BGC exploration

Daniel Martinez

Background

This would be the part where the **background** of the project goes!

Data exploration

Reading the data

Here we are going to read the data we have

```
library(tidyverse)
library(readr)
library(readxl)
library(cowplot)
```

```
# read the main dataset of BGCs
bgc = read_csv("../data/BGCs/antismash_summary.csv")

metadata = read_excel("../data/MAIN_metadata.xlsx",
                      sheet = "metadata")
```

Now that we have the data generated, we can take a look at it:

```
head(bgc)
```

```
# A tibble: 6 x 6
  genome cluster type        contig  start    end
  <chr>    <dbl> <chr>       <chr>   <dbl>  <dbl>
1 1.2          1 thiopeptide 3      118103 144396
```

```
2 1.2          1 NRPS         12       1323  45205
3 1.3          1 NRPS         5       16303  60185
4 1.3          1 NRPS|T1PKS   8       42969  98656
5 1.3          1 siderophore 77           0  13235
6 1            1 NRPS         18      49670  93552
```

From this table, the column type refers to...

phylogroup exploration

This is the part where we start exploring the phylogroups and its relevance for the BGCs.

```
phylo =
  metadata %>%
  filter(Origin %in% c('AUS', 'ECOREF')) %>%
  filter(Discard == 'No') %>%
  select(fasta, Broadphenotype, phylogroup) %>%
  distinct(fasta, .keep_all = TRUE) %>%
  mutate(genome = str_sub(fasta,
                          start = 1, end = -7)) %>%
  select(-fasta)


bgc_extended = bgc %>%
  left_join(phylo)
```

```
Joining, by = "genome"
```

```
bgc_extended %>%
  filter(start != 0 ) %>%
  # filter(type == 'NRPS') %>%
  mutate(gene_length = end - start) %>%
  mutate(ID = paste(genome, type, sep = '_')) %>%
  mutate(type = as.factor(type)) %>%
  ggplot(aes(y = phylogroup,
             x = gene_length,
             fill = phylogroup)) +
  geom_boxplot(show.legend = T) +
  geom_point(alpha = 0.5,
             show.legend = F) +
```

```
theme_cowplot(15)
```

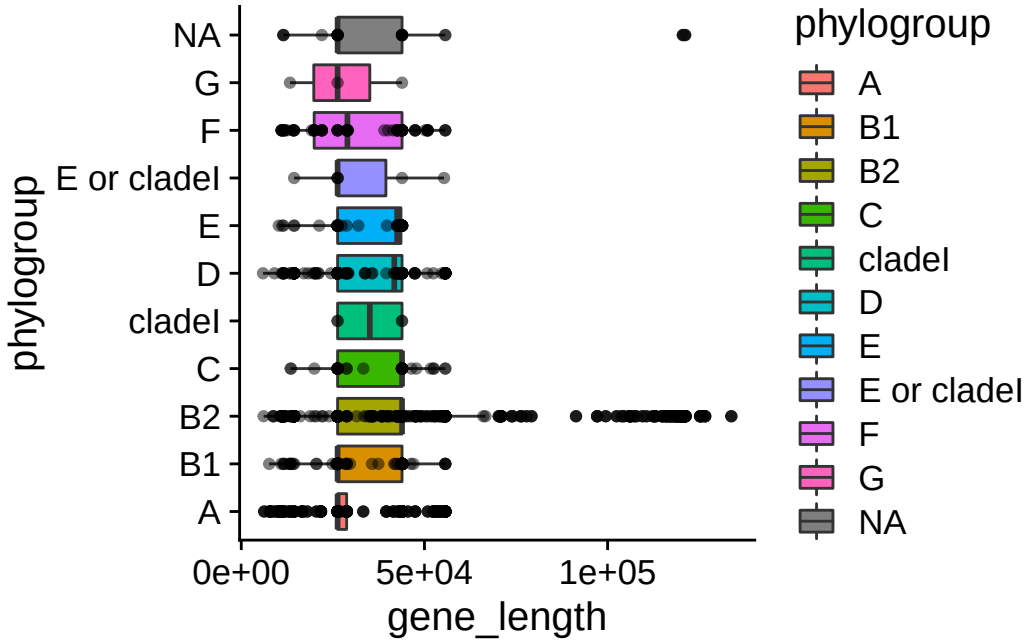

Figure 1: Gene length for phylogroup

As we see in Figure 1, we see that different phylogropus have different gene lenghts in the cohort.