Results of PangoVis

Devan Becker

2021-08-25

Load Packages and Data

```
# Packages that Art hates
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library(stringr)
library(here)
```

```
## here() starts at /mnt/BCC20BCCC20B8A3A/sup
```

```
dirich <- params$dirich

# Read in CSV files
csvs <- list.files(here("data/", "pangolineages"),
    pattern = ifelse(dirich, "*_d.csv", "*.csv"),
    full.names = TRUE)

# Remove any copies
csvs <- csvs[!grepl("-1", csvs)]

# Bring them into one data frame
lins <- bind_rows(lapply(csvs, read.csv))

# Taxon is encoded as _ACCSESSIONNUMBER.ID, split into ACCESSIONNUMBER and ID
lins <- lins %>%
    separate(col = "taxon", sep = "\\.",
        into = c("taxon", "sample")) %>%
    mutate(taxon = str_replace(taxon, "\\_", ""))

badlins <- table(lins$taxon)
badlins <- names(badlins[which(badlins < 4500)])
```

```r
cat(length(badlins), " runs were removed for having too few samples.")
```

```
## 0  runs were removed for having too few samples.
```

```r
lins <- filter(lins, !taxon %in% badlins)

write.csv(lins, here("data", "output", "lins.csv"),
    row.names = FALSE)

#### Visualize the uncertainty in the base calls ----
taxons <- sort(unique(lins$taxon))
length(taxons)
```

```
## [1] 118
```

Abstract Info

```r
summs <- lins %>%
    group_by(taxon) %>%
    summarise(
        maxperc = mean(lineage == names(sort(table(lineage),
            decreasing = TRUE)[1])),
        uniques = length(unique(lineage)),
        minpango = min(probability),
        maxpango = max(probability),
        menpango = mean(probability),
        max = names(sort(table(lineage), decreasing = TRUE))[1])

print("summary info")
print(summs)
1 - mean(summs$maxperc); 1 - mean(summs$menpango)
```

Stacked Bar Plots

```r
max_label <- 250
other_label <- 100

par(mfrow = c(17, 1), mar = c(0.05, 7.75, 0.05, 0.05))
if (exists("seq_info")) rm(seq_info)
for (i in seq_along(taxons)) {
    pang <- lins[lins$taxon == taxons[i], ]
    called <- pang$lineage[pang$sample == 0][1]
    pangtab <- sort(table(pang$lineage), decreasing = TRUE)

    # Prep the data for a nicely formatted table
    # Subtract one because of the conseq.
    seq_info_i <- data.frame(
        called = called,
        mode = names(pangtab)[1],
            mode_n = pangtab[1] - 1,
            perc = round(100 * (pangtab[1] - 1) / (sum(pangtab) - 1), 2),
        runner_up = names(pangtab)[2],
            ru_n = pangtab[2],
        unique = length(pangtab), atoms = sum(pangtab == 1))
    seq_info_i$taxon <- taxons[i]

    if (!exists("seq_info")) {
        seq_info <- seq_info_i
    } else {
        seq_info <- bind_rows(seq_info, seq_info_i)
    }
```

```
    colvec <- rep("grey", length(pangtab))
    colvec[which(names(pangtab) == called)] <- "red"

    n <- sum(pangtab > max_label)
    if (n > 1) {
        add_other <- FALSE
        if (sum(pangtab < other_label) > 10) {
            add_other <- TRUE
            other_count <- sum(pangtab <= other_label)
            pangtab <- c(pangtab[pangtab > other_label],
                c("other" = sum(pangtab[pangtab <= other_label])))
            colvec[which(names(pangtab) == "other")] <- "black"
        }
        barlabx <- c(0, cumsum(pangtab[1:(n - 1)])) +
            pangtab[1:n] / 2
        barlabels <- names(pangtab)[1:n]
        barlens <- sapply(gregexpr("\\.", barlabels), length)
        for (j in seq_along(barlabels)) {
            if (pangtab[j] < 400 & barlens[j] >= 2) {
                barsplit <- strsplit(barlabels[j], split = "\\.")[[1]]
                barn <- length(barsplit)
                half <- floor(barn / 2)
                barlabels[j] <- paste0(
                    paste(barsplit[1:half], collapse = "."),
                    ".\n",
                    paste(barsplit[(half + 1):barn], collapse = ".")
                )
            }
        }

        barplot(as.matrix(pangtab),
            col = colvec, hori = TRUE, axes = FALSE)
        text(barlabx, 0.7, barlabels, cex = 1.5)
        if (add_other) {
            text(x = sum(pangtab) - pangtab["other"] / 2,
            y = 0.7, col = "white", cex = 1.5,
            label = paste0("Others:\n", other_count))
        }
        mtext(side = 2, cex = 1, las = 1,
            text = paste(substr(taxons[i], 1, 3),
                substr(taxons[i], 4, 20), sep = "\n"))
        abline(v = seq(0, 10000, 1000), lty = 2)
        "pretty_labels <- seq(0, sum(pangtab),
                by = ifelse(sum(pangtab) < 2000, 100, 1000))
        mtext(side = 1,
            at = pretty_labels,
            text = pretty_labels,
            line = 0,
            cex = 0.75
        )"
    }
}

seq_info$taxon <- taxons
seq_info <- arrange(seq_info, mode, mode_n) %>%
    select(taxon, everything())
knitr::kable(seq_info, row.names = FALSE)
```

taxon	called	mode	mode_n	perc	runner_up	ru_n	unique	atoms
SRR12749715	A	A	4656	93.19	B.1	61	40	11
SRR12749716	A	A	4692	93.92	B.1	48	34	11
ERR5082598	AA.3	AA.3	4611	92.29	B.1.177.15	195	23	5
ERR5077713	AD.2	AD.2	4921	98.50	B.1	23	12	4
ERR4890693	AM.3	AM.3	4594	91.95	B.1.1	249	46	18
ERR4890771	AM.3	AM.3	4746	95.00	B.1.1	122	41	16
ERR5079699	AM.3	AM.3	4843	96.94	B.1.1	62	27	11
ERR4891444	B.1.1	B.1	1680	33.63	B.1.1	764	270	63
ERR4693865	B.1	B.1	2418	48.40	B.1.2	115	262	58
ERR4890531	B.1	B.1	2752	55.08	B.1.1	89	255	62
ERR4891415	B.1	B.1	2879	57.63	B.1.595	111	259	65
ERR5082578	B.1	B.1	3705	74.16	B.1.280	154	153	41
ERR4693801	B.1.1	B.1.1	2578	51.60	B.1	173	247	64
ERR4891178	B.1.1	B.1.1	2790	55.84	B.1.1.307	184	251	50
ERR4891497	B.1.1	B.1.1	2801	56.06	B.1.1.374	102	241	54
ERR4890881	B.1.1	B.1.1	3276	65.57	B.1.1.217	151	228	58
ERR4890572	B.1.1	B.1.1	3577	71.60	B.1.1.121	171	168	46
ERR4890926	B.1.1	B.1.1	3994	79.94	B.1.1.217	50	189	55
ERR4891572	B.1.1.10	B.1.1.10	4759	95.26	B.1.1	85	34	15

taxon	called	mode	mode_n	perc	runner_up	ru_n	unique	atoms
ERR5078897	B.1.1.240	B.1.1.240	4011	80.28	B.1.1	443	123	53
ERR4694498	B.1.1.310	B.1.1.310	4027	80.60	B.1.1	218	133	40
ERR4694380	B.1.1.37	B.1.1.37	4924	98.56	B.1.1.294	32	12	6
ERR5081836	B.1.1.434	B.1.1.434	4785	95.78	B.1.1	68	32	21
ERR4891493	B.1.1.51	B.1.1.51	4725	94.58	B.1.1	190	52	35
ERR4694330	B.1.1.58	B.1.1.58	3838	76.82	B.1.1.217	367	59	20
ERR5077924	B.1.1.7	B.1.1.7	4996	100.00	NA	NA	1	0
ERR5078863	B.1.1.7	B.1.1.7	4996	100.00	NA	NA	1	0
ERR5079000	B.1.1.7	B.1.1.7	4996	100.00	NA	NA	1	0
ERR5080131	B.1.1.7	B.1.1.7	4996	100.00	NA	NA	1	0
ERR5080504	B.1.1.7	B.1.1.7	4996	100.00	NA	NA	1	0
ERR5082214	B.1.1.7	B.1.1.7	4996	100.00	NA	NA	1	0
ERR5082673	B.1.1.7	B.1.1.7	4996	100.00	NA	NA	1	0
ERR4694010	B.1.13	B.1.13	4043	80.92	B.1	520	54	25
ERR5082710	B.1.160	B.1.160	3871	77.48	B.1.160.15	191	49	12
ERR5082346	B.1.160	B.1.160	4431	88.69	B.1.160.11	118	36	0
ERR4890974	B.1.160	B.1.160	4787	95.82	B.1	26	32	7
ERR5081077	B.1.177	B.1.177	3267	65.39	B.1.177.73	1104	70	11
ERR5082706	B.1.177	B.1.177	4046	80.98	B.1.177.44	75	66	9
ERR5078210	B.1.177	B.1.177	4391	87.89	B.1.177.21	37	65	7
ERR4891001	B.1.177	B.1.177	4429	88.65	B.1.177.25	34	59	8
ERR5082656	B.1.177	B.1.177	4849	97.06	B.1.177.68	58	14	1
ERR5082694	B.1.177	B.1.177	4876	97.60	B.1.177.21	35	17	7
ERR4891304	B.1.177	B.1.177	4893	97.94	B.1.177.73	35	30	17
ERR5082606	B.1.177	B.1.177	4895	97.98	B.1.177.68	31	13	5
ERR5081293	B.1.177	B.1.177	4897	98.02	B.1.258	27	15	7
ERR4891433	B.1.177	B.1.177	4919	98.46	B.1.177.68	28	11	5
ERR4891532	B.1.177	B.1.177	4956	99.20	B.1.177.68	18	9	1
ERR4891011	B.1.177.16	B.1.177.16	4897	98.02	B.1.177	84	6	2
ERR5080327	B.1.177.18	B.1.177.18	4480	89.67	B.1	276	37	22
ERR4891061	B.1.177.19	B.1.177.19	4821	96.50	B.1.2	50	14	5
ERR4890746	B.1.177.4	B.1.177.4	4949	99.06	B.1.177	28	13	8
ERR5079423	B.1.177.57	B.1.177.57	2486	49.76	B.1.177.56	2082	34	13
ERR5082708	B.1.177.57	B.1.177.57	4922	98.52	B.1.177.73	21	13	6
ERR5082622	B.1.177.58	B.1.177.58	4898	98.04	B.1.177	27	17	3
ERR5082654	B.1.177.65	B.1.177.65	4334	86.75	B.1.177	308	34	15
ERR5082702	B.1.177.65	B.1.177.65	4802	96.12	B.1.177	120	18	5
ERR5080159	B.1.177.7	B.1.177.7	4339	86.85	B.1.177.16	612	7	1
ERR4891103	B.1.177.9	B.1.177.9	4896	98.00	B.1.177	49	18	5
ERR4891037	B.1.258	B.1.258	4729	94.66	B.1.258.14	51	31	9
ERR4891235	B.1.258	B.1.258	4827	96.62	A	26	26	4
ERR4890609	B.1.258	B.1.258	4856	97.20	B.1.258.14	40	23	7
ERR4891261	B.1.258.12	B.1.258.12	4854	97.16	B.1.258.14	27	17	3
ERR5082600	B.1.258.5	B.1.258.5	4650	93.07	B.1.258	131	29	17
ERR5082700	B.1.36.39	B.1.36.39	4696	94.00	B.1	160	18	6
ERR4890820	B.1.391	B.1.391	4282	85.71	B.1	441	48	17
ERR4891675	B.1.523	B.1.523	4691	93.90	B.1.400	135	27	12
ERR4891238	B.4.8	B.4.8	4816	96.40	B.1	56	25	10
ERR4693495	B.40	B.40	4937	98.82	B.1	23	11	3
SRR12639961	B.41	B.41	3956	79.18	B.1.1	235	39	16
SRR13021017	B.4.6	None	4885	99.53	A	18	4	1
SRR13021008	A.2.2	None	4910	99.63	A	13	6	3
SRR13021020	A	None	4915	99.66	A	16	3	1
SRR13020991	A.1	None	4930	99.72	A	12	4	2
SRR13021124	B.4.6	None	4930	99.72	A	10	5	2
SRR13021131	B.1	None	4935	99.74	A	10	4	1
SRR13021013	B.1	None	4940	99.76	A	6	4	0
SRR11433882	B.1	None	4955	99.82	A	8	3	1
SRR13021011	A	None	4960	99.84	A	8	2	0
SRR11433888	B.1.413	None	4965	99.86	A	5	4	2
SRR11433893	B.1	None	4965	99.86	A	6	3	1
SRR13021109	A.1	None	4970	99.88	A	4	3	0
SRR13021111	B.58	None	4970	99.88	A	5	3	1
SRR13021113	B.1.1	None	4970	99.88	A	5	3	1
SRR13021099	A.1	None	4980	99.92	A	2	3	0
SRR13021104	B.1.1	None	4985	99.94	A	2	3	1
SRR13021130	A	None	4990	99.96	A	2	2	0
SRR13020998	B.1.1	None	4995	99.98	B.1.1	1	2	1
SRR13020999	B.1	None	4995	99.98	B.1	1	2	1
SRR13021003	A.1	None	4995	99.98	A.1	1	2	1
SRR13021010	A.2.2	None	4995	99.98	A.2.2	1	2	1
SRR13021022	A.1	None	4995	99.98	A.1	1	2	1
SRR13021052	B.1.1	None	4995	99.98	B.1.1	1	2	1
SRR13021053	B.1.1.71	None	4995	99.98	B.1.1.71	1	2	1
SRR13021059	B.1.13	None	4995	99.98	B.1.13	1	2	1
SRR13021061	A.1	None	4995	99.98	A.1	1	2	1
SRR13021067	A.2.2	None	4995	99.98	A.2.2	1	2	1
SRR13021072	B.1	None	4995	99.98	B.1	1	2	1
SRR13021073	A.1	None	4995	99.98	A.1	1	2	1

taxon	called	mode	mode_n	perc	runner_up	ru_n	unique	atoms
SRR13021077	A.1	None	4995	99.98	A.1	1	2	1
SRR13021084	B.1	None	4995	99.98	B.1	1	2	1
SRR13021090	B.1	None	4995	99.98	B.1	1	2	1
SRR13021093	A.2.2	None	4995	99.98	A.2.2	1	2	1
SRR13021098	A.1	None	4995	99.98	A.1	1	2	1
SRR13021107	A.2.2	None	4995	99.98	A.2.2	1	2	1
SRR13021115	B.1.1.10	None	4995	99.98	B.1.1.10	1	2	1
SRR13021133	A	None	4995	99.98	A	1	2	1
SRR13021134	B.4.6	None	4995	99.98	B.4.6	1	2	1
SRR13021135	A.2.2	None	4995	99.98	A.2.2	1	2	1
SRR13021143	A.2.2	None	4995	99.98	A.2.2	1	2	1
ERR4692420	None	None	4996	100.00	NA	NA	1	0
ERR4692568	None	None	4996	100.00	NA	NA	1	0
ERR4692877	None	None	4996	100.00	NA	NA	1	0
ERR4692945	None	None	4996	100.00	NA	NA	1	0
ERR4693014	None	None	4996	100.00	NA	NA	1	0
ERR4693019	None	None	4996	100.00	NA	NA	1	0
ERR4890819	None	None	4996	100.00	NA	NA	1	0
ERR5082599	None	None	4996	100.00	NA	NA	1	0
SRR13592146	None	None	4996	100.00	NA	NA	1	0