Genomic Islands: an overview of current software tools and future improvements

Siomar de Castro Soares[1,2*], Letícia de Castro Oliveira[2], Arun Kumar Jaiswal[2], Vasco Azevedo[2]

[1]Department of Microbiology, Immunology and Parasitology, Institute of Biological and Natural Sciences, Federal University of Triângulo Mineiro, Uberaba - MG, Brazil

[2] Laboratory of Cellular and Molecular Genetics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte - MG, Brazil

Summary

Microbes are highly diverse and widely distributed organisms. They account for ~60% of Earth's biomass and new predictions point for the existence of 10^{11} to 10^{12} species, which are constantly sharing genes through several different mechanisms. Genomic Islands (GI) are critical in this context, as they are large regions acquired through horizontal gene transfer. Also, they present common features like genomic signature deviation, transposase genes, flanking tRNAs and insertion sequences. GIs carry large numbers of genes related to specific lifestyle and are commonly classified in Pathogenicity, Resistance, Metabolic or Symbiotic Islands. With the advent of the next-generation sequencing technologies and the deluge of genomic data, many software tools have been developed that aim to tackle the problem of GI prediction and they are all based on the prediction of GI common features. However, there is still room for the development of new software tools that implements new approaches, such as, machine learning and pan-genomics based analyses. Finally, GIs will always hold a potential application in every newly invented genomic approach as they are directly responsible for much of the genomic plasticity of bacteria.

1 Living in the "Age of Bacteria"

Stephen Jay Gould, a renowned paleontologist, once said, "We live now in the 'Age of Bacteria.' Our planet has always been in the 'Age of Bacteria' ever since the first fossils, bacteria, of course, were entombed in rocks more than three and a half billion years ago" [1]. Microbes are highly diverse organisms responsible for approximately 60% of the Earth's biomass. They were the first organisms on Earth, they are distributed worldwide, from volcanos to salt water, and they play a pivotal role in several medical, biotechnological and industrial applications. Although their importance is widely known, less than 1% of the previously estimated 2-3 billion microbial species are identified so far [2]. Much of this lack of knowledge on microbes is due to the use of culture-dependent identification and characterization of microbes. Microbiological culture media are usually intended for selective growing and, thus, the microorganisms recovered using these methods are not representative of the microbial community inside the sample [3]. However, with the advent of the next-generation sequencing (NGS) technologies and the widespread of metagenomics methodologies, scientists are now able to determine the complete gene set off an entire community, transcending the idea of a single species genomics to a complete view of the

[*] To whom correspondance should be addressed. Email: siomars@gmail.com

doi:10.2390/biecoll-jib-2016-301

microbial population dynamics under a given environmental time and condition. By using estimates generated from global microbiota, for instance, one may now predict the total number of microbial species on Earth to be much larger than the previously estimated 2-3 billion species, ranging from 10^{11} to 10^{12} species [4].

Although we can predict the putative number of species, we are still very young for the identification of protein functions from non-cultivable organisms. Even cultivable well-studied microbes such as *Escherichia coli* present more than 35% of hypothetical proteins in their genomes, i.e., predicted genes with no assigned function due to the lack of experimental data. Much of those genes are typically located in regions acquired through horizontal gene transfer (HGT). These areas present low similarities with the genome where they are harboured in and may have originated from non-cultivable organisms, therefore, explaining the lack of information about their function [5]. More interestingly, those regions may transfer between all domains (Bacteria, Archaea, and Eukarya), in all possible directions, adding to the pool of genes that will be driven, by selection, to entirely new functions [6].

2 Horizontal Gene Transfer

HGT events may occur through diverse mechanisms, including plasmids, transposons and non-canonical classes of Mobile Genetic Elements (MGEs) [7,8]. The success in the spread of a given MGE depends on its arsenal of coding genes and how they affect the behavior of the acceptor organism in influencing the host cell or even the neighboring cells. For instance, MGEs harboring genes coding for an advantageous characteristic in a given environment are more prone to be fixed in the population and to spread to other organisms. Adaptive traits carried by MGEs may include virulence factors, antibiotic resistance, detoxifying agents and metabolic- and symbiotic-related genes [9].

MGEs carrying adaptive traits are usually classified as Genomic Islands (GI) and sub-classified in Pathogenicity Islands (PAI), Resistance Islands (RI), Metabolic Islands (MI) and Symbiotic Islands (SI). The term PAI was coined by Hacker and colleagues when they identified and experimentally validated the instability of the major genomic regions harboring hemolysin and fimbrial adhesin genes in the genome of *E. coli* [10]. Since then, the terms RI, MI and SI were created to accommodate other classes of GIs according to their effect on the fitness of the acceptor organism. In summary, GIs are characterized for being large genomic regions acquired through horizontal gene transfer, which presents anomalous G+C content and/or codon usage deviation, as they reflect the genomic signature of the DNA donor organism. Also, they may harbor transposases and tRNA flanking genes, which are important during the DNA insertion into the acceptor genome. Moreover, they are unstable, present mosaic structure and are usually absent from other closely related organisms [11,12]. Finally, the only feature differentiating the classes of GIs is the gene composition; PAIs, RIs, MIs and SIs are characterized by the prevalence of virulence factors and resistance-, metabolic- and symbiotic- related genes, respectively[13].

3 Prediction of GIs

3.1 Data quality

An important variable to be considered during the prediction of GIs is the quality of the genome sequence. With the advent of the NGS technologies and the generation of smaller sequencing reads as compared to the previous Sanger methodology, there was a huge increase in the total number of genome sequences and also draft genomes. Although draft genomes may be used for the prediction of GIs, the comparison between draft genomes in these analyses may take to false-positive or false negative results, due to the absence of regions in the query or reference genome caused by unresolved gaps [13]. Therefore, the prediction of GIs should be only performed using complete genome sequences. To circumvent this, researchers may take advantage of combined sequencing approaches, using PacBio or MinIon along with Illumina or Ion Torrent platforms [14]. In this scenario, the long-read sequencing technologies PacBio and MinIon would be helpful in the assembly of complete genomes, whereas Illumina and Ion Torrent would result in the base quality needed to achieve a good quality sequence.

Also noticeable, the high frequency of nucleotide substitutions and insertion/deletions by Ilumina and Ion Torrent platforms, respectively, may take to non-synonimous substitutions and pseudogeneization of genes, which will impact the codon usage, the G+C content and also the gene composition [15]. Thus, a high genome coverage coupled with a manual curation of the sequence using genome mapping visualization software tools is also desirable. Finally, the gene composition is also important in the prediction of GIs and, also, in the post prediction analyses to find biological correlations. Thus, it is also recommendable to perform manual curation of the whole genome annotation to avoid poor quality annotation.

3.2 Software tools

The first identification of a PAI was achieved using molecular biology approaches; however, this strategy is time and money consuming [10]. Nowadays, with the advent of next-generation sequencing technologies, some software tools have been developed to tackle the problem of GI identification from the genome sequence. The existent software tools mainly focus on the commonly shared GI features for the prediction, like identification of genomic regions with G+C and/or codon usage anomalies compared to the whole genome sequence (Table 1). However, because GIs present genes that are relevant to the bacterial fitness, the selective pressure will ultimately select mutations that adapt the codon usage of the gene to the one of the acceptor genome, increasing the translation efficiency. Also, the preference for GC-rich or AT-rich codons may also drive the G+C content of the genes in the genomic region, taking the whole region to have a more homogeneous G+C content overtime [16]. Therefore, software tools that predict GIs using only the genomic signature information (e.g., GI-SVM, IGIPT, PAI-IDA and SIGI-HMM) may fail in predicting GIs that were not acquired recently (Table 1). Alternatively, the use of other GI features, like the presence of flanking tRNAs, mobility genes, insertion sequences and specific factors may be helpful in identifying GIs with homogeneous genomic signature (e.g., EGID, Islander and Islandpath) (Table 1). However, the genomic comparison showing the absence of the region in a closely-related organism is one of the most important features, as previously reported [17]. Indeed, the more features the software tool uses to predict GIs, the more efficient it is in tackling the problem, highlighting the importance of using genomic signature, the comparative genomics analyses, and other additional features to achieve a better result (e.g., GIHunter, GI-POP, GIPSy, GIST, INDeGenIUS, IslandViewer, PAIDB, PIPS and RPGFinder) (Table 1). This scenario explains the appearing of ensemble software tools, which combine different software tools to achieve the goal of providing the user with a comprehensive analysis of all GI features (e.g., EGID, GIPSy, GIST, IslandViewer and PIPS) (Table 1).

Until recently, there was little information about the sub-classes of GIs others than PAIs. A quick search for RIs, MIs, and SIs in PubMed does not return genomic coordinates of these GIs. The specific prediction of these subclasses of GIs was partially addressed in the software InDeGenIUS, IslandViewer, and PAIDB (Table 1). However, the first software tool to be completely developed for the specific prediction of all 4 classes of GIs, individually, was only published recently [13]. Thus, there is still a huge urge for the widespread of information on other GIs.

Table 1: GI prediction tools and their methodologies.

Tool	Software tool/ database	Genomic signature	tRNA genes	Mobility genes	Comparative genomics/ clustering	Insertion sequences	Specific factors	Refe- rences
AlienHunter	SW	ON	-	-	+	-	-	[18]
EGID	ES	GC+DI+TRI+ON+CU	+	-	-	-	-	[19]
GC-Profile	SW	GC	-	-	-	-	-	[20]
GEMINI*	SW	-	-	-	+	-	-	[21]
GIHunter	SW	-	+	+	+	-	-	[22]
GI-POP	SW	GC+ON+CU	+	-	+	+	-	[23]
GIPSy	ES	GC+CU	+	+	+	-	VF+RF+MF+SF	[13]
GIST	ES	GC+DI+ON+CU	+	-	+	-	-	[24]
GI-SVM	SW	GC+CU	-	-	-	-	-	[25,26]
HGTector	SW	-	-	-	+	-	-	[27]
IGIPT	SW	GC+DI+CU	-	-	-	-	-	[28]
INDeGenIUS	SW	ON	-	-	+	-	VF+RF+MF+SF	[29]
Islander	DB	GC	+	+	-	+	-	[30]
Islandpath	DB	GC+DI	+	+	-	-	-	[31]
IslandPick	SW	-	-	-	+	-	-	[32]
IslandViewer 3	DB+ES	GC+DI+ CU	+	+	+	-	VF+RF	[33]
MSGIP	SW	GC	-	-	+	-	-	[34]
PAIDB	DB	GC+CU	+	-	+	-	VF	[35]
PAIDB v2.0:	DB	GC+DI+CU	+	+	+	-	VF+RF	[36]
PAI-IDA	SW	GC+DI+CU	-	-	-	-	-	[37]
PIPS	ES	GC+CU	+	+	+	-	VF	[17]
Pre_GI	DB	GC+ON	-	-	+	-	-	[38]
RGPFinder	SW	GC+CU+ON	+	+	+	+	-	[39]
SIGI-HMM	SW	CU	-	-	-	-	-	[40]
Zisland Explorer	SW	GC+CU	-	-	+	-	-	[41]

DB, database; SW, software tool; ES, ensemble software that combines different software tools; GC, G+C content; DI, dinucleotide frequency; TRI, trinucleotide frequency; ON, oligonucleotide; CU, codon usage *Gemini uses a genome segmentation and clustering approach

4 Future improvements

Future improvements in the area may involve the use of machine learning approaches for GI classification based on the concentration of all features in a genomic region, i.e., the concentration of genes with G+C content variation, codon usage deviation, transposase genes and so on [42]. GIs are mosaic regions in nature and each GI may or may not present a combination of different features [17,41]. For instance, a GI may have a G+C content deviation, harbors transposase genes and be flanked by tRNAs, while another one may only harbor a large number of virulence factors and present codon usage deviation. This mosaic structure may take to false-negative results even in ensemble methodologies that uses different software tools to cover all features. Hence, the implementation of machine learning approaches may be helpful in detecting all the possible scenarios during the classification of GIs using different features [42,43].

Also, one task that needs addressing is the prediction of the origin of the GIs [38,44]. Because GIs adapt their genomic signature with time, it is not always possible to predict their origin by comparing them with the genomic signature of other organisms [16]. Besides, two distantly related organisms may have the same codon usage, due to tRNA bioavailability [45]. Alternatively, the phylogenetic comparison of syntenic genes inside the GI with orthologous genes in other organisms could be the key to predicting their putative origin and also for the prediction of MGE data pools in bacterial populations from the comparison of GIs with available metagenomics data.

Another area that is constantly taking advantage from GI analyses nowadays is pan-genomics. The area was created by Tettelin *et al.* (2005) and consists in the identification of similarities and differences between a set of strains from the same species or a set of species from the same genus [46]. The term pan-genome is also used to define the non-redundant set of genes in the complete analyses. The approach normally makes use of the orthology prediction between all genes from all genomes in the dataset. Then, the approach identifies which genes are: commonly shared between all strains (core genome); shared between 2 or more strains, but not all ("shared genome"); and, unique to a single strain (singletons). The commonly shared genes in the core genome are important for vaccine and drug development. The genes in the shared genome and the singletons are normally responsible for differential adaptation to new environments and, hence, genes in GIs normally account for this dataset [47]. Future strategies in pan-genomics allied to GI analyses could aim firstly at identifying GIs in all strains and comparing the identified GIs to measure their degree of mosaicism. After, epidemiological analyses may be performed using phylogenomics-based approaches on those GIs throughout the strains. Then, the final step may include the identification of the origin of the GIs from gene synteny conservation between distantly related species.

The identification of the origin of the GIs allied with pan-genomics analyses may reveal the acquirement of blocks of genes influencing the adaptability of bacteria to new traits and hosts, which may be correlated to specific traits of the donor organism. Overall, this combined strategy may be helpful in tracing the origin of new clonal complexes, in epidemiological analyses, and also in the creation of new diagnostic methods for emerging pathogenic strains [48,49]. Finally, because GIs account for much of the genomic variability in bacterial species, for every new field created in comparative genomics there is a hidden potential for the creation of new GI comparison analyses.

5 Acknowledgements

Arun Kumar Jaiswal and Letícia de Castro Oliveira are grateful for financial support from CAPES and CNPq. This work was funded by CAPES, CNPq, and Fapemig.

References

[1] S. J. Gould. Full House: The spread of Excellence from Plato to Darwin. *New York: Harmony Books*, 1996.

[2] C. M. Fraser, J. A. Eisen and S. L. Salzberg. Microbial genome sequencing. *Nature*. 406:799-803, 2000.

[3] E. F. DeLong and N. R. Pace. Environmental diversity of bacteria and archaea. *Syst Biol*. 5(4):470-478, 2001.

[4] K. J. Locey and J. T. Lennon. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A*. 113(21):5970-5975, 2016.

[5] W. W. Hsiao, K. Ung, D. Aeschliman, J. Bryan, B. B. Finlay and F. S. Brinkman. Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet*. 1(5):e62, 2005.

[6] L. Boto. Horizontal gene transfer in evolution: facts and challenges. *Proc Biol Sci*. 277(1683):819-827, 2010.

[7] X. Bellanger, S. Payot, N. Leblond-Bourget and G. Guédon. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol Rev*. 38(4):720-760, 2014.

[8] U. Dobrindt and J. Hacker. Whole genome plasticity in pathogenic bacteria. *Curr Opin Microbiol*. 4:550-557, 2001.

[9] D. J. Rankin, E. P. Rocha and S. P. Brown. What traits are carried on mobile genetic elements, and why?. *Heredity (Edinb)*. 106(1):1-10, 2011.

[10] J. Hacker, L. Bender, M. Ott, J. Wingender, B. Lund, R. Marre and W. Goebel. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. *Microb Pathog*. 8(3):213-225, 1990.

[11] M. Juhas, J. R. van der Meer, M. Gaillard, R. M. Harding, D. W. Hood and D. W. Crook. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev*. 33(2):376-393, 2009.

[12] H. Schmidt and M. Hensel. Pathogenicity Islands in Bacterial Pathogenesis. *Clin Microbiol Rev*. 17(1):14-56, 2004.

[13] S. C. Soares, H. Geyik, R. T. Ramos, P. H. de Sá, E. G. Barbosa, J. Baumbach, H. C. Figueiredo, A. Miyoshi, A. Tauch, A. Silva et al.. GIPSy: Genomic island prediction software. *J Biotechnol*. 232:2-11, 2016.

[14] S. Goodwin, J. D. McPherson and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 17(6):333-351, 2016.

[15] F. L. Pereira, S. C. Soares, F. A. Dorella, C. A. Leal and H. C. Figueiredo. Evaluating the efficacy of the new Ion PGM Hi-Q Sequencing Kit applied to bacterial genomes. *Genomics*. 107(5):189-198, 2016.

[16] R. Hershberg and D. A. Petrov. General rules for optimal codon choice. *PLoS Genet*. 5(7):e1000556, 2009.

[17] S. C. Soares, V. A. Abreu, R. T. Ramos, L. Cerdeira, A. Silva, J. Baumbach, E. Trost, A. Tauch, R. Hirata Jr, A. L. Mattos-Guaraldi et al.. PIPS: pathogenicity island prediction software. *PLoS One*. 7(2):e30848, 2012.

[18] G. S. Vernikos and J. Pakhill. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics*. 22(18):2196-2203, 2006.

[19] D. Che, M. S. Hasan, H. Wang, J. Fazekas, J. Huang and Q. Liu. EGID: an ensemble algorithm for improved genomic island detection in genomic sequences. *Bioinformation*. 7(6):311-314, 2011.

[20] F. Gao and C. T. Zhang. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Res*. 34:W686-691, 2006.

[21] M. Jani, K. Mathee and R. K. Azad. Identification of Novel Genomic Islands in Liverpool Epidemic Strain of Pseudomonas aeruginosa Using Segmentation and Clustering. *Front Microbiol*. 7:1210, 2016.

[22] D. Che, H. Wang, J. Fazekas and B. Chen. An Accurate Genomic Island Prediction Method for Sequenced Bacterial and Archaeal Genomes. *J Proteomics Bioinform*. 7:214-221, 2014.

[23] C. C. Lee, Y. P. Chen, T. J. Yao, C. Y. Ma, W. C. Lo, P. C. Lyu and C. Y. Tang. GI-POP: a combinational annotation and genomic island prediction pipeline for ongoing microbial genome projects. *Gene*. 518(1):114-123, 2013.

[24] M. S. Hasan, Q. Liu, H. Wang, J. Fazekas, B. Chen and D. Che. GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences. *Bioinformation*. 8(4):203-205, 2012.

[25] B. Lu and H. W. Leong. GI-SVM: A sensitive method for predicting genomic islands based on unannotated sequence of a single genome. *J Bioinform Comput Biol*. 14(1):1640003, 2016.

[26] A. Tsirigos and I. Rigoutsos. A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res*. 33(12):3699-3707, 2005.

[27] Q. Zhu, M. Kosoy and K. Dittmar. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics*. 15:717, 2014.

[28] R. Jain, S. Ramineni and N. Parekh. IGIPT - Integrated genomic island prediction tool. *Bioinformation*. 7(6):307-310, 2011.

[29] S. Shrivastava, C. V. Reddy and S. S. Mande. INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *J Biosci*. 35(3):351-364, 2010.

[30] C. M. Hudson, B. Y. Lau and K. P. Williams. Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Res*. 43:D48-53, 2015.

[31] W. Hsiao, I. Wan, S. J. Jones and F. S. Brinkman. IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*. 19(3):418-420, 2003.

[32] M. G. Langille, W. W. Hsiao and F. S. Brinkman. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics*. 9:329, 2008.

[33] B. K. Dhillon, M. R. Laird, J. A. Shay, G. L. Winsor, R. Lo, F. Nizam, S. K. Pereira, N. Waglechner, A. G. McArthur, M. G. Langille et al.. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res*. 43(W1):W104-108, 2015.

[34] D. M. de Brito, V. Maracaja-Coutinho, S. T. de Farias, L. V. Batista and T. G. do Rêgo. A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm. *PLoS One*. 11(1):e0146352, 2016.

[35] S. H. Yoon, Y. K. Park, S. Lee, D. Choi, T. K. Oh, C. G. Hur and J. F. Kim. Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res*. 35:D395-400, 2007.

[36] S. H. Yoon, Y. K. Park and J. F. Kim. PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Res*. 43:D624-30, 2015.

[37] Q. Tu and D. Ding. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiol Lett*. 221(2):269-275, 2003.

[38] R. Pierneef, L. Cronje, O. Bezuidt and O. N. Reva. Pre_GI: a global map of ontological links between horizontally transferred genomic islands in bacterial and archaeal genomes. *Database (Oxford)*. 2015:bav058, 2015.

[39] J. C. Ogier, A. Calteau, S. Forst, H. Goodrich-Blair, D. Roche, Z. Rouy, G. Suen, R. Zumbihl, A. Givaudan, P. Tailliez et al.. Units of plasticity in bacterial genomes: new insight from the comparative genomics of two bacteria interacting with invertebrates, Photorhabdus and Xenorhabdus. *BMC Genomics*. 11:568, 2010.

[40] S. Waack, O. Keller, R. Asper, T. Brodag, C. Damm, W. F. Fricke, K. Surovcik, P. Meinicke and R. Merkl. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*. 7:142, 2006.

[41] W. Wei, F. Gao, M. Z. Du, H. L. Hua, J. Wang and F. B. Guo. Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties. *Brief Bioinform*. pii:bbw019, 2016.

[42] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafe, A. Perez et al.. Machine learning in bioinformatics. *Briefings in Bioinformatics*. 7(1):86-112, 2005.

[43] T. Manning, R. D. Sleator and P. Walsh. Biologically inspired intelligent decision making: a commentary on the use of artificial neural networks in bioinformatics. *Bioengineered*. 5(2):80-95, 2014.

[44] P. Wan and D. Che. A Computational Framework for Tracing the Origins of Genomic Islands in Prokaryotes. *Int Sch Res Notices*. 2014:732857, 2014.

[45] S. Karlin, J. Mrázek and A. M. Campbell. Codon usages in different gene classes of the Escherichia coli genome. *Mol Microbiol*. 29(6):1341-1355, 1998.

[46] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin et al.. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*. 102(39):13950-13955, 2005.

[47] S. C. Soares, A. Silva, E. Trost, J. Blom, R. Ramos, A. Carneiro, A. Ali, A. R. Santos, A. C. Pinto, C. Diniz et al.. The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains. *PLoS One*. 8(1):e53818, 2013.

[48] L. F. Chen, D. J. Anderson and D. L. Paterson. Overview of the epidemiology and the threat of Klebsiella pneumoniae carbapenemases (KPC) resistance. *Infect Drug Resist*. 5:133-141, 2012.

[49] O. Bezuidt, R. Pierneef, K. Mncube, G. Lima-Mendez and O. N. Reva. Mainstreams of horizontal gene exchange in enterobacteria: consideration of the outbreak of enterohemorrhagic E. coli O104:H4 in Germany in 2011. *PLoS One*. 6(10):e25702, 2011.