

# Initial data survey

*LJM*

*2019-11-19*

# Contents

<b>1</b>	<b>The data</b>	<b>2</b>
1.1	Which bins belong to which fOTUS? . . . . .	2
1.2	What bin does each cluster of orthologous genes (COG) belong to? . . . . .	3
1.3	Stats on all fOTUed bins . . . . .	3
1.4	Stats on representative bins . . . . .	4

# 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-1172"
```

## 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;COG_170311;COG_170310;COG_170309;COG_170308;COG_170307;COG_170306;COG_170305;COG_170304;COG_170303;COG_170302;COG_170301;COG_170300;COG_170299;COG_170298;COG_170297;COG_170296;COG_170295;COG_170294;COG_170293;COG_170292;COG_170291;COG_170290;COG_170289;COG_170288;COG_170287;COG_170286;COG_170285;COG_170284;COG_170283;COG_170282;COG_170281;COG_170280;COG_170279;COG_170278;COG_170277;COG_170276;COG_170275;COG_170274;COG_170273;COG_170272;COG_170271;COG_170270;COG_170269;COG_170268;COG_170267;COG_170266;COG_170265;COG_170264;COG_170263;COG_170262;COG_170261;COG_170260;COG_170259;COG_170258;COG_170257;COG_170256;COG_170255;COG_170254;COG_170253;COG_170252;COG_170251;COG_170250;COG_170249;COG_170248;COG_170247;COG_170246;COG_170245;COG_170244;COG_170243;COG_170242;COG_170241;COG_170240;COG_170239;COG_170238;COG_170237;COG_170236;COG_170235;COG_170234;COG_170233;COG_170232;COG_170231;COG_170230;COG_170229;COG_170228;COG_170227;COG_170226;COG_170225;COG_170224;COG_170223;COG_170222;COG_170221;COG_170220;COG_170219;COG_170218;COG_170217;COG_170216;COG_170215;COG_170214;COG_170213;COG_170212;COG_170211;COG_170210;COG_170209;COG_170208;COG_170207;COG_170206;COG_170205;COG_170204;COG_170203;COG_170202;COG_170201;COG_170200;COG_170199;COG_170198;COG_170197;COG_170196;COG_170195;COG_170194;COG_170193;COG_170192;COG_170191;COG_170190;COG_170189;COG_170188;COG_170187;COG_170186;COG_170185;COG_170184;COG_170183;COG_170182;COG_170181;COG_170180;COG_170179;COG_170178;COG_170177;COG_170176;COG_170175;COG_170174;COG_170173;COG_170172;COG_170171;COG_170170;COG_170169;COG_170168;COG_170167;COG_170166;COG_170165;COG_170164;COG_170163;COG_170162;COG_170161;COG_170160;COG_170159;COG_170158;COG_170157;COG_170156;COG_170155;COG_170154;COG_170153;COG_170152;COG_170151;COG_170150;COG_170149;COG_170148;COG_170147;COG_170146;COG_170145;COG_170144;COG_170143;COG_170142;COG_170141;COG_170140;COG_170139;COG_170138;COG_170137;COG_170136;COG_170135;COG_170134;COG_170133;COG_170132;COG_170131;COG_170130;COG_170129;COG_170128;COG_170127;COG_170126;COG_170125;COG_170124;COG_170123;COG_170122;COG_170121;COG_170120;COG_170119;COG_170118;COG_170117;COG_170116;COG_170115;COG_170114;COG_170113;COG_170112;COG_170111;COG_170110;COG_170109;COG_170108;COG_170107;COG_170106;COG_170105;COG_170104;COG_170103;COG_170102;COG_170101;COG_170100;COG_170099;COG_170098;COG_170097;COG_170096;COG_170095;COG_170094;COG_170093;COG_170092;COG_170091;COG_170090;COG_170089;COG_170088;COG_170087;COG_170086;COG_170085;COG_170084;COG_170083;COG_170082;COG_170081;COG_170080;COG_170079;COG_170078;COG_170077;COG_170076;COG_170075;COG_170074;COG_170073;COG_170072;COG_170071;COG_170070;COG_170069;COG_170068;COG_170067;COG_170066;COG_170065;COG_170064;COG_170063;COG_170062;COG_170061;COG_170060;COG_170059;COG_170058;COG_170057;COG_170056;COG_170055;COG_170054;COG_170053;COG_170052;COG_170051;COG_170050;COG_170049;COG_170048;COG_170047;COG_170046;COG_170045;COG_170044;COG_170043;COG_170042;COG_170041;COG_170040;COG_170039;COG_170038;COG_170037;COG_170036;COG_170035;COG_170034;COG_170033;COG_170032;COG_170031;COG_170030;COG_170029;COG_170028;COG_170027;COG_170026;COG_170025;COG_170024;COG_170023;COG_170022;COG_170021;COG_170020;COG_170019;COG_170018;COG_170017;COG_170016;COG_170015;COG_170014;COG_170013;COG_170012;COG_170011;COG_170010;COG_170009;COG_170008;COG_170007;COG_170006;COG_170005;COG_170004;COG_170003;COG_170002;COG_170001;COG_170000;COG_169999;COG_169998;COG_169997;COG_169996;COG_169995;COG_169994;COG_169993;COG_169992;COG_169991;COG_169990;COG_169989;COG_169988;COG_169987;COG_169986;COG_169985;COG_169984;COG_169983;COG_169982;COG_169981;COG_169980;COG_169979;COG_169978;COG_169977;COG_169976;COG_169975;COG_169974;COG_169973;COG_169972;COG_169971;COG_169970;COG_169969;COG_169968;COG_169967;COG_169966;COG_169965;COG_169964;COG_169963;COG_169962;COG_169961;COG_169960;COG_169959;COG_169958;COG_169957;COG_169956;COG_169955;COG_169954;COG_169953;COG_169952;COG_169951;COG_169950;COG_169949;COG_169948;COG_169947;COG_169946;COG_169945;COG_169944;COG_169943;COG_169942;COG_169941;COG_169940;COG_169939;COG_169938;COG_169937;COG_169936;COG_169935;COG_169934;COG_169933;COG_169932;COG_169931;COG_169930;COG_169929;COG_169928;COG_169927;COG_169926;COG_169925;COG_169924;COG_169923;COG_169922;COG_169921;COG_169920;COG_169919;COG_169918;COG_169917;COG_16
```

### 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.csv")
```

```
## 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"
```