Send Orders for Reprints to reprints@benthamscience.ae

Current Genomics, 2015, *16*, 245-252

Inside the Pan-genome - Methods and Software Overview

Luis Carlos Guimarães[1,2,*], Leandro Benevides de Jesus[1], Marcus Vinícius Canário Viana[1], Artur Silva[2], Rommel Thiago Jucá Ramos[2], Siomar de Castro Soares[3] and Vasco Azevedo[1,*]

[1]*Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Avenue Antônio Carlos, 6627, Belo Horizonte, Minas Gerais, Brazil;* [2]*Department of Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil;* [3]*Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences Federal University of Triângulo Mineiro, Uberaba, Minas Gerais, Brazil*

V. Azevedo

Abstract: The number of genomes that have been deposited in databases has increased exponentially after the advent of Next-Generation Sequencing (NGS), which produces high-throughput sequence data; this circumstance has demanded the development of new bioinformatics software and the creation of new areas, such as comparative genomics. In comparative genomics, the genetic content of an organism is compared against other organisms, which helps in the prediction of gene function and coding region sequences, identification of evolutionary events and determination of phylogenetic relationships. However, expanding comparative genomics to a large number of related bacteria, we can infer their lifestyles, gene repertoires and minimal genome size. In this context, a powerful approach called Pan-genome has been initiated and developed. This approach involves the genomic comparison of different strains of the same species, or even genus. Its main goal is to establish the total number of non-redundant genes that are present in a determined dataset. Pan-genome consists of three parts: core genome; accessory or dispensable genome; and species-specific or strain-specific genes. Furthermore, pan-genome is considered to be "open" as long as new genes are added significantly to the total repertoire for each new additional genome and "closed" when the newly added genomes cannot be inferred to significantly increase the total repertoire of the genes. To perform all of the required calculations, a substantial amount of software has been developed, based on orthologous and paralogous gene identification.

Keywords: Pan-genome, Core genome, Accessory genome, Species-specific genome, Comparative genome.

BACKGROUND

The advent of Next-Generation Sequencing (NGS) has allowed the reduction in the time and cost per genome sequenced [1-3]; with the use of this tool, we have observed an exponential increase in the number of whole genome sequences that have been deposited in public databases (http://www.genomesonline.org). In this context, the large number of genomes available boosted the development of comparative genomics and, consequently, the rise of the pan-genomic area [4, 5].

Comparative genomics is the direct comparison of the genetic content of an organism against another, and its main aim is to obtain a better biological understanding of many species [6]. This approach could help to determine gene function and coding region sequences of genomes as well as to characterize the frequency of evolutionary events, such as genome plasticity, and to establish phylogenetic relationships [7, 8]. Most of the comparative analyses have as an objective to identify similarities and differences among the organisms [9].

*Address correspondence to these authors at the Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Avenue Antônio Carlos, 6627, Belo Horizonte, Minas Gerais, Brazil; Tel/Fax: +55 (31) 3409-2610; E-mails: luisguimaraes.bio@gmail.com; vascoariston@gmail.com

A comparative genomics approach is used often in many different aspects of science, such as in the comparison of the *Drosophila melanogaster* (fruit fly - model organism) genes versus human genes, where 548 human genes were identified as homologous in the fly genome. All of these genes are linked to human diseases of different natures (cardiovascular, visual, auditory, endocrine and skeletal diseases) [10]. Thus, the finding of homologous genes that are commonly shared between humans and model organisms has opened the possibility of testing new therapies in model organisms [6].

Similarly, comparative genetics can be used in prokaryotic organisms, e.g., in the comparison of *Bacillus licheniformis,* which is a gram-positive bacterium of biotechnology and pharmaceutical interest and is used for the expression of proteins and antibiotic production, in two related species (*Bacillus subtilis* and *Bacillus halodurans*). The comparison among these three bacteria not only enabled the assembly of the *Bacillus licheniformis* genome but also helped in evolutionary studies and the identification of horizontal gene transfer between them [11]. Furthermore, comparative genomics analyses in related species have shown an extensive genomic intra-species diversity and highlighted the associated bacterial promiscuity [12].

However, comparative genomics can be used with a large number of bacteria with distinct lifestyles. A study that used three hundred and seventeen genomes was performed, aim-

1875-5488/15 $58.00+.00

ing to establish patterns among the organisms' lifestyles, their gene repertoires and the sizes of their genomes. As a result, the authors observed that intracellular pathogens are more prone to gene loss, or reductive genome evolution [13]. Thus, the availability of thousands of bacterial genomes in databases and the use of comparative genomics taking a variety of approaches have allowed the development of new terms such as pan-genome, core genome and accessory genome [14-16].

PAN-GENOME

The main goal of pan-genome is the genomic comparison of different strains of the same species, or even genus [17, 18]. Currently, the availability of a large number of genomes from different isolates of the same pathogen has opened the possibility of investigating several genomic characteristics that are intrinsic to one or more species [16]. One way to investigate these attributes is through the pan-genomic approach [15].

The first work that described the term pan-genome was conducted by Tettelin and colleagues (2005), who used eight different strains of *Streptococcus agalactiae*, a pathogenic species isolated from human. After this research, other studies were performed using pan-genomic analysis for different microorganisms, including *Bacillus cereus* [19], *Escherichia coli* [20], Sulfolobus islandicus [21], Streptococcus pneumoniae [22], Methanobrevibacter smithii [23], Corynebacterium diphtheriae [24], Corynebacterium pseudotuberculosis [25], and Pantoea ananatis [26], among others.

The idea of pan-genomic studies brings significant insights of the understanding of bacterial evolution, niche adaptation, population structure and host interaction as well as inferences in more applied issues, such as vaccine and drug design and the identification of virulence genes.

The term "pan-genome" reflects the total number of non-redundant genes that are present in a given dataset [16, 17]. It consists basically of three parts: i) core genome, formed by genes shared by all genomes and usually involved in essential cellular processes; ii) accessory or dispensable genome, composed of genes absent in some isolates; and iii) species-specific or strain-specific genes, which are those genes that are present in a single genome [16, 27] (Fig. **1**). Usually, genes that are present in accessory and species-specific or strain-specific genes are involved in niche adaptation [5, 28].

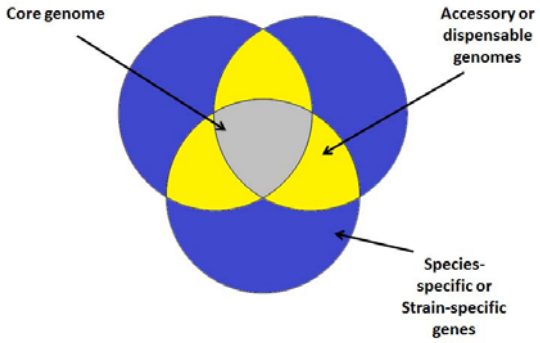

Fig. (1). Venn diagram plot that represents the three parts of the pan-genome. In gray: core genome; yellow: accessory or dispensable genome, and blue: species-specific or strain-specific genes. Adapted: Muzzi *et al.*, 2007.

In this review, we describe the different approaches to studying the pan-genome and its sub-products (core genome, accessory or dispensable genome and species-specific or strain-specific genes), and we discuss the impact that the pan-genome concept has on characterizing the bacterial life style.

CORE GENOME, ACCESSORY OR DISPENSABLE GENOME, AND SPECIES-SPECIFIC OR STRAIN-SPECIFIC GENES

Core Genome

The core genome is the subset of genes that are present in all of the genomes, and it can be determined by comparing the different genomes [15]. Lapierre and Gogarten (2009) said that over 250 gene families have been characterized as part of the bacterial core genome and these gene families constitute evidence that gene conservation highlights the conservative nature of evolution.

Normally, genes that are present in the core genome are associated with the maintenance of the basic aspects of the organism's biology and are mainly related to replication, translation and maintenance of cellular homeostasis [16, 28].

Moreover, the core genome undergoes significant selective pressure in relation to its function, which inhibits the occurrence of drastic changes [14]. The number of genes that compound the core genome could indicate the genetic diversity among the studied organisms; thus, the core genome becomes smaller when diversity increases among the organisms [29]. On the other hand, phylogenetically related genomes tend to share more genes and consequently present a larger core genome [29, 30].

Accessory or Dispensable Genome

The accessory or dispensable genome is the subset of genes that is shared by some organisms but is not present in all of the studied organisms, and it is represented by approximately 8000 gene families [14]. This subset includes genes that have specific functions that are related to survival in different niches, and usually they are associated with virulence or resistance to antibiotics and can be reflected in the organism lifestyle [27, 31].

The accessory or dispensable genome has been described as variations in gene sequences that can provide the emergence of new functions from the genes [14]. Although similar at the nucleotide level, they show a high diversity in their specificity for substrates. The accessory genome could have emerged by horizontal gene transfer and paraphyletic evolution, where it occurs as gene duplication followed by mutation. Additionally, the strain divergence can occur [32, 33]. For example, the ABC transporters gene family presents various types of substrate specificity, which is caused by nucleotide substitutions in its binding periplasmic sites [14].

Species-specific or Strain-specific Genes

Species-specific genes are present in a single species at the inter-species level, whereas strain-specific genes are only present in one strain and are at the intra-species level [15].

Normally, species-specific or strain-specific genes are obtained by horizontal gene transfer among species. According to Lefébure and colleagues [34], who conducted a pan-genomic study with the *Streptococcus* genus, the species-specific genes are represented by 139,000 gene families [34, 35]. The presence of these genes could confer an adaptive advantage over those strains that lack them. Moreover, studies have shown that those genes have a connection with virulence or pathogenicity in pathogenic organisms. [36, 37]. In non-pathogenic organisms, these genes could have a connection with metabolism and could be metabolic islands that are acquired by horizontal gene transfer [38].

In general, this group of genes is under relaxed mutational pressure, with mutations occurring constantly in its sequences, in contrast with genes that are present in the core genome, which have constant selective pressure to maintain their conserved sequence [39, 40]. When mutations occur successfully, raising bacterial adaptation to specific environments and conditions, the genes can be maintained in the genome and shared among species that are integrated into the accessory genome (bacterial evolution). On the other hand, mutations can lead to the creation of pseudogenes (unfunctional genes), which, during the evolutionary process, could be excluded from the genomes [14].

Jordan and colleagues (2001) made a study with strain-specific genes in which they analyzed 21 genomes, and they observed that strain-specific genes ranged from 5% to 35% per genome. They also observed that the majority of strain-specific genes were duplicated, i.e., most are paralogous genes that are arranged in tandem. Generally, these genes are considered to be virulence factors because they can encode surface-exposed proteins, which would confer on pathogenic bacteria the ability to bind to cell hosts [41, 42].

Open and Closed Pan-genome

To determine the number of genomes to be sequenced to obtain the complete gene repertoire of a given species or related organisms, it is necessary to determine how many extra genes are to be added for each newly sequenced genome [5, 16]. Thus, we have the concept of the open or closed pan-genome.

Tettelin and colleagues [16] used mathematical extrapolation of the data; as a result, they observed that the *S. agalactidae* pan-genome is enormous and that unique genes will always continue to be identified even after hundreds of genomes have been sequenced. In this case, we have an "open" pan-genome, which means that each new genome sequenced will provide novel genes. However, this "infinite" gene pool is clearly a mathematical extrapolation from the available sequenced genomes; however, it supports the fact that some species have extremely flexible genetic content [5, 27]. However, some species live in an isolated and restricted niche that would hamper the ability to obtain foreign genes by the lack of mechanisms for gene exchange and recombination. In this case, the gene pool is no longer expanding after two or three sequenced genomes; in this way, we can infer that these species have a "closed" pan-genome [28]. However, we must keep in mind that the closed pan-genome does not necessarily denote that all of the strains show the same phenotype because different nucleotide polymorphisms could confer singular features to the strains; for example, some *Buchnera* had its thermal tolerance amended by a single nucleotide mutation in a promoter region [43, 44].

Heap's Law is used to calculate whether the pan-genome is open or closed. Heap's Law is an empirical law that describes the number of distinct words in a document (or set of documents) as a function of the document length, and it is represented by the formula $n=k*N^{-\alpha}$ [45]. In a genetic context, n is the expected number of genes for a given number of genomes, N is the number of genomes, and the k and α ($\alpha = 1-\gamma$) are free parameters that are determined empirically [5].

According to Heap's Law, when $\alpha > 1$ ($\gamma < 0$), the pan-genome is considered to be closed, and the addition of new genomes will not increase the number of new genes significantly. On the other hand, when $\alpha < 1$ ($0 < \gamma < 1$), the pan-genome is open, and for each newly added genome, the number of genes will increase significantly [5].

PAN-GENOME STUDIES

In this section, we describe some pan-genomic studies that were performed with respect to the following species: *Pantoea ananatis* [26], *Lactobacillus rhamnosus* [46], *Corynebacterium pseudotuberculosis* [25], *Corynebacterium diphtheria* [24], and *Buchnera aphidicola* [26] (Table **1**).

Pantoea ananatis belongs to the *Enterobacteriaceae* family, which is frequently found in a wide variety of environments, such as rivers, soil samples, refrigerated beef and aviation fuel tanks, and frequently associated with plants and animals [47, 48]. Computing the pan-genome using eight strains of *P. ananatis* resulted in an open pan-genome in which approximately 106 new protein coding sequences would be added for each new genome. The *P. ananat* pan-genome consists of a core genome with 3,876 protein coding sequences and an accessory genome with 1,690 protein coding sequences [26].

Lactobacillus rhamnosus is a Gram-positive lactic acid bacteria species that covers a range of bodily habitats and is typically associated with certain fermented milk products. Isolates of *L. rhamnosus* are recognized as health-beneficial and are thus used as probiotics [49, 50]. The *L. rhamnosus* pan-genome study focused on the characterization of relevant surface-exposed proteins, such as the *spaCBA* operon, which encodes pili that have a muco-adhesive phenotype, an uncommon occurrence in this species [46].

Corynebacterium pseudotuberculosis is an important animal pathogen causative of several infectious and contagious chronic diseases, such as caseous lymphadenitis (CLA). This disease normally affects small ruminants (sheep and goat), causing significant economic loss [51]. The *C. pseudotuberculosis* pan-genome study resulted in an open pan-genome in which approximately 19 new protein coding sequences were added for each new genome. The core genome consists of 1,504 protein coding sequences. Analysis that was more detailed about the pan-genome revealed differences between the biovar *ovis* and *equi* strains, where the biovar *ovis* showed a more clonal-like behavior than the biovar *equi* strains [25].

Corynebacterium diphtheriae is an important human pathogen and the causative agent of classical diphtheria. This

Table 1. Pan-genome studies.

Organism	No of Genomes	Open/Closed Pan-genome	Pan-genome Size
Pantoea ananatis	8	open	5,566
Lactobacillus rhamnosus	13	open	4,893
Corynebacterium pseudotuberculosis	15	open	2,782
Corynebacterium diphtheriae	13	open	4,786
Buchnera aphidicola	6	closed	2,597

disease is an upper respiratory tract illness that is characterized by sore throat, low-grade fever, and the formation of an adherent membrane on the tonsils, pharynx, and/or nasal cavity [52, 53]. The A-B exotoxin called diphtheria toxin encoded by gene *tox* is the main virulence factor of toxigenic *C. diphtheriae* [54]. A pan-genomic study with thirteen strains showed an open pan-genome with 4,786 coding protein sequences, which was increasing at an average of 65 unique genes per newly sequenced strain. The core genome consists of 1,632 coding protein sequences. Analysis with the gene *tox* revealed that the strain *C. diphtheriae* 31A harbors a hitherto-unknown tox^+ corynephage [24].

Buchnera aphidicola is the obligate intracellular endosymbiont of aphids; they inhabit an isolated and limited niche that would impede the ability to acquire external genes, and in addition, they do not have mechanisms for gene exchange and recombination [16, 27]. Pan-genomic analyses with 4 genomes reveal that this bacteria has a closed pan-genome with an estimated number of approximately 2,600 genes [17].

Comparing the *B. aphidicola* pan-genome with others previously cited (*P. ananatis, L. rhamnosus*, *C. pseudotuberculosis* and *C. diphtheria*), we observed that only *B. aphidicola* has a closed pan-genome. This observation can be correlated with lifestyle because intracellular bacteria have a restricted niche, which could cause gene losses to occur; on the other hand, free-living and facultative intracellular bacteria inhabit several environments, receiving many external stresses. Moreover, free-living and facultative intracellular bacteria normally show a capacity to acquire foreign genes by horizontal transfer [13, 16, 27, 37].

METHODS AND SOFTWARE USED IN PAN-GENOME STUDIES

In this section, we describe some methods and tools that have been developed to calculate the pan-genome. All of these pan-genome software systems are based on orthologous and paralogous gene identification for posterior dataset (core genome, dispensable genome, and strain- or species-specific genome) prediction.

EDGAR (Efficient Database Framework for Comparative Genome Analyses Using BLAST Score Ratios)

EDGAR is a web-tool (available in: https: //edgar.computational.bio.uni-giessen.de/). This software performs homology analyses based on a specific cutoff that is automatically adjusted to the query data [55]. The orthology analysis to calculate pan-genome, core-genome, and singletons is performed using BLAST Score Ratio Values (SRV). This method divides the BLAST bit score by the maximum possible bit score, generating the SRV, and the cutoff is calculated using a sliding window instead of a fixed SRV threshold of 30, as proposed by Lerat *et al.* (2003).

The core genome is predicted through an iterative pairwise comparison using all of the selected genomes. One genome is selected as a reference, and its gene set (A) is compared with another gene set (B). Genes with a reciprocal best hit (the A and B gene sets) are filtered according to an orthology criterion based on the SRVs, and this new gene subset forms the core AB. Subsequently, this subset is compared with another gene set (C), and this comparison continues for all of the genome sets. The pan-genome is predicted in the same way, however, adding non-orthologous genes. One genome forms the pan-genome (A), and non-orthologous genes that are present in the other genome (B) are added to the pan-genome (A), forming the pan-genome (AB). This process continues until all of the genomes have been analyzed. The singletons are predicted using genes that are present in only one genome; in other words, the singletons are predicted using non-orthologous genes that are present in a single genome [55].

PGAT (Prokaryotic Genome Analysis Tool)

PGAT is a web-tool (available in: http://nwrce.org/pgat) that is used to compare multiple strains of the same species, to predict genetic differences. Its analyses include pan-genome, synteny, identification of genes present or absent in a dataset, comparison of SNPs (single-nucleotide polymorphism) in orthologous genes, comparison of genes in metabolic pathways and improvement of functional annotation [56].

The identification of present or absent genes is based on the ortholog assignments. This method is an improvement of the ortholog prediction method, which depends on the annotation that is derived from single genome processing [57]. However, the ortholog assignment removes the bias of the single genome annotation, where the genes are separated into groups and clustered by gene families that are determined through the BLAST protein [58]. Additionally, all of the groups are mapped, using all six-frame translations, and then, the homogenized set of orthologous genes is identified through all of the genomes [56]. The SNP identification is

made using MUSCLE [59], by multiple sequence alignment of orthologous genes. The metabolic pathways are predicted using KEGG [60].

PGAP – Pan-genome Analysis Pipeline

PGAP is a stand-alone tool (available in: http:// pgap.sf.net) developed to perform pan-genome analysis, genetic variation, evolution and function analysis of gene clusters. The software uses two methods to calculate all of the analyses: (i) the GF method to detect homologous genes, and (ii) the MP method to detect orthologous genes.

The GF method is based on the protein BLAST and MCL algorithms. All of the protein sequences are brought together, and protein BLAST is performed; the results are filtered and clustered using the MCL algorithm [58, 61].

The MP method is based on two algorithms: (i) Inparanoid to search orthologous and parologous genes using BLAST. Then, the pairwise ortholog clusters are moved to (ii) MultiParanoid, which was specifically developed to search for gene clusters among multiple strains [58, 62-64].

PanGP: A Tool for Quickly Analyzing Bacterial Pan-genome Profiles

PanGP is a stand-alone tool that was developed to perform pan-genome analysis for large-scale strains with an extremely low time cost. The program works with two algorithms, totally random (TR) and distance guide (DG), which are integrated in the software with a user-friendly graphic interface (available at http://PanGP.big.ac.cn) [65].

The basic difference between the TR and DG algorithm consists of estimating the sample size, where the TR algorithm repeats randomly the samples in non-redundant combinations for all possible combinations, and the DG algorithm has a variable amplification coefficient, which controls the sample size for evaluating the genome diversity of all of the combinations. Tests performed by the authors showed that the DG algorithm has better efficiency [65].

ITEP – Integrated Toolkit for the Exploration of Microbial Pan-genomes

ITEP is a collection of scripts that are written in Python, and BASH is integrated with the SQLite database. This software system is a stand-alone toolkit that is available for download at https://price.systemsbiology.net/itep. The ITEP toolkit was developed to predict protein families, orthologous genes, functional domains, pan-genome (core and variable genes), and metabolic networks for related microbial species [66].

The ITEP workflow consists of a three-step process: **Step 1 – Input data:** ITEP receives three different types of data: Genbank file format, organism file format, and groups file format, and all of the inputs require pre-processing before running the ITPEP toolkit (for more details, see the ITEP documentation); **Step 2 – Building a database (startup scripts):** In this step, scripts are run to predict the gene locations, BLAST results, and clustering results; **Step 3 – Analyses database:** Once the database is ready, the user can start the analyses with the following: core and variable genes, phylogenies, metabolic reconstructions and gene gain and loss patterns [66].

GET_HOMOLOGUES

GET_HOMOLOGUES is a stand-alone and open-source toolkit that was written in Perl and R that can be installed on personal machines. It was developed to perform pan-genome and comparative-genomic analysis of bacterial strains [67].

To build clusters of orthologous groups, the program starts using BLAST+ [58] and HMMER [68]. Then, the sequences, features, and intergenes are extracted, sorted, and indexed. The results are submitted to the bidirectional best hit (BDBH) algorithm, which sorts the genomes by size and takes the smallest as a reference and then identifies paralogous genes that arose by duplication after speciation. Subsequently, new genomes are added and compared with the reference genome, and their BDBHs are annotated; in the last step, clusters that comprise at least one sequence per genome are conserved [67]. Concomitantly, the results are submitted to OrthoMCL [69] version 1.4 and COGtriangles [70].

PanFunPro: PAN-genome Analysis Based on FUNctional PROfiles

PanFunPro is a stand-alone tool for pan-genome analysis using functional domains from HMM (Hidden Markov Models) to group homologous proteins into families based on their functional domain content [71, 72]. In addition to pan-genome analyses, the software performs homology detection and genome annotation using HMM, genome and proteome estimation as well as gene ontology (GO) information [72, 73].

PanFunPro has four steps: **Step 1 – Genome selection:** Submission of the data set can be accomplished using amino acid sequences for all of the encoded proteins. If the data set does not have annotation, then it should first be submitted to Prodigal software [72, 74] for protein prediction; **Step 2 – Prediction of functional domains:** Prediction of functional domains in proteins for a complete data set using PfamA, TIGRFAM, and Superfamily are all integrated into the InterProScan software [75-78]; **Step 3 – Construction of functional profiles and protein groupings:** Here, the software considers HMM hits with an E-value below 0.001 to create functional profiles and protein grouping; **Step 4: Pan, core and accessory genomes analyses:** In the last step, the pan-genome, core genome, and accessory genome are calculated from the GO terms [72].

Panseq – Pan-genome Sequence Analysis Program

Panseq is a freely available web-tool written in BioPerl [79], which is available at http://76.70.11.198/panseq. However, the users can download the BioPerl scripts by contacting the author [80].

In contrast to the other programs described here, Panseq defines the core and accessory genome based on the sequence identity and segmentation length and not on the predicted proteins. For this purpose, the NRF module (Novel Region Finder) was developed. The NRF module first splits the genome sequence into fragments with predefined sizes, and then, the MUMmer alignment program [81] identifies

the sequences and contiguous regions that are present or absent in the database [80]. Next, the CAGF module (Core and Accessory Genome Finder) compares a single sequence file and makes comparisons with all of the other sequences. If this single sequence fits with predefined parameters, then it is added to pan-genome, and then, the newly-added-to fragment sequence is used for subsequent comparisons, and the looping continues until all of the fragment sequences have been tested [80].

CONCLUSION

The amount of pan-genome software has increased since the first time that this term was used by Tettelin and colleagues [16] because the importance of pan-genome studies enables us to identify efficient target genes that can be used in vaccine and drug development through core-genome analyses. Moreover, analyses with genes that belong to the dispensable genome can help us to understand the different symptoms and infections in the hosts, niche adaptations, evolutionary studies development and diagnosis with respect to strains.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Mardis, E.R. Next-Generation DNA Sequencing Methods. *Annu. Rev. Genomics Hum. Genet.* **2008**, *9*, 387.

[2] Metzker, M. L. Emerging Technologies in DNA Sequencing. *Genome Res.* **2005**, *15*, 1767.

[3] Hert, D.G.; Fredlake, C.P.; Barron, A. E. Advantages and Limitations of next-Generation Sequencing Technologies: A Comparison of Electrophoresis and Non-Electrophoresis Methods. *Electrophoresis,* **2008**, *29*, 4618.

[4] Metzker, M.L. Sequencing Technologies [mdash] the next Generation. *Nat. Rev. Genet.,* **2010**, *11*, 31.

[5] Tettelin, H.; Riley, D.; Cattuto, C.; Medini, D. Comparative Genomics: The Bacterial Pan-Genome. *Curr. Opin. Microbiol.,* **2008**, *11*, 472.

[6] Sivashankari, S.; Shanmughavel, P. Comparative Genomics - A Perspective. *Bioinformation,* **2007**, *1*, 376.

[7] Ogier, J.-C.; Calteau, A.; Forst, S.; Goodrich-Blair, H.; Roche, D.; Rouy, Z.; Suen, G.; Zumbihl, R.; Givaudan, A.; Tailliez, P. Units of Plasticity in Bacterial Genomes: New Insight from the Comparative Genomics of Two Bacteria Interacting with Invertebrates, Photorhabdus and Xenorhabdus. *BMC Genomics,* **2010**, *11*, 568.

[8] Rust, A. G.; Mongin, E.; Birney, E. Genome Annotation Techniques: New Approaches and Challenges. *Drug Discov. Today,* **2002**, *7*, S70.

[9] Abby, S.; Daubin, V. Comparative Genomics and the Evolution of Prokaryotes. *Trends Microbiol.,* **2007**, *15*, 135.

[10] Reiter, L. T.; Potocki, L.; Chien, S.; Gribskov, M.; Bier, E. A Systematic Analysis of Human Disease-Associated Gene Sequences in Drosophila Melanogaster. *Genome Res.,* **2001**, *11*, 1114.

[11] Rey, M.W.; Ramaiya, P.; Nelson, B.A; Brody-Karpin, S.D.; Zaretsky, E. J.; Tang, M.; Lopez de Leon, A.; Xiang, H.; Gusti, V.; Clausen, I.G. Complete Genome Sequence of the Industrial Bacterium Bacillus Licheniformis and Comparisons with Closely Related Bacillus Species. *Genome Biol.,* **2004**, *5*, R77.

[12] Pallen, M. J.; Wren, B. W. Bacterial Pathogenomics. *Nature,* **2007**, *449*, 835.

[13] Merhej, V.; Royer-Carenzi, M.; Pontarotti, P.; Raoult, D. Massive Comparative Genomic Analysis Reveals Convergent Evolution of Specialized Bacteria. *Biol. Direct.,* **2009**, *4*, 13.

[14] Lapierre, P.; Gogarten, J. P. Estimating the Size of the Bacterial Pan-Genome. *Trends Genet.,* **2009**, *25*, 107.

[15] Muzzi, A.; Masignani, V.; Rappuoli, R. The Pan-Genome: Towards a Knowledge-Based Discovery of Novel Targets for Vaccines and Antibacterials. *Drug Discov. Today,* **2007**, *12*, 429.

[16] Tettelin, H.; Masignani, V.; Cieslewicz, M. J.; Donati, C.; Medini, D.; Ward, N. L.; Angiuoli, S. V; Crabtree, J.; Jones, A. L.; Durkin, A.S. Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial "Pan-Genome". *Proc. Natl. Acad. Sci. U. S. A.,* **2005**, *102*, 13950.

[17] Snipen, L.; Almøy, T.; Ussery, D. W. Microbial Comparative Pan-Genomics Using Binomial Mixture Models. *BMC Genomics,* **2009**, *10*, 385.

[18] Alcaraz, L. D.; Moreno-Hagelsieb, G.; Eguiarte, L. E.; Souza, V.; Herrera-Estrella, L.; Olmedo, G. Understanding the Evolutionary Relationships and Major Traits of Bacillus through Comparative Genomics. *BMC Genomics,* **2010**, *11*, 332.

[19] Rasko, D. a; Altherr, M. R.; Han, C. S.; Ravel, J. Genomics of the Bacillus Cereus Group of Organisms. *FEMS Microbiol. Rev.,* **2005**, *29*, 303.

[20] Rasko, D.A; Rosovitz, M. J.; Myers, G. S. a; Mongodin, E. F.; Fricke, W. F.; Gajer, P.; Crabtree, J.; Sebaihia, M.; Thomson, N. R.; Chaudhuri, R.. The Pangenome Structure of Escherichia Coli: Comparative Genomic Analysis of E. Coli Commensal and Pathogenic Isolates. *J. Bacteriol.,* **2008**, *190*, 6881.

[21] Reno, M. L.; Held, N. L.; Fields, C. J.; Burke, P. V; Whitaker, R. J. Biogeography of the Sulfolobus Islandicus Pan-Genome. *Proc. Natl. Acad. Sci. U. S. A.,* **2009**, *106*, 8605.

[22] Donati, C.; Hiller, N. L.; Tettelin, H.; Muzzi, A.; Croucher, N. J.; Angiuoli, S. V; Oggioni, M.; Dunning Hotopp, J. C.; Hu, F. Z.; Riley, D. R. Structure and Dynamics of the Pan-Genome of Streptococcus Pneumoniae and Closely Related Species. *Genome Biol.,* **2010**, *11*, R107.

[23] Hansen, E. E.; Lozupone, C. a; Rey, F. E.; Wu, M.; Guruge, J. L.; Narra, A.; Goodfellow, J.; Zaneveld, J. R.; McDonald, D. T.; Goodrich, J.A. Pan-Genome of the Dominant Human Gut-Associated Archaeon, Methanobrevibacter Smithii, Studied in Twins. *Proc. Natl. Acad. Sci. U. S. A.,* **2011**, *108 Suppl* , 4599.

[24] Trost, E.; Blom, J.; Soares, S. D. C.; Huang, I.-H.; Al-Dilaimi, A.; Schröder, J.; Jaenicke, S.; Dorella, F. A; Rocha, F. S.; Miyoshi, A. Pangenomic Study of Corynebacterium Diphtheriae That Provides Insights into the Genomic Diversity of Pathogenic Isolates from Cases of Classical Diphtheria, Endocarditis, and Pneumonia. *J. Bacteriol.,* **2012**, *194*, 3199.

[25] Soares, S. C.; Silva, A.; Trost, E.; Blom, J.; Ramos, R.; Carneiro, A.; Ali, A.; Santos, A. R.; Pinto, A. C.; Diniz, C. The Pan-Genome of the Animal Pathogen Corynebacterium Pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar Ovis and Equi Strains. *PLoS One,* **2013**, *8*, e53818.

[26] De Maayer, P.; Chan, W.Y.; Rubagotti, E.; Venter, S.N.; Toth, I.K.; Birch, P. R. J.; Coutinho, T.A. Analysis of the Pantoea Ananatis Pan-Genome Reveals Factors Underlying Its Ability to Colonize and Interact with Plant, Insect and Vertebrate Hosts. *BMC Genomics,* **2014**, *15*, 404.

[27] Mira, A.; Martín-Cuadrado, A.B.; D'Auria, G.; Rodríguez-Valera, F. The Bacterial Pan-Genome:a New Paradigm in Microbiology. *Int. Microbiol.,* **2010**, *13*, 45-57.

[28] Medini, D.; Donati, C.; Tettelin, H.; Masignani, V.; Rappuoli, R. The Microbial Pan-Genome. *Curr. Opin. Genet. Dev.,* **2005**, *15*, 589.

[29] Lawrence, J. G.; Hendrickson, H. Genome Evolution in Bacteria: Order beneath Chaos. *Curr. Opin. Microbiol.,* **2005**, *8*, 572.

[30] Lerat, E.; Daubin, V.; Moran, N.A. From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the Gamma-Proteobacteria. *PLoS Biol.,* **2003**, *1*, E19.

[31] Read, T. D.; Ussery, D. W. Editorial Overview : *Opening the Pan-Genomics Box*, **2006**, *9*, 1.

[32] Croll, D.; McDonald, B. a. The Accessory Genome as a Cradle for Adaptive Evolution in Pathogens. *PLoS Pathog.,* **2012**, *8*, e1002608.

[33] Grim, C.J.; Kotewicz, M.L.; Power, K.A; Gopinath, G.; Franco, A. a; Jarvis, K. G.; Yan, Q. Q.; Jackson, S. a; Sathyamoorthy, V.; Hu, L. Pan-Genome Analysis of the Emerging Foodborne Pathogen

Cronobacter Spp. Suggests a Species-Level Bidirectional Divergence Driven by Niche Adaptation. *BMC Genomics,* **2013**, *14*, 366.

[34] Lefébure, T.; Stanhope, M. J. Evolution of the Core and Pan-Genome of Streptococcus: Positive Selection, Recombination, and Genome Composition. *Genome Biol.,* **2007**, *8*, R71.

[35] Zhang, R.; Zhang, C.-T. Identification of Genomic Islands in the Genome of Bacillus Cereus by Comparative Analysis with Bacillus Anthracis. *Physiol. Genomics.,* **2003**, *16*, 19.

[36] Jordan, I. K.; Makarova, K. S.; Spouge, J. L.; Wolf, Y. I.; Koonin, E. V. Lineage-Specific Gene Expansions in Bacterial and Archaeal Genomes. *Genome Res.,* **2001**, *11*, 555.

[37] Soares, S. C.; Abreu, V. a C.; Ramos, R. T. J.; Cerdeira, L.; Silva, A.; Baumbach, J.; Trost, E.; Tauch, A.; Hirata, R.; Mattos-Guaraldi, A. L. PIPS: Pathogenicity Island Prediction Software. *PLoS One,* **2012**, *7*, e30848.

[38] Penn, K.; Jenkins, C.; Nett, M.; Udwary, D. W.; Gontang, E. A.; Mcglinchey, P.; Foster, B.; Lapidus, A.; Podell, S.; Allen, E. E. *Adaptation in Marine Actinobacteria*. **2010**, *3*, 1193.

[39] Lawrence, J. G.; Ochman, H. Amelioration of Bacterial Genomes: Rates of Change and Exchange. *J. Mol. Evol.,* **1997**, *44*, 383.

[40] Daubin, V.; Ochman, H. Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in E. Coli. *Genome Res.,* **2004**, *14*, 1036.

[41] Alm, R. a; Bina, J.; Andrews, B. M.; Doig, P.; Hancock, R. E.; Trust, T. J. Comparative Genomics of Helicobacter Pylori: Analysis of the Outer Membrane Protein Families. *Infect. Immun.,* **2000**, *68*, 4155.

[42] Forrellad, M.A; Klepp, L.I.; Gioffré, A.; Sabio y García, J.; Morbidoni, H. R.; de la Paz Santangelo, M.; Cataldi, A. a; Bigi, F. Virulence Factors of the Mycobacterium Tuberculosis Complex. *Virulence,* **2013**, *4*, 3.

[43] Mizuki, T.; Kamekura, M.; DasSarma, S.; Fukushima, T.; Usami, R.; Yoshida, Y.; Horikoshi, K. Ureases of Extreme Halophiles of the Genus Haloarcula with a Unique Structure of Gene Cluster. *Biosci. Biotechnol. Biochem.,* **2004**, *68*, 397.

[44] Dunbar, H. E.; Wilson, A. C. C.; Ferguson, N. R.; Moran, N. a. Aphid Thermal Tolerance Is Governed by a Point Mutation in Bacterial Symbionts. *PLoS Biol.,* **2007**, *5*, 1006.

[45] Egghe, L. Untangling Herdan ' S Law and Heaps ' Law : Mathematical and Informetric Arguments. *J. Am. Soc. Inform. Sci. Tech*., **2007**, *58*, 702-709.

[46] Kant, R.; Rintahaka, J.; Yu, X.; Sigvart-Mattila, P.; Paulin, L.; Mecklin, J.-P.; Saarela, M.; Palva, A.; von Ossowski, I. A Comparative Pan-Genome Perspective of Niche-Adaptable Cell-Surface Protein Phenotypes in Lactobacillus Rhamnosus. *PLoS One,* **2014**, *9*, e102762.

[47] Ercolini, D.; Russo, F.; Torrieri, E.; Masi, P.; Villani, F. Changes in the Spoilage-Related Microbiota of Beef during Refrigerated Storage under Different Packaging Conditions. *Appl. Environ. Microbiol.,* **2006**, *72*, 4663.

[48] Coutinho, T. a; Venter, S. N. Pantoea Ananatis: An Unconventional Plant Pathogen. *Mol. Plant Pathol.,* **2009**, *10*, 325.

[49] Martín, R.; Heilig, G.H.J.; Zoetendal, E. G.; Smidt, H.; Rodríguez, J.M. Diversity of the Lactobacillus Group in Breast Milk and Vagina of Healthy Women and Potential Role in the Colonization of the Infant Gut. *J. Appl. Microbiol.,* **2007**, *103*, 2638.

[50] Kankainen, M.; Paulin, L.; Tynkkynen, S.; von Ossowski, I.; Reunanen, J.; Partanen, P.; Satokari, R.; Vesterlund, S.; Hendrickx, A. P. a; Lebeer, S. Comparative Genomic Analysis of Lactobacillus Rhamnosus GG Reveals Pili Containing a Human- Mucus Binding Protein. *Proc. Natl. Acad. Sci. U. S. A.,* **2009**, *106*, 17193.

[51] Dorella, F. A.; Pacheco, L. G. C.; Oliveira, S. C.; Miyoshi, A.; Azevedo, V. Corynebacterium Pseudotuberculosis : Microbiology , Biochemical Properties , Pathogenesis and Molecular Studies of Virulence.*Vet. Res.*, **2006**, *37*, 201.

[52] Hadfield, T. L.; Mcevoy, P.; Polotsky, Y.; Tzinserling, V. A.; Yakovlev, A. A. The Pathology of Diphtheria. *J. Infect. Dis*., **2000**, *20306*, 3.

[53] Gomes, D. L. R.; Martins, C. a S.; Faria, L. M. D.; Santos, L. S.; Santos, C. S.; Sabbadini, P. S.; Souza, M. C.; Alves, G. B.; Rosa, A. C. P.; Nagao, P.E. Corynebacterium Diphtheriae as an Emerging Pathogen in Nephrostomy Catheter-Related Infection: Evaluation of Traits Associated with Bacterial Virulence. *J. Med. Microbiol.,* **2009**, *58*, 1419.

[54] Holmes, R. Biology and Molecular Epidemiology of Diphtheria Toxin and the Tox Gene. *J. Infect. Dis.,* **2000**, 156, 1.

[55] Blom, J.; Albaum, S. P.; Doppmeier, D.; Pühler, A.; Vorhölter, F.-J.; Zakrzewski, M.; Goesmann, A. EDGAR: A Software Framework for the Comparative Analysis of Prokaryotic Genomes. *BMC Bioinformatics,* **2009**, *10*, 154.

[56] Brittnacher, M.J.; Fong, C.; Hayden, H. S.; Jacobs, M. a; Radey, M.; Rohmer, L. PGAT: A Multistrain Analysis Resource for Microbial Genomes. *Bioinformatics,* **2011**, *27*, 2429.

[57] Salichos, L.; Rokas, A. Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade. *PLoS One,* **2011**, *6*, e18755.

[58] Altschup, S. F.; Gish, W.; Pennsylvania, T.; Park, U. Basic Local Alignment Search Tool. *J. Mol. Biol.,* **1990**, *215*, 403.

[59] Edgar, R. C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.,* **2004**, *32*, 1792.

[60] Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.,* **2000**, *28*, 27.

[61] Enright, a J.; Van Dongen, S.; Ouzounis, C.A. An Efficient Algorithm for Large-Scale Detection of Protein Families. *Nucleic Acids Res.* **2002**, *30*, 1575.

[62] Ostlund, G.; Schmitt, T.; Forslund, K.; Köstler, T.; Messina, D. N.; Roopra, S.; Frings, O.; Sonnhammer, E. L. L. InParanoid 7: New Algorithms and Tools for Eukaryotic Orthology Analysis. *Nucleic Acids Res.* **2010**, *38*, D196.

[63] Alexeyenko, A.; Tamas, I.; Liu, G.; Sonnhammer, E. L. L. Automatic Clustering of Orthologs and Inparalogs Shared by Multiple Proteomes. *Bioinformatics,* **2006**, *22*, e9.

[64] Remm, M.; Storm, C. E.; Sonnhammer, E. L. Automatic Clustering of Orthologs and in-Paralogs from Pairwise Species Comparisons. *J. Mol. Biol.,* **2001**, *314*, 1041.

[65] Zhao, Y.; Jia, X.; Yang, J.; Ling, Y.; Zhang, Z.; Yu, J.; Wu, J.; Xiao, J. PanGP: A Tool for Quickly Analyzing Bacterial Pan-Genome Profile. *Bioinformatics,* **2014**, *30*, 1297.

[66] Benedict, M. N.; Henriksen, J. R.; Metcalf, W. W.; Whitaker, R. J.; Price, N. D. ITEP: An Integrated Toolkit for Exploration of Microbial Pan-Genomes. *BMC Genomics,* **2014**, *15*, 8.

[67] Contreras-Moreira, B.; Vinuesa, P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Appl. Environ. Microbiol.,* **2013**, *79*, 7696.

[68] Finn, R. D.; Clements, J.; Eddy, S. R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.,* **2011**, *39*, W29.

[69] Li, L.; Stoeckert, C. J.; Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.,* **2003**, *13*, 2178.

[70] Kristensen, D. M.; Kannan, L.; Coleman, M. K.; Wolf, Y. I.; Sorokin, A.; Koonin, E. V; Mushegian, A. A Low-Polynomial Algorithm for Assembling Clusters of Orthologous Groups from Intergenomic Symmetric Best Matches. *Bioinformatics,* **2010**, *26*, 1481.

[71] Eddy, S. R. Hidden Markov Models. *Curr. Opin. Struct. Biol.,* **1996**, *6*, 361.

[72] Lukjancenko, O.; Thomsen, M. C.; Voldby Larsen, M.; Ussery, D. W. PanFunPro: PAN-Genome Analysis Based on FUNctional PROfiles. *F1000Research,* **2013**, *265*, 1.

[73] Gene, T.; Consortium, O. The Gene Ontology Project in 2008. *Nucleic Acids Res.,* **2008**, *36*, D440.

[74] Hyatt, D.; Chen, G.-L.; Locascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics,* **2010**, *11*, 119.

[75] Finn, R. D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J. Pfam: The Protein Families Database. *Nucleic Acids Res.,* **2014**, *42*, D222.

[76] Haft, D. H. The TIGRFAMs Database of Protein Families. *Nucleic Acids Res.,* **2003**, *31*, 371.

[77] Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: Protein Domains Identifier. *Nucleic Acids Res.,* **2005**, *33*, W116.

[78] Wilson, D.; Pethica, R.; Zhou, Y.; Talbot, C.; Vogel, C.; Madera, M.; Chothia, C.; Gough, J. SUPERFAMILY--Sophisticated Comparative Genomics, Data Mining, Visualization and Phylogeny. *Nucleic Acids Res.,* **2009**, *37*, D380.

[79] Stajich, J. E.; Block, D.; Boulez, K.; Brenner, S. E.; Chervitz, S. a; Dagdigian, C.; Fuellen, G.; Gilbert, J. G. R.; Korf, I.; Lapp, H. The

Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Res.,* **2002**, *12*, 1611.

[80] Laing, C.; Buchanan, C.; Taboada, E. N.; Zhang, Y.; Kropinski, A.; Villegas, A.; Thomas, J. E.; Gannon, V. P. J. Pan-Genome Sequence Analysis Using Panseq: An Online Tool for the Rapid Analysis of Core and Accessory Genomic Regions. *BMC Bioinformatics,* **2010**, *11*, 461.

[81] Kurtz, S.; Phillippy, A.; Delcher, A. L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S. L. Versatile and Open Software for Comparing Large Genomes. *Genome Biol.,* **2004**, *5*, R12.

Received on: January 26, 2015 **Revised on: April 20, 2015** **Accepted on: April 21, 2015**