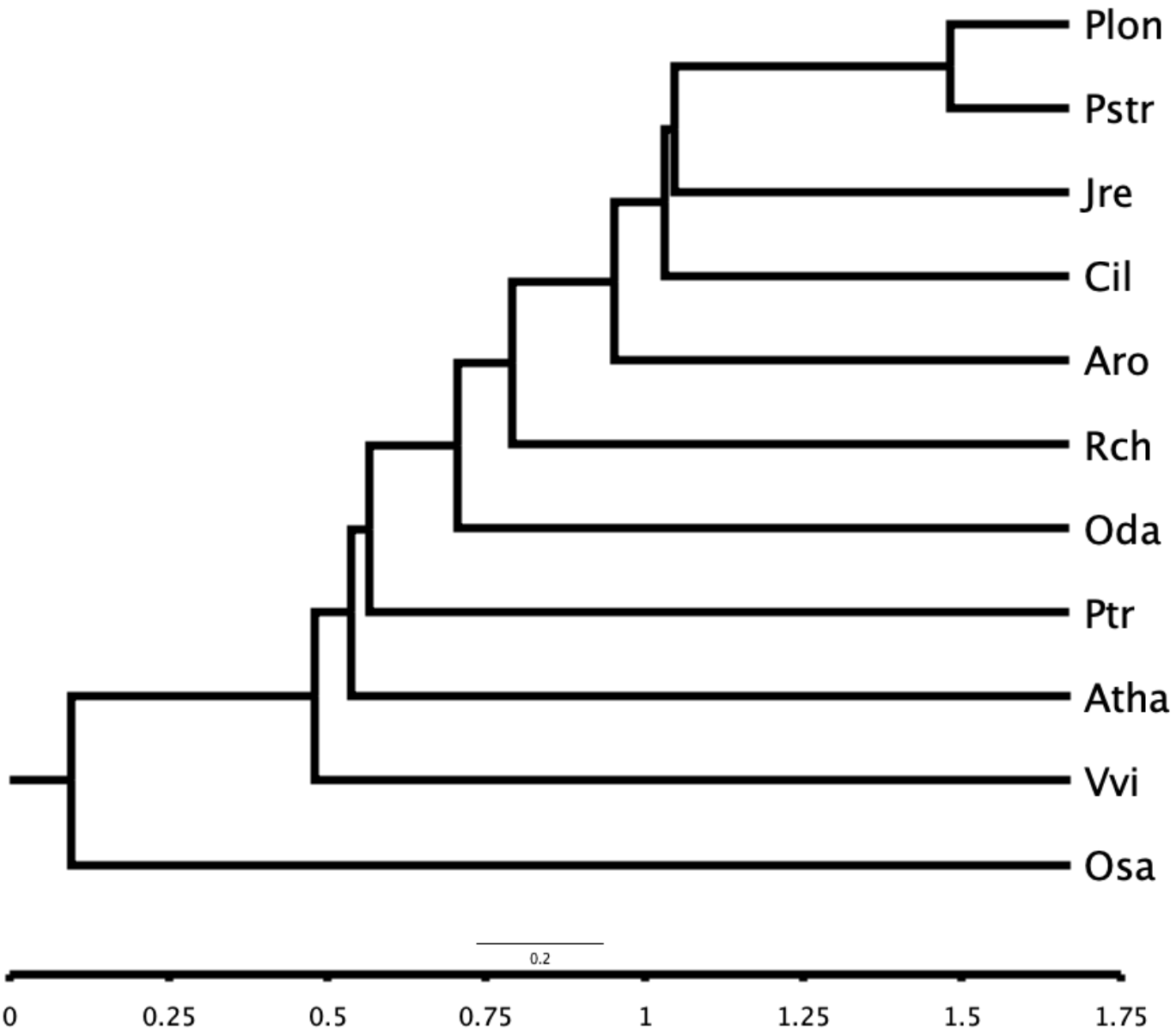
run_pipeline_MCMCtree.sh Bash 复制代码

```
#0. modify the format of cds's name
#Mru:
#sed -r "/>/s/>(\S+) \[.*\] \[.*\] \[.*\] \[.*\]/>\1/g" Mru.cds >Mru.cds.1
#sed -r "/>/s/(\S+)_cds_(\S+)_(\S+)/>\2/" Mru.cds.1 >Mru.cds.2
:<<!
ln -s /data/data/Juglandaceae/Platycarya/13_geneExCon/genome/*.ext.merge.cds ./
sed -r 's/rna-gnl\|WGS:JAEDWW\|//g' Cil.ext.merge.cds >Cil.ext.merge.cds.2
sed -r 's/rna\-gnl\|WGS\:RXIC\|mrna\.//g' Mru.ext.merge.cds >Mru.ext.merge.cds.2
!

#step 0 1 2
:<<!
for i in `cat list_11sp.index`
do
  l=`echo ${i%_*}`
  sp=`echo ${i#*_}`
  colnu=$[$l+1]
  #0. get the orthlist of each sp
# cut -f $colnu ./0_orthlist/Orthogroups_SingleCopyOrthologues.txt.2 |sed '1d' >./0_orthlist/orthlist_$sp
  #1. get singlecopy orthologues of each sp
# awk '/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0"\n":$0 }' ./1_genefa/$sp".ext.merge.cds" >./1_genefa/$sp".cds1"
# grep -A 1 -f ./0_orthlist/orthlist_$sp ./1_genefa/$sp".cds1" -w | sed "/^--$/d" > ./1_genefa/$sp".orth.cds"
  #2. modify name of cds to orthfroup's name
# perl 2_substitude_orthname.pl ./0_orthlist/Orthogroups_SingleCopyOrthologues.txt.2 $l ./1_genefa/$sp".orth.cds" &
done
!

#3. genefa to indfa
#perl 3_genefa2indfa.pl 0_orthlist/orthlist_OG ../list_11sp 1_genefa 2_indfa &

:<<!
#step 45 can be substitude by translatorx!!!
for i in `cat ../0_orthlist/orthlist_OG`
do
        translatorx_vLocal.pl -i $i".ind.fa" -o $i -p maFft
done
!

```

```
:<<!
#4. pal2nal
#nohup sh 4.1_mafft.sh orthlist 2_indfa &>log_mafft &
#sh 4.2_pal.sh orthlist 2_indfa
#5. partition
cd 2_indfa
mkdir 5_part_fa
sh ../5_run_codon123.sh ../orthlist
#creat gene123.list by:
cd 5_part_fa
mkdir bak
mv *12.fa bak
ls *.fa >../gene123.list

for i in `ls 5_part_fa/condon3/*.fa`
do
        Gblocks $i -t=DNA
done
!

:<<!
ls *.nt1_ali.fasta >condon1.list
ls *.nt2_ali.fasta >condon2.list
ls *.nt3_ali.fasta >condon3.list
perl ../5_join_part.pl condon1.list . geneposition1.txt align_condon1.phy
0
perl ../5_join_part.pl condon2.list . geneposition2.txt align_condon2.phy
0
perl ../5_join_part.pl condon3.list . geneposition3.txt align_condon3.phy
0
cat align_condon1.phy align_condon2.phy align_condon3.phy >align_condon12
3.phy
!

#7. mcmctree
##input file:   1:phylip format sequence file;
##    2:ctl file
##    3.tree file with fossil node time(CIs)

#mkdir 7_mcmctree
#nohup mcmctree mcmctree.ctl 1>log_mcmc_test_0831 2>err_mcmc_test_0831 &
:<<!
for i in {1..2}
do
  mkdir -p run$i
  cp mcmctree.ctl mcmctree.tree align_condon123.phy run$i/
  cd run$i
  nohup mcmctree mcmctree.ctl 1>log_mcmc_condon123 2>err_mcmc_condon123 &
```

```
    cd ../
done
!
```

Now you get the phylogeny topology, **remember** the branch length of tree in MCMCtree is in the unit of **100Ma**, which means you should transform it into 1Ma in order to match CAFE software.



4. Gene family expansion and contraction analysis using CAFE

run_pipeline_cafe.sh Bash 复制代码

```
#0. prepare the inputfile of cafe
awk 'OFS="\t" {$NF="" ;print $0}' Orthogroups.GeneCount.tsv > cafe.data.1
sed 's/^/null\t&/g' cafe.data.1 >cafe.data
#modify title of cafe.data:Desc Family ID Cil Mru Plon  Pstr  Rch
python cafetutorial_clade_and_size_filter.py -i cafe.data -o cafe.filter.data -s
##. modify cafetutorial_run.sh

#1.run cafe by shell script
nohup cafe cafetutorial_run.sh &
nohup cafe cafetutorial_run_filter.sh &
#2. summarize the result of cafe
#output the rapidly changing families on each nodes
python2 /data/data/Juglandaceae/Platycarya/13_geneExCon/python_scripts/cafetutorial_report_analysis.py -i resultfile.cafe -o summary_cafe_rapidChange >log_summary_cafe_rapidChange &
python2 /data/data/Juglandaceae/Platycarya/13_geneExCon/python_scripts/cafetutorial_report_analysis.py -i resultfile.largefilter.cafe -o summary_cafe_largefilter_rapidChange >log_summary_cafe_largefilter_rapidChange &
#ouput all changing families on each nodes
python2 /data/data/Juglandaceae/Platycarya/13_geneExCon/python_scripts/cafetutorial_report_analysis.py -i resultfile.cafe -o summary_cafe_allChange -r 0 >log_summary_cafe_allChange &
python2 /data/data/Juglandaceae/Platycarya/13_geneExCon/python_scripts/cafetutorial_report_analysis.py -i resultfile.largefilter.cafe -o summary_cafe_largefilter_allChange >log_summary_cafe_largefilter_allChange &

#3.1 visualize through cafe home script, the figures are quite ugly though :C
python3.6 /data/data/Juglandaceae/Platycarya/13_geneExCon/python_scripts/cafetutorial_draw_tree.py -i summary_cafe_rapidChange_node.txt -t '((((Plon:2,Pstr:2):62,Cil:64):23,Rch:87):10,Odav:97)' -d '((((Plon<0>,Pstr<2>)<1>,Cil<4>)<3>,Rch<6>)<5>,Odav<8>)<7>' -o summary_cafe_rapidChange_node_expand.png
python3.6 /data/data/Juglandaceae/Platycarya/13_geneExCon/python_scripts/cafetutorial_draw_tree.py -i summary_cafe_rapidChange_node.txt -t '((((Plon:2,Pstr:2):62,Cil:64):23,Rch:87):10,Odav:97)' -d '((((Plon<0>,Pstr<2>)<1>,Cil<4>)<3>,Rch<6>)<5>,Odav<8>)<7>' -y Contractions -o summary_cafe_rapidChange_node_contract.png

#3.2 visualize through cafe_fig
python3.6 ~/software/CAFE_fig-master/CAFE_fig.py resultfile.cafe -pb 0.01 -pf 0.01 --dump test/ -g pdf --count_all_expansions

#4. get orthologues from output of cafe
```

```
#rapidly changing OGs' list
sed -n '7p' summary_cafe_rapidChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "+"|cut -c1-9 >signif_expandOG_inPS.list
sed -n '6p' summary_cafe_rapidChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "+"|cut -c1-9 >signif_expandOG_inPL.list
sed -n '7p' summary_cafe_rapidChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "-"|cut -c1-9 >signif_contractOG_inPS.list
sed -n '6p' summary_cafe_rapidChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "-"|cut -c1-9 >signif_contractOG_inPL.list

#all changing OGs' list
sed -n '7p' summary_cafe_allChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "+"|cut -c1-9 >all_expandOG_inPS.list
sed -n '6p' summary_cafe_allChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "+"|cut -c1-9 >all_expandOG_inPL.list
sed -n '7p' summary_cafe_allChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "-"|cut -c1-9 >all_contractOG_inPS.list
sed -n '6p' summary_cafe_allChange_fams.txt |sed 's/,/\n/g;s/:/\n/g;s/\t//g'|grep "-"|cut -c1-9 >all_contractOG_inPL.list

#reform the OG-gene pair file
python split_with_one_gene.py Orthogroups.txt Orthogroups_genes.txt

#get genelist of changing OGs
grep -f signif_expandOG_inPS.list Orthogroups_genes.txt|grep "Pstr" >signif_expandOG_inPS.OGs.genes
grep -f signif_expandOG_inPL.list Orthogroups_genes.txt|grep "Plon" >signif_expandOG_inPL.OGs.genes
grep -f signif_contractOG_inPS.list Orthogroups_genes.txt|grep "Pstr" >signif_contractOG_inPS.OGs.genes
grep -f signif_contractOG_inPL.list Orthogroups_genes.txt|grep "Plon" >signif_contractOG_inPL.OGs.genes

cut -f2 signif_contractOG_inPL.OGs.genes >signif_contractOG_inPL.genes
cut -f2 signif_contractOG_inPS.OGs.genes >signif_contractOG_inPS.genes
cut -f2 signif_expandOG_inPL.OGs.genes >signif_expandOG_inPL.genes
cut -f2 signif_expandOG_inPS.OGs.genes >signif_expandOG_inPS.genes

grep -f all_expandOG_inPS.list Orthogroups_genes.txt|grep "Pstr" >all_expandOG_inPS.OGs.genes
grep -f all_expandOG_inPL.list Orthogroups_genes.txt|grep "Plon" >all_expandOG_inPL.OGs.genes
grep -f all_contractOG_inPS.list Orthogroups_genes.txt|grep "Pstr" >all_contractOG_inPS.OGs.genes
grep -f all_contractOG_inPL.list Orthogroups_genes.txt|grep "Plon" >all_contractOG_inPL.OGs.genes

cut -f2 all_contractOG_inPL.OGs.genes >all_contractOG_inPL.genes
```

```
cut -f2 all_contractOG_inPS.OGs.genes >all_contractOG_inPS.genes
cut -f2 all_expandOG_inPL.OGs.genes >all_expandOG_inPL.genes
cut -f2 all_expandOG_inPS.OGs.genes >all_expandOG_inPS.genes

```

As for the visualization of the output, the best way is to add the number of expand/contract genefamily (or genes included) to the phylogeny by yourself.

GenefamilyExpCon_11sps.2.pdf

And you can also explore the function of genefamilies which are expand in specific taxon (taxa), by doing GO enrichment or KEGG enrichment, or both.