E. coli pangenome metabolic capabilities

Daniel Martinez-Martinez

Quarto

Background

We are animals walking in a world of microbes. Since the first moment we set foot on this planet, it had been inhabited by a myriad of other beings that had already mastered the cycles of Earth. Our interdependence with microbes is not only evolutionary but also ecological and functional. They affect our daily life in more ways that we could have ever imagined, in health and disease. But then the opposite is also true. We, the hosts, also affect them and can control their populations at some extent. This endless dance between the two parts is a hard topic to study due to the massive number of elements at a play. To add an extra layer of complexity, microbes are really diverse, and we can find several hundreds of different species in the human gut.

In a recent paper by Almeida *et al.*[1], they cataloged a potential number of species that populate our guts, with around 4500 species characterized. More importantly is the fact that these species are divided in more than 200K different genomes, which points to the presence of more than one reference genome per species. This has profound implications in this context as within the same species we can find a huge genetic diversity that must be explored if we want to study microbe-host interactions in a natural context. The genetic diversity found within a species defines what is known by a **pangenome**, which can be described as the collection of different organisms within a single taxonomic level. The concept has been extensively used to describe the genetic diversity of species, describing how diverse it is. In this regard, we can distinguish between two ends of the spectrum: the open pangenomes, where the accessory genome (i.e., the genes that are strain-specific) is very large in comparison with the core genome (i.e., the genes that are shared by almost every strain); and the closed pangenome where the size of the accessory genome is much smaller. Although we can find some examples in-between, this remains a good way of classifying pangenomes.

This framework has been used to study the *E. coli* pangenome and determining that it is an open pangenome[2] . This species has been under study for decades and it is still a source of unknown features. The *E. coli* species can be further divided in different phylogenetic groups, named as phylogroups. There are more than 10 phylogroups currently defined for *E. coli*, although usually we can find between 8 or 9 (A, B1, B2, C, D, E, F, G are the most common)[3]. These groups have an evolutionary history that make them related also in their ecology, as for example the B2 group is usually enriched with pathogens. What this also defines is that their genomes and accessory genomes will be similar **within** phylogroups compared to **between** phylogroups.

[To follow with more info]

Data exploration

From the metabolic models generated by `gapseq`[4], we can gather all the info into a big table that has this look:

```
# A tibble: 20 x 9
   ID                  Name  Prediction Completeness VagueReactions KeyReactions
   <chr>               <chr> <lgl>             <dbl>          <dbl>        <dbl>
 1 |12DICHLORETHDEG-P~ 1,2-~ FALSE                25              0            1
 2 |14DICHLORBENZDEG-~ 1,4-~ FALSE                28              0            2
 3 |1CMET2-PWY|        fola~ TRUE                100              0            0
 4 |2AMINOBENZDEG-PWY| anth~ FALSE                 0              0            0
 5 |2ASDEG-PWY|        orth~ FALSE                25              1            0
 6 |2OXOBUTYRATECAT-P~ 2-ox~ FALSE                50              0            0
 7 |2PHENDEG-PWY|      phen~ FALSE                50              0            1
 8 |3-HYDROXYPHENYLAC~ 4-hy~ FALSE                42              0            1
 9 |4-HYDROXYMANDELAT~ 4-hy~ FALSE                50              0            1
10 |4TOLCARBDEG-PWY|   4-to~ FALSE                33              0            0
11 |6-HYDROXYCINEOLE-~ 1,8-~ FALSE                 0              3            0
12 |ACETOACETATE-DEG-~ acet~ TRUE                100              0            1
13 |ADENOSYLHOMOCYSCA~ L-me~ FALSE                66              0            0
14 |AEROBACTINSYN-PWY| aero~ FALSE                 0              0            0
15 |ALACAT2-PWY|       L-al~ TRUE                100              0            0
16 |ALADEG-PWY|        L-al~ TRUE                100              0            1
17 |ALANINE-DEG3-PWY|  L-al~ TRUE                100              0            0
18 |ALANINE-SYN2-PWY|  L-al~ TRUE                100              0            0
19 |ALANINE-VALINESYN~ L-al~ TRUE                100              0            0
20 |ALKANEMONOX-PWY|   two-~ TRUE                100              0            1
# ... with 3 more variables: KeyReactionsFound <dbl>, ReactionsFound <chr>,
#   Genome <chr>
```

The different columns from the table give us information about the status of each pathway by genome or strain. We can then see the distribution of how many complete pathways do we

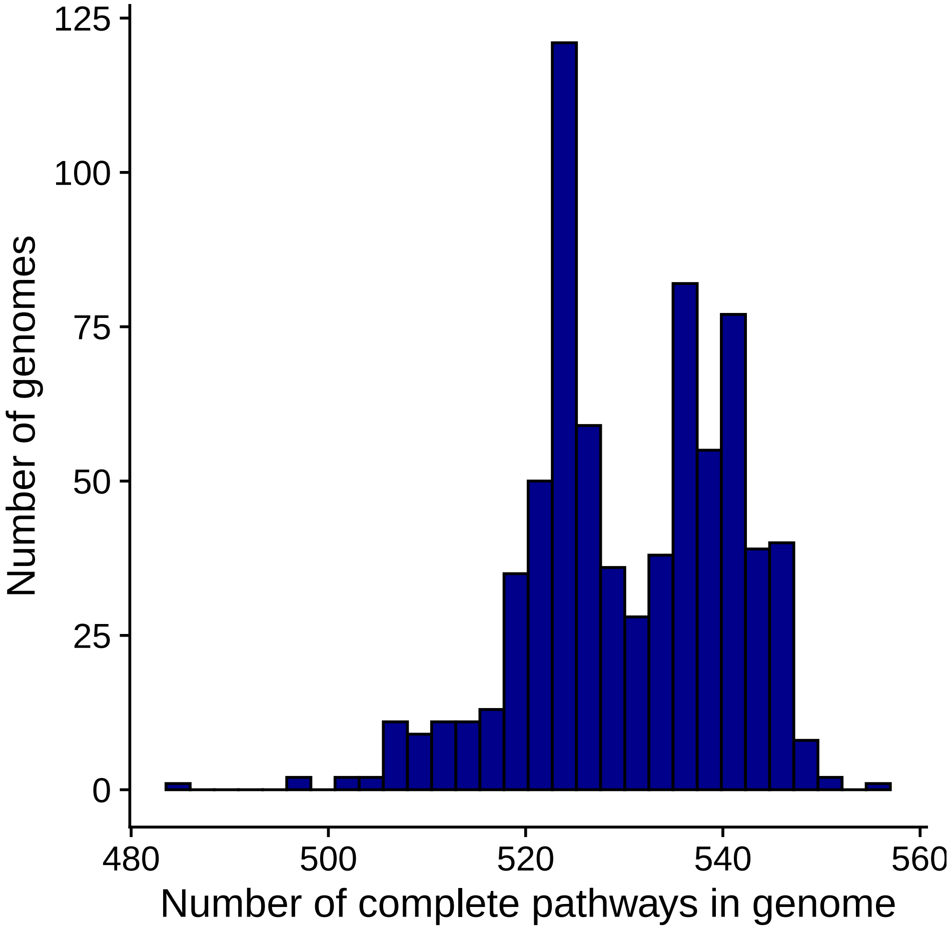

Figure 1: Histogram of complete pathways per genome in our cohort

have per strain, as seen in Figure 1 . By the look of this histogram, seems that we are dealing with a bimodal distribution.

1. Almeida, A. *et al.* A unified catalog of 204, 938 reference genomes from the human gut microbiome. *Nature Biotechnology* **39**, 105–114 (2020).

2. Touchon, M. *et al.* Phylogenetic background and habitat drive the genetic diversification of escherichia coli. *PLOS Genetics* **16**, e1008866 (2020).

3. Abram, K. *et al.* Mash-based analyses of escherichia coli genomes reveal 14 distinct phylogroups. *Communications Biology* **4**, (2021).

4. Zimmermann, J., Kaleta, C. & Waschina, S. Gapseq: Informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biology* **22**, (2021).