ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Comparative Genomic Analysis of *Helicobacter pylori* Strains CPY1124 and SJM180 Related to Gastric Disorder Using Next-Generation Sequence (NGS) Data

Kavitha Kannan[1*], Shuhalfarina S[2] , Swaminathan Venkataramanan[2], and Mohana Priya Arumugam[1],

[*1]School of Bio Sciences and Technology(SBST), VIT University, Vellore 632014, Tamil Nadu, India.

[2]Department of Diagnostic and Allied Health Sciences, Management and Science University MSU, Shah Alam 40100, Malaysia.

ABSTRACT

Prolong colonization of Helicobacter pylori in the stomach is the major cause that may lead to various gastric disorder including gastritis, gastric ulcer and gastric cancer. H. pylori are able to survive in high acidic environment by neutralizing the stomach acid. The neutralization may lead to several gastric disorders. By doing complete comparative genomic analysis of H. pylori strains, the specific region of gene that induces the infection of the stomach can be identified. Sequencing and comparing targeted genomic region with reference genome by using multiple genome alignment tools provide necessary foundation regarding the conserved regions. Visualization of genomic comparison helps in determination of the differences of genotype between the closely related strains. Genome sequences of H. pylori strains CPY1124 and SJM180, isolated from gastric ulcer and gastritis patients respectively, were chosen for this research analysis. These two strains were sequenced with reference genome of PeCan4, isolated from gastric cancer patient, by using Next-Generation Sequence (NGS) data. It is shown that CPY1124 have higher percentage of similarities to PeCan4. Therefore, higher percentage of cancer-causing gene of PeCan4 was predicted to be found in CPY1124 rather than SJM180. This concludes that gastric ulcer may lead to gastric cancer.

Keywords: Helicobacter pylori, Comparative genomic analysis, strain CPY1124, strain SJM180, strain PeCan4, Next-Generation Sequence.

**Corresponding author*

INTRODUCTION

Helicobacter pylori (*H. pylori*) bacterium is a type of microorganism that can lead to serious disease. This bacterium is able to live in high acidic stomach environment which they can be found around the stomach mucus lining. In United Kingdom, approximately 40% of the citizens have been infected with *H. pylori.* About 9 out of 10 individuals that have been infected are not showing any symptoms (1). Even though this bacterium has been discovered in many decades, it is not precisely clear on how *H. pylori* is passed from one person to another person and why just some of the infected individuals get the ulcer disease. In fact, how *H. pylori* elevate the risk of gastric cancer to happen in those infected individuals are still unclear. A clear view of the pathogenesis of *H. pylori* may help to increase the understanding on how these gastroduodenal diseases to occur in a general manner (2). Currently, the treatment for *H. pylori* infection is highly effective. However, this bacterium can become resistant to the common antimicrobial drugs. In order to overcome this infection, further drug development need to be done to stop this infection to happen again and by developing an alternative treatment when resistance occurred, as for example, developing a vaccine against *H. pylori*. The occurrence of *H. pylori* cases in UK have decreased, but more research work need to conducted as to help decreasing the *H. pylori*-infected cases in the nations that are more susceptible towards this bacterium. (3) The *RAST* server(http://rast.nmpdr.org) is free and available for complete annotation of the genomes in order to identify the segments of genes and any associated important information with those segments. (4)

Gastric disorder, especially gastric cancer, is the leading cause associated with cancer-related death in Far Eastern nations such as China, Japan and Korea (5). Moreover, the early prevention steps to prevent the gastric cancer to occur cannot be determined because the absence of initial symptoms and most patients are examined after the cancer has attacked the specific part that are promoted to cause the gastric malignant. (6,7) In many cases, gastric cancer started from a cell within internal mucosa lining of the stomach. As the gastric cancer started to develop, the cancer cell may lead to the production of tumor that may attack the internal part of the stomach lining until it by pass the stomach wall and invade the nearest organs, such as liver and pancreas. The tumor may spread to the small intestine and esophagus as well. Some of the tumor cells may fuse into the circulatory system or lymph channels and spread to adjacent lymph nodes or other areas of the body. (8,9) Visualization of comparative genomic presented by *BRIG* software allows the user to view and identify the genotypic differences between related genomes (10).

MATERIALS AND METHODS

EUROPEAN NUCLEOTIDE ARCHIVE(*ENA*) DATABASE

ENA database (http://www.ebi.ac.uk/ena) is Europe's repository that provides a public access to get primary nucleotide sequences data. *ENA* database is aimed to provide support and facilitate the usage of nucleotide sequencing to be part of the research work by enabling submission of data, archive, search and download features. *ENA* comprises of several databases including Sequence Read Archive (*SRA*), Trace Archive and European Molecular Biology Laboratory (*EMBL*) Bank. By uniting all of these three databases for draft sequence data, assembly data and functional annotation, it brings *ENA* towards an integrated and comprehensive resource for biological information. *ENA* collaborate closely with DNA Data Bank of Japan (DDBJ) and National Center for Biotechnology Information (NCBI) in the International Nucleotide Sequence Database Collaboration.

PATHOSYSTEMS RESOURCE INTEGRATION CENTER (*PATRIC*) SERVER

PATRIC server (http://www.patricbrc.org) is represented as the Bacterial Bioinformatics Resource Center. This information system is designed to handle the communities of biomedical research work based on diseases that are infected by bacteria. It integrates the major and important information of the bacteria that can be retrieved by numerous analysis tools provided by the server. It provides an interactive user interface to ease the discovery process of the data and comparative genomics analysis can be conducted with efficient. PATRIC provides bacterial genomic databases, data associated with genomic analysis and numerous tools to perform bioinformatics analysis.

VELVET DE NOVO ASSEMBLER

Velvet is a collection of algorithm that has been used to manage alignments of short read sequencing and *de novo* genome assembly. This algorithm package is accomplished through the control of *de Bruijn* graphs for sequence assembly of genome. *Velvet de novo* assembler manipulates *de Bruijn* graph in an effective way by simplifying and compressing the data into single nodes without any loss of the graph data. It recognizes and removes any error and corrected the overlapping regions that contain in biological sequences by using a particular algorithm of error correction which combines the regions together. All of the overlapping regions are removed from the biological sequence by using the repeat solver algorithm which is responsible to separate the overlaps. In order to build *Velvet de Bruijn* graph, the user need to choose the appropriate *k-mer* length, expected coverage, minimum coverage cut-off value and length of minimum coverage. The input for *Velvet* assembler is sequence reads in FASTQ format and the outputs is the assembled contigs in multi-fasta format.

RAPID ANNOTATION BY USING SUBSYSTEM TECHNOLOGY (*RAST*) ANNOTATED SERVER

RAST takes the ordered contigs as input in multi-fasta format to recognize open reading frames (ORF) based on subsystem techniques in order to compare between a sophisticated genes database and sequences of RNA to produce a high-quality of assembly annotation. It provides high quality genome annotations to these genomes over the whole phylogenetic tree. Features of *RAST* annotated server include identification of genes that encode for proteins, ribosomal RNA (rRNA) and transfer RNA (tRNA), determination of genes' functions and identification of possible subsystems within the genomes. Users are able to use this information to rebuild the metabolic network and *RAST* produces an output which is easily to be downloaded. Other than that, the output of *RAST* annotated server can be viewed and analyzed in the environment that allows the comparative analysis to be done as the annotated genomes are maintained in the *SEED* environment.

MAUVE MULTIPLE GENOME ALIGNMENT

Mauve software is one of the methods that can be used to generate an alignment of whole genome sequences and also can be used to evaluate the quality of alignments, and drawn comparison for suitable downstream phylogenetic or evolutionary analyses. *Mauve* genome alignment has the ability to determine the conserved regions and any inversions and rearrangements that occur within the conserved regions can be identified by *Mauve*. If any rearrangement occurs within multiple genomes, *Mauve* is able to recognize the definite breakpoints among the genomes. In addition, *Mauve* was the first genome alignment software that incorporates extensive scale of evolution and orthodox flow of multiple genomic sequence alignment. *Mauve* takes a set of genome assemblies as input and generate multiple of complete genome alignment as output. This multiple genome alignment method is capable to provide alignment of conserved regions that presence in genome rearrangement. Unlike other genome alignment methods, *Mauve* was use an irrelevant basis for genome discovery to increase the speed of alignment. Furthermore, based on the algorithm of irrelevant selection, it assumes that the genomes to be studied by using *Mauve* are collinear. Identification and alignment of local collinear regions can also be done in *Mauve.* The local collinear regions are called as locally collinear blocks (LCBs). Each of the LCB is represented as a homologous sequence region that is shared by minimum of two genomes. This LCB did not consist of any rearrangement within the homologous region.

BLAST RING IMAGE GENERATOR (*BRIG*) VISUALIZATION

BRIG leads to the new visualization and abstraction techniques that are needed to enhance the understanding, validation and relation of databetween related genomes. By using *BRIG*, the user is able to generate graphical comparison in a circular form, which is represented as the genomic ring. Visualizing the comparison between the target and reference genomes enable theidentical regionas well as the distinct region between them to be identifiedbased on the coloured genomic rings.*BRIG* application is profoundly adaptable as it has the capability in visualizing raw genome sequence data which consists of contigs boundaries, read coverage or mapping data.Furthermore, visualization of multiple graphs and annotations can be done by using *BRIG* in a simultaneous way. A whole raw genome that has been assembled into contigs sequences or scaffolds, which contains ordered contigs that are separated by gaps, in multi-fasta format is used as reference sequence. *BRIG* takes a reference sequence and target sequence as input to display the circular image that represent the high identical percentage among those sequences.

Retrieval of *Helicobacter Pylori* Strains

The genome sequences of *Helicobacter pylori* were downloaded from the ENA database and the visualization of each targeted genome were done using PATRIC server. *H. pylori* strains of CPY1124, SJM180 and PeCan4 were chosen for this study. All of the genome sequences were downloaded in FASTQ format. The accession number of strains CPY1124, SJM180 and PeCan4 were SRR400670, ERR351267 and ERR351243 respectively. The length of genome sequences for CPY1124 was 1,538,955bp, SJM180 was 1,644,768bp and PeCan4 was 1,629,557bp. These three strains were also available in DDBJ/EMBL/GenBank databases under the accession number of CP002074, CP002073 and NC_014555 respectively.

Genome Assembly

For genome assembly, draft quality of *H. pylori* genome sequences for each targeted strain was obtained. Whole complete genome sequences of *H. pylori* strains of CPY1124 and SJM180 were assembled by using *Velvetde novo* assembler. *Velvet*assembler was used to read sequences and set of overlapping sequence reads were removed without guidance from the reference genome in order to produce high quality of unique contigs sequences. Genome sequence reads of both *H. pylori* strains in FASTQ format are used as the data input in *Velvet* assembler. Parameters which are the *k-mer* length, expected coverage, minimum coverage cut-off value and length of minimum contigs were set to build the *Velvetde Bruijn* graph. The length of *k-mer*, expected coverage value and minimum coverage cut-off value were set as 31, 21 and 2.81 respectively. Next, the length of minimum contigs was set to 200bp as it was the minimum length of genome sequence to be submitted into GenBank.

Ordering Contigs of Assembled Sequence

The contigs file containing the unordered contigs sequences for both strains were used as the data input in *Mauve*. These unordered contigs sequences were compared with the genome sequence of strain PeCan4, which was used as a reference, to make the contigs sequences in ordered sequences. The PeCan4 genome sequence used in *Mauve* was in FASTA format.

Large-Scale of Genome Annotation

The ordered contigs sequences for both *H. pylori* strains of CPY1124 and SJM180 obtained in *Mauve* were used as the data input in *RAST. RAST* accepted the ordered contigs sequences in multi-fasta format. By using *RAST* server, the contigs sequences for both strains were annotated. The segments of open reading frames (ORF) and important information related to that particular ORF were identified. After both of the contigs sequences were annotated, the annotated sequences were downloaded in GenBank format to be used in further analysis.

Whole Genome Multiple Alignment

Whole multiple genome alignment was done by using *Mauve*. The annotated contigs sequences obtained from *RAST* for *H. pylori* strains of CPY1124 and SJM180 were used as the data input to be aligned with the reference strain. Contigs sequences of these two strains, together with the reference genome sequence of PeCan4, were uploaded in *Mauve*. By using *Mauve*, the complete genome alignments of these three strains were obtained to analyze on the similar and conserved regions between the query strains and reference strain.

Visualization of Comparative Genome

The comparison between the query *H. pylori* strains of CPY1124 and SJM180, to the reference strain of PeCan4 were visualized by using *BRIG* software. As the data input, the annotated contigs sequences obtained from *RAST* for both *H. pylori* strains of CPY1124 and SJM180 were used together with the reference genome sequence of PeCan4. The comparison was done and the similar regions between the query strains and reference strain were shown based on the colour-shaded rings.

RESULTS

In previous study, the extracted DNA samples of *H. pylori* strains CPY1124 and SJM180 are sequenced using 454 GSFLX Titanium and Illumina Hiseq 2000 sequencing technology respectively. The whole genome sequence of *H. pylori* strains CPY1124 and SJM180 is isolated from a patient with gastric ulcer and gastritis disease respectively. The general characteristic of those *H. pylori* strains was compared and summarized (Table 1)

	CPY1124	SJM180
Origin	Japan	Peru
Sequence Platform	454 GS FLX Titanium	IlluminaHiseq 2000
Disease	Gastric Ulcer	Gastritis
Genome Size	1, 538,955 bp	1,644,768 bp
Contigs Number	367	717
GC Content	38.9%	38.9%
Coding Sequence	1,563	1,560
Subsystems Number	252	239
RNA Number	47	46

Table 1: Genome characteristics of *H. pylori* strains

These two targeted strains were annotated by using the *SEED-Viewer* that generated by RAST that usually used the subsystem based assertions to construct the detailed of genomic metabolism process (Figure 1 and 2). The identification of identical and non-identical region among both *H. pylori* strains CPY1124 and SJM180 were done to identify the specific causative gene that may lead to gastric cancer disease. (Figure 3).

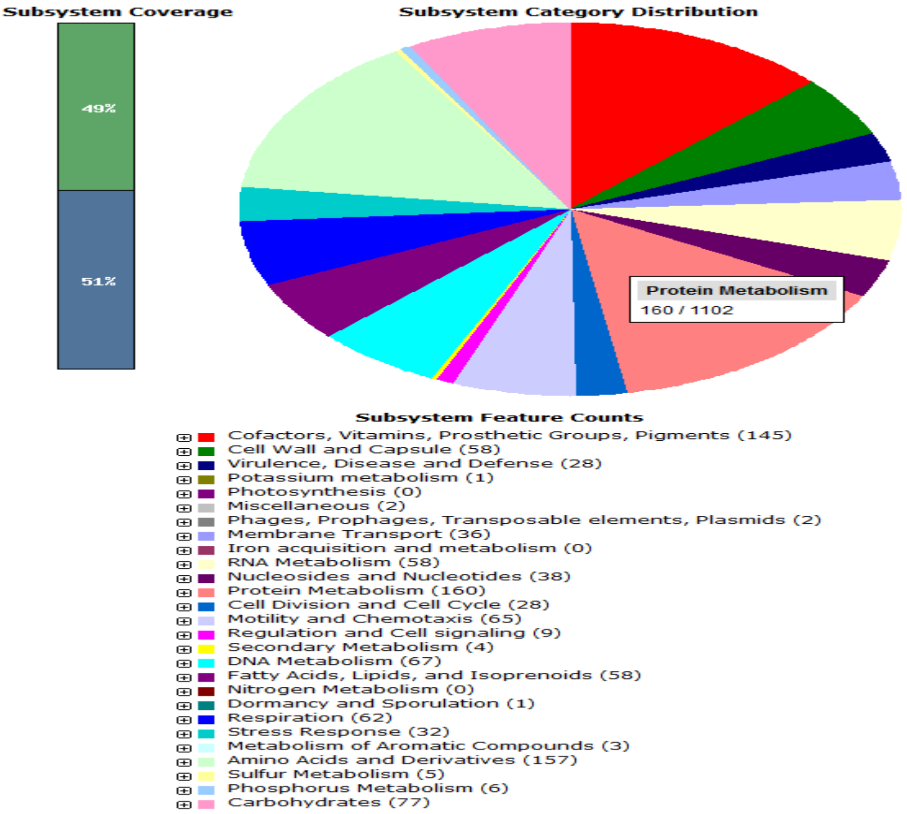

Figure 1: Subsystem distribution statistic of *H. pylori* CPY1124 strain based on genome annotation performed by RAST server.

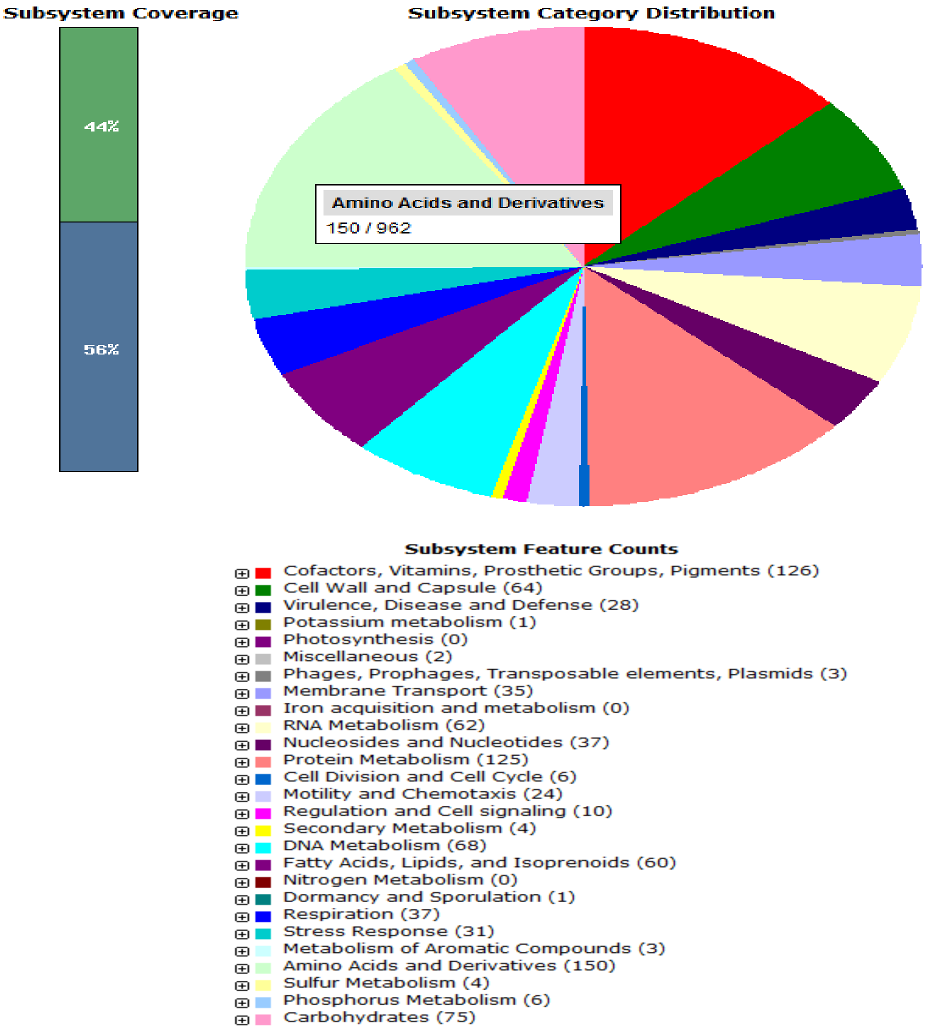

Figure 2: Subsystem distribution statistic of *H. pylori* SJM180 strain based on genome annotation performed by RAST server.

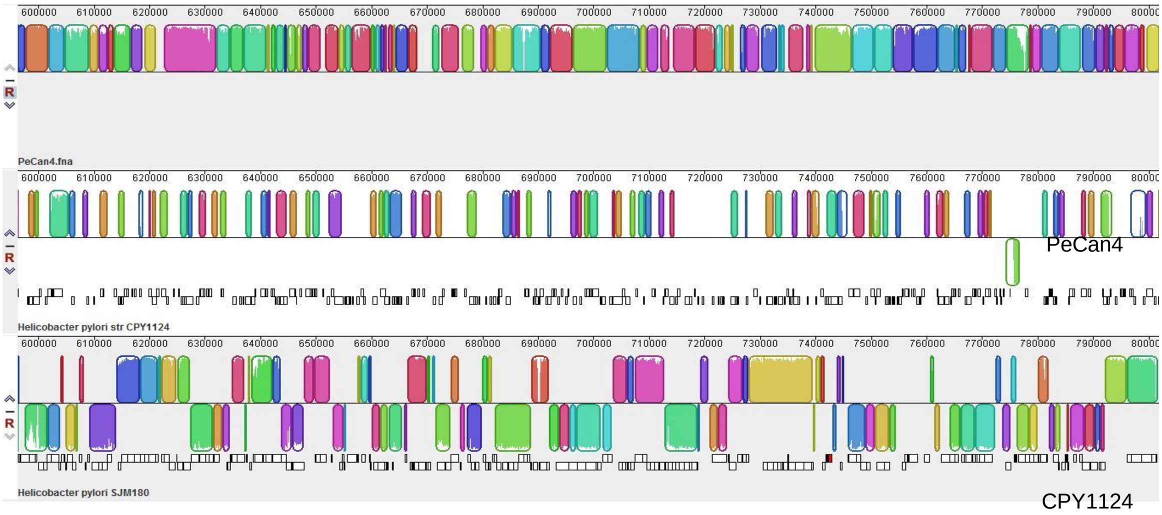

Figure 3: Result of multiple genomic comparisons between H. pylori strains CPY1124 and SJM180 after aligned with reference genome of strain PeCan4 by using Mauve. The same colour-shaded blocks indicate the conserved and unique gene sequence between those three genomes. The white gaps between the coloured blocks represent low coverage areas when compared with reference genome. Row 1 is represented by reference genome of strain PeCan4. Row 2 is represented by annotated contigs sequence of strain CPY1124. Row 3 is represented by annotated contigs sequence of strain SJM180.

SJM180

DISCUSSION

Based on this result, the targeted *H. pylori* strains CPY1124 and SJM180 had been compared with reference strain PeCan4. The genomic comparison between the two targeted strains with reference strain allows the visualization of the shared regions between them. These findings had proven that the gastritis and gastric ulcer disease have the possibility to cause gastric cancer or not, hence, determine which condition had the higher risk to develop the cancer. High genomic similarity between CPY1124 with PeCan4 leads to the idea which gastric ulcer disease may lead to gastric cancer. This may help in increasing the awareness of the medical health care practitioners as well as the *H. pylori* infected patients to be more cautious and prevent from the gastric cancer to occur.

CONCLUSION

In this study, comparative genomic analysis was done on the target sequences of *H. pylori* strains CPY1124 and SJM180, which these two strains were isolated from patients with gastric ulcer and gastritis disease, respectively. Comparative analysis of these two *H. pylori* strains derived from different clinical conditions had provide a foundation to understand in a clear manner on why *H. pylori* may be associated with gastric cancer. The alternative hypothesis that *H. pylori* strains CPY1124 have high identical region with reference strain, PeCan4 than strain SJM180, as based on the result, gastric ulcer sequence has shown the highest similarity with gastric cancer sequence. This concludes that gastric ulcer conditions may lead to the development of gastric cancer disease.

REFERENCES

[1] Fuccio L, Zagari RM, et al. "Meta-analysis: can *Helicobacter pylori* eradication treatment reduce the risk for gastric cancer?".*Ann Intern Med*151 (2): 121–8.

[2] Bytzer P, Dahlerup JF, Eriksen JR, Jarbøl DE, Rosenstock S, Wildt S "Diagnosis and treatment of Helicobacter pylori infection".*Dan Med Bull*58 (4): C4271. PMID 21466771. Retrieved 7 August 2013

[3] Touchman, J. "Comparative Genomics".*Nature Education Knowledge* 3 (10).

[4] Aziz, R. K., Bartels, D., et al The RAST Server: rapid annotations using subsystems technology. BMC Genomics, 9, 75. doi:10.1186/1471-2164-9-75.

[5] Zerbino, D. R.Genome assembly and comparison using de Bruijn graphs. Molecular Biology, 149. doi:10.1016/j.isprsjprs.2006.12.001.

[6] Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W.GenBank. Nucleic Acids Res., 38, D46–D51

[7] Blanchard, T G; Nedrud, J G"9. Helicobacter pylori Vaccines".In Sutton, Philip; Mitchell, Hazel.*Helicobacter Pylori in the 21st Century*. Mitchell, Hazel. CABI. pp. 167–189. ISBN 978-1-84593-594-8.Retrieved 7 August 2013.

[8] Moon JK, Kim JR, Ahn YJ, Shibamoto T"Analysis and anti-Helicobacter activity of sulforaphane and related compounds present in broccoli (Brassicaoleracea L.) sprouts". *J. Agric. Food Chem.*58 (11): 6672–7. doi:10.1021/jf1003573.

[9] Wroblewski, L. E., Peek, RM, & Wilson, KT.Helicobacter pylori and gastric cancer: Factors that modulate disease risk. *Clinical Microbiology Reviews*, *23*(4), 713–739. doi:10.1128/CMR.00011-10.

[10] Alikhan, N. F., Petty, N. K., et al. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics, 12(1), 402. doi:10.1186/1471-2164-12-402