* Upset plot for gene family

library(UpSetR)

library(tidyverse)

gene\_families <- read.table("top10.Orthogroups.binary",header = T, check.names = F)

gene\_single <- read.table("Orthogroups\_UnassignedGenes.tsv.binary",header = T,check.names = F)

gene\_all <- bind\_rows(gene\_families, gene\_single)

upset(gene\_all,sets = c('A-grandis','B-buxifolia','C-ichangensis','D-reticulata','E-fortunella','F-sinensis','G-medica','H-clementina','I-satsuma'),order.by = 'degree', keep.order = T,main.bar.color = '#259FDA', sets.bar.color = '#F4BD23', text.scale = c(1.2, 1.1, 1.2, 1.1, 1.2, 1.1), mainbar.y.label = "Number of Gene Family", sets.x.label = 'Number of Gene Family')

* PAVs back genome

/public/home/tahir/citrus-ppsPCP-pan/test\_back\_pavs

cat murraya.fa.fai | cut -f 1,2 > murraya.chr.length

shuf -n 100 murraya\_final.bed | awk 'BEGIN{n=1}{print $1"\t"$2"\t"$3"\t"$4"\_"n;n++}' > murraya\_random.pavs

bedtools flank -i murraya\_random.pavs -g murraya.chr.length -l 150 -r 150 | awk '{if(NR%2==1){print $0 > "murraya\_pavs\_5.bed"}else{print $0 > "murraya\_pavs\_3.bed"}}'

bedtools getfasta -fi murraya.fasta -bed murraya\_pavs\_5.bed -name > murraya\_pavs\_5.fa

bedtools getfasta -fi murraya.fasta -bed murraya\_pavs\_3.bed -name > murraya\_pavs\_3.fa

blat grandis.fa murraya\_pavs\_5.fa -t=dna -q=dna -minIdentity=80 murraya\_pavs\_5.psl

perl generate\_pair.pl region\_300.align.1.bed region\_300.align.2.bed | perl find\_pair.pl - > murraya\_300\_script.pav.location.txt

./run.sh grandis species.list &> run.log

cat \*\_on\_grandis.bed | sortBed | bedtools coverage -a ../pav\_back\_murraya2grandis/grandis.100Kb.bed -b - | cut -f 1-4 > grandis\_plot\_100Kb.middle.txt

In R

>ggplot(d\_low,aes(x=V2,y=V5,fill=PAVs)) + geom\_raster() + facet\_grid(V1~., scales = 'free') + scale\_fill\_viridis(option = 'C') + theme(axis.title.y = element\_blank(), axis.ticks.y = element\_blank(), axis.text.y = element\_blank(), plot.margin = unit(rep(1.5,4),'lines'), panel.background = element\_blank(), panel.grid = element\_blank(), axis.ticks.x = element\_line(size = 1), axis.text.x = element\_text(size = rel(1.2)), strip.text = element\_text(size = rel(1.2)), legend.text = element\_text(size = rel(1.2)), legend.title = element\_text(size = rel(1.5))) + xlab("") + scale\_x\_continuous(breaks = c(seq(0,50000000,5000000)), labels = c(paste(seq(0,50,5),'Mb', sep = '')))

> ci <- read.table("clipboard",header = T)

> ci$Species <- factor(ci$Species, levels = as.vector(ci[1:10,1]))

> ci$Type <- factor(ci$Type, levels = c('High','Mid','Low-1','Low-2','N\_flank','N\_mapped'))

>ggplot(ci,aes(x=Species,y=Num, fill=Type)) + geom\_bar(stat = 'identity', position = 'dodge') + geom\_text(aes(label=Num),position = position\_dodge(0.9),vjust=-0.5, size=3) + scale\_fill\_npg() + theme(plot.margin = unit(rep(1.5,4),'lines'), panel.background = element\_blank(),panel.grid = element\_blank(),axis.line = element\_line(size = 1),axis.ticks = element\_line(size = 1), axis.title = element\_text(size = rel(1.5)), axis.text = element\_text(size = rel(1.2)), legend.position = c(1,1), legend.justification = c(1,1), axis.title.y = element\_text(vjust = 7), legend.text = element\_text(size = rel(1.2)), legend.title = element\_text(size = rel(1.5))) + xlab("") + ylab("Number of PAVs")

* Gene length

grep '>' ../matrix\_t2/core\_gene.fa | sed 's/>//' | awk '{print $1}' | sort > core\_gene\_t2.list

perl citrus\_get\_shell\_length.pl ../sweet\_orange.gff3 core\_gene\_t2.list core > core\_gene\_t2.length

grep '>' ../matrix\_t2/shell\_gene.fa | sed 's/>//' | awk '{print $1}' | sort > shell\_gene\_t2.list

perl citrus\_get\_shell\_length.pl ../sweet\_orange.gff3 shell\_gene\_t2.list shell > shell\_gene\_t2.length

cat core\_gene\_t2.length shell\_gene\_t2.length > gene\_length.plot

In R

> ggplot(d,aes(x=V1,y=log10(V2),fill=V1)) + geom\_boxplot(width=0.6) + theme\_bw() + xlab("") + ylab("log10(Gene length(bp))") + theme(legend.position = "none",axis.title.y = element\_text(size = 15,vjust = 7),axis.line = element\_line(size = 1),axis.text = element\_text(size = 15),panel.grid = element\_blank(),panel.border = element\_blank(), plot.margin = unit(rep(1.5,4),'lines'), plot.tag = element\_text(size = rel(2))) + scale\_fill\_manual(values = c("#F4BD23","#259FDA")) + labs(tag = 'A')

* Exon number

/public/home/tahir/work/orange\_get\_homo/results/run\_2019-04-04/gene\_cds\_length\_plot

perl citrus\_get\_exon\_number.pl ../sweet\_orange.gff3 core\_gene\_t2.list core > core\_gene\_t2.exon

perl citrus\_get\_exon\_number.pl ../sweet\_orange.gff3 shell\_gene\_t2.list shell > shell\_gene\_t2.exon

cat core\_gene\_t2.exon shell\_gene\_t2.exon > exon\_number.plot

* CDS length

/public/home/tahir/work/orange\_get\_homo/results/run\_2019-04-04/gene\_cds\_length\_plot

perl citrus\_get\_cds\_length.pl ../sweet\_orange\_cds.fa core\_gene\_t2.list core > core\_gene\_t2.cds\_length

perl citrus\_get\_cds\_length.pl ../sweet\_orange\_cds.fa shell\_gene\_t2.list shell > shell\_gene\_t2.cds\_length

cat core\_gene\_t2.cds\_length shell\_gene\_t2.cds\_length > cds\_length.plot

> ggplot(d,aes(x=V1,y=log10(V2),fill=V1)) + geom\_boxplot(width=0.6) + theme\_bw() + xlab("") + ylab("log10(CDS length(bp))") + theme(legend.position = "none",axis.title.y = element\_text(size = 15,vjust = 7),axis.line = element\_line(size = 1),axis.text = element\_text(size = 15),panel.grid = element\_blank(),panel.border = element\_blank(), plot.margin = unit(rep(1.5,4),'lines'), plot.tag = element\_text(size = rel(2))) + scale\_fill\_manual(values = c("#F4BD23","#259FDA")) + labs(tag = 'B')

* Orthofinder

/public/home/tahir/work/orange\_pan\_genome/results/orthoFinder\_2019-05-30/sequences

for i in ../../../data/sequence/\*faa; do ln -s $i; done

bsub -J ortho -n 20 -R span[hosts=1] -o ortho\_2019-05-30.log -q normal "orthofinder -t 20 -a 20 -M msa -f sequences"

less Orthogroups.GeneCount.tsv | awk '/^O/{print }' | cut -f 2-12 | awk '{for(i=1;i<=11;i++){if($i!=0){printf("%d\t",1)}else{printf("%d\t",0)}};printf("\n")}' | sort | uniq -c | less

less Orthogroups.GeneCount.tsv | awk '/^O/{print }' | cut -f 2-12 | awk '{for(i=1;i<=11;i++){if($i!=0){printf("%d\t",1)}else{printf("%d\t",0)}};printf("\n")}' | sort | uniq -c | wc -l

less Orthogroups.GeneCount.tsv | awk '/^O/{print }' | cut -f 2-12 | awk '{for(i=1;i<=11;i++){if($i!=0){printf("%d\t",1)}else{printf("%d\t",0)}};printf("\n")}' | sort | uniq -c | sort -k1,1nr | head

less Orthogroups.GeneCount.tsv | awk '/^O/{print }' | cut -f 2-12 | awk '{for(i=1;i<=11;i++){if($i!=0){printf("%d\t",1)}else{printf("%d\t",0)}};printf("\n")}' | sort | uniq -c | sort -k1,1nr | head >

perl upset\_filter.pl Orthogroups.tsv.binary Top10.list > Top10.Orthogroups.binary

> genes <- read.table("Top10.Orthogroups.binary",header = T, check.names = F)

> upset(genes,sets = c('A-grandis','B-murraya','C-buxifolia','D-trifoliata','E-ichangensis','F-reticulata','G-fortunella','H-sinensis','I-medica','J-clementina','K-satsuma'),order.by = 'freq', keep.order = T)

* GO

/public/home/tahir/work/orange\_pan\_genome/results/go\_enrichment\_2019-06-18

cut -f 1 -d '[' ../gene\_family\_2019-06-12/fam\_details/B-murraya.exp.txt > B-murraya.exp.list

perl get\_All\_from\_orthofinder.pl B-murraya.exp.list ../orthoFinder\_2019-05-30/sequences/OrthoFinder/Results\_May30\_1/Orthogroups/Orthogroups.tsv > B-murraya.exp.gene.list

perl filter\_go.pl All.filter.go B-murraya.exp.gene.list | grep -cv '#' > B-murraya.exp.gene.filter.list

/public/home/tahir/softwares/miniconda3/envs/eggnog/opt/trinity-2.6.6/Analysis/DifferentialExpression/run\_GOseq.pl --genes\_single\_factor B-murraya.exp.gene.filter.list --GO\_assignments All.filter.go --lengths All.filter.length --background All.filter.genes

* get\_homologues-est

/public/home/tahir/work/orange\_get\_homo/results/run\_2019-04-04

pbssubmit -g log\_cluster\_t0 -j t0 -n 1 -p 2 -l "get\_homologues-est.pl -d sequence -M -t 0 -S 75"

pbssubmit -g log\_compare\_t0 -j t0 -l "compare\_clusters.pl -d sequence\_est\_homologues/A-grandislongest\_0taxa\_algOMCL\_e0\_S75\_ -o matrix\_t0 -m"

perl -F'\t' -ane '$r++;for(1 .. @F){$m[$r][$\_]=$F[$\_-1]};$mx=@F;END{for(1 .. $mx){for $t(1 .. $r){print"$m[$t][$\_]\t"}print"\n"}}' matrix\_t1/pangenome\_matrix\_t0.tab

pbssubmit -g log\_classify\_t0 -j t0 -l "parse\_pangenome\_matrix.pl -m matrix\_t0/pangenome\_matrix\_t0.tab -s"

pbssubmit -g log\_classify\_t1 -j t1 -l "parse\_pangenome\_matrix.pl -m matrix\_t1/pangenome\_matrix\_t0.tab -s"

pbssubmit -g log\_classify\_t2 -j t2 -l "parse\_pangenome\_matrix.pl -m matrix\_t2/pangenome\_matrix\_t0.tab -s"

pbssubmit -g log\_cluster\_t2\_c -j t2\_c -n 1 -p 20 -l "get\_homologues-est.pl -d sequence -c -z -M -t 2 -S 75 -n 20"

pbssubmit -g log\_growth\_core\_t2 -j t2 -l "plot\_pancore\_matrix.pl -i sequence\_est\_homologues/core\_genome\_2taxa\_algOMCL\_S75.tab -f core\_both -a growth\_core\_t2"

pbssubmit -g log\_growth\_soft-core\_t2 -j t2 -l "plot\_pancore\_matrix.pl -i sequence\_est\_homologues/soft-core\_genome\_2taxa\_algOMCL\_S75.tab -f core\_both -a growth\_soft-core\_t2"

pbssubmit -g log\_growth\_pan\_t2 -j t2 -l "plot\_pancore\_matrix.pl -i sequence\_est\_homologues/pan\_genome\_2taxa\_algOMCL\_S75.tab -f pan -a growth\_pan\_t2"

pbssubmit -g log\_nuc\_identity\_t2 -j t2 -n 1 -p 20 -l "get\_homologues-est.pl -d sequence -A -t 2 -M -n 20 -S 75"

plot\_matrix\_heatmap.sh -i sequence\_est\_homologues/A-grandislongest\_2taxa\_algOMCL\_e0\_S75\_Avg\_identity.tab -d 2 -t "CDS ANI matrix with t2" -o pdf &> log\_nuc\_identity\_plot\_t2

pbssubmit -g log\_tree\_t2 -j t2 -l "compare\_clusters.pl -o tree\_t2 -m -T -d sequence\_est\_homologues/A-grandislongest\_2taxa\_algOMCL\_e0\_S75\_"

/public/home/tahir/work/orange\_get\_homo/results/run\_2019-04-04/tree\_t2

plot\_matrix\_heatmap.sh -i pangenome\_matrix\_t0.tab -o pdf -r -t "Pangenome with t2" -k "genes per cluster" &> tree\_plot\_t2

/public/home/tahir/work/orange\_get\_homo/results/run\_2019-04-04/matrix\_t2

less pangenome\_matrix\_t0\_\_pangenes\_list.faa | grep '>' | awk '{print $1}' | sed 's/>//' > accessory\_gene.list

perl ../go\_annotation\_2019-06-17/filter\_go.pl All.filter.go accessory\_gene.list | grep -v '#' > accessory\_gene.filter.list

/public/home/tahir/softwares/miniconda3/envs/eggnog/opt/trinity-2.6.6/Analysis/DifferentialExpression/run\_GOseq.pl --genes\_single\_factor accessory\_gene.filter.list --GO\_assignments All.filter.go --lengths All.filter.length --background All.filter.genes

>sp <- read.table("J-clementina.accessory.filter.list.GOseq.enriched.fdr005", header = T,sep = "\t", quote = "", stringsAsFactors = F)

>sp$ratio <- sp$numDEInCat / total\_num

>sp <- sp[order(sp[,c("ratio")],decreasing = T),]

>sp$go\_term <- factor(sp$go\_term, levels = rev(sp$go\_term))

>ggplot(sp,aes(x=go\_term,y=ratio)) + geom\_point(aes(color=over\_represented\_pvalue, size=numDEInCat)) + coord\_flip() + theme\_bw() + scale\_color\_gradient(low = 'red',high = 'blue', guide=guide\_colorbar(reverse=TRUE)) + theme(axis.text = element\_text(size = 12), axis.title = element\_text(size = 15)) + labs(size='Number of Genes', color = 'p-value') + xlab("") + ylab("Gene ratio")

perl ../go\_annotation\_2019-06-17/filter\_go.pl All.filter.go shell\_gene\_t2.list | grep -v '#' > shell\_gene\_t2.filter.list

/public/home/tahir/softwares/miniconda3/envs/eggnog/opt/trinity-2.6.6/Analysis/DifferentialExpression/run\_GOseq.pl --genes\_single\_factor shell\_gene\_t2.filter.list --GO\_assignments All.filter.go --lengths All.filter.length --background All.filter.genes

perl ../go\_annotation\_2019-06-17/filter\_go.pl All.filter.go core\_gene\_t2.list | grep -v '#' > core\_gene\_t2.filter.list

/public/home/tahir/softwares/miniconda3/envs/eggnog/opt/trinity-2.6.6/Analysis/DifferentialExpression/run\_GOseq.pl --genes\_single\_factor core\_gene\_t2.filter.list --GO\_assignments All.filter.go --lengths All.filter.length --background All.filter.genes

* r8s

conda activate r8s

python ../gene\_family\_2019-05-27/cafetutorial\_prep\_r8s.py -i SpeciesTree\_rooted\_node\_labels.txt -s 23905 -p 'D-trifoliata,G-fortunella' -c '9.1' -o r8s\_ctl\_file.txt

r8s -b -f r8s\_ctl\_file.txt > r8s\_tmp.txt

tail -n 1 r8s\_tmp.txt | cut -c 16- > r8s\_ultrametric.txt

* CAFE

awk '/^\t/{print "Desc\tFamily ID"$0}!/^\t/{print "(null)\t"$0}' Orthogroups.GeneCount.tsv | sed 's/\_longest//g' | cut -f 1-13 > cafe.input

cafe.sh

#! cafe

version

date

load -i cafe.input -t 8

tree (((((((I-medica:6.215104,A-grandis:6.215104)N7:0.828895,E-ichangensis:7.043999)N5:0.602952,(((F-reticulata:4.384379,J-clementina:4.384379)N9:1.228121,K-satsuma:5.612500)N8:0.509188,H-sinensis:6.121688)N6:1.525263)N4:0.352253,G-fortunella:7.999203)N3:1.100797,D-trifoliata:9.100000)N2:4.195235,C-buxifolia:13.295235)N1:11.548976,B-murraya:24.844211)N0

lambda -s

report pan\_genome

bsub -J i -n 8 -R span[hosts=1] -o cafe.log -q normal "cafe cafe.sh"

python ../gene\_family\_2019-05-27/cafetutorial\_report\_analysis.py -i pan\_genome.cafe -r 0 -o summary

(((((((I-medica:6.2151,A-grandis:6.2151):0.828895,E-ichangensis:7.044):0.602952,(((F-reticulata:4.38438,J-clementina:4.38438):1.22812,K-satsuma:5.6125):0.509188,H-sinensis:6.12169):1.52526):0.352253,G-fortunella:7.9992):1.1008,D-trifoliata:9.1):4.19524,C-buxifolia:13.2952):11.549,B-murraya:24.8442)

# -d

(((((((I-medica<0>,A-grandis<2>)<1>,E-ichangensis<4>)<3>,(((F-reticulata<6>,J-clementina<8>)<7>,K-satsuma<10>)<9>,H-sinensis<12>)<11>)<5>,G-fortunella<14>)<13>,D-trifoliata<16>)<15>,C-buxifolia<18>)<17>,B-murraya<20>)<19>

python cafetutorial\_draw\_tree.py -i summary\_node.txt -t '(((((((I-medica:6.2151,A-grandis:6.2151):0.828895,E-ichangensis:7.044):0.602952,(((F-reticulata:4.38438,J-clementina:4.38438):1.22812,K-satsuma:5.6125):0.509188,H-sinensis:6.12169):1.52526):0.352253,G-fortunella:7.9992):1.1008,D-trifoliata:9.1):4.19524,C-buxifolia:13.2952):11.549,B-murraya:24.8442)' -d '(((((((I-medica<0>,A-grandis<2>)<1>,E-ichangensis<4>)<3>,(((F-reticulata<6>,J-clementina<8>)<7>,K-satsuma<10>)<9>,H-sinensis<12>)<11>)<5>,G-fortunella<14>)<13>,D-trifoliata<16>)<15>,C-buxifolia<18>)<17>,B-murraya<20>)<19>' -y Expansions -o summary\_tree\_expansions.png

perl parse\_fams.pl ../summary\_fams.txt

* dN/dS

pbssubmit -g run.log -j core -l "sh run.sh"

pbssubmit -g run.log -j shell -l "sh run.sh"

bsub -J core -n 1 -R span[hosts=1] -o core.log -q normal "sh run.sh pangenome\_matrix\_t4\_\_core\_list.txt"

bsub -J cloud -n 1 -R span[hosts=1] -o cloud.log -q normal "sh run.sh pangenome\_matrix\_t4\_\_cloud\_list.txt"

bsub -J shell -n 1 -R span[hosts=1] -o shell.log -q normal "sh run.sh pangenome\_matrix\_t4\_\_shell\_list.txt"

> core <- read.table("dnds.core",stringsAsFactors = F)

> shell <- read.table("dnds.shell",stringsAsFactors = F)

> dnds <- rbind(core,shell)

> dnds$ratio <- dnds$V3/dnds$V2

> dnds\_sub <- subset(dnds,dnds$ratio <= 1.5)

> ggplot(dnds\_sub,aes(x=V1,y=ratio, fill=V1)) + geom\_boxplot() + scale\_fill\_manual(values = c("#F4BD23","#259FDA")) + theme(legend.position = "none", panel.background = element\_blank(),axis.title.y = element\_text(size = 15,vjust = 7),axis.line = element\_line(size = 1),axis.text = element\_text(size = 15),panel.grid = element\_blank(),panel.border = element\_blank(), plot.margin = unit(rep(1.5,4),'lines')) + xlab("") + ylab("dN/dS")