Phylogeny Statistics

Background

After creating the community-level phylogeny of the angiosperm species found in Brazil, we wanted to get a better idea of what the phylogeny contained. Using proportion tables, we can get a numeric summary of the make-up of the phylogeny for multiple variables of interest. In Part 1, we look at the Family, Lifeform, Phytogeographic Domain and State (administrative region) make-up of the species that compose our phylogeny. In Part 2, we create a list of all the species which were not used to build the phylogeny and we also look at the average node support of the phylogeny to gauge the quality of the phylogeny.

Part I – Phylogeny Content Statistics

Packages

```
library(tidyverse) # for general data manipulation
```

Data

The data used for the descriptive statistics of the phylogeny comes from two sources: the overall dataset used to build the phylogeny and a tab-delimited table exported by Sequence Matrix v1.8 containing the names of the taxa used in the final alignment, along with the identity and length of the gene regions sequenced for those taxa. The two data frames are joined together using `dplyr::semi_join()` to create a single data frame containing information on the family, phytogeographic domain, lifeform and location (state) for the species found in the phylogeny. The composite taxa from outside of Brazil were removed for these statistics.

```
# tibble containing information for all species found in Brazil

(data <- read_csv('nordeste.csv'))

# tibble containing all taxa found in the phylogeny

(phylo <- read_csv('alignment_table.csv'))

# remove OUTSIDE taxa from the data; total of 248 OUTSIDE taxa removed
(phylo <- phylo %>%
  filter(str_detect(Taxon, 'OUTSIDE', negate = TRUE)) %>%
  mutate(Combination = str_replace(Taxon, '.*aceae\\s', '')))

### phylo$Combination <- sub('.*aceae ', '', phylo$Taxon)

# select only the species from the data which were used to create the phylogeny
(stats <- semi_join(data, phylo, by = 'Combination'))

# 6 taxa were renamed during the ncbi search and their synonym equivalent
# in the data could not be found; they are tacked on to the end of
# the stats data for completeness as they were used in the tree
(stats <- phylo %>%
  select(Combination) %>%
  setdiff(select(stats, Combination)) %>%
```

```
  mutate(Family = c('Bromeliaceae', 'Fabaceae', 'Fabaceae',
                    'Fabaceae', 'Poaceae', 'Poaceae')) %>%
  full_join(stats, .))
```

The tbl_func Function

The function tbl_func was created to count instances of a desired variable in both the overall set of data and the subset containing only the species found in the phylogeny. tbl_func also provides the percentage value of the variable to the whole in the data_Percentage and tree_Percentage columns. Lastly, tbl_func creates a column (total_Percentage) indicating the percentage of overall data that is found in the phylogeny for a particular variable.

```
tbl_func <- function(x, y, States = FALSE){ # function to create % tables

    # x -- variable from the overall set of data
    # y -- variable from the phylogeny subset
    # States -- whether or not the variable of interest is the location (state)

    # data frame containing counts and percentages of overall data
    tbl.x <- x
    res.x <- data.frame(cbind(row.names(tbl.x), tbl.x,
      if(States == TRUE) {
        round(tbl.x / nrow(data) * 100, 2)
      } else {
        round(prop.table(tbl.x) * 100, 2)}))
    colnames(res.x) <- c('id', 'data_Count', 'data_Percentage')

    # data frame containing counts and percentages of phylogeny data
    tbl.y <- y
    res.y <- data.frame(cbind(row.names(tbl.y), tbl.y,
      if(States == TRUE) {
        round(tbl.y / nrow(stats) * 100, 2)
        } else {
          round(prop.table(tbl.y) * 100, 2)}))
    colnames(res.y) <- c('id', 'tree_Count', 'tree_Percentage')

    # join the above data frames and add a column showing the
    # percentage of overall data that is found in the tree
    res <- full_join(res.x, res.y, by = 'id')

    # rename any empty id names 'Unknown'
    (res <- res %>%
      mutate(id = str_replace(id, '^$', 'Unknown')) %>%
      mutate_all(type.convert) %>%
      mutate_if(is.numeric, ~replace_na(., 0)) %>%
      # column showing the percentage of overall data that is found in the tree
      mutate(total_Percentage = round(tree_Count / data_Count * 100, 2)))
}
```

Descriptive Statistics

Family

We can use the Family columns of our data sets to illustrate the use of `tbl_func`. The two variables for the function are `table(data$Family)` and `table(stats$Family)` and these provide a count of the number of taxa belonging to each family in the data sets and the function does the rest.

As the table shows, in the overall data, there are 1,328 members of the Fabaceae family, comprising 8.12% of all taxa in the data set. There are 601 members of the Fabaceae family in the phylogeny, comprising 14.17% of all taxa in the tree. Of all 1,328 Fabaceae in the overall data set, only 45.26% of them were used in the creation of the phylogeny. The function has also been applied to the Lifeform, Phyotogeographic Domain and State columns below. Please note that the taxa from outside of Brazil were not included in these statistics.

```
tbl_func(table(data$Family), table(stats$Family)) %>%
  arrange(desc(tree_Percentage)) %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = c('repeat_header', 'striped', 'hold_position'),
                repeat_header_text = 'Family (continued)',
                repeat_header_method = 'replace',
                font_size = 9)
```

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
Fabaceae	1328	8.12	601	14.17	45.26
Orchidaceae	1174	7.18	388	9.15	33.05
Poaceae	722	4.42	299	7.05	41.41
Asteraceae	1272	7.78	289	6.81	22.72
Bromeliaceae	635	3.88	190	4.48	29.92
Apocynaceae	405	2.48	150	3.54	37.04
Malpighiaceae	306	1.87	124	2.92	40.52
Bignoniaceae	268	1.64	117	2.76	43.66
Melastomataceae	708	4.33	107	2.52	15.11
Cyperaceae	372	2.28	104	2.45	27.96
Myrtaceae	524	3.21	98	2.31	18.70
Rubiaceae	552	3.38	82	1.93	14.86
Gesneriaceae	72	0.44	64	1.51	88.89
Euphorbiaceae	496	3.03	63	1.49	12.70
Cactaceae	152	0.93	60	1.41	39.47
Malvaceae	399	2.44	60	1.41	15.04
Solanaceae	243	1.49	57	1.34	23.46
Lamiaceae	279	1.71	52	1.23	18.64
Passifloraceae	80	0.49	49	1.16	61.25
Araceae	168	1.03	46	1.08	27.38
Annonaceae	115	0.70	44	1.04	38.26
Piperaceae	206	1.26	44	1.04	21.36
Lauraceae	180	1.10	37	0.87	20.56
Lentibulariaceae	61	0.37	37	0.87	60.66
Convolvulaceae	266	1.63	36	0.85	13.53
Dioscoreaceae	75	0.46	36	0.85	48.00
Sapotaceae	95	0.58	35	0.83	36.84
Celastraceae	78	0.48	33	0.78	42.31
Arecaceae	118	0.72	30	0.71	25.42
Moraceae	92	0.56	26	0.61	28.26
Polygalaceae	117	0.72	25	0.59	21.37
Sapindaceae	205	1.25	25	0.59	12.20
Phyllanthaceae	75	0.46	24	0.57	32.00
Dilleniaceae	36	0.22	23	0.54	63.89
Alstroemeriaceae	28	0.17	22	0.52	78.57

Family (continued)

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
Chrysobalanaceae	82	0.50	21	0.50	25.61
Lecythidaceae	29	0.18	21	0.50	72.41
Violaceae	41	0.25	21	0.50	51.22
Iridaceae	80	0.49	19	0.45	23.75
Verbenaceae	164	1.00	18	0.42	10.98
Aristolochiaceae	48	0.29	17	0.40	35.42
Polygonaceae	51	0.31	17	0.40	33.33
Salicaceae	51	0.31	17	0.40	33.33
Marantaceae	100	0.61	16	0.38	16.00
Alismataceae	25	0.15	15	0.35	60.00
Cucurbitaceae	82	0.50	15	0.35	18.29
Eriocaulaceae	507	3.10	15	0.35	2.96
Urticaceae	38	0.23	15	0.35	39.47
Amaranthaceae	100	0.61	14	0.33	14.00
Amaryllidaceae	47	0.29	14	0.33	29.79
Boraginaceae	91	0.56	14	0.33	15.38
Pontederiaceae	18	0.11	14	0.33	77.78
Symplocaceae	25	0.15	14	0.33	56.00
Meliaceae	35	0.21	13	0.31	37.14
Podostemaceae	22	0.13	13	0.31	59.09
Anacardiaceae	29	0.18	12	0.28	41.38
Commelinaceae	65	0.40	12	0.28	18.46
Plantaginaceae	74	0.45	12	0.28	16.22
Campanulaceae	31	0.19	10	0.24	32.26
Menispermaceae	31	0.19	10	0.24	32.26
Ochnaceae	83	0.51	10	0.24	12.05
Onagraceae	33	0.20	10	0.24	30.30
Simaroubaceae	18	0.11	10	0.24	55.56
Acanthaceae	188	1.15	9	0.21	4.79
Aquifoliaceae	36	0.22	9	0.21	25.00
Burseraceae	27	0.17	9	0.21	33.33
Combretaceae	31	0.19	9	0.21	29.03
Droseraceae	23	0.14	9	0.21	39.13
Portulacaceae	13	0.08	9	0.21	69.23
Araliaceae	34	0.21	8	0.19	23.53
Cleomaceae	21	0.13	8	0.19	38.10
Hydrocharitaceae	12	0.07	8	0.19	66.67
Nyctaginaceae	33	0.20	8	0.19	24.24
Phytolaccaceae	19	0.12	8	0.19	42.11
Clusiaceae	33	0.20	7	0.17	21.21
Loasaceae	8	0.05	7	0.17	87.50
Lythraceae	120	0.73	7	0.17	5.83
Primulaceae	67	0.41	7	0.17	10.45
Quiinaceae	8	0.05	7	0.17	87.50
Rutaceae	102	0.62	7	0.17	6.86
Velloziaceae	193	1.18	7	0.17	3.63
Vochysiaceae	49	0.30	7	0.17	14.29
Caricaceae	6	0.04	6	0.14	100.00
Ericaceae	56	0.34	6	0.14	10.71
Loganiaceae	71	0.43	6	0.14	8.45

Family (continued)

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
Olacaceae	19	0.12	6	0.14	31.58
Apiaceae	31	0.19	5	0.12	16.13
Calophyllaceae	32	0.20	5	0.12	15.62
Ebenaceae	17	0.10	5	0.12	29.41
Rhamnaceae	27	0.17	5	0.12	18.52
Begoniaceae	61	0.37	4	0.09	6.56
Capparaceae	14	0.09	4	0.09	28.57
Hypericaceae	16	0.10	4	0.09	25.00
Linderniaceae	6	0.04	4	0.09	66.67
Myristicaceae	12	0.07	4	0.09	33.33
Potamogetonaceae	9	0.06	4	0.09	44.44
Turneraceae	98	0.60	4	0.09	4.08
Burmanniaceae	11	0.07	3	0.07	27.27
Cabombaceae	4	0.02	3	0.07	75.00
Cannaceae	3	0.02	3	0.07	100.00
Elaeocarpaceae	12	0.07	3	0.07	25.00
Erythroxylaceae	88	0.54	3	0.07	3.41
Escalloniaceae	6	0.04	3	0.07	50.00
Hernandiaceae	6	0.04	3	0.07	50.00
Icacinaceae	9	0.06	3	0.07	33.33
Lacistemataceae	6	0.04	3	0.07	50.00
Martyniaceae	3	0.02	3	0.07	100.00
Molluginaceae	4	0.02	3	0.07	75.00
Monimiaceae	18	0.11	3	0.07	16.67
Peraceae	9	0.06	3	0.07	33.33
Proteaceae	18	0.11	3	0.07	16.67
Rhizophoraceae	4	0.02	3	0.07	75.00
Rosaceae	10	0.06	3	0.07	30.00
Siparunaceae	7	0.04	3	0.07	42.86
Typhaceae	3	0.02	3	0.07	100.00
Xyridaceae	117	0.72	3	0.07	2.56
Asparagaceae	8	0.05	2	0.05	25.00
Bixaceae	5	0.03	2	0.05	40.00
Brassicaceae	2	0.01	2	0.05	100.00
Cannabaceae	5	0.03	2	0.05	40.00
Cardiopteridaceae	4	0.02	2	0.05	50.00
Cunoniaceae	7	0.04	2	0.05	28.57
Haemodoraceae	2	0.01	2	0.05	100.00
Juncaceae	4	0.02	2	0.05	50.00
Loranthaceae	62	0.38	2	0.05	3.23
Marcgraviaceae	11	0.07	2	0.05	18.18
Menyanthaceae	2	0.01	2	0.05	100.00
Nymphaeaceae	9	0.06	2	0.05	22.22
Oleaceae	9	0.06	2	0.05	22.22
Orobanchaceae	29	0.18	2	0.05	6.90
Oxalidaceae	51	0.31	2	0.05	3.92
Ranunculaceae	9	0.06	2	0.05	22.22
Santalaceae	42	0.26	2	0.05	4.76
Trigoniaceae	12	0.07	2	0.05	16.67
Vitaceae	29	0.18	2	0.05	6.90

Family (continued)

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
Achariaceae	8	0.05	1	0.02	12.50
Adoxaceae	1	0.01	1	0.02	100.00
Aizoaceae	1	0.01	1	0.02	100.00
Basellaceae	2	0.01	1	0.02	50.00
Bataceae	1	0.01	1	0.02	100.00
Bonnetiaceae	1	0.01	1	0.02	100.00
Caryocaraceae	6	0.04	1	0.02	16.67
Ceratophyllaceae	1	0.01	1	0.02	100.00
Chloranthaceae	1	0.01	1	0.02	100.00
Clethraceae	1	0.01	1	0.02	100.00
Connaraceae	31	0.19	1	0.02	3.23
Costaceae	6	0.04	1	0.02	16.67
Cyclanthaceae	7	0.04	1	0.02	14.29
Cymodoceaceae	2	0.01	1	0.02	50.00
Dichapetalaceae	6	0.04	1	0.02	16.67
Goodeniaceae	1	0.01	1	0.02	100.00
Goupiaceae	1	0.01	1	0.02	100.00
Haloragaceae	2	0.01	1	0.02	50.00
Heliconiaceae	12	0.07	1	0.02	8.33
Hypoxidaceae	3	0.02	1	0.02	33.33
Linaceae	4	0.02	1	0.02	25.00
Magnoliaceae	1	0.01	1	0.02	100.00
Mayacaceae	4	0.02	1	0.02	25.00
Myoporaceae	1	0.01	1	0.02	100.00
Pentaphylacaceae	5	0.03	1	0.02	20.00
Picramniaceae	14	0.09	1	0.02	7.14
Plumbaginaceae	1	0.01	1	0.02	100.00
Rapateaceae	4	0.02	1	0.02	25.00
Ruppiaceae	1	0.01	1	0.02	100.00
Schlegeliaceae	1	0.01	1	0.02	100.00
Stemonuraceae	1	0.01	1	0.02	100.00
Styracaceae	15	0.09	1	0.02	6.67
Surianaceae	1	0.01	1	0.02	100.00
Theaceae	1	0.01	1	0.02	100.00
Thymelaeaceae	8	0.05	1	0.02	12.50
Ulmaceae	3	0.02	1	0.02	33.33
Zingiberaceae	5	0.03	1	0.02	20.00
Zygophyllaceae	2	0.01	1	0.02	50.00
Apodanthaceae	2	0.01	0	0.00	0.00
Balanophoraceae	8	0.05	0	0.00	0.00
Berberidaceae	3	0.02	0	0.00	0.00
Calyceraceae	1	0.01	0	0.00	0.00
Canellaceae	1	0.01	0	0.00	0.00
Caprifoliaceae	6	0.04	0	0.00	0.00
Caryophyllaceae	3	0.02	0	0.00	0.00
Elatinaceae	1	0.01	0	0.00	0.00
Gentianaceae	57	0.35	0	0.00	0.00
Griseliniaceae	1	0.01	0	0.00	0.00
Humiriaceae	4	0.02	0	0.00	0.00
Hydnoraceae	1	0.01	0	0.00	0.00

Family (continued)

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
Hydroleaceae	1	0.01	0	0.00	0.00
Krameriaceae	4	0.02	0	0.00	0.00
Laxmanniaceae	1	0.01	0	0.00	0.00
Opiliaceae	1	0.01	0	0.00	0.00
Picrodendraceae	2	0.01	0	0.00	0.00
Putranjivaceae	2	0.01	0	0.00	0.00
Rhabdodendraceae	1	0.01	0	0.00	0.00
Sabiaceae	3	0.02	0	0.00	0.00
Schoepfiaceae	2	0.01	0	0.00	0.00
Scrophulariaceae	8	0.05	0	0.00	0.00
Smilacaceae	27	0.17	0	0.00	0.00
Taccaceae	1	0.01	0	0.00	0.00
Thismiaceae	1	0.01	0	0.00	0.00
Triuridaceae	4	0.02	0	0.00	0.00
Tropaeolaceae	1	0.01	0	0.00	0.00
Vivianiaceae	1	0.01	0	0.00	0.00
Winteraceae	1	0.01	0	0.00	0.00

Lifeform

```
tbl_func(table(data$Formas_de_Vida), table(stats$Formas_de_Vida))  %>%
  arrange(desc(tree_Percentage)) %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = c('repeat_header', 'striped', 'hold_position'),
                repeat_header_text = 'Lifeform (continued)',
                repeat_header_method = 'replace',
                font_size = 9) %>%
  column_spec(1, width = '10em')
```

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
Erva	5682	34.88	1567	37.19	27.58
Arvore	2155	13.23	768	18.22	35.64
Liana/voluvel/trepadeira	1644	10.09	505	11.98	30.72
Arbusto; Arvore	1174	7.21	375	8.90	31.94
Arbusto	2096	12.87	317	7.52	15.12
Subarbusto	1443	8.86	214	5.08	14.83
Arbusto; Subarbusto	823	5.05	107	2.54	13.00
Erva; Subarbusto	535	3.28	105	2.49	19.63
Arbusto; Liana/voluvel/trepadeira	171	1.05	74	1.76	43.27
Arbusto; Erva; Subarbusto	110	0.68	35	0.83	31.82
Arbusto; Arvore; Subarbusto	42	0.26	16	0.38	38.10
Erva; Liana/voluvel/trepadeira	54	0.33	16	0.38	29.63
Liana/voluvel/trepadeira Subarbusto	56	0.34	16	0.38	28.57
Arbusto; Arvore; Liana/voluvel/trepadeira	40	0.25	14	0.33	35.00
Erva; Suculenta	18	0.11	14	0.33	77.78

Lifeform (continued)

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
Erva; Subarbusto; Suculenta	12	0.07	10	0.24	83.33
Arbusto; Erva	30	0.18	9	0.21	30.00
Arbusto; Suculenta	24	0.15	7	0.17	29.17
Arvore; Liana/voluvel/trepadeira	16	0.10	6	0.14	37.50
Subarbusto; Suculenta	14	0.09	6	0.14	42.86
Arbusto; Subarbusto; Suculenta	13	0.08	5	0.12	38.46
Bambu	26	0.16	4	0.09	15.38
Arbusto; Arvore; Suculenta	6	0.04	3	0.07	50.00
Arvore; Subarbusto	10	0.06	3	0.07	30.00
Erva; Liana/voluvel/trepadeira Subarbusto	4	0.02	3	0.07	75.00
Arbusto; Erva; Liana/voluvel/trepadeira; Subarbusto	4	0.02	2	0.05	50.00
Arbusto; Liana/voluvel/trepadeira Subarbusto	14	0.09	2	0.05	14.29
Arbusto; Liana/voluvel/trepadeira; Suculenta	6	0.04	2	0.05	33.33
Desconhecida	28	0.17	2	0.05	7.14
Palmeira	31	0.19	2	0.05	6.45
Arbusto; Arvore; Liana/voluvel/trepadeira Subarbusto	1	0.01	1	0.02	100.00
Arbusto; Desconhecida	1	0.01	1	0.02	100.00
Bambu; Liana/voluvel/trepadeira	2	0.01	1	0.02	50.00
Dracenuide; Erva	1	0.01	1	0.02	100.00
Liana/voluvel/trepadeira Suculenta	1	0.01	1	0.02	100.00
Arvore; Suculenta	3	0.02	0	0.00	0.00

Phytogeographic Domain

```
# included but not evalutated due to a bug with knitr/kable
tbl_func(table(data$Dom_Fitogeografico), table(stats$Dom_Fitogeografico))  %>%
  arrange(desc(tree_Percentage)) %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = c('repeat_header', 'striped', 'hold_position'),
                repeat_header_text = 'Domain (continued)',
                repeat_header_method = 'replace',
                font_size = 9) %>%
  column_spec(1, width = '10em')
```

States

The State columns were arranged differently from the other variables of interest in the data set and so `tbl_func` acts slightly different for the states. Rather than using `table(x)` and `table(y)`, data frames containing column sums were created and used as the data sources for the function. Each column represented a single state and the elements inside were either a 1 (for the presence of the taxa in that state), a 0 (for the absence) or an NA (for no data). When summed together, this provided a count of how many taxa were found in each state. As taxa could be found in multiple states, the proportion of a particular state's contribution to the whole was determined by dividing the sum of that state by the total number of taxa (16,349 for the overall data and 4,241 for the phylogeny) in each data set.

For example, the Brazilian state of Minas Gerais (MG) contains 11,136 of the species in our data, accounting for 68.11% of overall species. In the phylogeny, there are 2,933 species that can be found in Minas Gerais, accounting for 69.16% of the overall tree. Of the species which can be found in Minas Gerais, only 26.34% were used in the creation of the phylogeny. Based on these numbers alone, Minas Gerais would be an excellent state to focus collections on because of the large number and high proportion of species which have not been collected and sequenced.

```
tbl_func(data.frame(colSums(data[10:36])),
         data.frame(colSums(stats[10:36], na.rm = TRUE)),
         States = TRUE)  %>%
  arrange(desc(tree_Percentage)) %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = c('repeat_header', 'striped', 'hold_position'),
                repeat_header_text = 'State (continued)',
                repeat_header_method = 'replace',
                font_size = 9)
```

id	data_Count	data_Percentage	tree_Count	tree_Percentage	total_Percentage
MG	11136	68.11	2933	69.16	26.34
BA	8887	54.36	2780	65.55	31.28
SP	5741	35.12	2123	50.06	36.98
RJ	5211	31.87	1926	45.41	36.96
ES	4255	26.03	1669	39.35	39.22
PR	4010	24.53	1626	38.34	40.55
GO	4299	26.30	1558	36.74	36.24
MT	3646	22.30	1557	36.71	42.70
PA	3084	18.86	1473	34.73	47.76
PE	2982	18.24	1355	31.95	45.44
AM	2535	15.51	1314	30.98	51.83
MA	2805	17.16	1282	30.23	45.70
SC	2932	17.93	1265	29.83	43.14
MS	2716	16.61	1235	29.12	45.47
DF	2795	17.10	1152	27.16	41.22
CE	2360	14.44	1074	25.32	45.51
RS	2227	13.62	999	23.56	44.86
AC	1521	9.30	900	21.22	59.17
AL	1808	11.06	892	21.03	49.34
RO	1639	10.03	892	21.03	54.42
PB	1826	11.17	889	20.96	48.69
TO	1987	12.15	856	20.18	43.08
RR	1541	9.43	842	19.85	54.64
PI	1967	12.03	839	19.78	42.65
AP	1413	8.64	785	18.51	55.56
SE	1610	9.85	782	18.44	48.57
RN	1237	7.57	630	14.85	50.93

Part II – Missing Data and Node Support Values

Packages

```
library(dplyr)
library(phylotools)
```

Data

Determining the amount of missing data in the alignment and the average node support values required the final, combined alignment and the phylogenetic tree data, respectively.

```
# alignment data

alignment <- read.phylip(file.path('..', 'results', 'alignments',
                                   'combined_alignment', 'final_combined_alignment'))
alignment$family <- factor(str_replace_all(alignment$seq.name, '_.*', ''))

# phylogenetic tree data

tree <- read.tree(file.path
        ('..', 'results', 'trees', 'combined_tree', 'combined_tree.tree'))
tree$node.label <- as.numeric(tree$node.label)
```

Missing Data

For our purposes, missing data was considered to be anything other than an A, T, C or G. This includes indels (both interior and exterior), abbreviations for uncertain nucleotides (N, R, Y, W, S, M, K, etc.) and questions marks (indicating a complete lack of a gene region sequence). The amount of missing data in the alignment was overall rather high with a mean of 90.16% of the alignment containing missing data. Missing data could potentially be reduced by manually pruning large gaps in the alignment caused by small numbers of taxa. While the average family contained a mean of 89.26% missing data, the alignments of some families proved to be more complete than others with a range of 31.86% separating the minimum and maximum missing data percentages.

```
# create a column displaying the number of missing values per species
alignment$missing <- map_dbl(alignment$seq.text, function(x) {
  length(unlist(str_extract_all(x, ('[^ATCG]'))))
}) # 20,376 total loci per taxa

summary(alignment$missing) / 20376 * 100 # summary of missing percentages
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  63.54   88.09   91.43   90.16   93.72   97.55
```

```
ali_mis <- alignment %>%
    group_by(family) %>%
    summarise(missing.perc = ((sum(missing) / length(family)) / 20376) * 100)
summary(ali_mis$missing.perc) # summary of missing percentages by family
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  65.43   85.86   89.26   88.19   92.42   97.29
```

Node Support Values

Node support values had a median of 70.0 and a mean of 63.8. A fairly large number of unsupported or minimally supported nodes, mostly near the root of the tree, are a possible cause for the skewed mean value. A total of 439 nodes out of 4488 (9.78%) have a bootstrap value of 10 or less, with 134 nodes having 0 support.

```
tree <- read.tree(file.path
                  ('..', 'results', 'trees', 'combined_tree', 'combined_tree.tree'))
tree$node.label <- as.numeric(tree$node.label)

# summary of node support values
summary(tree$node.label)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0    38.0    70.0    63.8    96.0   100.0       1
```

```
# a count of the 11 lowest bootstrap values
data.frame(table(tree$node.label), row.names = NULL) %>%
  rename(Node_Support = Var1, n = Freq) %>%
  slice(1:11) %>%
  kable(booktabs = TRUE, longtable = TRUE) %>%
  kable_styling(latex_options = c('repeat_header', 'striped', 'hold_position'),
                repeat_header_text = 'State (continued)',
                repeat_header_method = 'replace',
                font_size = 9)
```

Node_Support	n
0	134
1	54
2	40
3	21
4	23
5	33
6	43
7	23
8	20
9	22
10	26