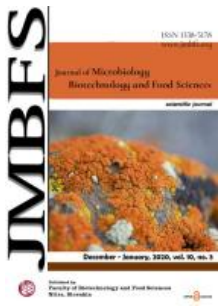

JMBFS

Journal of Microbiology, Biotechnology and Food Sciences

International peer-reviewed scientific online journal

Published by Faculty of Biotechnology and Food Sciences

APPLICATION OF SYSTEMS BIOLOGY APPROACH FOR INVESTIGATION OF PAN GENOME AND PHYLOGENETICS IN VARIOUS STRAINS OF ANAPLASMA PHAGOCYTOPHILUM

Pallavi Singh[1], *Dipanshi Verma*[2]

Address(es): Pallavi Singh, Ph.D,
[1,2] Department of Biotechnology, IILM CET, Greater Noida, Uttar Pradesh, India; Contact at:+91-8750406460

*Corresponding author: pallavi.singh@iilmcet.ac.in

doi: 10.15414/jmbfs.2020.10.3.445-448

ARTICLE INFO

Received 20. 5. 2020
Revised 4. 7. 2020
Accepted 23. 9. 2020
Published 1. 12. 2020

Regular article

OPEN ACCESS

ABSTRACT

Anaplasma phagocytophilum, has been observed as an emerging human pathogen of public health importance and is commonly transmitted to humans by tick bites. The bacterium Anaplasma phagocytophilum has been known from decades to cause the disease, tick-borne fever (TBF) in domestic ruminants and cattles in various areas in northern Europe, China, Russian border and United Kingdom. In recent years, outbreak of A.phagocytophilum infection has enhanced multifold and is widely reported in I. persulcatus and engorged as D. silvarum ticks in north eastern regions of China. However, few genome sequences have been completed so far, thus observations on biological, ecological, and pathological differences between genotypes of this bacterium,are yet to be elucidated by molecular and experimental infection studies.

In our current work, We have investigated 4 completely sequenced Genomic strains of A.phagocytophilum using various insilico tools SPINE and AGENT to characterize the percentage of Core Genome, Accessory genome and Pan Genome of these species. Further, we have tried to characterize the serotype and find the resistance genes observed in these four strains using MLST and ResFinder tools available at centre of Genetic epidemiology, Denmark Technical University server. By application of ClustAGE tool, we have made a comparative assessment of accessory genes across these strains. Heatmaps using expression map of these 4 genomes was constructed to infer the conserved genes and variable genes across these strains. Our study led to conclusion that core genome across these strains varies from 1.43 Mbps to 1.47 Mbps and accessory genomes varies from 0.0410 Mbps to 0.0734 Mbps. Comparison of the Gene clusters led to conclusion that gene clusters led to core genome value of 318 and Pan Genome value of 1035. Our analysis characterizes the dominance of accessory genes during evolution of Anaplasma phagocytophilum and lesser conservation of genes as there is a phylogenetic variation observed.

Keywords: *Anaplasma phagocytophilum*, pan genome, core genome, ClustAGE, SPINE

INTRODUCTION

Anaplasma phagocytophilum is a causative agent for human granulocytic anaplasmosis (HGA), a significant tick-borne zoonosis rising in the United States and other portions of the world (Xiong et al, 2019). *Anaplasma phagocytophilum*, influences a few types of wild and tamed warm-blooded animals. The specialty for *A. Phagocytophilum*, the neutrophil, demonstrates that the pathogen has special adjustments and pathogenetic mechanisms. HGA is progressively perceived as a significant and successive reason for fever after a tick bite in the Upper Midwest, New England, parts of the Mid-Atlantic States, and many parts of Europe, all territories where Ixodes ticks bite people (Mghirbi, Youmna ,2012).

A. phagocytophilum genomes are currently available, of which just four are complete (Dunning Hotopp JC, Lin M, Madupu R et al, 2006).Apart from Norway Variant 2, obtained from a Norwegian sheep, all genomes correspond to North American strains: human strains HZ, HZ2, and HGE1, Dog2 dog strain, MRK horse strain, JM rodent strain, and the tick (*Ixodes scapularis*) strains CRT38 and CRT35. The number of anaplasmosis cases reported to CDC has increased steadily since the disease became reportable, from 348 cases in 2000, to a peak of 5,762 in 2017 (CDC Reports, 2017). However, cases reported in 2018 were substantially lower. The case fatality rate (i.e., the proportion of anaplasmosis patients that reportedly died as a result of infection) has remained low, at less than 1%.Molecular modelling and computational chemistry approaches are applied to model the proteins. (Adejoro I.A et al, 2012). In silico prediction of biological activity using PASS in relation to the chemical structure of a compound is now a commonly used technique in drug discovery and development to predict the biological activity spectrum for a compound on the basis of its structural formula. (Valli G., Ramu K., Mareeswari P.,2012)

In South Korea, *Ixodes* spp. ticks are uncommon.(Park SW, Song BG, Shin EH, Yun SM,2014) However, *A. phagocytophilum* has been demonstrated in *Haemaphysalis longicornis* ticks, which are the most abundant species in South Korea (Kim CM, Kim MS, Park MS, Park JH, Chae JS, 2003) and this has led to growing concern about the possible emergence of HGA in South Korea. Actually, recent seroprevalence studies have shown that 1.8% of serum samples from febrile patients were positive for *A. phagocytophilum* in an immunofluorescence assay (IFA) test in 2002, and in 2003 the percentage was 8.9% from patients with symptoms of high fever suspected mainly scrub typhus. Many of the Thiohydantoine derivatives which are known to possess anti-tumorous activity have been tested against strains of Anaplasma. (Nillohit Mitra Ray, Rahul Singh, et al 2020)

Comparative genomic analyses lend insight into structural features such as variations related to genomic rearrangements, changes in the gene repertory, identification of horizontal gene transfer elements and prophage-related sequences, and hence expose particularities on the evolution in species.(Daniel Castillo et al, 2016)

In this work, we have used various insilico tools to investigate Core genome and Pan Genome of Anaplasma species and further applied ClustAGE to identify conserved accessory elements across these species.

METHODOLOGY

OBTAINING GENOMIC DATA OF STRAINS

Complete genomic sequences of 4 strains , Anaplasma phagocytophilum str HZ, RefSeq NC_007797.1,Genomic assembly GCA_000013125.1; Anaplasma phagocytophilum str HZ2, RefSeq NC_021879,Genomic assembly GCA_000439755.1; Anaplasma phagocytophilum str JM, RefSeq NC_021880.1,Genomic assembly GCA_000439775.1; Anaplasma phagocytophilum str. Norway variant2, RefSeq NZ_CP015376.1,Genomic assembly GCA_000689635.2 was downloaded from FTP site of National centre for Biotechnology information. Feature count and feature details of these strains were saved for analysis using opensource online servers. Genomic size of HZ

strain is 1.471 Mbps, consists of 1,155 coding genes and has GC content of 41.6 percent, Genomic size of HZ2 strain is 1.477 Mbps, consists of 1,154 coding genes and has GC content of 41.6 percent ; Genomic size of JM strain is 1.481 Mbps, consists of 1,162 coding genes and has GC content of 41.6 percent while Genomic size of Norway variant2 strain (NV) is 1.545 Mbps, consists of 1,179 coding genes and has GC content of 41.7 percent.

SEROTYPING AND RESISTANCE GENES PROFILING

Batch upload of assembled genomic sequences of these 4 strains of A.phagocytophilum in Fasta format was done at Bacterial analysis pipeline (available at https://cge.cbs.dtu.dk/services/cge/) tool available at Centre for Genomic epidemiology. At cut off percentage of 80%, resistance genes present in these strains and serotyping of these strains with known bacterial strains was done.ResFinder is a prominent tool to find known existing resistance genes in microbial species and MLST (Kleinheinz KA, Joensen KG, Larsen MV, 2014)is used to detect for typing the species based on local alignment of query sequence with known target Kmers stored in the database of CGE. Based on maximum HSP length and percentage similarity, Species typing results are obtained.

ANALYSIS OF CORE GENOME, ACCESSORY GENOME AND PAN GENOME

Genomic sequences of all the four strains of A.phagocytophilum were submitted to Spine tool available at http://vfsmspineagent.fsm.northwestern.edu. in order to identify the core genome and accessory genome of each individual strain and generation of Pan Genome data as a whole. SPINE is a script written in Perl to identify core genome of various from genomic DNA sequences. It uses NUCmer parameters (Ozer EA, Allen JP, and Hauser AR, 2014). Further, AGEnt tool was used to find out accessory genome of each sequence. CORE genome data obtained as an output of SPINE was used as an input in AGENT. Accessory genome data generated by AGENT was further processed as an input file for ClustAGE application (Ozer, E.A, 2018). ClustAGE application is freely available at this site http://vfsmspineagent.fsm.northwestern.edu/cgi-bin/clustage.cgi to construct the map of accessory genes of various strains. We need to use the output generated by AGENT for accurate generation of ClustAGE map.Further, there is limitation of maximum of 15 accessory genomes for web server of ClustAGE. We had run ClustAGE on accessory genome dataset of 4 strains in fasta format using the default settings of a minimum of 80% identity in nucleotide sequence and threshold of hit length of 100 bp at minimum. EDGAR application was used to find out relationship between orthologous gene clusters of these strains. Standard Decay function with fitted equation 1034.555.∗x^0.144 (α=0.586) was applied for Pan Genome analysis.

RESULTS AND DISCUSSION

Genomic data availability at NCBI FTP site showed 30 sequenced samples available of A.phagocytophilum and it included four whole genome sequences available for sequences under study. Dendrogram of Anaplasma strains (Figure 1) depicts close similarity and phylogenetic relationship between HZ,HZ2 and JM strains. However Norway Variant2 is most distant homologue as inferred from maximum phylogenetic distance in the lineage map.

Bacterial analysis pipeline results of CGE server show absence of any antibiotic resistance gene at threshold of 80% and 90% similarity in known resistance genes and A.phagocytophilum genomic sequences. However,HZ and HZ2 strains have shown species typing matched to ST161, JM strain to ST-64 and NV strain to ST-82.(Figure 2).

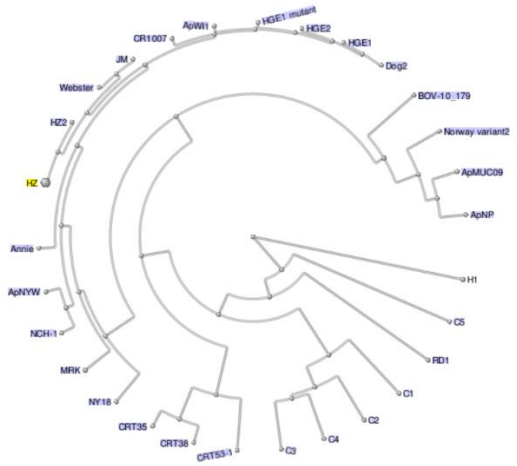

Figure 1 Dendrogram of various strains of Anaplasma phagocytophilum showing a close evolutionary lineage between the strains HZ and HZ2. Norway Variant2 (NV) has maximum phylogenetic distance from HZ strain

Sample Name	Species	MLST	Plasmids	pMLSTs	Resistance Genes	Virulence Genes
HZ strain	Anaplasma phagocytophilum	ST-161	NA	NA		NA
HZ2 strain	Anaplasma phagocytophilum	ST-161	NA	NA		NA
JM strain	Anaplasma phagocytophilum	ST-64	NA	NA		NA
NV strain	Anaplasma phagocytophilum	ST-82	NA	NA		NA

Figure 2 Bacterial Analysis Pipeline results of CGE server for A.phagocytophilum

Results obtained from SPINE and Agent show that core genome varies from 1.43 Mbp in HZ strain to 1.436 Mbp in HZ2 strain, 1.44 Mbp in JM and 1.47 Mbp in NV leading to interpretation of huge amount of conserved genomes observed across these species.(Table 1). Smallest output segment of core genome varies on an average from 77,748 bases to largest output segment of 7,62,043 bases across these species while Accessory genome varies from 10 bases to 2561 bases. (Figure 3).Maximum variation is observed across the A. phagocytophilum strain Norway variant2. Pangenome development plot (Figure 4) obtained from EDGAR shows approximately 1034 new genes are added to each new species and core genome plot (Figure 5) shows that as new species are added ,319 genes are conserved in these species showing substantial phylogenetically consensus patterns to be observed during addition of new genomes.

ClustAGE plot generated 153 bin gene clusters varying across these genomes and reveals that variable genes of size 5 kbp to 70 kbp are existing across strains HZ, HZ2,JM and NV variants.

Table 1 Results of SPINE and AGEnt depicting analysis of Core Genome and Accessory Genome composition for various strains of A.Phagocytophilum

A.phagocytophilum Strain Name	Accession No.	Type of Genome	Size in basepairs	Smallest Segment (Size in Bases)	Largest Segment (Size in bases)	Average length of the output segment
HZ	NC_007797.1	Accessory	41049	10	2359	220.69
		Core	1430214	79042	752744	1170
HZ2	NC_021879	Accessory	41343	10	2359	218.75
		Core	1436219	79042	744155	1163
JM	NC_021880.1	Accessory	41305	10	2359	225.71
		Core	1440262	78108	762043	1294
NV2	NZ_CP015376.1	Accessory	73440	11	2561	248.11
		Core	1471706	77748	477827	769.5
	Result	**BackBone**	**1430214**	**102**	**79042**	**7527.44**
		Pangenome	**1539423**	**10**	**1471282**	**5148.57**

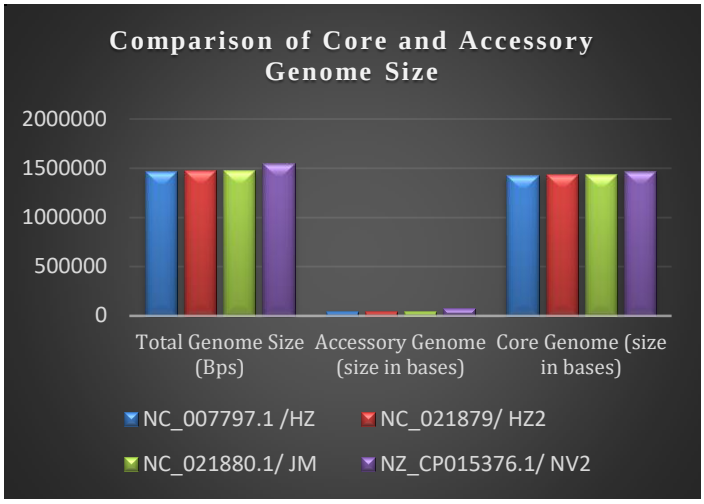

Figure 3 Comparison of Core Genome and Accessory Genome of given variants of A.phagocytophilum

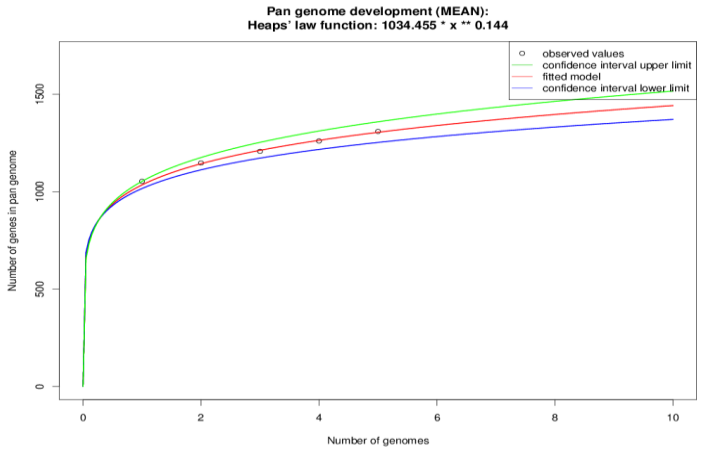

Figure 4 Pan genome development plot

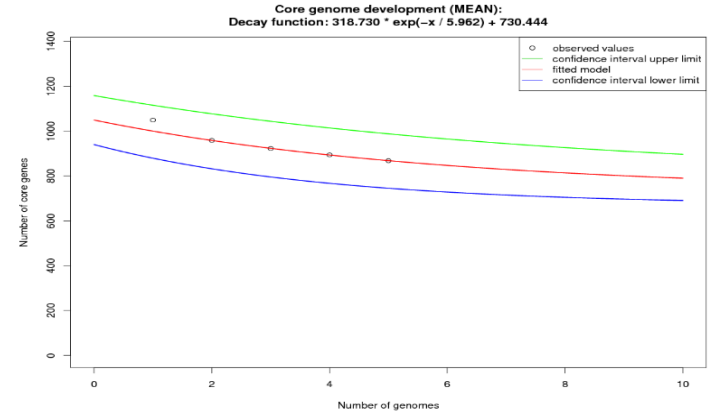

Figure 5 Core genome development plot

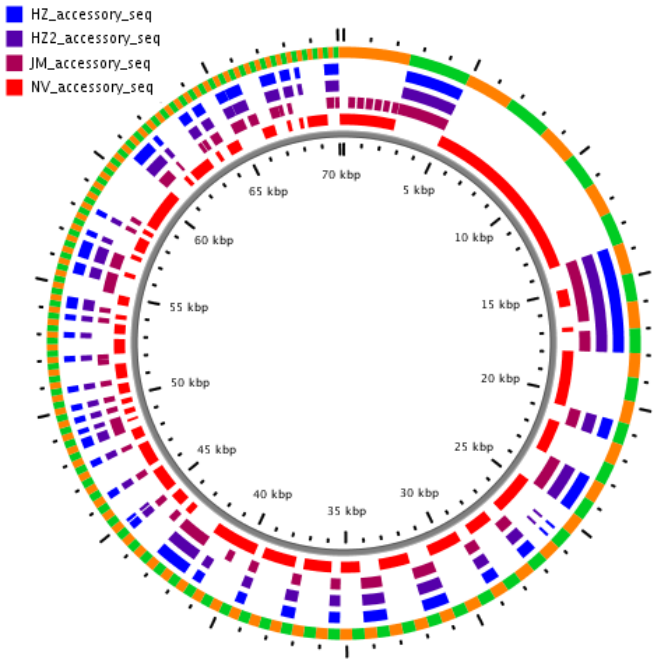

Figure 6 ClustAGE results for comparing the accessory genes across HZ,HZ2,JM and NV2 strains of *A.phagocytophilum*

CONCLUSION

Identification and analysis of Core Genome and accessory genomic elements in Anaplasma phagocytophilum variants is critical to understand the evolutionary relationship,niche adaptation, virulence factors and infectious potential. Our analysis of sequences NC_007797.1, NC_021879, NC_021880.1 and NZ_CP015376.1 has revealed that only 5% of genomes of this species contains accessory genetic elements and majority of genes are conserved during evolution. Core Genome of Anaplasma phagocytophilum is composed of 1.43 Mbp,1.43Mbp,1.44 Mbp,1.47 Mbp in total genomic size of 1.471 Mbp, 1.477 Mbp,1.481 Mbp and 1.545 Mbp across strains HZ, HZ2, JM and Norway Variant2 respectively. On analysis by AGENT, 179 accessory genetic elements were obtained between these species varying in size from 10 basepairs to 814 basepairs. Pan Genome analysis shows that on every new genome addition, 1035 genes will be added which is not a very huge number considering the average genome clusters to be 739 in these strains. Further, Species typing has led to interesting insights on matching of substantial portion of genome of this pathogen to Salmonella strains in CGE server. These insights are crucial for developing structure based drugs against Anaplasma species and targeting the infectious potential of this pathogen. High genetic lineage and similarity is a substantial benefit to be explored in developing therapies against this emerging pathogen.

Acknowledgements: Authors of this work express their gratitude to Director, IILM Engineering, Dr.Jyotsna Singh and Senior Director, IILM Engineering, Shri Ajay Pratap Singh for their constant motivation and valuable support in making this work possible. We extend our sincere thanks to entire IILM management for encouraging the faculty members and students in Engineering wing to pursue high quality research work consistently and create an impact in scientific domain.

REFERENCES

Adejoro I.A,Odiaka T.I and Akinyele O.F,2012. Molecular Modeling and Computational Studies of Dimethylpyridino-1,4-η-2-methoxycyclohexa-1,3-DieneIron tricarbonyl Complexes. Asian Journal of Research in Chemistry, Vol 5 (1), pg 146-152. https://doi.org/10.9734/ACSJ/2016/27275

Barbet AF, Al-Khedery B, Stuen S, Granquist EG, Felsheim RF, Munderloh UG, 2013.An emerging tick-borne disease of humans is caused by a subset of strains with conserved genome structure. Pathogens.,Vol2: 544-555. https://doi.org/10.3390/pathogens2030544

Bernard, G., Chan, C.X., Ragan, M.A., 2016. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetictransfer. Sci. Rep. 6, 28970. https://doi.org/10.1038/srep28970

Bhardwaj, T., Somvanshi, P., 2014. Plant Systems Biology: Insights and Advancements. In: Plant Omics: The Omics of Plant Science. Springer Publications, 978-81-322-2171-5, pp. 791–819. https://doi.org/10.1007/978-81-322-2172-2_28

Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter FJ, Zakrzewski M, Goesmann A 2009,EDGAR: a software framework for the comparative

analysis of prokaryotic genomes, BMC Bioinformatics, Vol 10:154 https://doi.org/10.1186/1471-2105-10-154
Centers for disease control and prevention, statistics and epidemiology of anaplasmosis. [http://www.cdc.gov/anaplasmosis/stats/]
Chang, G.W., Chang, J.T., 1975. Evidence for the B12-dependent enzyme ethanolamine deaminase in Salmonella. Nature 254, 150–151. https://doi.org/10.1038/254150a0s
Daniel Castillo,Rói Hammershaimb Christiansen, Inger Dalsgaard,Lone Madsen,Romilio Espejo,Mathias Middelboe, 2016.Comparative genome analysis provides insights into the pathogenicity of *Flavobacterium psychrophilum*. PLoS One 11 (4), e0152515. https://doi.org/10.1371/journal.pone.0152515
DelVecchio, V.G., Kapatral, V., Elzer, P., Patra, G., Mujer, C.V., 2002. The genome ofBrucella melitensis. Vet. Microbiol. 90, 587–592. https://doi.org/10.1016/s0378-1135(02)00238-9
Dunning Hotopp JC, Lin M, Madupu R, Crabtree J, Angiuoli SV, Eisen J, Seshadri R, Ren Q, Wu M, Utterback TR, Smith S, Lewis M, Khouri H, Zhang C, Niu H, Lin Q, Ohashi N, Zhi N, Nelson W, Brinkac LM, Dodson RJ, Rosovitz MJ, Sundaram J, Daugherty SC, Davidsen T, Durkin AS, Gwinn M, Haft DH, Selengut JD, Sullivan SA, 2006.Comparative genomics of emerging human ehrlichiosis agents. PLoS Genet., 2: 208-223. 10.1371/journal.pgen.0020208. https://doi.org/10.1371/journal.pgen.0020021
Emmanuelle Lerat ,Vincent Daubin, Nancy A Moran (2003). From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the gamma-Proteobacteria. PLoS Biol,1(1), .doi: 10.1371/journal.pbio.0000019
Gotoh, O. 1982. An improved algorithm for matching biological sequences.*J. molec. Biol.*162, 705–708.
Heo EJ, Park JH, Koo JR, Park MS, Park MY, Dumler JS, Chae JS, 2002. Serologic and molecular detection of *Ehrlichia chaffeensis* and *Anaplasma phagocytophila* (human granulocytic ehrlichiosis agent) in Korean patients. J Clin Microbiol 40: 3082–3085. https://doi.org/10.1128/jcm.40.8.3082-3085.2002
Herve´ Tettelin , David Riley , Ciro Cattuto and Duccio Medini,2008,Comparative Genomics: The Bacterial PAN Genome, Current opinion in Microbiology, Vol 12, Pg 472-477. https://doi.org/10.1016/j.mib.2008.09.006
https://doi.org/10.1016/0022-2836(82)90398-9
Kim CM, Kim MS, Park MS, Park JH, Chae JS, 2003. Identification of *Ehrlichia chaffeensis*, *Anaplasma phagocytophilum*, and *A. bovis* in *Haemaphysalis longicornis* and *Ixodes persulcatus* ticks from Korea. Vector Borne Zoonotic Dis 3: 17–26. https://doi.org/10.1089/153036603765627424
Kleinheinz KA, Joensen KG, Larsen MV. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*.2014;4(1):e27943. doi:10.4161/bact.27943
Mghirbi, Youmna ,2012 .*Anaplasma phagocytophilum in horses and ticks in Tunisia.*, Parasites and Vectors, Vol. 5.pg 86-91. https://doi.org/10.1186/1756-3305-5-180
Nillohit Mitra Ray, Rahul Singh, Joginder Singh, Shipra Bhati, Vikas Kaushik, 2020. Computational screening of Thiohydantoin Derivatives for antitumor activity. Research J. Pharm. and Tech ,Vol13(2):795-800.https://doi.org/10.5958/0974 360X.2020.00150.X
Ozer EA, Allen JP, and Hauser AR. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics* 2014 15:737
Ozer, E.A. ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. *BMC Bioinformatics* 19, 150 (2018). https://doi.org/10.1186/s12859-018-2154-x
Park JH, Heo EJ, Choi KS, Dumler JS, Chae JS, 2003. Detection of antibodies to *Anaplasma phagocytophilum* and *Ehrlichia chaffeensis* antigens in sera of Korean patients by western immunoblotting and indirect immunofluorescence assays. Clin Diagn Lab Immunol 10: 1059–1064.https://doi.org/10.1128/CDLI.10.6.1059-1064.2003
Park SW, Song BG, Shin EH, Yun SM, Han MG, Park MY, Park C, Ryou J, 2014.Prevalence of severe fever with thrombocytopenia syndrome virus in *Haemaphysalis longicornis* ticks in South Korea. Ticks Tick Borne Dis 5: 975–977. https://doi.org/10.1016/j.ttbdis.2014.07.020
Rocha EPC, 2004, The replication-related organization of bacterial genomes.Microbiology. Vol 150 :1609-1627. https://doi.org/10.1099/mic.0.26974-0
Tettelin, Hervé & Riley, David & Cattuto, Ciro & Medini, Duccio. (2008). Comparative genomics: The bacterial pan-genome. Current opinion in microbiology. 11. 472-7. 10.1016/j.mib.2008.09.006.
Tobias H Kloepperand Daniel H Huson, 2008, Drawing explicit phylogenetic networks and their integration into Splits Tree, BMC Evol Biol. 2008; 8: 22. https://doi.org/10.1186/1471-2148-8-22
Valli G., Ramu K., Mareeswari P.,2012, Salicylaldehyde Schiff bases Bioactivity Prediction by Insilico Approach.Asian Journal of Research in Chemistry, Vol 5 (4), pg 504-509.
Xiong, Qingming, et al ,*Infection by anaplasma phagocytophilum requires recruitment of low-density lipoprotein cholesterol by flotillins,2019,*American Society for Microbiology, Vol. 10 pg 1128-35. https://doi.org/10.1128/mBio.02783-18