Identification of horizontal genes transfer elements across strains inhabiting the same niche using pan-genome analysis

J. Schwarzerová[1,2] and D. Čejková[1]

[1] Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication (FEEC), Brno University of Technology (BUT), Brno, Czech

[2] Molecular Systems Biology (MOSYS), University of Vienna, Vienna, Austria

E-mail: Jana.Schwarzerova@vut.cz, cejkovad@vut.cz

Abstract—Tracing horizontal gene flux across strains in farm animals is one of the important steps for research focused on detection and genomic enzymology of genes conferring antibiotic resistance. In this study, we have built the comprehensive computational methodology for the detection of horizontal genes transfer elements via pan-genome analysis. In total, 133 anaerobes isolated from chicken gastrointestinal tract were examined for the presence of traits of horizontal transfer. The shared genes from all isolates, so called core genome genes were identified and characterised in order to assign the function to the gene within individual bacterial cell and within community of cells. This study provides an evidence that horizontal transmission frequently occurs not only between closely related bacteria, but also between distant taxonomical groups. Hence chickens are known primary reservoirs of antibiotic resistance genes, and the dissemination of these genes to other bacterial pathogens often leads to life-threatening infections, even within human population. Thus, the research on this subject, and the associated results are of a great importance for public health.

Keywords — Comparative Genomics, Core Genome, Antibiotics Resistance, Chicken Microbiome

1. INTRODUCTION

Horizontal gene transfer (HGT) plays major role in bacterial evolution. HGT is also known as 'the non-genealogical transmission of genetic material from one organism to another' [1]. HGT is an important driving force that modulate bacterial genomes, and thus plays a pivotal role in evolution of prokaryotes. HGT is often mediated via mobile genetic elements (MGE) which are able to acquire and harbor foreign genetic material, including antibiotic resistance genes. Bacteria carrying such cargo have evolutionary advantage if particular antibiotics is present in the niche, they overgrow other bacteria. Later, the MGE can be spread vertically and horizontally and additional genes can be acquired by the same bacteria. Bacteria harboring many antibiotic resistance genes are called multidrug resistance bacteria (MRB). MRB, often human pathogen associated with nosocomial infections, can be resistant to many known antibiotics; e.g. methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant *Enterococcus faecalis* represent one of major threats for public health [2]. Therefore, it is very important to trace such genes as well as their vehicles, MGE, in the bacterial community.

The study used whole-genome sequencing data to trace horizontal gene-flux across strains and provide insight into species evolution. One of the most used methods in multi-genome studies is pan-genome analysis. Pan-genome is the term that was inferred in the study by Tettelin et al. [3]. The authors discovered that different strains of *Streptococcus sp.* might differ substantially in their gene content and total gene pool of a species might be orders of magnitudes larger than the gene content of any single strain. Pan-genome analysis provides information on genomic diversity of the investigated bacteria, determining core (conserved, that can characterize the biological function in large microbial clades), accessory (dispensable genes in different species) and unique (strain-specific) gene pool of a species [4]. Our study aims to determine core and accessory sequences from 133 gastrointestinal tract (gut) anaerobes isolated from chicken caecum in pure cultures and to detect gene elements playing role as mediator in HGT.

2. MATERIALS

Dataset originated from the study by Medvedcky et al. [5]. In the study, 204 novel bacterial isolates from chicken caecum were isolated via cultivation on the Wilkins-Chalgren anaerobe agar under anaerobic growth conditions. The genomes were sequenced using Illumina NextSeq 500 platform [6]. Raw sequencing reads were quality trimmed using Trimmomatic [7] and assembled by IDBA-UD [8]. Our follow-up study used the annotated genomes data that are deposited in NCBI under accession number PRJNA377666 [5]. In total, it is 133 draft genomes from gut anaerobes isolated from chicken caecum in pure cultures.

3. METHODS

Methodology carried out in this study is based on pan-genome analysis. The whole analysis is performed on seven different strains such as Actinobacteria, Bacteroidetes, Elusimicrobia, Firmicutes, Proteobacteria, Synergistetes and Verrucomicrobia. Firstly, we divided these strains into two different groups. The first group includes Gram-positive, the second group includes Gram-negative bacteria group. Pan-genome analysis was performed for both groups. Throughout the study, gene and protein sequences were analyzed. However, the term core genes are widely used, so we will use names for core genes and core protein sequences interchangeably.

The pan-genome analysis was performed by Bacterial Pan Genome Analysis (BPGA) tool [9]. Besides defining the core, accessory and unique genome gene pools, BPGA also enables additional features for downstream analyses, core/pan/MLST (Multi Locus Sequence Typing) phylogeny, exclusive presence/absence of genes in specific strains, subset analysis; atypical G + C content analysis of core, accessory and unique genes, to name a few [9]. Thus, the BPGA tool shown new insight into the analysis of chicken gut genomes.

The identified core genome genes from each group were functionally annotated, i.e. superfamily and Clusters of Orthologous Groups (COG) were assigned to every gene to predict their function within bacterial cell and/or community. Firstly, the core genome gens were analysed using Batch CD Search tool [10]. The superfamilies were visualized using Venn diagram. The overlapping region containing shared superfamilies of both groups was analyzed in more detail using CDD Search tool [11]. The COG assignment was performed using eggNOG-mappper tool [12]. The results were analysed and visualized using R, expressly, R/ Biostrings [13], R/ seqinr [14] and R/ stats packages [15] were applied. The whole pipeline of methodology is shown in Figure 1.

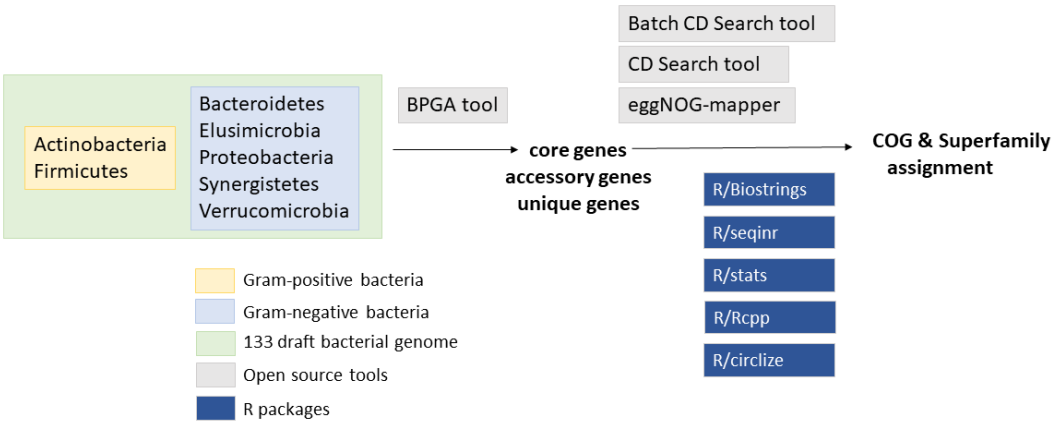

Figure 1: The methodology used in the study; in yellow are marked Gram-positive bacteria which represents Actinobacteria and Firmicutes; in blue are marked Gram-negative bacteria represented by Bacteroidetes, Elusimicrobia, Proteobacteria, Synergistetes and Verrucomicrobia phyla.

4. RESULTS & DISCUSSION

The original dataset was divided into two groups. Gram-positive bacterial represents first group and include 99 draft genomes. Gram-negative bacteria represented the second group and include 34 draft genomes. In total, 133 draft genomes were analyzed.

Pan-genome analysis

In total, 365,298 genes were identified, 259,514 genes were identified in Gram-positive and 105,784 were identified in Gram-negative bacteria. The highest number of core genes were in group of Gram-positive bacteria. It included 2,475 core genes, see Table I. The number of core genes from Gram-negative bacteria was significantly lower including 340 core genes. The low number of core genes in Gram-negative bacteria might be caused because the group was consisted of taxonomically distant bacterial genomes which were examined, so the members of this group likely do not share many genes. Therefore, they may bear only hundred gene in common. Thus, there is very different distribution based on standard deviation from average GC content, on contrary to conventional pan-genome analysis.

In addition, pan-genome analysis also identified accessory and unique genes. However, also accessory genes can be mediated by and/or associated with HGT. The accessory genes will be further analysed in follow-up studies. Summary of pan-genomes analysis is shown in Table I.

Table I: Number of sequences from pan-genome analysis using BPGA tool. The default setting for pan-genome analysis was used: 2 × standard deviation (SD) from average GC content.

Group of genomes	Core genes	Accessory genes	Unique genes
Gram positive	2,475	217,317	39,722
Gram negative	340	79,403	26,041

Determination of horizontally transferable gene elements

The core genes were used as input in Batch CD Search tool. In total, 10 superfamilies were assigned to 302 core genes in Gram-negative bacteria, as well as 10 superfamilies were assigned to 900 core genes in Gram-positive bacteria. This intersection represents 3 superfamilies, especially cl3508, cl40667 and cl35051, see Figure 2. Superfamily cl35085 represents molecular chaperone DnaK. cl40667 represents ribosomal protein L2 and cl35051 is an elongation factor EF-Tu. In total, 17 superfamilies were identified in all analysed draft genomes. The rest 613 genes belonged to gene of unknown function.

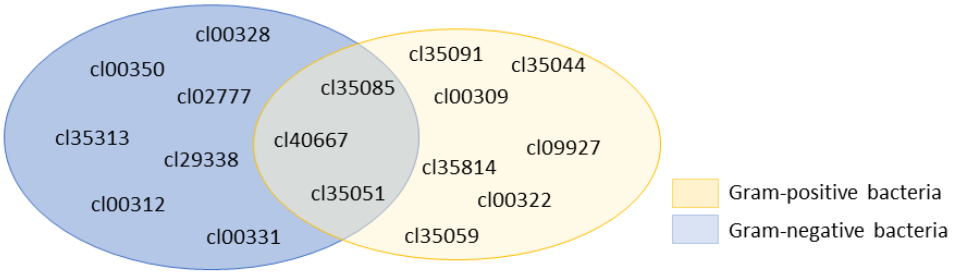

Figure 2: Venn diagram of identified superfamilies identified via Batch CD Search tool.

In more detail, we focused on the intersection set. The next analysis was performed by standard CD search tool using Blastp [16]. The verification of the associated superfamily and assigned gene function was performed using visualization from CD search tool.

In addition, the COG assignment was performed for core gene sequences using eggNOG-mapper tool. On summary, we identified 25 diverse clusters in Gram-positive bacteria and 10 diverse clusters in Gram-negative bacteria. In total, we determined 4 different COG categories, see Figure 3. The most common category is J, group of genes involved in translation, ribosomal structure and biogenesis function.

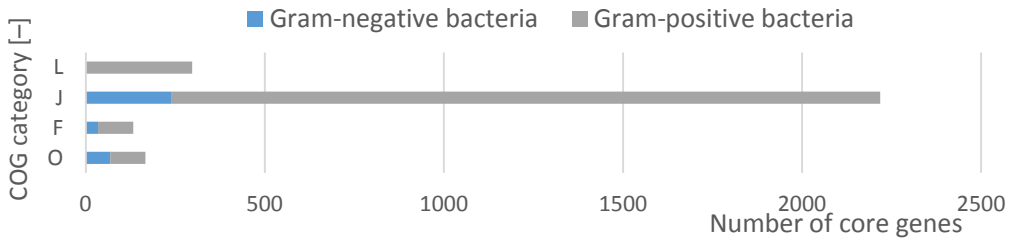

Figure 3: COG distribution using eggNOG-mapper. L: Replication and repair category, J: Translation, ribosomal structure and biogenesis, F: Nucleotide metabolism and transport O: Post-translational modification, protein turnover, and chaperone.

In the last, the association between individual core genes and COGs were visualised by heatmap. For clarity, illustration of the analysis containing first thirty core gene sequences are depicted in Figure 4. Considerable amount of determined elements was connected to horizontal gene transfer analysis. For example, *tuf* gene is identified as gene with possible HGT [17], or *dnaK* gene is also often mentioned in literature about HGT [18][19]. Moreover, *dnaK* gene is connected to research focused to multidrug resistance [20].

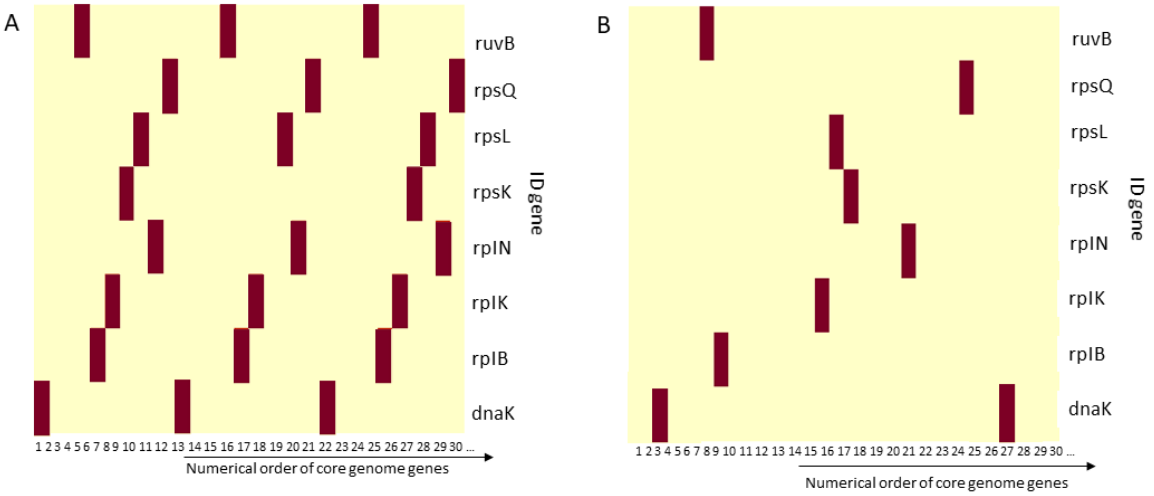

Figure 4: Heatmaps represent examples of identified genes obtained by eggNOG-mapper in chicken microbiome and their prevalence across core sequences. (A) represents example for Gram-negative bacteria; (B) represents Gram-positive bacteria.

5. CONCLUSION

Advantages of high-throughput low-cost sequencing technologies and metagenomic techniques have shift the research interest from single or few genome analyses to large-scale comparison studies. It allows us to pave new avenues for novel research challenges, such as the detection of horizontal gene transfer in diverse bacteria. One of such challenge comprised our study focusing on identification of transferable genes and elements in bacterial genomes, not between closely related but also between distant bacterial members inhabiting the same environment.

Our study used pan-genome analyses to determine core-genome of chicken microbiome for Gram-negative and Gram-positive bacteria. These core-genomes genes were used in subsequent analysis in which functional COG and superfamily categories were assignment to each gene. Especially, functional COG and superfamily categories were assignment to each gene. We identify 10 superfamilies in Gram-positive and Gram-negative bacteria using CD Batch tool. Nevertheless, we also identify 10 ID genes in Gram-negative bacteria and 25 ID genes in Gram-positive bacteria using eggNOG-mapper that is spaced across both of core genome genes. These identified genes include genes which connects to HGT, which is a promising way to open further studies.

ACKNOWLEDGMENT

This work has been supported by grant project GACR 22-16786S.

REFERENCES

[1] BOTO, Luis. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences*, 2010, 277.1683: 819-827.

[2] ARGEMI, Xavier, et al. Comparative genomic analysis of *Staphylococcus lugdunensis* shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC genomics*, 2018, 19.1: 1-16.

[3] TETTELIN, Hervé, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 2005, 102.39: 13950-13955.

[4] CHAUDHARI, Narendrakumar, et al. BPGA-an ultra-fast pan-genome analysis pipeline. *Scientific reports*, 2016, 6.1: 1-10.

[5] MEDVECKY, Matej, et al. Whole genome sequencing and function prediction of 133 gut anaerobes isolated from chicken caecum in pure cultures. *BMC genomics*, 2018, 19.1: 1-15.

[6] PAIJMANS, Johanna LA, et al. Sequencing single-stranded libraries on the Illumina NextSeq 500 platform. *arXiv preprint arXiv:1711.11004*, 2017.

[7] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

[8] PENG, Yu, et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 2012, 28.11: 1420-1428.

[9] CHAUDHARI, Narendrakumar M.; GUPTA, Vinod Kumar; DUTTA, Chitra. BPGA-an ultra-fast pan-genome analysis pipeline. *Scientific reports*, 2016, 6.1: 1-10.

[10] MARCHLER-BAUER, Aron, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research*, 2010, 39.suppl_1: D225-D229.

[11] MARCHLER-BAUER, Aron; BRYANT, Stephen H. CD-Search: protein domain annotations on the fly. *Nucleic acids research*, 2004, 32.suppl_2: W327-W331.

[12] CANTALAPIEDRA, Carlos P., et al. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*, 2021, 38.12: 5825-5829.

[13] PAGES, Hervé, et al. Biostrings: String objects representing biological sequences, and matching algorithms. *R package version*, 2016, 2.0: 10.18129.

[14] CHARIF, Delphine, et al. Package 'seqinr'. 2021.

[15] *Documentation for package 'stats' version 4.2.0: The R Stats Package* [online]. Available from: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html

[16] WHEELER, David L., et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 2007, 36.suppl_1: D13-D21.

[17] KACAR, Betül, et al. Functional constraints on replacing an essential gene with its ancient and modern homologs. *MBio*, 2017, 8.4: e01276-17.

[18] PETITJEAN, Céline, et al. Horizontal gene transfer of a chloroplast DnaJ-Fer protein to Thaumarchaeota and the evolutionary history of the DnaK chaperone system in Archaea. *BMC evolutionary biology*, 2012, 12.1: 1-14.

[19] HAGHI, Morteza, et al. Detection of heat shock protein (DnaK, DnaJ and GrpE) horizontal gene transfers among *Acanthamoeba polyphaga*, *Acanthamoeba polyphaga mimivirus* (APMV), amoeba-infecting bacteria and sputnik virophage. *Int J Adv Biotechnol Res*, 2016, 7: 1618-1622.

[20] MUJAWAR, Shama, et al. Variant analysis from bacterial isolates affirms DnaK crucial for multidrug resistance. In: *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer, Cham, 2020. p. 237-248.