trees

Davinder Singh

2023-04-13

install.packages("adegenet") install.packages("phangorn")

```
dna <- fasta2DNAbin(file="http://adegenet.r-forge.r-project.org/files/usflu.fasta")
```

```
##
##  Converting FASTA alignment into a DNAbin object...
##
##
##  Finding the size of a single genome...
##
##
##  genome size is: 1,701 nucleotides
##
## ( 30  lines per genome )
##
##  Importing sequences...
## ................................................................................
##  Forming final object...
##
## ...done.
```

```
dna
```

```
## 80 DNA sequences in binary format stored in a matrix.
##
## All sequences of same length: 1701
##
## Labels:
## CY013200
## CY013781
## CY012128
## CY013613
## CY012160
## CY012272
## ...
##
## Base composition:
##     a     c     g     t
## 0.335 0.200 0.225 0.239
## (Total: 136.08 kb)
```

```
annot <- read.csv("http://adegenet.r-forge.r-project.org/files/usflu.annot.csv", header=TRUE, row.names=
annot
```

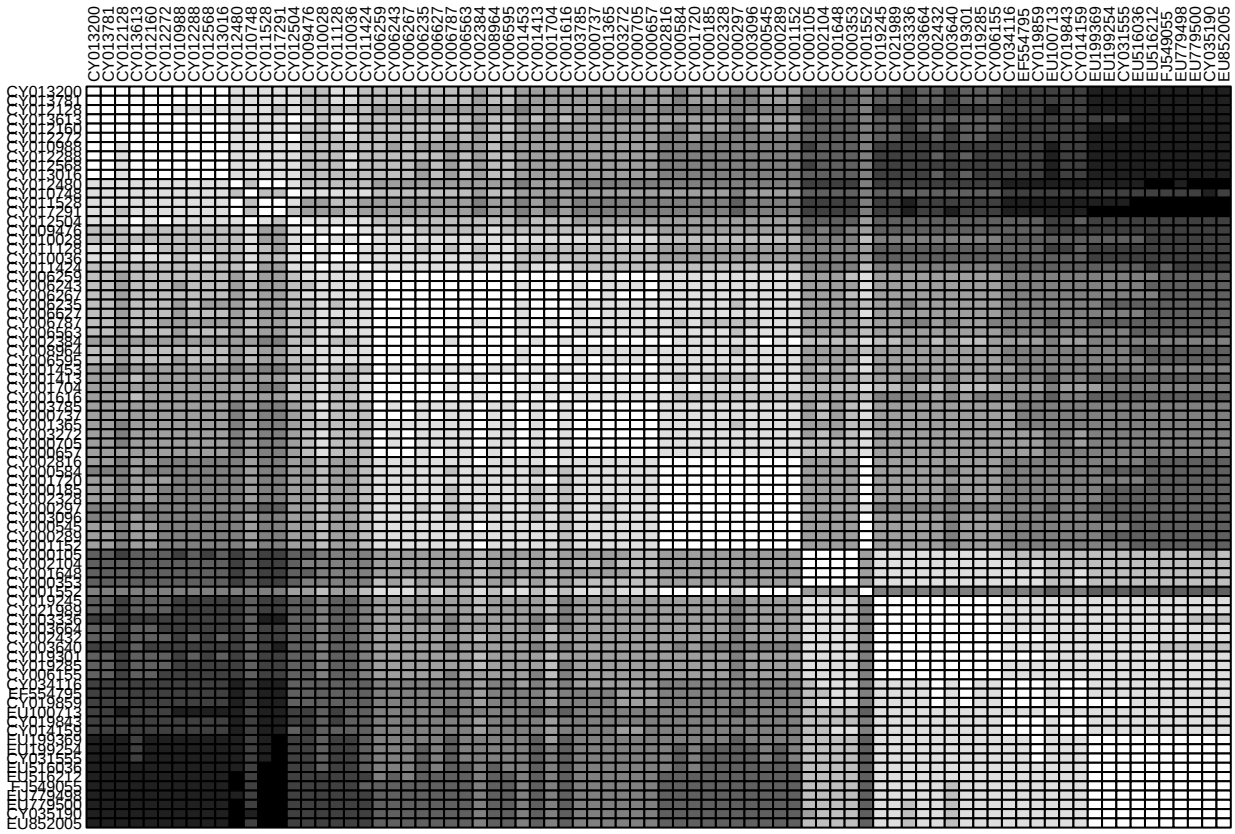
```
##    accession year                              misc
## 1   CY013200 1993       (A/New York/783/1993(H3N2))
## 2   CY013781 1993       (A/New York/802/1993(H3N2))
## 3   CY012128 1993       (A/New York/758/1993(H3N2))
## 4   CY013613 1993       (A/New York/766/1993(H3N2))
## 5   CY012160 1993       (A/New York/762/1993(H3N2))
## 6   CY012272 1994       (A/New York/729/1994(H3N2))
## 7   CY010988 1994       (A/New York/733/1994(H3N2))
## 8   CY012288 1994       (A/New York/734/1994(H3N2))
## 9   CY012568 1994       (A/New York/746/1994(H3N2))
## 10  CY013016 1994       (A/New York/750/1994(H3N2))
## 11  CY012480 1995       (A/New York/666/1995(H3N2))
## 12  CY010748 1995       (A/New York/648/1995(H3N2))
## 13  CY011528 1995       (A/New York/669/1995(H3N2))
## 14  CY017291 1995       (A/New York/681/1995(H3N2))
## 15  CY012504 1995       (A/New York/678/1995(H3N2))
## 16  CY009476 1996       (A/New York/565/1996(H3N2))
## 17  CY010028 1996       (A/New York/591/1996(H3N2))
## 18  CY011128 1996       (A/New York/599/1996(H3N2))
## 19  CY010036 1996       (A/New York/592/1996(H3N2))
## 20  CY011424 1996       (A/New York/577/1996(H3N2))
## 21  CY006259 1997       (A/New York/511/1997(H3N2))
## 22  CY006243 1997       (A/New York/508/1997(H3N2))
## 23  CY006267 1997       (A/New York/513/1997(H3N2))
## 24  CY006235 1997       (A/New York/505/1997(H3N2))
## 25  CY006627 1997       (A/New York/547/1997(H3N2))
## 26  CY006787 1998       (A/New York/506/1998(H3N2))
## 27  CY006563 1998       (A/New York/533/1998(H3N2))
## 28  CY002384 1998       (A/New York/330/1998(H3N2))
## 29  CY008964 1998       (A/New York/540/1998(H3N2))
## 30  CY006595 1998       (A/New York/542/1998(H3N2))
## 31  CY001453 1999       (A/New York/184/1999(H3N2))
## 32  CY001413 1999       (A/New York/263/1999(H3N2))
## 33  CY001704 1999       (A/New York/257/1999(H3N2))
## 34  CY001616 1999       (A/New York/265/1999(H3N2))
## 35  CY003785 1999       (A/New York/422/1999(H3N2))
## 36  CY000737 2000       (A/New York/180/2000(H3N2))
## 37  CY001365 2000       (A/New York/187/2000(H3N2))
## 38  CY003272 2000       (A/New York/437/2000(H3N2))
## 39  CY000705 2000       (A/New York/175/2000(H3N2))
## 40  CY000657 2000       (A/New York/169/2000(H3N2))
## 41  CY002816 2001       (A/New York/301/2001(H3N2))
## 42  CY000584 2001       (A/New York/127/2001(H3N2))
## 43  CY001720 2001       (A/New York/273/2001(H3N2))
## 44  CY000185 2001        (A/New York/83/2001(H3N2))
## 45  CY002328 2001        (A/New York/77/2001(H3N2))
## 46  CY000297 2002        (A/New York/96/2002(H3N2))
## 47  CY003096 2002       (A/New York/403/2002(H3N2))
## 48  CY000545 2002       (A/New York/115/2002(H3N2))
## 49  CY000289 2002        (A/New York/92/2002(H3N2))
```

```
## 50  CY001152 2002        (A/New York/74/2002(H3N2))
## 51  CY000105 2003       (A/New York/60A/2003(H3N2))
## 52  CY002104 2003           (A/Memphis/31/03(H3N2))
## 53  CY001648 2003       (A/New York/270/2003(H3N2))
## 54  CY000353 2003        (A/New York/21/2003(H3N2))
## 55  CY001552 2003       (A/New York/215/2003(H3N2))
## 56  CY019245 2004       (A/New York/908/2004(H3N2))
## 57  CY021989 2004       (A/New York/908/2004(H3N2))
## 58  CY003336 2004       (A/New York/354/2004(H3N2))
## 59  CY003664 2004       (A/New York/471/2004(H3N2))
## 60  CY002432 2004       (A/New York/362/2004(H3N2))
## 61  CY003640 2005       (A/New York/463/2005(H3N2))
## 62  CY019301 2005       (A/New York/918/2005(H3N2))
## 63  CY019285 2005       (A/New York/913/2005(H3N2))
## 64  CY006155 2005       (A/New York/258/2005(H3N2))
## 65  CY034116 2005       (A/Wisconsin/67/2005(H3N2))
## 66  EF554795 2006               (A/Ohio/2006(H3N2))
## 67  CY019859 2006       (A/New York/938/2006(H3N2))
## 68  EU100713 2006        (A/Maryland/09/2006(H3N2))
## 69  CY019843 2006       (A/New York/933/2006(H3N2))
## 70  CY014159 2006         (A/New York/7/2006(H3N2))
## 71  EU199369 2007       (A/Minnesota/08/2007(H3N2))
## 72  EU199254 2007           (A/Idaho/01/2007(H3N2))
## 73  CY031555 2007 (A/Kentucky/UR06-0571/2007(H3N2))
## 74  EU516036 2007         (A/Georgia/07/2007(H3N2))
## 75  EU516212 2007      (A/California/33/2007(H3N2))
## 76  FJ549055 2008        (A/Illinois/14/2008(H3N2))
## 77  EU779498 2008     (A/Mississippi/01/2008(H3N2))
## 78  EU779500 2008         (A/Indiana/02/2008(H3N2))
## 79  CY035190 2008 (A/Pennsylvania/PIT43/2008(H3N2))
## 80  EU852005 2008           (A/Texas/06/2008(H3N2))
```

```
D <- dist.dna(dna, model = "TN93")
length(D)
```

```
## [1] 3160
```

```
temp <- as.data.frame(as.matrix(D))
table.paint(temp, cleg=0, clabel.row=.5, clabel.col=.5)
```



```
tre <- nj(D)
class(tre)
```

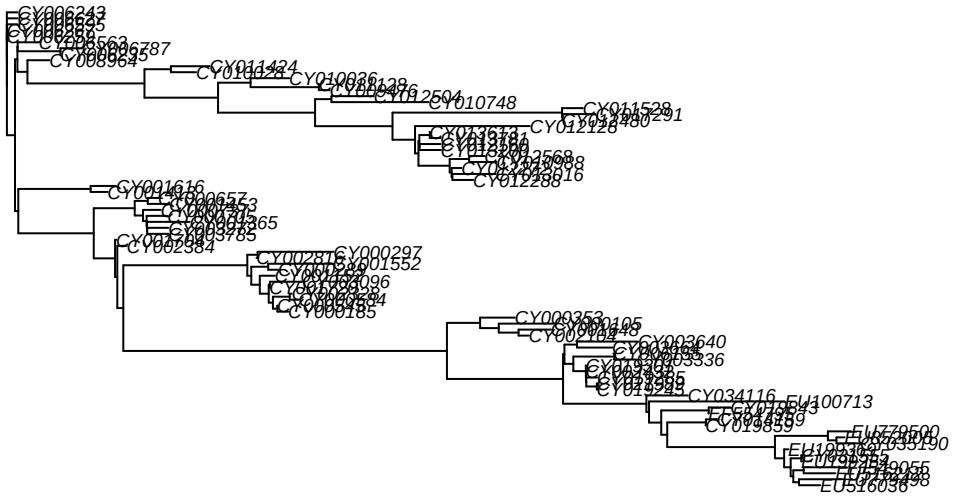
```
## [1] "phylo"
```

```
tre <- ladderize(tre)
tre
```

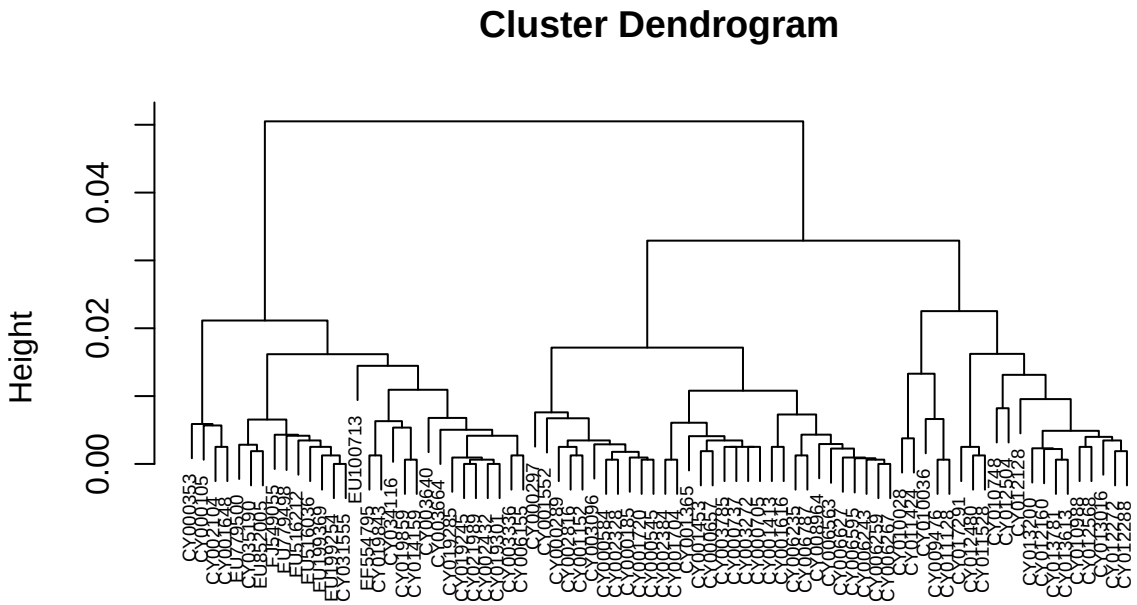
```
##
## Phylogenetic tree with 80 tips and 78 internal nodes.
##
## Tip labels:
##   CY013200, CY013781, CY012128, CY013613, CY012160, CY012272, ...
##
## Unrooted; includes branch lengths.
```

```
plot(tre, cex = 0.6)
title("A Simple NJ Tree")
```

A Simple NJ Tree

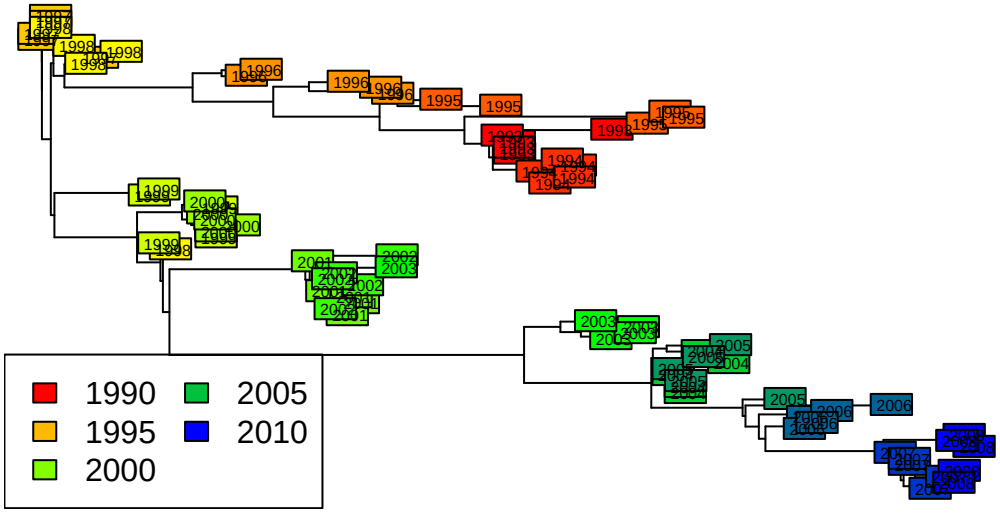



```
h_cluster <- hclust(D, method = "average", members = NULL)
plot(h_cluster, cex = 0.6)
```

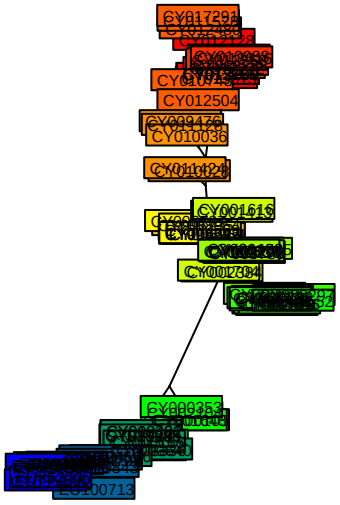


D
hclust (*, "average")

```
plot(tre, show.tip=FALSE) # gets rid of the labels on the end, refer to the first tree depicted above
title("Unrooted NJ tree")
myPal <- colorRampPalette(c("red","yellow","green","blue"))
tiplabels(annot$year, bg=num2col(annot$year, col.pal=myPal), cex=.5) #we use the annot dataset to get o
temp <- pretty(1993:2008, 5)
legend("bottomleft", fill=num2col(temp, col.pal=myPal), leg=temp, ncol=2)
```



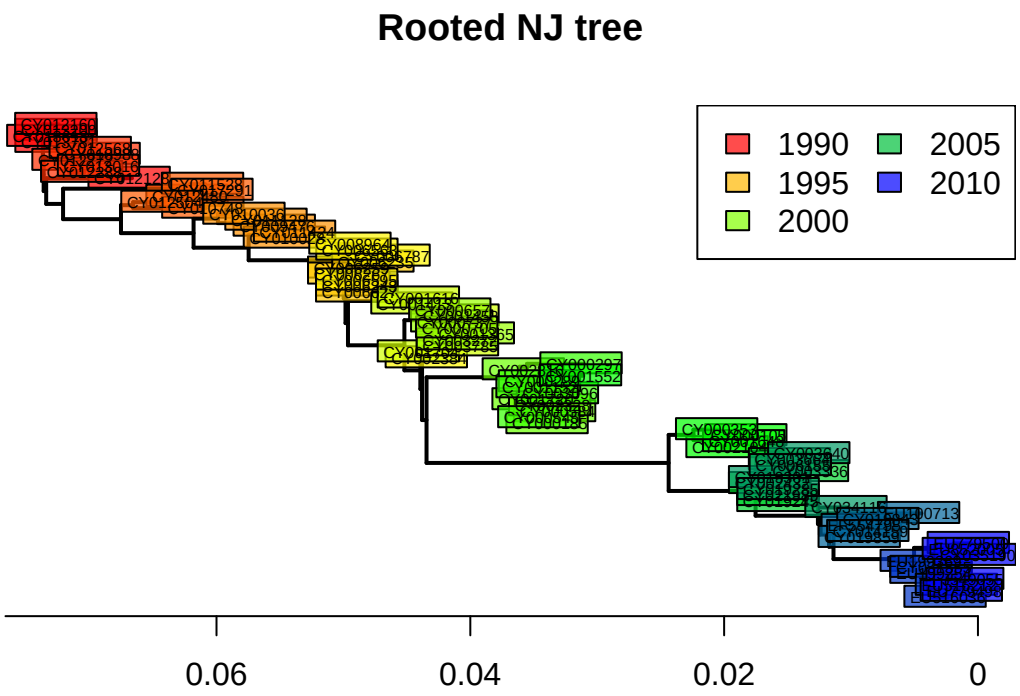

```
plot(tre, type = "unrooted", show.tip = FALSE)
title("Unrooted NJ Tree")
tiplabels(tre$tip.label, bg = num2col(annot$year, col.pal = myPal), cex = 0.5)
```



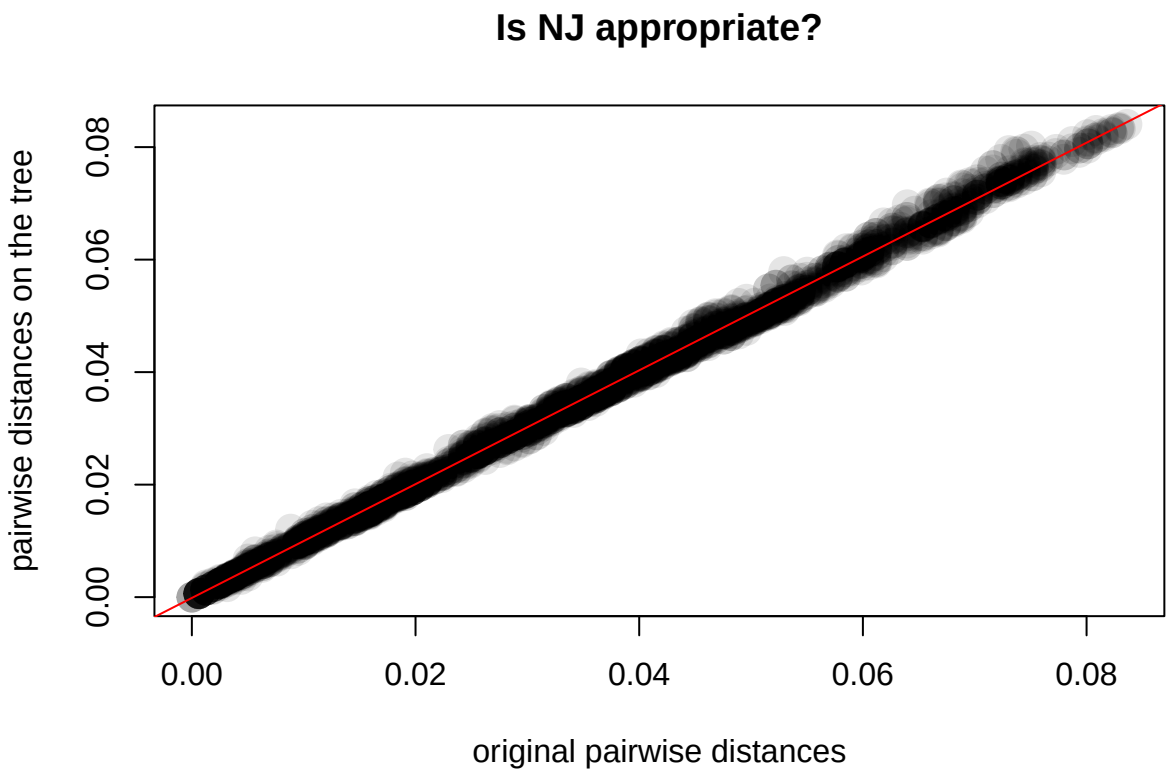

```
head(annot)
```

```
##   accession year                          misc
## 1  CY013200 1993 (A/New York/783/1993(H3N2))
## 2  CY013781 1993 (A/New York/802/1993(H3N2))
## 3  CY012128 1993 (A/New York/758/1993(H3N2))
## 4  CY013613 1993 (A/New York/766/1993(H3N2))
## 5  CY012160 1993 (A/New York/762/1993(H3N2))
## 6  CY012272 1994 (A/New York/729/1994(H3N2))
```

```
tre2 <- root(tre, out = 1)
tre2 <- ladderize(tre2)
plot(tre2, show.tip=FALSE, edge.width=2)
title("Rooted NJ tree")
tiplabels(tre$tip.label, bg=transp(num2col(annot$year, col.pal=myPal),.7), cex=.5, fg="transparent")
axisPhylo()
temp <- pretty(1993:2008, 5)
legend("topright", fill=transp(num2col(temp, col.pal=myPal),.7), leg=temp, ncol=2)
```



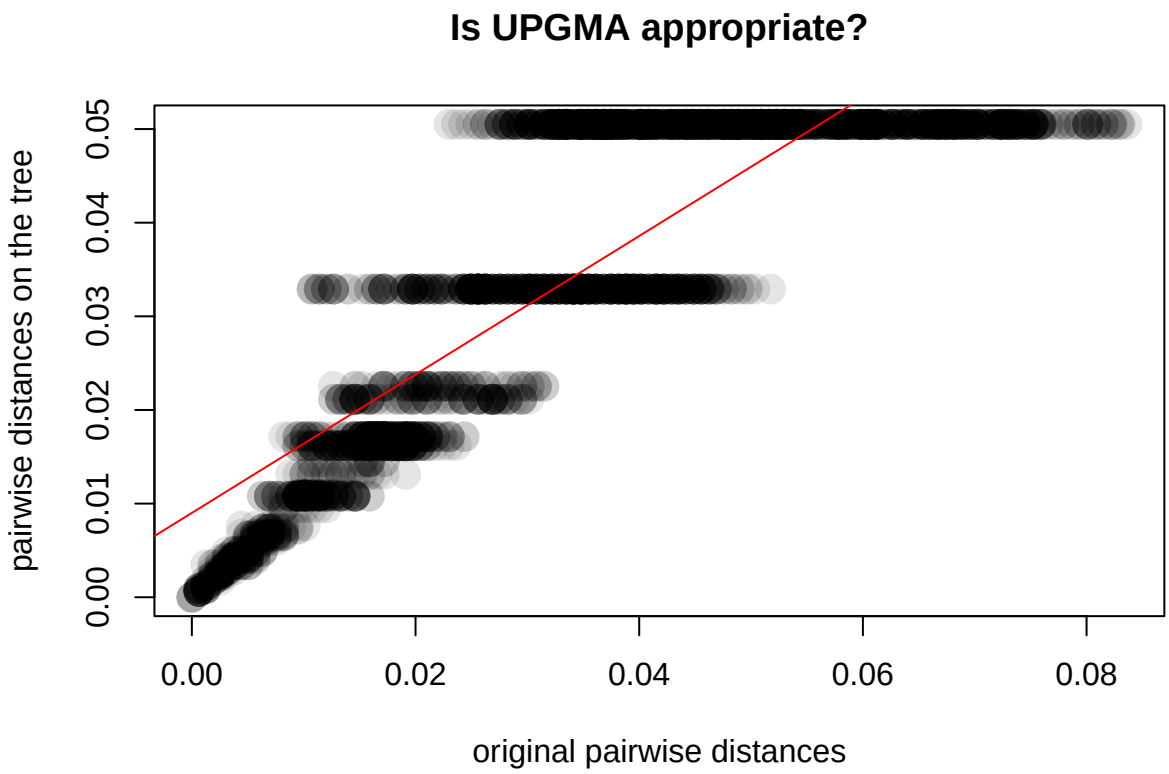

```
x <- as.vector(D)
y <- as.vector(as.dist(cophenetic(tre2)))
plot(x, y, xlab="original pairwise distances", ylab="pairwise distances on the tree", main="Is NJ appro
abline(lm(y~x), col="red")
```




```
cor(x,y)^2
```

```
## [1] 0.9975154
```

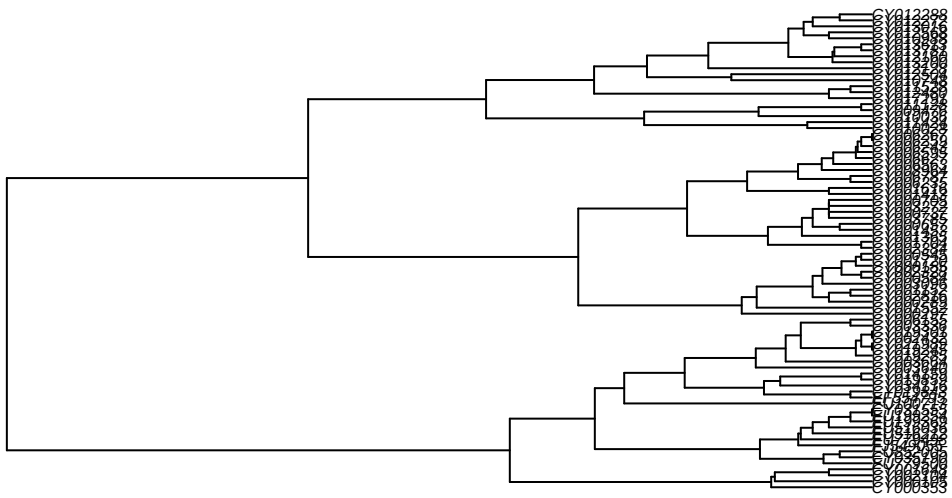
```
tre3 <- as.phylo(hclust(D,method="average"))
y <- as.vector(as.dist(cophenetic(tre3)))
plot(x, y, xlab="original pairwise distances", ylab="pairwise distances on the tree", main="Is UPGMA ap
abline(lm(y~x), col="red")
```




```
cor(x,y)^2
```

```
## [1] 0.7393009
```

```
plot(tre3, cex=.5)
title("UPGMA tree")
```



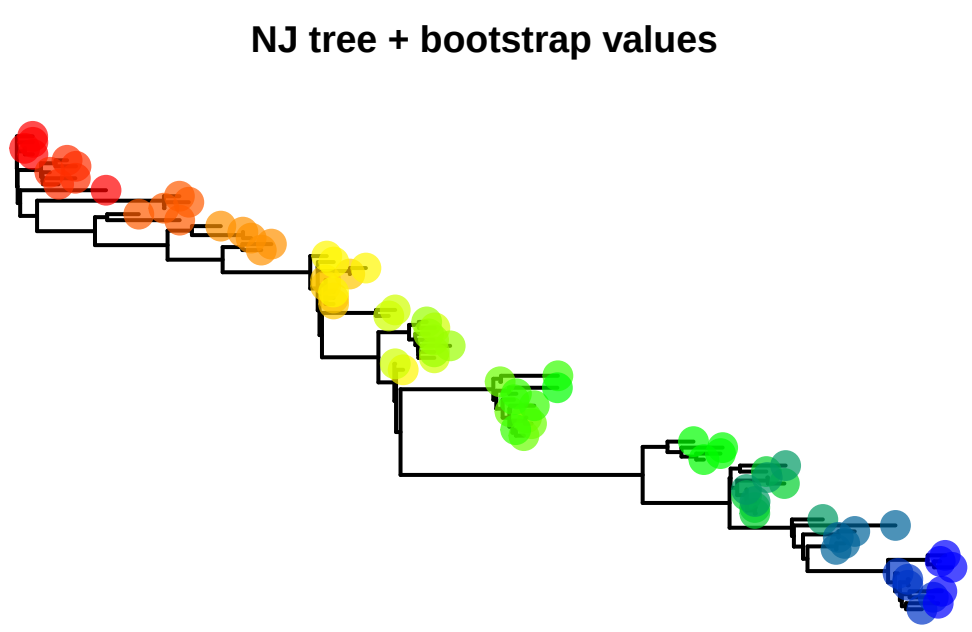

```
myBoots <- boot.phylo(tre2, dna, function(e) root(nj(dist.dna(e, model = "TN93")),1))
```

```
## Running bootstraps:       100 / 100
## Calculating bootstrap values... done.
```

```
myBoots
```

```
##  [1]  NA  38  34  27  73  44  64  51  47 100  99  70  20  54  93  58  34  16  15
## [20]  78  99 100  51  79  56  72  54  88  36  50  84 100  99  95 100  98  99 100
## [39]  92  80  67  50  26  65  92  43  44  86 100  99  86  90 100  39  64  75  94
## [58]  34  50  71  99 100  43  51  44  99  99 100  59  62  35  32  67  49  94  57
## [77] 100  65
```

```
plot(tre2, show.tip=FALSE, edge.width=2)
title("NJ tree + bootstrap values")
tiplabels(frame="none", pch=20, col=transp(num2col(annot$year, col.pal=myPal),.7), cex=3, fg="transparer
```

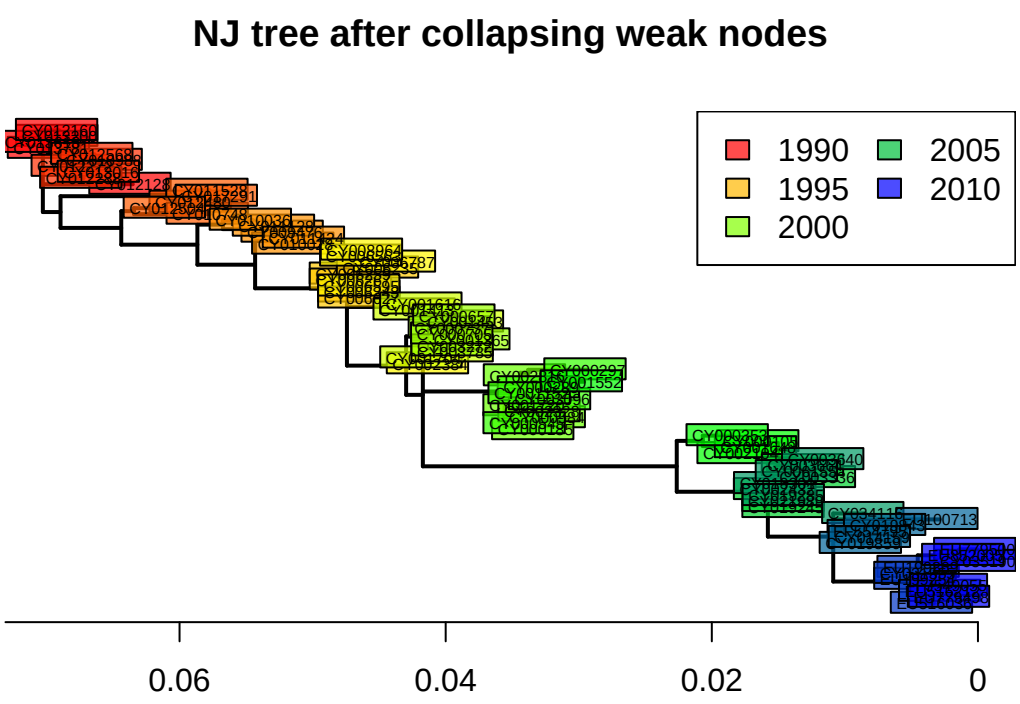



```
16
```

```
## [1] 16
```

```
#axisPhylo()
#temp <- pretty(1993:2008, 5)
#legend("topright", fill=transp(num2col(temp, col.pal=myPal),.7), leg=temp, ncol=2)
#nodelabels(myBoots, cex=.6)
```

```
temp <- tre2
N <- length(tre2$tip.label)
toCollapse <- match(which(myBoots<70)+N, temp$edge[,2])
temp$edge.length[toCollapse] <- 0
tre3 <- di2multi(temp, tol=0.00001)
plot(tre3, show.tip=FALSE, edge.width=2)
title("NJ tree after collapsing weak nodes")
tiplabels(tre3$tip.label, bg=transp(num2col(annot$year, col.pal=myPal),.7), cex=.5, fg="transparent")
axisPhylo()
temp <- pretty(1993:2008, 5)
legend("topright", fill=transp(num2col(temp, col.pal=myPal),.7), leg=temp, ncol=2)
```




```
dna2 <- as.phyDat(dna) #assign the original dna sequences data as a phyDat object...
class(dna2)
```

```
## [1] "phyDat"
```

```
dna2
```

```
## 80 sequences with 1701 character and 269 different site patterns.
## The states are a c g t
```

```
tre.ini <- nj(dist.dna(dna,model="raw"))
tre.ini
```

```
##
## Phylogenetic tree with 80 tips and 78 internal nodes.
##
## Tip labels:
##   CY013200, CY013781, CY012128, CY013613, CY012160, CY012272, ...
##
## Unrooted; includes branch lengths.
```

```
parsimony(tre.ini, dna2)
```

```
## [1] 422
```

```
tre.pars <- optim.parsimony(tre.ini, dna2)
```

```
## Final p-score 420 after  2 nni operations
```

```
tre.pars
```

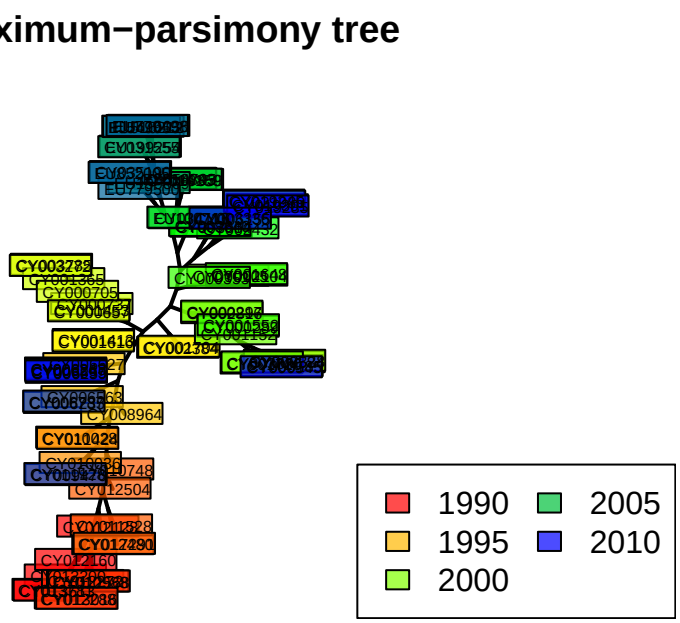
```
##
## Phylogenetic tree with 80 tips and 76 internal nodes.
##
## Tip labels:
##   CY013200, CY013781, CY012128, CY013613, CY012160, CY012272, ...
##
## Unrooted; no branch lengths.
```

```
parsimony(tre.pars, dna2)
```

```
## [1] 420
```

```
myPal <- colorRampPalette(c("red","yellow","green","blue"))
```

```
plot(tre.pars, type="unr", show.tip=FALSE, edge.width=2)
title("Maximum-parsimony tree")
tiplabels(tre.pars$tip.label, bg=transp(num2col(annot$year, col.pal=myPal),.7), cex=.5, fg="transparent
temp <- pretty(1993:2008, 5)
legend("bottomright", fill=transp(num2col(temp, col.pal=myPal),.7), leg=temp, ncol=2, bg=transp("white"
```



```
tre.ini <- nj(dist.dna(dna,model="TN93"))
pml(tre.ini, dna2, k=4)
```

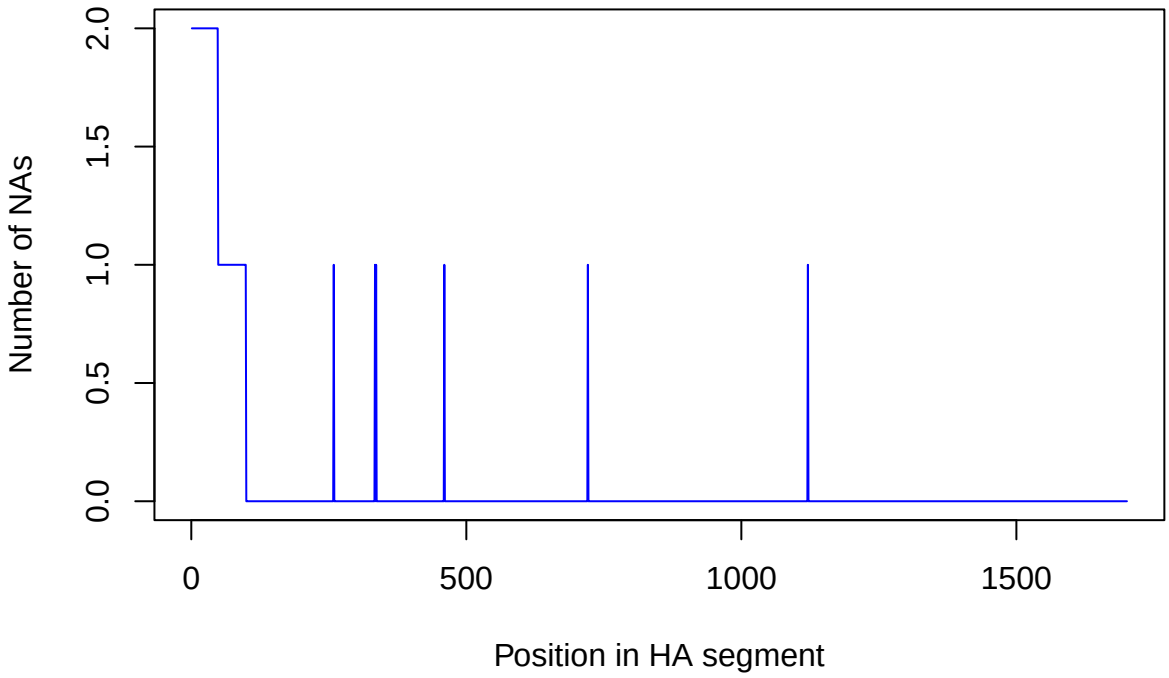
```
## model: JC+G(4)
## loglikelihood: -5641.785
## unconstrained loglikelihood: -4736.539
## Discrete gamma model
## Number of rate categories: 4
## Shape parameter: 1
##
## Rate matrix:
##   a c g t
## a 0 1 1 1
## c 1 0 1 1
## g 1 1 0 1
## t 1 1 1 0
##
## Base frequencies:
##    a    c    g    t
## 0.25 0.25 0.25 0.25
```

```
table(as.character(dna2))
```

```
##
##     -     a     c     g     k     m     r     s     t     w
##   147 45595 27170 30613     1     2     1     1 32549     1
```

```
na.posi <- which(apply(as.character(dna),2, function(e) any(!e %in% c("a","t","g","c"))))
```

```
temp <- apply(as.character(dna),2, function(e) sum(!e %in% c("a","t","g","c")))
plot(temp, type="l", col="blue", xlab="Position in HA segment", ylab="Number of NAs")
```




```
dna3 <- dna[,-na.posi]
table(as.character(dna3))
```

```
##
##     a     c     g     t
## 43402 25104 28828 30346
```

```
dna4 <- as.phyDat(dna3)
tre.ini <- nj(dist.dna(dna3,model="TN93"))
fit.ini <- pml(tre.ini, dna4, k=4)
fit.ini
```

```
## model: JC+G(4)
## loglikelihood: -5184.119
## unconstrained loglikelihood: -4043.367
## Discrete gamma model
## Number of rate categories: 4
## Shape parameter: 1
##
## Rate matrix:
##   a c g t
## a 0 1 1 1
## c 1 0 1 1
## g 1 1 0 1
## t 1 1 1 0
```

```
##
## Base frequencies:
##    a    c    g    t
## 0.25 0.25 0.25 0.25
```

```
fit <- optim.pml(fit.ini, optNni=TRUE, optBf=TRUE, optQ=TRUE, optGamma=TRUE)
```

```
## optimize edge weights:  -5184.094 --> -5166.996
## optimize base frequencies:  -5166.996 --> -5121.313
## optimize rate matrix:  -5121.313 --> -4933.871
## optimize shape parameter:  -4933.871 --> -4919.646
## optimize edge weights:  -4919.646 --> -4919.326
## optimize topology:  -4919.326 --> -4916.187  NNI moves:  2
## optimize base frequencies:  -4916.187 --> -4915.89
## optimize rate matrix:  -4915.89 --> -4915.868
## optimize shape parameter:  -4915.868 --> -4915.867
## optimize edge weights:  -4915.867 --> -4915.867
## optimize topology:  -4915.867 --> -4915.867  NNI moves:  0
## optimize base frequencies:  -4915.867 --> -4915.866
## optimize rate matrix:  -4915.866 --> -4915.866
## optimize shape parameter:  -4915.866 --> -4915.866
## optimize edge weights:  -4915.866 --> -4915.866
## optimize base frequencies:  -4915.866 --> -4915.866
## optimize rate matrix:  -4915.866 --> -4915.866
## optimize shape parameter:  -4915.866 --> -4915.866
## optimize edge weights:  -4915.866 --> -4915.866
## optimize base frequencies:  -4915.866 --> -4915.866
## optimize rate matrix:  -4915.866 --> -4915.866
## optimize shape parameter:  -4915.866 --> -4915.866
## optimize edge weights:  -4915.866 --> -4915.866
```

```
fit
```

```
## model: F81+G(4)
## loglikelihood: -4915.866
## unconstrained loglikelihood: -4043.367
## Discrete gamma model
## Number of rate categories: 4
## Shape parameter: 0.2829846
##
## Rate matrix:
##           a          c         g          t
## a 0.0000000  2.3836329 8.2983982  0.8563163
## c 2.3836329  0.0000000 0.1485362 10.0779972
## g 8.2983982  0.1485362 0.0000000  1.0000000
## t 0.8563163 10.0779972 1.0000000  0.0000000
##
## Base frequencies:
##         a         c         g         t
## 0.3415991 0.1953602 0.2243303 0.2387104
```

```
anova(fit.ini, fit)
```

```
## Likelihood Ratio Test Table
##   Log lik.  Df Df change Diff log lik. Pr(>|Chi|)
## 1  -5184.1 158
## 2  -4915.9 166         8        536.51  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(fit.ini)
```

```
## [1] 10684.24
```

```
AIC(fit)
```

```
## [1] 10163.73
```

```
tre4 <- root(fit$tree,1)
tre4 <- ladderize(tre4)
plot(tre4, show.tip=FALSE, edge.width=2)
title("Maximum-likelihood tree")
tiplabels(annot$year, bg=transp(num2col(annot$year, col.pal=myPal),.7), cex=.5, fg="transparent")
axisPhylo()
temp <- pretty(1993:2008, 5)
legend("topright", fill=transp(num2col(temp, col.pal=myPal),.7), leg=temp, ncol=2)
```