Initial data survey

LJM

2019-11-19

Contents

Chapter 1

The data

```
rm(list=ls())
library(tidyverse)
```

What is in the files?

1.1 Which bins belong to which fOTUS?

```
fOTUs <- read_csv(file = "../data/fOTUs.csv", col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   cogOTU_0 = col_character(),
##   `fMAG:Loc081215-8m_megahit_metabat_bin-2464;fMAG:Loc081215-5m_megahit_metabat_bin-1172
## )
```

```
glimpse(fOTUs)
```

```
## Observations: 2,810
## Variables: 2
## $ cogOTU_0
## $ `fMAG:Loc081215-8m_megahit_metabat_bin-2464;fMAG:Loc081215-5m_megahit_metabat_bin-1172
```

```
colnames(fOTUs)
```

```
## [1] "cogOTU_0"
## [2] "fMAG:Loc081215-8m_megahit_metabat_bin-2464;fMAG:Loc081215-5m_megahit_metabat_bin-117
```

1.2 What bin does each cluster of orthologous genes (COG) belong to?

```
bin2cogs <- read_csv(file = "../data/bin2cogs.csv", col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   `Loclat_megahit_metabat_bin-01178` = col_character(),
##   `COG_187551;COG_170313;COG_741993;COG_112861;COG_88993;COG_70779;COG_170312;COG_5458;CO
## )
```

```
glimpse(bin2cogs)
```

```
## Observations: 29,614
## Variables: 2
## $ `Loclat_megahit_metabat_bin-01178`
## $ `COG_187551;COG_170313;COG_741993;COG_112861;COG_88993;COG_70779;COG_170312;COG_5458;CO
```

```
colnames(bin2cogs)
```

```
## [1] "Loclat_megahit_metabat_bin-01178"
## [2] "COG_187551;COG_170313;COG_741993;COG_112861;COG_88993;COG_70779;COG_170312;COG_5458;
```

1.3 Stats on all fOTUed bins

```
stats_on_all_fOTUed_bins <- read_csv(file = "../analyses/stats_on_all_fOTUed_bins.csv", col_
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   length = col_double(),
##   nb_contigs = col_double(),
##   nb_proteins = col_double(),
##   coding_density = col_double(),
##   GC = col_double(),
##   vir_fract = col_double(),
##   fOTU = col_character()
## )
```

```
glimpse(stats_on_all_fOTUed_bins)
```

```
## Observations: 29,615
## Variables: 8
## $ X1             <chr> "Loc081215-8m_megahit_metabat_bin-2464", "Loc08...
## $ length         <dbl> 36092, 30213, 53243, 36091, 29849, 23184, 29990...
## $ nb_contigs     <dbl> 1, 1, 5, 1, 1, 1, 2, 2, 1, 5, 3, 6, 10, 4, 6, 8...
## $ nb_proteins    <dbl> 64, 55, 80, 65, 55, 39, 55, 12, 19, 27, 21, 23,...
## $ coding_density <dbl> 0.9361077, 0.9678284, 0.9031046, 0.9190103, 0.9...
## $ GC             <dbl> 0.4203425, 0.4260087, 0.4239431, 0.4202987, 0.4...
## $ vir_fract      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ fOTU           <chr> "cogOTU_0", "cogOTU_0", "cogOTU_0", "cogOTU_0",...
```

```
colnames(stats_on_all_fOTUed_bins)
```

```
## [1] "X1"             "length"         "nb_contigs"     "nb_proteins"   
## [5] "coding_density" "GC"             "vir_fract"      "fOTU"
```

1.4 Stats on representative bins

```
stats_on_representative_bins <- read_csv(file = "../analyses/stats_on_representative_bins.cs
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   length = col_double(),
##   nb_contigs = col_double(),
##   nb_proteins = col_double(),
##   coding_density = col_double(),
##   GC = col_double(),
##   vir_fract = col_double(),
##   fOTU = col_character(),
##   other_members = col_character()
## )
```

```
glimpse(stats_on_representative_bins)
```

```
## Observations: 2,430
## Variables: 9
## $ X1             <chr> "Loc081215-5m_megahit_metabat_bin-1172", "Loc08...
## $ length         <dbl> 30213, 13170, 44529, 14195, 23811, 70833, 24419...
## $ nb_contigs     <dbl> 1, 3, 9, 1, 5, 1, 3, 1, 3, 3, 2, 8, 19, 1, 5, 9...
## $ nb_proteins    <dbl> 55, 23, 71, 24, 29, 121, 30, 32, 16, 27, 19, 45...
## $ coding_density <dbl> 0.9678284, 0.8018223, 0.8892407, 0.9256781, 0.8...
## $ GC             <dbl> 0.4260087, 0.4763857, 0.6836219, 0.6211342, 0.5...
## $ vir_fract      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 83076, 0, 0...
## $ fOTU           <chr> "cogOTU_0", "cogOTU_1", "cogOTU_2", "cogOTU_3",...
## $ other_members  <chr> "fMAG:Loc081215-8m_megahit_metabat_bin-2464;fMA...
```

```
colnames(stats_on_representative_bins)
```

```
## [1] "X1"             "length"         "nb_contigs"     "nb_proteins"
## [5] "coding_density" "GC"             "vir_fract"      "fOTU"
## [9] "other_members"
```