MERGE DESJARDINS AND ASHTON TREES

Claudia Zirión-Martínez

Setup

```
library(RRphylo)
library(manipulate)
library(ape)
library(phytools)
library(ggtree)
library(tidyverse)
library(RColorBrewer)
library(ggnewscale)
library(patchwork)
setwd("/FastData/czirion/Crypto_Diversity_Pipeline/analyses/tree_duplications/scripts")
```

Metadata

Use the metadata table that has all the samples included in the final Crypto_Desjardins_Ashton dataset and H99 (n = 1056).

```
metadata <- read.delim(
    "../../data/processed/metadata_ashton_desj_all_fungalpop_H99.csv",
    header=TRUE,
    sep=",")
summary <- metadata %>%
    group_by(dataset, lineage) %>%
    summarize(count = n())
summary
```

dataset	lineage	count
Ashton	VNI	668
Desjardins	VNBI	122
Desjardins	VNBII	64
Desjardins	VNI	185
Desjardins	VNII	16
Reference	VNI	1

Make separate dataframes for each metadata field.

```
metadata$vni_subdivision <- factor(metadata$vni_subdivision,
                           levels = c("VNIa-4", "VNIa-5", "VNIa-32",
                           "VNIa-93", "VNIa-X", "VNIa-Y", "VNIb",
                           "VNIc", "VNIa-outlier"))

sublineage <- metadata %>%
                filter(lineage == "VNI")%>%
                select(strain, vni_subdivision)%>%
```

```
                column_to_rownames("strain")%>%
                droplevels()
lineage <- metadata %>%
            select(strain, lineage)%>%
            column_to_rownames("strain")
dataset <- metadata %>%
            select(strain, dataset)%>%
            column_to_rownames("strain")
source <- metadata %>%
            select(strain, source)%>%
            column_to_rownames("strain")
```

Make color vectors for all plots

```
dataset_colors <- c(brewer.pal(9, "Set1")[c(1, 2)], "white")
names(dataset_colors) <- levels(as.factor(dataset$dataset))

lineage_colors <- brewer.pal(8, "Dark2")[c(1, 2, 3, 4)]
names(lineage_colors) <- levels(as.factor(lineage$lineage))

sublineage_colors <- c(brewer.pal(12, "Set3")[c(1:9)])
names(sublineage_colors) <- levels(sublineage$vni_subdivision)

source_colors <- brewer.pal(11, "BrBG")[c(9, 3)] # 9, 3 are the colors for the two sources
names(source_colors) <- levels(as.factor(source$source))
```

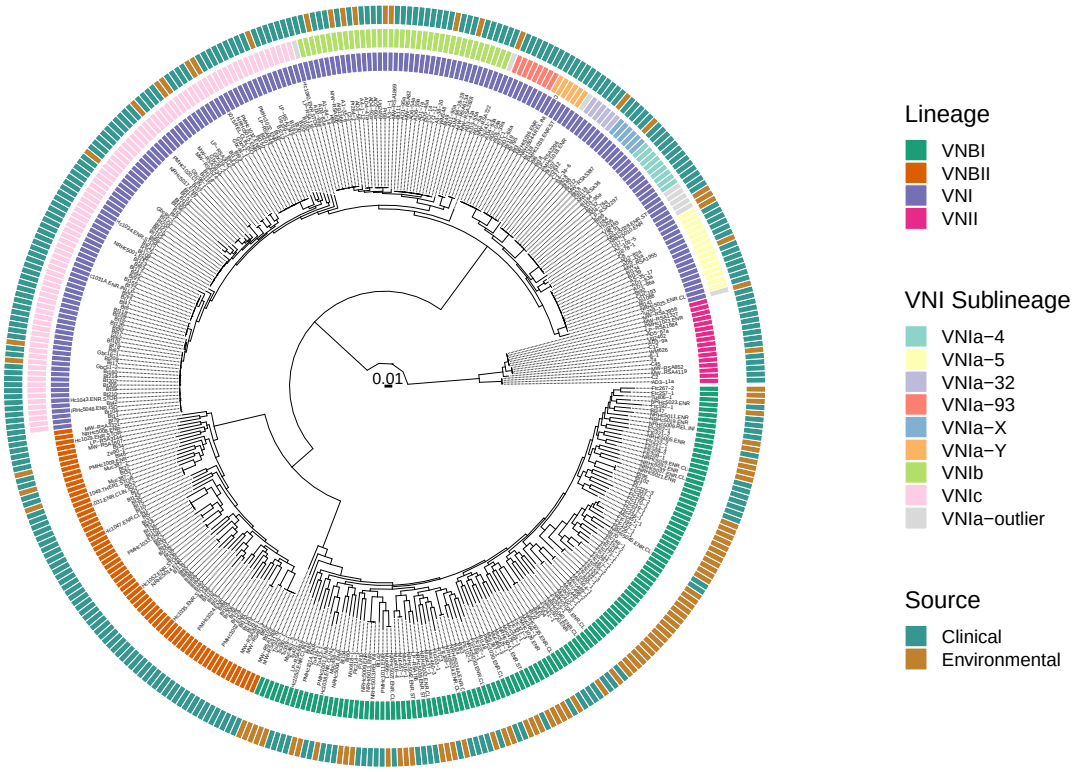
Desjardins tree

Import the raw Desjardins tree

```
desj_tree_path <- "../../data/raw/CryptoDiversity_Desjardins_Tree.tre"
desj_tree <- read.tree(desj_tree_path)
```

Reroot the tree at the middle of the branch leading to VNII

```
VNII_root <- getMRCA(desj_tree, c("C2","C12"))
edge_length <- subset(desj_tree$edge.length, desj_tree$edge[,2] == VNII_root)
desj_tree <- reroot(desj_tree, VNII_root, edge_length/2)
write.tree(desj_tree, file = "../../data/processed/desj_tree.newick")
```



Ashton tree

Import the raw Ashton tree

```
ashton_tree_path <- "../../data/raw/2017.06.09.all_ours_and_desj.snp_sites.mod.fa.cln.tree"
ashton_tree_unrooted <- read.tree(ashton_tree_path)
```

Rename tips to use strain names in the Desjardins samples (which have run accessions).

```
ashton_tree_unrooted$tip.label <- sapply(ashton_tree_unrooted$tip.label, function(x) {
    if (x %in% metadata$run) {
        metadata$strain[metadata$run == x]
    } else {
        x
    }
})
```

Get the samples that are present in the tree but absent from the metadata of the final dataset

```
tips_missing_from_final_dataset <- setdiff(ashton_tree_unrooted$tip.label, metadata$strain)
```

Compare the list of strains missing from metadata with the oringinal Ashton metadata

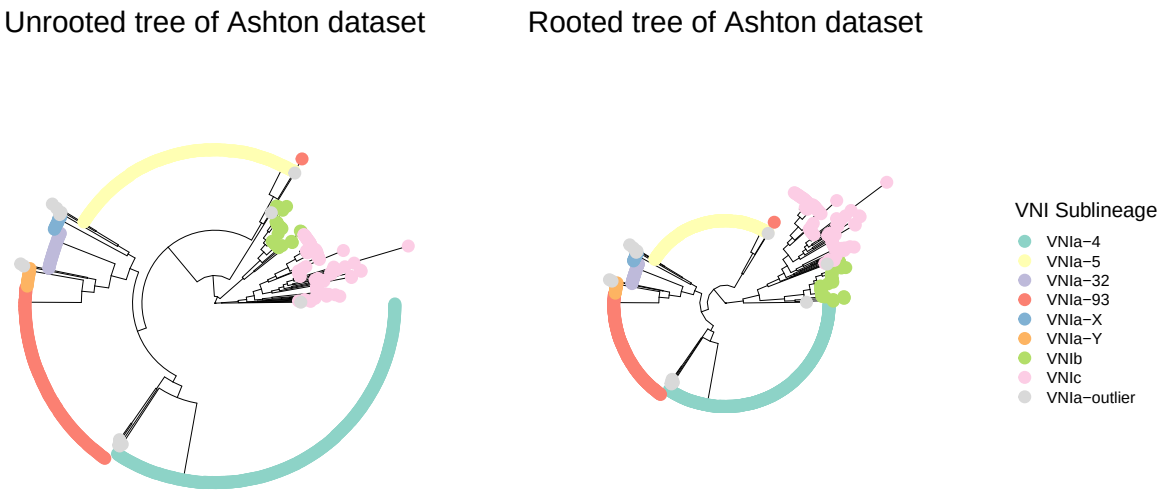
```
ashton_metadata <-read.delim(
    "../../../Crypto_Ashton/config/metadata_all_ashton_and_vni_desj.csv",
    header=TRUE, sep=",")
samples_missing_from_dataset <- ashton_metadata %>%
    filter(strain %in% tips_missing_from_final_dataset)%>%
    select(sample, strain, lineage, VNI_subdivision)
samples_missing_from_dataset
```

sample	strain	lineage	VNI_subdivision
ERS542414	15277_3#7	VNI	VNIa-4
ERS542415	15277_3#8	VNI	VNIa-4
ERS542595	15277_3#45	VNI	VNIa-4
ERS542403	15277_3#1	VNI	VNIa-4
ERS542456	15277_3#18	VNI	VNIa-4
ERS542410	15277_3#5	VNI	VNIa-5
ERS542411	15277_3#6	VNI	VNIa-5
	CNS_1465	VNI	VNIa-93
ERS542584	15277_3#42	VNI	VNIa-93
ERS542502	14893_1#16	VNI	VNIa-93

The CNS_1465 strain was not available for download and the rest had bad quality alignments.

Root Ashton tree at the middle of the branch leading to VNIa

```
VNIa_root <- getMRCA(ashton_tree_unrooted, c("AD3-95a","Tu259-1"))
edge_length <- subset(ashton_tree_unrooted$edge.length,
    ashton_tree_unrooted$edge[,2] == VNIa_root)
ashton_tree <- reroot(ashton_tree_unrooted, VNIa_root, edge_length/2)
write.tree(ashton_tree, file = "../../data/processed/ashton_tree.newick")
```



Rooted tree of Ashton dataset

Merge Desjardins and Ashton trees

Specify clades in Desjardins tree

```
VNI <- c("Bt92", "Bt79")
VNI_node <- getMRCA(desj_tree, VNI)
VNII <- c("C2","C12")
VNII_node <- getMRCA(desj_tree, VNII)
VNB <- c("Bt7", "Bt34")
VNB_node <- getMRCA(desj_tree, VNB)
```

Get the ages of the nodes from the original Desjardins tree. This is to attempt to have a calibrated tree, but the resulting branchlengths are not real.

```
edge_lengths <- node.depth.edgelength(desj_tree)
node_labels <- c(desj_tree$tip.label, desj_tree$node.label)
edge_length_mapping <- data.frame(
    node = node_labels,
    edge_length = edge_lengths,
    max_length = max(edge_lengths))
edge_length_mapping <- edge_length_mapping %>%
                        mutate(age = max_length - edge_length) %>%
                        rownames_to_column("node_id")
clade_ages <- edge_length_mapping %>%
                filter(node_id %in% c(VNI_node, VNII_node, VNB_node))
nodeages <- c("Bt92-Bt79" = clade_ages$age[clade_ages$node_id == VNI_node],
              "C2-C12" = clade_ages$age[clade_ages$node_id == VNII_node],
              "Bt7-Bt34" = clade_ages$age[clade_ages$node_id == VNB_node])
tip_ages <- edge_length_mapping %>%
                filter(node %in% metadata$strain)
tipages <- tip_ages$age
names(tipages) <- tip_ages$node
```

Remove VNI clade from Desjardins tree to use it as backtree

```
VNI_tips <- tips(desj_tree, VNI_node)
backtree <- drop.tip(desj_tree, VNI_tips)
```

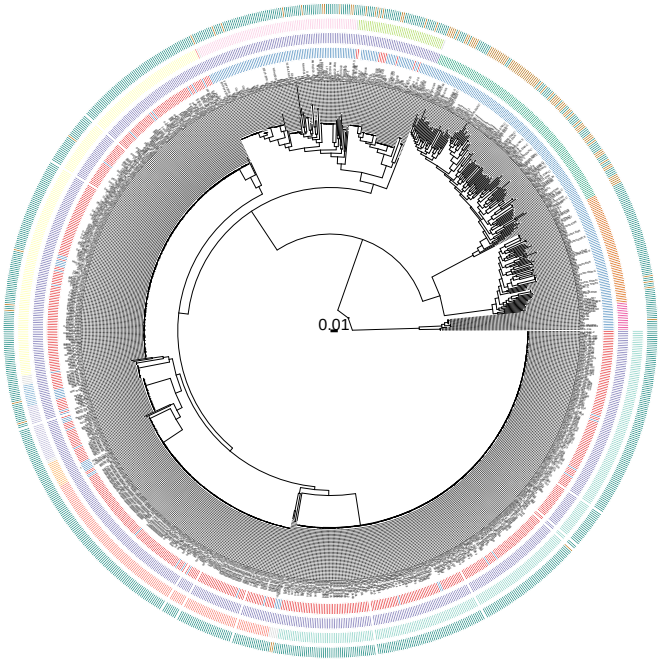
Create the reference tables

```
reference <- data.frame(bind=c("CNS_289-20427_2#4"),
                  reference=c("Bt7-Bt34"),
                  poly=c(FALSE))
```

Merge

```
merged <- tree.merger(backbone = backtree,
                        data=reference,
                        source.tree = ashton_tree,
                        plot=FALSE,
                        node.ages = nodeages,
                        tip.ages = tipages)
```

Merged tree with branchlengths (not real)



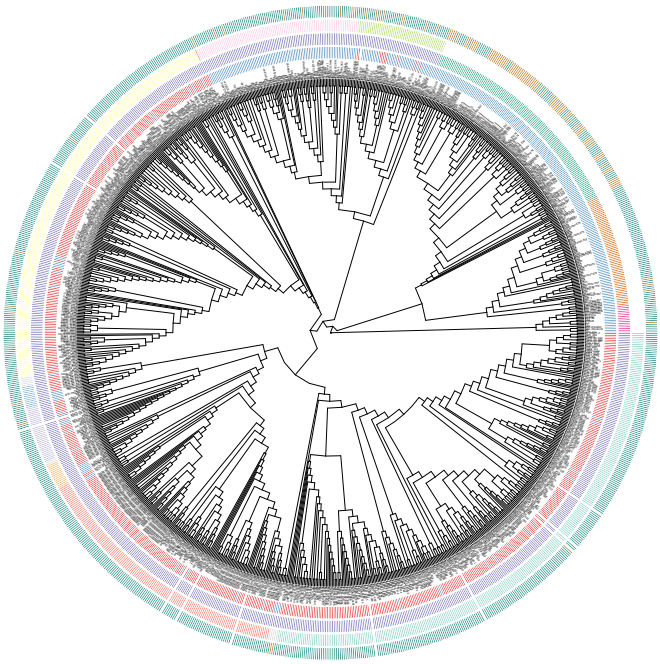
Dataset
- Ashton
- Desjardins
- Reference

Lineage
- VNBI
- VNBII
- VNI
- VNII

VNI Sublineage
- VNIa−4
- VNIa−5
- VNIa−32
- VNIa−93
- VNIa−X
- VNIa−Y
- VNIb
- VNIc
- VNIa−outlier

Source
- Clinical
- Environmental

Cladogram of merged tree



Dataset
- Ashton
- Desjardins
- Reference

Lineage
- VNBI
- VNBII
- VNI
- VNII

VNI Sublineage
- VNIa−4
- VNIa−5
- VNIa−32
- VNIa−93
- VNIa−X
- VNIa−Y
- VNIb
- VNIc
- VNIa−outlier

Source
- Clinical
- Environmental

Plot minimal version of the tree

Get one sample of each non-VNI lineage, VNI sublineage, and all VNIa-outlier

```
VNI <- metadata %>%
    filter(lineage == "VNI", vni_subdivision != "VNIa-outlier") %>%
    group_by(vni_subdivision) %>%
    slice(1) %>%
    ungroup()
VNIa_outlier <- metadata %>%
    filter(vni_subdivision == "VNIa-outlier")
VNII <- metadata %>%
    filter(lineage == "VNII") %>%
    slice(1) %>%
    ungroup()
VNBI <- metadata %>%
    filter(lineage == "VNBI") %>%
    slice(1) %>%
    ungroup()
VNBII <- metadata %>%
    filter(lineage == "VNBII") %>%
    slice(1) %>%
    ungroup()
tips <- rbind(VNI, VNIa_outlier, VNII, VNBI, VNBII)%>%
    select(strain)
```

Make a small version of the merged tree only with the tips in `tips`

```
small_tree <- drop.tip(merged, setdiff(merged$tip.label, tips$strain))
```

Minimal cladogram of merged tree with one strain per sublineage