**# Put all headers in one file**

sed -n -e '/^>/p' ABR2\_selected\_blast/ABR2\_blast\_all.fasta > ABR2\_blast\_all\_headers.txt

**# Keep species names and count how often each name occurs**

awk 'BEGIN{FS="|"};{print $1 "|" $2 "|" $3}' ABR2\_blast\_all\_headers.txt | uniq -c | sort -r > ABR2\_count\_paralogs.txt

**# Put all paralogs with the same number together**

sed -n -e '/^      2 /p' ABR2\_count\_paralogs.txt > ABR2\_headers\_2\_paralogs.txt

sed -n -e '/^      3 /p' ABR2\_count\_paralogs.txt > ABR2\_headers\_3\_paralogs.txt

sed -n -e '/^      4 /p' ABR2\_count\_paralogs.txt > ABR2\_headers\_4\_paralogs.txt

**# Remove ‘      2 >‘ of lines of file**

sed -e 's!       2 >!!' ABR2\_headers\_2\_paralogs.txt  > ABR2\_headers\_2\_paralogs.txt

sed -e 's!       3 >!!' ABR2\_headers\_3\_paralogs.txt  > ABR2\_headers\_3\_paralogs.txt

sed -e 's!       4 >!!' ABR2\_headers\_4\_paralogs.txt  > ABR2\_headers\_4\_paralogs.txt

**# Get original headers with 2 paralogs**

grep -Ff ABR2\_headers\_2\_paralogs.txt ABR2\_blast\_all\_headers.txt > ABR2\_complete\_headers\_2\_paralogs.txt

grep -Ff ABR2\_headers\_3\_paralogs.txt ABR2\_blast\_all\_headers.txt > ABR2\_complete\_headers\_3\_paralogs.txt

grep -Ff ABR2\_headers\_4\_paralogs.txt ABR2\_blast\_all\_headers.txt > ABR2\_complete\_headers\_4\_paralogs.txt

**# Add suffix ,1,1,#ff0000,1,1,3\_paralogs**

awk '{print $0, ",1,1,#ff0000,1,1,2\_paralogs"}' ABR2\_complete\_headers\_2\_paralogs.txt > ABR2\_headers\_itol\_u1\_2\_paralogs.txt

awk '{print $0, ",3,1,#00ff00,1,1,2\_paralogs"}' ABR2\_complete\_headers\_3\_paralogs.txt > ABR2\_headers\_itol\_u1\_3\_paralogs.txt

awk '{print $0, ",5,1,#0000ff,1,1,2\_paralogs"}' ABR2\_complete\_headers\_4\_paralogs.txt > ABR2\_headers\_itol\_u1\_4\_paralogs.txt

**DATASET\_SYMBOL example**

#Symbol datasets allow the display of various symbols on the branches of the tree. For each node, one or more symbols can be defined.

#Each symbol's color, size and position along the branch can be specified.

#lines starting with a hash are comments and ignored during parsing

#=================================================================#

# MANDATORY SETTINGS #

#=================================================================#

#select the separator which is used to delimit the data below (TAB,SPACE or COMMA).This separator must be used throughout this file.

#SEPARATOR TAB

#SEPARATOR SPACE

SEPARATOR COMMA

#label is used in the legend table (can be changed later)

DATASET\_LABEL,3 paralogs (green)

#dataset color (can be changed later)

COLOR,#00ff00

#=================================================================#

# OPTIONAL SETTINGS #

#=================================================================#

#=================================================================#

# all other optional settings can be set or changed later #

# in the web interface (under 'Datasets' tab) #

#=================================================================#

#Each dataset can have a legend, which is defined using LEGEND\_XXX fields below

#For each row in the legend, there should be one shape, color and label.

#Optionally, you can define an exact legend position using LEGEND\_POSITION\_X and LEGEND\_POSITION\_Y. To use automatic legend positioning, do NOT define these values

#Optionally, shape scaling can be present (LEGEND\_SHAPE\_SCALES). For each shape, you can define a scaling factor between 0 and 1.

#Optionally, shapes can be inverted (LEGEND\_SHAPE\_INVERT). When inverted, shape border will be drawn using the selected color, and the fill color will be white.

#Shape should be a number between 1 and 6, or any protein domain shape definition.

#1: square

#2: circle

#3: star

#4: right pointing triangle

#5: left pointing triangle

#6: checkmark

#LEGEND\_TITLE,Dataset legend

#LEGEND\_POSITION\_X,100

#LEGEND\_POSITION\_Y,100

#LEGEND\_SHAPES,1,2,3

#LEGEND\_COLORS,#ff0000,#00ff00,#0000ff

#LEGEND\_LABELS,value1,value2,value3

#LEGEND\_SHAPE\_SCALES,1,1,0.5

#LEGEND\_SHAPE\_INVERT,0,0,0

#largest symbol will be displayed with this size, others will be proportionally smaller.

MAXIMUM\_SIZE,5

#Internal tree nodes can be specified using IDs directly, or using the 'last common ancestor' method described in iTOL help pages

#=================================================================#

# Actual data follows after the "DATA" keyword #

#=================================================================#

#the following fields are required for each node:

#ID,symbol,size,color,fill,position,label

#symbol should be a number between 1 and 5:

#1: rectangle

#2: circle

#3: star

#4: right pointing triangle

#5: left pointing triangle

#6: checkmark

#size can be any number. Maximum size in the dataset will be displayed using MAXIMUM\_SIZE, while others will be proportionally smaller

#color can be in hexadecimal, RGB or RGBA notation. If RGB or RGBA are used, dataset SEPARATOR cannot be comma.

#fill can be 1 or 0. If set to 0, only the outline of the symbol will be displayed.

#position is a number between 0 and 1 and defines the position of the symbol on the branch (for example, position 0 is exactly at the start of node branch, position 0.5 is in the middle, and position 1 is at the end)

DATA

DOTHIDEO|jgi|VenspP74\_1|834318|gm1.13246\_g,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|VenspP74\_1|803797|estExt\_fgenesh1\_pg.C\_3\_t30062,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|VenspP74\_1|733324|estExt\_Genewise1Plus.C\_12\_t20196,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|Plesi1|504568|gm1.8350\_g,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|Plesi1|447466|fgenesh1\_pg.12\_#\_17,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|Plesi1|417893|estExt\_Genewise1.C\_220231,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|Lophiu1|608260|fgenesh1\_pm.6\_#\_86,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|Lophiu1|514277|e\_gw1.14.147.1,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|Lophiu1|526518|fgenesh1\_pg.5\_#\_326,3,1,#00ff00,1,1,2\_paralogs

DOTHIDEO|jgi|CocheC5\_3|81477|CocheC5\_1.estExt\_Genewise1.C\_30036,3,1,#00ff00,1,1,2\_paralogs