

Computational microbial genomics

Giacomo Fantoni

telegram: @GiacomoFantoni

Github: <https://github.com/giacThePhantom/computational-microbial-genomics>

March 7, 2022

Contents

1	<i>Escherichia Coli</i> general informations	2
1.1	<i>E. Coli</i> genomics	2
1.1.1	<i>E. coli</i> long-term evolution experiment	2
1.1.2	<i>E. coli</i> strains	2
1.1.3	PanPhlAn - strain detection and characterization	3
1.1.4	Genomics of <i>E. coli</i>	3

Chapter 1

Escherichia Coli general informations

1.1 *E. Coli* genomics

Escherichia Coli is a Gram-negative, facultative anaerobic, rodshaped, coliform bacterium, it pertains to the phylum of proteobacteria and to the family of Enterobacteriaceae. It can be grown easily and inexpensively. It has got genome with a length between 4.5 - 4.7 M bases, including about 4000-5000 genes, and about seven ribosomal RNA operons. Only the 38% of the genes of K-12 (one of the most studied bacterial strains of *E. coli*) were experimentally identified, overall 40-50% of the genes are to date without a known function. The original *E. coli* strain K-12 was obtained from a stool sample of a diphtheria patient in Palo Alto, CA in 1922.

1.1.1 *E. coli* long-term evolution experiment

The *E. coli* long-term evolution experiment led by Richard Lenski is one of the longest evolutionary experiments ever made (search "*The Longest-Running Evolution Experiment*"). The experiment started on 24th February 1988, and since that moment 12 populations of *E. coli* have been cultivated in the same environment. After each day (corresponding to the time of development of approximately 7 generations), a portion of bacteria from each flask was introduced in a new one, and let proliferating in it. Every 500 generation, it has been saved a sample of the bacteria of each flask, in order to track the evolutionary changes made. Today the experiment is on-going, and researchers reached approximately the 66000th generation. The study suggests a series of conclusions, to cite "long-term adaptation to a fixed environment can be characterized by a rich and dynamic set of population genetic processes, in stark contrast to the evolutionary desert expected near a fitness optimum" (Good et al 2017). In fact, despite of the fixed environment, some bacteria developed the capacity to aerobically grow on citrate, which is unusual in *E. coli* (around generation 31,000) and developed complex mutation patterns.

1.1.2 *E. coli* strains

E. coli could be found as commensal strains, pathogenic strains, or environmental strains. The pathogenic strains could pertain to these categories (which are not exclusive): enteropathogenic

(EPEC), enteroinvasive (EIEC), enterotoxigenic (ETEC), diffusely adherent (DAEC), adherent invasive (AIEC), shiga-toxin producing (STEC), enteroaggregative(EAEC), extraintestinal pathogenic (ExPEC). Resistances to antibiotics make even more difficult the process of categorization of *E. coli*. In 2011 in Germany, an outbreak of Stx-EAEC was responsible of the death of some people. An efficient counter-measure was found by sequencing the genome of those bacteria.

Shigella is *E. coli* with shiga toxin. It had been an issue for taxonomists.

Most of the genes are on plasmids, circular, additional to chromosome, and can be moved easily horizontally. Plasmids between different strains can be moved in enterobacteriaceae, this doesn't happen normally in other families. Some *E. coli* strains are even capable of causing tumors in humans: for example, colibactin-positive *E. coli* can cause colon and rectal cancer, by creating mutations which are responsible of the of the cancer onset.

several antigens can be used by taxonomists to categorize *E. coli* strains. In particular there are the O, H, K antigens, respectively related to the somatic, the flagella and the capsule. O antigens are 171, Ks are 80 and Hs are 56.

1.1.3 PanPhlAn - strain detection and characterization

Pangenome-based Phylogenomic Analysis (PanPhlAn) is a strain-level metagenomic profiling tool for identifying the gene composition and in-vivo transcriptional activity of individual strains in metagenomic samples. PanPhlAn's ability for strain-tracking and functional analysis of unknown pathogens makes it an efficient tool for culture-free infectious outbreak epidemiology and microbial population studies (PanPhlAn reference). This tool was for example used to study the strain responsible of an outbreak in Germany in 2011. This strain was a shiga-toxigenic *Escherichia coli* (STEC), and the study was conducted by Loman and colleagues in 2013. This method outlasted the traditional one.

sequencing means generally to sequence everything, it's normally difficult where to find it, although in *E. coli* is quite easy to understand the provenience. from all the world, populating diversity of *E. coli*, every time we sample we find points which are different from the reference genomes. points overlapping are people living together and share bacteria

1.1.4 Genomics of *E. coli*

In the genome of *E. coli* strains, it is possible to distinguish:

- **Core genome:** the set of all genes shared by all members of a bacterial species, it includes 1000 up to 3000 genes.
- **Accessory or dispensable genome:** the set of genes present in some but not all genomes within the same bacterial species. found on a single strain or in a subset of strains.
- **Pangenome:** core genome + accessory genome. set of all the genes foundable in the species strains. It is said to be "**closed**" when pangenome size tends to a maximum as number of genomes increases, instead it is "**open**" when pangenome keeps increasing as you add new genomes

Sequencing more organism of the same species means to lower the amount of genes in the core genome and augment the number of those in the pan-genome (figure 1.1). Because of technical problems the probability of getting a gene and not forget it is different from 0, so why probably the sequencing of other genomes would lead technically to a plummet to 0 of the pangenome. With some mathematical formulations we can predict a more probable plateau (Rasko David 2008).

Figure 1.1: Core- and pan- genome of *E. coli*

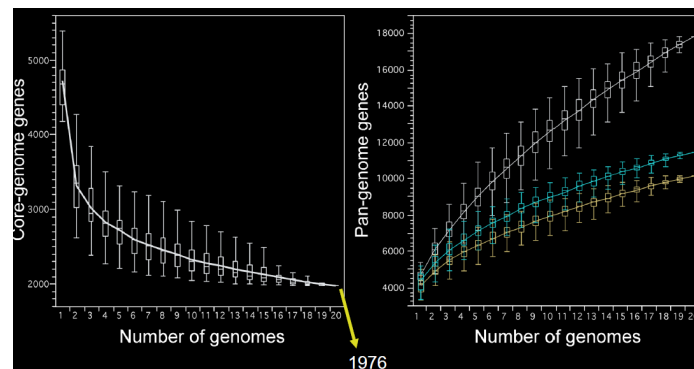
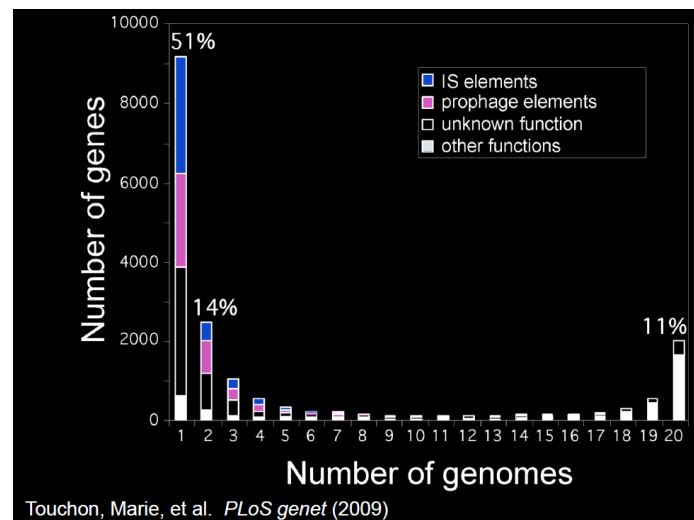


Figure 1.2: It can be seen that 51% of the genes are strain specific, and the other are shared between 2 to 20. of *E. coli*



each *E. Coli* genome contains in a balanced way genes of the core genome and of the pan genome, for a total amount of genes correspondant to about 4700 genes (figure 1.5). Core genomes' genes are responsible of some basic cellular functionalities and utilities to survive environment, while instead elements of the pangenome are quite usually specific to the single strains, they are not always functionally well known.

ratios of the pan-genome and the core-genome are not equal in other organisms behave differently

Figure 1.3: A different definition of orthologous genes can modify the number of pan genes of a species.

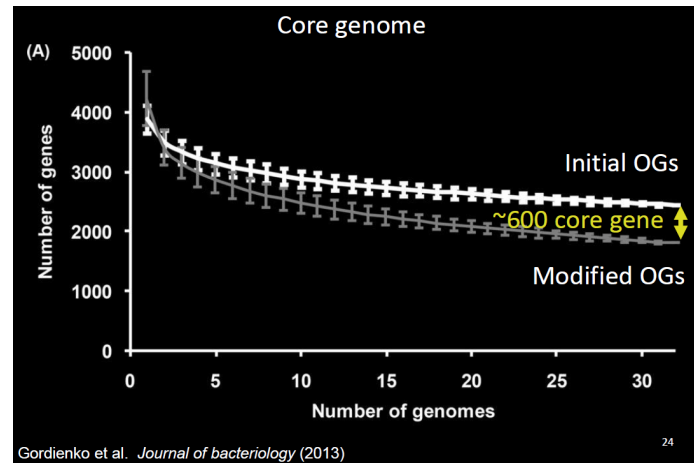


Figure 1.4: we predict a core genome size of 3079 genes for extraintestinal pathogenic *E. coli*

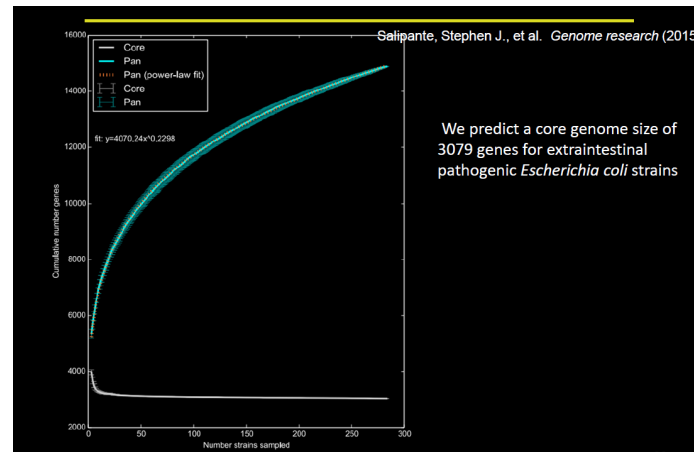


Figure 1.5: Balance between genes of the core- and of the pan-genome

