Phylogenetic_coverage_RMBL

Reed

3/28/2021

Althought the pie chart has mostly been relegated as a data visualization tool of the past, I think that is occasionally still does have it's uses. Acomplication with the size o

To show the extent of the available A353 sequence data, relative to the taxa at RMBL we will create a circular phylogenetic tree, and then place a circle plot indicating PRESENCE or ABSENCE of a sequence for the genus.

This will require 3 sets of data: A) The Current list of all PAFTOL sequence data. B) Our additional sequences C) The Vascular Plant Checklist for RMBL.

We will then make these data more coarse, to the resolution of genus. Naturally we will subset PAFTOL to the genera at RMBL to reduce extraneous tips hence branches.

Molecular import and wrangle.

```
non_biotic <- read.csv("../data/non_biotic_poll.csv")[,2]

rmbl_plants <- read.csv('../data/gothic_plant_list.csv') %>%
  group_by(genus) %>%
  mutate(no_species = n()) %>%
  dplyr::select(-species) %>%
  filter(!family %in% non_biotic) %>%
  ungroup() %>%
  mutate(taxid = name2taxid(binomial, db = 'ncbi'))

rmbl_genera <- rmbl_plants %>%
  distinct(genus) %>%
  pull(genus)

eastwood_summary <- read.delim('../data/eastwood_db_summary.csv', header = F,
                               col.names = c('prc_mini_map2_root', 'no_mini_root',
                                             'no_mini_map_clade', 'rank', 'taxid', 'taxon')
                               ) %>%
  mutate(taxon = str_trim(taxon)) %>%
  filter(rank == 'S',!taxid %in% c(7, 13055)) %>%
  separate(taxon, into = c('genus', 'species'), extra = "drop") %>%
  group_by(genus) %>%
  mutate(no_species = n()) %>%
  dplyr::select(-species) %>%
  distinct() %>%
  ungroup()

rm(non_biotic)
```

Morphological Import and wrangle

```
new <- read.csv("../data/Pollen_Label_box.csv") %>%
  separate(Taxon, into = c('genus', 'species'), extra = "drop") %>%
  select(genus, Family)

old <- read.csv("../data/existing_pollen_reference_slides.csv") %>%
  select('genus' = Genus, Family)

pollen_slides <- rbind(new, old)

rm(new, old)
```

```
obs <- read.csv("../data/Bombus_queen_observations_2015.csv")  %>%
  distinct(plant.species) %>%
  separate(plant.species, into = c('genus', 'species'), sep = '[.]' ) %>%
  mutate(genus = str_replace(genus, 'Distegia', 'Lonicera'),
         genus = str_replace(genus, 'Erythrocoma', 'Geum'),
         genus = str_replace(genus, 'Adenolinum', 'Linum'),
         genus = str_replace(genus, 'Ligularia', 'Hymenoxys')) %>%
  select(genus)
```

Retrieve OTT id's to request an Open Tree of Life Phylogeny. We will not be able to retrieve non-monophyletic genera, so we will try to pull out a single species from each genus and search for thata so we can at lest have some thing...

```
resolved_names <- tnrs_match_names(rmbl_genera, context = "Land plants")
resolved_names$in_tree <- is_in_tree(resolved_names$ott_id) # this will tell us whether the genus is in

non_monophyletic_genera <- resolved_names %>%
  filter(in_tree == "FALSE")  %>%
  pull(unique_name)
non_monophyletic_genera <- as.vector(non_monophyletic_genera)

species_to_q <- rmbl_plants %>%
  filter(genus %in% non_monophyletic_genera) %>%
  group_by(genus) %>%
  sample_n(1) %>%
  pull(binomial)

rmbl_genera1 <- rmbl_genera[!rmbl_genera %in% non_monophyletic_genera]
rmbl_queries <- c(rmbl_genera1, species_to_q)

rmbl_queries <- rmbl_queries %>%
  str_replace("Actaea", "actaea rubra") %>%  # I ran the code below, and came back to fix these by hand
  str_replace("Urtica", "urtica dioica") %>%  # fix a few records by hand...
  str_replace("Viola", "Viola praemorsa") %>%  # just threw in random names,
  str_replace("Heuchera", "heuchera cylindrica")

resolved_names <- tnrs_match_names(rmbl_queries, context = "Land plants")
resolved_names <- resolved_names %>% drop_na(ott_id)
resolved_names$in_tree <- is_in_tree(resolved_names$ott_id)
# our tree misses two genera of 298, not bad...
```

```
resolved_names <- resolved_names %>% filter(in_tree == "TRUE")

tree <- tol_induced_subtree(ott_ids = resolved_names$ott_id)

tree[["tip.label"]] <- sub("_.*", "",tree[["tip.label"]])
tip_label_order <- as.data.frame(sub("_.*", "",tree[["tip.label"]])) # we need these to annotate the tr
colnames(tip_label_order) <- "genus"

ape::write.tree(tree, file = "../data/rmbl_tree", append = FALSE, digits = 10, tree.names = FALSE)

rm(non_monophyletic_genera, rmbl_queries, rmbl_genera1, resolved_names, species_to_q)
```

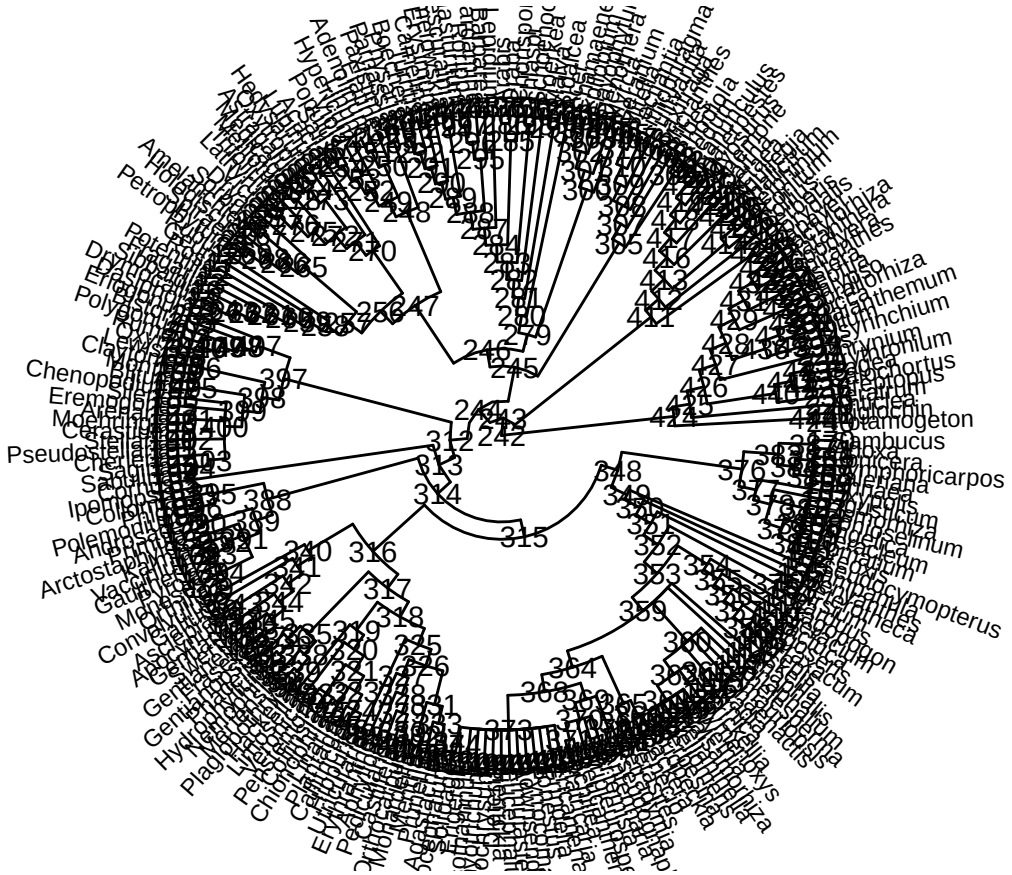
```
## Warning in rm(non_monophyletic_genera, rmbl_plants, rmbl_genera): object
## 'non_monophyletic_genera' not found
```

We can plot the entirety of the taxa for which we have the tree with names and the nodes labelled.

```
tree <- ape::read.tree(file = "../data/rmbl_tree")
tree[["tip.label"]] <- sub("_.*", "",tree[["tip.label"]])
tip_label_order <- as.data.frame(sub("_.*", "",tree[["tip.label"]])) # we need these to annotate the tr
colnames(tip_label_order) <- "genus"

rmbl_tree <- ggtree(tree, size=1.0, branch.length='none', layout='circular')
rmbl_named <- ggtree(tree, branch.length='none', layout='circular') +
  geom_tiplab(size = 3) +
  geom_text(aes(label=node))

rotate(rmbl_named, 348)
```




```
ggsave('../results/named_tree.png', device = 'png', width = 24, height = 24, units = 'in')

rm(rmbl_named)
```

Annotate tree and things

```
sequenced <- eastwood_summary %>%
  dplyr::select(genus) %>%
  distinct() %>%
  mutate(SEQUENCED = 'sequenced')

grouptreeSEQ <- left_join(tip_label_order, sequenced, by = "genus")%>%
  mutate(SEQUENCED = replace_na(SEQUENCED, replace = 'lacking')) %>%
  dplyr::select(SEQUENCED)
rownames(grouptreeSEQ) <- tree$tip.label

rm(eastwood_summary, sequenced)
```

```
pollen_slides_prepped <- pollen_slides %>%
  dplyr::select(genus) %>%
  distinct() %>%
  mutate(SLIDE = 'slide')

grouptreeMORPH <- left_join(tip_label_order, pollen_slides_prepped, by = "genus")%>%
  mutate(SLIDE = replace_na(SLIDE, replace = 'lacking')) %>%
```

```
  dplyr::select(SLIDE)

rm(pollen_slides, pollen_slides_prepped)
```

```
grouptreeOBS <- obs %>%
  mutate(OBSERVATION = 'observed') %>%
  left_join(tip_label_order, ., by = "genus") %>%
  mutate(OBSERVATION = replace_na(OBSERVATION, replace = 'lacking')) %>%
  distinct(genus, .keep_all = T) %>%
  dplyr::select(OBSERVATION)
```

```
node_dist <- data.frame(
  node = c(
    c(349:375, 105:156), # Asterales
    c(444, 240:241), # Alismatales
    c(428:439, 221:233), # Asparagales
    c(377:382, 157:165), # Apiales
    c(335:339, 90:95), # Boraginales
    c(282:298, 35:52), # Brassicales
    c(397:410, 186:205), #Caryophyllales
    c(255,8:9),  # Celastrales
    c(185), # Cornales
    c(383:387, 166:171),  # Dipsacales
    c(388:397, 172:184), # Ericales
    c(270:278, 24:33), # Fabales
    c(341:347, 97:104), # Gentianales
    c(55), # geraniales
    c(318:335, 70:89), # Lamiales
    c(440:443, 234:238, 220), # Liliales
    c(249:254, 1:7), # Malpighiales
    c(299, 53:54), # Malvales
    c(301:304, 56:60), # Myrtales
    c(411:423, 206:219), # Ranunculales
    c(257:269, 10:23), # Rosales
    c(34), # sapindales
    c(305:311, 61:69), # saxifragaceae
    c(96), # solanales
    c(221) # zingiberales
  ),
  order = c(
    rep("Asterales", each = length(c(349:375, 105:156))),
    rep("Alismatales", each = length(c(444, 240:241))),
    rep("Asparagales", each = length(c(428:439, 221:233))),
    rep("Apiales", each = length(c(377:382, 157:165))),
    rep("Boraginales", each = length(c(335:339, 90:95))), # #
    rep("Brassicales", each = length(c(282:298, 35:52))),
    rep("Caryophyllales", each = length(c(397:410, 186:205))),
    rep("Celastrales", each = length(c(255,8:9))),
    rep("Cornales", each = length(c(185))),
    rep("Dipsacales", each = length(c(383:387, 166:171))),
    rep("Ericales", each = length(c(388:397, 172:184))),
    rep("Fabales", each = length(c(270:278, 24:33))),
    rep("Gentianales", each = length(c(341:347, 97:104))),
```

```
    rep('Geraniales', each = length(c(55))),
    rep("Lamiales", each = length(c(318:335, 70:89))),
    rep("Liliales", each = length(c(440:443, 234:238, 220))),
    rep("Malpighiales", each = length(c(249:254, 1:7))),
    rep("Malvales", each = length(c(299, 53:54))),
    rep("Myrtales", each = length(c(301:304, 56:60))),
    rep("Ranunculales", each = length(c(411:423, 206:219))),
    rep("Rosales", each = length(c(257:269, 10:23))),
    rep("Sapindales", each = length(c(34))),
    rep('Saxifragales', each = length(c(305:311, 61:69))),
    rep("Solanales", each = length(c(96))),
    rep('Zingiberales', each = length(c(221)))
  )
)

upper <- data.frame(
  node = c(
    243,
    c(242,424:427),
    244,
    245:246,
    c(247:248, 256:257),
    c(279:281, 300),
    c(312:313),
    c(314:315),
    c(316:317, 340),
    c(348,376)),
  order = c(
    'Early',
    rep('Monocots', each = length(c(242,424:427))),
    'Eudicots',
    rep('Superrosids', each = length(c(245:246))),
    rep('Fabids', each = length(c(247:248, 256:257))),
    rep('Malvids', each = length(c(279:281, 300))),
    rep('Superasterids', each = length(c(312:313))),
    rep('Asterids', each = length(c(314:315))),
    rep('Lamiids', each = length(c(316:317, 340))),
    rep('Campanulids', each = length(c(348,376)))
  )
)

node_dist <- bind_rows(node_dist, upper)
APalG <- read.csv('../data/APG-hexCodes.csv')

node_dist <- left_join(node_dist, APalG, by = 'order')

rm(upper, tip_label_order)
```

Finally create the tree here

```
labDataDF <- cbind(grouptreeOBS, grouptreeMORPH, grouptreeSEQ)

rm(grouptreeOBS, grouptreeMORPH, grouptreeSEQ)
```

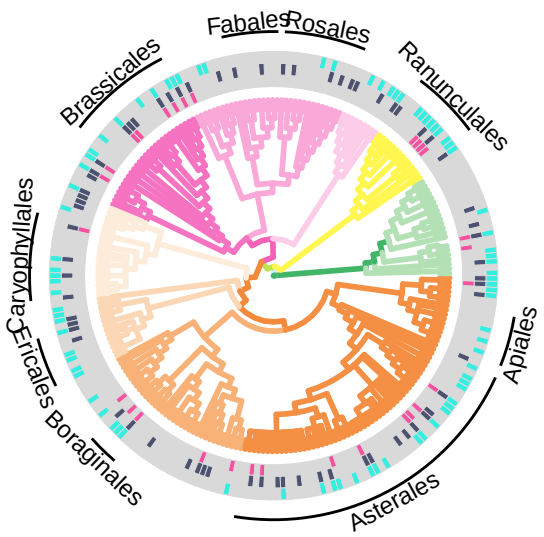
```
ob <- gheatmap(rmbl_tree, labDataDF, offset=.8, width=.2,
               colnames_angle=95, colnames_offset_y = .25, colnames = F, color = 'grey85') +
  scale_fill_manual('Status', values = c('grey85', '#f652a0', '#36eee0', '#4c5270') ) +

  theme(plot.title = element_text(hjust = 0.5), legend.title.align=0.5,
        legend.position="bottom") +
  ggtitle("Biotically pollinated plant genera \n with morphological or molecular data")
```

```
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```

```
ob1 <- ob %<+%
  node_dist +
  aes(color=I(color))

ob2 <- rotate(ob1, 246)
rotate(ob2, 348) +
  geom_cladelab(node=412, label="Ranunculales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=282, label="Brassicales",  angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=270, label="Fabales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=257, label="Rosales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=398, label="Caryophyllales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=389, label="Ericales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=335, label="Boraginales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=377, label="Apiales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5) +
  geom_cladelab(node=350, label="Asterales", angle='auto', color='white', fontsize = 3.0,
               offset=7, offset.text = 1, align = T, horizontal = F, hjust = 0.5)
```




```
ggsave('../results/rmbl_draft_tree.png', device = 'png', width = 5, height = 5, units = 'in')
```

```
rm(APalG, ob, ob1, ob2, obs, node_dist, labDataDF, rmbl_tree)
```