Bacterial pangenomes

Moritz Buck

December, 2019

Bacterial pangenomes are constrained by genome size

Understanding the population structure of wild microbial populations has been hard, most studies have been done on isolates, and more specifically isolates of a medical nature. The population of these will significantly differ from wild aquatic populations, where microbes have to survive in much poorer environment and are subjected to less bottlenecking events. Recent advances in sequencing and bioinformatics allow us finally to have better insight into a wild aquatic communities through metagenomics. Similarly In this analysis we have used a set of 15000 metagenomic assembled genomes (MAGs) clustered into around 1500 metagenomic operational taxonomic units (e.g. species) from a study involving 400 samples from 40 different freshwater bodies.

This set of MAGs has been used to compute core genomes and pangenomes for each of these bacterial species. This analysis comes to the conclusion that the number of genes available for a species in it's pangenome is constrained by the size of the genomes, allowing larger genomes to access to even more niches then previously thought.

Main

Size and diversity of microbial pangenomes is an open question, the data and the tools necessary have not been available for very long.

Computing core-genomes for has been limited in taxonomic scope, as most methods require a number of high-quality genomes of closely related organisms to have an accurate estimate of presence abscence in the population of interest. We present here a novel bayesian method for the computation of core genomes relying on the presence-abscence of genes/COGs/annotations in sets of draft (or complete) genomes. The method is wrapped in a tool available at this webpage: www.github.com/moritzbuck/mOTUlizer.

Each genome (we will use genome as a shorthand for any set of nucleotide sequences belonging to the same biological entity, e.g. draft genome, complete genome, or Metagenome Assembled Genome), is described as a set of traits (in the case of this analysis a set of COGs, but mOTUlizer is agnostic to the type of traits). And each genome it self is part of a set of genomes which we will call an mOTU (metagenomic Operational Taxonomic Unit, due to the nature of the data we analyse here). mOTUlizer uses an iterative bayesian approach to classify each trait of the genome in an mOTU as a "core"-trait or "auxiliary"-trait based on a likelihood ratio. For each of the two hypotheses (core-trait or auxiliary-trait) a probability is computed assuming a certain completeness values for the each genome. Whicever of these is more likely is picked as class for that trait. Using this new classification we update the completeness estimate and recalculate the likelyhood ratios and repeat the process until convergence.

To compute the probability of a distribution of a specific trait in an mOTU under the assumption that it is in the core, we will simply multiply the completeness c_g of the genomes g that have that COG, with the inverse probability $1 - c_p$ for the genomes that to not have that trait, e.g. equations 1 and 2.

(1)

$$p_{\text{trait}|\text{core}} = \prod_{\substack{\text{g in mOTU} \\ \text{if trait in g}}} p_{\text{trait in g}|\text{core}} \prod_{\substack{\text{g in mOTU} \\ \text{if trait not in g}}} (1 - p_{\text{trait in g}|\text{core}})$$

(2)

$$p_{\text{trait in g}|\text{core}} = c_g$$

For the probability under the assumption that it is in the auxiliary fraction of the genome, we will have to make some assumptions on the structure of the pangenome. We have assumed that the traitsin the pangenome that are not in the core, are independent, and each trait has a frequency $\frac{|\text{trait}|}{|G|}$ where |trait| is the number of genomes in the mOTU that have that trait, and $|G|$ the total size of the traits-pool. To "fill" the auxiliary fraction of a genome, we draw "$|g| - c_g|\text{core}_{\text{mOTU}}|$"-times, which is the number of spots in the auxiliary part of the genome assuming a genome with $|g|$ traits, core size $|\text{core}_{\text{mOTU}}|$ and completeness c_g. Resulting in equations (3) and (4).

(3)

$$p_{\text{trait}|\text{aux}} = \prod_{\substack{\text{g in mOTU} \\ \text{if trait in g}}} (1 - \bar{p}_{\text{trait in g}|\text{aux}}) \prod_{\substack{\text{g in mOTU} \\ \text{if trait not in g}}} \bar{p}_{\text{trait in g}|\text{aux}}$$

(4)

$$\bar{p}_{\text{trait in g}|\text{aux}} = (1 - \frac{|\text{trait}|}{|G|})^{|g| - c_g|\text{core}_{\text{mOTU}}|}$$

For practical reasons, these computations are all done in log-space. resulting into a log-likelyhood ratio:

(5)

$$LLHR = \log{(p_{\text{trait}|\text{core}})} - log(p_{\text{trait}|\text{aux}})$$

,

if this is positive, the trait is considered core, if negative, auxiliary. Using this classification we recompute for each genome an updated completeness:

(6)

$$c_g = \frac{|\text{core}_{\text{mOTU}}\text{in g}|}{|\text{core}_{\text{mOTU}}|}$$

.

And rerun the likelihood computation. This is repeated until convergence, to obtain a final set of core-traits and and auxiliary-traits. Unique convergence depending on initial completeness scores has been tested (Supp ?) and stable convergence is obtained if the number of genomes is larger then 5.

We used this new developed statistical method for the computation of core genomes to estimate core genomes and auxiliary genomes for ~1200 (meta-)genomic Operational Taxonomic Units (mOTUs) from public repositories and ~300 mOTUs from our own dataset of freshwater metagenomes.

For these selected ~1500 mOTUs we have computed core genomes using cluster of orthologous genes (COGs) as sequence based traits. To account for possible errors both in gene-prediction and metagenomic binning, all COGs appearing only once have been removed. So a core-COG will be a COG that is expected to be present in every genome of that mOTU, and an auxiliary-COG is a COG that is present in a fraction of the genomes of that mOTU, but at least two different genomes.

The core genomes obtained correspond to a large variety of genome sizes (Fig. 1). It is noticeable that the GTDB dataset has two strong peaks corresponding to the classes Bacilli (phylum Firmicutes) and Gammaproteobacteria (phylum Proteobacteria), which are heavily represented in the database. In general, core genomes of the genomes in the database is larger then in our environmental dataset (GTDB : mean 2844.31, sd 1307.647; Anoxic MAGs : mean 2357.898, sd 990.5579), which exhibits a general bias in of public databases to larger genomes (typically soil and disease associated microbes).

The converse fraction of the core genome, the variable fraction, which we define in our analysis as the COGs of a genome that are not part of the core, but are found in at least one other genome of this mOTU (singeltons are removed to remove spurious noise that could have been introduced by errors in binning of MAGs or gene-prediction, in general however results hold if they are kept). The gene-pool associated with all the variable fractions of the genomes of an mOTU, we will call auxiliary genome of an mOTU, as opposed to the pangenome which is the set of all the genes (COGs) that appear in a species.

The size of the variable genome of an mOTU correlates with the size of the core-genome (Fig 2a), larger genomes have more auxiliary COGs. However the actual fraction of the genome that is variable does not change relative to the genome size. Indicating an implicit limitation in the fraction of the genome that can be variable, e.g no matter how many genes you have in your genome only X out of Y genes can be from the auxiliary genome, more specifically about one in 15 COGs.

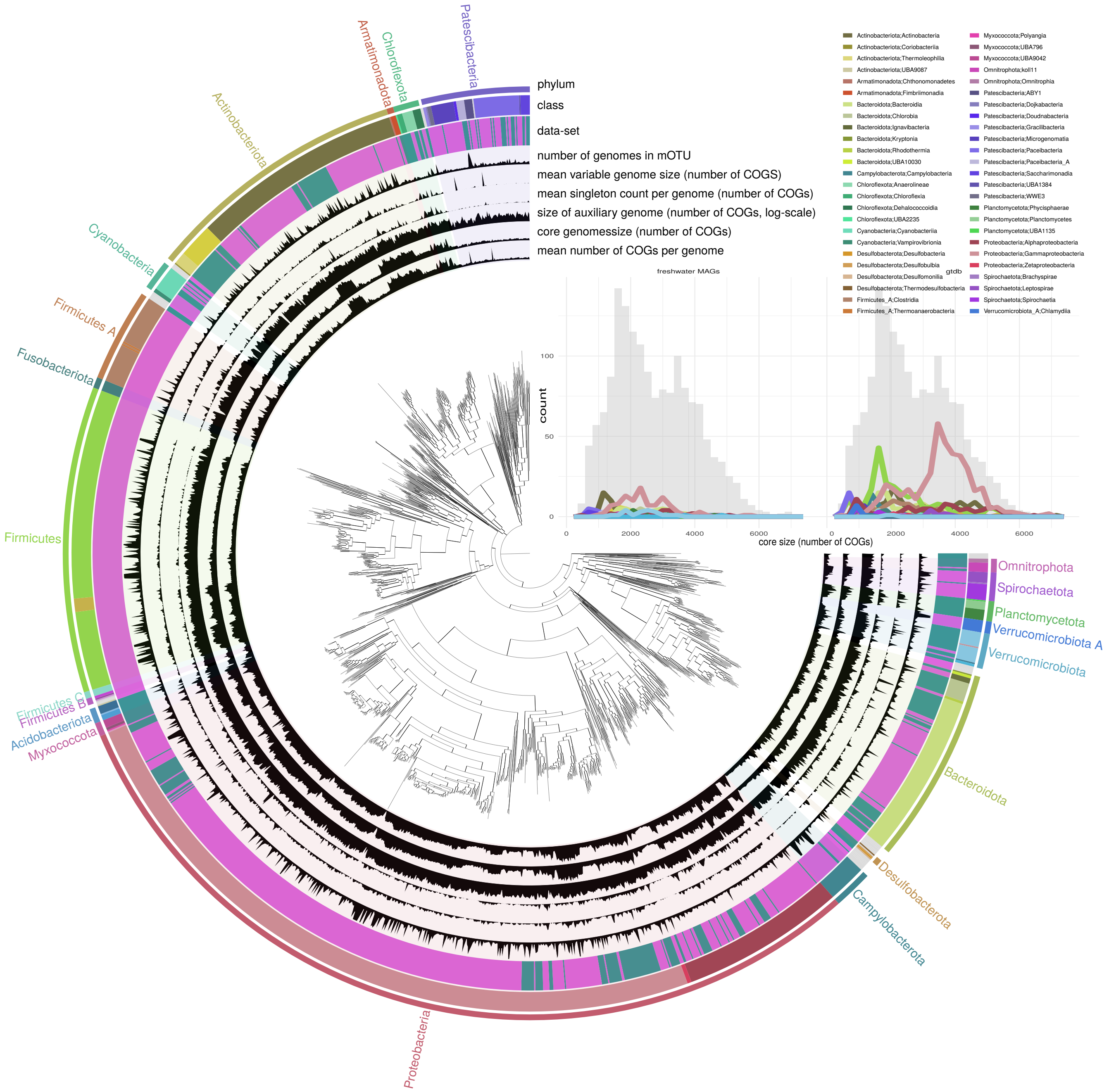

Figure 1: Tree of representative genomes of mOTUS, (corner panel) distribution of core genome sizes (number of COGs)

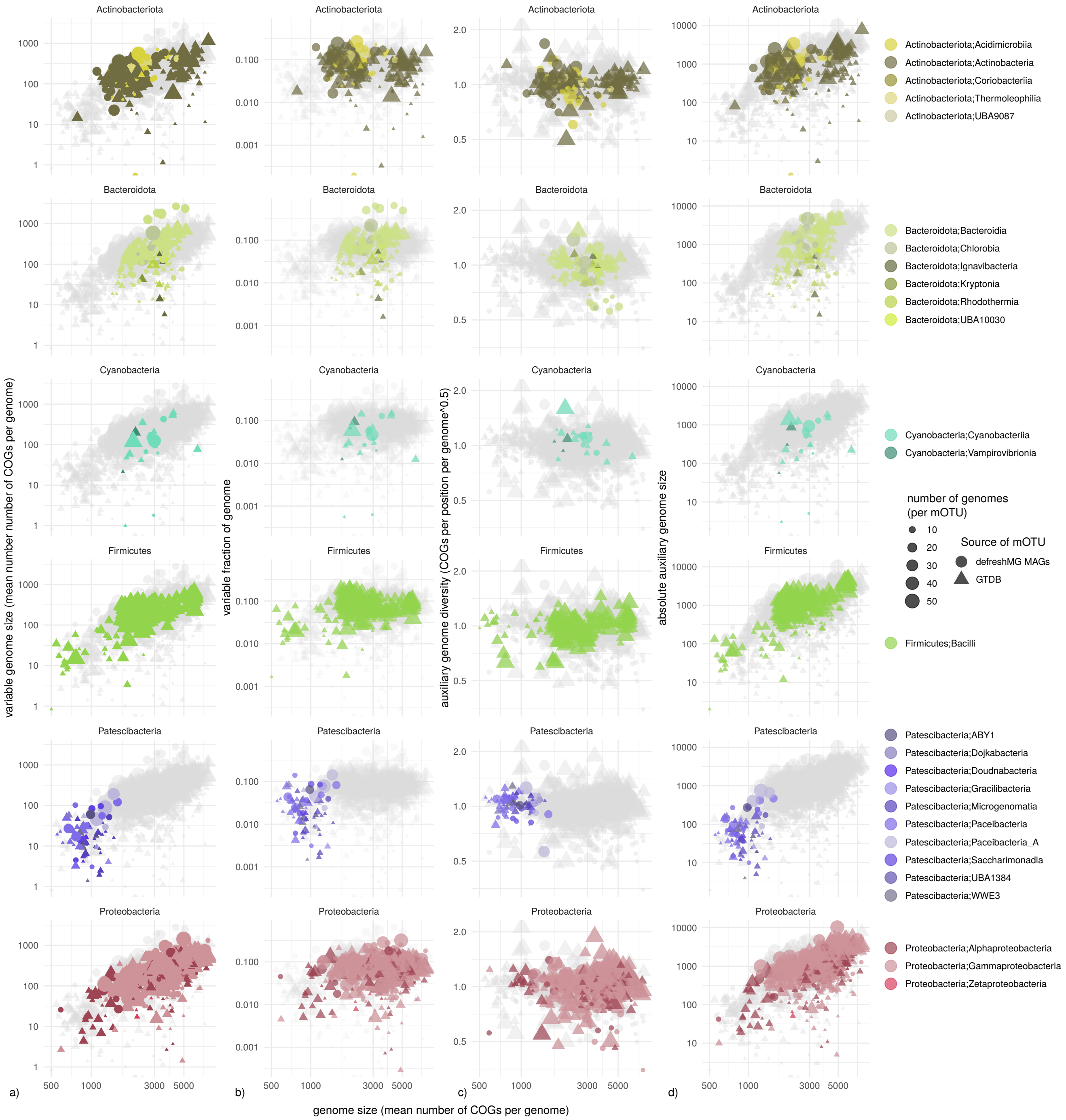

Figure 2: Variable genome structure and diversity, a-b) variable-genome correlates with genome size as fraction of variable genome is the same for all genome sizes, c)

Supplemental figures

Supplemental data

Core genomes