phylogeny\_snprelate

2023-05-18

library(SNPRelate)  
library(gdsfmt)  
library(ape)  
library(ggtree)  
library(ggplot2)

Download multi.norm.vcf.gz file from course github, unzip, and rename epi.vcf

# load normalized and filtered multisample VCF file  
# and convert VCF to GDS format  
snpgdsVCF2GDS(vcf.fn="epi.vcf", out.fn="epi.gds",   
 method="biallelic.only")

## Start file conversion from VCF to SNP GDS ...  
## Method: extracting biallelic SNPs  
## Number of samples: 6  
## Parsing "epi.vcf" ...  
## import 2111 variants.  
## + genotype { Bit2 6x2111, 3.1K } \*  
## Optimize the access efficiency ...  
## Clean up the fragments of GDS file:  
## open the file 'epi.gds' (20.0K)  
## # of fragments: 46  
## save to 'epi.gds.tmp'  
## rename 'epi.gds.tmp' (19.7K, reduced: 312B)  
## # of fragments: 20

# print GDS file summary  
snpgdsSummary("epi.gds")

## The file name: C:\Projects\snp\_relate\_workflow\epi.gds   
## The total number of samples: 6   
## The total number of SNPs: 2111   
## SNP genotypes are stored in SNP-major mode (Sample X SNP).

# open GDS file  
genofile <- snpgdsOpen("epi.gds")  
class(genofile)

## [1] "SNPGDSFileClass" "gds.class"

genofile

## File: C:\Projects\snp\_relate\_workflow\epi.gds (19.7K)  
## + [ ] \*  
## |--+ sample.id { Str8 6 LZMA\_ra(340.0%), 109B }  
## |--+ snp.id { Int32 2111 LZMA\_ra(14.7%), 1.2K }  
## |--+ snp.rs.id { Str8 2111 LZMA\_ra(4.64%), 105B }  
## |--+ snp.position { Int32 2111 LZMA\_ra(54.1%), 4.5K }  
## |--+ snp.chromosome { Str8 2111 LZMA\_ra(0.61%), 161B }  
## |--+ snp.allele { Str8 2111 LZMA\_ra(15.9%), 1.3K }  
## |--+ genotype { Bit2 6x2111, 3.1K } \*  
## \--+ snp.annot [ ]  
## |--+ qual { Float32 2111 LZMA\_ra(87.4%), 7.2K }  
## \--+ filter { Str8 2111 LZMA\_ra(4.64%), 105B }

# Set seed and run clustering  
set.seed(100)  
  
# Run identity-by-state analysis on genotypes; this step calculates  
# dissimilarity matrix and perform hierarchical clustering  
# on the dissimilarity matrix  
epi\_clust <- snpgdsHCluster(snpgdsIBS(genofile, remove.monosnp=TRUE,  
 num.thread=2,   
 autosome.only=FALSE))

## Identity-By-State (IBS) analysis on genotypes:  
## Excluding 139 SNPs (monomorphic: TRUE, MAF: NaN, missing rate: NaN)  
## # of samples: 6  
## # of SNPs: 1,972  
## using 2 threads  
## IBS: the sum of all selected genotypes (0,1,2) = 8804  
## Fri May 19 00:49:58 2023 (internal increment: 65536)  
## [..................................................] 0%, ETC: --- [==================================================] 100%, completed, 0s  
## Fri May 19 00:49:58 2023 Done.

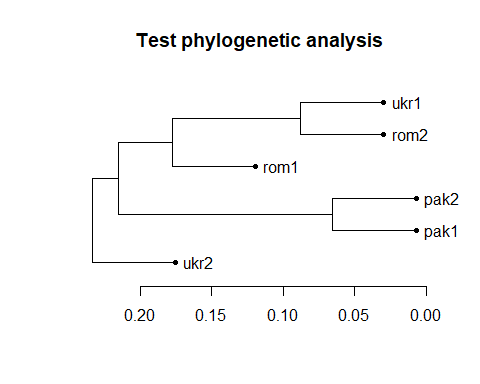
epi\_clust

## $sample.id  
## [1] "pak1" "pak2" "rom1" "rom2" "ukr1" "ukr2"  
##   
## $hclust  
##   
## Call:  
## hclust(d = as.dist(dist), method = "average")  
##   
## Cluster method : average   
## Number of objects: 6   
##   
##   
## $dendrogram  
## 'dendrogram' with 2 branches and 6 members total, at height 0.233688   
##   
## $dist  
## pak1 pak2 rom1 rom2 ukr1 ukr2  
## pak1 0.00000000 0.06543274 0.2334025 0.21142857 0.20696828 0.2288622  
## pak2 0.06543274 0.00000000 0.2292746 0.20586708 0.20451011 0.2257307  
## rom1 0.23340249 0.22927461 0.0000000 0.17979540 0.17582136 0.2545360  
## rom2 0.21142857 0.20586708 0.1797954 0.00000000 0.08808423 0.2324857  
## ukr1 0.20696828 0.20451011 0.1758214 0.08808423 0.00000000 0.2268255  
## ukr2 0.22886221 0.22573069 0.2545360 0.23248573 0.22682548 0.0000000  
##   
## attr(,"class")  
## [1] "snpgdsHCClass"

# Determine sub groups of individuals using a specified   
# dendrogram from hierarchical cluster analysis  
rv\_epi\_clust <- snpgdsCutTree(epi\_clust)

## Determine groups by permutation (Z threshold: 15, outlier threshold: 5):  
## Create 1 groups.

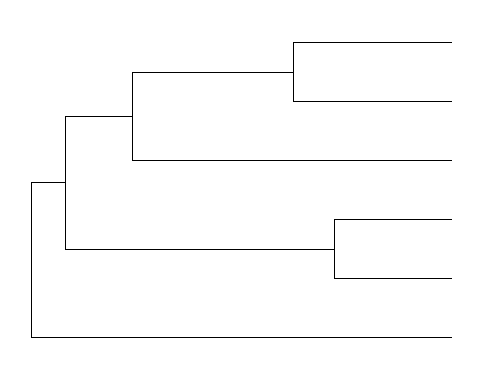
# Save dendrogram to a new variable  
epi\_tree <- rv\_epi\_clust$dendrogram  
  
# Plot the dendrogram  
plot(epi\_tree, horiz=T,   
 main ="Test phylogenetic analysis" )



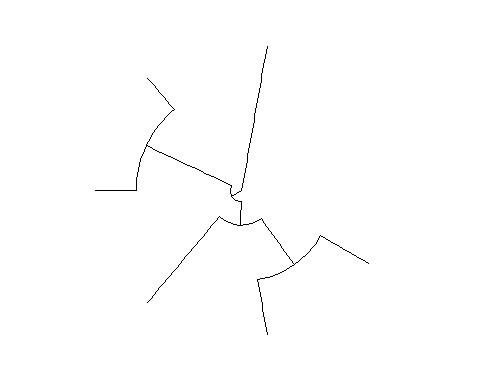
snpgdsClose(genofile)  
  
# ---------------------------------------------------- #  
# Customize phylogenetic analysis with ape and ggtree  
# Convert dendrogram object from snpRelate to   
# phylo object from ape  
epi\_phylo <- as.phylo(as.hclust(epi\_tree))  
epi\_phylo

##   
## Phylogenetic tree with 6 tips and 5 internal nodes.  
##   
## Tip labels:  
## pak1, pak2, rom1, rom2, ukr1, ukr2  
##   
## Rooted; includes branch lengths.

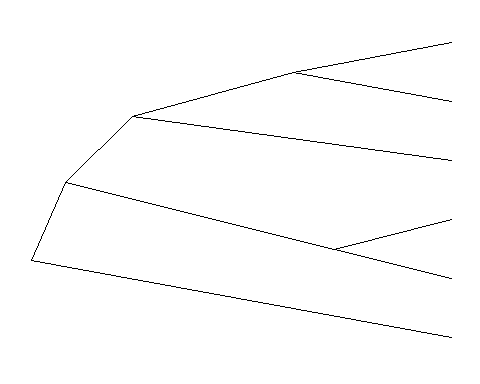
ggtree(epi\_phylo)



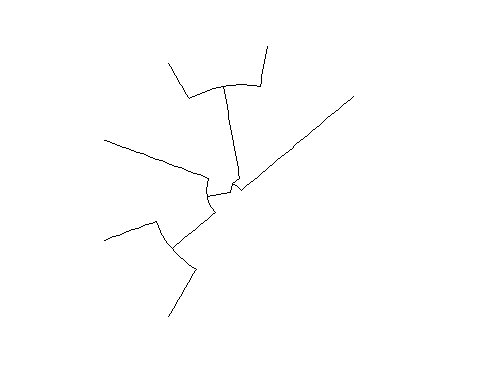
# Check out different layouts  
  
ggtree(epi\_phylo, layout="circular")



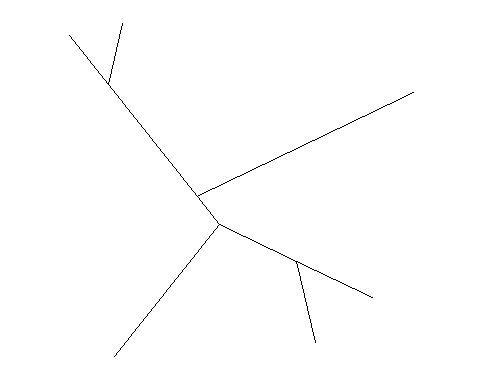
ggtree(epi\_phylo, layout="slanted")



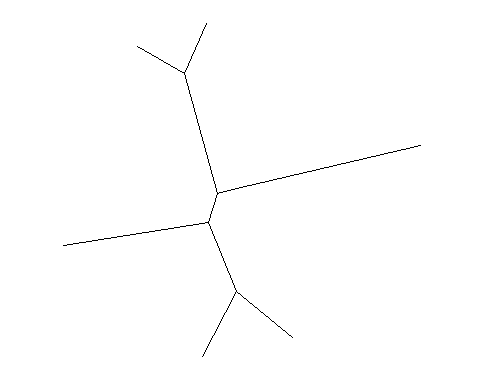
ggtree(epi\_phylo, layout="fan", open.angle = 120)



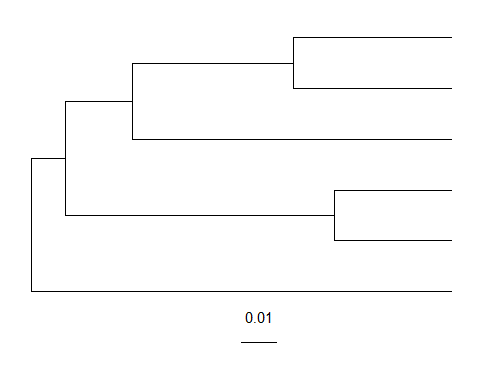
ggtree(epi\_phylo, layout="equal\_angle")



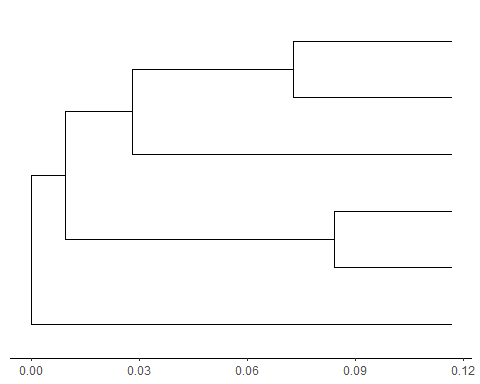
ggtree(epi\_phylo, layout="daylight")



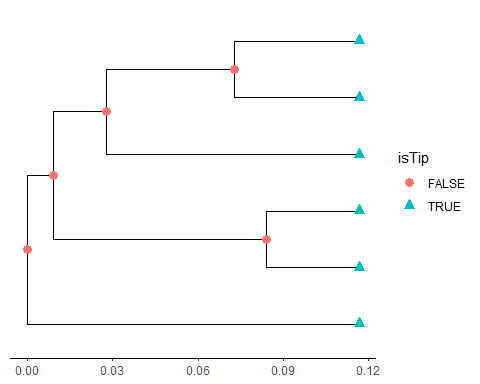
# Display tree scale (evolutionary distance)  
ggtree(epi\_phylo) + geom\_treescale()



# Display the scale by adding X-axis  
ggtree(epi\_phylo) + theme\_tree2()



# Adding nodes and tips  
ggtree(epi\_phylo) + theme\_tree2() +   
 geom\_point(aes(shape=isTip, color=isTip), size=3)



# Add labels  
ggtree(epi\_phylo) + theme\_tree2() +   
 geom\_point(aes(shape=isTip, color=isTip), size=3) +  
 geom\_tiplab(size=3, color="purple", vjust=-1, hjust = 1)

