Protoplasma (2011) 248:799–804
DOI 10.1007/s00709-010-0255-0

ORIGINAL ARTICLE

In silico prediction of drug targets in *Vibrio cholerae*

Pramod Katara · Atul Grover · Himani Kuntal · Vinay Sharma

Received: 26 September 2010 / Accepted: 7 December 2010 / Published online: 21 December 2010

Abstract Identification of potential drug targets is the first step in the process of modern drug discovery, subjected to their validation and drug development. Whole genome sequences of a number of organisms allow prediction of potential drug targets using sequence comparison approaches. Here, we present a subtractive approach exploiting the knowledge of global gene expression along with sequence comparisons to predict the potential drug targets more efficiently. Based on the knowledge of 155 known virulence and their coexpressed genes mined from microarray database in the public domain, 357 coexpressed probable virulence genes for *Vibrio cholerae* were predicted. Based on screening of Database of Essential Genes using blastn, a total of 102 genes out of these 357 were enlisted as vitally essential genes, and hence good putative drug targets. As the effective drug target is a protein which is only present in the pathogen, similarity search of these 102 essential genes against human genome sequence led to subtraction of 66 genes, thus leaving behind a subset of 36 genes whose products have been called as potential drug targets. The gene ontology analysis using Blast2GO of these 36 genes revealed their roles in important metabolic pathways of *V. cholerae* or on the surface of the pathogen. Thus, we propose that the products of these genes be evaluated as target sites of drugs against *V. cholerae* in future investigations.

Keywords *Vibrio cholerae* · Virulence genes · Essential genes · Drug targets · Coexpressed genes

Handling Editor: Reimer Stick

Pramod Katara and Atul Grover contributed equally to the manuscript.

P. Katara (✉) · A. Grover · H. Kuntal · V. Sharma
Department of Bioscience and Biotechnology,
Banasthali University,
Banasthali 304022, India
e-mail: pmkatara@gmail.com

A. Grover
e-mail: iatulgrover@gmail.com

H. Kuntal
e-mail: himanikuntal@gmail.com

V. Sharma
e-mail: vinaysharma30@yahoo.co.uk

Present Address:
A. Grover
Defence Institute of Bio Energy Research,
Defence Research and Development Organization,
Ministry of Defence,
Haldwani 263139, India

Introduction

Predicting virulence of bacterial pathogens and their ability to cause diseases is necessary for microbial risk assessment. Information on virulence is important for the analysis of qualitative and quantitative description of health outcomes. Thus, identification and characterization of virulence genes from whole genome sequences have become an important thrust area in recent years with an ultimate aim of identifying novel drug targets especially in the light of pathogens fast acquiring drug resistance (Hasan et al. 2006).

The experimental approach of considering gene function relating to an organism's pathogenicity has its limits (Kuruvilla et al. 2002). Bioinformatics analysis can reveal virulence potential of a genome-sequenced strain (Garg and Gupta 2008). A gene's contribution to phenotype is determined by the context of other genes present in the genome (Groth et al. 2008). Integration of expression data into sequence-based comparative analyses could thus

potentially provide new insights into the relation between genomic sequence and its function. Thus, if the gene expression patterns and associated gene networks are understood, we can better predict genes coding for virulence (Sridhar et al. 2007). It may even be possible to identify bacterial species that are not yet pathogenic but have the correct genetic repertoire to become so if particular genes or gene functions were acquired (Knell 2006). Gene expression pattern and network identification may thus become an important component for identification and characterization of microbial hazards, including emerging pathogens, in the context of microbial risk assessment. The availability of microarray data for several bacterial organisms (Hubble et al. 2009) and the completion of the human genome project have evolved the field of host–pathogen interactions and the field of drug discovery against threatening human pathogens. Those genes that share similar gene expression patterns are assumed to be similar at functional and metabolic level, play same type of metabolic function, or involve in similar metabolic processes (van-Noort et al. 2003; Bergmann et al. 2004). In fact, computational methods are already in place to predict the role of the products of different genes as drug targets (Gray and Keck 1999; Sakharkar et al. 2004; Li and Lai 2007; Perumal et al. 2007; Sakharkar et al. 2008). The present study aims to find new drug targets in *Vibrio cholerae* O1, the causative agent of cholera, considering that drug targets must be essential for the growth and viability of the pathogen and highly selective against the pathogen with respect to the human host (Galperin and Koonin 1999). The use of such assumptions in predicting a drug target in itself is novel, which has been tested in *V. cholerae*, a Gram-negative human pathogen, consisting of two circular chromosomes of sizes 2,961,146 bp and 1,072,314 bp. *V. cholerae* O1 subgroup in particular has displayed dynamism in its incidence pattern in Indian subcontinent (Sur et al. 2006).

Methods

Resources

Virulence genes available in public domain were cataloged from the published reports (Higgins et al. 1992; Lee et al. 1999; Camara et al. 2002; Zhu et al. 2002) and using available virulence factor database (VFDB; http://www.mgc.ac.cn/VFs/; Yang et al. 2008). The gene sequences of the virulence genes were downloaded from GenBank database of NCBI (ftp://ftp.ncbi.nlm.nih.gov). Microarray data (cDNA) for gene expression pattern analysis were downloaded from Stanford Microarray Database (SMD; http://genome-www5.stanford.edu/). A list of essential genes of the pathogen was obtained from database of essential genes (DEG; http://tubic.tju.edu.cn/deg/; Zhang and Lin 2009).

Raw data processing and clustering

To predict virulence genes, we used gene expression data from different experiments available at SMD. Data corresponding to the control sequences and the open reading frames (ORFs) whose expression values across the time points and various conditions had mean lower than 25% were filtered out from the analysis. Analogous filtering schemes included accepting only those ORFs that show at least a twofold change in expression level (Tamayo et al. 1999; Ulm et al. 2004). These ORFs were clustered based on their gene expression using K-mean clustering algorithm through the cluster tool (Eisen et al. 1998). As a rule of thumb, $K=500$ for genes was used. Clusters sharing at least one well-documented virulence gene were selected for further analysis, with all participating genes considered as probable virulence genes.

Identification of drug targets

These selected genes were subjected to BLASTX (Altschul et al. 1997) against the Database of Essential Genes. A random expectation value (*E* value) cutoff of 0.001 and a minimum bit-score cutoff of 100 were used as the baseline to identify the essential genes in the pathogen. These essential genes belonging to the pathogen were subjected to BLASTX against the human genome in the NCBI server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The homologs were excluded, and the lists of nonhomologs were compiled. The finally selected genes were further analyzed using Blast2GO (Conesa et al. 2005) to predict the potential targets, i.e., surface proteins and participants of important metabolic pathways (Overington et al. 2006).

Results and discussion

The present pharmaceutical scenario is under constant stress of discovering new antimicrobials due to the threat of resistance rapidly being developed in target microbes. Identification of microbe-specific proteins for directing drug discovery and to designing new drugs to previously known targets are the two popular means to combat this resistance. Modern tools of computational biology greatly enhance the speed and reliability of antimicrobial discovery. With an objective of identifying proteins potentially useful as drug targets, we have relied on the use of genomic data and a subtractive genomic approach. The results obtained

Table 1 Putatively selected drug targets after eliminating those vitally essential virulence genes that shared homology to human genes

Genes	Function
VC0051	Purine ribonucleotide biosynthesis
VC0061	Biosynthesis of cofactors, prosthetic groups, and carriers: thiamine
VC0215	Biosynthesis of cofactors, prosthetic groups, and carriers: pantothenate and coenzyme A
VC0233	Cell envelope: biosynthesis and degradation of surface polysaccharides and lipopolysaccharides
VC0293	Protein synthesis: ribosomal proteins: synthesis and modification
VC0307	Transcription: transcription factors
VC0315	Fatty acid and phospholipid metabolism: biosynthesis
VC0405	Cellular processes: pathogenesis
VC0438	Hypothetical: conserved
VC0467	Hypothetical: conserved
VC0786	Energy metabolism: amino acids and amines
VC0961	Hypothetical: conserved
VC0971	DNA metabolism: DNA replication, recombination, and repair
VC1138	Amino acid biosynthesis: histidine family
VC1174	Amino acid biosynthesis: aromatic amino acid family
VC1325	Transport and binding proteins: Carbohydrates, organic alcohols, and acids
VC1422	Transport and binding proteins: amino acids, peptides, and amines
VC1680	Transport and binding proteins: amino acids, peptides, and amines
VC1701	Hypothetical: conserved
VC1721	Regulatory functions
VC1847	DNA metabolism: DNA replication, recombination, and repair
VC1886	DNA metabolism: DNA replication, recombination, and repair
VC1911	Purines, pyrimidines, nucleosides, and nucleotides: pyrimidine ribonucleotide biosynthesis
VC2119	Hypothetical: conserved
VC2152	Amino acid biosynthesis: aspartate family
VC2272	Hypothetical: conserved
VC2320	DNA metabolism: DNA replication, recombination, and repair
VC2322	DNA metabolism: DNA replication, recombination, and repair
VC2424	Cell envelope: surface structures
VC2425	Cell envelope: surface structures
VC2437	Unknown: general
VC2688	Unknown: general
VC2732	Protein fate: protein and peptide secretion and trafficking
VCA0692	DNA replication, transcription, translation, and cell-wall biosynthesis and pathogenicity (for example, toxins, surface antigens, and adhesins)
VCA0872	Energy metabolism: electron transport
VCA1092	Cellular processes: chemotaxis and motility

on using sequence and expression data of *V. cholerae* are presented here below.

Gene expression data and virulence genes

Initial microarray dataset was constituted of 5,760 ORFs, which was reduced to 5,222 after the control dataset and the ORFs not showing any variation across time points were eliminated. Further filtering based on the cutoffs described above reduced the number of ORFs to 4,169 at 16 time points at the conclusion of filtering stage. ORFs were arranged into clusters based on the gene expression repertoire. All available ORFs could be divided in 500 clusters. Only those clusters were further selected which shared the presence of at least one well-documented virulence gene. A list of 155 virulence genes prepared on the basis of previously published reports or on mining VFDB was used as a reference for analysis of gene expression patterns in the clusters that contained them. K-mean clustering helped identification of those genes which

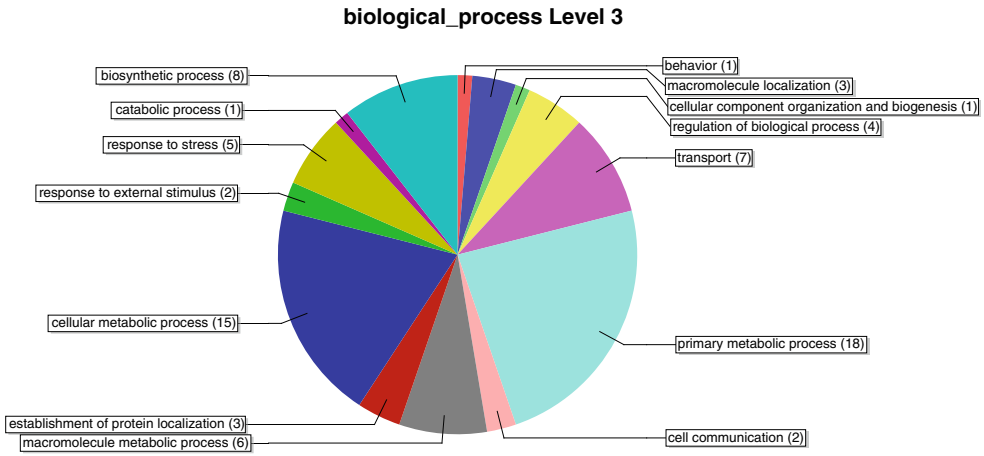

Fig. 1 Summary of the biological processes in which products of the predicted virulence genes are involved in

shared similar type of gene expression pattern in different experimental conditions and were thus considered as probable virulence genes (van-Noort et al. 2003). This approach led to identification of an additional 357 probable virulence genes. Thus, a total of 512 virulence-related genes were shortlisted. Each of the individual genes from these clusters was subjected to blastx against database of essential genes.

Virulence genes as drug targets

Similarity search for virulence related genes in DEG led us to identify 102 of the total 512 also as essential genes for *V. cholerae* which is quite large in number as compared to experimentally reported essential genes (Higgins et al. 1992; Lee et al. 1999; Judson and Mekalanos 2000; Camara et al. 2002; Zhu et al. 2002). These genes were considered as potential drug targets because of their essentiality for the growth and survival of the organism (Roemer et al. 2003).

To predict the role of these 102 virulence genes as drug target, we compared them with human proteome and found that 66 genes among the 102 essential genes showed considerable similarities. Thus, only 36 genes (Table 1), which did not show homology with human genes, could be considered as drug targets (Sakharkar et al. 2004). The 36 genes that were included in the gene ontology analysis using Blast2GO analysis were mapped and annotated to have important functions in the cell and critical locations inside or on the surface of the cell (Overington et al. 2006). More than 50% of the shortlisted genes were found coding for the proteins involved in metabolic processes of biopolymers including nucleic acids and proteins (Fig. 1). Another eight of the gene products were found involved in biosynthetic processes. The rest of the gene products were also found associated with vital metabolic processes of the

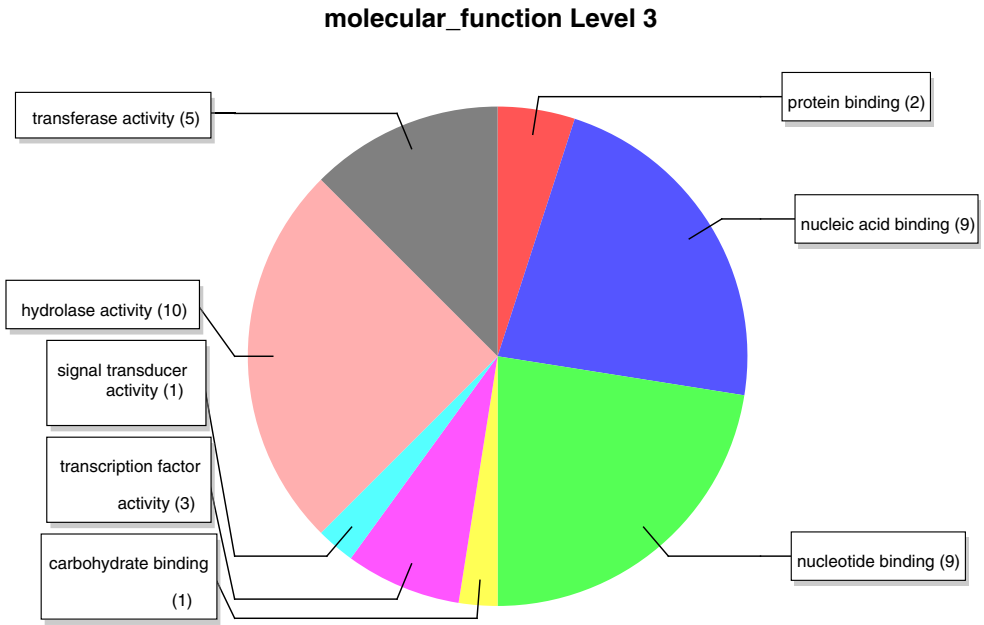

Fig. 2 Molecular functions of the products of the predicted virulence genes

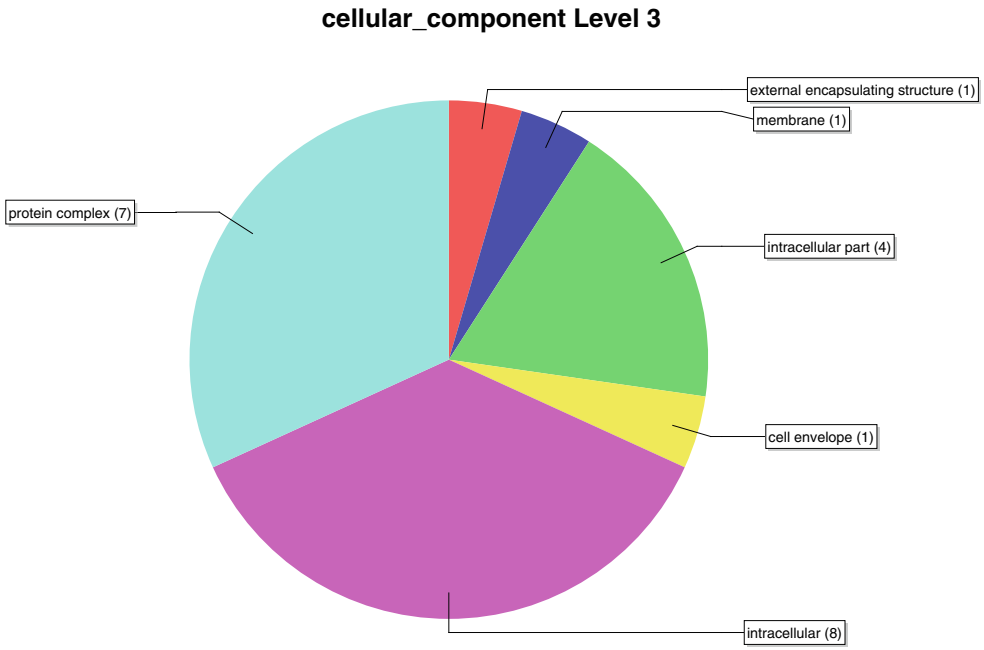

Fig. 3 Summary of the cellular components where products of the predicted virulence genes are localized

cell. Blocking the primary metabolic processes or synthesis of macromolecules has direct implication in discovery of new treatment strategies (Sharma et al. 2008; Sridhar et al. 2007). Furthermore, highest numbers of gene products were found carrying hydrolase activity. In fact, hydrolase proteins and nucleic acid-binding proteins together constituted more than 50% of the present dataset (Fig. 2). Significantly, ten gene products were found associated with nucleotide, nucleobase, and nucleic acid metabolism processes which corresponded to nine nucleotide-binding proteins (Fig. 2). Similarly, most of the gene products were mapped to intracellular locations, to some organelles, or to the enzyme complexes (Fig. 3).

Clearly, all functions played by these genes are very important for the growth and surveillance of *V. cholerae*. The gene products of VC2424, VC2425, and VCA0692 (Table 1) were considered most appropriate drug targets because of their location at cell surface and participation in cell wall synthesis (Hasan et al. 2006; Overington et al. 2006). The latter of these has been implicated in most vital processes of the cell (Table 1) like replication, transcription, cellular defense (cell wall biosynthesis), and even pathogenicity. The other two (VC2424 and VC2425) encode surface antigens and are likely to be involved in invasion of the host and establishment of virulence. All of these as present on cell surface become easy therapeutic targets.

The drug target identification is an important and sensitive first step of the drug discovery process that must need to satisfy various selection criteria to pass for next stage (Lipinski et al. 2001; Hefti 2008). Our results thus provide a starting material for future discovery of drug discovery against *V. cholera*, and we recommend these 36 targets may be experimentally validated. To the best of our knowledge, this is the first report on use of both sequence analysis and gene expression data to identify putative drug targets in a pathogenic species, and both results and methods are likely to find importance among the scientists actively participating in pharmaceutical research.

Conclusion

Various computational methods are available in scientific field for the prediction of virulence genes and drug targets, but all these methods mostly depend on the sequence-based homology or gene expression pattern analysis, and none of them is self-sufficient for such purposes.

In this work, we utilized the information regarding virulence genes and used it for prediction of other probable virulence genes by using gene expression pattern, and then predict their role as drug target using subtractive genomic approach. We found that the combination of these two approaches (gene expression pattern and sequence pattern) provide a very good method to find out the virulence genes and the role of their products as drug target. We tested this approach on the sequence and gene expression data of *V. cholerae* and efficiently shortlisted 36 genes. The number of genes with drug target potential is low, but this set of 36 genes is likely to prove highly potent and convenient targets for newer drugs to be discovered against this pathogen.

Acknowledgments The authors acknowledge the DBT center for bioinformatics facility at Department of Bioscience and Biotechnology, Banasthali University, Banasthali, India for providing essential facilities for completion of this research work. The authors also wish to thank Mr. Manish Roorkiwal, Guru Gobind Singh Indraprastha University, Delhi for critically reviewing the manuscript.

Disclosure statement No competing financial interests exist.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2:e9

Camara M, Hardman A, Williams P, Milton D (2002) Quorum sensing in *Vibrio cholerae*. Nat Genet 32:217–218

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868

Galperin MY, Koonin EV (1999) Searching for drug targets in microbial genomes. Curr Opin Biotechnol 10:571–578

Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinform 9:62

Groth P, Weiss B, Pohlenz HD, Leser U (2008) Mining phenotypes for gene function prediction. BMC Bioinform 9:136

Gray CP, Keck W (1999) Bacterial targets and antibiotics: genome-based drug discovery. Cell Mol Life Sci 56:779–787

Hasan S, Daugelat S, Rao PS, Schreiber M (2006) Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. PLoS Comput Biol 2:e61

Hefti FF (2008) Requirements for a lead compound to become a clinical candidate. BMC Neurosci 9:S7

Higgins DE, Nazareno E, DiRita VJ (1992) The virulence gene activator ToxT from *Vibrio cholerae* is a member of the *AraC* family of transcriptional activators. J Bacteriol 174:6974–6980

Hubble J, Demeter J, Jin H, Mao M, Nitzberg M, Reddy TB, Wymore F, Zachariah ZK, Sherlock G, Ball CA (2009) Implementation of GenePattern within the Stanford Microarray Database. Nucleic Acids Res 37:D898–D901

Judson N, Mekalanos JJ (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. Nat Biotechnol 18:740–745

Knell RJ (2006) New gene, new disease. Heredity 97:315

Kuruvilla FG, Shamji AF, Sternson SM, Hergenrother PJ, Schreiber SL (2002) Dissecting glucose signaling with diversity-oriented synthesis and small-molecule microarrays. Nature 416:653–657

Lee SH, Hava DL, Waldor MK, Camilli A (1999) Regulation and temporal expression patterns of *Vibrio cholerae* virulence genes during infection. Cell 99:625–634

Li Q, Lai L (2007) Prediction of potential drug targets based on simple sequence properties. BMC Bioinform 8:353

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46:3–26

Overington JP, Al-Lazikani B, Hopkins A (2006) How many drug targets are there? Nat Rev 5:993–996

Perumal D, Lim CS, Sakharkar KR, Sakharkar MK (2007) Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. In Silico Biol 7:0032

Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, Martel N, Veronneau S, Lemieux S, Kauffman S, Becker J, Srorms R, Boone C, Bussey H (2003) Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. Mol Microbiol 50:167–181

Sakharkar KR, Sakharkar MK, Chow VTK (2004) A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. In Silico Biol 4:0028

Sakharkar KR, Sakharkar MK, Chow VT (2008) Biocomputational strategies for microbial drug target identification. Meth Mol Med 142:1–9

Sharma V, Gupta P, Dixit A (2008) In silico identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*. In Silico Biol 8:0026

Sridhar P, Song B, Kahveci T, Ranka S (2007) Mining metabolic networks for optimal drug targets. Pac Symp Biocomput 13:291–302

Sur D, Sarkar BL, Manna B, Deen J, Datta S, Nivoqi SK, Gosh AN, Deb A, Kanunqo S, Palit A, Bhattacharya SK (2006) Epidemiological, microbiological & electron microscopic study of a cholera outbreak in a Kolkata slum community. Indian J Med Res 123:31–36

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907–2912

Ulm R, Baumann A, Oravecz A, Mate Z, Adam E, Oakeley EJ, Schafer E, Naqv F (2004) Genome-wide analysis of gene expression reveals function of the bZIP transcription factor HY5 in the UV-B response of Arabidopsis. Proc Natl Acad Sci USA 101:1397–1402

Van-Noort V, Snel B, Huynen MA (2003) Predicting gene function by conserved co-expression. Trends Genet 19:238–242

Yang J, Chen LH, Sun L, Yu J, Jin Q (2008) VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res 36:D539–D542

Zhang R, Lin Y (2009) DEG 50, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res 37:D455–D458

Zhu J, Miller MB, Vance RE, Dziejman M, Bassler BL, Mekalanos JJ (2002) Quorum-sensing regulators control virulence gene expression in *Vibrio cholerae*. Proc Natl Acad Sci USA 99:3129–3134