Biojournal of Science and Technology

Research Article

Identification of the positively selected genes governing host-pathogen arm race in Vibrio sp. through comparative genomics approach

Atai Rabby[1], Sajib Chakraborty[1*], Atiqur Rahman[1], Shamma Shakila Rahman[1], Shahjalal Soad[1], Kaniz Fatima Chanda[1], Rajib Chakravorty[2]

[1] *Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka-1000, Bangladesh.*
[2] *Department of EEE, University of Melbourne, National ICT Australia, Victoria 3010, Australia*

***Corresponding author**
Sajib Chakraborty
Assistant Professor, Department of Biochemistry and Molecular Biology, Faculty of Biological Sciences, University of Dhaka, Dhaka-1000, Bangladesh. Email: schak.du@gmail.com

Published: 24-07-2015
Biojournal of Science and Technology Vol.2:2015
Academic Editor: Dr. Md. Shahidul Islam

Received: 01-06-2015
Accepted: 29-06-2015
Article no: m140008

Abstract

Bacterial evolution is due to the adaptive nature of the core bacterial genomes that plays critical role in diversification, fitness and adaptation of the species to different environment and host. Since Vibrio cholerae represents an appropriate model organism for studying the interplay of environment and host driven factors shaping the microbial genome structure and function, the current study aims to identify genes that are under these strong forces in V. cholerae. Here, we employed a comparative genomics approach to identify genes that are under positive selection in ten strains of Vibrio sp. including four pathogenic V. cholerae strains. From the available genome sequence data, a total of 422 orthologous genes were identified by reciprocal BLAST best-hit method, recombination breakpoint frequency analysis and tree comparison method. These 422 genes, representing the core genome of Vibrio sp., constituted the dataset to be analyzed for evolutionary selections. The analysis of natural selection, based on Maximum Likelihood method on synonymous and non-synonymous substitution rate, confirms the hypothesis that the bacterial core genomes are mostly under purifying selection with a few positively selected regions. However, our finding also reveals that positively selected sites in the Vibrio genome occur in a wide range of different genes encompassing diverse functional pathways including cell surface proteins (e.g. outer membrane-specific lipoprotein transporter/assembly proteins etc.), cell motility proteins (e.g. flagellar motor switch proteins, flagellar hook and assembly proteins), nutrient acquisition (e.g. amino acid, carbohydrate and phosphate ABC transporters), DNA repair and transcription related proteins. Interestingly, these positively selected gene products are directly involved with host-pathogen interactions and fitness in gastrointestinal environment. Therefore, the collective evidences of these positively selected genes spanning several pathways raise the possibility of their involvement in evolutionary arms races with other bacteria, phages, and/or the host immune system. This finding points to the natural selections which is the responsible factor for the diversification of Vibrio genus.

Keywords: positive selection, Vibrio cholerae, genome wide selection, molecular evolution

INTRODUCTION

Host-pathogen interactions and environmental settings are evolutionary forces that confer the adaptation of bacterial genomes (Petersen et al. 2007). The growing number of sequenced bacterial genomes coupled with the comparative genomics techniques provides the solid platform to investigate the nature of the genome-shaping natural selection process in bacteria. Particularly, the genes under positive selection have drawn much of the attention since these genes correlate with the functional adaptations in response to environmental settings and host selection pressure.

A number of studies have been conducted to elucidate the positively selected genes in different bacterial genomes including pathogenic *Escherichia coli* (Chen et al. 2006, Petersen et al. 2007), *Campylobacter* sp. (Lefebure and Stanhope 2009) and *Burkholderia pseudomallei* (Nandi et al. 2010). Here, we attempted to investigate the positively selected genes in *V. cholerae* in genome wide manner. The extensive studies on *V. cholerae* in terms of natural habitat, pathogenicity as well as it's interactions with host along with the availability of genome sequences for a number of different strains of *V. cholerae* have made it as an appropriate model organism for studying the interplay of environment and host driven factors that tend to alter the microbial genome structure and functions. In this study, we aimed to identify the genes under positive selection in *V. cholerae* genome as a trace of ongoing interplay of environment and host driven factors with bacterial genome.

V. cholerae is a gram negative gammaproteobacteria. Some of its strains are pathogenic and responsible for one of the most prominent life threatening disease, cholera. The evolution of *V. cholerae* genes throughout its entire genome will give us insights regarding its pathogenicity and host-pathogen interaction patterns.

V. cholerae lives both in aquatic environment as well as human gut and also infected by Vibriophages. Thus, evolutionary pressures are constantly acting on the genome. Studies have shown that multiple quorum-sensing circuits function in parallel to control virulence and biofilm formation in *V. cholerae*.(Hammer and Bassler 2003) Motility of *Vibrio* group is mediated by a virulent factor which has also been established previously.(BERRY 1975) These data indicates to the possible targets for natural selections as the host will try to mimic the virulence factors and the protein products of the selected genes come to direct or indirect host-pathogen interactions. Comparative genomics data is ideal for identifying genes that are affected by selection pressure (Chen et al. 2006).

Here, we scanned positive selection in genome wide manure by employing a comparative genomics approach. By comparing the genes and their corresponding proteins of different organisms, the rate of synonymous and non-synonymous substitution (pressure of selection) could be identified. The statistics used to identify positive selection in these studies is the ratio of non-synonymous to synonymous substitution rate, ω (d_N/d_S).(Hurst 2002, Yang and Bielawski 2000) Our particular interest was in identifying positive selection because it provides evidence for adaptive changes in function. Moreover, we further aimed to investigate the selection pressure on different functional pathways of *V. cholerae* and to perform clustering statistics to find out highly selected genes that have evolved only in *V. cholerae* genome (species-specific evolution).

METHODS AND MATERIALS

Construction of primary data files with orthologous DNA sequences

Genome sequences of four *V. cholerae* strains and six other *Vibrio* sp. were retrieved from the NCBI FTP server (ftp://ftp.ncbi.nih.gov). All protein coding annotated gene sequences within ten selected species were blasted using reciprocal blast best hit method (Moreno-Hagelsieb and Latimer 2008) (Table-1). The hits with an E-value cut off of 10^{-6} and minimum query coverage was set to 75%

to be considered as positives and subsequently selected for further analysis. A meta-database was constructed by incorporating the positive hits representing the annotated genes from ten selected genomes (Table-1), excluding hypothetical, putative and predicted genes. When more than one gene sequence from a particular genome was found (gene duplication events could possibly account for such phenomena), only one sequence was chosen based on their query coverage and E-value. A total of 886 homologous gene sets from 10 *Vibrio* sp. were obtained as primary datasets. Bioedit tool was used for local blast and meta-database construction during homologous gene search (Hall 1999). The evolutionary relationship among the *Vibrio* sp. was established by analyzing 16srRNA phylogenetic tree (Figure 1).

Table-1: List of Vibrio Strains used in the analysis

Starin	RefSeq Accession
V. cholerae O395	NC_009456, NC_009457
V. cholerae O1	NC_002505, NC_002506
V. cholerae MJ	NC_012667, NC_012668
V. cholerae M66	NC_012578, NC_012580
V. splendidus LGP32	NC_011744, NC_011753
V. vulnificus CMCP6	NC_004459, NC_004460
V. vulnificus YJ016	NC_005128, NC_005139, NC_005140
V. fischeri ES114	NC_006840, NC_006841
V. harveyi ATCC	NC_009777, NC_009783
V. parahaemolyticus RIMD	NC_004603, NC_004605

Detection of Paralogous and Xenologous genes

Since the presence of paralogous and xenologous genes in the primary dataset could potentially interfere with the detection of positive selection, recombination branch point test was used to detect the recombination and gene duplication events using Recombination detection program (RDPversion3).(Martin and Rybicki 2000) In brief, RDP3 examines nucleotide sequence alignments and attempts to identify recombination breakpoints using ten published recombination detection algorithms, including RDP, geneconv and chimaera that we employed in this analysis.(Martin and Rybicki 2000, Padidam et al. 1999, Posada and Crandall 2001) The genes showing evidence of significant recombination in all three methods (P≤ 0.01) were considered as paralogue and removed from further analyses. Comparative phylogenetic method was used to detect horizontally transferred genes. In short, 16s rRNA tree was compared with species tree using T-Rex tool (Li et al. 2005). Genes showing significant evidence (bootstrap score) of Lateral gene transfer were removed from further analysis(Boc and Makarenkov 2003). After removing paralogoues and xenologoues, remaining genes were classified according to their COG categories found in NCBI genome annotation Report(Tatusov et al. 2000).

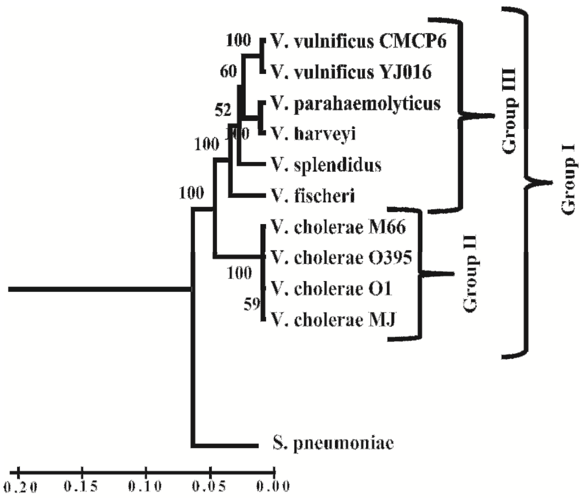

Figure 1. ***Species Tree of Vibrio cholerae Using 16s rRNA Sequence Alignment.*** *Sequence of 16s rRNA of O395 strain was retrieved from NCBI nucleotide database by data mining, then the sequence was used in to find closely related species. 10 closely related strains were under a single common ancestor were found and selected for analysis.*

Test for Positive selection

Ten selected *Vibrio* spps. were distributed into three different groups according to their host and habitat -Group I: all species, Group II: All *V.*

cholerae (VC) –M66, 0395, O1 and MJ, Group III: all non-*cholerae* Vibrio sp. (Supplementary Table 1). For each group a total of 422 orthologous gene sets were subjected to BLASTx and their corresponding amino acid sequence alignments were used to create the codon wise alignment of each orthologous gene by online tools: Revtrans 1.4 and Pal2Nal (Suyama et al. 2006, Wernersson and Pedersen 2003). HyPhy as implemented in MEGA 5.03 software package was used to detect codon wise nonsynonymous/synonymous substitution ratio (ω) (Tamura et al. 2011). Briefly, it involves the estimation of synonymous (S) and nonsyonymous (N) sites using the joint Maximum Likelihood reconstructions of ancestral states under a Muse-Gaut model (Muse and Gaut 1994) of codon substitution and Felsenstein 1981 model (Felsenstein 1981) of nucleotide substitution. A positive value for the test statistic indicates an overabundance of nonsynonymous substitutions. The probability of rejecting the null hypothesis of neutral evolution was also calculated (Kosakovsky Pond and Frost 2005, Suzuki and Gojobori 1999) and the P-value less than 0.05 were considered significant. Additionally, the normalized dN-dS values were also obtained using the total number of substitutions in the alignment (measured in expected substitutions per site). The entire orthologous gene sets and their corresponding codon alignments of other two groups were used to determine dN/dS ratio (ω) in each group.

GO enrichment score

GO enrichment score was calculated to identify the enriched GO terms in the positively selected genes in all four groups. To calculate enrichment score target list of positively genes was compared to the background set of 422 orthologous gene set via a standard approach that utilize hypergeometric distribution.(Sealfon et al. 2006) GO enrichment score was calculated by using the following equation

$$Prob(X \geq b) = HCT(b; N, B, n) = \sum_{i=0}^{min(n,N)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}}$$

Given a total number of genes N, with B of these genes associated with a particular GO term and n of these genes in the target set, then the probability that b or more genes from the target set are associated with the given GO term is given by the hyper geometric tail.

RESULT

Identifying highly selected genes using K-means unsupervised clustering method

Genes were classified as high, moderate and low using K-means unsupervised clustering method (J. A. Hartigan 1979, Kimura 1983, MacQueen 1967) according to their number of positively selected codons. 92 genes were found belonging to ‘high positive selection’ category as they harbor relatively high percentages of positively selected sites (%PS). When compared with the %PS of House-keeping genes, these 92 genes showed significant higher substitution rate (Supplementary Figure 1) suggesting that the genes with higher rate of selection are truly under influence of natural selection. These highly positively selected 92 genes were found belonging to 14 COG categories among the total 18 COG categories representing the core genome (Figure 2).

Core genome is largely under purifying selection while only a small portion is under positive selection.

Positive selection reflecting the traces of adaptive evolutionary process is determined by calculating the ration (ω) between non- synonymous (dN) and synonymous (dS) codon substitution rates. The ratio ω (dN/dS) > 1 refers to positive selection whereas ω=1 indicates purifying selection.

There are three types of evolutionary selection process in terms of codon substitution: positive selection, purifying selection and neutral selection (Hahn 2008, Montoya-Burgos 2011, Yang and

Bielawski 2000). Individual codons were analyzed in each of the 422 genes to observe the frequency distribution of these orthologous genes under the three types of selection process. For majority of the genes 75%-90% codons were found to be under purifying selection (Figure 3a). In case of neutral selection most number of genes harbor 10%-35% of neutrally selected codons (Figure 3b) whereas highest number of genes were found with only 4%-10% positively selected codons (Figure 3c). The highest percentage of positively selected codons was found to be 38%. These data suggested that purifying selection is the most favorable selection process, probably due to the codon adaptation or random mutation events (Kimura 1983). However, positive selection occurring only in a minor portion of gene is specific and reflects reminiscent of evolutionary pressure.

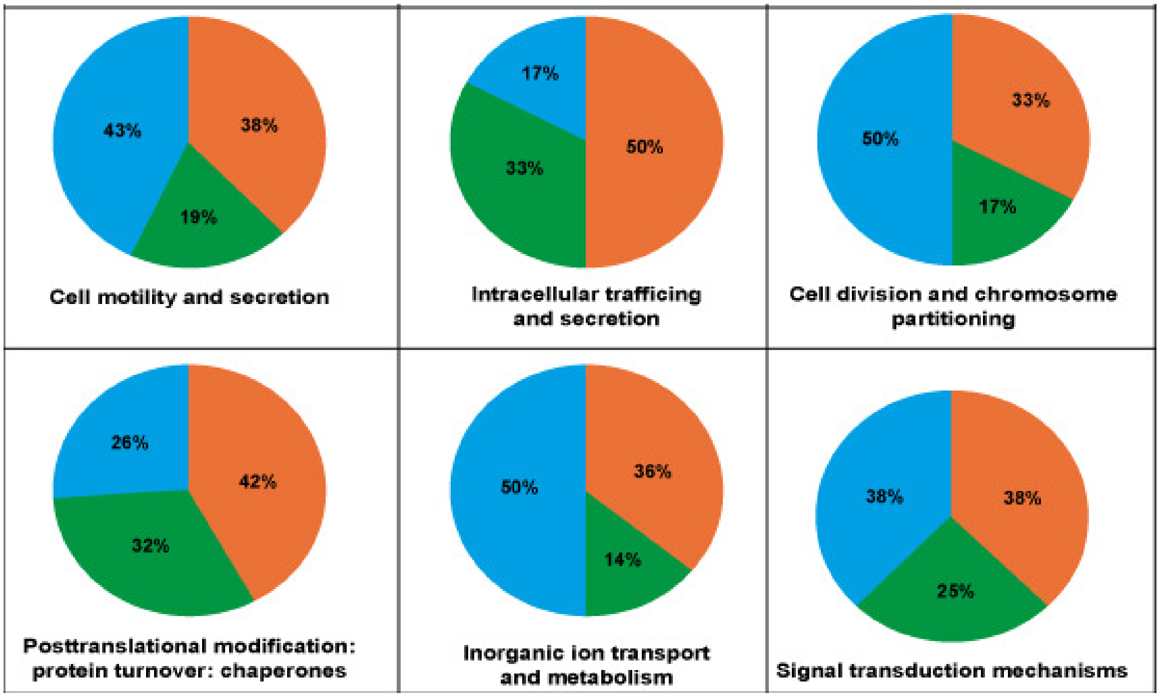

Figure 2. Classification of Orthologous genes according to COG category. *NCBI COG database was used to identify the COG category of every gene included as orthologous. All of the Vibrio cholerae genes were fallen into eighteen categories.*

Eight Cluster of Orthologous Genes(COG) pathway categories are enriched with highly selected genes.

422 genes were found as orthologous representing the core genome of the ten selected *Vibrio* sp. These genes were then classified into 18 categories according to their COG annotation as found in NCBI genome annotation Report (Tatusov et al. 2000). Distribution of these orthologous genes into different functional pathways is shown in a pie chart (Figure 2). Eight categories, namely- 1) amino acid transport and metabolism 2) translation and ribosomal biogenesis 3) carbohydrate transport and metabolism 4) energy production and conversion 5) DNA replication/recombination and repair 6) nucleotide transport and metabolism 7) transcription and 8) coenzyme metabolism, comprise 70% of the core genome where the first two categories (amino acid transport and metabolism, and transcription and ribosomal biogenesis) with highest number of genes representing 13% and 12% of the core genome respectively. Among the remaining categories cell motility and secretion (5%), Inorganic ion transport (4%) and Cell envelope and biogenesis (4%) were prominent.

There are three types of evolutionary selection process in terms of codon substitution: positive selection, purifying selection and neutral selection (Hahn 2008, Montoya-Burgos 2011, Yang and Bielawski 2000). Individual codons were analyzed in each of the 422 genes to observe the frequency distribution of these orthologous genes under the three types of selection process. For majority of the genes 75%-90% codons were found to be under purifying selection (Figure 3a). In case of neutral selection most number of genes harbor 10%-35% of neutrally selected codons (Figure 3b) whereas highest number of genes were found with only 4%-10% positively selected codons (Figure 3c). The highest percentage of positively selected codons was found to be 38%. These data suggested that purifying selection is the most favorable selection process, probably due to the codon adaptation or random mutation events (Kimura 1983). However, positive selection occurring only in a minor portion of gene is specific and reflects reminiscent of evolutionary pressure.

Eight Cluster of Orthologous Genes(COG) pathway categories are enriched with highly selected genes.

422 genes were found as orthologous representing the core genome of the ten selected *Vibrio* sp. These genes were then classified into 18 categories according to their COG annotation as found in NCBI genome annotation Report (Tatusov et al. 2000). Distribution of these orthologous genes into different functional pathways is shown in a pie chart (Figure 2). Eight categories, namely- 1) amino acid transport and metabolism 2) translation and ribosomal biogenesis 3) carbohydrate transport and metabolism 4) energy production and conversion 5) DNA replication/recombination and repair 6) nucleotide transport and metabolism 7) transcription and 8) coenzyme metabolism, comprise 70% of the core genome where the first two categories (amino acid transport and metabolism, and transcription and ribosomal biogenesis) with highest number of genes representing 13% and 12% of the core genome respectively. Among the remaining categories cell motility and secretion (5%), Inorganic ion transport (4%) and Cell envelope and biogenesis (4%) were prominent.

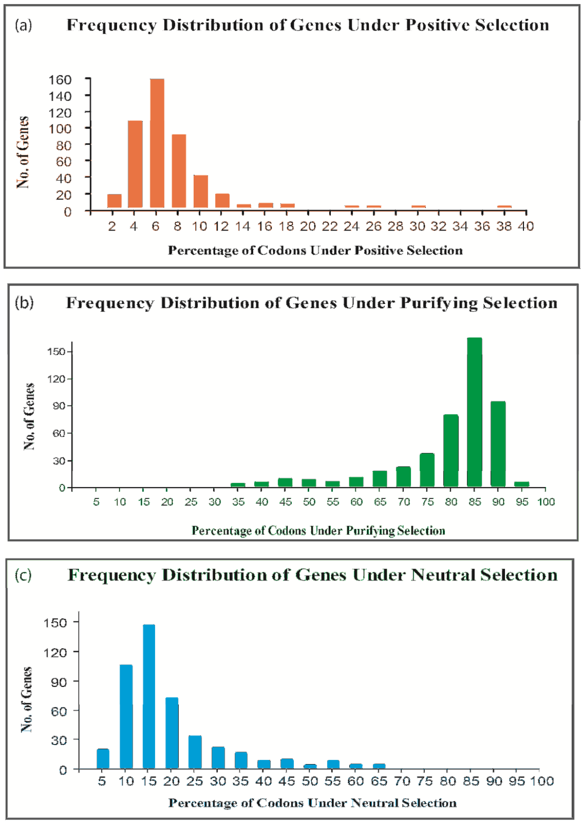

Figure 3. Frequency distribution of genes under three different type of codon selection. *Percentage of codons under positive, purifying and neutral selection were calculated and bar diagrams were build to indentify the distribution of genes under these selection process. Red, Green and blue bars are indicating Positive selection, Purifying selection and Neutral selection respectively.*

GO enrichment score provided the most important pathway to look insights for codon wise selection

To identify the functional gene categories enriched with 92 positively selected genes a GO enrichment scoring approach described by Eden et al published in BMC bioinformatics, 2009 was applied. In order to calculate enrichment score, GO categories of all of the 422 genes of core genome were used as a background dataset and an equation described by Sealfon *et. al.* was used. The pathways were then ranked based on the P- values and GO enrichment scores. Six functional pathways showing low P- value and higher GO enrichment score were considered as pathways with highly enriched positively selected genes (Figure 4). These categories are 1) cell motility and secretion 2) intracellular trafficking 3) post-translational modifications 4) inorganic ion transport and metabolism 5) signal transduction mechanism and 6) cell division and chromosome partitioning. Remaining other twelve categories harbor mostly low and moderately selected genes therefore not included in the analysis.

To understand the functional constraints' on the genes of flagellar assembly, functional domains were indentified with their corresponding protein sequence by InterProScan Sequence Search online tool.(Mulder and Apweiler 2007) Positions of these domains and motifs were indentified in the codon alignments with their corresponding selection rate (Figure 5). In few cases the domains are in between high selection rate (i.e. flgE, flgH, flgD, fliF) suggesting slow evolution of these functional domains. Therefore, it seemed that though the genes for flagellar assembly are in strong evolutionary pressure but their function is essential for the survival of *Vibrio cholerae*. However, it is one of the limitations for computational substitution approach that we cannot always correlate function of the gene the corresponding codon substitutions.

Identifying Vibrio cholerae specific positively selected genes using data map

In order to investigate the positively selected genes those played critical role in determining the fitness only selected in the *Vibrio cholerae* sp., three different groups described earlier and indicated in Figure 1 were assessed for their orthologous genes using dN/dS ratio. These three groups are – 1) Group I: All ten *Vibrio* species 2) Group II: *Vibrio cholerae* only 3) Group III: Non-*cholerae* Vibrio sp. The aim of the classification was to determine the genes that were under positive selection only in Vibrio sp. By excluding the positively selected genes found within Vibrio cholereae and and non-cholerae Vibrio species from the core positively selected gene set harboring 92 genes, we can identify the genes that undergone selection pressure during speciation of *Vibrio cholerae*.

Using the classification status of these three groups a data-map was built and the genes that were found to be under high positive selection in group I but not in group II and III were included as positive. 49 genes were found through the analysis (Table 2), where 12 of these genes were associated with the six GO categories that were enriched with highly selected genes. Three selected genes in flagellar assembly namely FlgE, FliF and MotB were found to be critical for speciation *Vibrio cholerae* from Vibrio co-ancestor.

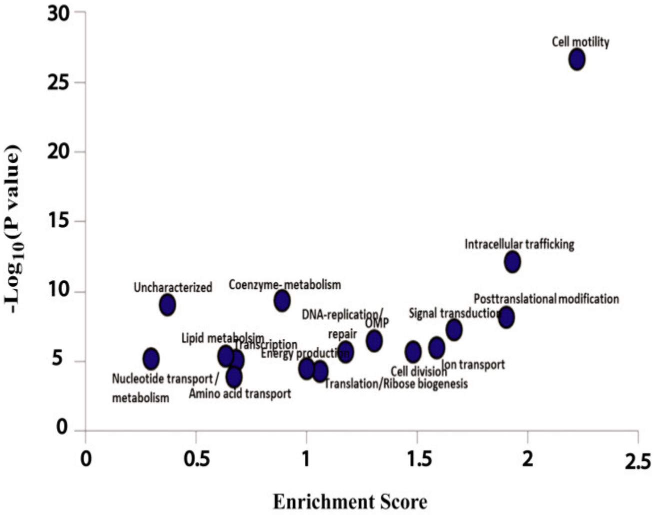

Figure 4: GO enrichment score of different pathways of Vibrio cholerae according to COG categories. *A GO enrichment score were calculated using the described algorithm and assigned for each category. Pathways or COG categories with higher enrichment score and lower P-value were thought to be evolving with higher selection rate.*

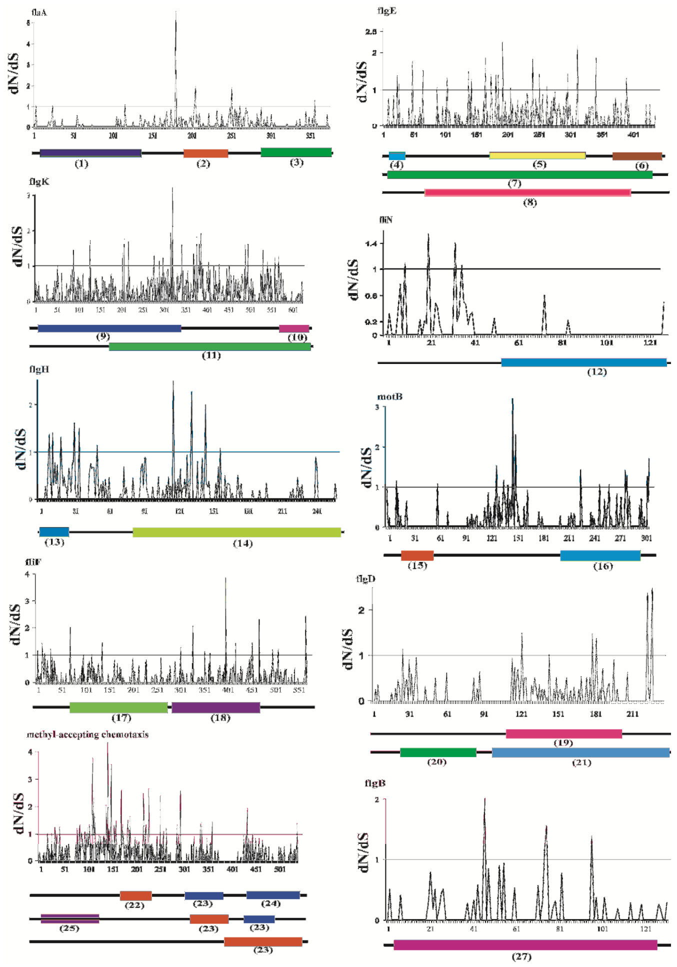

Figure 5. Individual substitution rate of codons with their corresponding domain position. *Corresponding protein sequence of codon alignment were used in INTERPRO scanner to find out their domains with positions. The labeling of the domains used in this figure is following :- (1) Bacterial flagellin N-terminal helical region[Flagellin_N (PF00669)] (2) Flagellin hook IN motif [Flagellin_IN (PF07196)] (3) Bacterial flagellin C-terminal helical region [Flagellin_C (PF00700)] (4) Flg_bb_rod (PF00460) (5) Flagellar basal body protein FlaE(PF07559) (6) Flagellar basal body rod FlgEFG protein C-terminal (7) FlgEFG_subfam(TIGR03506) (8) Flagellar hook protein flgE superfamily (9) flgK_ends(TIGR02492) (10) Flagellar basal body rod FlgEFG protein C-terminal [Flg_bbr_C (PF06429)] (11) Phase 1 flagellin superfamily(SSF64518) (12) Surface presentation of antigens (SPOA) (13) PROKAR_LIPOPROTEIN (PS51257) (14) FlgH (PF02107) (15) Membrane MotB of proton-channel complex MotA/MotB [MotB_plug (PF13677)] (16) OmpA domain [OmpA (PF00691)] (17) Secretory protein of YscJ/FliF family[YscJ_FliF (PF01514)] (18) Flagellar M-ring protein C-terminal [YscJ_FliF_C (PF08345)] (19) FlgD Tudor-like domain(PF13861) (20) Flagellar hook capping protein - N-terminal region (PF03963) (21) FlgD Ig-like domain(PF13860)(20) Flagellar hook capping protein - N-terminal region (PF03963) (22) Cache domain [Cache_1 (PF02743)](23) HAMP (PF00672,PS50885,SM00304) (23) HAMP (PF00672, PS50885, SM00304) (24) Methyl-accepting chemotaxis protein (MCP) signalling domain [MCPsignal (PF00015)] (25) Vibrio chemotaxis protein N terminus [MCP_N (PF05581)] (26) Methyl-accepting chemotaxis-like domains (chemotaxis sensory transducer) [MA (SM00283)] (27) Flagella basal body rod protein superfamily (IPR006300).*

DISCUSSION

A total of 92 genes were found under high positive selection out of the core genome of selected Vibrio sp. consisting of 424 orthologous genes. GO enrichment analysis showed that six pathways composed of 78 orthologous genes from the core genome are particularly enriched with highly positively selected genes. 31 genes highly positively selected genes belong to these GO categories out of 78 genes. Among the GO categories 'Cell motility and secretion' is the most enriched with high positively selected genes followed by intracellular trafficking as shown by the enrichment score and P value. Out of 20 orthologous genes belonging to Cell motility pathway ten genes were found to be under high positive selection.

Ten high positively selected genes out total 20 orthologous genes associated with Cell motility facilitates this category to emerge as the most enriched one with a log(P) value of -26. When the positively selected genes of the three different groups (Group I: all sp, Group II: Only VC strain, Group III: Non cholerae Vibrio sp) were compared 41 genes out of the 92 genes were found to only 31 positively selected genes of the six pathways (Supplementary Table 2) and 46 highly selected genes only in *Vibrio cholerae* species (Table 2) were found in this study. The selected genes are from different categories, performed various functions and lies in the different portions of the genome. Thus it is concluded that positive selection in *Vibrio cholerae* species is a genome wide phenomena. A possible explanation for this nearly genome-wide positive selection pressure, as well as its even distribution across the functional elements of the genome, may be the result of an evolutionary arms race, or macro-evolutionary version of the Red Queen Hypothesis, between competing species within the mammalian and/or vertebrate gastrointestinal tract.(Clay and Kover 1996) This habitat is known to harbor vast species diversity (Frank and Pace 2008, Ley et al. 2008), and thus competition will constantly exist between species for resources. According to the Red Queen Hypothesis, species involved in competition for resources can maintain their fitness relative to other competing species only by improving their specific fitness, and this could ultimately lead to extensive levels of positive selection signature across the genome (Clay and Kover 1996, Lefebure and Stanhope 2009). Every gene that has been found as Orthologous was annotated according to NCBI COG database and 18 pathways had been found containing all Orthologous genes of *Vibrio cholerae* O395. After estimating dN/dS for each orthologous gene sets, 92 genes of 16 different pathways were found to be highly selected and six pathways were found to be highly selected in vibrio genome (Figure 3). These pathways are (1) Intracellular trafficking and secretion (2) Posttranslational modification, protein turnover, chaperones (3) Cell motility and secretion (4) Signal transduction mechanisms (5) Inorganic ion transport and metabolism and (6) Cell division and chromosome partitioning.

Table 2. Genes that showed evidence of Positive selection

Gene Symbol	% PS[€]	COG[¥]	GO Term
MotA/TolQ/ExbbB	14.19	COG0811U	Intracellular trafficking and secretion
VC0395_0070	14.02	COG0840NT	Signal Transduction
mshH	12.39	COG2200T	Signal Transduction
VC0395_A1807	11.86	COG2217P	Inorganic ion transport and metabolism
FkpA	11.72	COG0545O	Post-translational modification, protein turnover, chaperone functions
CcmF	11.18	COG1138O	Post-translational modification, protein turnover, chaperone functions
FliN	10.85	COG1886NU	Cell motility and secretion
Tig	10.49	COG0544O	Post-translational modification, protein turnover, chaperone functions
flgB	9.92	COG1815N	Cell motility and secretion
PstS	9.89	COG0226P	Inorganic ion transport and metabolism
MotB*	9.77	COG1360N	Cell motility and secretion
FlaA	9.33	COG1344N	Cell motility and secretion
flgE*	9.26	COG1749N	Cell motility and secretion

flgD	8.94	COG1843N	Cell motility and secretion
glnE	8.9	COG1391OT	Signal Transduction
FexB	8.49	COG0642T	Signal Transduction
SppA*	8.28	COG0616OU	Post-translational modification, protein turnover, chaperone functions
ptsP*	8.17	COG3605T	Signal Transduction
metN*	8.14	COG1135P	Inorganic ion transport and metabolism
SmpB*	8.07	COG0691O	Post-translational modification, protein turnover, chaperone functions
mukB*	8.02	COG3096D	Cell division and chromosome partitioning
SecF*	7.94	COG0341U	Intracellular trafficing and secretion
flgK	7.85	COG1256N	Cell motility and secretion
lepB*	7.72	COG0681U	Intracellular trafficing and secretion
HscB	7.6	COG1076O	Post-translational modification, protein turnover, chaperone functions
VC0395_0557	7.45	COG0229O	Post-translational modification, protein turnover, chaperone functions
ftsZ*	7.42	COG0206D	Cell division and chromosome partitioning
cysJ*	7.37	COG0369P	Inorganic ion transport and metabolism
FliF*	7.34	COG1766N	Cell motility and secretion
metQ	7.06	COG1464P	Inorganic ion transport and metabolism
flgH	7.00	COG2063N	Cell motility and secretion

Note: COG categories were retrieved from table of Protein Details for Vibrio cholerae O395 using NCBI whole genome search. (* Genes that are selected only in *Vibrio cholerae* genomes; € %PS : Percentage of Positively Selected sites; ¥ COG : Cluster of Orthologous genes)

However, the most interesting observation came out from our study is the positive selection event of cell motility and secretion pathway. Evidently genes involved in flagellar assembly are under severe influence of natural adaptation in vibrio genome as they play roles in different bacterial process. Besides motility, flagella are also involved in chemotaxis and signaling mechanism. It is the outer membrane portion and directly interacts with host during adherence with intestinal epithelial cells. (Attridge and Rowley 1983) It was hypothesized earlier that, as motility is one of the important virulent factor for *Vibrio cholerae* thus host will try to mimic the activity of this system and proteins associated with motility will be important target for natural selection. Our study clearly reveals that the whole flagellar assembly is under strong evolutionary pressure and gene associated with this system is evolving together with different selection rate (Figure 6).

CONCLUSION

There is a possibility that we have missed some genes during this process as many of them were discarded due to the recombination event or evidence for Horizontal transfer. False positive or false negative results were no assessed here with alternative approach as we considered substitution rate of every codon site and finally the percentage of positively selected sites. Although highly selected genes were compared with house-keeping genes for their selection rate, which clearly indicates the validity of our method but still there is a chance that we have missed some genes or included some false positives genes as positively selected genes.

Among 18 COG categories only six were found to be under severe influence of evolutionary pressure. In brief, genes that are selected in vibrio genome are mostly involved in motility, chemotaxis, ion or nutrition transport, signaling, protein modification, multiplication of cell and intracellular trafficking. All these evidence suggest that genes that come with direct host-pathogen interaction, involved with nutritional status or environmental adaptation are under evolutionary pressure and shows high rate of natural selection. The results also clearly support Red Queen hypothesis.

We also concluded that in Vibrio Spp. Natural selection is a genome wide phenomena and purifying selection is the dominant selection Process.

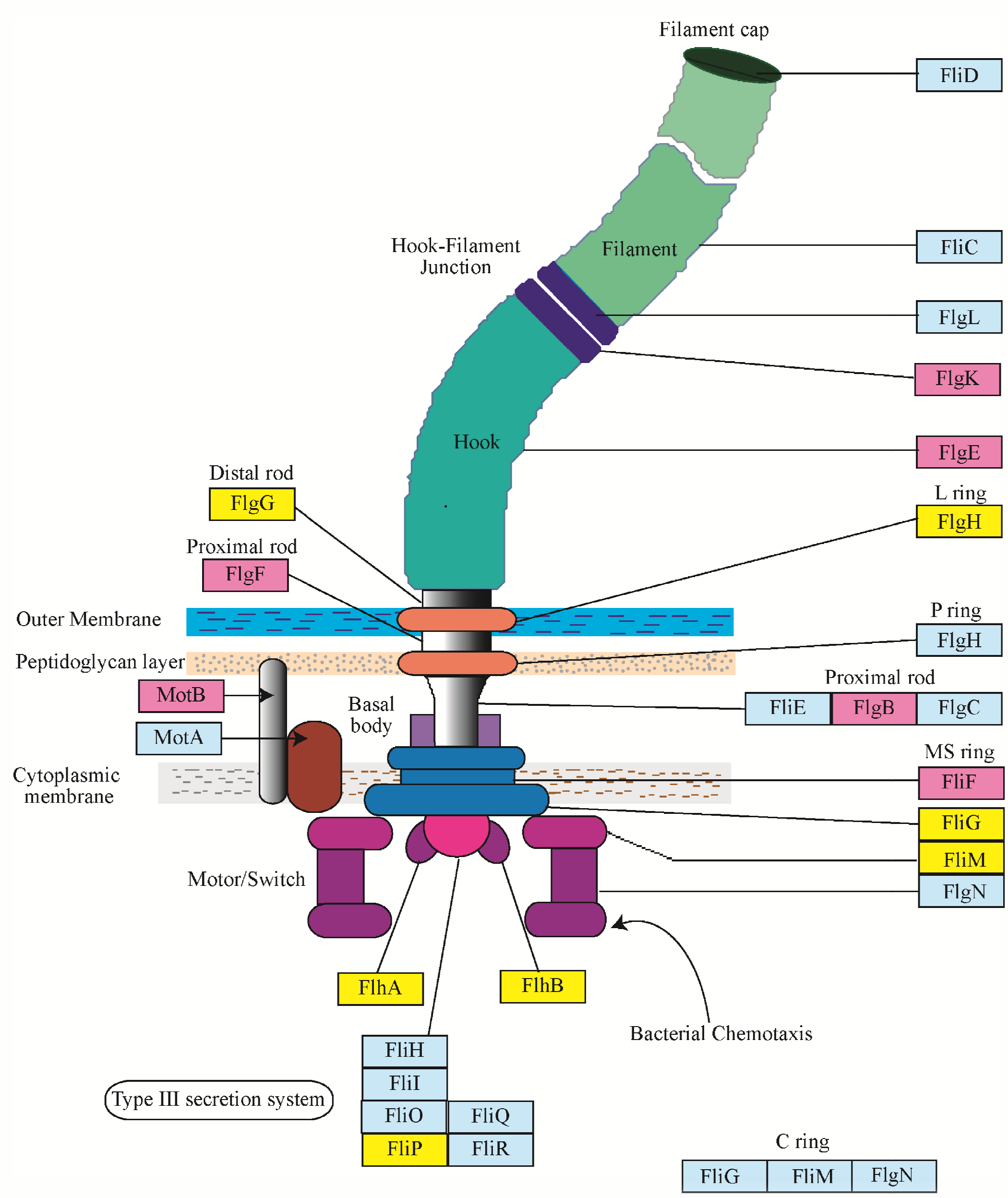

Figure 6. Genes involved in Flagellar Assembly. *A flagellar proteins assembly is shown here with their corresponding genes. Genes highlighted with Red and Yellow is highly & moderately selected respectively. Both inner membrane and outer membrane proteins coding genes are evolving with higher selection rate.*

ACKNOWLEDGEMENT

We thank Mahmuda Khatun and Hadisur Rahman for their help during data mining.

REFERENCE

1. Attridge SR, Rowley D. 1983. The Role of the Flagellum in the Adherence of Vibrio cholerae. Journal of Infectious Diseases 147: 864-872.
2. BERRY MNGALJ. 1975. Motility as a Virulence Factor for Vibrio cholerae. INFECTiON AND IMMUNITY 11: 890-897.
3. Boc A, Makarenkov V. 2003. New Efficient Algorithm for Detection of Horizontal Gene Transfer Events. Pages 190-201 in Benson G, Page RM, eds. Algorithms in Bioinformatics, vol. 2812 Springer Berlin Heidelberg.
4. Chen SL, et al. 2006. Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. Proc Natl Acad Sci U S A 103: 5977-5982.
5. Clay K, Kover PX. 1996. The Red Queen Hypothesis and plant/pathogen interactions. Annu Rev Phytopathol 34: 29-50.
6. Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368-376.
7. Frank DN, Pace NR. 2008. Gastrointestinal microbiology enters the metagenomics era. Curr Opin Gastroenterol 24: 4-10.
8. Hahn MW. 2008. Toward a selection theory of molecular evolution. Evolution 62: 255-265.
9. Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95-98.
10. Hammer BK, Bassler BL. 2003. Quorum sensing controls biofilm formation in Vibrio cholerae. Mol Microbiol 50: 101-104.
11. Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18: 486.
12. J. A. Hartigan MAW. 1979. A K-Means Clustering Algorithm. Applied Statistics 28: 100--108.
13. Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
14. Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208-1222.
15. Lefebure T, Stanhope MJ. 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus Campylobacter. Genome Res 19: 1224-1232.
16. Ley RE, et al. 2008. Evolution of mammals and their gut microbes. Science 320: 1647-1651.
17. Li Z, Wang L, Zhong Y. 2005. Detecting horizontal gene transfer with T-REX and RHOM programs. Brief Bioinform 6: 394-401.
18. MacQueen JB. 1967. Some Methods for Classification and Analysis of MultiVariate Observations. Pages 281-297 in Cam LML, Neyman J, eds. Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability: University of California Press.
19. Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16: 562-563.
20. Montoya-Burgos JI. 2011. Patterns of Positive Selection and Neutral Evolution in the Protein-Coding Genes of <italic>Tetraodon</italic> and <italic>Takifugu</italic>. PLoS ONE 6: e24800.
21. Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24: 319-324.
22. Mulder N, Apweiler R. 2007. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396: 59-70.
23. Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715-724.
24. Nandi T, et al. 2010. A genomic survey of

positive selection in Burkholderia pseudomallei provides insights into the evolution of accidental virulence. PLoS Pathog 6: e1000845.

25. Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265: 218-225.
26. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in Escherichia coli. Genome Res 17: 1336-1343.
27. Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A 98: 13757-13762.
28. Sealfon RS, Hibbs MA, Huttenhower C, Myers CL, Troyanskaya OG. 2006. GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. BMC Bioinformatics 7: 443.
29. Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34: W609-612.
30. Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16: 1315-1328.
31. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.
32. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33-36.
33. Wernersson R, Pedersen AG. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 31: 3537-3539.
34. Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496-503.