Transposable Elements in the Era of Data Science

Hand-On Session/Demo: Evolutionary Dynamics of LTR Retrotransposons

Romain Guyot (IRD)

Introduction

Project Description

Compare large and small genomes from the same genus to investigate TE diversity and characterize lineage specific expansions or contraction of TEs. We propose to study species from the Coffea genus with the large genome of *Coffea canephora* (700 Mb) and the small genome of *Coffea humblotiana* (400 Mb).

The divergence of *Coffea canephora* and *Coffea humblotiana* is about 8 My (https://doi.org/10.1038/s41598-021-87419-0)

In plant, genome size variations are mainly due to expansion/contraction of LTR retrotransposons.

The projects objectives is to identify Reverse Transcriptase domains in the 2 genomes, classify them and perform a comparative phylogenetic analysis at the family level, and identify expansions/contraction of these families.

RT Genome annotation

Download *C. canephora* and *C. humblotiana* genomes.

`wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/036/785/865/GCA_036785865.1_ASM3678586v1/GCA_036785865.1_ASM3678586v1_genomic.fna.gz`

`wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/023/065/735/GCA_023065735.1_IRD_Cohum_1.0/GCA_023065735.1_IRD_Cohum_1.0_genomic.fna.gz`

Identify LTR retrotranspoposon RT domains from both genome using Censor (https://www.girinst.org/downloads/software/censor/). RepeatMasker can also be used.

RT Genome annotation

```
CM041208.1    339760 339960  RT_Hopscotch#RLC+NA      11 79      d   0.3690  0.5500  100 0.0027.06
CM041208.1    339961 340691  Alesia.Ty1-RT__REXdb_ID1793#RLC+ALE-RETROFIT       2 256     d   0.4530  0.6700  576 0.0099.61
CM041208.1    745019 745607  Ikeros.Ty1-RT__REXdb_ID2697#RLC+IKEROS-TORK      41 238     c   0.6743  0.8252  709 0.0076.74
CM041208.1    745610 745729  RT_Poco#RLC+ORYCO       3 43      c   0.4634  0.6800  109 0.0016.40
CM041208.1    795476 795541  RT_Tnt-1#RLC+TORK     231 252     c   0.5375  0.7170  59 0.008.66
CM041208.1    795542 796224  Tork.Ty1-RT__REXdb_ID4820#RLC+TORK       1 226     c   0.7197  0.8265  848 0.0086.92
CM041208.1    909358 910113  TAR.Ty1-RT__REXdb_ID4619#RLC+TAR       1 256     c   0.5573  0.7400  719 0.00100.00
CM041208.1    915700 916572  RT_PVCV#Caulimovirus+NA      12 300     c   0.4742  0.6200  636 0.0096.01
CM041208.1    935897 936768  RT_PVCV#Caulimovirus+NA      12 300     d   0.4859  0.6450  677 0.0096.01
CM041208.1    986865 987622  TAR.Ty1-RT__REXdb_ID4619#RLC+TAR       1 256     c   0.5664  0.7498  755 0.00100.00
CM041208.1   1089022 1089795 Ikeros.Ty1-RT__REXdb_ID2697#RLC+IKEROS-TORK       1 258     c   0.5814  0.7500  801 0.00100.00
CM041208.1   1089796 1089843 RT_Hopscotch#RLC+NA      70 84      c   0.4516  0.7000  37 0.005.88
CM041208.1   1175752 1176535 Tork.Ty1-RT__REXdb_ID4820#RLC+TORK       1 260     c   0.6807  0.8086  907 0.00100.00
CM041208.1   1260113 1260808 RT_ToRTL1#RLC+SIRE       6 237     d   0.5388  0.7200  661 0.0093.55
CM041208.1   1265180 1265911 RT_Diaspora#RLG+Athila       1 243     d   0.6885  0.8200  887 0.00100.00
CM041208.1   1287550 1288230 Bianca.Ty1-RT__REXdb_ID2360#RLC+BIANCA      36 262     c   0.6740  0.8200  811 0.0086.64
CM041208.1   1303263 1303403 RT_Sto-4#RLC+TORK     207 253     c   0.5532  0.7700  145 0.0018.58
CM041208.1   1417518 1418167 RT_Gloin#RLG+REINA      12 222     c   0.5051  0.6458  478 0.0095.05
CM041208.1   1418202 1418269 RT_BsCVBV#Pararetrovirus+NA      55 75      c   0.3149  0.5100  21 0.006.69
CM041208.1   1896356 1896412 RT_Gloin#RLG+REINA     203 222     c   0.5850  0.7351  57 0.009.01
CM041208.1   1896413 1897007 RT_Gimli#RLG+REINA       1 198     c   0.6193  0.7826  661 0.0088.00
CM041208.1   1990319 1990384 RT_Hopscotch#RLC+NA      64 84      d   0.3551  0.5900  36 0.008.24
CM041208.1   1990385 1991151 TAR.Ty1-RT__REXdb_ID4620#RLC+TAR       1 256     d   0.6274  0.7593  828 0.00100.00
CM041208.1   2275533 2276204 Alesia.Ty1-RT__REXdb_ID1793#RLC+ALE-RETROFIT       1 224     d   0.4241  0.6400  485 0.0087.50
CM041208.1   2298964 2299638 Alesia.Ty1-RT__REXdb_ID1793#RLC+ALE-RETROFIT      22 246     d   0.4602  0.6800  525 0.0087.89
CM041208.1   2312980 2313147 RT_ToRTL1#RLC+SIRE     146 200     d   0.4821  0.7000  135 0.0022.18
CM041208.1   2316511 2317283 Ikeros.Ty1-RT__REXdb_ID2698#RLC+IKEROS-TORK       1 258     d   0.5426  0.7563  770 0.00100.00
CM041208.1   2437469 2438200 RT_Athila4-1#RLG+ATHILA       1 244     d   0.6885  0.8300  945 0.00100.00
CM041208.1   2527454 2528137 RT_Athila4-1#RLG+ATHILA       1 229     c   0.5375  0.6732  636 0.0093.85
CM041208.1   2528138 2528212 RT_CLNV+NA        39 62      c   0.3289  0.5100  26 0.007.48
CM041208.1   2528213 2528311 RT_ComYMV#Pararetrovirus+NA       1 31      c   0.2826  0.4600  21 0.009.87
CM041208.1   2615942 2616010 RT_Hopscotch#RLC+NA      64 85      d   0.4750  0.6200  49 0.008.63
CM041208.1   2616011 2616778 Alesia.Ty1-RT__REXdb_ID1793#RLC+ALE-RETROFIT       1 256     d   0.5039  0.7100  710 0.00100.00
CM041208.1   2632006 2632736 RT_Athila4-1#RLG+ATHILA       1 244     d   0.6133  0.7698  804 0.00100.00
CM041208.1   2645355 2646059 RT_Tork4#RLC+TORK       8 242     c   0.6553  0.8100  814 0.0092.16
CM041208.1   2646060 2646119 RT_Hopscotch#RLC+NA      69 87      c   0.4519  0.6100  36 0.007.45
CM041208.1   2701047 2701778 RT_Diaspora#RLG+Athila       1 243     c   0.6885  0.7900  876 0.00100.00
CM041208.1   2701779 2701937 RT_TaBV+NA       253 304     d   0.2914  0.5000  47 0.0016.61
CM041208.1   2727238 2727969 RT_Diaspora#RLG+Athila       1 243     c   0.7049  0.8200  927 0.00100.00
CM041208.1   2727973 2728089 RT_EVCV#Pararetrovirus+NA     264 301     d   0.2927  0.4800  30 0.0012.46
CM041208.1   2746177 2746950 Ikeros.Ty1-RT__REXdb_ID2697#RLC+IKEROS-TORK       1 258     d   0.6705  0.8200  947 0.00100.00
CM041208.1   2766037 2766765 RT_Diaspora#RLG+Athila       1 242     c   0.6420  0.8000  854 0.0099.59
CM041208.1   2766766 2766885 RT_DBV+NA        264 302     d   0.3158  0.5100  38 0.0012.42
CM041208.1   3110632 3110694 RT_Hopscotch#RLC+NA      65 84      d   0.3364  0.5600  28 0.007.84
CM041208.1   3110695 3111462 TAR.Ty1-RT__REXdb_ID4618#RLC+TAR       1 256     d   0.5451  0.7173  739 0.00100.00
CM041208.1   3485261 3485972 RT_Tork4#RLC+TORK      12 248     c   0.6055  0.7745  742 0.0092.94
CM041208.1   3506627 3507370 RT_Tork4#RLC+TORK       8 255     c   0.6201  0.7953  786 0.0097.25
```

Identification and phylogenetic analysis of RT domains

From each .map files, extract RT domains and filter domains with a minimum of 200 aa.

Use the python script censor_2_phylo.py to :

1_ filter map results with a minimum of 50% of id and 50% of coverage

2_ extract RT domain using Genewise,

3_ perform alignment using mafft

4_ phylogenetic tree using FastTree (4time of execution 48m12s, 12 CPUs).

You need to install emboss, mafft, fastTree and Genewise

```
python censor_2_phylo.py genome RTlibrary genome.map output_folder final_name_base %id %cov
```

Results for each genome:

A list of RT domains extracted `final_name_base.clean.fasta`

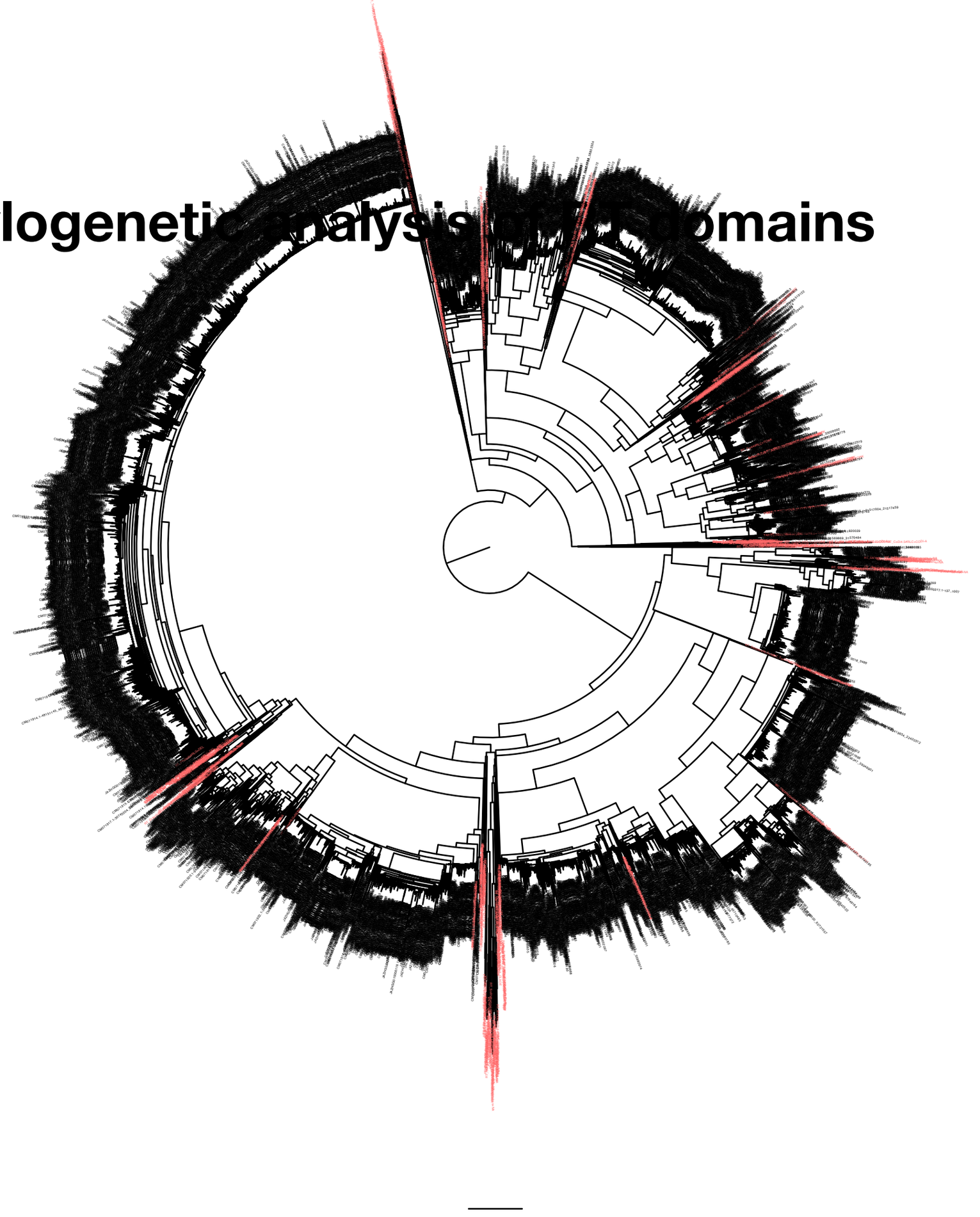
An alignment file of RT domains and reference domains: `final_name_base.clean.aln`

A phylogenetic tree: `final_name_base.tree`

you may display the tree file with figtree

Identification and phylogenetic analysis of RT domains

10524 RT domains



Comparative phylogenetic analysis of RT domains

First you need to merge RT domains from each genome, aligned them and compute a phylogenetic tree with mafft and FastTree

```
mafft --thread 12 ALL.fas > ALL.fas.maaft
FastTree ALL.fas.maaft > ALL.fas.maaft.tree
```

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

1_check the node number for the tree

2_Identify the reference RT sequence and clade

3_Draw the tree with each identified group

Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

```
You can:
1_check the node number in the tree
#Step 1 Get node number
################################################################################
library(ggtree)
library(phangorn)
library(ape)
library(dplyr)
# Newick tree
tree <- read.tree("mytree.tree")
#
rooted_tree <- midpoint(tree)

# see circular treewith node number
p <- ggtree(rooted_tree, layout = "circular") +
  geom_text(aes(label = node), color = "red", size = 0.5, hjust = -0.3) +
  geom_tree(color = "gray", size = 0.05, linewidth = 0.05)

# Register the picture in PDF
ggsave("mytree.NODE_NUMBER.pdf", plot = p, width = 9, height = 11, dpi = 1200)
```

Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

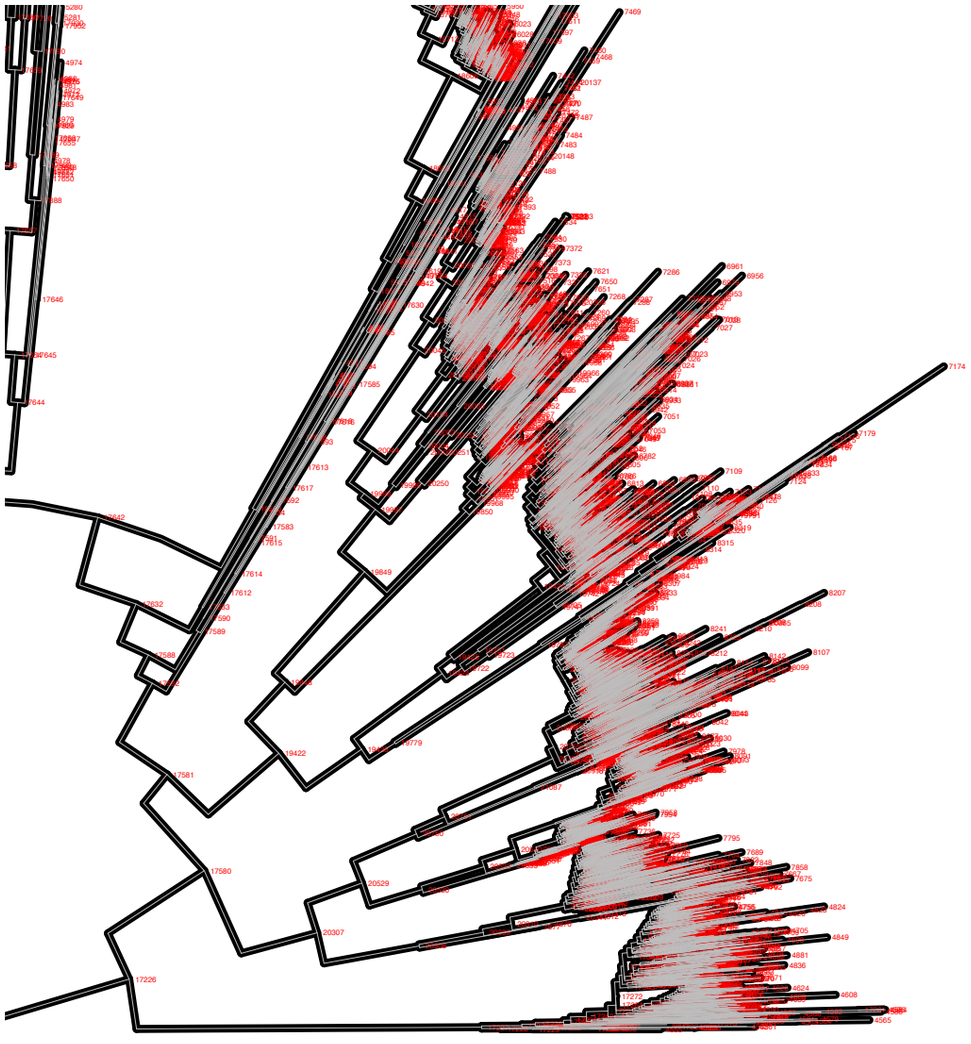
1_check the node number in the tree

Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

1_check the node number in the tree



Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

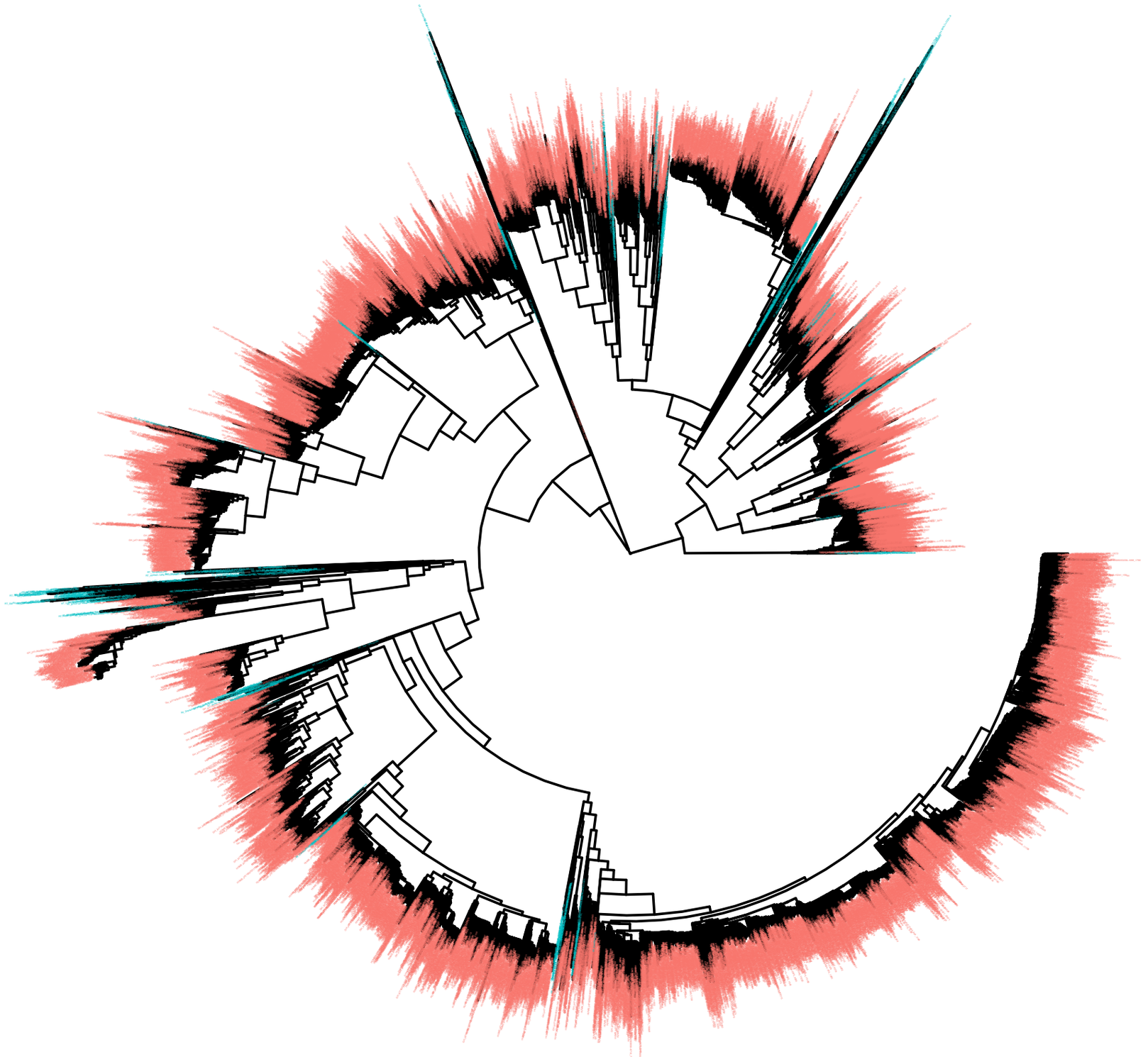
```
You can:
2_Identify the reference RT sequence and clade
#Step2 See  reference RT
################################################################################
install.packages("ggrepel")
library(ggtree)
library(ggrepel)
library(dplyr)

#  Newick tree
tree <- read.tree("Gmytree.tree")
rooted_tree <- midpoint(tree)
rooted_tree$node <- as.factor(rooted_tree$node)
rooted_tree$parent <- as.factor(rooted_tree$parent)

# circular tree
p <- ggtree(rooted_tree, layout = "circular") +
  geom_tree(linewidth = 0.01) +
  #
  geom_tiplab(aes(color = case_when(
    grepl("RT_", label) ~ "red",
    TRUE ~ "black"
  ), label = label), size = 0.2, offset = 0.02, align = FALSE) +
  theme(legend.position = "none")  # Supprimer la légende

# Register picture  in  PDF
ggsave("mytree.with_ref.pdf", plot = p, width = 9, height = 11, dpi = 1200)
```

Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

2_Identify the reference RT sequence and clade

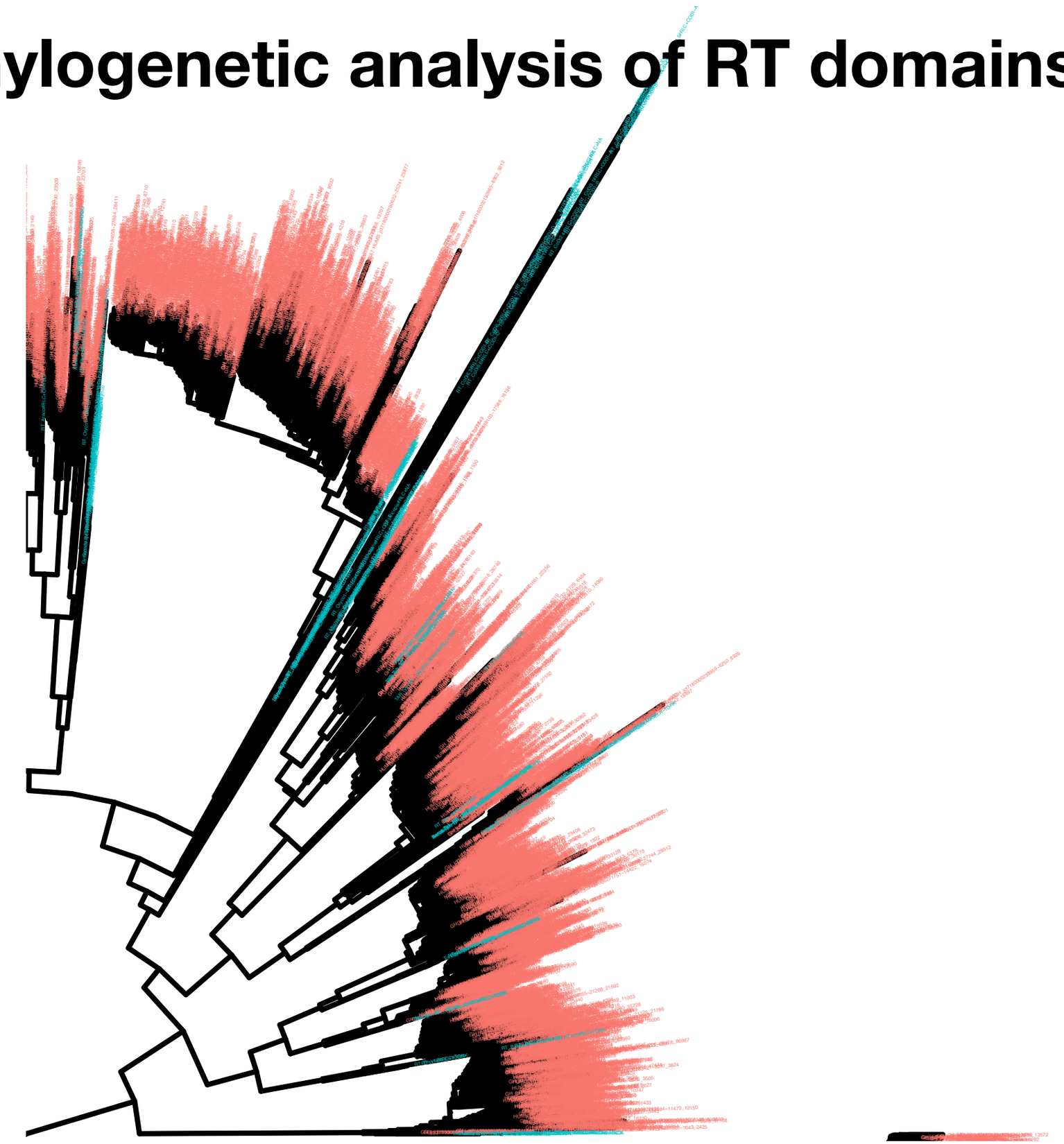


Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

2_Identify the reference RT sequence and clade



Day 4

Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

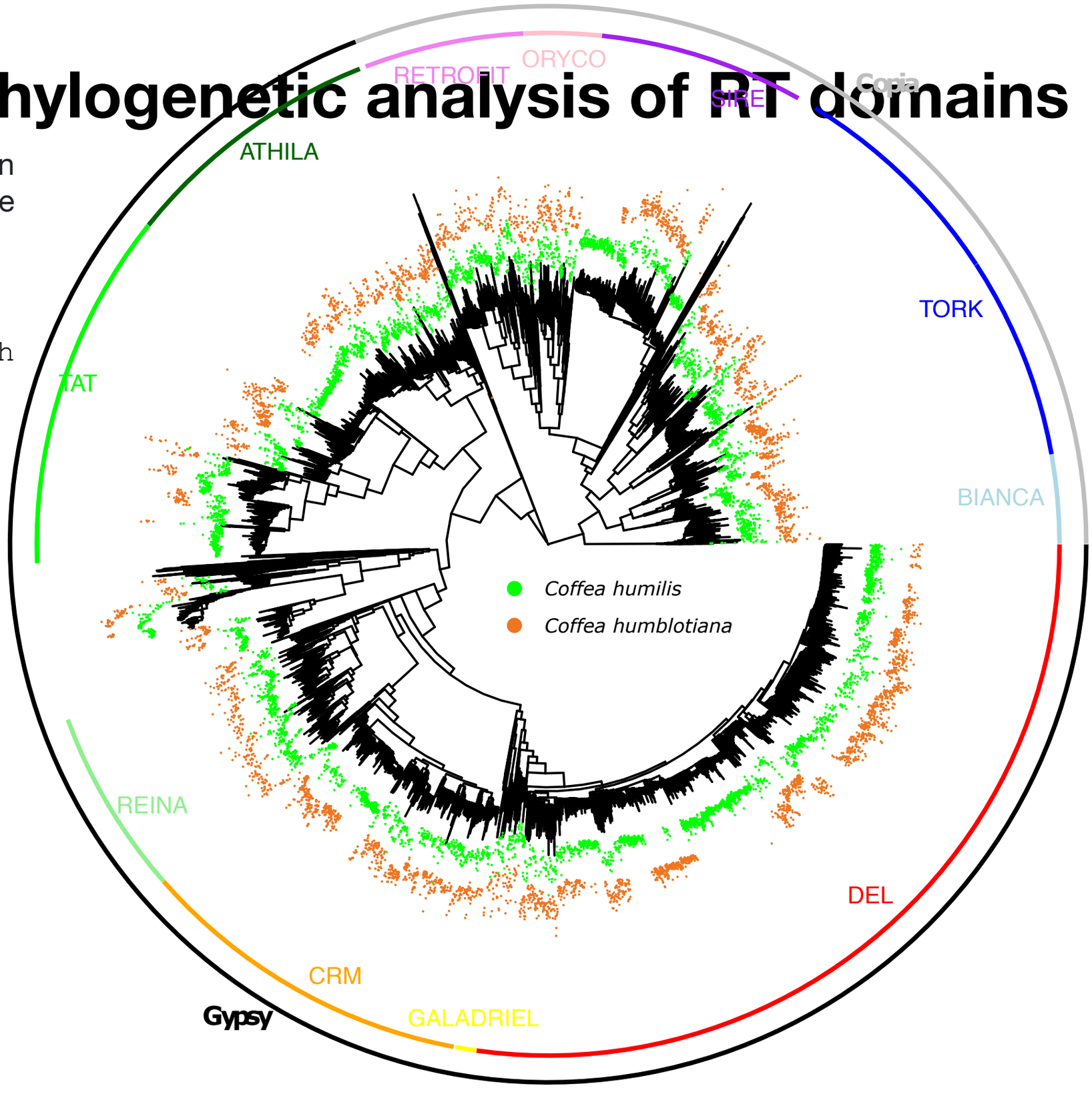
```
You can:
3_Draw the tree with each identified group
3. compare different genomes
################################################################################
library(ape)
library(ggtree)
library(ggrepel)
#  Newick tree
tree <- read.tree("mytree.tree")
rooted_tree <- midpoint(tree)
# convert in data frame
fortified_tree <- fortify(rooted_tree)
# 2 genomes (A, B)
legend_data <- data.frame(
  x = rep(0, 3),  # Position X fictive pour la légende
  y = rep(0, 3),  # Position Y fictive pour la légende
  category = c("A", "B")
  color = c("green", "pink"),
  shape = c(16, 16)
)
# circular tree
p <- ggtree(rooted_tree, layout = "circular") +
  geom_tree(linewidth = 0.05) +  # Lignes plus fines
  # A
  geom_point(aes(x = x + 0.2, y = y),
             shape = 16, size = 0.2, color = "green",
             data = subset(fortified_tree, grepl("^A", label))) +
  # B
  geom_point(aes(x = x + 0.4, y = y),
             shape = 16, size = 0.2, color = "pink",
             data = subset(fortified_tree, grepl("^B", label))) +

  # legende
  geom_point(aes(x = x, y = y, color = category, shape = category),
             size = 3, data = legend_data, show.legend = TRUE) +
  # légende
  scale_color_manual(values = setNames(legend_data$color, legend_data$category)) +
  scale_shape_manual(values = setNames(legend_data$shape, legend_data$category)) +
  labs(color = "Category", shape = "Category") +
  theme(legend.position = "right")


# Clade  labels
p <- p + geom_cladelabel(node = 28362, label = "Copia", offset = 1, color = "gray",
barsize = 1, align = TRUE)+
  geom_cladelabel(node = xx, label = "Gypsy", offset = 1, color = "black", barsize = 1,
align = TRUE)+
#  geom_cladelabel(node = xx, label = "Gypsy", offset = 1, color = "black", barsize = 1,
align = TRUE)+
  geom_cladelabel(node = xx, label = "TAT", offset = 0.8, color = "green", barsize = 1,
align = TRUE) +
  geom_cladelabel(node = xx, label = "ATHILA", offset = 0.8, color = "darkgreen", barsize
= 1, align = TRUE) +
  geom_cladelabel(node = xx, label = "DEL", offset = 0.8, color = "red", barsize = 1,
align = TRUE) +
  geom_cladelabel(node = xx, label = "GALADRIEL", offset = 0.8, color = "yellow", barsize
= 1, align = TRUE) +
  geom_cladelabel(node = xx, label = "CRM", offset = 0.8, color = "orange", barsize = 1,
align = TRUE) +
  geom_cladelabel(node = xx, label = "REINA", offset = 0.8, color = "lightgreen", barsize
= 1, align = TRUE)+
  geom_cladelabel(node = xx, label = "SIRE", offset = 0.8, color = "purple", barsize = 1,
align = TRUE) +
  geom_cladelabel(node = xx, label = "ORYCO", offset = 0.8, color = "pink", barsize = 1,
align = TRUE) +
  geom_cladelabel(node = xx, label = "RETROFIT", offset = 0.8, color = "violet", barsize
= 1, align = TRUE) +
  geom_cladelabel(node = xx, label = "TORK", offset = 0.8, color = "blue", barsize = 1,
align = TRUE) +
  geom_cladelabel(node = 32255, label = "BIANCA", offset = 0.8, color = "lightblue",
barsize = 1, align = TRUE)


# PDF
ggsave("my_tree_final.pdf", plot = p, device = "pdf", width = 10, height = 10, dpi =
1200)
```

Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

3_Draw the tree with each identified group

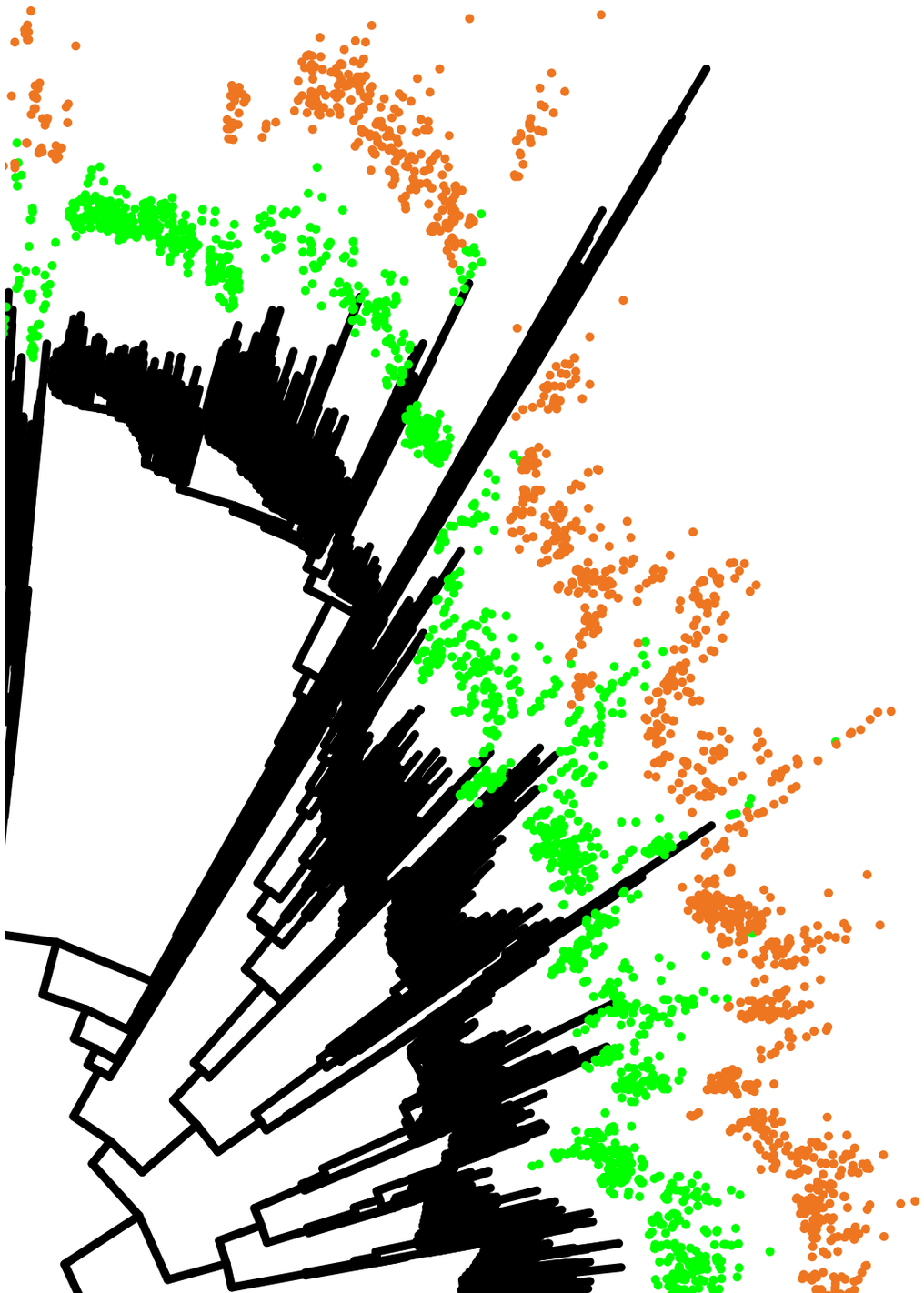


Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

3_Draw the tree with each identified group

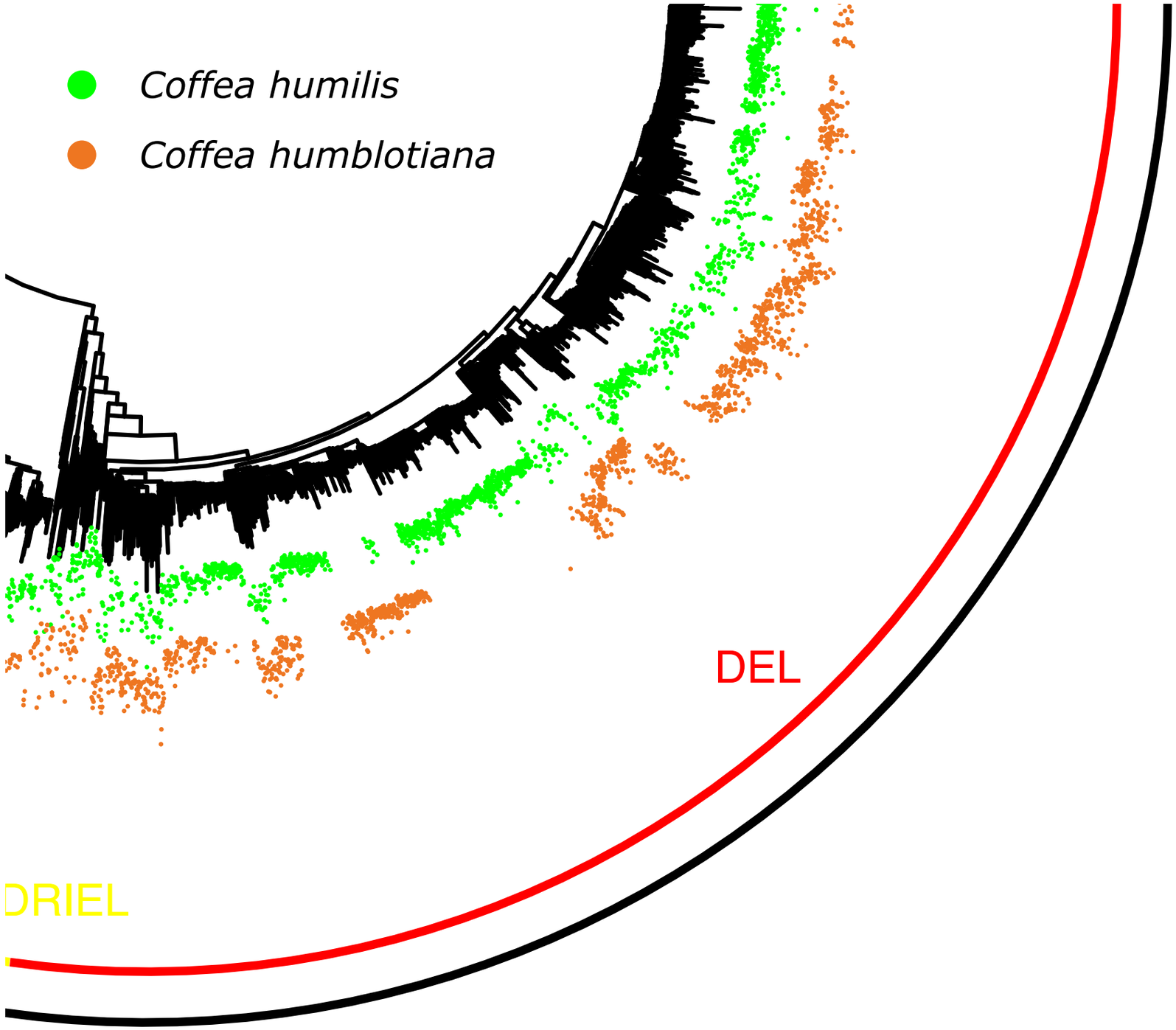


Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can:

3_Draw the tree with each identified group



Comparative phylogenetic analysis of RT domains

Using R and R studio you can analyze and edit comparative phylogenetic trees

You can count or extract RT domain names from a node in the tree for downstream analysis.