Phylogenetic tree for the taxa in sPlot 2.1

Oliver Purschke

01 Oktober, 2017

Abstract

This document describes the workflow (with contributions from Brody Sandel) that was used to generate the phylogenetic tree for the taxa in the global vegetation plot database sPlot (v. 2.1).

Contents

1 Load required packages

```
library(ape)
library(geiger)
library(dplyr)
```

2 Select taxon names in sPlot that are not in the reference phylogeny

2.1 Load the sPlot taxonomic backbone

```
load("../backbone.splot2.1.try3.is.vascular.Rdata")
```

```
head(backbone.splot2.1.try3[,c(2,32,34,35)])
```

```
##                     names.sPlot.TRY name.short.correct is.vascular.species
## 1                                 ?              <NA>                   NA
## 2                                 0              <NA>                   NA
## 3         [1269 Chlorophytum platt]       Chlorophytum                TRUE
## 4               [1284 Echinochloa]        Echinochloa                TRUE
## 5 [1285 Indigofera lange Bl-Stiele]         Indigofera                TRUE
## 6                   [1304 Polygala]           Polygala                TRUE
##   sPlot2.1.TRY
## 1            S
## 2            S
## 3            S
## 4            S
## 5            S
## 6            S
```

```
tail(backbone.splot2.1.try3[,c(2,32,34,35)])
```

```
##                               names.sPlot.TRY     name.short.correct
## 130597                    Chlorocyperus longus   Chlorocyperus longus
## 130598      Glyceria fluitans subsp. poiformis      Glyceria fluitans
## 130599                 Hydrodictyon utriculatum                   <NA>
## 130600 Bryum pseudotriquetrum var. duvalioides Bryum pseudotriquetrum
## 130601                   Hylocomium squarrosum  Hylocomium squarrosum
## 130602                    Carex hornschuchiana   Carex hornschuchiana
##        is.vascular.species sPlot2.1.TRY
## 130597                TRUE            S
## 130598                TRUE            S
## 130599                  NA            S
## 130600                  NA            S
## 130601                  NA            S
## 130602                TRUE            S
```

2.2 Select taxon names in sPlot

Because the taxonomic backbone includes combines and standardizes names from both, the sPlot and TRY databases, we need to exclude those names that are only found in the TRY database:

```
splot.nam <- backbone.splot2.1.try3[backbone.splot2.1.try3$sPlot.TRY != "T", c(4,32)]
```

How many entries in the backbone are only found in sPlot (S) as well as in both, sPlot and TRY (ST)?

```
table(splot.nam$sPlot.TRY)
```

```
##
##     S    ST
## 70329 24796
```

Generate vector of unique names.

```
splot.nam.2 <- unique(splot.nam$name.short.correct)
```

Check for `NA` entries and get rid of them:

```
any(is.na(splot.nam.2))
```

```
## [1] TRUE
```

```
na.ind <- which(is.na(splot.nam.2))
splot.nam.3 <- splot.nam.2[-na.ind]
any(is.na(splot.nam.3))
```

```
## [1] FALSE
```

```
length(splot.nam.3)
```

```
## [1] 61214
```

There are 61,214 unique names in the backbone that belong to sPlot as well to sPlot and TRY. However, some of them are non-vascular names and/or are only found in the previous version of sPlot (`sPlot 2.0`) but not in `sPlot 2.1`, which will be dealt with the last section.

Insert underscores to make names match with the `tip.labels` of the phylogeny.

```
splot.nam.4 <- gsub(" ", "_", splot.nam.3)
```

2.3 Select taxa in sPlot that are not in the phylogeny

I started with the fully resolved reference phylogeny for 31,749 vascular plants by Zanne *et al.* (2014), that was generated by David Tank and folks and therefore is referred to as `tank tree` in the subsequent text.

```
tank.tree <- read.tree("/home/oliver/Dokumente/PhD/PostPhD/IDiv/sDiv/sPlot/Analyses/Data/Phylogeny/2/Phy
```

```
ind.miss <- splot.nam.4 %in% tank.tree$tip.label
```

How many taxon names in sPlot are found/not found in the tank tree?:

```
table(ind.miss)
```

```
## ind.miss
## FALSE  TRUE
## 45784 15430
```

Only 25.2% of the taxa in sPlot are in the tank tree.

Before grafting the missing taxa onto the reference phylogeny, we might inspect those taxa, to see whether there are taxa that obviously should be in the tree.

Generate vector of the 45,784 names in sPlot that missing from the tree.

```
splot.nam.miss <- splot.nam.4[ind.miss==F]
```

Before grafting the missing species onto the reference phylogeny, we might inspect missing species to see whether there are species that obviously should be there.

```
write.csv(sort(splot.nam.miss), file = "splot.nam.miss.csv")
```

```
length(splot.nam.miss)
```

```
## [1] 45784
```

3 Grafting missing taxa onto the reference tree

3.1 Tree grafting function (R-code by Brody Sandel, Dec 2, 2014)

The following function `addCongeners` takes a tree and species list, each containing "Genus_species" or "Genus" formatted names. For each species in the list, if it has a congener on the tree but is not itself on the tree, add it next to a randomly selected congener at a random position along the edge returns a larger tree (see also the approach that was used to generate the BIEN-Phylogeny (Maitner *et al.*) here). Such approach has been demonstrated to introduce less bias into subsequent analysis than adding missing species as polytomies to the respective genera (Davies *et al.* 2012). I did not add species based on taxonomic information above genus level.

```
addCongeners = function(tree,speciesToAdd){
    resample <- function(x, ...) x[sample.int(length(x), ...)]
    staGenus = unlist(lapply(strsplit(speciesToAdd,"_"),function(i){i[[1]]}))
    gtree = tree
    for(i in sample(1:length(speciesToAdd),length(speciesToAdd),replace = F)){
        gtreeGenera = unlist(lapply(strsplit(gtree$tip,"_"),function(i){i[[1]]}))
        #If the species isn't on the tree but a congener is
        if(!speciesToAdd[i] %in% tree$tip.label & staGenus[i] %in% gtreeGenera){
            branchName = gtree$tip[resample(which(gtreeGenera == staGenus[i]),1)]
            newtree = sim.bdtree(n=2)
            newtree$tip.label = c(branchName,speciesToAdd[i])
            edgeL = gtree$edge.length[which.edge(gtree,branchName)]
            #Splice in at a random depth between 0 and 1
            depth = runif(1,0,1)
            newtree$edge.length = depth*newtree$edge.length*edgeL/max(newtree$edge.length)
            whereToGraft = which(gtree$tip == branchName)
            gtree = bind.tree(gtree,newtree,where = whereToGraft,position = edgeL*depth)
```

```
            #The grafting process duplicates the branchName tip. Drop one of them.
            gtree = drop.tip(gtree, which(gtree$tip == branchName)[1])}
    }
    return(gtree)
}
```

3.2 Run addCongeners function

```
tank.tree.added <- addCongeners(tank.tree, splot.nam.miss)
splot2.1.tank.tree.70287 <- tank.tree.added
```

Took 4.5 hours (on the RStudio-server) to add the missing 45,784 taxa in sPlot that were not included in the tank tree. This resulted in an extended tank tree with 70,287 taxa.

3.2.1 Read in the full extended tree

... and check which names could not be added:

```
splot2.1.tank.tree.70287 <- read.tree("/home/oliver/Dokumente/PhD/PostPhD/IDiv/sDiv/sPlot/Analyses/Data/
ind.miss.2 <- splot.nam.miss %in% splot2.1.tank.tree.70287$tip.label
splot.nam.miss.2 <- splot.nam.miss[ind.miss.2==F]
write.csv(sort(splot.nam.miss.2), file = "splot.nam.miss.2.csv")
length(splot.nam.miss.2)
```

```
length(splot.nam.miss.2)
```

```
## [1] 7246
```

Because they had no congeners in the reference phylogeny, 7,246 species could not be added (11.8% of all resolved taxa in sPlot). This resulted in a phylogeny with 53,967 names from a total of 61,214 standardized taxa in sPlot. Anyway, lichens, mosses as well as species that are not found in sPlot 2.1 are still included in that sPlot species list. We will take care of that in a later section.

3.3 Graft species onto the updated Qian & Jin (2016) tree

The reference phylogeny by Zanne *et al.* (2014) has recently been updated by Qian & Jin (2016), subsequently referred to as the `Qian tree`. So we will use that tree to generate our final phylogeny.

3.3.1 How many taxa are missing from the Qian tree?

```
phyto.phylo.tree <- read.tree("/home/oliver/Dokumente/PhD/PostPhD/IDiv/sDiv/sPlot/Analyses/Data/Phyloge
ind.miss <- splot.nam.4 %in% phyto.phylo.tree$tip.label
```

```
table(ind.miss)
```

```
## ind.miss
## FALSE  TRUE
## 45784 15430
```

16,121 names in sPlot are found in the Qian tree.

Generate vector of missing names:

```
splot.nam.miss.phyto.phylo <- splot.nam.4[ind.miss==F]
```

3.3.2 Run addCongeners on the Qian tree

```
phyto.phylo.tree.added <- addCongeners(phyto.phylo.tree, splot.nam.miss.phyto.phylo)
phyto.phylo.splot2.1.69335 <- phyto.phylo.tree.added
```

This resulted in an extended Qian tree with 69,335 taxa.

```
phyto.phylo.splot2.1.69335 <- read.tree("/home/oliver/Dokumente/PhD/PostPhD/IDiv/sDiv/sPlot/Analyses/Da

pdf("phyto.phylo.tree.splot.69335.pdf", height = 150, width = 150)
plot(phyto.phylo.splot2.1.69335, "f", cex = .1)
dev.off()

ind.miss.3 <- splot.nam.miss.phyto.phylo %in% phyto.phylo.splot2.1.69335$tip.label
```

How many taxa in sPlot could/could not be added to the Qian & Jin (2016) tree?:

```
table(ind.miss.3)
```

```
## ind.miss.3
## FALSE  TRUE
##  7147 37946
```

```
splot.nam.miss.3 <- splot.nam.miss.phyto.phylo[ind.miss.3==F]
write.csv(sort(splot.nam.miss.3), file = "splot.nam.miss.3.csv")
```

3.3.3 Prune tree to taxa in sPlot

```
pruned.tree <- drop.tip(phyto.phylo.splot2.1.69335, setdiff(phyto.phylo.splot2.1.69335$tip.label, splot

phyto.phylo.splot2.1.54067 <- pruned.tree

write.tree(phyto.phylo.splot2.1.54067, file = "phyto.phylo.splot2.1.54067.tre")

phyto.phylo.splot2.1.54067 <- read.tree("phyto.phylo.splot2.1.54067.tre")
```

```
16121/54067
```

Out of the total of 54,067 species that are now in the sPlot2.1 tree, 29.8% are actually found in the Qian tree. The remainder was added as congeners.

3.3.4 Plot the phylogeny

```
pdf("phyto.phylo.tree.splot.54067.pdf", height = 220, width = 220)
plot(phyto.phylo.splot2.1.54067, "f", edge.width = .4, label.offset = .03, cex = .1)
dev.off()

png("phyto.phylo.tree.splot.54067.png", height = 10000, width = 10000, pointsize = 1, res = 4000)
plot(phyto.phylo.splot2.1.54067, "f", edge.width = .1, label.offset = .03, cex = .1)
dev.off()
```

3.3.5 Tree for only vascular taxa in the most recent sPlot version (sPlot 2.1)

Select the respective set of species from the backbone.

Exclude `NA` entries from the backbone:

```
ind2.1 <- !is.na(backbone.splot2.1.try3$sPlot2.1.TRY)
back2.1 <- backbone.splot2.1.try3[ind2.1, ]
```

Generate a version of the backbone that only includes the unique resolved names in `name.short.correct`. For the non-unique names, keep the first row of duplicated names:

```
back2.1.uni <- back2.1[!duplicated(back2.1$name.short.correct), ]
back2.1.uni <- back2.1.uni[-1, ]
```

```
length(unique(back2.1.uni$name.short.correct))
```

```
## [1] 86760
```

Select vascular species only and exclude `NA` entries:

```
df.uni <- back2.1.uni %>%
    dplyr::filter(is.vascular.species == TRUE, !is.na(name.short.correct))
```

```
length(df.uni$name.short.correct)
```

```
## [1] 83677
```

```
table(df.uni$sPlot2.1.TRY)
```

```
##
##     S    ST     T
## 34105 20414 29158
```

3.3.6 Select names in sPlot 2.1

```
df.uni.splot <- df.uni %>%
    dplyr::filter(is.vascular.species == TRUE, !is.na(name.short.correct), df.uni$sPlot2.1.TRY!= "T")
```

```
length((df.uni.splot$name.short.correct))
```

```
## [1] 54519
```

```
splot_2.1_vasc_nam <- df.uni.splot$name.short.correct
length(splot_2.1_vasc_nam)
```

```
## [1] 54519
```

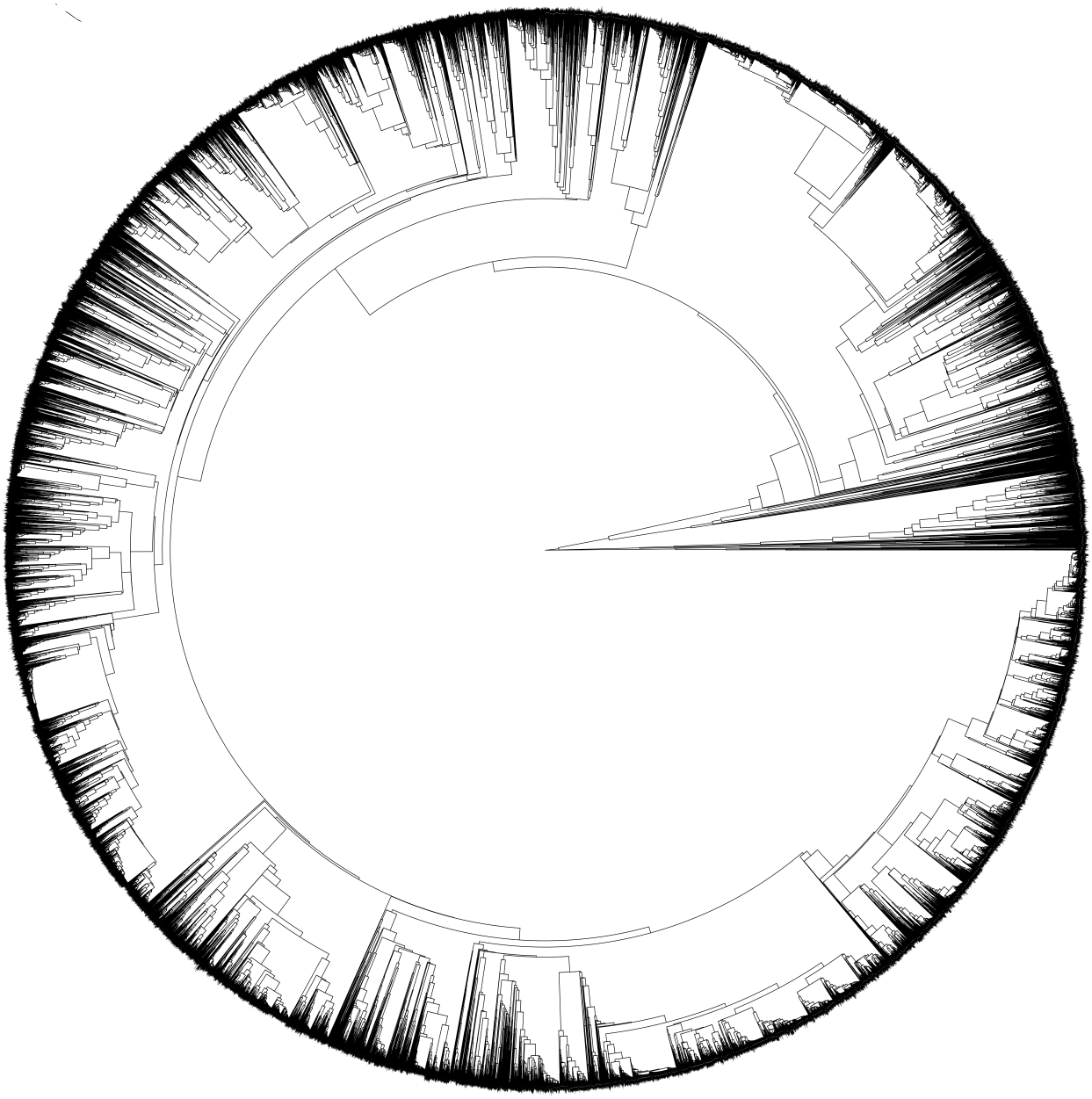
```
splot_2.1_vasc_nam_2 <- gsub(" ", "_", splot_2.1_vasc_nam)
```

4 Final tree

4.1 Prune tree to vascular plant taxa in sPlot 2.1

```
pruned.tree.splot2.1 <- drop.tip(phyto.phylo.splot2.1.69335, setdiff(phyto.phylo.splot2.1.69335$tip.labe

phyto.phylo.splot2.1.50167.vasc <- pruned.tree.splot2.1

write.tree(phyto.phylo.splot2.1.50167.vasc, file = "phyto.phylo.splot2.1.50167.vasc.tre")

phyto.phylo.splot2.1.50167.vasc <- read.tree("phyto.phylo.splot2.1.50167.vasc.tre")

pdf("phyto.phylo.splot2.1.50167.vasc.pdf", height = 220, width = 220)
plot(phyto.phylo.splot2.1.50167.vasc, "f", edge.width = .4, label.offset = .03, cex = .1)
dev.off()

png("phyto.phylo.splot2.1.50167.vasc.png", height = 10000, width = 10000, pointsize = 1, res = 4000)
plot(phyto.phylo.splot2.1.50167.vasc, "f", edge.width = .1, label.offset = .03, cex = .1)
dev.off()
```

```
knitr::include_graphics("/home/oliver/Dokumente/PhD/PostPhD/IDiv/sDiv/sPlot/Analyses/Code/sPlot_Phyloge
```



4.2 Some stats

From a total of 54,519 vascular taxon names in sPlot 2.1, 50,167 names (92%) are contained in the extended version of the Qian tree (extended by the names in sPlot 2.1).

How many vascular species in sPlot 2.1 are found in the Qian-tree?:

```
ind.miss <- splot_2.1_vasc_nam_2 %in% phyto.phylo.tree$tip.label
table(ind.miss)
```

```
## ind.miss
## FALSE  TRUE
## 39759 14760
```

There are 14,760 vascular names in sPlot 2.1 that are also found in the Qian reference tree (which contains 31,389 names), which are 47% from the total number of taxa in the Qian tree.

Of the in total 54,519 vascular taxon names in sPlot 2.1, 14,760 names (27.1%) are contained in the Qian tree.

---

5 R-settings

```
sessionInfo()
```

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.3 LTS
##
## Matrix products: default
## BLAS: /usr/lib/openblas-base/libblas.so.3
## LAPACK: /usr/lib/libopenblasp-r0.2.18.so
##
## locale:
##  [1] LC_CTYPE=de_DE.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_DE.UTF-8        LC_COLLATE=de_DE.UTF-8
##  [5] LC_MONETARY=de_DE.UTF-8    LC_MESSAGES=de_DE.UTF-8
##  [7] LC_PAPER=de_DE.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_DE.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] png_0.1-7     bindrcpp_0.2  dplyr_0.7.3   rmarkdown_1.6 geiger_2.0.6
## [6] ape_4.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.12     knitr_1.17       bindr_0.1        magrittr_1.5
##  [5] MASS_7.3-47      lattice_0.20-35  R6_2.2.2         rlang_0.1.2
##  [9] subplex_1.4-1    stringr_1.2.0    tcltk_3.4.1      tools_3.4.1
## [13] parallel_3.4.1   nlme_3.1-131     coda_0.19-1      htmltools_0.3.6
## [17] yaml_2.1.14      rprojroot_1.2    digest_0.6.12    assertthat_0.2.0
## [21] tibble_1.3.4     glue_1.1.1       deSolve_1.20     evaluate_0.10.1
## [25] stringi_1.1.5    compiler_3.4.1   backports_1.1.1  mvtnorm_1.0-6
## [29] pkgconfig_2.0.1
```

References

Davies, T.J., Kraft, N.J., Salamin, N. & Wolkovich, E.M. (2012) Incompletely resolved phylogenetic trees inflate estimates of phylogenetic conservatism. *Ecology*, **93**, 242–247.

Maitner, B.S., Boyle, B., Casler, N., Condit, R., Donoghue, J., Durán, S.M., Guaderrama, D., Hinchliff, C.E., Jørgensen, P.M., Kraft, N.J., McGill, B., Merow, C., Morueta-Holme, N., Peet, R.K., Sandel, B., Schildhauer,

M., Smith, S.A., Svenning, J.-C., Thiers, B., Violle, C., Wiser, S. & Enquist, B.J. The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods in Ecology and Evolution*, n/a–n/a.

Qian, H. & Jin, Y. (2016) An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *Journal of Plant Ecology*, **9**, 233–239.

Zanne, A.E., Tank, D.C., Cornwell, W.K., Eastman, J.M., Smith, S.A., FitzJohn, R.G., McGlinn, D.J., O'Meara, B.C., Moles, A.T., Reich, P.B. & others. (2014) Three keys to the radiation of angiosperms into freezing environments. *Nature*, **506**, 89.