Taxonomic assignment of endophytic isolate E14504F

Dan Spakowicz

May 11, 2017

Introduction

This is the script used to build a tree for Nneoma's fungus and assign its taxonomy.

I started by rolling through the databases from the recent review http://jcm.asm.org/content/55/4/1011.full to check if any would be useful for this project.

- BOLD systems http://v4.boldsystems.org/index.php/IDS_OpenIdEngine only have ITS identification.
- Looks like this is a good place for morphological features https://aftol.umn.edu/ and can even make a nexus file to include in the tree – but aftol has been lost? (goes to godaddy website…)
- BROAD doesn't have an identification portal
- EZBioCloud doesn't have a fungal id section
- FungiDB is just genomics
- UNITE is just ITS
- IndexFungorum doesn't have an id search (but could be useful for morphology)
- CBS can be searched directly for LSU and there are lots of good hits. However, I'd rather find a paper that has gone through the effort of identifying isolates with multiple loci
- SILVA has an LSU search https://www.arb-silva.de/
- Identity: 43.61, LCA tax SILVA: None
- SSU Iden: 99.37, LCA tax. SILVA: None
- RDP http://rdp.cme.msu.edu/classifier/
- E14504F-LSU Root(100%) Fungi(100%) Basidiomycota(100%) Agaricomycetes(100%) Cantharellales(100%) Ceratobasidiaceae(100%) Thanatephorus(100%)

The RDP result is strong, with 100% confidence in the genus Thanatephorus. The CBS searches also found organisms of either Thanatephorus (telomorph) or Rhizoctonia (anamorph). This will very likely be the genus to which E14504F belongs. In addition, Nneoma and I found a few papers that deal with isolates of Rhizoctonia/Thanatephorus:

- [Gonzalez et al.] has a bunch of Thanatephorus isolates with genbank accession numbers for ITS and 28S, but nothing outside the genus (which is necessary to demonstrate the circumscription in this case).
- [Tupac Otero et al.] just have ITS and have several genera that were isolated from orchids. It's more orchid-centric than fungus-centric.
- [López-Chávez et al.] defines a Thanatephorus isolate using ITS alone. The tree shows weak node support separating Thanatephorus from Ceratobasidium, but clearly their isolate is closest to a Thana.
- [Gónzalez et al.] does a really nice job of creating a multi-locus tree. This should be the model going forward.

Methods

I converted the table of genbank accession numbers from [Gónzalez et al.] to a google spreadsheet.

Here are the files that Nneoma created using the Staden package (pregap & gap). As soon as these have genbank accession numbers I'll add them to the table so that they can be pulled with the other sequences from the table and remove this code block and the merge code block.

```
files <- list.files(path = "~/Dropbox/Rainforest project/E14504F sequences/", pattern = "*.fast
ofastas <- list()
for (f in files) {
  tmp <- readLines(f)
  tmp[2:length(tmp)] <- toupper(tmp[2:length(tmp)])
  ofastas[[f]] <- tmp
}

# Adjust names of list
CleanPaths <- function(names){
  tmp <- gsub(".*E14504F? (.*)\\.fasta", "\\1", names)
  tmp <- gsub(" partial| FULL| ver2", "\\1", tmp)
  return(tmp)
}
names <- CleanPaths(names(ofastas))
names(ofastas) <- names

# Cat ITS1 and ITS2 sep by 100 N's and remove ITS2
its <- grep("ITS", names(ofastas))
ofastas[[its[1]]] <- c(ofastas[[its[1]]], paste(rep("N", 100), collapse = ""), ofastas[[its[2]]
ofastas <- ofastas[-its[2]]
names(ofastas)[grep("ITS", names(ofastas))] <- "ITS"
```

```
# Load table into dataframe
sheet <- gs_title("E14504F")
x <- gs_read(sheet)

# Convert hyphen-only columns to NA
x <- data.frame(apply(x, 2, function(x) gsub("^-$", NA, x)), as.is = TRUE)

# Take in a character vector of genbank accession numbers and return a fasta file in ape forma
RetrieveSequencesAsString <- function(charvec){
  try({
    string <- entrez_fetch(db = "nucleotide", id = charvec, rettype = "fasta")
    string <- unlist(strsplit(string, split = "\n"))
    return(string)
  }, silent = TRUE)
}

# Retrieve all sequence into a list
loci <- colnames(x[,5:11])
```

```r
gfastas <- list()
for (i in loci) {
  gfastas[[i]] <- RetrieveSequencesAsString(x[,grep(i, colnames(x))])
}
# Remove those without any sequences
gfastas <- gfastas[-(which(sapply(gfastas, class) == "try-error"))]
```

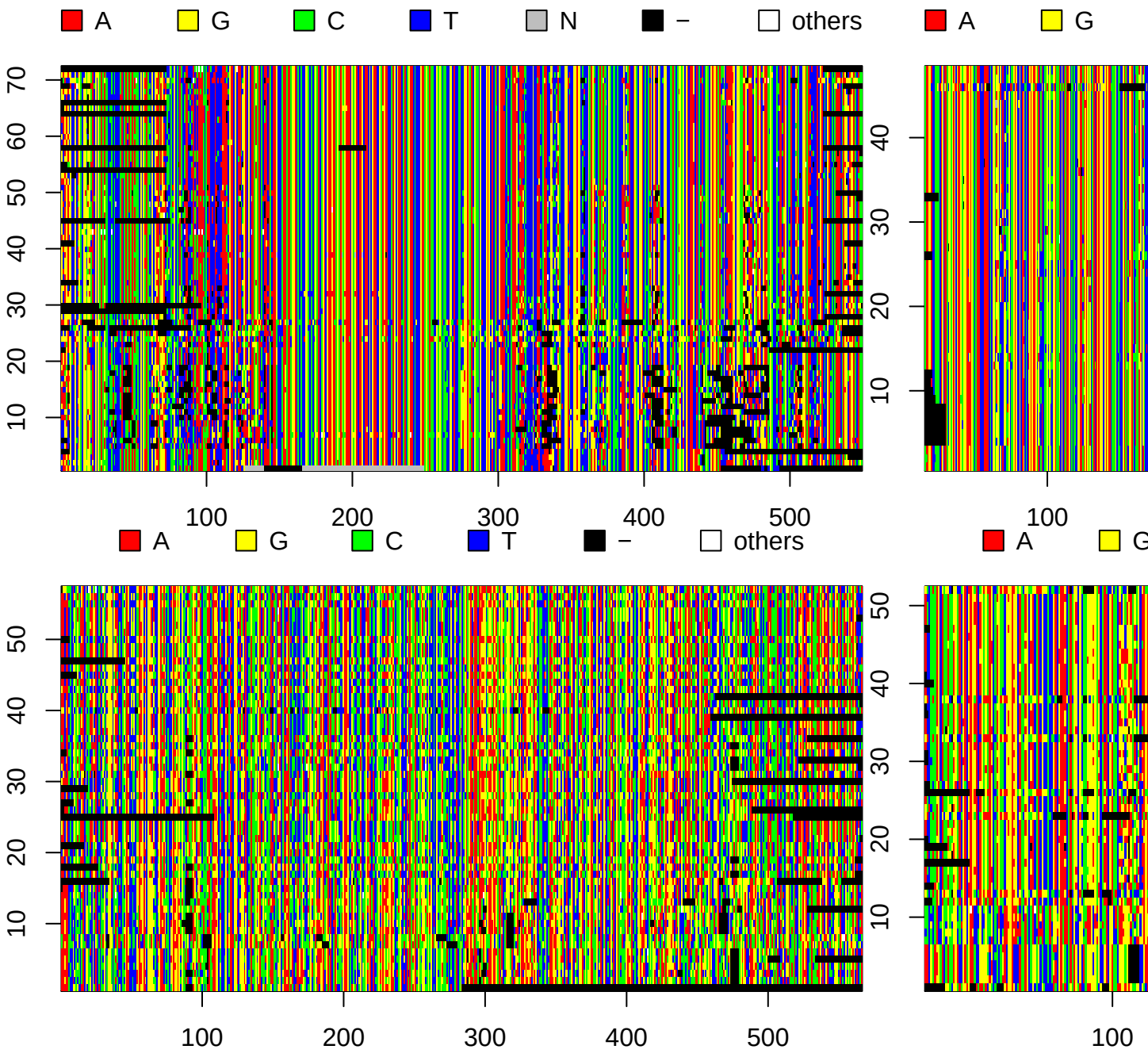
```r
bfastas <- list()
for (n in names(gfastas)){
  if (n %in% names(ofastas)) {
    bfastas[[n]] <- c(gfastas[[n]], ofastas[[which(names(ofastas) %in% n)]])
  } else {
    bfastas[[n]] <- gfastas[[n]]
  }
}
```

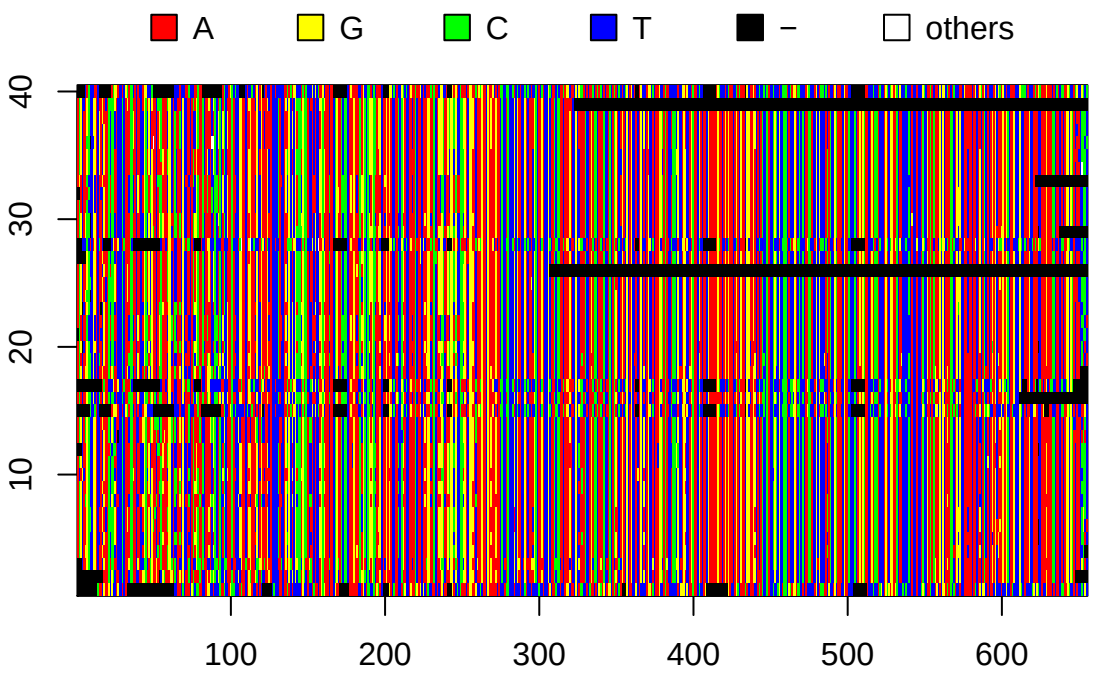
```r
# Read in each string as a fasta file
fastas <- list()
for (i in 1:length(bfastas)) {
  temp <- tempfile()
  write(bfastas[[i]], temp)
  fastas[[i]] <- read.FASTA(file = temp)
}

# Align sequences
alns <- lapply(fastas, ape::muscle)

# Remove gaps present in 30% of the columns
rmgaps <- lapply(alns, function(x) del.colgapsonly(x, threshold = 0.3))
names(rmgaps) <- names(bfastas)
```

```r
lapply(rmgaps, function(x) image.DNAbin(x, show.labels = FALSE))
```


A
G
C
T
N
−
others
A
G
70
60
50
40
30
20
10
100
200
300
400
500
40
30
20
10
100
A
G
C
T
−
others
A
G
50
40
30
20
10
100
200
300
400
500
50
40
30
20
10
100




```
## $ITS
## $ITS$rect
## $ITS$rect$w
## [1] 565.4039
##
## $ITS$rect$h
## [1] 10.82707
##
## $ITS$rect$left
## [1] -7.951949
##
## $ITS$rect$top
## [1] 85.99098
##
##
## $ITS$text
## $ITS$text$x
## [1]  23.35984 103.01355 182.66725 262.32096 341.97467 421.62837 501.28208
##
## $ITS$text$y
## [1] 80.57745 80.57745 80.57745 80.57745 80.57745 80.57745 80.57745
##
##
##
## $LSU
## $LSU$rect
## $LSU$rect$w
## [1] 670.4516
##
```

```
## $LSU$rect$h
## [1] 7.218045
##
## $LSU$rect$left
## [1] -9.475808
##
## $LSU$rect$top
## [1] 57.51256
##
##
## $LSU$text
## $LSU$text$x
## [1]  27.65347 122.10623 216.55898 311.01174 405.46450 499.91725 594.37001
##
## $LSU$text$y
## [1] 53.90353 53.90353 53.90353 53.90353 53.90353 53.90353 53.90353
##
##
##
## $RPB2
## $RPB2$rect
## $RPB2$rect$w
## [1] 500.7916
##
## $RPB2$rect$h
## [1] 8.571429
##
## $RPB2$rect$left
## [1] 32.85418
##
## $RPB2$rect$top
## [1] 68.19196
##
##
## $RPB2$text
## $RPB2$text$x
## [1]  65.13555 147.25577 229.37598 311.49620 393.61641 475.73663
##
## $RPB2$text$y
## [1] 63.90625 63.90625 63.90625 63.90625 63.90625 63.90625
##
##
##
## $TEF1
## $TEF1$rect
## $TEF1$rect$w
## [1] 376.0361
##
```

```
## $TEF1$rect$h
## [1] 7.819549
##
## $TEF1$rect$left
## [1] 24.73194
##
## $TEF1$rect$top
## [1] 62.25896
##
##
## $TEF1$text
## $TEF1$text$x
## [1]  48.97148 110.63419 172.29689 233.95960 295.62231 357.28501
##
## $TEF1$text$y
## [1] 58.34918 58.34918 58.34918 58.34918 58.34918 58.34918
##
##
##
## $ATP6
## $ATP6$rect
## $ATP6$rect$w
## [1] 579.538
##
## $ATP6$rect$h
## [1] 6.015038
##
## $ATP6$rect$left
## [1] 37.98099
##
## $ATP6$rect$top
## [1] 48.01975
##
##
## $ATP6$text
## $ATP6$text$x
## [1]  75.3384 170.3715 265.4046 360.4377 455.4708 550.5040
##
## $ATP6$text$y
## [1] 45.01223 45.01223 45.01223 45.01223 45.01223 45.01223
```

I can't figure out how to add titles in `image.DNAbin()`, but the order is "ITS" "LSU" "RPB2" "TEF1" "ATP6". The ITS looks to have a few orgs that might be revcomps, maybe three. I'm not sure how best to identify those and replace them in this format. It looks like the N length is short by ~ 20nt.

Alternative methods

- retreive Treebase file and add E14504F sequences to those alignments
- This is becomeing increasingly attractive given the seemingly large fraction of reverse complements observed. Particularly now that the ITS1 and ITS2 sequences are available, all sections except atp6 would be usable. It's worth taking the time to explore how to add one more sequence onto an existing alignment in R.

```
search_treebase()
```

Results and Discussion

References

Dolorez Gonzalez, Donald E. Carling, Shiro Kuninaga, Rytas Vilgalys, and Marc A. Cubeta. Ribosomal DNA systematics of ceratobasidium and thanatephorus with rhizoctonia anamorphs. 93(6): 1138–1150. ISSN 0027-5514. doi: 10.2307/3761674. URL http://www.jstor.org/stable/3761674.

Dolores Gónzalez, Marianela Rodriguez-Carres, Teun Boekhout, Joost Stalpers, Eiko E. Kuramae, Andreia K. Nakatani, Rytas Vilgalys, and Marc A. Cubeta. Phylogenetic relationships of rhizoctonia fungi within the cantharellales. 120(4):603–619. ISSN 1878-6146. doi: 10.1016/j.funbio.2016.01.012. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013834/.

Mariana Yadira López-Chávez, Karina Guillén-Navarro, Vincenzo Bertolini, Sergio Encarnación, Magdalena Hernández-Ortiz, Irene Sánchez-Moreno, and Anne Damon. Proteomic and morphometric study of the in vitro interaction between oncidium sphacelatum lindl. (orchidaceae) and thanatephorus sp. RG26 (ceratobasidiaceae). 26(5):353–365. ISSN 1432-1890. doi: 10.1007/s00572-015-0676-x.

J. Tupac Otero, James D. Ackerman, and Paul Bayman. Diversity and host specificity of endophytic rhizoctonia-like fungi from tropical orchids. 89(11):1852–1858. ISSN 0002-9122. doi: 10.3732/ajb.89.11.1852.